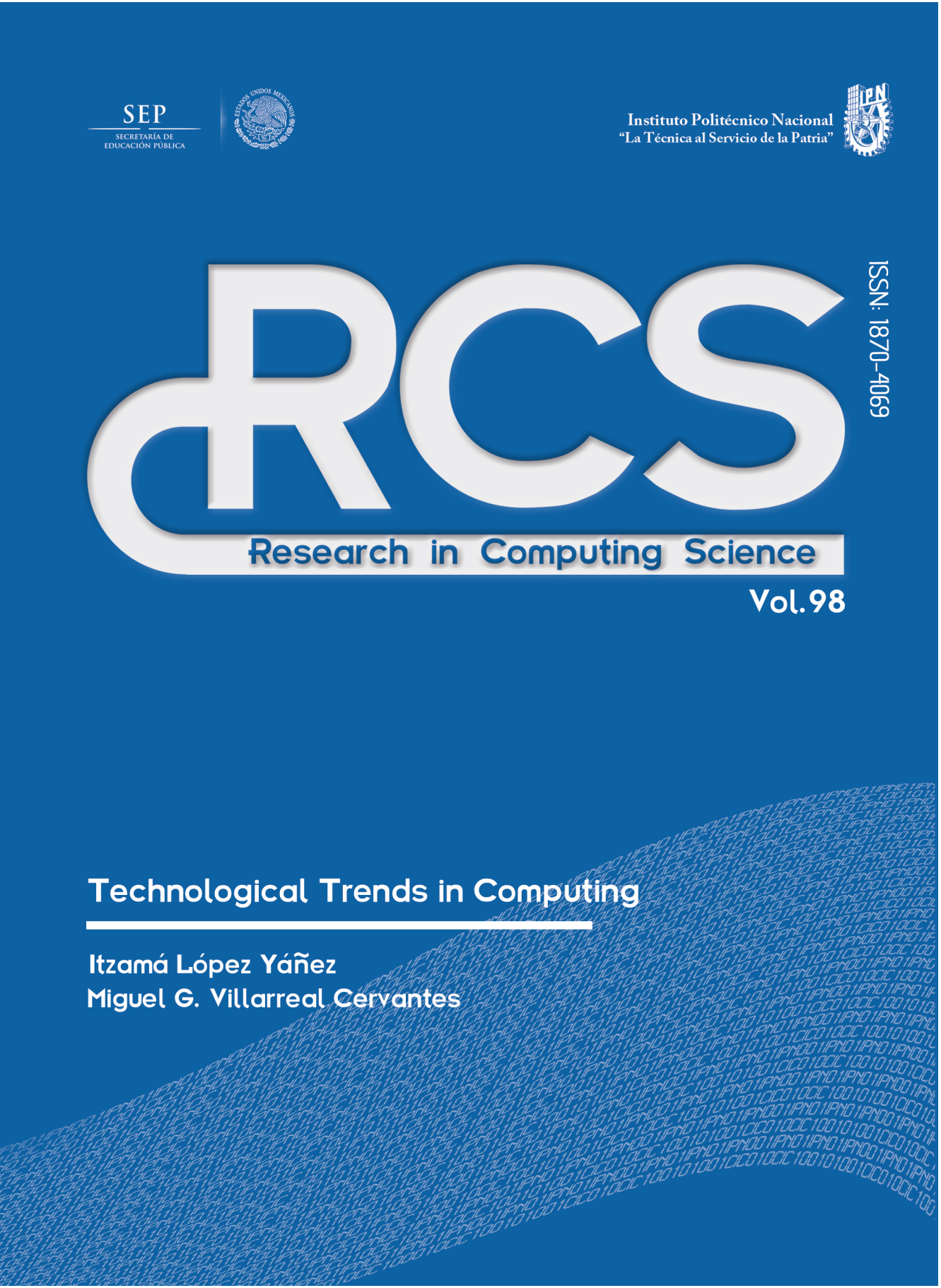
SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

Instituto Politécnico Nacional
"La Técnica al Servicio de la Patria"



ISSN: 1870-4069

Vol. 98

Technological Trends in Computing

Itzamá López Yáñez
Miguel G. Villarreal Cervantes

Technological Trends in Computing

Research in Computing Science

Series Editorial Board

Research in Computing Science es una publicación trimestral, de circulación internacional, editada por el Centro de Investigación en Computación del IPN, para dar a conocer los avances de investigación científica y desarrollo tecnológico de la comunidad científica internacional. **Volumen 98,** octubre 2015. Tiraje: 500 ejemplares. *Certificado de Reserva de Derechos al Uso Exclusivo del Título* No. : 04-2005-121611550100-102, expedido por el Instituto Nacional de Derecho de Autor. *Certificado de Licitud de Título* No. 12897, *Certificado de licitud de Contenido* No. 10470, expedidos por la Comisión Calificadora de Publicaciones y Revistas Ilustradas. El contenido de los artículos es responsabilidad exclusiva de sus respectivos autores. Queda prohibida la reproducción total o parcial, por cualquier medio, sin el permiso expreso del editor, excepto para uso personal o de estudio haciendo cita explícita en la primera página de cada documento. Impreso en la Ciudad de México, en los Talleres Gráficos del IPN – Dirección de Publicaciones, Tres Guerras 27, Centro Histórico, México, D.F. Distribuida por el Centro de Investigación en Computación, Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othón de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738, México, D.F. Tel. 57 29 60 00, ext. 56571.

Editor responsable: *Grigori Sidorov, RFC SIGR651028L69*

Research in Computing Science is published by the Center for Computing Research of IPN. **Volume 98**, October 2015. Printing 500. The authors are responsible for the contents of their articles. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior permission of Center for Computing Research. Printed in Mexico City, in the IPN Graphic Workshop – Publication Office.

Volume 98

Technological Trends in Computing

Itzamá López Yáñez
Miguel Gabriel Villarreal Cervantes (eds.)

Instituto Politécnico Nacional, Centro de Investigación en Computación
México 2015

ISSN: 1870-4069

Instituto Politécnico Nacional (IPN)
Centro de Investigación en Computación (CIC)
Av. Juan de Dios Bátiz s/n esq. M. Othón de Mendizábal
Unidad Profesional "Adolfo López Mateos", Zacatenco
07738, México D.F., México

http://www.ipn.mx
http://www.cic.ipn.mx

Indexed in LATINDEX and Periodica / Indexada en LATINDEX y Periódica

Printing: 500 / Tiraje: 500

Printed in Mexico / Impreso en México

Preface

The appeareance, development and success of computational tools and processes have brought about innovation in manifold topics of science and technology worldwide. Science and technology advances to fulfil mankind needs, have lately been triggered by innovation and development in computing science. The accelerated development pace of computing research has impinged on the increasing variety of topics on computational science.

This special issue of *Research in Computing Science* is titled Technological Trends in Computing. It contains 14 articles that were selected and based on a rigorous review process conducted by members of our Editorial Board. It is noteworthy that the content of this issue is a comprehensive set of results related to research projects and technological development. In this volume we present scientific contributions on topics of interest to the community involved in computing, engineering and related areas.

October, 2015

Itzamá López Yáñez
Miguel G. Villarreal Cervantes

Table of Contents

Feature Selection on Associative Models using Single Board Computer Paradigm

R. Ramírez-Rubio, E. Solórzano-Alor, M. Aldape-Pérez and C. Yáñez-Márquez

Instituto Politécnico Nacional, IPN
Centro de Innovación y Desarrollo Tecnológico en Cómputo, CIDETEC
México D.F., México
rogelioramirezr@hotmail.com; eduardosolorzano22@hotmail.com;
maldape@gmail.com; coryanez@gmail.com

Abstract. In this paper feature selection and associative models are addressed. Classification accuracy is evaluated using six medical datasets, widely used in Machine Learning. In order to obtain the optimal subset of features, an associative model was implemented on a Single Board Computer (SBC).

Keywords: Feature Selection, Single Board Computer, Associative Model.

1 Introduction

Most of the times the initial set of features consists of a large number of potential attributes that constitute an obstacle not only to the accuracy but to the efficiency of algorithms. In high dimensional spaces features often tend to be correlated, in this way a successful subset of features provided to a classifier can increase its accuracy, save computation time, and simplify its results [9]. The design of pattern classifiers has at least three goals: (1) to reduce the cost of extracting features, (2) to improve classification accuracy, and (3) to improve performance reliability [8].

This paper presents *Hybrid Classification and Masking* algorithm (HCM) implemented on a *Single Board Computer* (SBC). The main idea is to compare the performance of HCM with other classification algorithms to determine a level of competitiveness for different applications. The main idea is that a strong classifier could be obtained when a hybrid associative memory is used during the learning phase thus additional advantages appear. Since one of the associative memories properties is to establish links between patterns and classes to which they belong on a one-shot basis, iterative complexity is eliminated. This means that one and only one classifier is computed during the whole classification and masking process.

Another clear advantage during the second phase (recalling) arises when any kind of prior knowledge is not needed in the sense that no distributional

or functional assumptions are considered. Moreover the optimal mask search algorithm is applied only to those patterns that were previously considered during the first phase (learning), which means that no additional patterns are required to increase classifier accuracy.

2 Main Concepts

An associative memory M is a system that relates input patterns and output patterns as follows: $x \longrightarrow \boxed{\mathbf{M}} \longrightarrow y$ with x and y, respectively, the input and output pattern vectors. Each input vector forms an association with its corresponding output vector. For each k integer and positive, the corresponding association will be denoted as: (x^k, y^k). Associative memory M is represented by a matrix whose ij-th component is m_{ij} [10]. Memory M is generated from an *a priori* finite set of known associations, called the fundamental set of associations. If μ is an index, the fundamental set is represented as: $\{(x^\mu, y^\mu) \mid \mu = 1, 2, ..., p\}$ with p as the cardinality of the set. The patterns that form the fundamental set are called fundamental patterns. If it holds that $x^\mu = y^\mu \ \forall \mu \in \{1, 2, ..., p\}$ M is auto-associative, otherwise it is heteroassociative; in this case it is possible to establish that $\exists \mu \in \{1, 2, ..., p\}$ for which $x^\mu \neq y^\mu$. If we consider the fundamental set of patterns $\{(x^\mu, y^\mu) \mid \mu = 1, 2, ..., p\}$where n and m are the dimensions of the input patterns and output patterns, respectively, it is said that $x^\mu \in A^n$, $A = \{0, 1\}$ and $y^\mu \in A^m$. Then the j-th component of an input pattern is $x_j^\mu \in A$. Analogously, the j-th component of an output pattern is represented as $y_j^\mu \in A$.

2.1 Lernmatrix

Lernmatrix is a heteroassociative memory that can easily work as a binary pattern classifier if output patterns are appropriately chosen [13]. It accepts binary patterns suchlike $\mathbf{x}^\mu \in A^n$, $A = \{0, 1\}$ as input and returns binary patterns suchlike $\mathbf{y}^\mu \in A^m$ as output; it is worth pointing out that there are m different classes, each one coded by a simple rule: class $k \in \{1, 2, ..., m\}$ will be represented by a column vector which components will be assigned by $y_k^\mu = 1$, so $y_j^\mu = 0$ for $j = 1, 2..., k-1, k+1, ...m$.

The following matrix will keep the pattern association values after the Learning Phase for the Steinbuch's *Lernmatrix* is done:

$$\begin{array}{|c|c|c|c|c|c|c|} \hline & x_1^\mu & x_2^\mu & \cdots & x_j^\mu & \cdots & x_n^\mu \\ \hline y_1^\mu & m_{11} & m_{12} & \cdots & m_{1j} & \cdots & m_{1n} \\ \hline \vdots & \vdots & \vdots & & \vdots & & \vdots \\ \hline y_i^\mu & m_{i1} & m_{i2} & \cdots & m_{ij} & \cdots & m_{in} \\ \hline \vdots & \vdots & \vdots & & \vdots & & \vdots \\ \hline y_m^\mu & m_{m1} & m_{m2} & \cdots & m_{mj} & \cdots & m_{mn} \\ \hline \end{array} \quad (1)$$

Each one of the m_{ij} components of $\mathbf{M}$ is initialized with zero and will be modified by the following rule: $m_{ij} = m_{ij} + \Delta m_{ij}$ where:

$$\Delta m_{ij} = \begin{cases} +\varepsilon & if \quad y_i^\mu = 1 = x_j^\mu \\ -\varepsilon & if \quad y_i^\mu = 1 \quad and \quad x_j^\mu = 0 \\ 0 & otherwise \end{cases} \tag{2}$$

and ε a positive constant, previously chosen.

The Recalling Phase for the Steinbuch's *Lernmatrix* consists of finding the class which an input pattern $\mathbf{x}^\omega \in A^n$ belongs to. Finding the class means getting $\mathbf{y}^\omega \in A^m$ that corresponds to $\mathbf{x}^\omega$; accordingly to the construction method of all $\mathbf{y}^\mu$, the class should be obtained without ambiguity. The i-th component of y_i^ω is obtained according to the following rule, where $\vee$ is the *maximum* operator:

$$y_i^\omega = \begin{cases} 1 & \text{if} \quad \sum_{j=1}^n m_{ij}.x_j^\omega = \bigvee_{h=1}^m \left[\sum_{j=1}^n m_{hj}.x_j^\omega\right] \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

2.2 Linear Associator

Lets consider the fundamental set as $\{(\mathbf{x}^\mu, \mathbf{y}^\mu) \mid \mu = 1, 2, ..., p\}$ with

$$\mathbf{x}^\mu = \begin{pmatrix} x_1^\mu \\ x_2^\mu \\ \vdots \\ x_n^\mu \end{pmatrix} \in A^n \quad \text{y} \quad \mathbf{y}^\mu = \begin{pmatrix} y_1^\mu \\ y_2^\mu \\ \vdots \\ y_m^\mu \end{pmatrix} \in A^m$$

The *Learning Phase* is done in two stages.

1. Consider each one of the p associations $(\mathbf{x}^\mu, \mathbf{y}^\mu)$, so an $m \times n$ matrix is obtained by $\mathbf{y}^\mu \cdot (\mathbf{x}^\mu)^t$

$$\mathbf{y}^\mu \cdot (\mathbf{x}^\mu)^t = \begin{pmatrix} y_1^\mu \\ y_2^\mu \\ \vdots \\ y_m^\mu \end{pmatrix} \cdot (x_1^\mu, x_2^\mu, ..., x_n^\mu) = \begin{pmatrix} y_1^\mu x_1^\mu & \cdots & y_1^\mu x_j^\mu & \cdots & y_1^\mu x_n^\mu \\ y_2^\mu x_1^\mu & \cdots & y_2^\mu x_j^\mu & \cdots & y_2^\mu x_n^\mu \\ \vdots & \vdots & \vdots & & \vdots \\ y_i^\mu x_1^\mu & \cdots & y_i^\mu x_j^\mu & \cdots & y_i^\mu x_n^\mu \\ \vdots & \vdots & \vdots & & \vdots \\ y_m^\mu x_1^\mu & \cdots & y_m^\mu x_j^\mu & \cdots & y_m^\mu x_n^\mu \end{pmatrix} \tag{4}$$

2. $\mathbf{M}$ memory is obtained by adding all the p matrices

$$\mathbf{M} = \sum_{\mu=1}^{p} \mathbf{y}^\mu \cdot (\mathbf{x}^\mu)^t = [m_{ij}]_{m \times n} \tag{5}$$

in this way the ij-th component of $\mathbf{M}$ memory is expressed as:

$$m_{ij} = \sum_{\mu=1}^{p} y_i^{\mu} x_j^{\mu} \tag{6}$$

The *Recalling Phase* for the *Linear Associator* is done by operating the $\mathbf{M}$ memory with an input pattern $\mathbf{x}^{\omega}$, where $\omega \in \{1, 2, ..., p\}$; operate $\mathbf{M} \cdot \mathbf{x}^{\omega}$ as follows:

$$\mathbf{M} \cdot \mathbf{x}^{\omega} = \left[\sum_{\mu=1}^{p} \mathbf{y}^{\mu} \cdot \left(\mathbf{x}^{\mu}\right)^{t} \right] \cdot \mathbf{x}^{\omega} \tag{7}$$

Lets expand 7, we obtain:

$$\mathbf{M} \cdot \mathbf{x}^{\omega} = \mathbf{y}^{\omega} \cdot \left[\left(\mathbf{x}^{\omega}\right)^{t} \cdot \mathbf{x}^{\omega} \right] + \sum_{\mu \neq \omega} \mathbf{y}^{\mu} \cdot \left[\left(\mathbf{x}^{\mu}\right)^{t} \cdot \mathbf{x}^{\omega} \right] \tag{8}$$

Expression 8 lets us know about which restrictions have to be observed thus perfect recalling is achieved. These restrictions are:

a) $\left[\left(\mathbf{x}^{\omega}\right)^{t} \cdot \mathbf{x}^{\omega} \right] = 1$

b) $\left[\left(\mathbf{x}^{\mu}\right)^{t} \cdot \mathbf{x}^{\omega} \right] = 0$ whenever $\mu \neq \omega$

Given an arbitrary chosen index ω , $\forall \omega \in \{1, 2, ..., p\}$, means that input pattern $\mathbf{x}^{\mu}$ should be orthonormal. This restriction is expressed as:

$$\left(\mathbf{x}^{\mu}\right)^{t} \cdot \mathbf{x}^{\omega} = \begin{cases} 1 & \text{if} \quad \mu = \omega \\ 0 & \text{if} \quad \mu \neq \omega \end{cases} \tag{9}$$

If condition 9 is met, then a perfect recalling is expected. So 8 is expressed as:

$$\mathbf{M} \cdot \mathbf{x}^{\omega} = \mathbf{y}^{\omega}.$$

Nevertheless if orthonormality condition is not met, two situations appear:

· Factor $\left[\left(\mathbf{x}^{\omega}\right)^{t} \cdot \mathbf{x}^{\omega} \right]$ is not equal to 1

· Term $\sum_{\mu \neq \omega} \mathbf{y}^{\mu} \cdot \left[\left(\mathbf{x}^{\mu}\right)^{t} \cdot \mathbf{x}^{\omega} \right]$ is not equal to 0

This term is known as *cross-talk*, it represents some kind of noise that comes from input patterns interaction. As a consequence perfect recalling is not achieved, except in those cases where the number of stored patterns is rather small compared to n . Some researchers' results confirm that this number should be between $0.1n$ and $0.2n$ [2],[5], [11].

2.3 Hybrid Associative Classifier with Translation

Hybrid Associative Classifier with Translation (CHAT for its acronym in Spanish) combines *Linear Associator* learning phase and *Lernmatrix* recalling phase [6]. The algorithm of the *Hybrid Associative Classifier with Translation* [12] model is as follows:

1. Let n be the dimension of each input pattern in the fundamental set, grouped in m different classes.
2. Each one of the input patterns belongs to a k class, $k \in \{1, 2, ..., m\}$, represented by a column vector which components will be assigned by $y_k^\mu = 1$, so $y_j^\mu = 0$ for $j = 1, 2..., k-1, k+1, ...m$.
3. The learning phase is carried out as a *Linear Associator* , according to expression 4, 5 and 6.
4. The recalling phase is carried out as a *Lernmatrix* according to expression 3.

2.4 Hybrid Classification and Masking Approach

Hybrid Classification and Masking technique (HCM) is presented as a new feature selection approach to provide a mask that identifies the optimal subset of features without having to compute a new classifier at each step. This method allows us to identify irrelevant or redundant features for classification purposes [1]. In order to explain how optimal mask is found, some definitions are required.

Definition 1. *Let* $\mathbf{f}$ *be the number of features from the original set of data.*

Definition 2. *Let* $\mathbf{r}$ *be an index where* $r \in \{1, 2, ..., (2^f - 1)\}$

Definition 3. *Let* $\mathbf{e}^r$ *be a masking vector of size* n *represented as:*

$$\mathbf{e}^r = \begin{pmatrix} e_1^r \\ e_2^r \\ \vdots \\ e_n^r \end{pmatrix} \in B^n \tag{10}$$

where $B = \{0, 1\}$

Definition 4. *Let* $\dashv$ *be a new operation called* $IntToVector$ *which takes* $r \in \{1, 2, ..., (2^f - 1)\}$ *and returns a column vector* $\mathbf{e}^r$ *with* r *value expressed in its binary form. From a register transfer level perspective (RTL) this can be expressed as* $bin(r) \rightarrow [\mathbf{e}^r]$ *. For example: If* $r = 11$ *then* $\dashv \mathbf{e}^r$ *returns a column vector with* r *value in its binary form so the obtained vector is:*

$$\mathbf{e}^{11} = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \end{pmatrix}$$

where e_n^r is the Least Significant Bit (LSB)

Definition 5. *Let $\|$ be a new operation called $MagVector$ which takes a column vector $\mathbf{e}^r$ of size n and returns an integer and positive value according to the following rule:*

$$\| \mathbf{e}^r = \sum_{j=1}^{n} \left(e_j^r \wedge 1 \right) \tag{11}$$

Where $\wedge$ is the logical AND operator.

Another relevant thing to mention is that the *Recalling Phase* is dramatically different from the previous models; it is carried out by the following rule:

$$y_i^\mu = \begin{cases} 1 & \text{if} \quad \sum_{j=1}^{n} m_{ij} . \left(x_j^\mu . e_j^r \right) = \bigvee_{h=1}^{m} \left[\sum_{j=1}^{n} m_{hj} . \left(x_j^\mu . e_j^r \right) \right] \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

$$\text{where } \mu \in \{1, 2, ..., p\} \text{ and } r \in \{1, 2, ..., (2^f - 1)\}$$

It is said that the *Recalling Phase* is dramatically different from the previous models because a masking vector $\mathbf{e}^r$ of size n masks each input vector x^μ of size n. This is where the masking technique comes into view. Using the previous definitions and the clear advantages that inherit from the *Hybrid Associative Classifier* model, it is possible to enunciate The HCM algorithm.

1. Create a classifier using (4), (5) and (6).
2. Use the $IntToVector$ operator to get the r-th masking vector as in (10)
3. The recalling phase is carried out according to expression (12) so an r-th accuracy parameter is obtained
4. Store both parameters (the r-th accuracy parameter and the r-th masking vector) so feature selection can be evaluated in step (6)
5. Compare the r-th accuracy parameter with the $(r-1)$-th accuracy parameter. The best accuracy value is stored thus accuracy improvements are achieved with each iteration
6. The same applies to the r-th masking vector. Feature selection can be evaluated using expression (11). So the smaller this number is, a better mask is obtained.
7. The new subset of features is obtained by a mask value represented by a column vector, where accuracy and feature selection are optimal

2.5 Single Board Computer

The *SBC paradigm* is very effective for specific applications, for which it was designed [4], this paradigm can be used as a tool for tasks requiring precision control and data acquisition, which is consider scientific applications [7] and industrial applications such as remote monitoring and gas tank filling control [3]. The Parallella board is a high performance computing device, which can be used as a standalone computer (SBC) or as an embedded system. The Parallella platform includes a low-power consumption pro dual core ARM A9 processor , which is able to work with different Linux distributions, it makes it attractive by the versatility of having many work environments and give some facilities to the end users. Also this board contain a 1 GB DDR3 RAM and a slot for micro SD memory. One of the applications that can be given to Parallella board is image processing and analysis, a clear example is face detection, this issue has been extensively studied with low-cost computing and innovative algorithms [16].

3 Datasets

The following datasets were used because of their confidence and wide study in this area. These datasets have been used to probe and validate new classifiers as well as classical classification techniques.

Breast Cancer: This database was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg and was donated by Olvi Mangasarian. This dataset contains periodical samples of clinical cases. Breast cancer dataset consists of 699 instances belonging to two different classes (458 benign cases, 241 malign cases). Each instance consists of 10 attributes, including the class attribute.

Heart Disease: This database comes from the Cleveland Clinic Foundation and was supplied by Robert Detrano, M.D., Ph.D. of the V.A. Medical Center, Long Beach, CA. The purpose of the dataset is to predict the presence or absence of heart disease given the results of various medical tests carried out on a patient. This dataset consists of 270 instances belonging to two different classes: presence and absence (of heart-disease). Each instance consists of 14 attributes, including the class attribute.

Hepatitis Disease: This dataset was donated by the Jozef Stefan Institute, former Yugoslavia, now Slovenia. The purpose of the dataset is to predict the presence or absence of hepatitis disease in a patient. Hepatitis disease dataset consists of 155 instances belonging to two different classes (32 die cases, 123 live cases). Each instance consists of 20 attributes, 13 binary, 6 attributes with discrete values and a class label.

Liver Disorders: This database was created by BUPA Medical Research Ltd and was donated by Richard S. Forsyth. This dataset contains cases from a study that was conducted on liver disorders that might arise from excessive

alcohol consumption. Liver disorders dataset consists of 345 instances belonging to two different classes. Each instance consists of 7 attributes, including the class attribute.

Parkinson Disease: This database was created by Max Little of the University of Oxford in collaboration with the National Centre for Voice and Speech, Denver, Colorado, who recorded the speech signals. The original study published the feature extraction methods for general voice disorders. This dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). Each column in the table is a particular voice measure, and each row corresponds one of 195 voice recording from these individuals ("name" column). The main aim of the dataset is to discriminate healthy people from those with PD, according to "status" column which is set to 0 for healthy and 1 for PD.

Pima Indians Diabetes: This database was originally owned by the National Institute of Diabetes and Digestive and Kidney Diseases, U.S. This dataset contains cases from a study that was conducted on female patients at least 21 years old of Pima Indian heritage. This dataset consists of 768 instances belonging to two different classes (500 the patient tested positive for diabetes cases, 268 the patient tested negative for diabetes cases). Each instance consists of 9 attributes, including the class attribute.

4 Experimental Phase

The experimental phase was conducted with twenty - one algorithms all of them are different pattern classification methods, executed in WEKA enviroment [15], except for the proposed algorithm. Six medical datasets obtained from the UCI: Machine Learning Repository [14] were used in order to test the efficiency (the percentage of correct classification) and average of each one of them. In addition, algorithms proposed were compared, to view the execution times as well as the improved of efficiency, the algorithms were executed in a Parallella Board.

Table 1 shows the classification accuracy of the algorithms.

Table 3 shows the execution time comparison of CHAT before and after feature selection.

As can be seen in Table 4, classification accuracy and execution time were improved in CHAT, after the feature selection.

Table 1: Efficiency (%) of classification of each algorithm

Algorithm \ Dataset	Breast	Heart	Hepatitis	Liver	Parkinson	Pima	Average
AdaBoostM1	95.60	81.85	60.00	63.73	85.12	75.65	76.99
Bagging	95.90	81.11	68.38	70.43	87.69	74.73	79.71
BayesNet	97.36	**83.70**	70.96	63.18	80.00	75.78	78.50
Dagging	96.92	82.96	63.87	57.68	85.12	74.21	76.79
DecisionTable	95.75	82.96	**73.54**	63.18	83.58	74.86	78.98
DTNB	97.07	80.74	69.67	63.18	85.12	74.86	78.44
FT	97.21	80.37	69.03	**71.59**	84.61	76.69	79.92
LMT	96.19	**83.70**	66.45	69.85	86.15	77.08	79.90
Logistic	96.77	82.96	69.67	68.40	86.66	**77.73**	**80.36**
MultiClassClassifier	96.77	82.96	69.67	68.40	86.66	**77.73**	**80.36**
NaiveBayes	96.19	82.96	71.61	55.94	69.23	75.65	75.26
NaiveBayesSimple	96.33	**83.70**	70.96	55.36	69.23	75.26	75.14
NveBayesUpdateable	96.19	82.96	71.61	55.94	69.23	75.65	75.26
RandomCommittee	96.63	82.22	62.58	68.11	**90.76**	75.39	79.28
RandomForest	96.48	82.59	60.64	67.82	**90.76**	74.60	78.81
RandomSubSpace	95.90	81.11	67.09	68.69	88.7	73.30	79.13
RBFNetwork	96.19	81.85	71.61	64.63	84.10	73.69	78.68
RotationForest	**97.51**	79.62	64.51	70.72	90.25	76.56	79.86
SimpleLogistic	96.48	**83.70**	65.16	69.27	84.61	77.34	79.43
SMO	97.07	82.96	69.67	57.97	87.17	76.82	78.61
CHAT	**97.51**	64.07	66.45	55.36	68.72	61.45	68.93

Table 2: Classification efficiency (%) of CHAT and CHAT with Feature Selection

Algorithm \ Dataset	Breast	Heart	Hepatitis	Liver	Parkinson	Pima	Average
CHAT	97.51	64.07	66.45	55.36	68.72	61.45	68.93
CHAT(FS)	97.80	83.70	85.16	65.50	70.76	70.96	**78.98**

Table 3: Classification Time (Seconds) of the complete dataset using CHAT and CHAT with Feature Selection

Algorithm \ Dataset	Breast	Heart	Hepatitis	Liver	Parkinson	Pima	Average
CHAT	0.006673	0.003724	0.002654	0.002747	0.003317	0.007171	0.004381
CHAT(FS)	0.005364	0.001954	0.000793	0.002120	0.001641	0.005143	**0.002835**

Table 4: Improved classification accuracy and time saved (%) by using feature selection in CHAT compared with original features

Algorithm \ Dataset	Breast	Heart	Hepatitis	Liver	Parkinson	Pima	Average
Time	19.61	47.52	70.12	22.82	50.52	28.28	**39.81**
Classification	0.2928	19.62	18.70	10.14	2.05	9.50	**10.05**

5 Results Analysis and Conclusion

As shown in Table 1 the associative model CHAT shows competitive an performance compared with other classification algorithms, which makes it an option for classification. When added the feature selection technique, it is observed that the CHAT improved considerably in classification as shown in Table 2, also the classification obtained in less time, as it is shown in Table 3. Table 4 shows CHAT with feature selection, and improve the CHAT efficiency in time and classification accuracy.

Now, according with the previous work we can conclude that as shown in Table 1, CHAT proves competitive performance against other classification algorithms present in the state of the art either individually or averaged.

With respect to Table 2, it can be concluded that classification efficiency has undergone CHAT feature selection to the data set used, considerably improves from 68.93% to 78.98% on average.

In Table 3, the classification time of complete sets are shown, which in principle are low. It is worth noting that classification time for each instance not only is good but in average is acceptable considering that datasets that were used in the experimental phase range from nine to twenty-two features (excluding the class label).

Finally in Table 4, the time savings are observed for each database and the average of them is 39.81%. Likewise classification improvement of each database is shown and the average of them is 10.05%.

It is noteworthy that feature selection has its respective cost, but it must be paid only once. Moreover this is compensated with subsequent classifications as more reliable and less time consuming results are obtained. These advantages will be used in each and every one of the instances to be classified in the future.

Acknowledgments The authors of the present paper would like to thank the following institutions for their economical support to develop this work: Science and Technology National Council of Mexico (CONACYT), SNI, National Polytechnic Institute of Mexico (COFAA, SIP, CIDETEC, and CIC).

References

1. M. Aldape-Perez, C. Yanez-Marquez, and L.O. Lopez Leyva. Feature selection using a hybrid associative classifier with masking techniques. In *Artificial Intelligence, 2006. MICAI '06. Fifth Mexican International Conference on*, pages 151–160, Nov 2006.
2. James A. Anderson and Edward Rosenfeld, editors. *Neurocomputing: Foundations of Research*. MIT Press, Cambridge, MA, USA, 1988.
3. L. Benussi, M. Bertani, S. Bianco, F.L. Fabbri, P. Gianotti, V. Lucherini, E. Pace, N. Qaiser, and S. Sarwar. Design and implementation of an automatic gas-cylinder inversion system based on an embedded computer. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 461(13):98 – 99, 2001. 8th Pisa Meeting on Advanced Detectors.
4. Paul A. Braun, Randy H. Moss, and Paul D. Stigall. The single board computer: A practical alternative to a custom designed microsystem. *Computers & Electrical Engineering*, 12(34):109 – 118, 1986.
5. Mohamad H. Hassoun. *Fundamentals of Artificial Neural Networks*. MIT Press, Cambridge, MA, USA, 1st edition, 1995.
6. G. Dreyfus I.G.L. Personnaz. Information storage and retrieval in spin glass like neural networks. *Journal of Physical Letters*, 46:359 – 365, 1985.
7. Haidong Kim and Matthew J. Zabik. Design and evaluation of a versatile single board computer for embedded applications in scientific instrumentation. *Computers & Chemistry*, 16(3):261 – 263, 1992.
8. Mineichi Kudo and Jack Sklansky. Comparison of algorithms that select features for pattern classifiers. *Pattern Recognition*, 33(1):25 – 41, 2000.
9. Mark Last, Abraham Kandel, and Oded Maimon. Information-theoretic algorithm for feature selection. *Pattern Recognition Letters*, 22(67):799 – 811, 2001.
10. Gunther Palm, Friedhelm Schwenker, Friedrich T Sommer, and Alfred Strey. Neural associative memories. *Associative processing and processors*, pages 307–326, 1997.
11. G.X. Ritter, P. Sussner, and J.L. Diza-de Leon. Morphological associative memories. *Neural Networks, IEEE Transactions on*, 9(2):281–293, Mar 1998.
12. R. Santiago-Montero. Hybrid Associative Classifier Based on Lernmatrix and Linear Associator (In Spanish. Master's thesis, Center for Computing Research, México, 2003.
13. K. Steinbuch. Die lernmatrix. *Kybernetik*, 1(1):36 – 45, 1961.
14. UCI. Machine learning repository. Accessed on 16-05-2015 a url-http://archive.ics.uci.edu/ml/, 2007.
15. Brian Hardyment Liang Liu Vladimir Petko Duncan Willcock. Machine learning group at the university of waikato. Accessed on 16-05-2015 a url-http://www.cs.waikato.ac.nz/ml/weka/downloading.html, 1993.
16. Ming Yang, James Crenshaw, Bruce Augustine, Russell Mareachen, and Ying Wu. Adaboost-based face detection for embedded systems. *Computer Vision and Image Understanding*, 114(11):1116 – 1125, 2010. Special issue on Embedded Vision.

A Business Intelligence Model to support the engineering student life cycle analysis Recruitment, Retention-performance, and Graduate process

A. Santoyo-Sanchez[1], José Luis Reyna-Gascón[2],
L. I. Aguirre-Salas[3], Griselda Pérez-Torres[1], and C. A. De Jesús-Velásquez[4]
[1]Department of Computing, Universidad de Guadalajara – CUCEI, Guadalajara, Jalisco, México
[2]IT BI & Analytics department, Tlaquepaque, Freescale Semiconductor, Tlaquepaque, Jalisco, México
[3]Department of Engineering, Universidad de Guadalajara – CUCSUR, Autlán de Navarro, Jalisco, México
[4]Compatibility Validation, Intel Tecnología de México S.A., Tlaquepaque, Jalisco, México
alejandra.santoyo@cucei.udg.mx; jose.reyna@freescale.com; laguirre@cucsur.udg.mx

Abstract. Taking smart decisions during the proper time as well as respond with flexibility to any customer demand, market opportunity or threat, through Business Intelligence (BI), causes that the increase of designing multidimensional solutions become a great solution and alternative to support smart decisions in an organization. In this paper, we present an alternative approach of analyzing the life cycle of the engineering students during their stay at CUCEI, which is a thematic center (campus) of the University of Guadalajara in Mexico dedicated to higher education (Bachelor, Master and PhD Degree). This campus has one of the largest populations within the University of Guadalajara, serving more than 13,000 students [3]. Meaning a way to understand since the student is recruited, and then lives its studies looking to get the expected grades and be part of the graduate's pool; the proposed multidimensional model tries to cover this life cycle.

Keywords: Business Intelligence, Multidimensional Modeling, Students life cycle

1 Introduction

This paper presents a brief overview of how to model the student's life cycle when joining college. This paper intends to be a practical case for the public University of Guadalajara, being focused on one of their campuses, CUCEI which handles all the engineering careers.

The end users typically collects many reports stored in different locations in order to place them in a single repository, sometimes a spreadsheet, and apply their own calculations and also their own charts and metrics. The idea of modeling the business using multidimensional methodologies empower the end user to not waste time on gathering information from many types of sources. The work of the Business Intelligence Analyst is to understand the business and its needs in order to create a semantic layer, from the transactional and operative data that is used to track the day

to day student activities, which translates all these business requirements into a main repository that contains the Business measures and metrics to modeled into a multidimensional environment.

2 Background

The University of Guadalajara is a set of University Network in Jalisco, Mexico, with an historical tradition of more than 2 centuries. It has more than 240,000 students and has presence in all regions of the state, being the most important institution of higher education in western Mexico and the second of the country. Its high school education system covers 101 cities around the state and its higher education (college) is divided in all regions of the state thru 6 main campuses in the metropolitan area and 9 regional campuses around the state [2].

During 2012-2013 47 out of 100 college students studied at University of Guadalajara. This situation has been the trigger for social mobility in all regions of the state, stimulating the local economy, sparking scientific and cultural activities, and moreover, linking every day of hundreds of researches and teachers solving problems of their communities [2].

The main idea of this paper is to obtain a data analysis model to support smart decisions and provide a big picture of the situation, to the key users, that predominates during the students life cycle, which is when the student is recruited, passes an exam to be part of the University, lives its studies looking to get the expected grades and be part of the graduates pool, but turns out that during that period, some students decides to leave their studies due to some reasons. The objective of the Multidimensional model developed in this paper is to support predicting and providing student's situation snapshots during their studies in order to act proactively on students that might give up their career path.

3 Basic Concepts

This section presents the basic concepts related to BI used in this paper. For more details, consult [4-7].

Definition 1. *Business Intelligence* (BI) is the art and science of preparing companies for the future by way of systematic knowledge management process. It is creating knowledge from openly available information by use of a systematic process involving planning, collection, analysis, communication and management, which result in decision-maker action [8].

According to [9] the above definition emphasizes the following: (1) Creation of knowledge from business and project information - availability of information is important in knowledge creation, which will be intelligently used in project management decision making process. (2) The useful application of created knowledge results in better project management decisions. More specifically, BI solutions combine data gathering, data storage, and knowledge management with analytical tools to present complex and competitive information to planners and decision makers

[10]. According to [9] a BI solution has four major components, namely, a Data warehouse, Business Analytics (BI technologies), Business Performance Management (BPM) and a user interface. BI solution proposed is based in the follows components.

First, to providing an integrated historical data, data mart was the reference tool for storage of data to support and to feed the BI system.

Definition 2. *Data warehouse* is a database containing the consolidated data from all available resources, optimized for reporting, analysis and archiving. The Data warehouse integrates and stores data from both internal and external sources [12].

Definition 3. *Data Mart* is basically another form of a Data Warehouse, but it is a subset of information customized based on the business needs, with data summarized, aggregations, metadata, and specific features needed for the analysis [13].

Second, the business analysis was made using Online Analytical Processing (OLAP) systems. These tools are considered as a type of software aiming at fast exploring and analysis of data through a multidimensional approach with several levels of aggregation. The online analysis is a subject oriented, non-volatile and time variant tool with an optimized approach for decision support. The idea of using OLAP solutions is to have it as a long term information asset. This tool does not substitute the transactional and operational systems (OLTP); it is intended to be an additional solution that contributes the decisions support. The OLTP are the source data where OLAP will consult.

Definition 4. *OLAP system* is delivered by Microsoft as part of the Microsoft Dynamics AX solution. The OLAP system collects the performance information and generates reports on measurements. This kind of system is used generally in a business environment as an executive information system detailing the actual states of the subject to be analyzed [14].

Third, BI tools are used to determine customer's current business value (or profile). Then, "BI designer" will apply their knowledge and expertise to offer the best possible option to this customer, while taking into account organizational rules and policies, which is the case of the BI solution proposed. Current BPM comprises modeling, analyzing and monitoring of business processes.

Definition 5. *BPM system* is a system for sensing environmental stimulus, interpreting perceived data, adjudicating the data to be business situations, and making decisions about how to respond the situations [16].

According with [16] in general, there are five representative categories of services in a BPM system: Sense, Detect, Analyze, Decide and Execute [16].

1. "***Sense***" is the stage when a BPM system interacts with business solutions and provides data extraction, transformation, and loading capabilities for the sake of preparing qualified data (metrics) that is to be further monitored and analyzed.

2. "***Detect***" is the stage of detecting business situations and/or exception occurring in the business solutions. An example of situation detection could be lower-than-expected cycle time performance in certain manufacturing process.

3. "***Analyze***" is the stage when a BPM system performs business analytics such as risk-based analysis of resolving business exceptions. The output of this stage often comes with recommendation of potential resolutions to decision makers.
4. "***Decide***" is the stage when a decision maker will make decision about what to respond to business situations. A decision maker can be either human or software agent.
5. "***Execute***" is the stage when a BPM system carries out actions for the purpose of enforcing the decisions made by decision makers. Actions can be of many forms such as sending email or as complicated as a sub-process invocation.

The BI solution proposed in this paper covers the set of categories described in Definition 5, which will be explained in the next chapters, but the idea of contributing with a BI solution has to consider stages since the data extraction, transformation and loading till the analytic solution which will become the tool where decisions makes will consult and trust.

Finally, the interface of the final solution is presented using ad-hoc tools that will allow the user to create their own dashboards and pivots based on a single and unique repository, which will provide accurate information. The user will love to validate and play with the data generated, more over that this information will not require dealing with multiple data sources, the OLAP Data base will provide that unique repository that the user will browse and query. The interface used in this project was Microsoft Power Pivot, a BI tool for Microsoft.

4 Methodology

This section presents our methodology to develop the Business Intelligence model. This consists in two main aspects: Problem definition, source data set preparation and exploring; all this in order to guarantee quality for the BI solution developed, which must be included the follows.

1) Communication. The objective is to involve the client into the project development, generating trust and letting it know any concern or issue during the BI project development. All communication is intended to be in writing means using electronic and physical documents when requested, both by project managers and key users. Therefore, any verbal agreement was reinforced by a written document prepared by agreement of all parties. The meetings always ended to send a draft of the agreement and/or revised conclusions and commitments.

2) Formalization. Models formalize the representation and syntax of the things being modelled. They provide commonality where each component of the model has one and only one meaning.
3) Abstraction. Models allow us to work at multiple levels of abstraction, ranging from very general levels to highly specific. The modelling semantics help to ensure across the levels.

4) Prototypes control. This is in order to re-use any prototype in case a new BI Model needs to be implemented, so everything will be accessible and easy to re-use, these could be Metadata, Cubes or Data marts, and even queries used with parameters.

5) Information Flow. The purpose of this stage is to have all the standard documents of the company in order to check with the BI implemented model and compare if the BI model satisfies the standards company.

6) Control Version. This stage consist on tracking any change done on the code, scripts, BI model, requirements changes, deliveries dates, deliveries priorities.

A. Problem Definition

Today, engineering schools worldwide have a relatively high attrition rate. Typically, about 35% of the first-year students in various engineering programs do not make it to the second year. Of the remaining students, quite often drop out or fail in their second or third year of studies [22]. The purpose of this investigation is to identify the factors that serve as good indicators of whether a student will drop out or fail the program. In order to establish early warning indicators/measures thru the usage of analytics tools being fed by a multidimensional model that can cover the student life cycle.

Next, building a strong semantic layer and a multidimensional model that allow decision makers and users have an unified and consolidated information repository where they can be served, as a self-service instance, with accurate information and obtain consolidated data about many factors and measures such as the status in terms of new students accepted into the campus, performance of the students during their studies and a graduates situation, in each period cycle and career. All these in order to get a big picture of the features that can affect the academic performance of each student by career, looking forward to avoid student's desertion.

B. Data set preparation and exploring.

The idea of this step of the methodology is to know and get familiar with the current transactional reports and information that is currently being in use in CUCEI.
These reports provide information about the status of new student's recruitment until they are graduated, having information and measures about the overall number of all the high-school graduates that applied to be part of the college campus, other measures such as current assignments that the student is having during a cycle and some others such as the status and situation of the recent and non-recent graduates.

This stage of the methodology also looks to define extraction, depuration and data cleaning, eliminating non-needed or adding needed attributes that the user will be using as navigators thru the information. The Figure 1 shows these steps mentioned. The access granted by the University was thru their administrative information tool called SISECA (Tracking System for Quality Educational Program)[17-18], which contains a repository where the reports can be retrieved and used to analyze what time of measures it is used and the dimensions that the information is navigated.

Approximately 50% of the information gathered, to build the BI solution proposed, was obtained using the SISECA repository, being accessed thru SIIAU information system (Integrated Information System and University Administration). The other 50% of the information obtained came from a repository where the student's tutorial activities are tracked (which consist on a set of meetings with his/her tutor and follow up their career path), this repository is called SIT, which is the information system where the tutorials activities are tracked [21-23].

The next stage during the execution of the methodology is to focus on exploring the source data sets, meaning that figuring out what it is needed to retrieve in each one of the reports used at SISECA and SIT information systems, getting familiar with them and eventually consider them as a data validator against the multidimensional solution implemented.

At the end of the design phase objectives were examined and detailed scope of the system, the choice of architecture, and resolution of major risks.

The following logical data flow represents and resumes the steps mentioned on this part of the methodology, the stages of each design phase: Sources, Business Intelligence (ETL, Semantic Layer, and Analytics) and the BI Partner community that will be served with the information obtained.

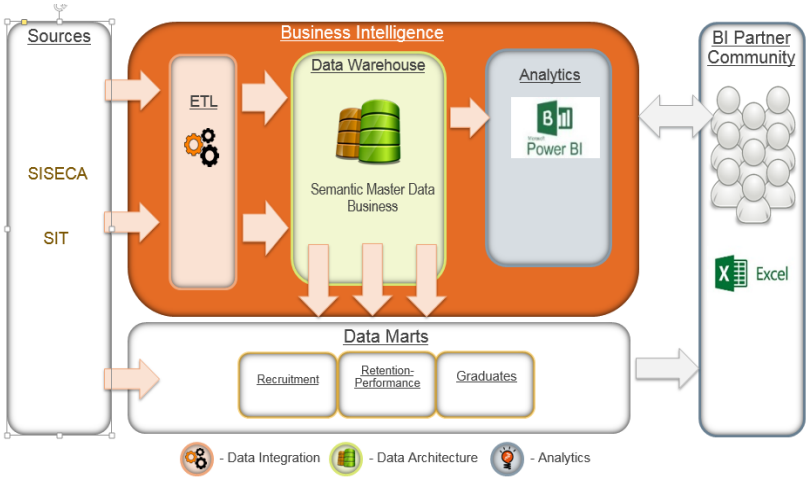


Fig. 1. Business Intelligence Logical Data Flow.

5 Practical Case

Based on the transactional reports provided by SISECA and SIT, the idea of design a Multidimensional solution under a constellation approach was created, including 3 starts schemas that could cover the life cycle of the engineering student, stated under the background section of this paper, for more details about Multidimensional solutions consult [5].

The idea of having a multidimensional solution is to focus all the information to be analyzed in one single repository that can be easy to obtain and retrieve, cross joining

measures that normally are presented or showed separately during the transactional reports and which possible can be difficult to compare using this type of solutions. And also get shared dimensions and measures to obtain better decisions based on the information offered in the model.

And it is because of the 3 stages that a student lives during their stay at the campus, (recruitment, retention-performance, graduates), it was decided to create 3 main fact tables. The Table 1 summarizes the semantic model implemented. Note that each one of the fact tables is joined with share dimensions, which basically are the attributes where the end user will navigate and obtain the needed information to take the proper decision. While as the Figure 2 show the constellation schema (multidimensional model design) which includes the 3 semantic layers described under table 1, that act as 3 stars schemas, resulting a one constellation schema (the union of multiple stars). The idea is that these 3 Fact tables can have shared dimensions as well as information unified.

Table 1. Fact tables descriptions.

FACT	PURPOSE
Recruitment (Fact_Recruitment)	DataMart focused in knowing the situation in the beginning of a cycle by campus and the number of students, prospects students and grades obtained during the admission test.
Retention Performance. (Fact_RetentionPerformance)	DataMart focused on obtaining the right metrics related with each student situation in regards its performance during each cycle time, campus and the different careers.
Graduates (Fact_Graduates)	DataMart focused in knowing the situation of the graduated students who obtain or did not obtain their grades.

Some interesting statistical numbers to be shared, which were obtained from the BI solutions are the following.

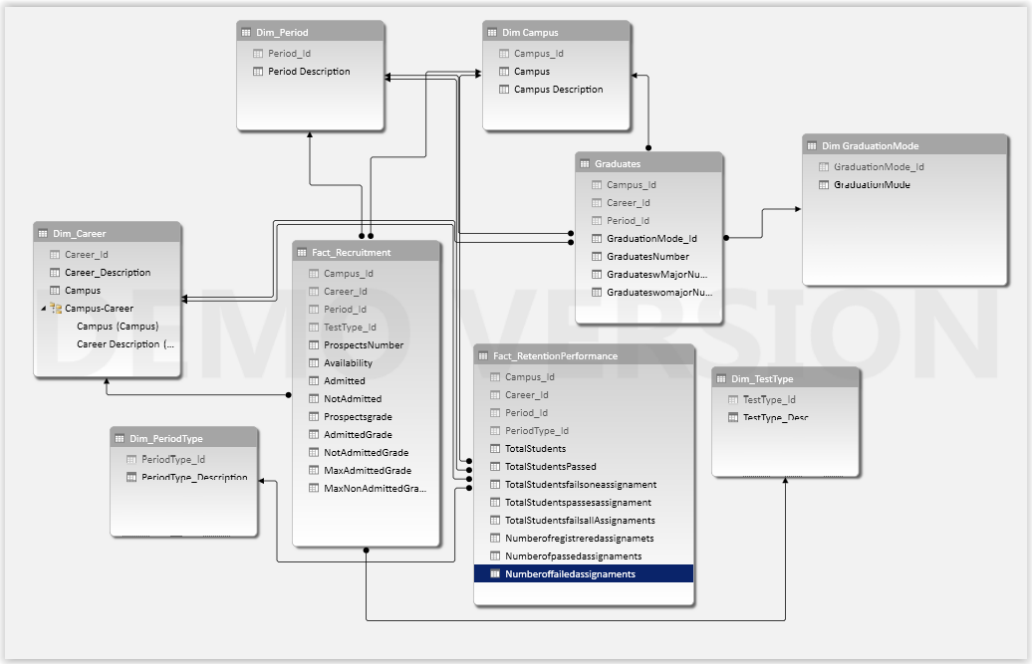


Fig. 2. Constellation Schema implemented in the Project.

- In every period, CUCEI has an admission rate for engineering careers; in overall the admission rate to the campus for all of them is 50% [3] see Figure 3a.
- The highest admission rate is from careers such as: Computer Engineering, Chemistry, Electronic Engineering, among others [3] sees Figure 4.
- The score average of the applicants at CUCEI was 59.7 out of 100, but the score average of the admitted applicants was 68.7 out of 100 where the highest of the non-admitted applicants was 81.7 out of 100 [3] see Figure 3b and 3c.
- Only 52.4% of the students in CUCEI obtain acceptable scores in all of their assignments, 3.7% of the students fail all their assignments and 43.9% of the students fail some assignments [3] see Figure 5.
- The average of enrolled assignments per students is 5.9 and 2.1 assignments are failed per students [3]. 29.6% of the students who did not get an acceptable score and were removed, asked for a new opportunity to enroll again and take the assignment. This occurred between the 1st and the 3rd semester. [3]. 35.9% occurred during the 4th and 6th semester. 22.7% occurred during the last stage of the career.

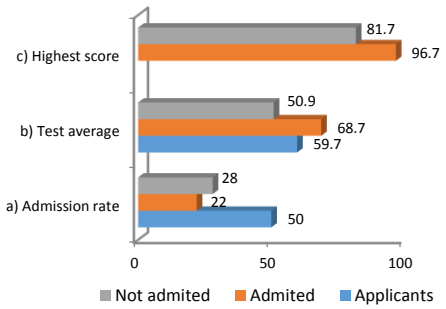


Fig. 3. a) Admission Rate average every period, b) Test average score per type of applicant, and c) The highest score per type of applicant.

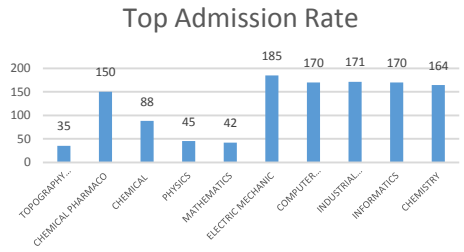


Fig. 4. Top admissions rate.

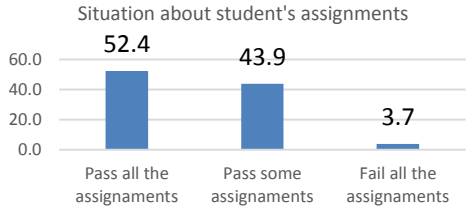


Fig. 5. Average situation about student's assignments.

All the reasons and statistical numbers obtained thru the current transactional reports, surveys and currently the BI system, are tracked at SIT software application system, as an user interface where the tutor can save all their inputs and comments in regard an specific student where economical and psychological aspects are tracked and are not obtained directly from SISECA (main transactional system). Physic and mathematical assignments were the most difficult ones to obtain an acceptable grade and that happened during the first stage of the student career. Some students mention that they do not feel prepared enough in the first semesters of the career due to the assignments commented above.

Additionally, other reasons that affect student's desertion are, in descendent order. 1) Economical and family situation. 2) Family dysfunction. 3) Healthy family problems and students. 4) The long distance the campus is away from the student residence. 5) Not the correct chosen career.

Therefore it is necessary to intensify the work to meet a good performance in each student, for instance the usage of learning objects, design and develop teaching materials, among other things aiming to support vulnerable groups of students to succeed in their career path.

The numbers presented here, from the BI environment results, lead us to propose some strategies that are aligned to CUCEI in order to meet the particular needs improving current conditions in the field of teaching and learning. 1) Systematize studies of relevance and quality. 2) Establish a formal CUCEI graduates monitoring

program and consulting employers. 3) Having evaluated educational programs, with satisfactory results of the assessment bodies. 4) Establish a program of renovation and certification for teachers in teaching and learning environments. 5) Increase the number of teachers with desirable profile. 6) Update and improve infrastructure and sources of information for teaching. 7) Improve infrastructure to support teaching, particularly laboratories and computer services. 8) Establish systems and procedures to give awards, certificates, and assessments to teachers in a timely manner. 9) Establish specific support programs for groups with different conditions. 10) Improve individual student support. 11) Implement the tutorial staff takes place in its various forms. 12) Seek for sport activities.

6 Conclusion and Future Work

Information technology teams need to be aligned into the business looking to be a strategic partner that could contribute on transforming and growing the business and more over making the life of the users easy in terms of information systems.

Analytics skills are needed by the Information Technologies teams in order to provide foundation for delivery of BI applications. Analysis concentrates on understanding business needs for data and information. Designing focuses on translating business information needs into data structures that are adaptable, extensible and sustainable. The core skills to develop include information needs analysis, specification of business metrics and data modeling. A solid understanding of data warehousing concepts, architectures and processes is also essential.

With all these said the end user as a third party validates the BI proposed architecture. Future work will be the use of mathematical models that Petri nets for validation, verification and performance analysis. Also expand the range of use of business process modeling.

Acknowledgment
This work was supported by Instituto Jalisciense de la Juventud (IJJ) in the scope of the project 2011A2986, period April to December 2014, between Universidad de Guadalajara, Jalisco, México and the IJJ, Jalisco, México.

The scholarship obtained by IJJ allowed José Luis Reyna Gascón to obtain his Master Degree in Informatics, sponsored by Jalisco Government. The pay back, due to this support, was to collaborate with CUCEI in creating a multidimensional modeling for their current workflow in terms of Recruitment, Retention-performance, and Graduate process of the students, which is explained thru this paper.

References

1. CUCEI crece en investigación, obras y equipos de laboratorios, Universidad de Guadalajara, consulting in 27-11-2014, http://www.udg.mx/es/noticia/cucei-crece-en-investigacion-obras-y-equipo-delaboratorios
2. Plan de Desarrollo Institucional 2014-2013, Universidad de Guadalajara, consulting in 27-11-2014, http://www.copladi.udg.mx/sites/default/files/pdifinal1_0.pdf
3. Reportes de Admisión a pregrado 2014A, Coordinación de Control Escolar, Universidad de Guadalajara. Internal Document stored in SISECA, www.siseca.udg.mx.

4. Curko K., Pejic Bach M. Business Intelligence and Business Process Management in Banking Operations, Proceedings of the ITI 29th Int. Conf. on Information Technology Interfaces, 2007, 57-62.
5. Kimball, R., The Data Warehouse Toolkit, Wiley Computer Publishing, United States of America, 2002.
6. Liautaud, B., Hammond, M.: e-Business Intelligence, Turning Information into Knowledge into Profit, McGraw-Hill, New York, (2001).
7. Osterfelt, S.; Business Intelligence: The Intelligent Customer, DM Review [http://wdmreview.com], 2000.
8. J.L. Calof, and B. Skinner, "Competitive intelligence for government officers: a brave new world", Optimum, vol. 28, issue 2. pp. 38-42, 1998.
9. L. Cheng, and P. Cheng, " Integration: Knowledge management and business intelligence", IEEE Computer Society, Fourth International Conference on Business Intelligence and Financial Engineering, pp. 307-310, 2011.
10. Foulquier, T.; Perreault, L.; Caron, C.; Levesque, J., "Combining BI Technologies with Microgeomatics at a Clothing Retailer", Proc. on 44th Hawaii International Conference on System Sciences, HICSS, 2011, pp. 1 – 10.
11. B.S.Sahay, and J. Ranjan, "Real time business intelligence in supply chain analytics". Information Management & Computer Security, vol. 16, issue 1, pp. 28-48, 2008.
12. Tvrdiková, M.; Koubek, O., "Support of e-business by business intelligence tools and data quality improvement", IEEE Proc. on International Multiconference on Computer Science and Information Technology (IMCSIT), 2010, pp.271 – 278.
13. Sung Ho Ha ; Sang Chan Park, "Data modeling for improving performance of data mart", IEEE Proc. on International Conference on Engineering and Technology Management, Pioneering New Technologies: Management Issues and Challenges in the Third Millennium, IEMC, 1998, pp. 436 – 441.
14. Selmeci, A.; Orosz, I.; Gyorok, G.; Orosz, T., "Key Performance Indicators used in ERP performance measurement applications", Proc. on 10th IEEE Jubilee International Symposium on Intelligent Systems and Informatics, SISY, 2012, pp. 43 – 48.
15. Yamamoto, R.; Yamamoto, K.; Ohashi, K.; Inomata, J., "Development of a Business Process Modeling Methodology and a Tool for Sharing Business Processes", Software Engineering Conference, Asia-Pacific, Notes in Computer Science, 679-687, 2005.
16. Jun-Jang Jeng, "Service-Oriented Business Performance Management for Real-Time Enterprise", Proc. on the 8th IEEE International Conference on E-Commerce Technology, and Enterprise Computing, E-Commerce, and the 3rd IEEE International Conference on E-Services, 2006, pp. 28.
17. Kin Fun Li; Rusk, D.; Song, F., "Predicting Student Academic Performance", Proc. on Seventh International Conference on Complex, Intelligent, and Software Intensive Systems, CISIS, 2013, pp.2733.
18. CIEP-Universidad de Guadalajara, consulting 11-27-2014, http://ciep.cga.udg.mx/contenidos/54/
19. SISECA-Universidad de Guadalajara, consulting 11-27-2014, www.siseca.udg.mx.
20. SIIAU-Universidad de Guadalajara, consulting 11-27-2014, http://www.siiau.udg.mx/siiau
21. CIEP-Universidad de Guadalajara, consulting 11-27-2014, http://ciep.cga.udg.mx/contenidos/15/
22. SIT-Universidad de Guadalajara, consulting 11-27-2014, http://tutoria.udg.mx/
23. Sistema Integral de Tutorías, CUEI- Universidad de Guadalajara, consulting 11-27-2014, http://tutorias.cucei.udg.mx/

Authentication Protocol for VoIP based on Elliptic Curve Cryptographic Primitives

Zamudio S. Luis[1], Gallegos-García. Gina[2] and Aguilar T. Gualberto[3]

[1, 2] Instituto Politécnico Nacional. ESIME Culhuacan. México
lzamudios0800@egresado.ipn.mx, ggallegosg@ipn.mx
[3] Comisión Nacional de Seguridad México
autg79y@yahoo.com

Abstract. This article presents the design and development of a new authentication protocol for the Voice over Internet Protocol service, on mobile devices. The proposal is different from the actual ones in four key aspects: uses Elliptic Curve Cryptography primitives; works independently from the SIP protocol messages; authentication is not centralized and is not necessary additional hardware to basic VoIP infrastructure. As a consequence, the advantages of the proposed protocol are the following: lower computational cost, no storage of information related to session keys on any device, the minimum infrastructure is needed and the messages exchanged for authentication are transmitted by a different channel from the signaling one. Due to SIP protocol is conditioned upon verifying digital signature and key exchange, generated independently by mobile devices, the results show that a session is not established unless the devices authenticate each other and that the messages for authentication are recognized as CISCO-CSP by a sniffer.

Keywords: Authentication, Diffie – Hellman, Digital Signature, Elliptic Curve Cryptography, Session Initiation Protocol

1 Introduction

During the last decade, communications have boomed and society is becoming more and more interconnected. At the same time, the growing rate in the use of mobile devices and overlapping services is increasing. This is the case of Voice over Internet Protocol (VoIP), a service that grows fairly rapidly and it is believed that in future, will completely replace the traditional Public Switched Telephone Network service (PSTN) [1].

One of the protocols used in VoIP, for signaling, monitoring and session control, is the Session Initiation Protocol (SIP), which was chosen, by the Third Generation Partnership Project (3GPP) [2], as a multimedia protocol for mobile applications and as a proof of the great use and popularity it has gained. However, due to SIP is a protocol for generic use [3] and not specific to VoIP, it is exposed to a variety of security threats, such as: attacks of fake user authentication and impersonation, two issues that have raised various solutions,

since it is expected that the VoIP service provides the equivalent security and privacy level than the PSTN [4].

The SIP protocol is a protocol that sends messages in plain text and uses Hyper Text Transfer Protocol (HTTP) digest authentication scheme, which on one hand has a very high computational cost and on the other hand, it is also vulnerable to man in the middle attacks and false authentication [5]. Considering the aforementioned it is necessary to develop an efficient and secure way of authentication, applicable in VoIP.

In that sense, many schemes for authentication, based on cryptographic algorithms, have been developed. Most of these schemes [6] use centralized servers for authentication, store pre-configured security and key session information and the messages exchanged to accomplish authentication are transmitted within SIP requests through the same signaling channel.

In this paper, a decentralized protocol, but controlled by mobiles devices, is proposed. It makes use of digital signatures and key exchange algorithms on elliptic curves, subjecting the SIP protocol to the successful key exchange and signature verification.

The reminder of this article is as follows: Section 2 gives an introduction to Elliptic Curve Cryptography and the SIP protocol within the VoIP service. In Section 3, related work that has addressed the problem of SIP authentication using additional infrastructure, pre shared security parameters or authentication messages within the SIP protocol using the same signaling channel, is analyzed. Our proposed protocol is presented in Section 4. In Section 5, the proofs are detailed. In section 6 the analysis and discussion are presented. Finally, Section 7 shows conclusions and future work.

2 Background

In this section, we explain the SIP structure and its original authentication scheme, which is based on the HTTP digest [21]. It is also important to understand the Elliptic Curve Cryptography (ECC); particularly the Elliptic Curve Diffie Hellman (ECDH) and the Elliptic Curve Digital Signature Algorithms (ECDSA).

2.1 SIP Protocol, Structure and Original Authentication Scheme.

The Session Initiation Protocol (SIP) is a signaling protocol designed for handling and managing multimedia sessions between two or more parties in a VoIP call. In 2002, it was improved for last time by the Internet Engineering Task Force (IETF), becoming the current RFC3261 standard [5]. To understand the use of the SIP protocol [9], Figure 1 shows the standard flow of a VoIP call that uses SIP to manage the session.

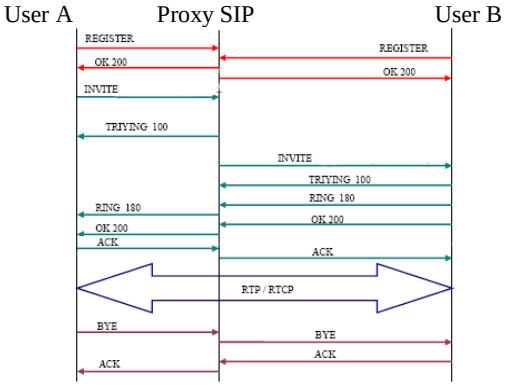


Fig. 1. The flow establishment of VoIP call

The SIP is located in the seventh layer of OSI reference model and runs over TCP transport layer to ensure session establishment [8]. It handles location of users, service quality negotiation for the call, and manages the transfer, modification and completion of sessions [7]. The SIP processing call is based on the client-server architecture and the request-response messages are similar to HTTP or the Simple Mail Transfer Protocol (SMTP) syntax.

Although the original SIP protocol authentication scheme relays on application layer and is based on the HTTP authentication [5], the authentication can be enabled at different layers, including application layer, transport layer and network layer. Along with HTTP, a challenge-response mechanism [5] and Secure/Multi Internet Mail Extensions (S/MIME) protocol are used to provide authentication to SIP Protocol in application layer. Finally there is the Secure Internet Protocol (IPSec) and Transport Layer Security (TLS) that encrypt the messages in transmission layer [9].

2.2 Elliptic Curve Cryptography, ECDH and ECDSA

Elliptic Curve Cryptography (ECC) has extended its use in the past 30 years. Due to its large number of applications, it has been used in mobile devices to meet security requirements in a more robust way. ECC theory is underlined in elliptic curves, defined on finite fields. The main impact of elliptic curves in cryptography is the substantial key size reduction and the lower computational cost due to the operations rely on additive groups [10]. Because of this, the computational benefits, such as reduced memory usage, faster processing and saving bandwidth, ECC fits perfectly in wireless environments and mobile devices. ECC has the versatility to be used in various applications. Figure 2 shows the layers of operation on ECC [11].

The Elliptic Curve Digital Signature Algorithm (ECDSA) [12] is the elliptic curve analogue of the Digital Signature Algorithm (DSA). The Elliptic Curve Diffie Hellman Algorithm (ECDHA) [13] is the adaptation to elliptic curves from the original Diffie Hellman Algorithm.

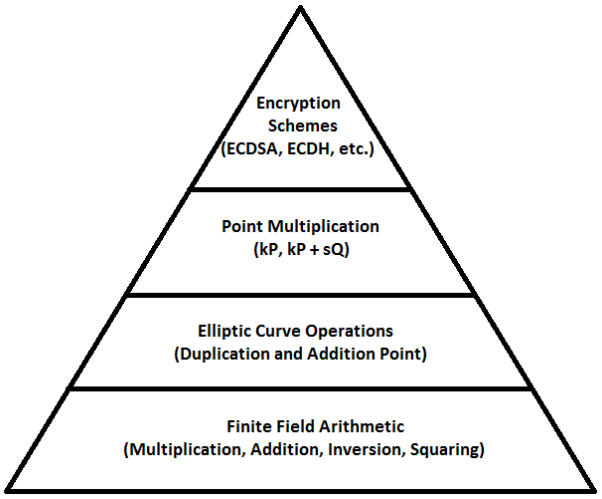


Fig. 2. Branches of Elliptic Curve Cryptography

3 Related Work

The focus of the researches has been divided into four main aspects [14]; proposals based on the Password Authenticated Key Exchange (PAKE); HASH and Symmetric Cipher; Public Key Cryptography (PKC) and ID based schemes. For present work, only the PAKE schemes are subject of study. Within PAKE schemes, there are six main schemes, the first is developed by Yang [15], the second one is proposed by Durlanik and Sogukpinar [16], the third one, proposed by Wu is detailed in [17], the forth scheme is developed by Yoon and Yoo [18], the fifth one is proposed by Qiong Pu and Shuhua Wu in [19] and the last one is shown in [20] and proposed by Shaohui Zhu.

In [15], Yang proposed a scheme with a new way of authentication based on the DH key exchange protocol. This scheme is very close to the original EKE scheme, with the only difference that the encryption part is replaced with an exclusive OR operation and it involves the hash value of the pre-shared password. It means that in step 1 and 2 of EKE in [14], the $Epwd(tx)$ information is replaced with $tx \oplus F(pwd)$. Moreover, in step 2, instead of encrypting a challenge with the session key *K*, the hash value of the session key, together with tx expressed as $F(tx, K)$, is sent by the server.

Authentication is made when the client compare received hash value against a calculated hash value, which the client is capable of calculating after step 2. The advantage of this scheme is the low computational cost in comparison to the EKE scheme. It is due to the encryption and decryption operations are replaced with exclusive OR operations.

Durlanik and Sogukpinar in [16] proposed a new authentication scheme based on ECDH. This is an extension of the EKE and a much closed approach to the Yang´s scheme. The

assumption for this scheme is that both, the server and the client, have a pre-shared password and a public key pair generated by an elliptic curve algorithm. The exclusive OR and classical encryption operations are replaced with elliptic curve encryption operations. The advantage of this scheme is the decreasing total execution time and memory requirement which is good for mobile devices. It is vulnerable to Denning Saco Attack, Offline Guessing Password Attack, Replay Attack and Stolen Verifier Attack.

In [17], Wu proposed the New Authenticated Key Exchange Protocol (NAKE) based on elliptic curve cryptography. The scheme is very similar to Durlanik and Sogukpinar´s scheme, but Wu´s scheme assumes that the client and the server have already agreed upon a password, a hash function and a base point. In step 2 and 3, it uses a random number and two parameters that represent the identities of the server and client. It also uses a session identifier. In addition to that it is vulnerable to Offline Guessing Password Attack and Brute Force attack.

Yoon and Yoo proposed in [18] an authentication scheme based on ECDLP using ECDH. The scheme consists of three phases: system setup, registration and authentication. The assumption for this scheme is that the client and the server have already exchanged messages over a secure channel to establish a shared secret, which is used to compute an elliptic curve point as secret value. This proposal is vulnerable to the Offline Guessing Password Attack.

Qiong Pu´s and Shuhua Wu´s Scheme in [19] show an improvement of [18]. The first two phases are very similar except for there is no shared secret, it is a full domain hash function and also a public parameter. The scheme is vulnerable to Online Guessing Password Attack.

Shaohui Zhu proposes in [20] a scheme that consists of three phases: authentication key agreement, voice data encryption and session key update. The key agreement is made by using a modification of ECDH along with time stamp validation.

4 Proposed Protocol

The proposed protocol differs from the related work in three main aspects: the use of a different channel from the signaling one for authentication messages exchange, there is no necessity of additional infrastructure to basic VoIP infrastructure and the use of a pre shared patron instead of pre shared passwords or configuration tables. All of them are numerated as follows.

4.1 Proposed Protocol Features

1. - The authentication is made independently from the SIP requests inside signaling channel. The exchanged messages for the authentication are transmitted by an UDP socket when the INVITE SIP message is sent by the caller. After this, the SIP requests are blocked and only the required messages for authentication are exchanged. When the authentication is completed, the SIP requests continue the normal flow as in a VoIP call.

2. - There is no need of additional infrastructure. The calculations and operations are made at each end mobile device, which do not storage additional configuration or information for each SIP session. Each session uses a different key. The proposed protocol uses End to End architecture.

3. - The use of a pre shared patron, not a pre shared password or configuration tables. It is used to verify the signature and the validity of the key agreement.

4.2 Protocol Description

The proposed protocol can be divided into three phases: key generation for ECDSA and ECDH, signature generation of ECDH and signature verification.

4.2.1 Key Generation

1) The INVITE SIP message, works as a trigger. Once this message is detected, user A named as the caller catches it and keeps it in stand by while the proposed protocol works. It is important to mention that the initial parameters for the ECDSA and ECDH are defined by the user who initiates the call (user A).

2a) User A executes ECDSA [12] in the following order using the initial domain parameters $D = (q, a, b, P, n)$, pre shared message *m* and the private key d_{A1} randomly selected in the interval [1,*n*−1]. The public key P_{A1} is a point in the elliptic curve equation $Eq(a, b)$ and is computed according to Equation (1).

$$P_{A1} = d_{A1}\, x\, P \tag{1}$$

2b) Initially user A calculates the HASH value of the pre shared message $e = H(m)$ and selects a cryptographically secure random integer $k \in R$ [1,*n*−1], which later, is used for calculation of the curve point kP according to Equation (2).

$$kP = (x1, y1) \text{ and convert } x1 \text{ to an integer } \overline{x1}. \tag{2}$$

2c) Computes the first component, *r* of the signature in Equation (3).If *r* = 0 then select another *k*

$$r = \overline{x1} mod\ n. \tag{3}$$

2d) Computes the second component *s* of the signature according to Equation (4).
If *s = 0* then select another *k*

$$s = k^{-1}(e + d_{A1} r)\ mod\ n. \tag{4}$$

If *s* and$r \neq 0$, then the signature is the pair (r,s)**.** After generation of keys and signature of ECDSA, user A then executes ECDH [13] in the following order using initial domain parameters $D = (q, a, b, P, n)$,. Calculates its private keyd_{A2}, which is an integer less than *n* using the private key d_{A1} as seed.

2e) Calculates the public key P_{A2} as is shown in Equation (5), which is a point in $Eq\ (a, b)$.

$$P_{A2} = d_{A2}\ x\ P \tag{5}$$

4.2.2 Diffie – Hellman Signing

1) User A signs the result of its ECDH calculation according to Equation (6) and sends it to user B named as the callee, along with both public keys $P_{A1,}\ P_{A2.}$

$$S(DH)\ d_{A1} \tag{6}$$

4.2.3 Signature Verification

1) User B receives the messages sent by user A, and verifies the signature using the pre shared message and the public key P_{A1} along with the domain parameters of the curve, but first, user B must verifies that P_{A1} is a valid curve point, by extracting the initial parameters of the curve definition and checking three conditions:

- P_{A1} is not equal to the identity element O , and its coordinates are otherwise valid
- Check that P_{A1} lies on the curve
- Check that *n* $P_{A1} = O$

2) After that, user B computes the Hash value of the message $e = H(m)$ and verifies that *r* and *s* are integers in the interval [1,*n*−1]. If any verification fails then return "Reject the signature"

2a) Calculates the inverse and modular operation *n* of the first component of the signature *s* according to Equation (7)

$$W = s^{-1} mod\ n. \tag{7}$$

2b) Then Computes the *x*-coordinate as seen in Equation (9) using the two components $u1, u2$ from Equation (8), public key P_{A1} and base point .

$$u1 = ew\ mod\ n;\ u2 = rw\ mod\ n. \tag{8}$$

$$X = u1P + u2\ P_{A1}. \tag{9}$$

2c) If X =∞ then return "Reject the signature"; and the VoIP call finishes. Otherwise, convert the *x*-coordinate $x1$ of X to an integer $\overline{x1}$, which is used to calculate the verification *v* of the second component of the signature according to Equation (10)

$$v = \overline{x1}\ mod\ n. \tag{10}$$

2d) If $v = r$ then return "Accept the signature" and the VoIP session continues by User B doing the same actions as User A in steps 1, 2 and 3. If else, return "Reject the signature".

Assuming the valid signature, user B executes Key Generation and Diffie – Hellman Signing phases in the same way as user A. For Key Generation phase (section 4.2.1), executes

ECDSA and ECDH to generate its private keys d_{B1}, d_{B2} and public keys P_{B1}, P_{B2}. At this point, user B has already calculated the Key Agreement.

For Diffie – Helmman Signing phase (section 4.2.2), user B encrypts the pre shared message with the generated key agreement and also signs the DH calculation, then send them to user A along with its public keys P_{B1} and P_{B2}.

Finally, user A receives and verifies the signature of the DH result of user B in the same way user B verified the signature (section 4.2.3). If the signature is not valid, the VoIP call is finished. If the signature is valid, then generates the key agreement and decrypts the pre shared message with its key agreement to validate it is correct.

If the key agreement is correct, the SIP message INVITE, continues the normal flow of a VoIP call. If the key agreement is corrupt the VoIP call finishes.

5 Tests

In order to proof our proposed protocol, a softphone was developed for Android mobile devices, compatible with versions from ICS 4.0 to Kit Kat 4.4.

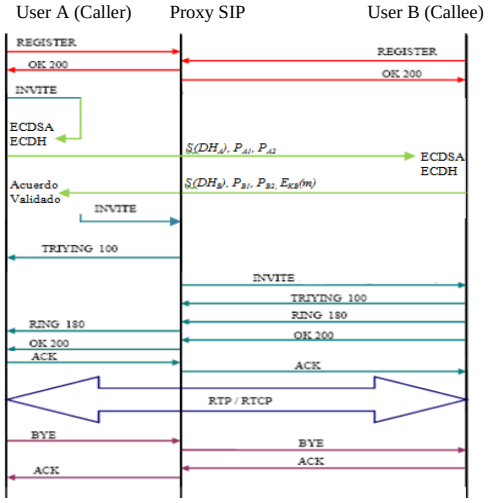


Fig. 3. The flow of the Proposed Protocol for a VoIP call

5.1 Test Scenario

The mobile devices used for testing were: Motorola X, Samsung S5, Sony Xperia SP and HTC One. Figure 4 shows the scenario proposed for the test; it is a self-hosted infrastructure, composed by a PBX server running Elastix, one CISCO router working also as an access point, four mobile devices and one attacker monitoring and eavesdropping the network.

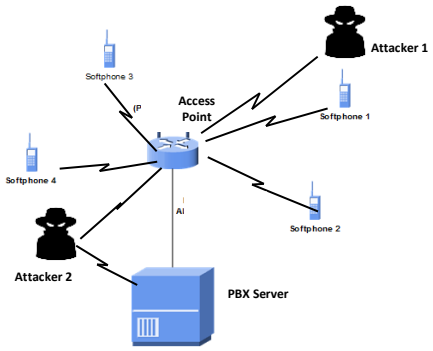


Fig. 4. Self-Hosted Scenario for testing

For the proposed scenario, two mobile devices were installed with a commercial softphone, and two other mobile devices with the developed softphone that includes the proposed protocol. Is assumed that the attacker is able to capture and eavesdrop the traffic.

For the first test, two calls were made. Figure 5 shows the flow of a VoIP call between two devices using our proposed protocol. In this case, due to the users authenticate correctly, the session is established and the VoIP call flows in a normal way. For the second test, it is supposed, that an attacker has already stolen the registration password of another user and try to make a call using its own softphone, as seen in Figure 6, the call is not established. The SIP session is finished after the TRYING and RINGING messages because the authentication protocol is not completed successfully.

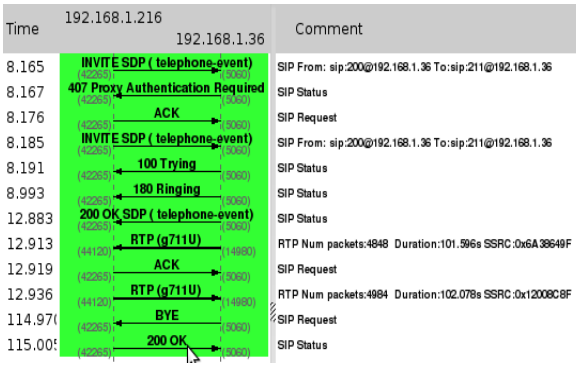


Fig. 5. SIP Session between devices using the proposed protocol

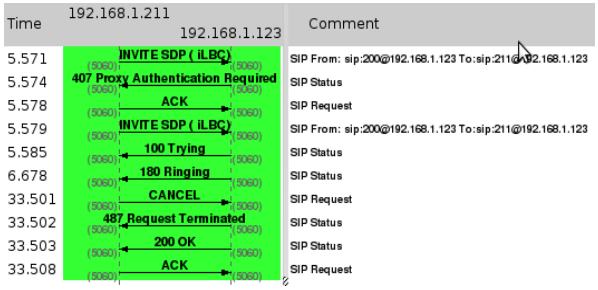


Fig. 6. SIP Session between a device using our proposed protocol and other device using a commercial softphone.

For the third test, the behavior of the two softphones, before, during and after the session was analyzed. Using the Nmap scanner, the following ports and services where identified. In Figure 7, the SIP Ports for the commercial softphone is opened before and during the session establishment. In Figure 8, the ports used for SIP communication, are closed before the session establishment but open only after the session has already been established.

```
# Nmap 6.47 scan initiated Sun Apr 19
14:27:58 2015 as: nmap -vv -sS -p 5060,5061
-oN
/root/Desktop/IdentificaciÃ³nPuertosMIS123T
CPSIP.txt 192.168.1.123
Nmap scan report for 192.168.1.123
Host is up (0.0053s latency).
Scanned at 2015-04-19 14:27:58 CDT for 3s
PORT     STATE    SERVICE
5060/tcp open sip
5061/tcp filtered sip-tls
MAC Address: 00:24:1D:7D:15:67 (Giga-byte
Technology Co.)
Read data files from:
/usr/bin/../share/nmap
# Nmap done at Sun Apr 19 14:28:01 2015 --
1 IP address (1 host up) scanned in 2.44
seconds
```

```
# Nmap 6.47 scan initiated Sun Apr 19
14:37:58 2015 as: nmap -vv -sS -p 5060,5061
-oN
/root/Desktop/IdentificaciÃ³nPuertosMIS123T
CPSIPON.txt 192.168.1.123
Nmap scan report for 192.168.1.123
Host is up (0.0053s latency).
Scanned at 2015-04-19 14:27:58 CDT for 3s
PORT     STATE    SERVICE
5060/tcp open sip
5061/tcp filtered sip-tls
MAC Address: 00:24:1D:7D:15:67 (Giga-byte
Technology Co.)
Read data files from:
/usr/bin/../share/nmap
# Nmap done at Sun Apr 19 14:40:25 2015 --
1 IP address (1 host up) scanned in 147.66
seconds
```

Fig. 7. SIP ports on commercial softphone before/during session establishment

```
# Nmap 6.47 scan initiated Sun Apr 19
14:06:32 2015 as: nmap -vv -sS -p 5060,5061
-oN
/root/Desktop/IdentificaciÃ³nPuertosMIST216
TCPSIP.txt 192.168.1.216
Nmap scan report for 192.168.1.216
Host is up (0.094s latency).
Scanned at 2015-04-19 14:26:48 CDT for 0s
PORT     STATE  SERVICE
5060/tcp closed sip
5061/tcp closed sip-tls
MAC Address: F8:E0:79:5D:8E:54 (Motorola
Mobility)
Read data files from:
/usr/bin/../share/nmap
# Nmap done at Sun Apr 19 14:18:06 2015 --
1 IP address (1 host up) scanned in 694.53
seconds
```

```
# Nmap 6.47 scan initiated Sun Apr 19
14:24:51 2015 as: nmap -vv -sS -p
5060,5061 -oN
/root/Desktop/IdentificaciÃ³nPuertosMIST21
6TCPSIP.txt 192.168.1.216
Nmap scan report for 192.168.1.216
Host is up (0.094s latency).
Scanned at 2015-04-19 CDT14:26:48  for 0s
PORT     STATE  SERVICE
5060/tcp open sip
5061/tcp closed sip-tls
MAC Address: F8:E0:79:5D:8E:54 (Motorola
Mobility)
Read data files from:
/usr/bin/../share/nmap
# Nmap done at Sun Apr 19 14:25:05 2015 --
1 IP address (1 host up) scanned in 15.02
seconds
```

Fig. 8. SIP ports on developed softphone before/during session establishment

6 Discussion

From the proposed protocol we can highlight some key aspects in comparison with Yang [15], Durlanik and Sogukpinar [16], Wu [17], Yoon and Yoo [18], Qiong Pu and Shuhua Wu [19] and Shaohui Zhu [20] schemes.

6.1 Key Aspects of the proposed protocol

Related work use cryptographic algorithms, SIP requests are used for sending and receiving authentication messages. In other words, the information is transmitted within the SIP messages in plain text and is vulnerable for eavesdropping. As a consequence it can be used for guessing the passwords or redirecting the key information. Also, some of these schemes need additional infrastructure like an Authentication Server. This server storages configuration of pre shared passwords or key configuration tables, causing not only false authentication but also adding vulnerabilities like off line passwords attacks or replays attacks. Table 1 summarizes a comparison of these key aspects.

Table 1. Comparison with proposed protocol and actual proposed schemes

Feature	Our Proposed protocol	Related Work
Authentication messages transmission	Independent from SIP. Using an UDP Socket.	Dependent on SIP Using the signaling channel
Authentication Process	Before setting session with SIP protocol.	During session establishment with SIP protocol.
Authentication infrastructure	No needed	Authentication servers
Authentication Information	Not pre-configured or pre shared	Use of configuration and key tables

6.2 Analysis and Discussion

Since authentication and key exchange messages in our proposed protocol are transmitted independently from SIP protocol requests, if an attacker is listening to the channel to capture SIP communication, he will not find relevant information within SIP communication that lets him guess the session key or modify information on digital signatures, on the contrary he would have to analyze all UDP traffic obtained from isolation to SIP.

It is important to mention that method used for authentication and key agreement is known as Signed Diffie – Hellman for multiplicative groups, and according to [22] it is vulnerable to a replay attack. Since user identity is not known, an active attacker could replace the signature of a user with its own signature. In additive groups there is no certainty that this could happen. In [12] is mentioned that even though and attacker be able to obtain signatures of messages from the legitimate signer, if he has not request the new message and obtain the signature he will not be able to produce a valid signature used for fake authentication.

7 Conclusion

We conclude that for an attacker to success on impersonating a valid user, he will need to: analyze all UDP Traffic instead of only SIP, identify the port used for authentication and

request the original signer to send him a new message, so that he could generate a new valid signature. Even if he gets a new valid signature, he would need to get the pre shared patron for verifying the signature, which means hacking the mobile device, not the network traffic. Also, by using only basic VoIP infrastructure and conditioning SIP functionality, the replay and message modification attacks are not feasible.

For future work there is the need of modeling different scenarios of attacking to identify vulnerabilities of the protocol and make the necessary hardening. Also this authentication protocol could also be the base for a secure key establishment and then both parties of the call could encrypt the flow of RTP to provide confidentiality of the call.

Acknowledgments
The authors thank the Instituto Politecnico Nacional and the Consejo Nacional de Ciencia y Tecnologia. The research for this paper was financially supported by Project Grant No. SIP-2014-RE/123, CONACyT 216533.

References

1. ITU. (2010). Measuring the Information Society [Online]
2. IETF. (2005). RFC 4083, Input 3GPP 5 requirements on the SIP. [Online]
3. Pierre Lascuyer, *"Evolved UMTS Architecture" in* Evolved Packet System, the LTE and SAE Evolution of 3G UMTS, 1st Ed, Ed Willey, 2008.
4. Travis Russell, "*Security in a SIP Network" in* Session Initiation Protocol (SIP): Controlling Convergent Networks, 1st Ed, Ed, Mc Graw Hill: 2008.
5. IETF. (2002). RFC 3261, Session Initation Protocol. [Online]
6. H. Hakan Kilinc and Tugrul Yanik, "A Survey of SIP Authentication and Key Agreement Schemes". IEEE. 2013
7. *Specialized Forum on VoIP Telephony*, Foro VoIP: http://voipforo.com/
8. Larry Chaffin, "*SIP Architecture" in* Building a VoIP Network with Nortel´s Multimedia Server 5100, 1st Ed. Ed. Syngress, 2006, pp345 -384.
9. IETF. (2003)RFC 3665, Session Initiation Protocol Basic Call Flow Examples. [Online].
10. Julio Lopez and Ricardo Dahab, "An overview of Elliptic Curve Cryptography", Citeseer, 2000.
11. Lawrence C. Washington. "*Elliptic Curve Crptography*" in Elliptic Curves, Number Theory and Cryptography, 2nd Ed. Ed. CRC – Press. 2008, pp. 169-187.
12. Darrel Henkerson, *"Cryptographic Protocols" in* Guide to Elliptic Curve Cryptography, 1st Ed. Ed. New York: Springer, 2004, pp 153-196.
13. William Stallings, Cryptography and Network Security, 5th Ed. Ed. Prentice Hall, 2011.
14. S. M. Bellovin and M. Merritt, "Encrypted Key Exchange: Password based Protocols Secure against Dictionary Attacks". IEEE. 1992

15. C.-C. Yang *et al*, "Secure Authentication Scheme for Session Initiation Protocol" Elsevier. 2004.
16. A. Durlanik and I. Sogukpinar, "SIP Authentication Scheme using ECDH" World Academy of Science, Engineering and Technology, 2005.
17. L. Wu, *et al*, "A New Provably Secure Authentication and Key agreement protocol for sip using ecc," National Natural Science Foundation of China, 2007.
18. E.-J. Yoon and K.-Y. Yoo, "A New Authentication Scheme for Session Initiation Protocol" IEEE. 2009.
19. Qiong Pu and Shuhua Wu, "Secure and Efficiente SIP Authentication Scheme for Converged VoIP Networks", The international Arab Journal of Information Technology, 2010
20. Shaohui Zhu, *et al* "ECC-based Authenticated Key Agreement Protocol with Privacy Protection for VoIP Communications", IEEE. 2013
21. IETF (2002). Enhanced Usage of HTTP Digest Authentication for SIP [Online]
22. Hugo Krawczyk, (2003), "SIGMA: The 'SIGn-and-MAc' Approach to Authenticated Diffie-Hellman and Its Use in the IKE Protocols", *Advances in Cryptology: 23rd Annual International Cryptology Conference,* [Online], pp. 400-425.

Semantic Indoor Routes in 3D for Emergency Services applied to Buildings

Miguel-Félix Mata-Rivera[1], Alejandro Martínez-Rosales[1], Roberto Zagal-Flores[2], Joel-Omar Juárez-Gambino [2], Consuelo-Varinia García-Mendoza[2]

[1]Interdisciplinary Professional Unit on Engineering and Advanced Technologies, UPIITA-IPN,
Av. IPN 2580, Barrio La Laguna Ticomán, 07340, Mexico City, Mexico
[2] Superior School of Computer Science, ESCOM-IPN,
Lindavista, G.A. Madero, Mexico City, 07738, Mexico

mmatar@ipn.mx, mrosales81@gmail.com, rzagalf@ipn.mx,
omarjg82@gmail.com, cvgarcia@ipn.mx

Abstract. We present an approach to assist emergency services during a contingency. Particularly when they require access to the interior of buildings of more than one level, to mitigate a fire. It is important not only to find the shortest or fastest route, but also identify areas of risk and entry points as well as to identify the location of services and supplies of gas, water and electricity, among others. Knowing these data is possible to define the strategy and route of access more quickly and with greater security. This knowledge can be captured into Ontology. The system provides virtual 3D maps of buildings, displaying the location of gas pipelines, water and electrical wiring superimposed on the virtual building by layers. The routes generated on virtual scenarios unfold in the Z axis, i.e. vertically based on Dijkstra algorithm. Routes show points of access any of the floors of the building. In addition, the system offers the possibility of generating a priori route (via Web) or in situ (while accident is happening). Our case study was made within the Campus UPIITA-IPN using web browsers and Android phones.

Keywords: Semantic Indoor Routes, Emergency Services applied to Buildings

1 Introduction

Emergency services such as firefighters have the difficult task of having to face different types of emergencies, including those having to do with the attention of fires or leaks in buildings with more than one floor. This implies to generate action plans in manual way based on experience and semantic information; in order to identify possible access point

and exits in one level or from one floor to other. These routes can be generated automatically using algorithms routes [12] but modifying them using mobile devices for working in vertical way and considering risks. Another common problem facing a firefighter, is that the indoor building is unknown, and in the same way to location of gas pipes, water, and electricity grid, doors, windows and stairs. In this context the approaches to interior can be adapted to be an alternative solution to this problem [13].

Without considering that in the access points, the brigade may encounter obstacles or locked doors, among other factors hindering rapid intervention in case of disasters. This other problem can be solved through the virtual models approach [12] where having a virtual 3D representation of the building, points of risk and possible access points overlap. The 3D virtual modeling and deployment, currently can already offer interaction and good performance in apps for cell phones [14], which would allow for a mobile, will be deployed on a virtual model the possible entry points and possible routes (vertical) and thereby assist in the action plan to deal with emergencies.

Moreover, having information of a building such as: number of people living and/or working in it, type of building and area (commercial, industrial, rural, hospital, etc.) would allow analyze semantically and reduce various risks, such as damages to people and surrounding infrastructure, range of fire risk at some point (e.g. a gas pipeline). Then, plan of action would be more effective. This would be possible if this information (knowledge) is captured in ontology, a similar approach is described in [15], but no ontologies were neither used nor mobile devices. Thus, in our approach a plan of action is obtained from ontology exploration and associate it with semantic rules, so that together with an algorithm of route, can offer access routes for indoor buildings to meet a contingency.

Then, considering the scenario described, an integrated approach is proposed where through the use of ontologies, virtual 3D models, and a route algorithm is generated an action plan with a set of indoor routes for buildings in an emergency event.

A prototype system with this approach was developed. In which a virtual model of a particular building is displayed, including the location of stairs, emergency stairs, elevators, gas pipes, water pipes and mains. In addition to doors and windows in order to generate routes and support the action plan in case of fire. The system has two versions: Web and mobile. The web version allowed to aided firefighters to establish routes from before the arrival of the team brigade and during the attention of the emergency. The aim is to reduce the brigade's reaction time and the risk margin.

The remainder of the paper is organized as follows. Section 2 presents the state of art while section 3 describes the modules that composes the system, design of ontology and 3D building modeling and the experimental interfaces developed so far. Section 4 shows the tests and obtained results. Finally section 5 concludes the paper and outlines further work.

2 State of Art

Nowadays, research for indoor has attained increasing attention during the last years because today a person an average spends about 90% inside a building [1]. Early research works on indoor environments appears in [2]. For example, the work in [3] uses an airport as case of study to indoor environment and presents an agent-based indoor way-finding simulation. But the Emergency Service Problems require an approach that integrate multiple technologies, in a framework for Supporting Urban Search and Rescue is used for to assist decision making during and after emergencies.

Indoor navigation is not a new concept, but in contrast to outdoor (or car) navigation, it has more and new challenges to deal with; some of them are: shape diversity, degrees of freedom in movement, granularity, and network type [5]. In the case of approaches to routes generation the history says that routing is based on graphs since road networks can easily be described as sets of nodes and edges. One popular approach to solve this problem is to focus on the topology of rooms and build a graph to represent this topology [4]. For guiding rescuers to move quickly from various entrances (within a building) to the disaster site, in[16] the authors developed a spatiotemporal optimal route algorithm for micro-scale emergency situations using real-time data for a GIS-based emergency response system. For another hand, research for routes generation combined with ontologies is focused on network analysis or in the planning process based on impedance or weight [6,7].

Ontologies have been signaled as a tool to assist in tasks of indoor environments, including the example of firefighters as possible application of the approach [9]. Moreover, navigation ontology for outdoor and indoor space is being built based on the exploration of the relationships between the two spaces [10]. In this context, SEMA4A[18] is an Ontology that includes concepts and relationships for provide emergency notifications (accessibility, emergency, communications). Now, the use of 3D models has been used to support route planning, in [8] an approach to route planning is presented according to users' widths, heights and requirements (e.g. avoid stairs). Thereby trends on indoor research suggest to explore integral approaches like is presented in this paper.

3 Framework

The architecture of the solution is composed of four modules: 1) First one is the semantic module, this contains the ontology, semantic rules, spatial and non-spatial information of the building (type, location, neighborhood, people live or work on building, among other data). This module provides the requirements to generate a route. It generates the parameters to send to the action plan module. While 2) Virtualization model handles the creation and rendering of 3D model for each building, the scenery, and displaying. In addition, it generates the queries for sending to semantic module. 3) Action plan module is the third one of the system; it manages the options of visualization, call to algorithm to generate the

route (based on the requirements obtained from ontology). 4) Finally the fourth module manages the displaying option and the interaction with user. In Figure 1 is shown the general architecture with the modules described.

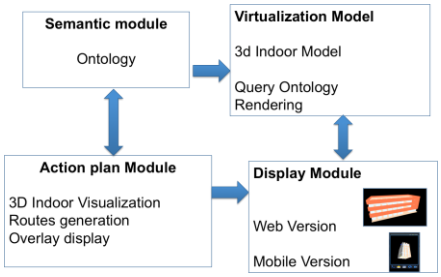


Fig. 1. Architecture of system

The general functionality of approach is based on the following scenery with a sequence of steps that show the attention process when an emergency is presented:
1) A situation emergency is presented, 2) the report of the emergency (by telephonic line) is received, 3) the report is sent to the emergency squad, who comes out to attend the emergency 4) System is accessed via web and mobile to generate an action plan (strategy) to address the incident, 5) possible plans of action are sent and displayed to the fire brigade that is in charge of the emergency.

3.1 Design of Ontology

Ontology was designed based on common procedures of fireman and brigade fire, concepts and classes are identified in order to analyze the indoor structure of building and find the access nodes and risk nodes. These data are used as an input parameters of action plan module.

The process of ontology exploration is based on the nature of B-Tree used in databases [11]. Exploration of ontology allows knowing considerations and restrictions required on a path for a particular building. The result of ontology exploration is a list of nodes. These node represent risk points and free points (trusted points), they are inserted into a vector, and these vector is send as an input parameter to the action plan module, to generate the possible indoor routes using and adaptation of Dijkstra.

Structure of ontology uses semantic relations (well-known in ontology domain) they are listed below:

- Hyponymous relationships ("is a" relation or hyponym-hyperonym), generic relation, genus-species relation: a hierarchical subordinate relation. (A is kind of B; A is subordinate to B; A is narrower than B.

- Instance-of relation. Designates the semantic relations between a general concept and individual instances of that concept.
- Locative relation: A semantic relation in whom a concept indicates a location of a thing designated by another concept.
- Meronymy, partitive relation (part-whole relation): a relationship between the whole and its parts.

Then, ontology is an upper- ontology domain, where concepts and classes capture the knowledge required to analyze a building and indicate the parameters to consider in the path generation. A snippet of Ontology is shown in Figure 2.

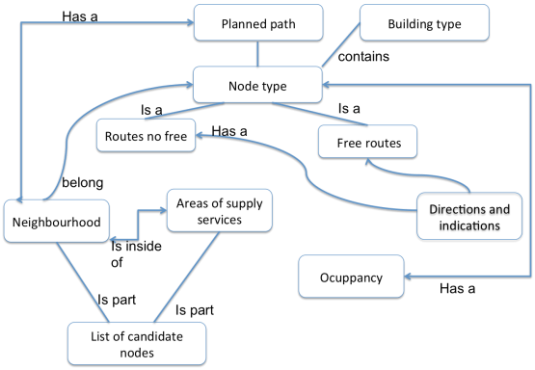


Fig. 2. Snippet of Ontology

As it can see, ontology contains concepts and classes related to analyze the possible paths the node type (hazardous, no dangerous) level of occupancy, use of building (building type) and other properties such as schedule, identification of areas where exist supply services. Semantic relations allow expressing queries in order to obtain the knowledge and help in the decision-making, this process is explained in the next section.

3.2 Semantic Query Processing

The semantic module receives the descriptive information of the building (address or location). This information is used to explore the ontology and answer semantic queries, we define some queries in order to help in case of emergency. One of the first aspects to know in the case of emergency is to know the location of hazard materials o dangerous areas. It is expressed in the query 1.

Q_1= {Find places contains hazardous materials inside the building}

Ontology is explored with an adaptation of the algorithm to explore B-Tree. When the class and concepts are found, the neighborhood nodes of them are extracted and stored into a vector. This vector is send to the virtualization model in order to display and generate the model of building with risk point and secure points. As a result the areas or spaces containing hazardous materials in the building are shown with a dot red line in its contour. It is shown in Figure 3.

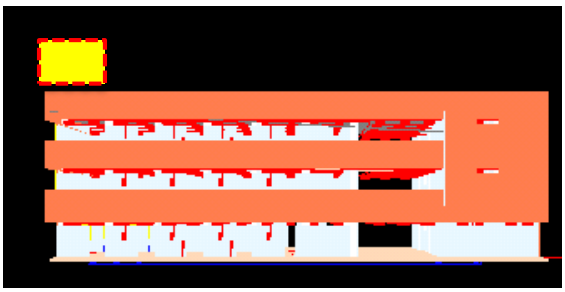


Fig. 3. Hazardous places (dot red line) identified in building

Here, it is possible retrieving semantic information such as: It is located in level 3 , in the surrounding areas is found a laboratory with 18 person working throughout academic hours, the hazard material type is gas. Other required data is identify all risk points and exit points, this is made using the query 2, and the result is shown in Figure 4 where exit points are displayed using green squares.

Q_2= {Retrieve all risk points and exit doors}

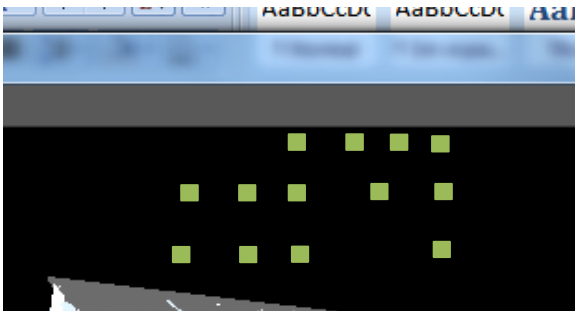


Fig. 4. Hazardous places (dot red line) identified in building

While, in query 3 is possible to know an estimated number of people in the building. In addition, use of the space and people involve in it. This information is useful to know all possible entry point and planning for evacuation.

Q_3= *{Retrieving occupancy information}*

In that way, is semantically processed the information, and displayed on a virtual model in 3D. In the next section is explained how is build the 3D model.

3.3 Buildings Modeling

The construction of 3D model is based on two steps: the first one is to take architectonic drawing of building as a basis of model 3D construction. In Figure 3, side view of a building (semi-automatically generated from building drawings) indicating vertical routes to access the different levels of the building is shown. Secondly, all vertices that make up the building are loaded: external walls, internal walls, columns, stairs, etc. which will become property of a scenario in which the model (building) will be displayed and additionally will allow the capture of events, so that the user can interact with the model. Figure 5 shows the result of this step.

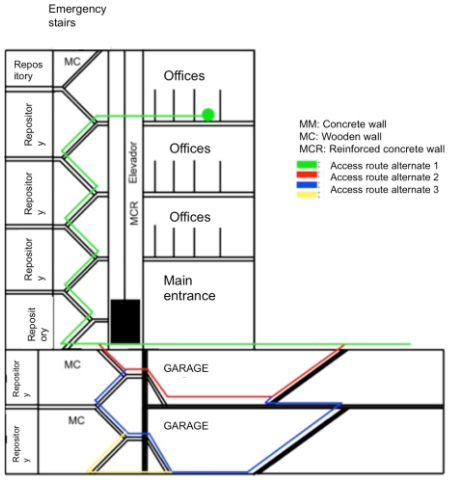


Fig. 5. Lateral View Building and possible routes (in red, green and yellow)

Each view model requires three classes that are deploying the model, the first one allow interaction with elements of itself, the second is the one who draws and rendering the model; the third is one that detects the angle from which displays the model. The resulting model is shown in Figure 6 where elements are seen in yellow (gas pipelines and gas stationary) elements in blue represent water pipelines and the water tank). Moreover, elements in red, representing electrical wiring and power stations.

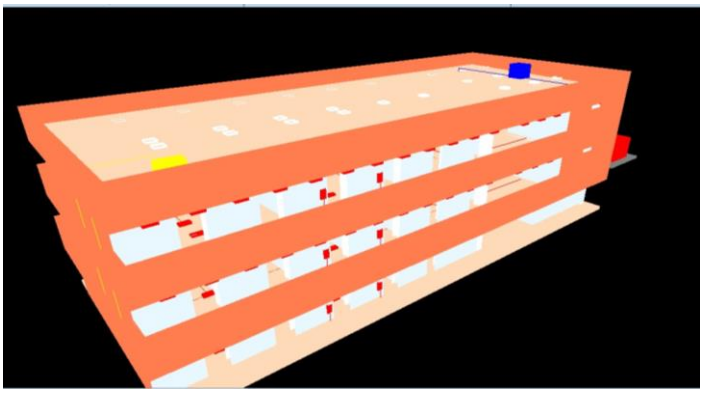


Fig. 6. Model of Building in 3D obtained from drawings

As explained overall functionality of the modules is now proceeds to show the tests and results in the next section.

4 Results

In this section, is shown the interfaces developed, for a simulated case of fire event, and evacuation plan as an emergency. Testing was made using a cellular phone with android 2.3 for mobile version. While, the web version was tested using Google Chrome version 42.0.2311.152 (64-bit). Firstly, the model of building is displayed with all layers (a layer contains a service supply) with a checkbox is possible to overlay each layer in the model. This is shown in Figure 7

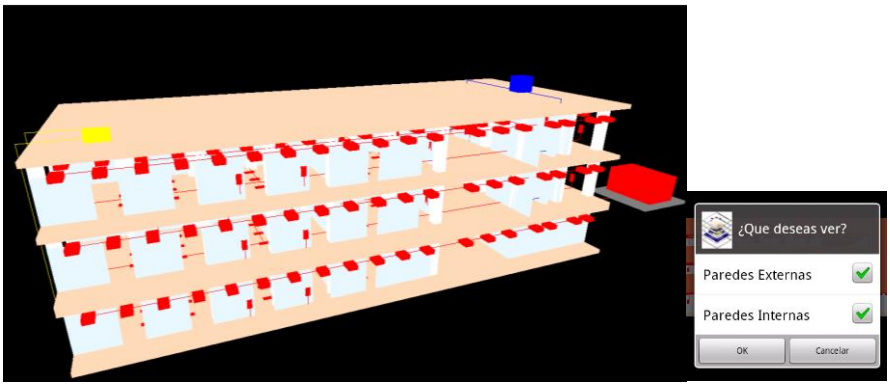


Fig. 7. Views supplies (electric mains, water and gas pipelines) in the building

Figure 7 display all the supplies services found in a building, in this case, in red appears electric mains, in blue color the water pipelines and finally in yellow the gas pipelines as displayed. The system allows selection through the layers, filtering the elements or services that are displayed. In addition, system shows the action plan or the routes as an instructions in text mode. It is shown in Figure 8a, while in Figure 8b, a possible evacuation route when an emergency has occurred is shown with a green line (in this case the danger is present in the rooftop, while people is located in first level, at west side, then the evacuation route is shown from west to east, and from first level to ground floor. These results are shown using mobile devices.

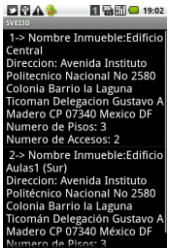


Fig. 8a routes instruction in mobile version

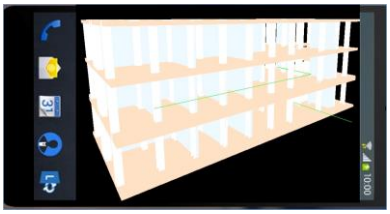


Fig. 8b routes in mobile version

Routes generation is shown when a fire is not present, it is shown in Figure 9a, while when danger is present the Figure 9b shows the trusted routes.

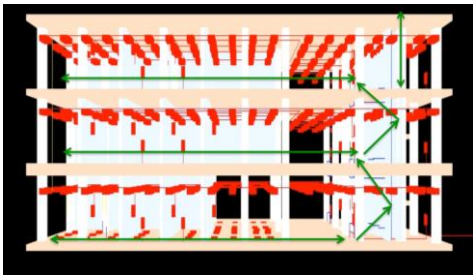


Fig. 9a trusted routes (in green color)

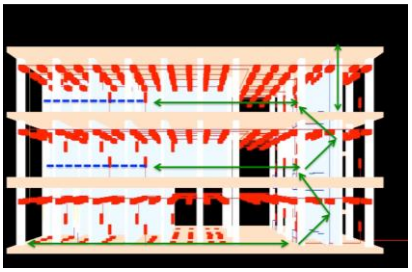


Fig. 9b risk routes (in blue color)

5 Conclusions and Future work

Currently, fire brigades and have minimal knowledge of the interior of a building, even though there are tools and maps to assess around before reaching the site of the disaster. This means, much time will be lost in the generation and evaluation of the plan of action. In fact, having this information a priori or to generate automatically. It will increase the chances of rescuing civilians, while rescuers are less exposed to serious situations inside. The 3D models are a good choice to make a visual analysis, while ontologies can generate the action plan based on recommendations and procedures. The integration of ontologies, 3D model, route algorithm, indoor approach and mobile technology, allow offering a new possibility based on mobile devices making tasks to assist in real time in cases of fire o evacuation plan. In these cases, precise information is vital to avoid injury in people and damage in infrastructure and surrounding areas when an emergency occurs. The tests show that the approach is promising and useful. The future work is to integrate command voice, and interactions using augmented reality, in particular it could be useful when the brigade is in situ and helping in real time.

Acknowledgment
The authors of this paper want to thank to SIP (Secretaría de Investigación y Posgrado), IPN (Instituto Politecnico Nacional), COFAA (Comisión de Operacion y Fomento a las Actividades Academicas del IPN) and EDI for their support.

References

1. Klepeis, N., Nelson, W., Ott, W., Robinson, J., Tsang, A., Switzer, P., Behar, J., Hern, S., Engelmann, W.: The national human activity pattern survey (nhaps. Journal of Exposure Analysis and Environ-mental Epidemiology 11(3) (2001)231-252.
2. Raubal, M., and Worboys, M.: A formal model of the process of wayfinding in built envi ronments. In: Spatial information theory. Cognitive and Computational Foundations of Geographic Information Science, Lecture Notes in Computer Science 1661 (1999) 381-399
3. Raubal, M.: Ontology and epistemology for agent-based wayfinding simulation. International Journal of Geographical Information Science 15(7) (2001) 653-665.
4. Horst Steuer. High precision 3D indoor routing on reduced visibility graphs. In Progress in Location-Based Services, pages 265–275. Springer, 2013.
5. E. Stoffel. Hierarchical graphs as organizational principle and spatial model applied to pedestrian indoor navigation. PhD thesis, Ludwig Maximillian's University Munich, 2009.
6. Chunithipaisan, S., James, P., and Parker, D. Online Network Analysis from Heterogeneous Datasets – Case Study in the London Train Network, Map Asia Conference 2004. China, 2004.

7. Abolghasem Sadeghi-Niaraki, Kyehyun Kim, and Cholyoung Lee. 2008. Ontology drive road network analysis based on analytical network process technique. In *Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology* (CSTST '08). ACM, New York, NY, USA, 613-619. DOI=10.1145/1456223.1456349
8. W. Yuan and M. Schneider. 2010. Supporting 3D route planning in indoor space based on the LEGO representation. *ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness* (ISA '10). ACM, New York, 16-23. DOI=10.1145/1865885.1865890
9. Liping Yang and Michael Worboys. 2011. A navigation ontology for outdoor-indoor space: (work-in-progress). In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness (ISA '11). ACM, New York, NY, USA, 31-34. DOI=10.1145/2077357.2077364 http://doi.acm.org/10.1145/2077357.2077364
10. L. Yang and M. Worboys. Similarities and differences between outdoor and indoor space from the perspective of navigation. Accepted poster. Conference on Spatial Information Theory (COSIT 2011), September, Belfast, ME
11. I. Jaluta, S. Sippu, and E. Soisalon-Soininen. 2006. B-tree concurrency control and recovery in page-server database systems. *ACM Trans. Database Syst.* 31, 1 (March 2006), 82-132. DOI=10.1145/1132863.1132866 http://doi.acm.org/10.1145/1132863.1132866
12. H. Tashakkori, A. Rajabifard, M. Kalantari. A new 3D indoor/outdoor spatial model for indoor emergency response facilitation. Building and Environment. Vol. 89, 2015, P. 170–182.
13. Y. Zhou, T. Hong, J. Thill, E. Delmelle, Enhanced 3D visualization techniques in support of indoor location planning. Computers, Environment and Urban Systems, Vol. 50, 15–29.
14. M. Tsai, N. Yau. Enhancing usability of augmented-reality-based mobile escape guidelines for radioactive accidents. Environmental Radioactivity. Volume 118, 2013, P. 15–20.
15. T. Onorati, A. Malizia, P. Diaz, I. Aedo. Modeling an ontology on accessible evacuation routes for emergencies. Expert Systems with Applications. Volume 41, 2014, P. 7124–7134
16. Y. Chena, F. Peña-Mora b, P. Plans, J. Mehtad, Z. Aziz. Supporting Urban Search and Rescue with digital assessments of structures and requests of response resources. Advanced Engineering Informatics, volume 26, pages 833–845
17. J. Lee, M. Kwan, Spatiotemporal Routing Analysis for Emergency Response in Indoor Space. Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography. Vol. 32, No. 6, 637-650, 2014 http://dx.doi.org/10.7848/ksgpc.2014.32.6.637
18. A. Malizia, T. Onorati , P. Diaz , I. Aedo , F. Astorga-Paliza. SEMA4A: An ontology for emergency notification systems accessibility. International Journal of Expert Systems with Applications. Volume 37, Issue 4, April 2010, Pages 3380–3391

Improving Parameters of the Gamma Associative Classifier using Differential Evolution

Antonio Ramírez-Ramírez[1], Itzamá López-Yáñez[2],
Yenny Villuendas-Rey[3], and Cornelio Yáñez-Márquez[4].

1, 4 CIC, Instituto Politécnico Nacional, México
2 CIDETEC, Instituto Politécnico Nacional, México
3 Facultad de Informática de la Universidad de Ciego de Ávila, Cuba.

Abstract. This paper presents an effective method to improve some of the parameters in an associative classifier, thus increasing its performance. This is accomplished using the simplicity and symmetry of the differential evolution metaheuristic. The Gamma associative classifier, which is a novel associative model for pattern classification, is formed by some parameters, that when modified, it has been found to be more efficient in the correct discrimination of objects; experimental results show that applying evolutionary algorithms models, the desired efficiency and robustness of the classifier model is achieved. In this first approach, improving the Gamma associative classifier is achieved by applying the differential evolution algorithm.

Keywords: pattern classification, metaheuristics, Gamma Associative Classifier, differential evolution

1 Introduction

The standard paradigm of computing, which peaked with countless algorithms for almost all computing tasks has been overtaken by new metaheuristic hybridization techniques. Thus, neural networks, cellular automata, evolutionary strategies and quantum computation [1] [2]; all unconventional computing tasks, have gained new strength to be enhanced with the benefits of metaheuristics. In applying these algorithms in the automated search for values that are most suitable for the parameters of Gamma Associative Classifier (CAG, for its acronym in Spanish), it could maximize it to obtain more accurate classification results. Today the computer systems are facing a major problem is access to large data sets, therefore, the application of new non-traditional techniques of performing the tasks of programming and measurement is required. Such is

the case of several evolutionary algorithms, in particular, the differential evolution algorithm (DE). The latter is an algorithm recently used in unconventional computing area, and has proven to be a good solution for some of the optimization problems facing the paradigms of traditional computing, such as processing large volumes of information, as chemical and biological systems of our environment [3] [4] [5]. DE is a metaheuristic developed by R. Storn and K. Price [6], this algorithm is based on vectors and is cataloged within evolutionary algorithms, since it makes use of cross and mutation operators [7].
DE is used especially in problems involving large search spaces. Unlike genetic algorithms, DE is based on simple addition and subtraction operations, enabling you to simply and exploration of the search space more efficiently. Because of this and their own features as vector components, result in a better solution to specific problems [8].
On the other hand, in the area of pattern recognition (RP) they have handled four major tasks traditionally: clustering, regression, recovery and classification, the latter being one of the best known and addressed subjects [9] - [12]. In academia, have been studied these four tasks in different approaches and areas of application, such as the probabilistic-statistical approach [13], the deterministic approach or based on metric [14], the structural syntactic approach [15], focus or neuronal area [16] and the associative approach [10] [17] [18], among some others. Within the associative approach, approximation models have been developed that allow not only make recovery and pattern classification, but also achieved good results in the regression task, and even address some of the problems of prediction. Recently, the CAG has been applied successfully in some relevant issues like time series prediction in the oil wells production context [19] and the hybrid associative classifier with translation (CHAT by its acronym in Spanish) in medical applications [20].
This article addresses the one-parameter setting, of one of the best associative classifiers, the CAG and the results of the application of differential evolution algorithm (DE) over the parameters of CAG, are also presented [6], in order to adjust these parameters automatically and thus verify the improved performance of model. The increased performance of the classifier was demonstrated when was adjusted its parameters with the metaheuristic.

The rest of the paper is structured as follows: At the Associative Gamma Classifier section, is described the origin and basis of the classifier. In the Parameters subsection of CAG, it briefly describes its operation. In CAG parameter adjustment section it explains and exemplifies a way to manually look for different values for one of its parameters, in order to improve their performance. The central proposal of this work is presented in section looking for the solution vector with (DE). The algorithm description and some

examples are presented on Experimental results section. Finally conclusions are presented.

2 Gamma Associative Classifier

2.1 Gamma operators

The CAG emerged as a need to combine the power and versatility of the Alpha and Beta operators, developed at the Center for Computer Research since more than ten years ago and are the source and foundation of Alfa-Beta associative memories, which have been implemented in several issues of interest in the recovery pattern and pattern classification [21]. These two operators are defined in tabular form, considering the *A={0,1}* and *B={0,1,2}* sets, as shown in **Table 1**.

Table 1. Definition of the Alpha and Beta operators.

α: A×A → B

x	y	α(*x*,*y*)
0	0	1
0	1	0
1	0	2
1	1	1

β: B × A → A

x	y	β(*x*,*y*)
0	0	0
0	1	0
1	0	0
1	1	1
2	0	1
2	1	1

These operators, combined with the association model that is based on patterns coding method, known as modified Johnson-Möbius code [23], gave rise to Alfa-Beta bidirectional associative memories [10], the new association method, based one-hot vector classifier [20], as well as the CHAT-OHM method [20]. All these techniques are associative models, along with the CAG, with which they share common foundational elements: both they are based on associative memories and they operate thanks to the Alpha and Beta operators.

In recent years the CAG has been used in significant research as predicting of production of oil wells [19] and in the prediction of atmospheric pollutants [22], among others. This was made possible by its simple but efficient construction, as discussed in [22], along with the original operators Alpha and Beta, the design and operation of the CAG two

additional original operators was involved: the u_β operator, and Gamma operator of similarity generalized.

The unary u_β operator, receives as input an n-dimensional binary vector and delivers an integer number using the following expression:

$$u_\beta = \sum_{i=1}^{n} \beta\,(x_i, x_i) \tag{1}$$

The generalized similarity Gamma operator γ_g receives as input tow binary vectors $\mathbf{x} \in A^n$ and $\mathbf{y} \in A^m$, where $n, m \in \mathbb{Z}^+, n \leq m$, as well a non-negative number θ and delivery as result a binary digit, which is calculated as:

$$\gamma_g(x, y, \theta) = \begin{Bmatrix} 1 & if \quad m - u_\beta[\alpha(x, y) mod 2] \leq \theta \\ 0 & otherwise \end{Bmatrix} \tag{2}$$

That is, γ_g outputs 1 if both vectors differ by a max θ bits, and outputs 0 otherwise. This means that the CAG works on a set of patterns associated with a class called the fundamental set. These patterns $x^i \neq x^j$ meet ∀ i, j ∈ *{1, 2, ..., p}, i ≠ j,* which means that each of these patterns, it is unique and is associated with a single class, depending on the definition of problem addressed.

2.2 Gamma parameters

CAG algorithm consists of several parameters, among which can be highlighted a weights vector called *w*, this vector indicate how much each variable contributes to the decision to discriminate between input patterns. By modifying the values of this array, changes are obtained in the classification results. Other classifier parameters susceptible of modification is θ_0 (initial theta) representing the value that starts to evaluate how different can be two patterns and that gamma operator considers similar instead. This parameter is subject to ρ parameter (stop), which refers to the maximum value allowed to θ and able the algorithm to continue the search for a disambiguation near a border patterns; when theta reaches the value set to stop$(\theta = \rho)$, the CAG will cease iterating, assigning or placing in a class to the input pattern. In addition to the parameter ρ unemployment, CAG also has a parameter ρ_0 (pause), which helps the program to undertake an assessment of the pattern to classify, to determine their belonging to the unknown class.

Other parameters are the input variable d, which is evaluated to decide whether to classify the pattern belongs to an unknown class or to any of the known classes and finally the parameter u, which is the threshold against which d parameter is compared to decide which class the pattern belongs to class. Currently, there are empirical recommendations by the author of the CAG, to assign values to some of their parameters, without there so far exist, some automated method to determine the appropriate values for these parameters. In the flow diagram of Figure 1, we can appreciate the CAG algorithm model, which has highlighted in a different color, the elements (parameters) that may be adjusted.

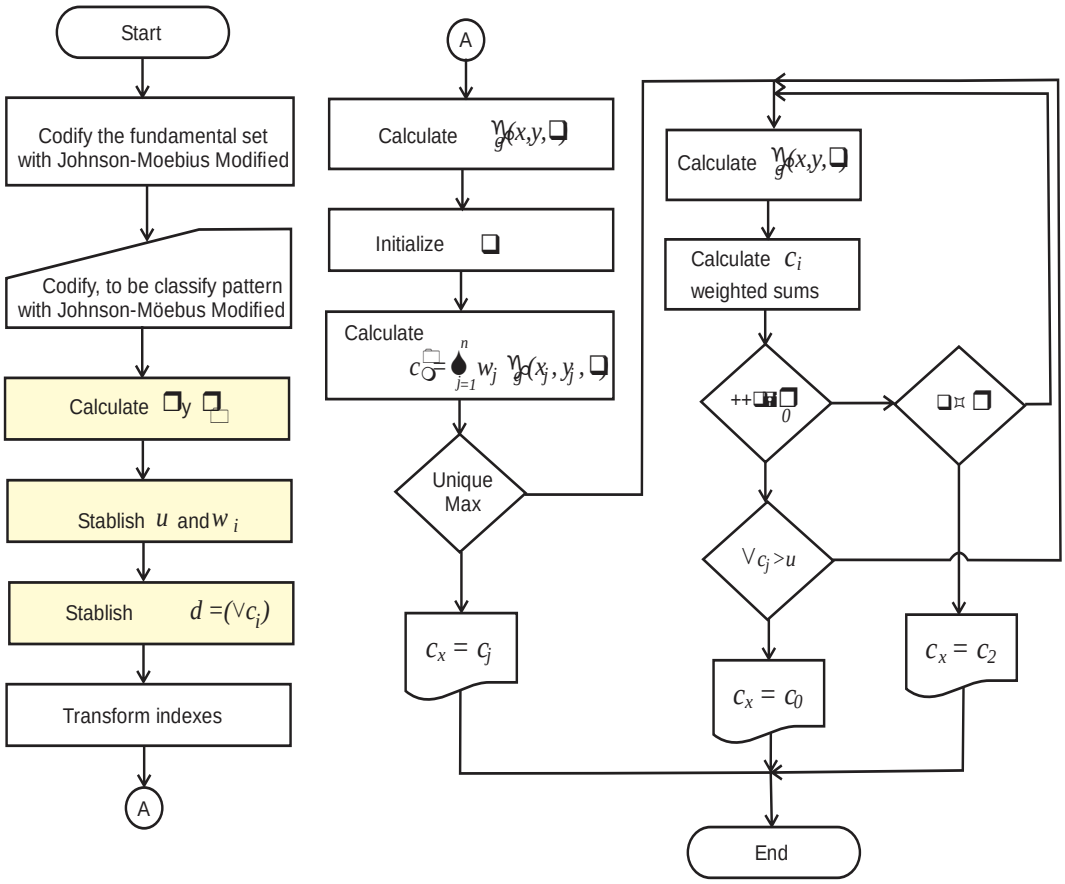


Fig. 1. Flowchart of the learning phase of the CAG.

This model widely discussed in [35], was modified to achieve an improved performance of it. It was decided to apply the metaheuristic on search of better w_i parameter weighting.

2.3 CAG parameters adjustment

In an initial experiment, a series of initial values assigned to the parameters w, θ, ρ, d and u was proposed; for better performance CAG like in [35]. The defaults that have been used for these parameters and have delivered good results in the various problems that have been successful using the CAG are shown in Table 2.

Table 2: Default values of CAG parameters in [35]

$w_i = 1, i = 1, 2, \ldots, n$ θ initialized on 0. $\rho = \bigvee_{j=1}^{n}\left(\bigvee_{i=1}^{p} x_j^i\right)$	$\rho_0 = \bigwedge_{j=1}^{n}\left(\bigvee_{i=1}^{p} x_j^i\right)$ $d = \vee c_i$ $u = 0$

In this initial recommendation proposed by the author of the CAG in [35], it raises a number of initial values that can be checked increased efficiency classifier for some problems. In recent years has been used successfully CAG and there have been some case studies, which have emerged, some empirical rules for the allocation of initial values for these parameters, This values definitely depend on the characteristics of data set to be evaluated. In developing these experiments, it was found that changes improve the performance or in the worst cases, revealed the same conclusions, which had no changes. In Table 3 it can be seen that when modifying the weights of two different data sets (Data Set Iris Wisconsin Breast Cancer and Machine Learning UCI Repository [38]), these changes happen.

Table 3: case 1: Default values, case 2: suggested values in [35].

Data sets	Cases	Weights	Errors	Performance
Iris	1	[1.0,1.0,1.0,1.0]	11	92.67%
	2	[0.5 0.5 1.5 1.0]	7	95.33%
Breast Cancer	1	[1.0,1.0,1.0,1.0, 1.0,1.0,1.0,1.0,1.0]	59	91.55%
	2	[0.5,1.0,1.0,1.0, 1.0,1.5,1.0,1.0,1.5]	59	91.55%

Although apparently it's a very simple example, you can observe that in fact, when the weight vector *w* from CAG, you apply small changes, resulting in a different outcome and check if that change does not improve the original solution, at least retains the level performance. But the question then arises: will exist, and if so, could be automatically find a combination of weights such that depending on the characteristics of the data set needed to evaluate, provide better performance CAG?

3 Looking for the solution vector with DE

The assumptions underlying the central proposal of this article is that by applying a metaheuristic in the search for more precise w_i parameter from CAG, it can automate that search in order to find the most appropriate weighting to achieve better results in pattern classification problems where the AGC is applied.

A metaheuristic that of a simple and low computational cost, would find solutions naturally vectors is Differential Evolution (DE) [6], which has been applied to solve various problems of pattern classification in different tasks [29-34]. This metaheuristic is a direct search method that is robust, easy implementation and has fast convergence [24] [25] and it is highly parallelizable [26-28]. In [36] for example, dramatic results that demonstrate the effectiveness and efficiency of DE, when applied in 18 different problems arise. In these results some variants of the original algorithm, as proposed in [37], wherein the algorithm is improved harmonic DE include search. In the flow chart in Figure 2, it shows the basic operation model of DE algorithm.

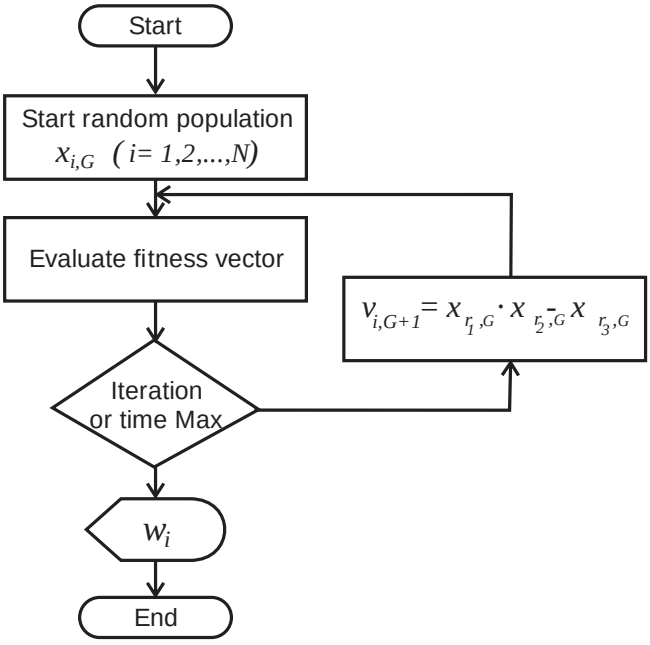


Fig. 2. Flowchart of basic Differential Evolution model.

Basically DE algorithm generates a population of N d-dimensional vectors, considered one generation G; that generation vectors are denoted by:

$$x_{i,G}, i = 1,2,\cdots,N \quad (3)$$

where N do not change across the process and the initial vectors population is automatic selected by randomly. The new vectors of parameters are generated by a process called mutation, it consist on for each element on $G, x_{i,G}, i = 1,2,\cdots,N$ what is an target vector, is determinate one value $F\epsilon(0,2] \subseteq \mathbb{R}$ and three different integer indexes are chosen randomly in pairs, and result on a mutant vector which agree with the expression:

$$v_{i,G+1} = x_{r_1,G} + F \cdot (x_{r_2,G} - x_{r_3,G}) \quad (4)$$

The parameters of the mutated vector is mixed with parameters another predetermined vector (*target vector*) by a process called *cross*, and hence a *vector test*, in which the cost function associated with the problem to be solved is evaluated is obtained. The target, the mutant and test, in the process of crossing, three vectors are involved through a random process. The 'cross' occurs between pairs of real values in the closed interval [0,1] and is determined randomly, evolving a vector of it, generating the weight vector w_i. If the test vector leads to a lower value on the cost function that the target vector, the vector replacing the target test vector in the next generation, thus leading to the *selection*.

There are so many works that offers detailed explanation for implementation of this metaheuristic in literature, as in [6, 26-28, 29-34, 36] and some others. Nevertheless, as commented paragraph above, the metaheuristic is applied on weights vectors population, and then generate an evolved vector that could make better CAG performance. In the Flowchart fragment in Figure 3, change appears on original model that enable apply the metaheuristic to find a better w_i vector and achieve the improved performance of CAG.

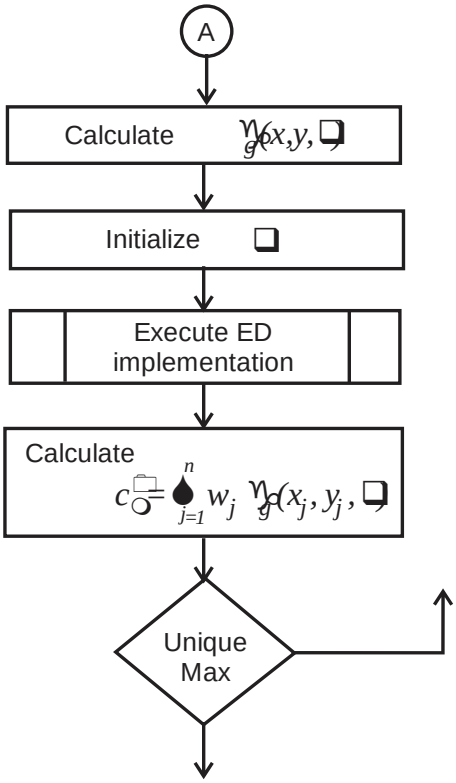


Fig. 3. Flowchart fragment from DE implementation, over CAG model.

4 Experimental Results

When the metaheuristic is performed over one data set, the outputs are a combinations series that conform a vectors list, to be used by CAG and to obtain a better result. The Table 4 shows a vectors list generated in order to apply the CAG on Glass data set.

As can be seen, the values between [0 - 1] for each on vector element, represent the weight of this attribute, that takes to account the CAG to make the classification. That is, on what measure the feature, is considered a weight attribute for this classification.

In the above work it has already demonstrated the feasibility of the algorithm DE, able to find multiple solutions that are at least as good as the best known. In Tables 5 and 6, these results can be seen and check the improvement achieved with relatively low cost, as populations of 100 individuals were used and convergence is reached in less than 50 iterations. In both cases, line 1 and 2 shows the CAG performance with empirical

suggestions as in [35] and the remaining lines, shows performance with w_i vector improved by DE algorithm.

Table 4: Weight vector delivered by DE to apply the CAG.

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9
0.677304	0.473165	0.519946	0.020773	0.950655	0.171773	0.581708	0.241777	0.153159
0.407815	0.460470	0.589114	0.689878	0.433206	0.124820	0.694805	0.838947	0.704812
0.003967	0.933099	0.735842	0.963115	0.756653	0.114569	0.874673	0.617319	0.496038
0.958731	0.854590	0.794889	0.766650	0.754062	0.806594	0.468336	0.231333	0.284859
0.689434	0.087731	0.135545	0.320545	0.786423	0.039731	0.336544	0.019812	0.818385
0.419945	0.075037	0.204713	0.989650	0.268974	0.992778	0.449641	0.616982	0.370039
0.762775	0.775271	0.040309	0.610759	0.558764	0.595403	0.558132	0.913882	0.522149
0.105413	0.229669	0.604417	0.347079	0.298745	0.917939	0.911512	0.019474	0.692385
0.836116	0.462810	0.945073	0.900974	0.331106	0.151076	0.779720	0.807953	0.225912
0.432075	0.689603	0.820313	0.289422	0.104743	0.860735	0.204476	0.395016	0.035265

Table 5: Results comparison from table 3 (iris data set) vs results of parameters modified by DE.

Data Set	Try	Weights	Errors	Performance
(previous result with suggested values)	1	[1.0 1.0 1.0 1.0]	11	92.67%
	2	[0.5 0.5 1.5 1.0]	7	95.33%
Iris Data Set Instances: 50 Classes: 3 Dimensions: 4	1	[0.1 0.1 1.5 1.5]	8	94.67%
	2	[0.0 0.0 1.5 1.0]	6	96.00%
	3	[0.0 0.0 2.5 1.0]	5	96.67%
	4	[0.0 0.0 2.5 1.0]	5	96.67%
	5	[0.0 0.0 4.0 2.0]	5	96.67%

Table 6: Results comparison from table 3 (breast cancer Wisconsin data set) vs results of parameters modified by DE.

Data Set	Try	Weights	Errors	Performance
(previous result with suggested values)	1	[1,1,1,1,1,1,1,1,1]	59	91.5594%
	2	[0.5,1,1,1,1,1.5,1,1,1.5]	59	91.5594%
Breast Cancer Wisconsin. Instances: 699 Classes: 2 Dimensions: 9	1	[1,2,2,1.5,1,2,1,1,0.5]	49	92.9900%
	2	[0.5,2,2,1,0.5,2,0.5,0.5,0]	47	93.2761%
	3	[0,2,2,1,0,2,0,0,0]	46	93.4192%
	4	[0.5,4,4,2,0.5,4,0.5,0.5,0]	46	93.4192%
	5	[0,2,2,1,0,2,0,0,0]	46	93.4192%

Although when to apply the same databases that worked in [35], proposal methodology provides good results, the classification was performed improved CAG with on other databases, this way we noted the increased performance of CAG, when combined with the metaheuristic used. In Table 7, one can see the improvement in the classification performed on another three databases by comparing the result of applying the CAG in its original format and with the modified DE.

Table 7: Comparison of results from all five processed data sets with the original CAG and modified by DE

Data set	Features	Instances	Classes	CAG outputs	Weights by DE
iris	4	150	3	0.9533	0.9667
breast-cancer-wisconsin	9	699	2	0.9156	0.9667
diabetes	8	768	2	0.6641	0.6782
glass	9	214	4	0.5318	0.5550
wine	13	178	3	0.8536	0.8592

Conclusions

This paper has presented an analysis that supports the hypothesis that through appropriately modify values of the parameters of the CAG, is possible improve the overall performance of this associative model that is currently at the forefront in terms of recovery methods, classification and regression. The original suggestion, proposed in [35] shows work specifically on these data sets, but nowadays is working in several others data sets with free access to demonstrate the truthiness and performance of this implementation. The substantial contribution lies in the application of a metaheuristic on searching of the correct weight vector instead of manually trying, this improvement was

made possible in this case, through the application of a metaheuristic that continues to work in several worldwide laboratories, who are also at the forefront of research and is known as Differential Evolution.

Acknowledgements The authors would like to thank the Instituto Politécnico Nacional (Secretaría Académica, COFAA, SIP, CIC, and CIDETEC), the CONACyT, and SNI for their economical support to develop this work, as well as the Informatics Faculty of the University Of Ciego de Ávila Cuba.

References

1. Yaoyao Hea, Qifa Xua, Shanlin Yanga, Aoyang Hanc, Li Yang d, "A novel chaotic differential evolution algorithm cascaded", International Journal of Electrical Power & Energy Systems, Elsevier, Vol. 61, pp. 455-462, China 2014.
2. Eleanor Rieffel and Wolfgang Polak, "An Introduction to Quantum Computing for Non-Physicists", ACM Computation Surveys, FX Palo Alto Laboratory, USA 2000.
3. Adamatzky, et al., "Unconventional Computing 2007", ISBN-13: 978-1-905986-05-7, Luniver Press, USA 2007.
4. K. L. Du, M.N.S. Swamy, "Neural Networks in a Softcomputing Framework", Springer Verlag London Limited, ISBN 1-84628-303-5, London 2006.
5. Enrique Alba, "Parallel Metaheuristics, a New Class of Algorithms" @Zeicience, A John Wiley & Sons, Inc., Publication, España 2005.
6. Rainer Storn and Kenneth Price, "Differential Evolution - A simple and efficient Heuristic for global optimization", Kluwer Academic Publishers, Journal of Global Optimization, Vol. 11, pp. 341–359, Netherlands 1997.
7. Carlos Coello, et al., "Un nuevo algoritmo evolutivo para la optimización de una o varias funciones objetivo sujetas a restricciones", Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería, Vol. 20, pp. 139-167, España 2004.
8. Vasileios K. Karakasis, and Andreas Stafylopatis, "Efficient Evolution of Accurate Classification Rules Using a Combination of Gene Expression Programming and Clonal Selection", IEEE Transactions on Evolutionary Computation, Vol. 12, No. 6, USA 2008.
9. Vidar V. Vikjord, and Robert Jenssen, "Information theoretic clustering using a k-nearest neighbors approach", Pattern Recognition, Elsevier Vol. 47 pp. 3070–3081, USA 2014.
10. María E. Acevedo-Mosqu eda, Cornelio Yáñez-Márquez, Itzamá López-Yáñez, "Alpha–Beta bidirectional associative memories: theory and applications", Neural Processing Letters, Vol. 26 pp 1–40, USA 2007.
11. Kötschau, et al., "Element Pattern Recognition and Classification in sunflowers (Helianthus annuus) grown on contaminated and non-contaminated soil", Microchemical Journal, Elsevier, Vol. 114 pp. 164–174, USA 2014.
12. Nileshkumar H. Jadhav, "Subset selection in multiple linear regression in the presence of outlier and multicollinearity", Statistical Methodology, Elsevier Vol. 19 pp. 44–59, USA 2014.
13. Richard O. Duda, Peter E. Hart and David G. Stork, "Pattern Classification", 2ª. Ed., John Wiley & Sons, Canada 2001.

14. Thomas M. Cover, Peter E. Hart, "Nearest Pattern Classification", IEEE Transaction Information Theory, Vol. 13, no. 1, pp. 21-27, USA 1967.
15. S. Watanabe, S., "Pattern Recognition: Human and Mechanical", Wiley 1st Ed., ISBN: 978-0471808152, New York 1985.
16. J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities", Proceedings of the National Academy of Sciences of the United States of America, Vol. 79, No. 8, 2554-2558, USA 1982.
17. Karl Steinbuch, "Adaptive Networks Using Learning Matrices", Kybernetik Band 2, Germany 1964.
18. Humberto Sossa, R. Barrón, R. A. Vázquez, "New Associative Memories to Recall Real-Valued Patterns", A. Sanfeliu et al. (Eds.): CIARP, LNCS 3287, pp. 195–202, Springer-Verlag, Berlin 2004.
19. Itzamá López Yáñez, L. Sheremetov, C. Yáñez-Márquez, "A novel associative model for time series data mining", Pattern Recognition Letters 41, Elsevier B. V., USA 2014.
20. Abril V. Uriarte A., et al., "One-Hot Vector Hybrid Associative Classifier for Medical Data Classification", Plos One, www.plosone.org, Vol. 9, Issue 4, USA 2014.
21. Yáñez-Márquez, J. Luis Díaz de L., "Memorias Asociativas basadas en relaciones de orden y operaciones binarias" Computación y Sistemas Vol. 6 No.4 pp. 300 - 311, *ISSN 1405-5546,* Centro de Investigación en Computación – Instituto Politécnico Nacional, México 2003.
22. Itzamá López Yáñez, Amadeo J. Argüelles-Cruz, Oscar Camacho-Nieto, and Cornelio Yáñez-Márquez., "Pollutants Time-series Prediction using the Gamma Classifier", International Journal of Computational Intelligence Systems, Vol. 4, No. 4, México 2011.
23. Rolando Flores Carapia, "Memorias asociativas Alfa-Beta basadas en código Johnson Möbius modificado", Tesis de Maestría, Centro de Investigación en Computación del Instituto Politécnico Nacional, México 2006.
24. Javier Apolloni, Enrique Alba, "Empirical evaluation of distributed Differential Evolution", standard benchmarks", Applied Mathematics and Computation, Vol. 236, pp. 351-366, Elsevier Inc., España 2014.
25. Yu SUN, Yuanxiang Li, Jun Liu, Gang Liu, "An Improved Differential Evolution Algorithm with Ensemble of Population Topologies", Journal of Computational Information Systems, 8667–8674, Binary Information Press, China 2012,
26. Marina S. Ntipteni, Ioannis M. Valakos and Ioannis K. Nikolos, "An Asyncronous parallel differential evolution algorithm", Department of Production Engineering and Management, Technical University of Crete, Grece 2007.
27. D. Zaharie, D. Petcu, Parallel implementation of multi-population differential evolution, in: Concurrent Information Processing and Computing, pp. 223–232, USA 2005.
28. Marco Tomassini, "Parallel and Distributed Evolutionary Algorithms: A Review" Institute of Computer Science, University of Lausanne, 1015 Lausanne, Switzerland 1999.
29. Kötschau, et al., "Element Pattern Recognition and Classification in sunflowers (Helianthus annuus) grown on contaminated and non-contaminated soil", Microchemical Journal, Elsevier, Vol. 114 pp. 164–174, USA 2014.
30. Eleanor Rieffel and Wolfgang Polak, "An Introduction to Quantum Computing for Non-Physicists", ACM Computation Surveys, FX Palo Alto Laboratory, USA 2000.
31. John R. Koza, "Genetic Programming on the Programming of Computers by Means of Natural Selection", A Bradford Book of the MIT Press, Cambridge, Massachusetts, England 1998.
32. D. B. Fogel and J. W. Atmar, "Comparing Genetic Operators with Gaussian Mutations in Simulated Evolutionary Processes Using Linear Systems", Biological Cybernetics, Vol. 63, pp. 111-114, Springer-Verlag, USA 1990.
33. J. M. García Nieto, E. Alba Torres, G. J. Luque Polo, "Algoritmos basados en Cúmulos de Partículas Para la Resolución de Problemas Complejos" Tesis doctoral, Universidad de Málaga, España 2006.

34. Xin-She Yang, Mehmet Karamanoglu, Tao Luanb, Slawomir Koziel, "Mathematical modelling and parameter optimization of pulsating heat pipes", Journal of Computational Science, Elsevier Vol. 5 pp. 119–125, China 2014.
35. Itzamá López Yáñez, "Teoría y aplicaciones del Clasificador Asociativo Gamma", Tesis Doctoral, Centro de Investigación en Computación del Instituto Politécnico Nacional, México 2011.
36. Huizhi Yi, et al., "Three improved hybrid metaheuristic algorithms for engineering design" Applied Soft Computing Vol. 13 pp. 2433–2444, Elsevier Ltd. USA 2013.
37. T. Warren Liao, "Two hybrid differential evolution algorithms for engineering design optimization", Applied Soft Computing, Elsevier Vol. 10, pp. 188–1199, USA 2010.
38. Bache, M. Lichman, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. University of California, School of Information and Computer Science, Irvine, CA 2013

Markov Models and their Application to Support Police Chase

Fernando Martínez[1], Víctor Hernández[1], Luis González[1], and Gabriel González[2]

[1]Universidad Autónoma de Chihuahua, Facultad de Ingeniería, 31125, Chihuahua, MX
{fmartine, vhernandez, lcgonzalez}@uach.mx
[2]Centro Nacional de Investigación y Desarrollo Tecnológico, Palmira, 62490, Cuernavaca, MX
gabriel@cenidet.edu.mx

Abstract. For any country the safety of citizens is of high priority. Social security should be provided opportunely and efficiently without exposing the lives of people. Developed countries have a large scale of technology to monitor criminal activities in urban spaces. This technology helps central station to coordinate police patrols when on the chase of a car. Developing countries on the other side often have technological limitations that constraint the police possibilities to chase law offenders. This work presents a vehicular network platform to help police departments particularly with situations of chasing a car on the run. We highlight the work done with the utilization of Markov chains to predict future location of the offender's car. At the current stage the algorithm can predict the escape route with 40% of success. SUMO simulator is being used to test the route predictor.

Keywords: vehicular networks, police chasing, urban computing, security and police assistance.

1 Introduction

One of the main concerns for a government is to offer a secure urban ambient for citizens, a task that can be challenging. The European Forum for Urban Security (Efus) issued the "Manifesto of Aubervilliers and Saint-Denis" in which member cities are called for the establishment of policies for crime prevention. For instance, surveillance cameras are employed in developed countries to monitor criminal activities. This information enable police department to offer quick responses. Other technologies such as sensor networks complement video monitoring and allows for tracking criminal on the move. On the other hand, the lack of technology in developing countries compromises the secure environment offered to citizens. Mobile technology together with the cloud services help explore solutions to support and assist the police department activities in the streets, An example of this situation could be, the logistic for the creation of a virtual barrier in order to constraint

the space of navigation for a car being driven by a criminal. In order to clarify our application scenario we offer the following application context:

Immediately after committing a crime the criminal jumps into his vehicle and starts the escape. A police officer observes the offender and starts a persecution. By means of the available technology in the car, including 3G communication capabilities to a web service, the officer signals his vehicle as the leader of the persecution. Using geo-location data the system running in the car continuously calculates possible future locations for the car being chased. This information is shared via web services with other police's vehicles that are around in order to signal and bring them together to the virtual barrier to be created to restrict escape's roads.

With this scenario in mind we explore two settings for applications. First, the development of a technology platform that integrates a vehicular network with the cloud hence automating the creation of the virtual barrier. Second, a statistical model based on Markov chains for the prediction of navigation routes to support police chasing. The rest of this work is structured as follows. In section two we review the application of VANETs-like technology as well as the application of artificial intelligence techniques to support intelligent transportation systems. Section three describes the implementation of the two components of the system, the route predictor and the cloud-based communication services. Section four discusses results from experiments, and section five presents the conclusions and future work.

2 Motivation

Patterson et al. [1] presents a probabilistic framework that seeks to infer and predict the mode of transportation a person uses, bus, foot or car. Position and velocity information, obtained from GPS readings, were processed by means of Bayesian filters in order to track the location and transportation mode of a person. In addition to the identification of a mode of transportation, Liao et al. [2] work aims to infer whether the location of a person matches with his/her daily goals (home/workplace), if there was an opportunistic event that modifies the person's routines (why the person missed the bus), and what would the next user movement be (if the person goes to the workplace s/he will turn left on the next corner); Rao-Blackwellised particle filters process GPS data and to build a predictive model that could help guidance systems for cognitively-impaired individuals. Simmons et al. [3] considers the context of driving routines and the information can be draw to predict route trajectories and destination intent. By training Markov models with past performance of drivers it was possible to make predictions of his/her destination with 98% of accuracy. Mao et al. in [4] equipped taxi vehicles with GPS sensors to collect information about the car's trajectories and with this data build a map of the city with points of interest (POI's). By means of the POI's map the weighted Markov model offers

interesting results for driving direction prediction, for instance, to identify roads with the most flow traffic.

By looking up onto discrete road's segments the car has just been driven its near-term future path can be predicted [5], including the anticipation of what direction the driver should turn in order to prepare for the next intersection [6]. The prediction Markov model was trained with the driver's past history and on this basis offered different destinations options, for instance, when there is the situation of slow traffic or when the car needs to fill gas. The implementation of accurate travel route prediction systems is challenging because its nature dependency with, for instance, the urban architecture and its associated traffic conditions [7].

As observed there is a great interest for implementing services that track and predicts individuals' mobility. Most of the proposed works seek to build systems that could assist persons with special needs but no one, as far as our understanding, has considered a police chase context, an urgent need for support in developing countries. An extra word in this regard is that we have talked to a police department and have found together the opportunity to contribute with a system that can offer support to the efforts they do to offer social security.

3 Supporting Police Activities

As mentioned before the proposed system aims at exploring the integration of a vehicular network with the cloud technology in contexts that supports police chasing, figure 1. In this work, however, we focus our attention to the route prediction Markov model.

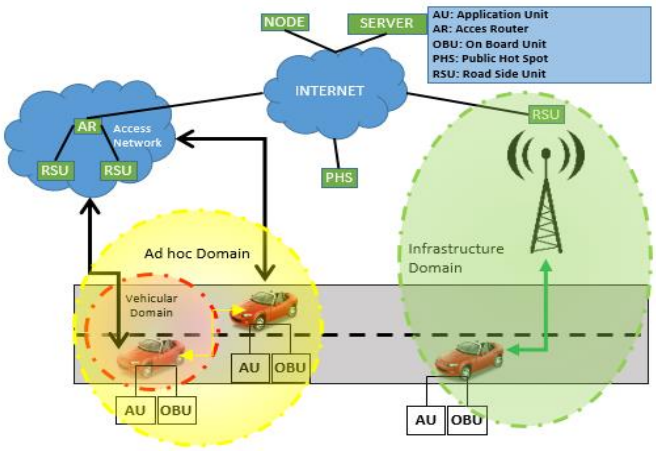


Fig. 1. General architecture for a system that support police chasing.

3.1 Route prediction

Markov chains are commonly referenced in the literature as a mathematical model that fit well for the prediction of future location for mobile targets. For the context of police chasing we are applying Markov chains of first and second order. The car being hunted moves through different adjacent intersections (as in a "chain"). The intersection that will be taken next depends only on the current state of the system, i.e. what intersection the car is coming from. In our police's chasing context the changes of intersections the car performs are called transitions. The probabilities associated with various state changes are called transition probabilities, represented in a transition matrix. Generally speaking, the transition matrix offers the conditional probability distribution space from which it is possible to predict a chain for different states given an initial state.

The probability distribution $P[X(n+1)]$ for the next intersection is independent of all, except $X(n)$, the current intersection. Then:

$$P[X(n+1)|X(n), X(n-1), \dots, X(0)] = P[X(n+1)|X(n)]$$

Thus, on the first hand the first order Markov model allows the prediction of what intersection the car being hunted would direct to in the very near future if the current intersection the car is in or has passed through is known. The probability distribution for such a case is given by $[X(n+1)|X(n)]$. That is, the probability distribution is defined considering all of the next immediate intersections available with respect to the direction the car is being driven to. The calculation of the first order Markov's transition probabilities considered that any of the periphery nodes are available is done by means of a heuristic like the one shown in figure 2.

```
Start (calculation of probabilities first order)
i, j are indices of the nodes in the transition matrix
For each i
      For each j
            p_ij = V(i,j) / N_i
      EndFor
EndFor
End
```

Fig. 2. Pseudo-code for the calculation of the transition matrix for a first order Markov model.

Where p_{ij} is the transition probability from node i to node j, $V(i,j)$ is the number of outgoing paths from node i to node j and N_i is the number is the total number of outgoing lanes of node i**.** Verifying that $\sum_{j \in E} p_{ij} = 1 \quad \forall\, i \in E$, i.e., sums per row is 1.

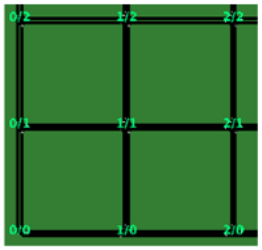
```
Start (Calculation of probabilities second order)
i, j, k are indices of the nodes in the transition matrix
For each i
      For each j
            For each k
                  If i ≠ k Then
                        p_{k|j.i} = V_jk / (S_j − Q_ij)
                  EndIf
            EndFor
      EndFor
EndFor
End
```

Fig. 3. Pseudo-code for the calculation of the transition matrix for a second order Markov model.

On the second hand, the second order Markov model considers the last but one intersection the car was seen, $P[X(n+1)|X(n), X(n-1)]$. Using the second order allows the consideration of a couple of nodes, hence a street segment, during the calculation of the transitions matrix. Figure 3 presents the heuristic can help for the calculation of the transition probabilities. There we must consider that $p_{\{k|j,i\}}$ is the conditional probability of going to the node k given nodes j, i, V_{jk} is the number of outbound lanes from node j to node k, S_j is the total outgoing lanes of node j, Q_{ij} is the number of lanes connecting node i to node j. Verifying that $\sum_{k \in E | k \neq i} p_{\{k|i,j\}} = 1 \quad \forall\, i, j \in E$, i.e., sums per row is 1.

Considering an urban area of nine intersections for which the transition matrix is the one given by figure 4, we can traverse through the following intersections tracing the following chain. Starting from intersection 0/0 the next probable transition can be either intersection 0/1 or intersection 1/0, because both intersections have the same weight. If it moves to the intersection 0/1 the car can go to either the intersection 0/2 or the intersection 1/1, or even return to intersection 0/0. In general, if the current state (intersection) for a mobile object is known then we can seek up into the transition matrix to know the potentially near future transition for that object.

map segment



transition matrix

	0/0	0/1	0/2	1/0	1/1	1/2	2/0	2/1	2/2
0/0	0.00	0.50	0.00	0.50	0.00	0.00	0.00	0.00	0.00
0/1	0.33	0.00	0.33	0.00	0.33	0.00	0.00	0.00	0.00
0/2	0.00	0.50	0.00	0.00	0.00	0.50	0.00	0.00	0.00
1/0	0.33	0.00	0.00	0.00	0.33	0.00	0.33	0.00	0.00
1/1	0.00	0.25	0.00	0.25	0.00	0.25	0.00	0.25	0.00
1/2	0.00	0.00	0.33	0.00	0.33	0.00	0.00	0.00	0.33
2/0	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.50	0.00
2/1	0.00	0.00	0.00	0.00	0.33	0.00	0.33	0.00	0.33
2/2	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.50	0.00

Fig. 4. The segment of a map with nine nodes and the correspondant Markov model transition matrix.

Two important elements worthy to note from the transition matrix in figure 4, are that the transitions sets to zero, and the transitions are equally weighted. Transitions with a value of zero imply that it is not possible to create an immediate preceding chain between these transitions and therefore these can be ignored by the system. For instance, a car at the intersection 0/1 cannot move to the intersection 2/1 if it has not been at intersection 1/1 first, therefore, we would say that the transition from 0/1 to 2/1 is impossible. On the other hand, intersections with same weight implies that in this context a car can transit on any of the probably roads at the same speed and without any trouble. Thus, in order for the Markov model to predict escape routes it has to seek up into the transition matrix and based on intersection weights identify the probable routes of trajectories.

The previous scenario considers a city with roads that may have not traffic congestion, without accidents or where road maintenance is not needed. If on the contrary we live in a city with such kinds of events there is necessary to count on a route predictor that listens for these changes, update the transition matrix accordingly, and to count on efficient search algorithms. Once opportunistic events are detected the predictor must first reassign the adequate weights to the lanes ahead. To do that, the first task is the identification of the available roads that are allowed for vehicle traffic, which is done with the heuristic presented in figure 5:

```
Start (change event)
i, j  Are indices of the nodes in the transition matrix
      get s_node and d_node from loc
      For each i
            For each j
                  If s_node == i and d_node == j Then
                        S_i = S_i − R_i
                  EndIf
            EndFor
      EndFor
      Remove transition corresponding  s_node and d_node
End
```

Fig. 5. Pseudo-code that helps with the calculation of the new roads' weights given the fact that an accident is blocking any road.

For the pseudo-code presented in figure 5, consider that loc is the is the location where an accident occurred event, s_{node} and d_{node} correspond to the source node and the destination node on the location respectively, S_i is the total outgoing lanes of node i and R_i is the number of affected lanes node i to d_{node} .

Once the update of roads' weights is done, the re-calculation of the probabilities space can be done with either the pseudo-code presented in figure 2, first order Markov model, or the one presented in figure 3, second order Markov model.

4 Results

As described in previous sections the main component of the proposed system to support police chasing is the route predictor algorithm. In order to test the first and second order Markov chains implemented for route prediction (RP) four simulations were conducted: two RP_L (route prediction with location data) and two RP_LRO (route prediction with location data and random obstacles). Before dealing with the simulation results we first describe the setting up for these experiments.

4.1 Experiment set up

The first component required for testing the route predictor is a geo-referenced map. A map of a city in the United States was downloaded from www.openstreetmap.org. The map was manually configured, for instance, to not allow for U-turns and to remove traffic lights. These considerations are based on the run on escape context in which, we assume, a theft would not respect traffic lights and would never drive in the opposite direction over the same lane is being hunted. In addition, street categories were also properly configured, for instance, to differentiate the number of lanes that are part of a road and the maximum speed allowed for that road. For example, a road with category 1 has 4 lanes and an allowed speed of up to 110 km/hr. whereas a road with category 3 has one lane and an allowed speed of up to 50 km/hr. The lowest the category the highest the priority assigned to a car being driven on a road.

The second component that is part of the test experiment and that need some configurations is the simulation of the urban mobility platform, SUMO [8]. This tool can be configured with the number of cars that could travel around in the map, their direction, and their navigation speed among others. The car being hunted in our experiment, for instance, has an acceleration value of 2.9 m/s^2 and a deceleration value of 7.5 m/s^2, and the maximum speed it can reach is 250 km/hr. These parameters are important because these have a great impact with the decisions made by the simulator to create navigation trajectories for the car on the run. Another important configuration for the simulator is the random generation of obstacles on roads. These obstacles consist on the sudden presence of broken cars that block or alter the mobility on roads, which force the simulator to look up for alternate escape trajectories. Once the setup is done the next step is the evaluation of the route predictor system, see figure 6.

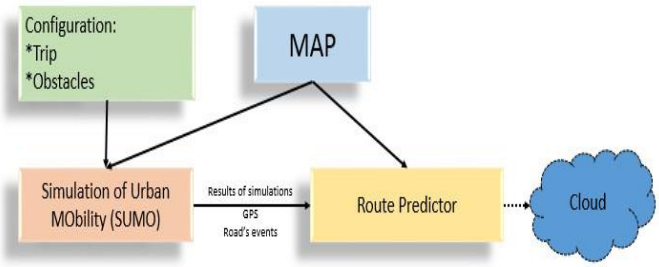


Fig. 6. Events synchronization between the simulation tool and the route predictor.

4.2 Route prediction evaluation

Four experiments were carried out in order to test the performance of the service that predicts potentially escapes routes. For the sake of space we discuss only the fourth of the experiments. For evaluation purposes only five square kilometers of the city of Denver in the US are considered, area that comprises 281 intersections and around 930 different roads, see figure 7.

a) RP_L: Calculates the probabilities for a car to take available roads based solely on a geo-referenced map. When running the RP_L experiment with a first order Markov chain the prediction route takes into account the GPS location of the car and the roads' priorities to seek up into the transition matrix and to predict the next road a car will probably take. Basically, the probability weight for every intersection at the vicinity of the lane used by the mobile node is 33.33%.

For the second experiment we considered the same road and city map configurations but applying a second order Markov model. The first thing noteworthy is the fact that for a second order Markov the state of previous transition is important, it is not possible for a car to return to the same road on which it is escaping, as it is the number of lanes offered by the routes ahead of the mobile node. The predictor offers three nodes for continuing escaping. The road with two lanes was assigned 50% of probability whereas the other two roads got 25% of the probability space each. The prediction hence is that the car would use the two lanes road for escaping.

Fig. 7. The geo-referenced map used for the route prediction model corresponds to five square kilometers of a US city.

b) RP_LRO: With a first order Markov the model calculates the probabilities for a car to take available escape routes based on a geo-referenced map and random presence of accidents on roads. The simulation tool is in charge of inserting the obstacles into the city map. An obstacle on the road means that it is not possible for a vehicle to get through it. Remember that the simulation tool reports the location of the car and any event occurring on a particular road. Based on that information the predictor must update the transition probabilities for the city map. From the figure 8 a), it is possible to identify that when the car reaches the node 178 there are four probable routes but the ones with the higher probability are the way out to node 125 (33%) or to node 177 (33%). However when the car is moving towards intersection 177 the predictor realizes there is a car's accident on node 55, therefore, indicating that there are three probable routes to continue the travel being the node 178 (50%) the best road for escaping.

The fourth experiment consisted on repeating the third experiment (RO_LRO) but using a second order Markov model. If there is not an obstacle the predictor offers three probable roads for continue driving. When the car travels on road 160 (labeled as edge 160), for instance, the predictor outcomes three probable routes, road 159, road 328 and road 793. However, when the car is being driven on road 159 (labeled as edge 160) the predictor offer only two probable routes because it is taken into account that the road 158 is blocked. Observe that because in this case both routes (road 265 and road 732) have the

same number of lanes and the allowed travel speed is also the same, the two of routes got the same probability of 50%, see figure 8 b).

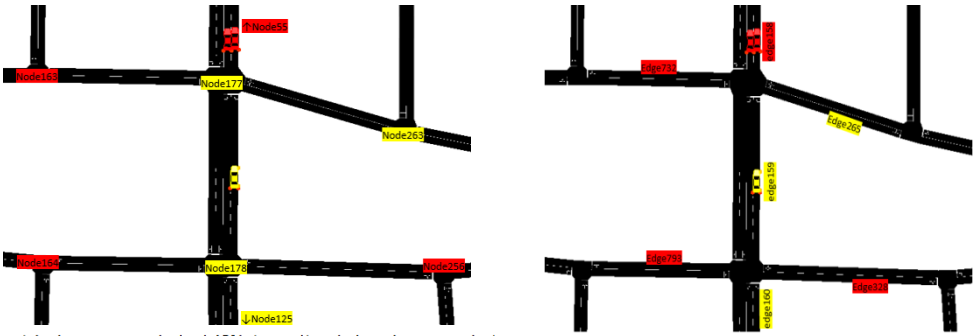


Fig. 8. First (a) and second (b) order Markov route prediction with the random presence of obstacles. The Markov simulation performance is graphically shown whereas the prediction outcomes are in text format.

5 Conclusions and Future Work

In this paper we have presented the application of Markov models for the context of police chase. The aim is to predict the probable escape routes that a criminal driving a car could use. If the leader patrol shares its location and prediction data with other patrol vehicles it might be possible to constraint the escape options by means of the creation of a virtual barrier, i.e. the strategic and coordinated colocation of the police vehicles at different roads and streets to limit or block flow traffic. A particular focus was given to demonstrate how the route prediction is done using first and second order Markov models. Two applications scenarios have been tested: streets with free traffic flow and streets with obstacles (cars' accidents). In any case it has been shown that the performance of the second order Markov model outperforms that of one of first order. A Markov model of second order takes into account previous locations, which provide a car's direction and help im-

prove the predictor performance. Although we are already considering opportunistic events that can occur on the streets, e.g. cars' accidents, there are others that need to be included in the route prediction model, e.g. traffic lights and road maintenance, in order to build a robust system that offers support to the police activities. For future development we have considered to increase the order of the Markov model and to integrate other probabilistic techniques such as Bayesian networks to improve the updating of weights for the Markov probability space.

Acknowledgment
The authors wish to express their gratitude to reviewers for their invaluable feedback. We would also like to thank the PROMEP for funding "Platform for the Experimentation with Mobile Technologies" (PEMT) project, at the Autonomous University of Chihuahua.

References

1. D. J. Patterson, L. Liao, D. Fox and H. Kautz, "Inferring High-Level Behavior from Low-Level Sensors," in UbiComp 2003: Ubiquitous Computing, 2003.
2. L. Liao, D. J. Patterson, D. Fox and H. Kautz, "Learning and inferring transportation routines," Artificial Intelligence, vol. 171, no. 5, pp. 311-331, 2007.
3. R. Simmons, B. Browning, Y. Zhang and V. Sadekar, "Learning to Predict Driver Route and Destination Intent," in Intelligent Transportation Systems Conference, 2006. ITSC'06. IEEE, 2006.
4. B. Mao, J. Cao, Z. Wu, G. Huang and J. Li, "Predicting driving direction with weighted Markov model," Springer Berlin Heidelberg, vol. 7713, pp. 407-418, 2012.
5. J. Krumm, "A Markov model for driver turn prediction," in SAE World Congress & Exhibition, Detroit, MI USA, 2008.
6. J. Krumm, "Where Will They Turn: Predicting Turn Proportions At Intersections," Personal and Ubiquitous Computing, vol. 14, no. 7, pp. 591-599, 2010.
7. U. Nagaraj and N. Kadam, "Vehicular Route Prediction In City Environment Based On Statistical Models," Journal of Engineering Research and Applications, vol. 3, no. 5, pp. 701-704, 2013.
8. Simulation of Urban MObility (SUMO). Homepage. http://www.sumo.dlr.de

Multi-objective Artificial Bee Colony Algorithm *with external memory* for the Optimization of Engineering Problems

Carlos A. Duchanoy[a] ⋆, Marco A. Moreno-Armendáriz[a], Carlos A. Cruz-Villar[b] and Hiram Calvo[a]

[a] *Instituto Politécnico Nacional, Centro de Investigación en Computación, Av. Juan de Dios Bátiz s/n, México D.F. 07738*
[b] *Sección de Mecatrónica, CINVESTAV, Av. Instituto Politécnico Nacional 2508, Apdo Postal 14740, México D.F., México*

Abstract. In this paper, a memory module for the Multi-Objective Artificial Bee Colony (MOABC) algorithm is proposed. The inclusion of a memory has the objective of reducing the number of evaluations required by the MOABC algorithm. The proposed memory stores the fitness values for every food source, thus avoiding to evaluate more than once these sources. As case study, we present a damper optimal design. With the objective of carrying out an in-depth analysis of the suspension system, the full behavior of the suspension is considered. This is accomplished using a full car model that includes the kinematics, dynamics and geometry of the suspension. However, simulation of the car model requires a lot of computing power which leads to long simulation times. The simulation times motivates the modifications made to the MOABC algorithm. As result of the modifications the number of evaluations needed by the MOABC is reduced to the half.

Keywords: Multi-Objective Artificial Bee Colony, Evolutionary optimization, Suspension System

1 Introduction

Several computational methods have been proposed to find the optimal solution to a problem. In [7] a comparison between methods such as Particle Swarm Optimization (PSO), Genetic Algorithms (GA), Differential Evolution (DE), Evolutionary Algorithms (EA) and Artificial Bee Colony algorithms (ABC) is presented. As conclusion the authors determine that the ABC algorithm

⋆ Corresponding author. Email: duchduchanoy@gmail.com

performs better than the mentioned algorithms and can be efficiently employed to solve multimodal engineering problems with high dimensionality. More studies on the ABC algorithm behavior are enlisted in [8], which is a survey of the ABC algorithm an its applications.

Optimization problems with more than one criterion to be considered are common in many engineering problems. The goal is to find a set of solutions that represents a tradeoff among these criteria. These type of problems are called multi-objective and some authors have proposed modifications to the artificial bee colony algorithm (ABC) in order to solve them. The original ABC algorithm was modified for solving multi-objective problems in [6]. This algorithm is Pareto based and was named as Multi-Objective Artificial Bee Colony (MOABC). A multi-objective ABC algorithm for scheduling problems is introduced in [2]. Finally an improved version of the MOABC algorithm is presented in [1].

Of all the above-mentioned algorithms, which has been applied more to solve the different problems of engineering is the MOABC algorithm [[3] to [7]]. Other example is [5], where the authors proposed the application of a MOABC to solve the motif discovery problem in DNA sequences. Also, in [3] a MOABC for optimization of power and heating system is introduced.

2 Multi-Objective Artificial Bee Colony Algorithm *with external memory*

In this section, the proposed Multi-Objective Artificial Bee Colony Algorithm *with external memory* (MOABCM) is explained in detail. It represents the foraging behavior of honey bees.The colony has three types of bees: employed, onlooker and scout. This algorithm assumes a fixed number of food sources; each food source represents a solution to the problem, these food sources have been found by employed bees and will be presented in the dance area. The onlooker bees wait on the dance area to choose a food source by its quality. When the food source could not be improved after some cycles, a scout bee will do a random search in order to find a new food source.

The MOABCM is a Pareto based algorithm with an external archive used to store non-dominated solutions and a memory that stores the aptitude values of each food source, in order to reduce the number of evaluations of the objective function. This algorithm requires as parameters: Population size, Max-trial, maximum number of iterations. Population size is the number of food sources, employed and onlooker bees, Max-trial is the value that the algorithm uses to find out which food sources should be abandoned. The maximum number of iterations indicates when the algorithm must be stopped (end condition).

2.1 Proposed memory module

The memory has the objective of eliminating the need of evaluating twice a food source. As a result, the number of evaluations of the objective function is considerably reduced. The novel memory stores: All the food parameters (vector ϕ for our case study), the fitness of all the objectives and the number of food sources dominated by each solution. In order to simplify the code this memory is divided into three smaller ones: The first one stores the parameters of all the food sources, it is named $food$, the second stores the fitness for all the objectives and is named *apt* and the last one stores the number of food sources dominated by each solution, this one is called *Dom*. A schematic of the memory is shown in figure 1.

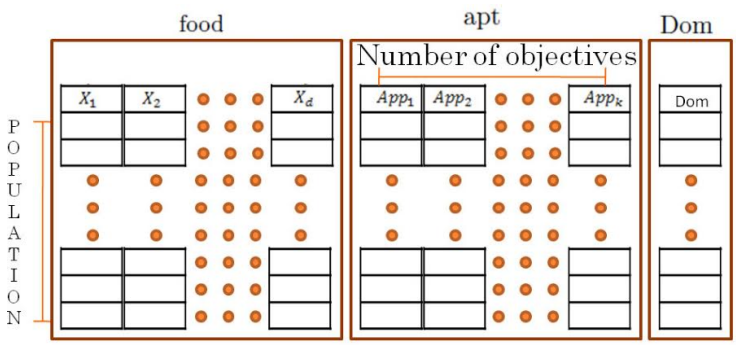


Fig. 1: Schematic of the proposed memory module for MOABCM.

2.2 Update archive

The MOABCM algorithm uses a different structure for the external archive. The new archive holds the best solutions ever found, saving the food parameters in a entity called Non-dom and fitness of the non dominated solutions in a entity called *arch*. The pseudocode for updating the archive is given in algorithm 1; the algorithm makes use of the Dominate function that calculates the Pareto dominance for the given data. This function uses the food memory and the aptitude matrix (apt) and it returns the non-dominated data matrix (Non−dom), their aptitude matrix (arch), and a domination matrix (Dom) in which the program indicates which solution dominates whom; then it saves this information into an external archive.

The update-archive function proposed in this paper also differs from the one proposed in [1] because in the original MOABC the update archive function

Algorithm 1 [arch, Non−dom, Dom]=update−archive(food, apt)

1: [arch, Non−dom, Dom]**Dominate**(apt, food)
2: **Save**(arch, Non−dom)

performs the evaluation of all the food sources while in the MOABCM, the update archive function does not perform evaluations of the fitness function. Instead, the fitness values stored in the memory are used.

2.3 Initialization

In the initialization phase, a new matrix named *food* is defined to preserve the possible solutions. The pseudocode of the initialization module is given in Algorithm 2. It uses the same parameters as the original MOABC. The initialization algorithm uses the population number which is equal to the number of food sources, it also needs the dimension of the problem which is the number of parameters that will be optimized. In this function Lev_{down} and Lev_{up} are the lower and upper bound matrices respectively.

Algorithm 2 [food, trial]=Initialization(dimension, population)

1: **For** i=1:dimension
2: **For** j=1:population
3: $Food(j,i) = Lev_{down}(i) + Rand()(Lev_{up}(i) - Lev_{down}(i))$
4: $trial(i) = 0$
5: **End For**
6: **End For**

2.4 Evaluate

The evaluate function calculates the aptitude of each food source using the fitness function of the problem. This function returns the aptitude value and stores it in the proposed memory, the food memory and the population number.

2.5 Employed−bee

The pseudocode of Employed−Bee function is given in Algorithm 3. The employed bee function improves food sources using the external archive since it contains the best solutions found so far. It is similar to the one proposed in the MOABC algorithm, but differs in line 8 because in order to make the comparison of the original food source with the updated one, it calculates the fitness of both

sources and compares them. The proposed MOABCM instead of evaluating the fitness of the original food source, uses the value stored in the *apt* memory and compares it with the fitness of the updated food source.

Algorithm 3 [food, apt, trial]=Employed−bee(food, apt, trial, population, dimension, Non−dom)

For i=1:population
  D =**Randi**($dimension$)
  N =**Randi**(Non−dom)
  $v = food(i, D) + w_1$(**Rand**()($food(i, D)$−Non−dom(N, D))
  $test = food(i, :)$
  $test(D) = v$
  [apd] =**Evaluate**($test, 1$)
  If apd is better than $apt(i)$
    $food(i, :) = test$
    $apt(i) = apd$
  Else
    $trial(i) = trial(i) + 1$
  End If
End For

The Employed−bee function uses the **Randi**(k) function, this returns a random integer between 0 and $k > 0$ and uses **Rand**() function which returns a random number between 0 and 1.

2.6 Onlooker−bee

After all employed bees optimized their food source by considering the information provided by the external archive, they will share their information with onlooker bees which optimize food sources advertised by employed bees. The pseudocode of the Onlooker−bee module is given in Algorithm 4.

In order to choose which food source advertised by the employed bees will be optimized, the algorithm uses a probabilistic method. For this purpose, the probability for each food source advertised by the corresponding Employed−bee will be calculated with equation (1).

$$P_k = \frac{fit(x_k)}{\sum_{m=1}^{population} fit(x_m)} \tag{1}$$

Where $fit(x_k)$ is the probability of the food source proposed by Employed−bee k which is proportional to the quality of food source. In MOABC, the quality

Algorithm 4 [food, apt, trial]=Onlooker−bee(food, apt, trial, population, dimension, Dom)

1: **For** i=1:population
2: D =**Randi**($dimension$)
3: N =**Roulette**(Dom)
4: $v = food(i, D) + w_1(\textbf{Rand}()(food(i, D) - food(N, D))$
5: $test = food(i, :)$
6: $test(D) = v$
7: $[apd]$ =**Evaluate**($test, 1$)
8: **If** apd is better than $apt(i)$
9: $food(i, :) = test$
10: $apt(i) = apd$
11: **Else**
12: $trial(i) = trial(i) + 1$
13: **End If**
14: **End For**

of a food source depends on the number of food sources dominated by the food source k. This can be formulated using the equation (2).

$$fit(x_k) = \frac{Dom(k)}{population} \tag{2}$$

Where $Dom(m)$ is the number of food sources dominated by food source m. After the probabilities are calculated, the onlooker bees use roulette wheel to chose the food source. After that, each Onlooker−bee optimizes its food source.

The pseudocode of Algorithm 4 is different from the one proposed in MOABC in two parts: the first one is that in line 3, the MOABC in order to calculate the probability of the food source, evaluates the fitness of all food sources and the MOABCM uses the values stored in the Dom memory. The second change is in line 8, the proposed MOABCM instead of evaluating the fitness of the original food source uses the value stored in the apt memory and compares it with the fitness of the updated food source.

2.7 Scout−bee

The pseudocode of Scout−bee module is given in Algorithm 5. This algorithm will find if an employed be should abandon his food source and replace it with a new one. In the MOABCM line 4 of the Algorithm 5, is added in order to save the fitness of the new food source.

Algorithm 5 [food, apt, trial]=Scout−bee(food, apt, trial, population, dimension)

1: **For** i=1:population
2: **If**$(trial(i) >= maxtrial)$
3: $[food(i), trial(i)]$ =**Initialization**$(1, dimension)$
4: $[apt(i)]$ =**Evaluate**$(food(i, 1))$
5: **End If**
6: **End For**

3 Comparison between algorithms

The MOABCM algorithm changes were proposed with the purpose of reducing the number of evaluations needed by the algorithm with very similar performance. Table 1 presents a comparison of the number of needed evaluations in the MOABC and in the MOABCM. This table describes the evaluations made by each module and in the iteration row it enlists the number of evaluations per iteration.

Module	MOABC [1]	MOABCM
update-archive	Population	0
Employed-bee	2 Population	Population
Onlooker-bee	3 Population	Population
Scout-bee	0	0≤x≤Population
Evaluations per iteration	6 Population	max 3 Population

Table 1: Number of evaluations of the fitness function

It is necessary to consider that the number of evaluations made in the scout-bee can be a number between zero and the population number because it only evaluates new food sources. As shown in the last row of Table 1 the MOABCM algorithm in the worst case scenario has to evaluate three times the population while the original MOABC needs six times population evaluations per iteration. This is a considerable improvement because the number of evaluations is reduced to a half.

4 Case study: Optimal Suspension Problem

The suspension system performs two main functions. The first one is to guarantee the correct contact of tires during a travel and its second function is to isolate the passengers from road disturbances. During the design process, these

objectives are in conflict. For this reason, the objective is to design a damper that provides the best compromise between these objectives. In order to perform the optimization process, a complex mathematical model of the vehicle is required. For our case, we select the full car model developed by our research group in [4]. This full car model consists of the front suspension, rear suspension, the tires model and their interaction with the chassis.

The mathematical model is programmed in Matlab Simulink© in order to simulate the behavior of the vehicle. A terrain in which the performance of the suspension system will be measured, also was simulated. The terrain selected for this experiment is conformed by a section of caps, see figure 2.

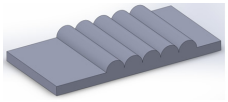


Fig. 2: Schematic diagram of the caps terrain.

The MOABCM algorithm uses the evaluate function to calculate the fitness of the possible solutions. Each time the function calls the simulator and waits for the result. As result the simulator generates a plot of the displacement of the cockpit (x_c) and four plots of contact area of the tires, one for each tire ($A_{ct}(fr), A_{ct}(fl), A_{ct}(rr), A_{ct}(rl)$). The results are used to calculate the fitness of the solutions using equations (3 and 4). The objective of this case study is to design a passive damper which improves the performance of the suspension in two factors: vehicle's safety and comfort. So, we define safety and comfort as follows: it should be understood by comfort C, a reduction in the amplitude of the cockpit displacement and as safety S, an increase of the contact area of the tires. The problem is formulated as the minimization of the chassis displacement and the maximization of the contact area of all tires. The problem is expressed in equations (3) to (9).

$$\min_{\text{under } \phi} C = \int_0^T L_1(x(t), z(t), \theta, \phi, t)dt \tag{3}$$

$$\max_{\text{under } \phi} S = \int_0^T L_2(x(t), z(t), \theta, \phi, t)dt \tag{4}$$

$$\phi = [K_d(fr), K_d(fl), K_d(rr), K_d(rl), B_d(fr), B_d(fl), B_d(rr), B_d(rl)] \tag{5}$$

Subject to:

$$\dot{x} = f(t, x(t), z(t), \theta, \phi) \tag{6}$$
$$0 = g(t, x(t), z(t), \theta, \phi) \tag{7}$$

Where:

$$L_1(x(t), z(t), \theta, \phi, t) = z_c \tag{8}$$
$$L_1(x(t), z(t), \theta, \phi, t) = A_{ct}(fr) + A_{ct}(fl) + A_{ct}(rr) + A_{ct}(rl) \tag{9}$$

We propose as design variables the stiffness of the front suspension damper $K_d(fr), K_d(fl)$, the damping of the front suspension damper $B_d(fr), B_d(fl)$, the stiffness of the rear suspension $K_d(rr), K_d(rl)$ and the damping of the rear suspension damper $B_d(rr), B_d(rl)$. The design variables vector ϕ is expressed in equation (5).

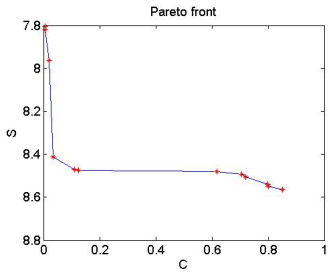
5 Experiments and results

The optimization of the passive dampers for the suspension was made trough the modification of the design variables vector ϕ (5), applying the algorithm described in section 2. The method was performed using the following parameters: Population=60, max-trial=10, max-iterations=40, and $w_1 = 0.8$ (an explanation about how to select this values is founded in [4]). The execution of the algorithm spent around of 30 hrs, with a total of 1420 evaluations of the fitness function and used half of time of the original MOABC (on average, obtained by running the experiment 10 times). The Pareto front obtained as solution for this problem is shown in figure 3a. The algorithm finds 14 possible solutions for the problem that for a numeric optimization are very few, but for our case of study are good enough since they provide a good set of different behaviours to test the optimized suspension.

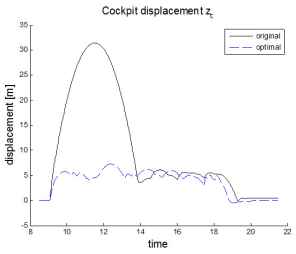
As an example, we select a solution on the middle of the Pareto front (best compromise between comfort C and safety S) and the results are shown in in figures 3b to 3d. Notice that the optimal solution improves the cockpit displacement and reduces the time where the tires do not have contact with the ground.

6 Conclusion and Future Work

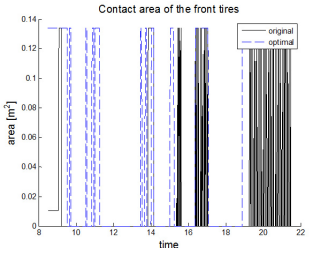
The model proposed for the vehicle is able to more accurately represent its behavior. It is important to highlight that more detailed models consume longer simulation times, which motivates the proposed modifications to the MOABC algorithm. To solve this problem, we present the MOABCM algorithm which is



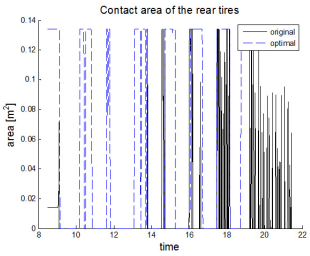
(a) Pareto front.



(b) Cockpit displacement.



(c) Contact area of the front tires.



(d) Contact area of the rear tires.

Fig. 3: Dynamic behavior of the Vehicle.

capable to reduce the number of evaluations of the fitness function. This in turn helps reducing the optimization time to less than half of the time needed by the original MOABC algorithm.
To test the performance of the algorithm a case study of a passive suspension optimization was presented. As shown in Section 5 we could appreciate that the behavior of the two parameters were optimized by the proposed methodology.
As future work remains the construction and testing of the suspension designed with this methodology.

Acknowledgments The authors would like to thank the team members of UPIITA SAE Baja 2009 who made the construction and design of the vehicle in which this work was based. The authors appreciate the support of Mexican Government (SNI, SIP-IPN, COFAA-IPN, BEIFI-IPN and CONACyT).

References

1. Akbari, R., Hedayatzadeh, R., Ziarati, K., Hassanizadeh, B.: A multi-objective artificial bee colony algorithm. Swarm and Evolutionary Computation 2, 39–52 (Feb

2011)
2. Arsuaga-Rios, M., Vega-Rodriguez, M.a., Prieto-Castrillo, F.: Multi-Objective Artificial Bee Colony for scheduling in Grid environments. In: 2011 IEEE Symposium on Swarm Intelligence. pp. 1–7. IEEE (Apr 2011)
3. Atashkari, K., NarimanZadeh, N., Ghavimi, A., Mahmoodabadi, M.J., Aghaienezhad, F.: Multi-objective Optimization of Power and Heating System Based on Artificial Bee Colony. In: International sympo- sium on innovations in intelligent systems and applications (INISTA). pp. 64–68 (2011)
4. Duchanoy, C.A., Moreno Armendaríz, M.A., Cruz-villar, C.A.: Nonlinear Full-car Model for Optimal Dynamic Design of an Automotive Damper. In: Multibody Mechatronic Systems, pp. 489–500. Springer International Publishing (2015)
5. González-Álvarez, D.L., Vega-Rodríguez, M.A., Gómez-Pulido, J.A., Sánchez-Pérez, J.M.: Finding Motifs in DNA Sequences Applying a Multiobjective Artificial Bee Colony (MOABC). Evolutionary computation, machine learning and data mining in bioinformatics. Lecture notes in computer science, 6623, 89–100 (2011)
6. Hedayatzadeh, R., Hasanizadeh, B., Akbari, R., Ziarati, K.: A multi-objective Artificial Bee Colony for optimizing multi-objective problems. In: 3rd International Conference on Advanced Computer Theory and Engineering(ICACTE). vol. 2, pp. 277–281. Ieee (Aug 2010)
7. Karaboga, D., Basturk, B.: On the performance of artificial bee colony (ABC) algorithm. Applied Soft Computing 8(1), 687–697 (Jan 2008)
8. Karaboga, D., Gorkemli, B., Ozturk, C., Karaboga, N.: A comprehensive survey: artificial bee colony (ABC) algorithm and applications. Artificial Intelligence Review pp. 1–37 (Mar 2012)

Gamma classifier based instance selection

[1]Jarvin A. Antón Vargas, [1]Yenny Villuendas-Rey, [2]Itzamá López-Yáñez

[1]Departamento de Ciencias Informáticas, Universidad de Ciego de Ávila, Cuba
[2] Centro de Innovación y Desarrollo Tecnológico en Cómputo, Instituto Politécnico Nacional, México
janton@unica.cu; yenny@cav.uci.cu; itzama@gmail.com

Abstract. Filtering noisy and mislabeled instances is an important problem in Pattern Recognition. This paper introduces a novel similarity function based on the Gamma operator, and use it for instance selection for improving the Gamma classifier. Numerical experiments over repository databases show the high quality performance of the proposal according to classifier accuracy and instance retention ratio.

Keywords: Gamma Classifier, Instance Selection, Supervised Classification.

1 Introduction

The training dataset plays a key role for supervised classification. Training data allows building classifiers able to estimate the label or class of a new unseeing instance. Several researchers have pointed out that if the dataset has high quality instances, the classifier can produce predictions that are more accurate [1]. However, in several real-world applications, it is not possible to obtain a training set without noising and mislabeled instances. To overcome this problem, several instance selection and generation algorithms have been proposed [1, 2].

The Gamma classifier [3, 4] is a recently proposed supervised classifier, ant it has been applied successfully to several prediction tasks, such as air quality [5], pollutant time series [6] and development effort prediction of software projects [7]. Despite of the excellent performance of the Gamma classifier, it is noted that it is affected by noisy or mislabeled instances.

However, most of the instance selection algorithms are designed for the Nearest Neighbor (NN) classifier [8]. Little work has been done for selecting instance to other supervised classifiers, such as ALVOT [9, 10] and Neural Networks [11], and there are no directly applicable to the Gamma classifier.

This paper proposes a similarity function based on the Gamma operator of the Gamma classifier, and use it for similarity comparisons in NN instance selection algorithms. The

thorough experimental study carried out shows the significant performance gains of the proposed approach.

2 Gamma Classifier and Gamma Based Similarity

The Gamma classifier is based on two operators named Alpha and Beta, which are the foundation of the Alpha-Beta associative memories [12]. The Alpha and Beta operators are defined in a tabular form considering the sets $A = \{0, 1\}$ and $B = \{0, 1, 2\}$, as shown in figure 1.

$\alpha : A \times A \rightarrow B$		
x	y	$\alpha(x, y)$
0	0	1
0	1	0
1	0	2
1	1	1

$\beta : B \times A \rightarrow A$		
x	y	$\beta(x, y)$
0	0	0
0	1	0
1	0	0
1	1	1
2	0	1
2	1	1

Fig. 1. Operators Alpha and Beta.

In addition to the Alpha and Beta operator, the Gamma classifier also uses two other operators: the u_β operator and the generalized gamma similarity operator, γ_g. The unary operator u_β receives as an input a binary n-dimensional vector, and returns a number $\mathrm{p} \in \mathbb{Z}^+$ according to the following expression:

$$u_\beta = \sum_{i=1}^{n} \beta\,(x_i, x_i) \tag{1}$$

The generalized gamma similarity operator receives as input two binary vectors x and y and also a non-negative integer θ, and returns a binary digit, as follows:

$$\gamma_g(x, y, \theta) = \begin{cases} 1 & if\ m - u_\beta[\alpha(x, y) mod 2] \leq \theta \\ 0 & otherwise \end{cases} \tag{2}$$

That is, the γ_g operator returns 1 if the input vectors differentiates at most in θ bits, and returns zero otherwise.

The Gamma classifier is designed for numeric patterns, and assumes that each pattern belongs to a single class. However, as the generalized gamma similarity operator receives as input two binary vectors, the Gamma classifier codifies numeric instances using a modified Johnson-Möbius code [3]. In figure 2 we show a simplified schema of the Gamma classifier.

According to the classification strategy of the Gamma classifier, we propose a similarity function to compare pairs of instances, regarding the θ parameter. This allows us to detect noisy or mislabeled instances.

The proposed Gamma based similarity (GBS) uses the generalized gamma operator, but it considers the standard deviation of the feature instead of the θ parameter. Let be X and Y to instances, the Gamma based similarity between them is computed as:

$$GBS(X,Y) = \sum_{i=1}^{p} \gamma_g\,(x_i, y_i, \sigma_i) \tag{3}$$

where p is the amount of features describing the instances, σ_i is the standard deviation of the i-th feature, and x_i and y_i are the binary vectors associated with the i-th feature in instances X and Y, respectively.

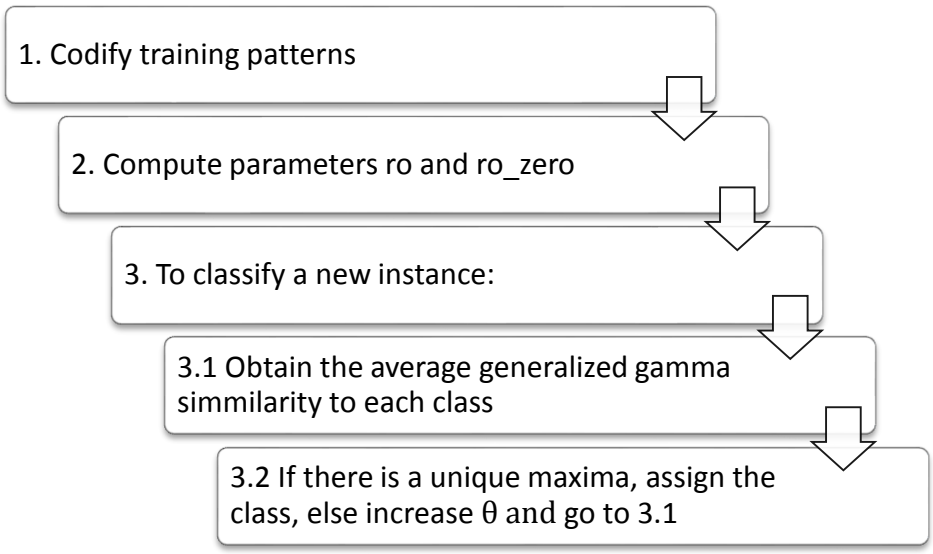


Fig. 2. Simplified schema of the classification process with the Gamma classifier.

Considering this novel similarity, we are able to apply several instance selection algorithms which were designed for the Nearest Neighbor classifier, and test its performance in the filtering of noisy and mislabeled instances for the Gamma classifier.

3 Experimental Results

We select some of the most representative instance selection algorithms and perform the test over seven databases from the Machine Learning repository of the University of California at Irvine [13]. Table 1 shows the characteristics of the selected databases.

We selected error-based editing methods due to their ability of smoothing decision boundaries and to improve classifier accuracy. The selected methods are the Edited Nearest Neighbor (ENN) proposed by Wilson [14], the Gabriel Graph Editing method (GGE) proposed by Toussaint [15] and the MSEditB method, proposed by García-Borroto et al. [16].

The ENN algorithm (Edited Nearest Neighbor) is the first error-based editing method reported [14]. It was proposed by Wilson in 1972 and it consist on the elimination of the objects misclassified by a 3-NN classifier. The ENN works by lots, because it flags the misclassified instances and then simultaneously deltetes them all, which guaranteed order independence. The ENN has been extensively used in experimental comparisons, showing very good performance [1].

Table 1. Databases used in the experiments.

Databases	Objects	Attributes	Classes
balance-scale	625	4	3
breast-w	699	9	2
ecoli	336	7	8
heart-statlog	270	13	2
ionosphere	351	34	2
iris	150	4	3
vehicle	846	18	4

The GGE algorithm is based on the construction of a Gabriel graph. A Gabriel graph is a directed graph such that two instances $x \in U$ and $y \in U$ form an arc if and only if $\forall z \in U\ (d((x + y/2), z) > d(x, y)/2)$, where d is a dissimilarity function. That is, two instances x and y are related in a Gabriel graph if there is no object in the hypersphere centered in the middle point of x and y, and with radius the distance between x and y.

The GGE algorithm consist in deleting those instances connected to others of different class labels. It deletes borderline instances, and keep class representatives ones.

The MSEditB algorithm [16] uses a Maximum similarity graph to select the objects to delete. A Maximum similarity graph is a directed graph such that each instance is connected to its most similar instances. Formally, let be S a similarity function, an instance $x \in U$ form

an arc in a Maximum similarity graph with an instance $y \in U$ if and only if $d(x,y) = \max_{z \in U} d(x,z)$.

The MSEditB algorithm deletes an instance if it has a majority of its predecessors and successors instances not of its class.

All algorithms were implemented in C# language, and the experiments were carried out in a laptop with 3.0GB of RAM and Intel Core i5 processor with 2.67HZ. We cannot evaluate the computational time of the algorithms, because the computer was not exclusively dedicated to the execution of the experiments.

To compare the performance of the algorithms, it was used the classifier accuracy. The classifier accuracy is measure as the ratio of correctly classified instances. It was also computed the Instance retention ratio (IRR) for every algorithm, in order to determine the amount of selected instances. Table 2 and 4 show the results according to classifier accuracy and instance retention ratio, respectively.

In table 2, we show the accuracy of the Gamma classifier without selecting instances (Gamma) and the accuracy of the Gamma classifier trained using the instances selected by ENN, GGE and MSEditB, respectively.

Table 2. Accuracy of the gamma classifier before and after the selection of instances.

Databases	Gamma	Instances selected by		
		ENN	GGE	MSEEditB
balance-scale	0.838	0.598	0.810	**0.891**
breast-w	0.907	**0.908**	**0.908**	**0.908**
ecoli	**0.708**	0.560	0.474	0.536
heart-statlog	0.819	**0.833**	0.826	0.830
ionosphere	**0.749**	**0.749**	**0.749**	**0.749**
iris	**0.913**	0.813	0.907	0.867
vehicle	0.573	0.576	**0.582**	0.568

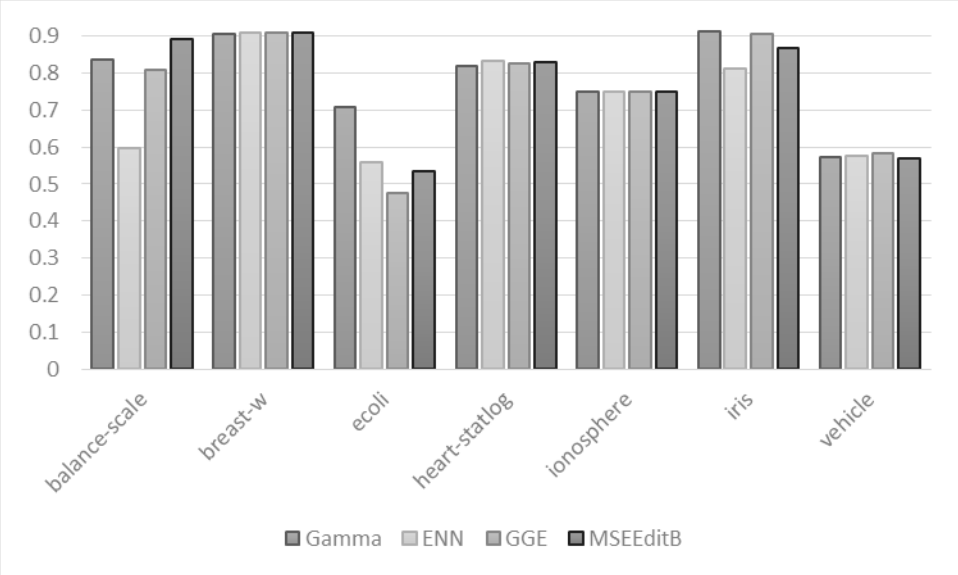


Fig. 3. Accuracy of the Gamma classifier using selected instances.

As shown, the instance selection algorithms were able to improve the Gamma classifier accuracy in four databases, and to obtain the same accuracy with fewer instances in one database. Still, for the ecoli and iris datasets, no improvement were obtained.

However, to determine the existence or not of significant differences in algorithm´s performance it was used the Wilcoxon test [17]. It was set as null hypothesis no difference in performance between the gamma classifier without instance selection (Gamma) and the gamma classifier with instance selection algorithms, and as alternative hypothesis that latter had better performance. It was set a significant value of 0.05, for a 95% of confidence. Table 3 summarizes the results of the Wilcoxon test, according to classifier accuracy.

Table 3. Wilcoxon test comparing classifier accuracy.

Original Gamma vs	**ENN**	**GGE**	**MSEditB**
wins-looses-ties	3-3-1	3-3-1	3-3-1
probability	0.345	0.600	0.735

The Wilcoxon test obtains probability values greater than the significance level, and thus, we do not reject the null hypothesis. These results confirm the proposed approach is able to preserve classifier accuracy, using a small amount of instances.

Table 4. Instance retention ratio obtained by the selection of instances.

Databases	ENN	GGE	MSEEditB
balance-scale	0.799	**0.723**	0.837
breast-w	0.973	0.968	**0.963**
ecoli	0.812	0.884	**0.792**
heart-statlog	0.889	**0.712**	0.807
ionosphere	0.915	**0.842**	0.854
iris	0.726	**0.689**	0.852
vehicle	0.659	0.751	**0.541**

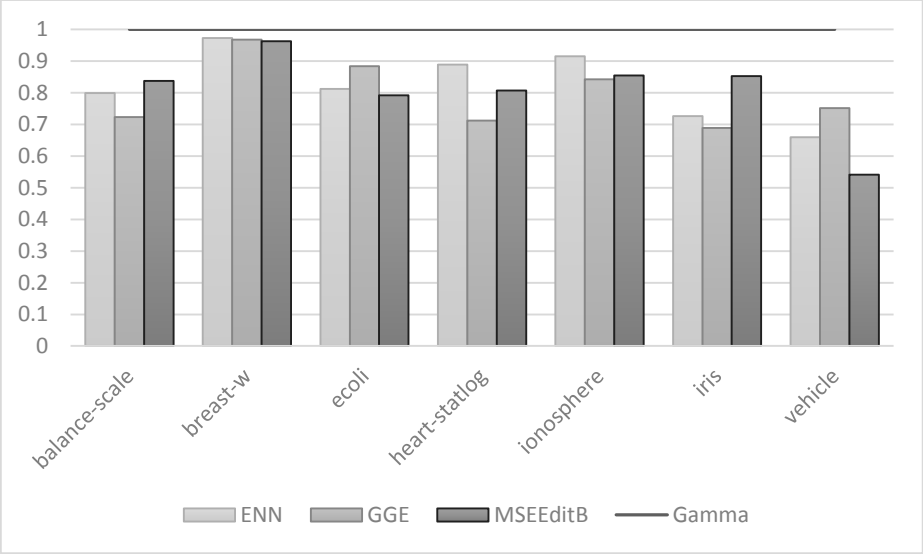


Fig. 4. Instance retention ratio obtained by the algorithms.

As shown in table 4, all instance selection methods are able to delete among the 40% and 4% of the data, without decreasing the classifier accuracy. These results confirm the proposed approach is able to obtain an adequate training set for the Gamma classifier, without losing representative objects.

Table 5. Wilcoxon test comparing instance retention ratio.

Original Gamma vs	**ENN**	**GGE**	**MSEditB**
wins-looses-ties	0-7-0	0-7-0	0-7-0
probability	**0.018**	**0.018**	**0.018**

According to instance retention ratio, the Wilcoxon test rejects the null hypothesis in all cases. That is, the number of selected objects using ENN, GGE and MSEditB with the proposed gamma based similarity function, was significantly lower than the original amount of instances in the training set.

The experimental results carried out show that selecting instances by using a similarity function based on the Gamma operator maintains classifier accuracy, and also reduces the cardinality of the training sets, diminishing the computational cost of the Gamma classifier.

4 Conclusions

Instance selection is an important preprocessing step for learning with most supervised classifiers. In this paper, a novel similarity measure is introduced, based on the Gamma operator of the Gamma classifier. We used the proposed similarity to select relevant instances for this classifier. Experimental results carried out over several repository data show that using the proposed similarity function for instance selection preserves classifier accuracy, and decreases the computational cost of the Gamma classifier.

References

1. S. García, J. Derrac, J. R. Cano, and F. Herrera, "Prototype Selection for Nearest Neighbor Classification: Taxonomy and Empirical Study," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, pp. 417-435, 2012.
2. I. Triguero, J. Derrac, S. Garcia, and F. Herrera, "A taxonomy and experimental study on prototype generation for nearest neighbor classification," Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, vol. 42, pp. 86-100, 2012.
3. I. López Yáñez. Clasificador automático de alto desempeño (MS dissertation, Instituto Politécnico Nacional-Centro de Investigación en Computación), 2007.
4. I. López-Yáñez, L. Sheremetov and C. Yáñez-Márquez, "A novel associative model for time series data mining," Pattern recognition Letters, vol 41, pp. 23-33, 2014.
5. C. Yánez-Márquez, I. López-Yánez and G.D. Morales, "Analysis and prediction of air quality data with the gamma classifier," Progress in Pattern Recognition, Image Analysis and Applications, pp. 651-658, 2008.
6. I. Lopez-Yanez, A.J. Argüelles-Cruz, O. Camacho-Nieto and C. Yanez-Marquez, "Pollutants time-series prediction using the Gamma classifier," International Journal of Computational Intelligence Systems, vol 4, no 4, pp. 680-711, 2012.
7. C. López-Martın, I. López-Yánez and C. Yánez-Márquez, "Application of Gamma Classifier to Development Effort Prediction of Software Projects," Appl. Math, vol 6 no 3, pp. 411-418, 2012.
8. T. M. Cover and P. E. Hart, "Nearest Neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, pp. 21-27, 1967.

9. M. A. Medina-Pérez, M. García-Borroto, Y. Villuendas-Rey and J. Ruiz-Shulcloper, “Selecting objects for ALVOT,” Progress in Pattern Recognition, Image Analysis and Applications, pp. 606-613, 2006.
10. M.A. Medina-Pérez, M. García-Borroto and J. Ruiz-Shulcloper, “Object selection based on subclass error correcting for ALVOT,” Progress in Pattern Recognition, Image Analysis and Applications, pp. 496-505, 2007.
11. H. Ishibuchi, T. Nakashima and M. Nii, “Learning of neural networks with GA-based instance selection,” IFSA World Congress and 20th NAFIPS International Conference, vol. 4, pp. 2102-2107, 2001.
12. C. Yáñez-Márquez and J. Luis Díaz de L., “Memorias Asociativas basadas en relaciones de orden y operaciones binarias,” Computación y Sistemas vol 6 no 4, pp. 300 - 311, ISSN 1405-5546, Centro de Investigación en Computación – Instituto Politécnico Nacional, México 2003.
13. A. Asuncion and D. Newman, UCI machine learning repository, 2007.
14. D. L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data," IEEE Transactions on Systems, Man and Cybernetics, vol. SMC-2, pp. 408-421, 1972.
15. G. T. Toussaint, "Proximity Graphs for Nearest Neighbor Decision Rules: Recent Progress," in 34 Symposium on Computing and Statistics INTERFACE-2002, Montreal, Canada, 2002, pp. 1-20.
16. M. García-Borroto, Y. Villuendas-Rey, J. A. Carrasco-Ochoa, and J. F. Martinez Trinidad, "Using Maximum Similarity Graphs to edit nearest neighbor classifiers," Lecture Notes on Computer Science, vol. 5856, pp. 489-496, 2009.
17. J. Demsar, "Statistical comparison of classifiers over multiple datasets," The Journal of Machine Learning Research, vol. 7, pp. 1-30, 2006.

New unsupervised classification scheme for datasets with presence of noise

José Felix Cabrera Venegas, Jarvin Antón Vargas, Yenny Villuendas Rey, Cornelio Yáñez-Márquez

Abstract. In this work an unsupervised classification algorithm, capable of handling data sets that have noisy elements is proposed. The proposed algorithm is based on a strategy to eliminate noise and then it applies a hierarchical agglomerative algorithm to obtain the groups. To determine the performance of the proposed algorithm, numeric experiments were performed and the quality of the proposal with respect to other algorithms were analyzed by using synthetic data sets developed specifically for this purpose; and also numerical databases from the Machine Learning Repository of the University of California at Irvine. The experimental results show that the new proposal has a stable performance and a good performance in noise elimination.

Keywords: Unsupervised clasification, noise detection.

1 Introduction

In recent years Pattern Recognition (PR) has gained some popularity thanks to automation of solutions to many problems of real life. In this area an important role is played by the Unsupervised Classification (UC) methods. The purpose of UC techniques is to reveal the structure of a set of data by creating groups of objects (clusters), associating each object with their most similar so that the objects in a cluster are different from the objects of other clusters [1].

In the literature, many UC algorithms have been proposed, but most of these attempts to group the data taking into account the presence or absence of certain features in the database. Many of the existing databases contains instances with rare or erratic behavior and many clustering algorithms are sensitive to such data, generating groups of poor quality. A noisy data (outlier) is an observation that deviate much from other observations and encourages suppose that was generated by different mechanisms [2]. Although there are proposals address this problem, many of them still exhibit the same deficiencies that must be settled.

Many UC algorithms try to minimize the influence of outliers or remove them all together, however this may result in a significant loss of information in certain situations. Fraudulent transactions such as credit card outliers are typical examples that may indicate fraudulent activity to be detected. That is why many applications of RP focus on the detection of noise, which is the main

result of the analysis of the data [3, 4] and not as such the classification of all data, on the other hand there are UC techniques engaged in the labeling of objects into groups and allow dealing with noise marking also with a label [5].

The state of the art techniques for the detection of noise is classified into several categories: distribution based techniques, based on depth, based on distances, based on density and based on groups, etc. [3]. Methods based on distribution and depth needs to adjust the data to statistical distributions or to assign depth to the data in a space of representation; and both become inefficient with the growing number of data. For these reasons there have been proposed methods that belong to other categories.

In [6], Knorr and Ng first proposed the notion of noise based on distance, which is a non-local approach and cannot find local outliers in databases with complex structure. Later Breuning et al. [7] presented the local outliers detection method based on density, Local Outlier Factor (LOF), supported by the distance of a point to their k nearest neighbor, declaring the n greater distance points as outliers. These proposals have the disadvantage of requiring the number of outliers to identify.

In addition to the actual noise detection techniques, there exist UC algorithms that allow detection of noise as a consecuence of the clustering process, as in the case of DBSCAN proposed by Ester et al. in [5]. DBSCAN discover the groups in dense regions of objects in the data space which are separated by regions of low density (elements of the latter regions represent noise). This method makes it possible to find groups of arbitrary shapes automatically generating the number of groups. However, it has the limitation of being sensitive to the initial value of its parameters, which determines the density of the groups to find and it does not works well finding groups of different densities and high-dimensional data.

Besides DBSCAN, other approaches have been published for noise management, one of the most recent is the APCOD (Automatic PAM Clustering algorithm for Outlier Detection) published in 2012 [4]. This method uses the classical clustering algorithm based on particional k-medioids PAM [8] and an internal index for automatic determination of the number of groups. Then using the LOF concept [7] are determined noisy data. This method has the disadvantage of not function properly with databeses of non-convex groups, as well as most centroids based methods. In addition, the number of outliers (noisemakers) to search is a parameter of the algorithm.

Actually, none of the above algorithms can group each data accurately, so some special data can be noise and be label as belonging to a group. Moreover, many of these and other proposals are limited in terms of assumptions about how to find groups, the dimensionality of the data and the density of the groups to obtain.

2 Proposed solution and experimentation planning.

We introduce a method with the ability to group data with noise by combining a strategy of noise detection and subsequent application of a classical clustering technique. Some clustering algorithms such as DBSCAN [5] allows to group data with noise but they require input parameters which are often difficult to determine. Moreover, there are other methods as APCOD [4] that first performs the grouping and then remove the noise, this often gives good results in the detection of noise but flawed in obtaining accurate clusters. The new proposal presented below is based on the design of a strategy to remove noise and the application of an agglomerative clustering algorithm. Unlike previous approaches, the proposed method allows greater accuracy in the production of "natural groups"of a data set with the presence of noise.

2.1 New proposal for data clustering in the presence of noise.

The proposed new method consists of two stages: noise elimination and clustering finding. For detection and noise elimination was designed a new strategy based on the concept of density using the k-nearest neighbors. In the case of obtain the clusters, it was adopted a classic strategy of UC as is the agglomerative single-link.

Outliers or noisy objects can be considered as objects or small groups located in regions of low density in contrast to the dense and larger structure formed by objects in the cluster. In this sense, there is an approach to identify outliers considering the number of existing objects around.

To identify regions of low density is necessary to determine for each object p the density $\delta_{Hr}(p)$ around it by counting the number of objects in the $H_{Hr}(p)$ hyperspace limited by certain hiperradio Hr.

This argument is not new; it has been used in the algorithms based on density, but their application needs of parameters such as the size of the neighborhood or hyperspace (Hr) and the threshold for which an object is considered noise. The novelty of the proposed method is the automatic determination of these parameters using the average nearest neighbor distance of all the objects d^{-}_{1-NN} and the average density for all δ^{-}_{Hr} objects using an iterative algorithm that remove the noise. The formulas for calculating these measures are offered in the following:

$$H_{Hr}(p) = \{q \in D \setminus \{p | d(q,p) \leq Hr\}\} \tag{1}$$

$$\delta_{Hr}(p) = |Hr(p)| \tag{2}$$

$$d^{-}_{1-NN} = \sum_{i=1}^{|D|} \frac{d_{1-NN}(i)}{|D|} \tag{3}$$

$$\delta^{-} = \sum_{i=1}^{|D|} \frac{\delta_{Hr}(i)}{|D|} \tag{4}$$

After having set noise-free data, we proceed to perform clustering using the single-link agglomerative strategy. The pseudo code of the proposed method is given:

Algorithm 1 Proposed Algorithm

Require: D: Dataset.
K: K-nearest neighbor.
G: Amount of non empty groups to obtain.
Ensure: C: Structuring in groups
First Stage: Noise detection
1: **repeat**
2: Calculate the nearest neighbor distance for all objects.
3: Calculate the average nearest neighbor distance for all objects.
4: **until** visit all K neighbors
5: Identifies and removes all the noisy elements (where $\delta_{Hr}(i) < \frac{1}{3}\delta^{-}_{Hr}$)
Secund Stage: Grouping data
6: Get the new free noise data base.
7: Run Agglomerative Algorithm Single Link.

Thus our proposal works in two stages, the first stage regarded to the noise detection and elimination and the second stage dedicated ti obtain clusters of the remaining data.

2.2 Experimental protocol

To carry out the tests we used nine synthetic databases, created by the user in a (CreatorData2D) tool implemented in Matlab (see the characteristic of each one in Table 1 and Fig. 2 shows the graphic representation) and eight actual databases from machine learning Repository of the University of California at Irvine (UCI Machine Learning Repository) [8], see table 2. The datasets of UCI repository it is known that in many cases have datasets with outliers that do not have labeled, it is for this we used this datasets too.

The choice of these datasets is that these are labeled databases, which facilitates the evaluation of the quality of the clusters obtained by the algorithms, using external validation indexes.

The algorithms against which the proposed method is compared are DBSCAN [5] and APCOD [4], being superior to other conventional techniques for detecting noise as LOF [7]. One of the most important design aspects of the experiments are the parameters needed to execute the methods. Like all

Table 1. Features synthetic data bases.

Datasets	Attributes	Groups	Class	Instances
ele	2	2	2	176
ruido	2	3	3	444
onda	2	2	2	306
anillo	2	2	2	806
base 1	2	3	3	910
base 2	2	3	3	472
base 3	2	3	3	518
base 4	2	3	3	637
base 5	2	3	3	451

Table 2. Features real data bases.

Datasets	Attributes	Groups	Class	Instances
iris	4	3	3	110
lymphography	18	4	4	148
cmc	9	10	10	1473
haberman	3	3	3	306
hayes-roth-train	4	3	3	110
liver-disorders	7	2	2	345
tae	5	3	3	151
wine	13	3	3	178

algorithms require knowledge of the number of groups to be formed, the value assigned to this parameter will match, for each database, with the number of classes. This allows taking labels classes of each database as a ground truth and; then compare the latter with the grouping obtained by the algorithms.

The rest of the parameters required by DBSCAN and APCOD are shown in Table 3 below. They were selected because these values are reported in literature as giving better results.

A Toshiba Laptop with the following characteristics was used for implementing and testing algorithms: Operating System Windows 7 Ultimate, 64-bit Microprocessor Intel Core i3 at 2.40GHz, with 2.65 GB RAM DDR3 memory at 665 MHz Dual-Channel, Motherboard Vendor PSKC8C-08G00R. It was also used MATLAB 7.0 Software [9], for numerical calculations with vectors, matrices, both real and complex scalar numbers, strings and other complex data structures, due to it has a basic code and several specialized libraries (toolboxes).

We use the Rand [10] index to compare the performance of the algorithms, over both synthetic and real databases.

Table 3. Parameters used in each algorithm.

Algorithms	Params
DBSCAN	Eps: radio of the the vicinity of each point - 0.3
	MinPts: minimum number of points in a neighborhood - 2
APCOD	α y β parameters for large and small groups alfa 0.75 y beta - 5

To process the data and draw conclusions on the performance of the algorithms the Wilcoxon test for related samples was used. It allows to stablished whether or not exist significant differences in the performance of each of the methods, with a 0.05 significance value. It should be noted that this test is recommended by Demsar [11] for such comparisons. In the next section the results of experimentation are described.

3 Experimentation and Results

In order to show the performance of the algorithms compared with the proposal (APCOD and DBSCAN), each of them in the test databases. Then the resulting clustering were evaluated using the Rand external index [10].

Rand is one of the indices of external validation most commonly used in the comparison of clustering algorithms. This index seeks to maximize the number of objects in the same group in the clusting to evaluate (CE), the ground truth clustering (GC) and the number of objects in different groups of CE and GC. Therefore, while higher is the value of Rand, more similar are clustering to the classes in the database analyzed. The Rand index is given by the following equation:

$$Rand = \frac{a+d}{\frac{N(N-1)}{2}} \tag{5}$$

where a is the number of pairs of objects in the same clusters in GC and CE, d is the number of pairs of objects in different clusters in GC and CE, N is the total number of objects.

The results of Rand index of the proposed algorithm over the others algorithms using synthetic databases are shown in Table 4 and real databases results are shown in Table 5.

Table 4 shows that the method of grouping DBSCAN have better performance (higher value of Rand) in six of the nine databases used in experimentation. The proposed algorithm wins in the database (base2) and ties in the database (ele) being secondly after DBSCAN while APCOD get third place.

To determine if the results are statistically significant Wilcoxon test is used for two related samples, which is recommended by Demsar [11]. For this, it was

Table 4. Rand index values of the HCOD algorithm and different methods applied to synthetic data bases.

Synthetic data bases	Proposed	APCOD	DBSCAN
ele	1.00000	0.70688	1,00000
ruido	0.98698	0.98191	0.99897
anillo	0.99504	0.49953	0.99998
onda	0.57495	0.51223	1,00000
base 1	0.92787	0.67001	0.99473
base 2	0.93999	0.63637	0.31098
base 3	0.97609	0.79633	0.98003
base 4	0.62388	0.56639	0.66244
base 5	0.85726	0.82193	0.92052

compared the proposed algorithm with each other. In each of the cases, it is set to null hypothesis that there are no significant differences between the performance of HCOD and the other algorithms. It is formulated as an alternative hypothesis, that the proposed algorithm has higher value of the Rand index than the other algorithms. The value of meaning that is adopted is 0.05 for 95% confidence. The Table 5 shows the results of the test.

Table 5. Wilcoxon test statistics for the Rand index of the algorithms (synthetic data bases)

	Algorithms	**Asymptotic significance**
PROPOSED ALGORITHM vs	APCOD	0.008
	DBSCAN	0.161

As shown in Table 5, the proposed algorithm outperforms the APCOD method. It is inferred from the value of asymptotic significance, which is less than 0.05, allowing reject the null hypothesis. However, there are no significant differences in performance of the proposed method with respect to DBSCAN algorithm, due to significance value of 0.161.

With regard to the actual databases in Table 6 Rand index values calculated are shown. There can be seen that the proposed algorithm has higher Rand index value than other methods in two databases (iris, hayes-roth-train) APCOD Rand has greater value in three databases (Lymphography, cmc, wine) than other algorithms and DBSCAN beats other methods in databases (haberman, liver-disorders, tae). DBSCAN and APCOD being tied by three in the first place with better performance and the proposed algorithm finishes second.

Table 6. Rand index values of the HCOD algorithm and different methods applied to real data bases.

Synthetic data bases	Proposed	APCOD	DBSCAN
iris	0.92727	0.85004	0.92344
Lymphography	0.28838	0.63164	0.23396
cmc	0.47599	0.55803	0.34430
haberman	0.57270	0.49182	0.61571
hayes-roth-train	0.65903	0.57819	0.34525
liver-disorders	0.49870	0.50266	0.50563
tae	0.51859	0.54623	0.62155
wine	0.62141	0.69009	0.65930

Applying the Wilcoxon test to these databases and using the same assumptions as used in comparison with previous methods, the table below shows the results obtained:

Table 7. Wilcoxon test statistics for the Rand index of the algorithms (real data bases)

	Algorithms	**Asymptotic significance (bilateral)**
PROPOSED ALGORITHM vs	APCOD	0.674
	DBSCAN	0.674

Table 7 shows that in the algorithms no significant differences are evident. Observed a good performance in these three methods.

4 Conclusions

In this paper, a new clustering algorithm which has the ability to handle noisy data sets of numerical type is proposed. The new proposal presented is based on the design of a strategy to remove noise and then applying an agglomerative clustering algorithm to obtain groups. Experimental results show that the proposed algorithm has good performance in synthetic databases and real data using as validation the Rand index. The proposed algorithm is stable and in some cases it has superior behavior removing noise with respect to other techniques. This method can be defined as a clustering algorithm in the presence of noise that can efficiently manipulate numeric data types obtaining efficient performance.

References

1. Xu, R. and D. W. II: Survey of Clustering Algorithms. IEEE Transactions on Neural Networks. 16, 645-678. (2005)
2. Hawkins, D. M.: Identification of outliers. New York, USA: Chapman and Hall. (1980)
3. Dianmin, Y., et al.: A Survey of Outlier Detection Methods in Network Anomaly Identification. Computer Journal. 54, 570-588. (2011)
4. Lei, D., et al.: Automatic PAM Clustering Algorithm for Outlier Detection. Journal of Software. 7, 1045-1051. (2012)
5. Ester, M., et al.: A density-based algorithm for discovering clusters in large special databases with noise. 2nd International Conference on Knowledge Discovery and Data Mining (KDD96). Portland, OR: AAAI Press. (1996)
6. Knorr, E. M. and R. T. Ng.: Algorithms for Mining Distance-Based Outliers in Large Datasets. 24rd International Conference on Very Large Data Bases. New York, USA. (1998)
7. Breunig, M. M., et al.: LOF: identifying density-based local outliers. ACM SIGMOD International Conference on Management of Data. New York, USA. (2000)
8. Asuncion, A. and D. J. Newman, : UCI machine learning repository. http://archive.ics.uci.edu/ml. University of California. Irvine, CA. (2007)
9. Jalon, J. G., J. I. Rodrguez, and J. Vidal: Aprenda MATLAB 7.0 como si estuviera en primero. 1-128. (2005)
10. Jain, A. K. and R. C. Dubes: Algorithms for Clustering Data. Prentice Hall. (1988)
11. Demsar, J.: Statistical comparison of classifiers over multiple datasets. The Journal of Machine Learning Research. 7, 1-30. (2006)

Segmentation and Identification of Peripheral Blood Cells Using Image Analysis

Carlos Eymard Zamacona López[1], Raúl Bruno Sánchez[1], Arnulfo Catalán Villegas[1], Gustavo A. Alonso[1], Ma. Elena Moreno-Godínez[2], Antonio Alarcón-Paredes[1]

Universidad Autónoma de Guerrero, Unidad Académica de Ingeniería, Chilpancingo, México[1]
Universidad Autónoma de Guerrero, Unidad Académica de Ciencias Químico-Biológicas, Chilpancingo, México[2]

{cezamacona[1], rbruno[1], catalanvillegas[1], gsilverio[1], aalarcon[1]}
@uagro.mx
emoreno20[2]@hotmail.com

Abstract. One of the preferred techniques to measure genetic damage is the Micro-Nucleus (MN) Test in peripheral blood cells. The counting and identification of micronucleus in binucleated cells, is traditionally performed manually and directly on a microscope slide by an expert, which results in a highly time consuming task. Although it is a standard method to monitor the genetic damage, there are some significant differences in laboratories which perform the MN test. One of the main differences lie in the use of different compounds to stain the cells, and the use of sophisticated hardware and cameras to obtain high-detailed images. In this paper, a method for automatic segmentation of peripheral blood cells is presented and can serve as an alternative to the issues before mentioned. Results show that the proposed algorithm is able to identify MN in images obtained with a mobile phone camera, which results in a simple and flexible method.

Keywords: Cell segmentation, Micronucleus Test, Image analysis, Mathematical Morphology

1 Introduction

For several years, there is a growing industrial activity in several different areas. Recently, some of the growing industrial activity involves the use of genotoxic agents that endanger and threaten the genetic integrity in human beings.

In addition to this, there are different factors that influence the genetic damage, such as: feeding habits, medical therapies that involve radiation, as well as the increase of solar radiation to which we are exposed due to uncontrollable climatic change.

Whatever the factor of genetic damage is, it is important to quantify and determine an acceptable level of damage in human population[1].

A technique to measure genetic damage is the MicroNucleus (MN) Test in peripheral blood cells. These studies are useful when there is a bad or deficient cell division and there is a break or chromosome loss [2] giving as a result the generation of MN. For that reason, it is crucial to provide a reference range for a "normal" quantity of MN present in a blood cell of a person.

The counting and identification of MN in binucleated cells, is traditionally performed manually by an expert.

Different automatic methods for counting (or scoring) MN exist, whose results appear to be quite good and can be consulted in [1, 3]. However, many of these works involve the acquisition of specialized hardware equipment; in addition, the commercial software code is proprietary and it is tuned to work only with that hardware [4-6].

Although this test has been consolidated as a standard method to monitor the level of damage in chromosomes, there are some differences when it is carried out. In [7] a study called "HUMN PROJECT" was made, in this study 25 different laboratories in 16 countries were involved, where significant differences were found in methods used among the participating laboratories to perform the MN Test.

Some of the most significant differences lie in the use of different compounds to stain cells to be analyzed which hinders an algorithm for automatic processing to function properly in different samples whose staining is very different; as well as the use of specialized cameras that allow image acquisition whose quality or resolution is very different from the equipment that can be found in laboratories with low budget.

In this paper, a method of automatic segmentation in peripheral blood cells for scoring MN is presented; it can serve as an alternative to the issues which are mentioned in the previous paragraph. The proposed method uses images taken with a mobile phone camera pointing in the microscope, and with a compound of staining which differs from those clear and detailed images presented in the literature [5, 8].

The rest of this paper is organized as follows: In Section 2 the necessary techniques to get the best information from imaging microscopy are described, in Section 3 the proposed Algorithm is presented. Section 4 shows the results obtained by the use of the proposed algorithm. The conclusions and ideas for future work of the developed system are shown in Section 5. Finally, acknowledgement and references are shown.

2 Methods

In this section the techniques used to perform the Algorithm to obtain image information are described.

2.1 Otsu's Binarization Method

Binarization is an image processing technique which consists in an image information quantization into two values: 0 (black) and 255 (white). The main goal is to obtain an appropriate binarization threshold; a good option is to use the Otsu Method.

The Otsu Method is a nonparametric and unsupervised method of automatic threshold optimal selection [9], the Otsu Method can discriminate the pixels of an image in several classes $C_0 = \{1,...,k_1\}$, $C_1 = \{k_1,...,k_{2\},...,}$ $C_n = \{k_n,...,L\text{-}1\}$ where L is the maximum value of a pixel and each class includes all the pixels to a threshold detected by Otsu Method, and $U = \{k_1,k_2,...,k_n\}$ the thresholds set detected.

Probability distribution of the normalized gray levels could be expressed as $p_i = n_i/N$ where p_i is the probability of occurrence, n_i is the gray level intensity, and N is the number of pixels.

The zeroth- and the first-order cumulative moments are defined as $\omega_0 = \sum_{i=1}^{k} pi = \omega(k)$ and $\omega_1 = \sum_{i=k+1}^{L} pi = 1 - \omega(k)$, respectively, with $\mu_0 = \mu(k) / \omega(k)$ and $\mu_1 = \mu T - \mu(k) / 1 - \omega(k)$ where $\omega(k) = \sum_{i=1}^{k} p_i$ and $\mu(k) = \sum_{i=1}^{k} ip_i$.

The total mean level of the original picture is obtained with the equation defined as $\mu_{T =} \mu(L) = \sum_{i=1}^{L} ip_i$ and $\sigma_B^2 = \omega_0 (\mu_0 - \mu_T)^2 + \omega 1 (\mu_1 - \mu_T)^2$.

The optimal threshold in a search for the maximum value in σ_B^2 is

$$\sigma_B^2(k^*) = \max_{1 \le k < L} \sigma_B^2(k)$$

For problems involving multilevel threshold, it is assumed that measurement criteria σ_B^2 are based on multiple variables $k_1, k_2, ..., and\ k_n$, and an optimum set of threshold $k_1^*, k_2^*, ..., and\ k_n^*$ is selected maximizing σ_B^2

$$\sigma_B^2(k_1^*, k_2^*, \dots, k_n^*) = \max_{1 \le k1 < k2 < \cdots < kn < L} \sigma_B^2(k_1, k_2, \dots, k_n)$$

2.2 Morphological Image Processing

Mathematical morphology has simple techniques for extracting components in an image, which are useful in the representation and description of region shapes. The two main operations, dilation and erosion, are performed over a set representation of an image A using a structuring element denoted by B, which defines the shape and size of the neighborhood of the pixel that will be analyzed subsequently to modify its value [10, 11]. The morphological operations used in this paper are described next.

- **Dilation**. Dilation of A by B is the Minkowski sum [12] of A and B, it is defined as:

$$A \oplus B = \{x = a + b \in X \mid a \in A \wedge b \in B\}$$

- **Erosion**. Erosion of A by B is the Minkowski subtraction [13], and it is defined as:

$$A \ominus B = \{x \in X \mid x + b \in A, \wedge \forall b \in B\}$$

- **Opening**. Opening of A by B is the erosion of A by B, followed by dilation of the result by B, it is defined as:

$$A \circ B = (A \ominus B) \oplus B$$

- **Region Filling**. Region Filling is based on a set of dilations, complements and intersections, it is defined as:

$$Xk = (Xk - 1 \oplus B) \cap Ac\ k = 1, 2, 3, \ldots$$

Where $X_0 = \mathrm{p}$, and B is a symmetrical structural element. The algorithm ends in step k if $x_k = X_k - 1$.

2.3 Distance transform

For each pixel in a binary image, this transformation assigns a number that is the distance between the pixel and the nearest nonzero pixel, the Euclidean distance is used, it is defined as:

$$d(\mathrm{x}, \mathrm{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

2.4 Watershed Transform

In mathematical morphology, Watershed is a technique where an image is considered as a topographic relief [14]. There are several different methods of segmentation that can be used to perform a Watershed Transform, however, the method used in this paper is carried out as follows:

1. A pre-processing filter for noise removal is done, and then the image is binarized.
2. A distance transform is performed and its complement is obtained.
3. The areas with the minimum values are identified; these areas are placed as markers in the binarized image.
4. A Morphological Reconstruction is performed.
5. Finally Watershed Transform is carried out.

3 Algorithm

A Data flow diagram of the proposed algorithm is shown in Figure 8. This implements the methodology mentioned in the previous section, and it is able to perform cell segmenta-

tion in order to count the amount of MN present in a microscope slide, by means of a mobile phone camera. The algorithm is described as follows:
1. Obtain an RGB picture of the peripheral blood cell slide, and get only the green plane, as shown in Figures 1a and 1b.

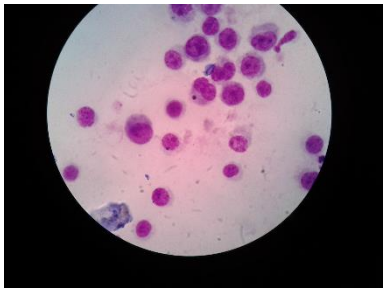


Fig. 1a. Original RGB image

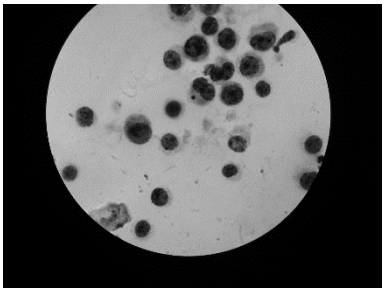


Fig. 1b. Green plane

2. Obtain four thresholds values in the green plane image by means of Otsu multi-threshold technique. Then binarize the image using the third threshold, and get the negative image, as shown in Figure 2.
3. Here, a cleaning step is performed over the binarized image. This step consists in applying an algorithm to fill holes [15] in the detected objects, as well as the application of an opening function in order to delete several small objects as showed in Figure 3.
4. The distance transform is now performed over the complement of the Figure 3. This algorithm computes the (euclidian) distance between every zero-valued pixel and the nearest non-zero pixel. The complement of the resulting image is a gray-scaled gradient where the innermost pixel in the objects, has the darkest value, as shown in Figure 4.

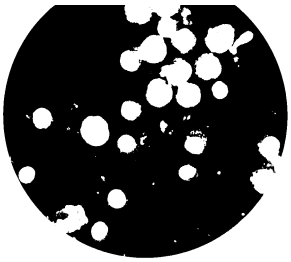


Fig. 2. Binarized

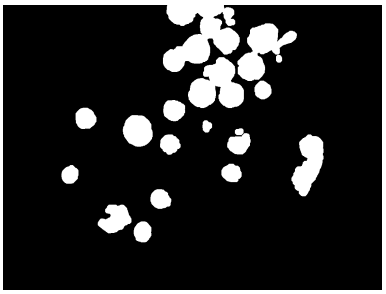


Fig. 3. Image after cleaning process

5. Obtain the morphological reconstruction of the negative distance transformed image. In this case the morphological reconstruction selects the darkest regions of the distance trans-

formed image, and masks the gray scale image so these regions are the minima regions of reconstructed image. The darkest regions could be used as markers representing the basins, and the reconstructed image is the topographic surface to be flooded with the watershed transform. Figures 5a and 5b show the result of this process.

Fig. 4. Result of applying the distance transform

6. Compute the watershed transform to the image resulting of the step 5 and could be seen in Figure 6.
7. Obtain the object labeling using an 8-connected neighbourhood (see Figure 7).

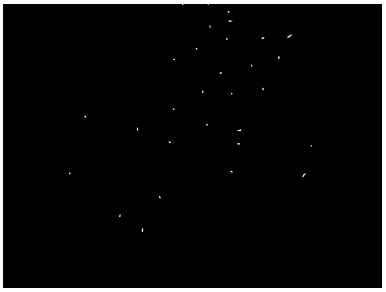


Fig. 5a. Markers (basins)

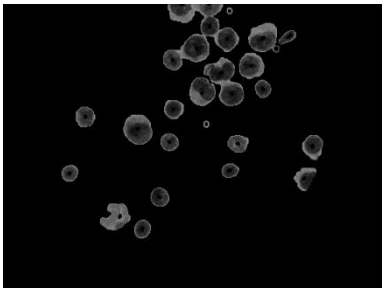


Fig. 5b. Topographic surface (morphological reconstruction)

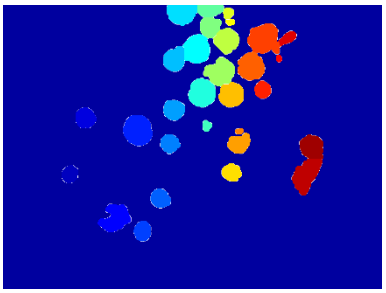


Fig. 6. Results of watershed transformation

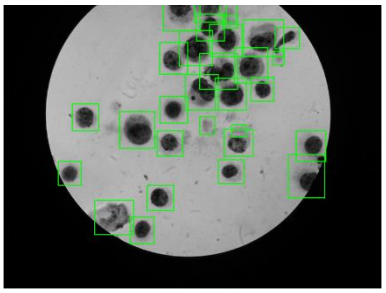


Fig. 7. Object labeling

4 Results

The proposed algorithm was implemented using Octave on a computer with an Intel core i3 processor and 4 GB of RAM, running on Windows 8.1, and was applied to a set of 16 peripheral blood cell images obtained with a cell phone camera. The cellular phone used to capture the images was a Samsung gt-i8190l, running over Android 4.1 OS with a 5.0 MP camera resolution, with image size of 2560 pixels width and 1920 pixels length.

The segmentation provides very good results in cell identification, since the images are obtained using camera phones. In Figure 9, a visual example of the results obtained is shown. In addition, Figure 10 shows the data results of the cells identified by the proposed method versus manual identification. Nonetheless there are some differences between the manual and automatic scoring processes, often caused due to border located cells and some waste of the stain compound, this problem will be overcome in future work by means of a further step to discriminate between true cells and the spurious ones.

In addition, images immerse in the literature have clearer boundaries than ours, which yields to an easier image segmentation. This is caused due to the use of a more expensive compound to stain the cells, and the hardware they use to obtain the image since it is very robust and permits increasing the zoom of the pictures, so the images to be analyzed have much more details.

Here, the results show that despite of the drawbacks of having compounds that could not make the boundaries of cells too much clear, it is possible to segment the blood cells in a microscope slide image taken with a phone camera. Results of segmentation of some other cells are shown in Figures 11 to 14.

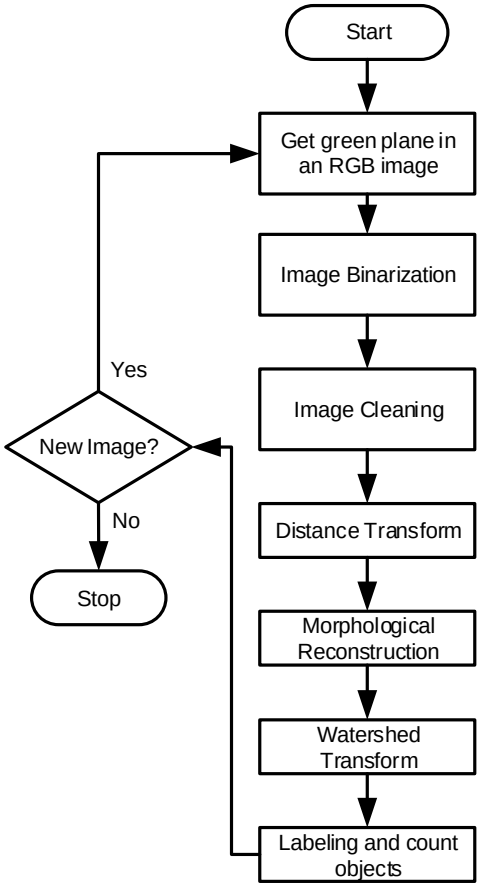


Fig. 8. Dataflow diagram of proposed algorithm

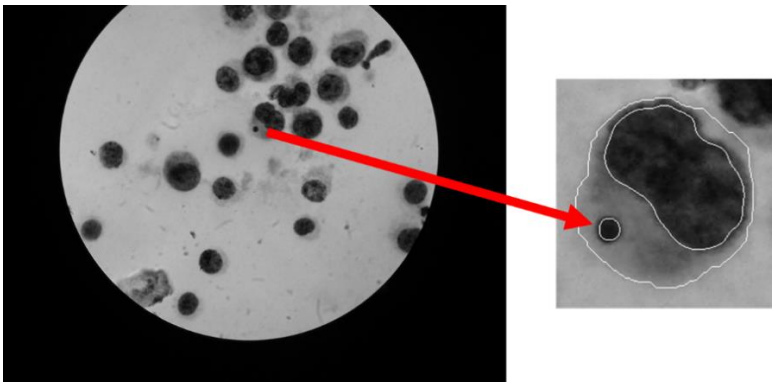


Fig. 9. A micronucleus within a binucleated cell, is shown.

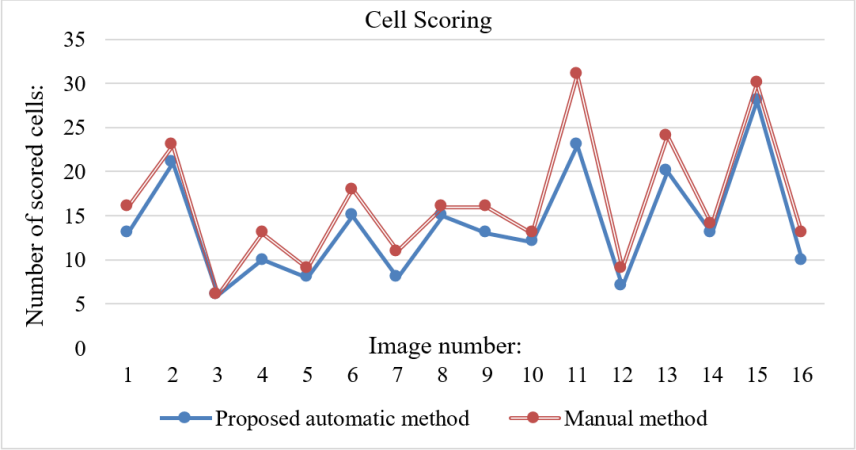


Fig. 10. Comparison of the manual vs automatic identified cells

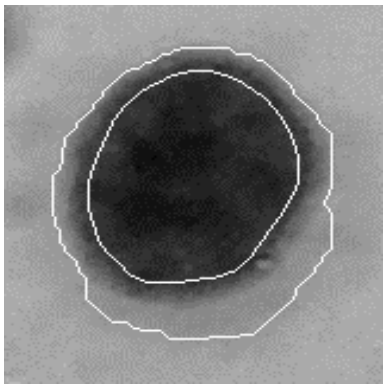


Fig. 11. Segmented cell

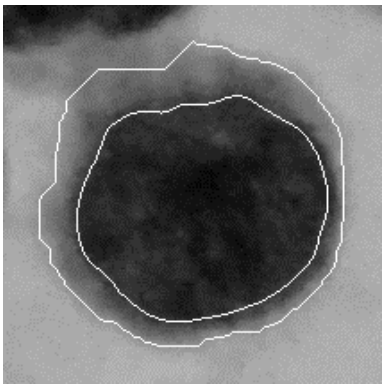


Fig. 12. Segmented cell

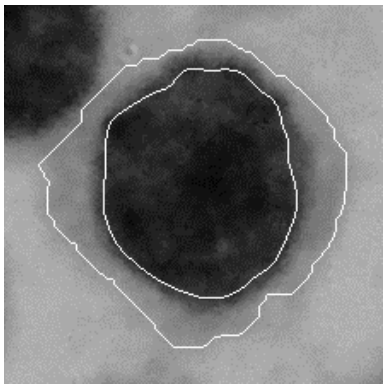


Fig. 13. Segmented cell

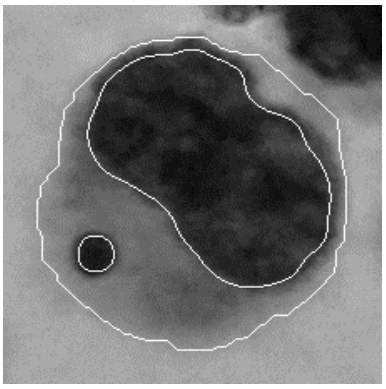


Fig. 14. Segmented cell with a MN

5 Conclusions

The automation of cell segmentation allows laboratories with no sophisticated and inexpensive equipment, and also without the need for an expert, to perform the MN test in human populations which are exposed to genotoxic compounds, to help in the genetic damage diagnosis.

The proposed algorithm, results in a reliable method for peripheral blood cell segmentation, since the results obtained using images with poor distinguishable boundaries between cells are competitive. It is worth to mention that images in the state of the art have a clear cytoplasmic boundary due to specific compound used to stain the slides, which rep-

resents a difference in the images used among laboratories worldwide. For this reason, we can conclude that if this method is applied over images in the literature, the results would be even better.

In contrast with the use of sophisticated hardware for image acquisition in the related works, here, a mobile phone camera was used to obtain the blood cell images, which represents that proposed algorithm is a very flexible method and could be used in laboratories where the high technology devices for image acquisition are not affordable.

The results presented in this work are assumed to be improved by means of merging some other segmentation methods. With the aim of enhance the hardware and time consumption, a migration from Octave to another programming language such as C or C++ are also expected.

Acknowledgments

The authors of this work appreciate the support given by CONACYT for their economic support to develop this work.

References

1. Fenech, M.: The cytokinesis-block micronucleus technique: a detailed description of the method and its application to genotoxicity studies in human populations. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 285, 35-44 (1993)
2. Schmid, W.: The micronucleus test. Mutation Research/Environmental Mutagenesis and Related Subjects 31, 9-15 (1975)
3. Fenech, M.: The in vitro micronucleus technique. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 455, 81-95 (2000)
4. Böcker, W., Müller, W.U., Streffer, C.: Image processing algorithms for the automated micronucleus assay in binucleated human lymphocytes. Cytometry 19, 283-294 (1995)
5. Frieauff, W., Martus, H., Suter, W., Elhajouji, A.: Automatic analysis of the micronucleus test in primary human lymphocytes using image analysis. Mutagenesis 28, 15-23 (2013)
6. Rodrigues, M., Beaton-Green, L., Kutzner, B., Wilkins, R.: Automated analysis of the cytokinesis-block micronucleus assay for radiation biodosimetry using imaging flow cytometry. Radiation and environmental biophysics 53, 273-282 (2014)
7. Bonassi, S., Fenech, M., Lando, C., Lin, Y.-p., Ceppi, M., Chang, W.P., Holland, N., Kirsch-Volders, M., Zeiger, E., Ban, S.: HUman MicroNucleus project: international database comparison for results with the cytokinesis-block micronucleus assay in human

lymphocytes: I. Effect of laboratory protocol, scoring criteria, and host factors on the frequency of micronuclei. Environmental and molecular mutagenesis 37, 31-45 (2001)
8. Fenech, M., Chang, W.P., Kirsch-Volders, M., Holland, N., Bonassi, S., Zeiger, E.: HUMN project: detailed description of the scoring criteria for the cytokinesis-block micronucleus assay using isolated human lymphocyte cultures. Mutation Research/Genetic Toxicology and Environmental Mutagenesis 534, 65-75 (2003)
9. Otsu, N.: A threshold selection method from gray-level histograms. Automatica 11, 23-27 (1975)
10. Díaz-de-León-Santiago, J.L., Yáñez-Márquez, C.: Introducción a la morfología matemática de conjuntos. Instituto Politécnico Nacional (2003)
11. Matheron, G., Serra, J.: Image analysis and mathematical morphology. London: Academic Press (1982)
12. Minkowski, H.: volumen and Oberfläche. Math. Ann., vol. Vol. 57, pp. 447-495. (1903)
13. Hadwiger., H.: Vorslesunger über Inhalt, Oberfläche und Isoperimetrie. Berlin: Springer (1957)
14. Beucher, S., Meyer, F.: The morphological approach to segmentation: the watershed transformation. OPTICAL ENGINEERING-NEW YORK-MARCEL DEKKER INCORPORATED- 34, 433-433 (1992)
15. Soille, P. Morphological image analysis: principles and applications. Springer Science & Business Media. (2013)

Platforms and Graphication Tools for a mobile application: Simulation of Planetary Objects' Trajectories

Analuz Silva Silverio[1,*], Mario A. De Leo Winkler[2], Antonio Alarcón Paredes[1]

[1]Universidad Autónoma de Guerrero, Unidad Académica de Ingeniería, Chilpancingo, Mexico
sisianlight@gmail.com, aalarcon@uagro.mx
[2]University of California, Physics and Astronomy Department, Riverside, USA
mario.deleo-winkler@ucr.edu

Abstract. During the development of mobile applications some of the first things that should be taken under consideration are the platform, code and development tools to be used. To do this, it is important to keep in mind compatibility of devices and operating systems, the focus audience, the project's budget, and other factors that end up being crucial for the successful completion of a project of this type. Because this particular case we want to develop applications for scientific outreach in astronomy, we investigate and present trends in developing mobile applications as well as propose a number of tools for our particular project, themed on planetary objects in a bi-dimensional simulation of positions, as well as the charted and tabulated data.

Keywords: Physics computing, Mobile computing, Native app, Application software, Open source software, Software tools, Programming environments, Computer science education, Astronomy.

1 Introduction

In mobile applications' (apps) development, one of the first things to take into account is the type of application that is being developed. For this reason, it becomes necessary to know positive and negative aspects in the trends of apps development.

Nowadays, the guidelines of app development have a wider scope than the initial ones where only two paradigms existed: native apps and web apps. Native applications are developed on the native language of an operating system (OS). A web application consists of a website specifically optimized with an interface and a set of functions to be used in mobile devices. Recently, with the arose of HTML5, the concept of hybrid applications has emerged and comes with added features and functionalities related to direct hardware control and the possibility to be implemented in different OS. Those new features have made hybrid apps become a new trend of development.

We analyzed the possibility of having a greater exposure of astronomy through the use of mobile apps for scientific outreach where we reviewed currently available astronomical apps for devices running the Android OS and we defined through that review the astronomical themes that needed further development or improvement [1]. It is worth to mention that there are no works with a similar intention in Spanish and through this, we contribute to the creation of a scientific outreach tool for observers, amateur astronomers and public education in general.

The final objective of selecting a specific kind of tool for development is to create a set of apps focused to create outreach for astronomy in Spanish, under a GNU license, completely free charge for users and an open source code commented for its future use or improvement among the Spanish-speaking population interested in this subject. Our app will simulate the position of the major moons around the gaseous planets in our Solar System; such a simulation serves three purposes: it creates a useful tool for amateur astronomers to plan their telescopes/binocular observations, our app is paired with important information on the satellites and planets themselves creating a educational tool, the source code is commented and open sourced so it can be used as an educational tool for programmers of astronomical apps.

In this paper we discuss some of the trends in platforms and graphic tools for the development of mobile apps. The paper is organized as follows: Section 2 has a review of apps development guidelines, in which we explain a research we made for selecting the IDEs, frameworks, trends on graphics building, charting libraries, and the arguments of choosing the ones we used. In the Results Section, the products of using those tools are shown. Section 4 deals with the discussion of the proposed work, and finally, the conclusions and references are presented.

2 Method

We must mention that in every cases compatibility issues could occur, mainly related to the hardware of the device and the OS version, which is also a latent issue when developing apps.

2.1 Review of the Context on Mobile Applications' Development Trends

Describing the main characteristics of the paradigms of the guidelines of apps development is important for based on that information be able to select one of them.

Web applications. This type of apps properly constitutes a website, the programming languages are the same used in common websites (e.g. HTML, CSS, and JavaScript (JS)), so the content could be anything supported on a website. It can be reached by an URL in a web browser and optimized for hardware specifications of mobile devices [2]. Since a

browser allows the creation of a website shortcut on the desktop, the web apps are not truly installed, so the "installation" step is the creation of a shortcut.

Adapting a website to a mobile format is a good solution in many cases, although there are some drawbacks such as the lack of accessing the device's hardware, publishing on official application stores (at least before HTML5) and the poor response time and user experience in contrast with common web navigation.

Native applications. These applications are developed on native languages of each operating system (OS). The programming languages for this purpose are Java (Android), Objective C (iOS), C# and Visual Basic .NET (Windows), and C++ (BlackBerry) [3].

Nevertheless, native apps are the option with the best implementation [3], it depends on a good app design at all levels; according to the above, the development of native apps may grant full access to mobile device hardware, push notifications, publishing the app at Google Play, App Store, or Windows Phone Store [3], and data synchronization for offline use.

The drawbacks are mainly found in economic and time consumption terms, such as maintenance, functionality and OS updates which are frequently required. In addition, the impossibility to reuse code among different OS programming languages makes necessary that a single developer needs to know different programming languages or to have a bureau of skilled programmers for each OS and for some less-used languages among developers (compared to web languages).

Hybrid applications. For developing a hybrid application, a web programming language-based framework (such as HTML5, CSS3 or JS) is needed. In these apps, the OS integration level depends on the selected framework, and the OS itself. These apps let the developer have access to the hardware of the mobile device and, often, to some OS libraries. However this method has still not reached the same response time and experience as native apps. Nonetheless, they have been evolving to more complex processes where, in some cases, they could act like a web app and, in other cases, natively.

Hybrid applications are a very economic option and could be very useful to get a great number of users in a variety of different platforms and devices.

2.2 Selection of the Type of Application

In order to develop an astronomy app, and due to the highest percentage of users it has, the Android OS could be selected. But if we choose to develop a native app, it would run only on Android devices [1].

On the other hand, the major problem with web and hybrid apps is their high energy consumption when calling JS, Canvas, and some other graphical resources. However, this

drawback is overcome by the automatic release of the resources while the device is not using the application.

So, there is no clear advantage between the native and hybrid apps, but the trend is to use even more hybrid HTML5 apps over the native ones. Nonetheless, it is important to note the remarkable use of HTML5 in app development, as can be seen in Fig. 1; it is one of the preferred platforms for developers due to its versatility, new features and capabilities [4].

The selection of the application type has to do with the app's functions, and the use we must give it. Based on that, in this project we chose to develop a hybrid app, since the graphical resource consumption to represent a planetary simulation will not be as exhaustive as for a game app, for example. On the other hand, since hybrid applications have become easy to develop due to the use of some tools such as PhoneGap Developer App [5], we think the development time and resources needed, as well as the project costs will be minimum. Moreover, the app will not be limited to Android users only, and the contents could even be part of a website, making our app a wide and very helpful tool for scientific outreach.

2.3 Selecting the Integrated Development Environment (IDE)

Most of the IDEs have a smart code autocomplete function, a class navigator, object finder, class hierarchical diagrams for object-oriented software, and probably also an integrated version control system.

For a long time, one of the tools most used for developing web pages (and now apps) was Adobe Dreamweaver, but its use requires a purchased licence.

Aptana Studio is an alternative for developing websites that has become very popular. It is an open source IDE based on Eclipse, which can run under Windows, Mac and Linux, provides support for languages like PHP, Python, Ruby, CSS, Ajax, HTML, Adobe AIR, and the use of plugins for new languages is also possible. Its main characteristics are a code assistant for HTML and JS, Ajax libraries (jQuery, prototype, scriptaculous, Ext JS, dojo, YUI, and Spry, etc.), data base tools, syntax verifier, compatibility with Eclipse plugins or extensions (more than 1000).

Thus, Aptana Studio seems to be a very versatile IDE for the needs of our project. It also has a very good compatibility with most used frameworks for app development. So that, in this work we select Aptana Studio as our IDE.

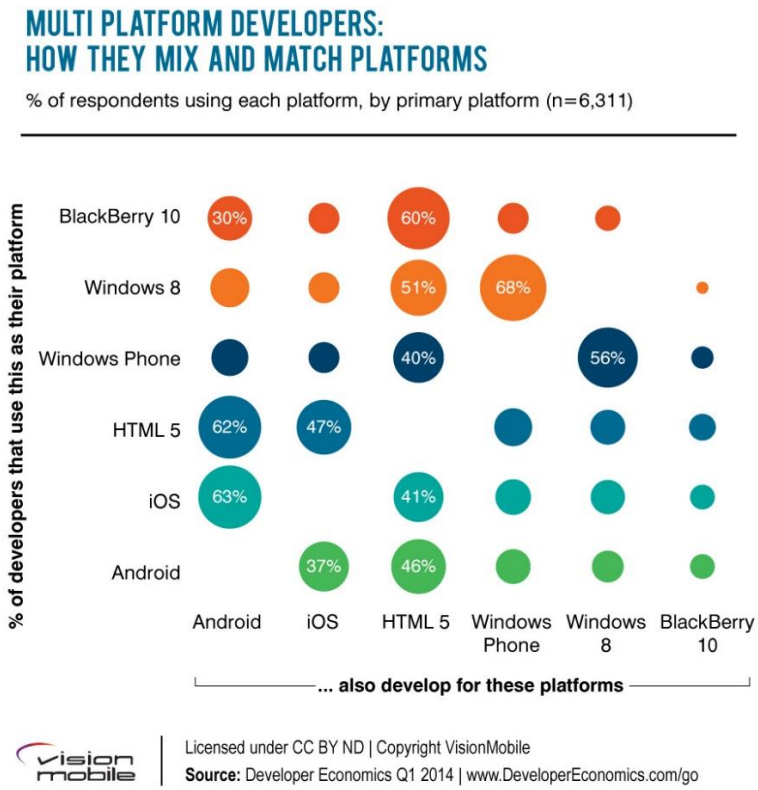


Fig. 1. Multi Platform Developers: How They Mix And Match Platforms. The third column shows how HTML5 is one of the preferred platforms for developers to match with other platforms due to its versatility, new features and capabilities [4].

2.4 Frameworks and Libraries

Seeking among the most used frameworks to create apps for Android and iOS with HTML5 [6], we found those for developing web apps and hybrid apps that require no financial cost to use or it is not requested unless the developer will profit from the end product. Those frameworks are:

Chrome Dev Editor. Created by Google to develop web apps and Chrome Apps, it can be used from devices using Chrome OS (Chromebooks) or the Chrome browser. This allows to create projects with templates such as Web JS apps and combining them with Polymer, Dart apps and Chrome Apps, for desktop and mobile, which are easily executable in Android and even packable in an executable Android application package (APK) or

Web Starter Kit. It includes Dart compiler and launches an embedded service to implement projects from the local machine. It is also able to directly upload the Chrome Apps to the Web Store [7].

PhoneGap. Being one of the best known, PhoneGap is the preferred framework, and also, the one that concentrates the greatest number of developers worldwide [8]. Based on Apache Cordova license, it allows packaging code as an application for iOS, Android, Blackberry and Windows Phone. PhoneGap allows programming through its API native SDK functionalities. For distribute a project among different platforms, a native application for each one is built with the help of Phonegap, and using the same source code it generates the executable for the different OS [6].

Icenium. This framework removes the concern of each platform's SDK. Through its own IDE creates apps using web standards (such as HTML5 and CSS) and publishes them directly to the app stores through a simple step by step. This is the main advantage of Icenium compared to PhoneGap. Due to its environment Icenium Mist can be accessed from any browser and makes the publication by selecting a desired platform (Android or iOS). Unlike PhoneGap, this framework is not available for free [6].

Appcelerator. Just as Icenium, it has its own IDE, allowing code common functionality for different platforms. One of the most interesting qualities of Appcelerator is the easy connection of applications with different data sources using their customized connected services with node.js. Although it is required to work within the platform and use its own functionalities. Appcelerator offers features that easily integrate information from multiple services into applications [6].

For the development of this project we chose to work with PhoneGap and its direct tester app "PhoneGap Developer App" [5], as it proves easy to work offline, and to test applications without internet access. Because offline use is one of the objectives of this work, the possibility to use them in the middle of the field without suffering problems in the use of the application due to the absence of internet. On the other hand, as shown in Fig. 2, PhoneGap is the most sought tool in this tenor. It should be mentioned that if a migration to other frameworks is required, this is possible because they work under the same type of scheme.

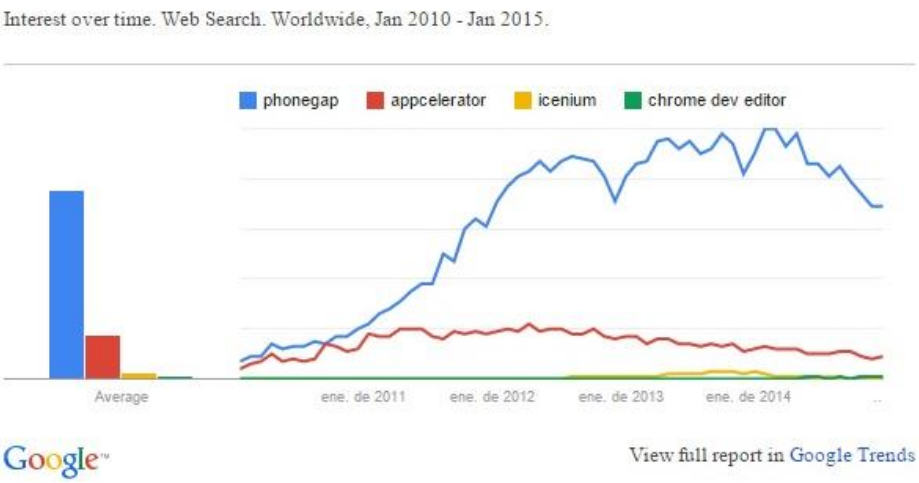


Fig. 2. Interest over time of frameworks for hybrid applications development [9].

Once we have chosen the framework we proceed to review trends of web graphics usage. HTML5 Canvas, SVG and Adobe Flash are trendy in Web graphic representations today. Arising from the creation of HTML5, Canvas, rendered through a bitmap, requires a single element to be added to the DOM, the canvas element in HTML5. Hence, we can use JS to access the context and then draw things. This offers a wide variety of options to be deployed, without making a page or app much heavier, making this element one of the most innovative resources of HTML5 [10].

Since SVG and VML require DOM elements for each form added to the document, the model document can quickly overflow for complex drawings, and the application can become slow or even crash. While Canvas is relatively new, it has proved to be very efficient and have a better performance mainly because its strongest point is the independence of complements for its handling. Then, if an application is rendered in 2D, it needs to be fast and does not require direct events on the canvas from the user interface, it is recommended to create an app using Canvas [10].

We have found several libraries to create charts, but few of them handled Canvas as its deployment element and, among those offering this option, few had a wide variety of charts or configuration possibilities. Through searches we performed, we found Highcharts was among the most popular and complete of the libraries. Highcharts is a graphics library written in JS, which provides an easy way to add interactive graphics to web sites or web applications. For this project we opted to use this library as it covers our developing needs with the following features [11]:

- It works in all modern browsers, on iOS and Android; it provides multitouch support and it is free for non-commercial uses.

- It is solely based on native browser technologies and doesn't require client side plugins like Flash or Java, it allows downloads of the source code and to make your own edits and offers a simple configuration syntax.
- Through a full API it can add, remove and modify series points, or modify axes at any time after chart creation and all options can be set individually, including reversing, styling and position, additionally, has datetime axis and offers inverted chart or reverse axis. Upon hovering the chart can display a tooltip text for information on each point following the movement of the mouse over the graph.
- Users can export the chart to PNG, JPG, PDF or SVG format at the click of a button, or print the chart directly from the web page.

3 Results

For the development of this project we took some decisions related to specific requirements trying to give the astronomical theme we selected the widest possible exposure, for instance: free distribution, have the opportunity of development by a single developer and develop within a limited budget. During the development we verified that the simulations building, graphics in HTML5 Canvas and the Highcharts library offered the versatility needed for this project.

Fig. 3. Result of the bi-dimensional simulation of Jupiter made in HTML5 Canvas. In this capture the moons of Jupiter can be seen given a particular date in the text field below together with controls for changing the positions and speed of reproduction of animation.

In the deployment of the developed simulation, Canvas includes several manipulable JS functions allowing freedom of handling and an outstanding design. The functions used to generate the positions of the planetary objects were based on a code made in HTML [12], and the algorithms were explained in the book "Astronomical Algorithms" [13]. As shown in Fig. 3, we used buttons, sliders and text fields to redraw the Canvas element as the data is taken automatically from the system date or manually from a date entered, as well as the activation and de-activation of the animation and the increase or decrease of the speed of simulation. The way in which the positions of the objects are drawn just re-

quired to load the data in advance in the page before calling the canvas element so that no errors occur during the deployment.

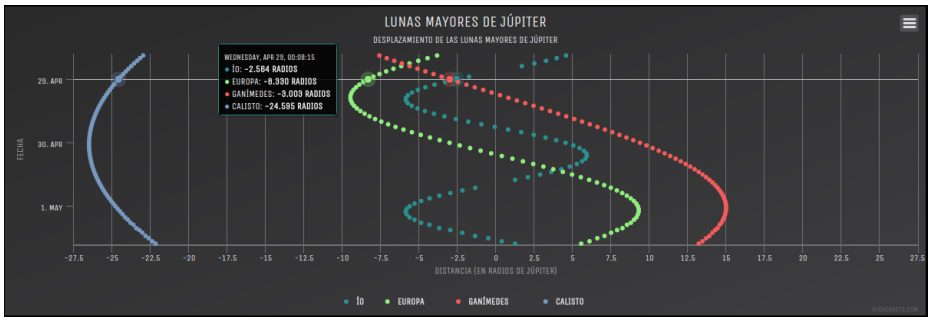


Fig. 4. Result of the chart made with Highcharts. In this image we can see a graphic generated for the positions of the trajectory of the Galilean satellites for the following 72 hours.

Highcharts, offered the possibility of translating charts with inverted axes required to capture the sinusoids formed by the changing position of the satellites with respect to time, with high quality graphics design, and a range of settings for units and options in the design of the chart, which is well adjusted to the needs of the project. As seen in Fig 4, the deployment was adjusted to the design needs of the project. The axes are inverted, a recurrent representation in apps that present satellites trajectories such as the app "Moons of Jupiter" [14]. Among the options worked in Highcharts was the setting of time units for showing the displacements during a time span of three days. Unlike other charts we tested, this one allows to draw points corresponding to each satellite, even when lacking the data to complete the sequence. This is important when a satellite is not visible (when it passes behind the planet), this allows the discontinuity in the display, making it a visible educational tool in the chart.

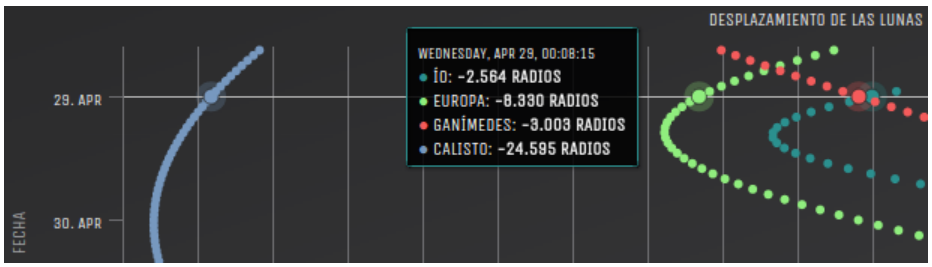


Fig. 5. Labels and points in the chart created with Highcharts. Upon hovering, the chart can display a tooltip text with information on each point and series.

Highcharts also allowed us to show dynamic labels with the name, time and location of each satellite. The tooltip follows the user's moves of a mouse or touchpad over the graph, it sticks to the nearest point and makes it easy to read a point that is below another one. We can also configure the shape, size or color, as shown in Fig 5.

Another of the most interesting options is the generating if an instant chart capture in PNG, JPEG, PDF, SVG format or send it to print, as can be seen in Fig 6. This useful feature for users to save the chart at any given time, either for an interesting event reflected in the graph, or to save a record of the positions movement for various ends.

Fig. 6. Menu of options of capture in the chart. Users can export the chart to PNG, JPG, PDF or SVG format at the click of a button, or print the chart directly from the web page.

4 Discussion

From the results obtained during the time that we have been working with these tools, we can consider our choice of tools as appropriate. The search done among available tools yielded solid arguments to decide in favor of the best for graphics and simulations. The choices were tested and configured to meet the needs of our project. Some tools offered unexpected features that were quite useful and enriched the nature of the application, for example, capture options on charts generated by Highcharts.

Better tools are available, but they require some sort of license payment to use and this project is budget-restricted. Our selection has secured the best results for a public and free application which is focused in offering an educational tool on astronomy for the general Spanish speaking public.

5 Conclusions

For this project we choose some of the best free tools and open resources. By these and the publishing of our code under GNU license, we assure the possibility for its future use or improvement among the Spanish-speaking population interested in this issue by providing them our source code. By using these tools, we can conclude that this work provides useful features and solutions to the needs of the project, producing satisfactory results in order to develop a free app of the major moons of Jupiter, Saturn, Uranus and Neptune

and only fewer problems were encountered. We are creating more tools for outreach/education in astronomy that are not currently available in Spanish (the app and the open source code) which we hope will capture the interest of enthusiast, amateur astronomers and the general public.

References

1. Silva, A., De Leo, M., Cuevas, R.: Análisis del actual desarrollo de aplicaciones móviles sobre astronomía para dispositivos basados en Android. Tlamati Sabiduría. 5, 96–103 (2014).
2. Stark, J.: Building Android apps with HTML, CSS, and JavaScript. O'Reilly, Sebastopol, CA (2010).
3. Tipos de apps: nativas, híbridas y Web Apps, http://appio.es/tipos-de-apps/.
4. Voskoglou, C.: Getting your priorities straight, http://www.developereconomics.com/report/q1-2014-getting-your-priorities-straight/.
5. Brooks, Michael: PhoneGap Developer App, http://phonegap.com/blog/2014/04/23/phonegap-developer-app/, (2014).
6. Prego, I.: PhoneGap, Appcelerator e Icenium: Crea apps para iOS y Android con HTML5, CSS y Javascript | ivanprego.com, http://ivanprego.com/diseno-web/css/desarrollar-aplicaciones-para-ios-y-android-con-html5-css-y-javascript/, (2013).
7. Rodríguez, T.: Chrome Dev Editor, el IDE de Google para programar desde Chrome, http://www.genbetadev.com/herramientas/chrome-dev-editor-el-ide-de-google-para-programar-desde-chrome.
8. Michalakos, D., Vakulenko, M., Constantinou, A.: How can HTML5 compete with Native?, https://www.developereconomics.com/reports/can-html5-compete-native/.
9. Google Trends - Web Search interest - Worldwide, 2004 - present, https://www.google.com/trends/explore?q&hl=en-US.
10. Rowell, E.: Web Graphics Trends in 2013, http://www.html5canvastutorials.com/articles/web-graphics-trends-in-2013/.
11. Highcharts product, http://www.highcharts.com/products/highcharts.
12. akkana/scripts, https://github.com/akkana/scripts.
13. Meeus, J.: Astronomical Algoriths. Willmann-Bell, Inc., Richmond, Virginia (1998).
14. Moons of Jupiter - Aplicaciones de Android en Google Play, https://play.google.com/store/apps/details?id=pl.bizcalc.jupitergalileanmoons&hl=es.

Unconventional Computing to Estimate Academic Performance in University Freshmen Students

Carolina Fócil Arias, Amadeo José Argüelles Cruz, and Itzamá López Yáñez

[1] Centro de Investigación en Computación
[2] Centro de Innovación y Desarrollo Tecnológico en Cómputo

focil.carolina@gmail.com, jamadeo@cic.ipn.mx, ilopez@ipn.mx

Abstract. Tracking academic performance of students with different levels is a topic of actual relevance. For undergraduate students, in the specific field of Mathematics, different techniques of pattern recognition have been applied to estimate their academic performance. In this study, we propose the use of a unconventional model called Gamma classifier for estimating academic performance, in freshmen student, applied to a dataset provided by the Technological University of Pereira, located in Colombia. The results revealed that Gamma classifier is a competitive algorithm, helping to determine whether a freshman student will pass Mathematics course.

Keywords: Academic Performance, Gamma Classifier, Machine Learning, Wilson Editing, Data Cleaning

1 Introduction

The little interest for the studies, the lack of abilities for developing solutions to reasoning problems, among others academic characteristics, are some factors that affect students performance in Mathematics course, which is related with student desertion. In spite of all services and programs efforts to avoid student desertion, only half of them get a bachelors degree [7].

Technological University of Pereira (located in Colombia) is an example of student desertion, it has records of roughly 52% of dropout between 1994 and 2004, where, 18.3% of students left the university due to poor academic performance in Mathematics in the first year[4].

This research is oriented as an alternative, applied with other techniques based on data and evidence, to reduce the dropout in freshmen engineering study and take actions over needed interventions in Mathematics course. We propose the use of a dataset provided by Technological University of Pereira and an unconventional algorithm called Gamma classifier which has been used in several applications [13][12]. Then, results are compared with those obtained in

[4] and other seven classifiers provided by a system called WEKA [16], with the purpose to determine whether a student will pass or no Mathematics course.

At the same time, data cleaning and Wilson editing techniques are also being applied to improve classifier performance, since dataset has many missing values.

The rest of this paper is structured as follows. Section 2 describes related work using patterns recognition approach. Section 3 describes the techniques used for determining whether a student will pass or not Mathematics course. Section 4 summarizes the results of using Gamma classifier and the techniques described in section 3. Finally, conclusions and future works are shown in section 5.

2 Related Work

Patterns recognition have been applied in several studies which have allowed to improve education in several levels and areas. In this section, we present some related studies with education and the use of several approaches to evaluate academic performance. For example, University of Singhaniya located in India [3] used a classification task to evaluate student's performance at the end of the semester in accordance with quiz, seminar, attendance and assiggnment of each student. The results obtained is based on decision tree approach.

University Technological of Pereira in Colombia [4] applied a Logistic Multiple Regresion model as proposal to identify the factors that essentially allow to predict whether a freshman engineering student will pass or not Mathematics course, in essence social and academic factors were used.

Another researches such as [7][14][2] have used Support Vector Machines into classification area. The proposed project by [7] is able to predict freshmen student attrition according to the past and present educational success. A dropout method for e-learning courses is reported by [14] which determines whether a student will pass or not Networks and Web design courses using the methods Neuronal Networks, Support Vector Machines and Probabilist approach. The School of Physical Education and Sports at Cukurova University located in Turkey [2] used Support Vector Machines to predict whether a candidate will be admitted in University of Cukurova based on a physical test.

A software called MUSKUP (Mugla University Student Knowledge Discovery Unit Program) was developed by [11] to identify essentially facts that affect the success in a student using the method of Decision Tree.

One of the most popular data mining techniques used to improve educational standards are association rules [16]. The study [1] extracted from a dataset some association rules with the aim of determining how many students are inscribed in a program but they are not interested on it. Another work [6] identified why some students did not finish their career in a period of years not less than 6 years.

3 Material and Methods

In this study, we have used an unconventional model called Gamma classifier and two preprocessing techniques: Data Cleaning and Wilson Editing with the aim of determining with higher precision whether a freshman engineering student will pass or not Mathematics course. To validate Gamma classifier performance a stratified k-fold cross validation was used. In stratified k-fold cross validation each fold contains approximately the same proportion of predictor labels as the original dataset [7].

3.1 Dataset

The dataset used in this study came from Technological University of Pereira located in Colombia with an enrollment of 834 freshmen engineering students and it was recorded during a period of 2005 to 2007. The dataset has two classes for identifying when a student pass or not Mathematics course. In this case, the class takes the value of 0 whether a student did not pass Mathematics course and it takes the value of "1" whether a student passed Mathematics course. For better interpretation, the dataset has 481 students that did not pass and 353 students that passed Mathematics course.

The variables used in this study are related with academic performance and social behavior. Table 1 shows the variables used in this experimental study and their selection is based on [4].

Table 1: Description of the variables used in this study

Variables	Description
ICFES	Points obtained from Colombian Institute for the Promotion of Higher Education
CodProg	Program code
Cabstract	Abstract Logical Thinking

According to Technological University of Pereira, the academic programs used in this study are Electrical Engineering, Industrial Engineering, Mechanical Engineering, Electrical Technology, Industrial Technology, Mechanical Technology, Chemical Technology, Engineering Systems and Computing, Physical engineering, Systems Engineering and Industrial Engineering.

3.2 Data Preprocessing: Data Cleaning and Wilson Editing

Data preprocessing is a task that improves the quality of the data before they are used to their analysis [15][9]. Due to the fact that dataset has many missing

values and based on type of variable, we have decided to use the attributes mean and mode to fill them. This means that whether the variable is numerical, we will use mean and we will use mode on nominal variables. For example, ICFES is a numeric variable then we will use mean by class. Using this approach, we have replaced 201 patterns by meand and mode.

Another preprocessing technique used in this study is Wilson Editing. This technique eliminates all patterns that are misclassified using KNN algorithm where the value of K is 3. Wilson Editing allows to increment the classifier's performance and to eliminate outliers. Wilson editing algorithm is shown below [10][8].

Initialization;
$S \leftarrow X$
for $x \in X$ **do**
  if *it is misclassified using the KNN rule with prototypes in* $X - \{x_i\}$ **then**
    $S \leftarrow S - \{x_i\}$;
  end
end

Algorithm 1: Wilson Editing Algorithm

3.3 Gamma classifier

Gamma classifier is an unconventional algorithm which is based on alpha and beta operators and its operation is completely based on binary representation into the modified Johnson Moebius.[13][12]

Definition 1: The operators Alpha and Betha are defined in table 2, given the sets A={0,1} and B={0,1,2}.

Table 2: Alpha and Beta operators

(a) Alpha Operator

$\alpha : A \times B \rightarrow B$		
x	y	$\alpha(x, y)$
0	0	1
0	1	0
1	0	2
1	1	1

(b) Beta Operator

$\beta : B \times A \rightarrow A$		
x	y	$\beta(x, y)$
0	0	0
0	1	0
1	0	0
1	1	1
2	0	1
2	1	1

1. Code the fundamental set into the code Modified Johnson Moebius and to obtain the value of e_m for each component.

$$e_m = \bigvee_{i=1}^{\rho} x_j^i \tag{1}$$

2. Compute the stop parameter.

$$\rho = \bigvee_{j=1}^{n} e_m(j) \tag{2}$$

3. Code the test pattern y with the Modified Johnson Moebius. If y_j is greater than e_m, Gamma operator will use y_j instead of e_m.
4. Transform the index of all patterns into two indices, one for their class and another for their position in the class.
5. Define the weight of each dimension Suggested empirical values are detailed below:

   a Within the range [1.5, 2] to features that are separable.
   b Within the range (0,0.5] to features that are not separable.

6. Initialize $\theta = 0$
7. Compute $\gamma_g(x_j^{i\omega}, y_j, \theta)$ for each component of the fundamental patterns in each class, according to:

$$\gamma_g(x_j^{i\omega}, y_j, \theta) = \begin{cases} 1 & \text{if} \quad m - u_\beta[\alpha(x,y) mod 2] \leq \theta \\ 0 & \quad Otherwise \end{cases} \tag{3}$$

8. Compute the weighted sum c_i for each class, as follows:

$$c_i = \frac{\sum_{\omega=1}^{k_i} \sum_{j=1}^{n} \gamma_g(x_j^{i\omega}, y_j, \theta)}{k_j} \tag{4}$$

9. If there is more than a unique maximum among c_i, increment θ by 1 and repeat steps 7 and 8 until to find a unique maximum or the stop parameter is fulfilled with the condition $\theta \leq \rho$.
10. If there only one maximum, assign y to the class which correspond such maximum.

$$C_y = C_j such that \bigvee_{i=1}^{m} c_i = c_j \tag{5}$$

11. Otherwise, assign y to the class of the first maximum.

4 Experimental Results

The aim of this study is to determine with higher precision whether a student will pass or not Mathematics course and overcome the results presented in [4] using the method of Gamma classifier and in accordance with the obtained results, we have considered seven classifiers provided by a system which has a collection of machine learning used basically to data mining task called WEKA (Acronym for Waikato Environment for Knowledge Analysis)[16] to compare the results. For our purpose, we used WEKA 3.6.10 version software and the algorithms are listed below.

1. Naive Bayes
2. Bayes Net
3. Support Vector Machines
4. Simple Logistic
5. Logistic
6. KNN where K=3
7. Tree Decision: J48 algorithm

Table 3: Confusion Matrix

Predicted class	**True class**	
	Positive	**Negative**
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

The elements can be detailed as follows [2]:

- True Positive (TP) is the number of correct predictions that an instance is positive.
- False Positive (FP) is the number of incorrect predictions that an instance is negative.
- False Negative (FN) is the number of incorrect predictions that an instance is positive.
- True Negative (TN) is the number of correct predictions that an instance is negative.

Using the elements described above, we can obtain the following measures:

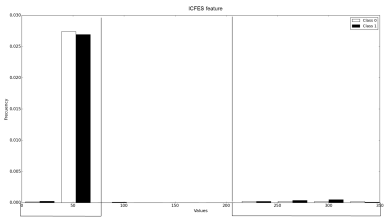
$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$precision = \frac{TP}{TP + FP} \tag{7}$$

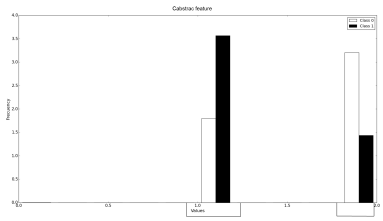
$$recall = \frac{TP}{TP + FN} \tag{8}$$

$$F - Measure = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} \tag{9}$$

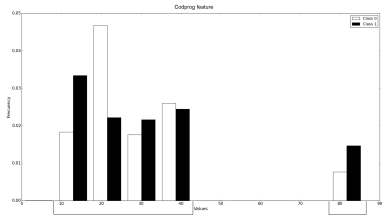
First, we applied data cleaning and Wilson Editing to dataset provided by Technological University of Pereira [4]. Then, based on Gamma algorithm, a combination of weights were tested to each feature. The combination of weights is arranged according to the histograms and were selected when Gamma classifier was run and provided the best results. Figure 1 shows the feature's histograms.



(a) ICFES



(b) Cabstrac



(c) CodProg

Fig. 1: Histograms of each feature used in this study

Based on the histograms, we can see that the classes are not separated, this means that the features have almost the same values, thus we have decided to assign a weight of 0.1 instead of 0. Table 4 details a final combination of these weights for each feature with the aim of obtaining better results.

Table 4: Proposed weights

Variables	Weight
ICFES	0.5
CodProg	0.1
Cabstract	0.1

The following step is to validate Gamma classifier a stratified k-fold cross validation was used in this study. Finally,Gamma classifier was run and it was compared with the algorithms provided by WEKA [16].The results of this experiment are detailed in table 5.

Table 5: Performance Measures

Classifier	% Acc.	Precision		Recall		F-measure	
		0	1	0	1	0	1
KNN	94.48	0.932	0.885	0.913	0.91	0.922	0.897
J48	91.14	0.893	0.836	0.875	0.859	0.884	0.847
Bayes Net	81.78	0.81	0.832	0.892	0.718	0.849	0.771
Gamma	**77.93**	**0.8576**	**0.7629**	**0.7376**	**0.8353**	**0.7931**	**0.7634**
SMO	77.60	0.857	0.698	0.732	0.835	0.789	0.761
Simple Logistic	77.26	0.856	0.694	0.726	0.835	0.785	0.758
Logistic	76.93	0.855	0.689	0.72	0.835	0.782	0.755
Naive Bayes	76.92	0.847	0.693	0.729	0.824	0.784	0.753

Some works have been made by Technological University of Pereira using the same dataset that is being used in this research and the purpose of the table 6 is to show the results obtained by them [4] [5].

Table 6: A comparison between Gamma Classifier and Logistic Multiple Regresion

Classifier	% Performance
Gamma classifier	**77.93%**
Logistic Multiple Regresion	70.40%
Logistic Multiple Regresion	61.8%

A unconventional classifier to determine whether a student will pass or not Mathematics course was applied and in order to compare its result, we have considered to use some algorithm provided by WEKA [16] according with the aim

of the study. The results for this experiment shows that KNN produced the best result followed by J48 and Bayes Net with a performance of 94.4816%,91.1371% and 81.7726%, respectively. However, Gamma classifier produced a result of 77.9264% and it is evident that it overcame the results published in [4] (Details are given in table 6) as well as got better result than SMO with a classification rate of 77.592% followed by Simple Logistic with a classification rate of 77.2575%, Logistic with 76.9263% and finally, Naive Bayes with a classification rate of 76.9231%.

With regards to the results in this experiment, Gamma classifier can be used like a border to determine algorithms with high performance and algorithms competitive. It can be seen from the table 3 that Gamma classifier has competitive results in front of some algorithms provided by WEKA [16] and it can be used like a solution for determining whether a student will pass or not the Mathematics course.

5 Conclusions

This study presents a unconventional algorithm called Gamma classifier and two data preprocessing: Data cleaning and Wilson Editing approach for determining whether student will pass or not Mathematics course. The result obtained is competitive with a classification rate of 77.9264% and outperformed the results published by [4] with a classification rate of 70.4% and [5] with a classification rate of 61.8/

However, some considerations should be taken to determine more accurately the situation of a student in the Mathematics course. Firstly, adding social, academic, health risk and personal variables which can be very useful to obtain better result. Secondly, applying techniques for creating automatic weights. Finally, balancing methods for unbalanced dataset should be applied to improve the rate of classification.

References

1. Z. Abdullah, T. Herawan, N. Ahmad, and M. M. Deris. Mining significant association rules from educational data using critical relative support approach. *Procedia Soc. Behav. Sci*, 28:97–101, 2011.
2. M. Acikkar and M. F. Akay. Support vector machines for predicting the admission decision of a candidate to the school of physical education and sports at cukurova university. *Expert Syst. Appl.*, 36(3):7228–7233, 2009.
3. B. Baradwaj and S. Pal. Mining educational data to analyze student's performance. *Int. J. of Adv. Comput. Sci. Appl*, 2(6):63–69, 2012.
4. P. Carvajal, J. C. Mosquera, and I. Artamonova. Modelos de predicción del rendimiento académico en matemáticas i en la universidad tecnológica de pereira. *Sci. Tech.*, 43:258–263, 2009.

5. P. Carvajal, J. C. Mosquera, and I. Artamonova. Rendimiento en matemáticas i en la universidad tecnológica de pereira. *Sci. Tech*, 41:379–383, 2009.
6. M. Chalaris, I. Chalaris, C. Skourlas, and A. Tsolakidis. Extraction of rules based on students' questionnaires. In *2nd International Conference on Integrated Information*, volume 73, pages 510–517, 2013.
7. D. Delen. A comparative analysis of machine learning techniques for student retention management. *Decis. Support Syst*, 49(4):498–506, 2010.
8. C. F. Eick, N. Zeidat, and R. Vilalta. Using representative-based clustering for nearest neighbor dataset editing. In *IEEE International Conference on Data Mining (ICDM-04)*, pages 375–378, 2004.
9. C. Hernández G. and J. Rodríguez R. Preprocesamiento de datos estructurados. *Vínculos*, 4(2):27–48, 2008.
10. S. L. Donghai Guan, Weiwei Yuan, and Young-Koo Lee. Semi-supervised nearest neighbor editing. In *IEEE International Joint Conference on Neural Networks (IJCNN 2008)*, volume 8, pages 1183–1187, 2008.
11. H. Guruler, A. Istanbullu, and M. Karahasan. A new student performance analysing system using knowledge discovery in higher educational databases. *Comput. Educ*, 55(1):247–254, 2010.
12. C. López-Martín, I. López-Yáñez, and C. Yáñez-Márquez. Application of gamma classifier to development effort prediction of software projects. *Nat. Sci. Publ.*, 418(3):411–418, 2012.
13. I. López-Yáñez. Clasificador automático de alto desempeño (in spanish). Master's thesis, Center for Computing Research, National Polytechnics Institute, 2007.
14. I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, and and V. Loumos G. Mpardis. Dropout prediction in e-learning courses through the combination of machine learning techniques. *Comput. Educ*, 53(3):950–965, 2009.
15. D. F. Nettleton. Data mining of social networks represented as graphs. *Comput. Sci. Rev*, 7:134, 2013.
16. I. H. Witten and E. Frank. *Data Mining Practical Machine Learning Tools and Techniques*. Morgand Kaufmann Publishers, 2nd edition, 2005.

Collaborative Reputation Mechanism for Cloud Storage Service

Gerardo García-Rodríguez and Francisco de Asís López-Fuentes

Departament of Information Technology
Universidad Autónoma Metropolitana-Cuajimalpa (UAM-C)
Av. Vasco de Quiroga 4871, Cuajimalpa, Santa Fe
05348 Mexico City, México
{flopez}@correo.cua.uam.mx

Abstract. Cloud computing technology has emerged during the last years as a promising solution to transform the IT industry. Cloud computing is able to offer storage service, computing power and flexibility to end-users. Nevertheless, cloud computing technology still faces several challenges such as privacy, robustness, security and throughput. Distributed infrastructures as P2P networks have emerged as promising solutions for the management and storage all data. This paper proposes a collaborative mechanism based on reputation for cloud storage services. The proposed mechanism is implemented on a P2P infrastructure, which is used as an alternative platform for deploying storage services. Our solution integrates a qualified storage mechanism based on reliability indices. All peers collaborate to build the individual reputation of each peer in the storage system.

Keywords: P2P networks, Cloud computing, Distributed systems.

1 Introduction

Recently, cloud computing has become more popular, and many users and companies use these services to obtain several benefits such as more store capacity or computing power. Cloud computing is defined by M. Armbrust et al [2], as the applications delivery and services over the Internet, as well as the hardware and system software in the datacenters that provide these services. Based on this definition we can deduce that cloud computing is a model that allows access to files, applications or services in a ubiquitous and pervasive way through network in order to share a set of configurable computing resources. These resources can be servers, storage, applications and services, which can be rapidly provisioned and released with a minimal effort in service management or interacting with the provider. Therefore, cloud computing provides the illusion of unlimited and on-demand scalability. Essential characteristics of a cloud are [3]: on-demand self-service, network access, resource pooling, elasticity and measured service. In cloud computing two concepts are very basic: abstraction and virtualization [6].

Abstraction means that the implementation details of the system users and developers are abstracted. Therefore, applications run on physical systems that are not specified, the files are stored in places where users do not know their actual location, the system can be managed via outsourcing, and clients can access to the system in a ubiquitous manner. On the other hand, resources of the systems are pooled and shared in a virtualized way. Regarding virtualization, Sosinsky, B. states in [6] that systems and storage can be provisioned as needed from a centralized, costs are assessed on a metered basis, multi-tenancy is enabled, and resources are scalable with agility.

Cloud computing introduces several benefits such as massive computing power, storage capacity and great flexibility; however this new computing paradigm still faces several challenges. Most popular cloud systems are centralized and its structure is based on the client-server paradigm. A centralized structure introduces several limitations such as storage dependence, privacy, scalability, privacy locally or connectivity. Peer-to-peer (P2P) networks have emerged as a promising distributed information management platform. A P2P network is a distributed network formed by a group of nodes, which build a network abstraction on top of the physical network, known as an overlay network. A peer can take the role of both, a server and of a client at the same time. An important advantage of these networks is that all available resources such as processing, memory and bandwidth are provided by the peers. Thus, when a new peer arrives to the P2P system the demand is increased, but the overall capacity too. This is not possible in a distribution infrastructure based on the client server model with a fixed number of servers. P2P paradigm allows that a system distributes its load and duties between all participating peers. Several examples of cloud computing platform based on P2P networks are reported in the literature [4-5], [7-9]. In this paper, we describe operation of our collaborative reputation mechanism for a storage service based on P2P cloud system [12]. Reputation systems has been proposed with the purpose to ensure that peers obtain reliable information about the quality of the resources they are receiving [11]. The main characteristic of our proposed mechanism is its reputation strategy which is based on based on reliability indices. These indices are totally transparent to the user as it is in the centralized cloud computing. Our project aims to integrate shared resources as volunteer computing to be used as a source of computing power and storage in social communities or research groups.

The remainder of this paper is organized as follows. In Section 2, we introduce our proposed model for a P2P cloud storage service. We present our collaborative mechanism in Section 3, and we explain its operation. This paper concludes in Section 4.

2 Architecture

A distributed storage system is an infrastructure that allows to store files in nodes, which are connected through a computer network. These systems are characterized by their wide

range of applications such as backup files, sharing of files in network and edition of documents from different locations. This section introduces our proposed model. We propose a distributed storage service based on cloud computing, which takes advantage of the benefits introduced by the P2P networks. These networks have the advantage that all available resources are provided by the peers [1]. Peers can greatly benefit from the capacity of other requesting peers via collaboration. Thereby, file storage is realized in a decentralized manner and its location is completely transparent to the users. Our proposal differs from traditional distributed file systems which are equally used to share resources, but they have a lower level of transparency, and users know about the file system. A P2P Cloud system consists of peers which contain the client application. Peers come together to share their available hard disk capacity in order to create a total storage space together. The contents are split and each of these fragments is distributed within the peers. If content is not split, then it only is sent by the tracker to any of the nodes that is part of the cloud. Thus, a user can realize a replication of files within the cloud.

Our proposed architecture is shown in Figure 1. Reputation levels play an important role in our model, because they allow the system to define the way how files are stored within the peers. Our proposal considers five levels of reliability (or reputation) which are poor, fair, good, very good and excellent. These levels are recorded in each peer using values from 1 to 5, where 1 means that a peer has a poor reputation, while 5 means that a peer has an excellent reputation. Fair, good and very good scores are represented by the values 2, 3 and 4, respectively. When a tracker is created, it must contain the table of confidence (trust) levels which consists of an average of the levels of trust, dynamism and availability of each peer in the system.

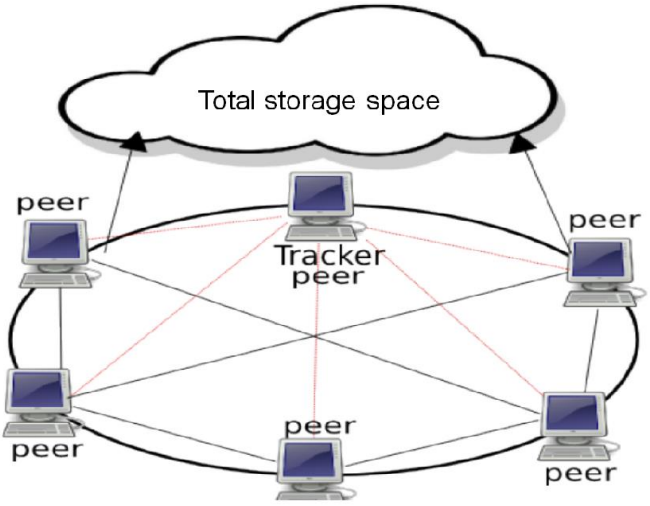


Fig. 1. Proposed architecture

The tracker decides where to store the files according to the level of reputation of each peer. For example, if the peer P1 has a level 5, then it can store its content on peers with same reputation level. In contrast, if P1 has a low reputation level its content is sent to peers with the same level of reputation. We use this strategy for the benefit of the cloud, and thus any type of contingency in the cloud as file access failure by peers with greater reliability. Figure 1 illustrates the tracker operation in the cloud. The tracker or server routes data and establishes communication between peers. These communications are based on the confidence level of each peer and contain the information required for each.

3 Mechanism

Our mechanism is composed by the following entities:

Peer. is an application hosted on each node in the P2P network, and it is responsible that each computer to be involved in the P2P cloud. Since a peer is an entity that receive and send files at the same time, we need to implement an application to perform these two tasks simultaneously. These tasks are realized through the peer application. Peer application also is responsible for monitoring each peer and reports its shared resources to tracker. Parameters to be monitored in each peer are its store capacity (trust), the number of disconnections during a day (dynamicity), its availability (number of storage requests that are rejected by a peer), and if a peer is cheating or no respect to a stored file. These three metrics are averaged in order to obtain the reliability level in each peer. After all, reliability is the metric used by a tracker to route content between peers. Peer application is formed by two parts: a server and a client. Server part always is listening in order to attend to other peers. In this case, to store files in the host computer. On the other hand, client part realizes different functions such as uploading files, display files and exit. Additionally, client program is who communicates with tracker in order to report all information about this computer in the cloud system. Figure 2 shows how an application hosted on a peer reports its data to the tracker while monitoring its host peer.

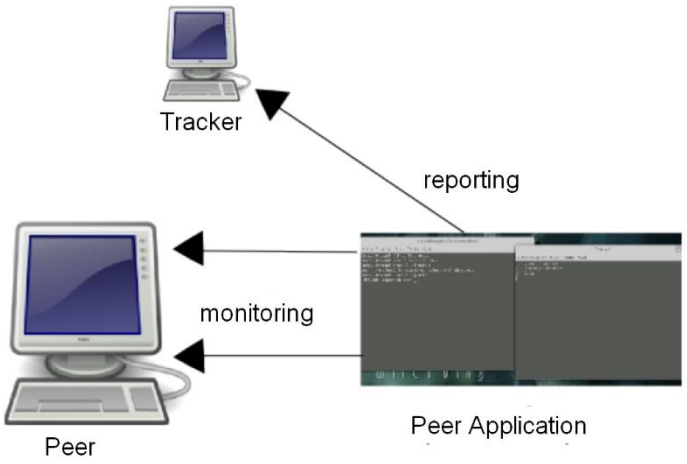


Fig. 2. Monitoring and reporting of the peer application

Tracker is responsible for system initialization. This application establishes the communications among peers. Figure 3 shows this scenario. Tracker applications also manipulates the database in which the reports generated by each peer and its contents are registered. Management of the database realized by the tracker is illustrated in Figure 4. It is important to note that the tracker is not a storage server, so the tracker never manage files. Tracker only redirects the contents received from a peer to another depending on levels of reliability of each peers. However, tracker offers localization transparency to all users in the systems. In this way, a peer sends its content to tracker, whom decides in which peer will be placed this content. This location is transparent for the requesting peer. When a requesting peer whishes recovery its content, it is requested via the tracker. Then, content is be addressed from the host peer to the requested peer. To allocate contents, the tracker queries the levels of dynamism, reliability and availability of each peer in the database and calculates an average integer value. The addressing is realized as follows. For example, if a peer P1 wishes to store a file in the cloud, it must submit a request to tracker, which selects the host peer. Peer P1 can receive an IP address to upload its file to the host peer by itself, or the file can be addressed by the tracker directly. Different host peers can have the same availability (or reputation) level. In this case, tracker selects peer with the largest storage space.

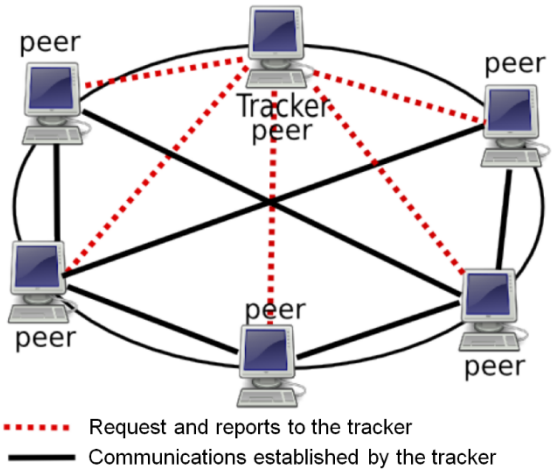


Fig. 3. Communication between nodes based on indications received from the tracker peer

Database, is designed to register and control all information received by the tracker from the peers. The database in this cloud system is based on a relational model. In its design we consider three entities: reputation metrics, peers and files. Recorded and monitored data in each peer are: physic address, IP address, date and time of the last connection, available space in disk, number of rejections, availability, dynamicity, true and reliability. Our design is based on the relational model and we used Wokbench [10] as a unified visual tool to produce a detailed data model of the database and tables. Visual database design is shown in Figure 5. Our scheme was developed in Spanish language. In this case, "confiabilidad" corresponds to reliability, while "dinamicidad", "disponibilidad" and "confianza" correspond to dynamicity, availability, and trust, respectively. Our database also captures information related to the behavior of each peer in the system. That is, if a node is cheating or not with respect to a stored file. Data recorded for each file are its name, size and location. Files are localized using the peer's physical address. Relational model is useful to maintain a right relationship between file, proprietary and storage place. Figure 6 shows how the peer P5 uploads a file to the cloud. We can see how the tracker is responsible for routing the file to a peer with the same level of reliability as P5. In this case tracker selects peers P3 to store the file of peer P5, because both peers have same level of reliability.

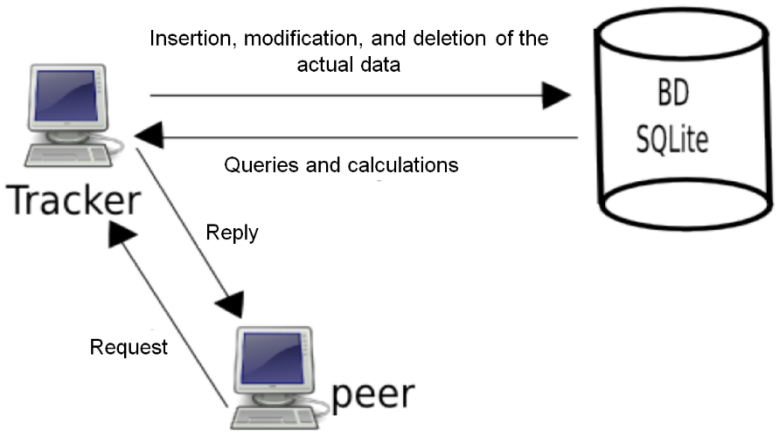


Fig. 4. Management of the database realized by the tracker peer

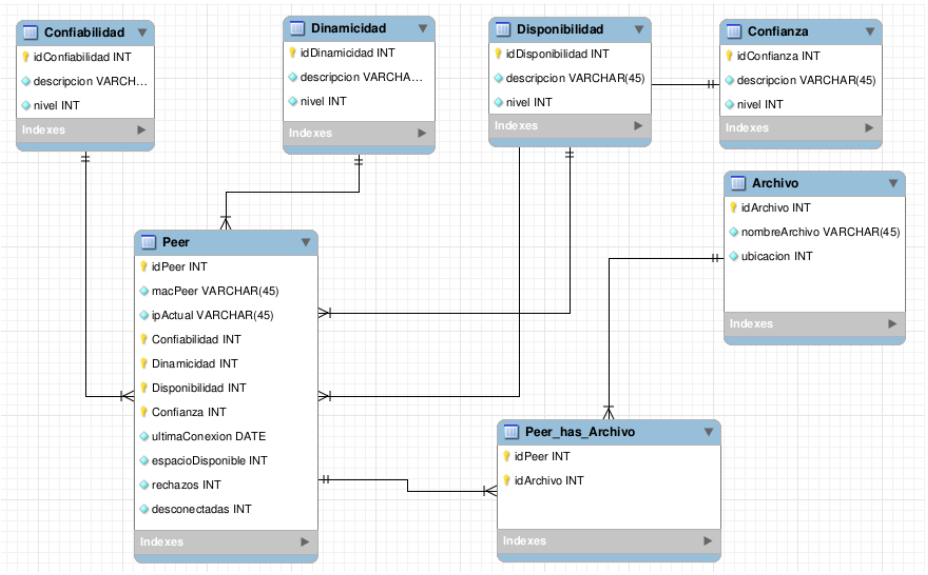


Fig. 5. Relational model of the database where all the information of peers and its contents are stored

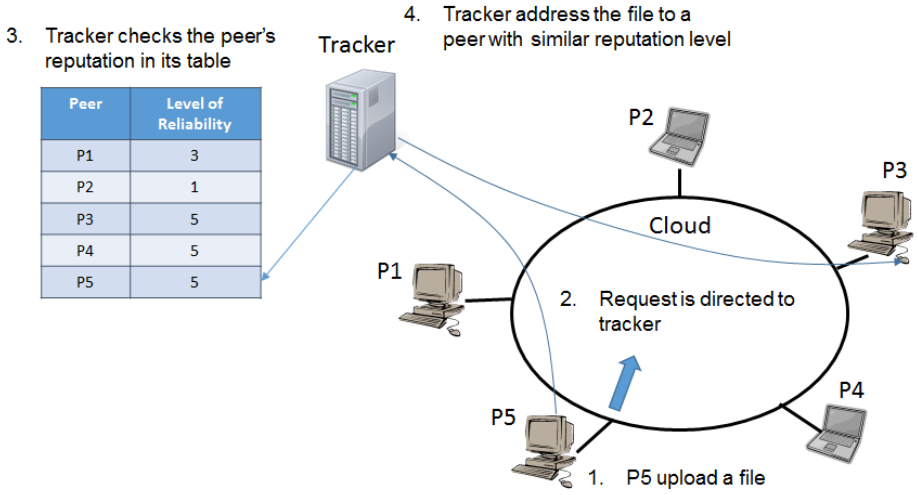


Fig. 6. System operation during a content distribution [12]

We implemented a small and experimental prototype of our proposal mechanism in our lab. This prototype has been developed on Linux Fedora version 19 using language C/C++. Our peers are communicated via TCP. We use this protocol because TCP is reliable, ordered and provides error-checked delivery between applications running on Internet.

4 Conclusions

Cloud computing have emerged as an ideal solution for storage service during the last years. This paper proposes a collaborative mechanism for cloud storage service based on P2P networks. Our solution exploits several characteristics introduced by the P2P networks such as collaboration, flexibility and scalability. Our proposed mechanism considers different approaches such as reputation and collaborative storage. Compared with other storage models such as cloud storage based on client-server, distributed storage or P2P distributed storage, our proposed model offers different benefits such as confidentiality, file sharing, data replication, data management, quality of service, decentralization, and transparency. A limitation of our proposed model is file fragmentation.

Although our experimental prototype was implemented using a reduced number of nodes, it can be scaled due to the properties of the P2P paradigm. In the future we plan to integrate a fragmentation policy in order to manipulate large multimedia contents.

Specifically, we are interested in adding a heterogeneous fragmentation policy in peers with heterogeneous reputation in our prototype.

References

1. Milojicic, D., Halogeraki, V., Lukose, R., Nagaraja, K., Pruyne, J., Richard B., Rolling, S. and Xu, Z.: Peer-to-Peer Computing. In: HP Labs Technical Report HPL-2002-57, (2002).
2. Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R. H., Konwinski, A., Lee, G., Patterson, D. A., Rabkin, A., Stoica, I. and Zanaria, I.: Above the Clouds: A Berkeley View of Cloud. In: Technical Report No. UCB/EECS-2009-28, University of California at Berkeley, Berkeley, CA, USA (2009).
3. Mell, P. and Grance, T.: The NIST Definition of Cloud Computing. In: Special publication 800-145 (draft), Gaithersburg, MD, USA (2011).
4. Lakshman, A. and Malik, P.: Cassandra - A Decentralized Structured Storage System. In: Proc. of the Workshop on Large-Scale Distributed Systems and Middleware (LADIS'09), Big Sky, MT, USA (2009).
5. Babaoglu, O., Marzolla, M. and Tamburini, M.: Design and Implementation of a P2P Cloud System. In: Proc. of the 27th Annual ACM Symposium on Applied Computing (SAC'12), pp. 412–417, Trento, Italy (2012).
6. Sosinsky, B.: Cloud Computing Bible. In: Wiley Publishing Inc., Indianapolis, IN, USA (2011).
7. Cunsolo, V., Distefano, S., Puliafito, A. and Scarpa, M.: Cloud@home: Bridging the gap between volunteer and cloud computing. In: Proc. of the 5th Int. Conference on Emerging Intelligent Computing Technology and Applications (ICIC'09), Ulsan, South Korea (2009).
8. Trajkovska, I., Salvachua Rodriguez, J and Mozo-Velasco, A.: A Novel P2P and Cloud Computing Hybrid Architecture for Multimedia Streaming with QoS Cost Functions. In: Proc. of the ACM International Conference on Multimedia (MM'10), pp. 1227-1230, Firenze, Italy (2010).
9. Xu, K., Song, M., Zhang, X., and Song, J. A.: Cloud Computing Platform based on P2P. In: Proc. of the IEEE International Symposium on IT in Medicine Education, pp. 427-432, Shandong, China, (2009).
10. MySQLWorkbench. Web site: http://www.mysql.com/products/workbench/.
11. Kamvar, S. D., Schlosser, M. T. and H. García-Molina, H.: The EigenTrust algorithm for reputation management in P2P networks. In: Proc. 12th Int. World Wide Web Conference, Budapest, Hungary (2003).
12. García-Rodríguez G., and López-Fuentes, F.: A Storage Service based on P2P Cloud System. In: Research in Computing Science 76, pp.89-96 (2014).
13. García-Rodríguez G.: Cómputo en la Nube basado en Redes Peer-to-Peer enfocados al Almacenamiento. Proyecto Terminal, Licenciatura en Tecnologías y Sistemas de la Información, UAM-Cuajimalpa, (2014).

Embedded System to Solve the Laplace's Equation

Miguel Ángel San-Pablo-Juárez[1], Eduardo Morales-Sánchez[1], and Fernando Ireta-Moreno[2]

[1] Centro de Investigación en Ciencia Aplicada y Tecnología Avanzada unidad Querétaro, Instituto Politécnico Nacional
miguelangel.sanpablo@gmail.com
[2] División de Ingenierías campus Irapuato-Salamanca, Universidad de Guanajuato
fireta@ugto.mx

Abstract. In this work is solved the Laplace's equation in two dimensions with Dirichlet boundary conditions using the Finite Element Method. Solution algorithms were designed and were loaded into a 32-bit microcontroller STM32. The programmed embedded system may be used in different applications such as solution potential in electrostatics, temperature distribution in heat equation, bubble theory or in image processing. The aim of this work is to solve the Laplace'e equation using an embedded system, instead of using a typical PC.

Keywords: STM32, Laplace's equation, Finite Element Method, Potential

1 Introduction

We present a numerical solution for solving Laplace's equation on two-dimensional irregular regions, using Dirichlet boundary conditions, the solution is implemented on an embedded system that is very fast in time for numerical computations. This approach uses a finite element discretization and avoid the use of a Personal Computer. Solutions for Laplace's equation have been studied in works as[1], based on iteratively solving integral equations; in [2], for mixed boundary value conditions; [3] using a Finite Volume Method and [4] where is used a regularized integral equation method. Studies about Poisson's equation was made too in [5], showing a finite-volume discretization or in [6], where is used a finite difference method. But note that no one have implemented a solution into a microcontroller embedded system yet.

1.1 Laplace's Equation

The partial differential equation that involves $u_{xx}(x, y) + u_{yy}(x, y)$ is called an elliptic equation. A particular elliptic equation we will consider is known as the

Laplace's equation [7]

$$\frac{\partial^2 u}{\partial x^2}(x,y) + \frac{\partial^2 u}{\partial y^2}(x,y) = 0 \tag{1}$$

Laplace's equation is often mentioned as a potential equation [8], because the primary task of electrostatics is to find the electric field of a given stationary charge distribution [9] and the best strategy for finding the electric field is first to calculate the potential, V. In this case, equation (1) reduces to Laplace's equation:

$$\boldsymbol{\nabla}^2 V = 0 \tag{2}$$

Where potential function $V(x,y) = u(x,y)$, ∇^2 represents the Laplacian, V the potential, with a charge density $\rho = 0$. This formula is so fundamental to the subject that one might almost say electrostatics *is* the study of Laplace's equation [9]. Some of the most used methods to solve Laplace's Equation are: Separation of Variables, Finite Difference Method and the Finite Element Method.

1.2 The Finite Element Method

The Finite Element Method (FEM) is useful in solving differential equations [10]. The finite element analysis of any problem involves four steps: (a) discretizing the solution region into a finite number of subregions or elements, (b) deriving governing equations for a typical element, (c) assembling of all elements in the solution region, and (d) solving the system of equations obtained [10]. FEM is the most commonly method used to solve the problem, as it is very easy to apply to problems with any type geometries [7].

Finite Element Discretization The solution region is divided into a number of finite elements as in Fig. 1, where the region is subdivided into six elements (all triangular) and seven nodes. We seek an approximation for the potential V_e within an element e and then interrelate the potential distributions in various elements such that the potential is continuous across interelement boundaries [10]. The approximate solution for the whole region is [7]:

$$V(x,y) \simeq \sum_{e=1}^{N} V_e(x,y) \tag{3}$$

where N is the number of triangular elements into which the solution region is divided. The most used approximation for V_e within an element is a polynomial approximation of the form [7]

$$V_e(x,y) = a + bx + cy \tag{4}$$

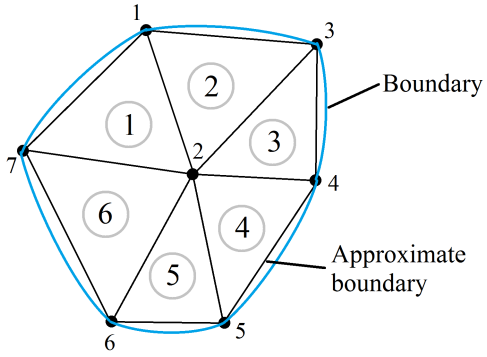


Fig. 1. A typical finite element subdivision used for an irregular domain

in this case for a triangular element. Notice that our assumption of linear variation of potential within the triangular element as in equation (4) is the same assuming that the electric field is uniform within the element; that is, [10]

$$\mathbf{E}_e = -\boldsymbol{\nabla} V_e = -(b\mathbf{a}_x + c\mathbf{a}_y) \tag{5}$$

Element Governing Equations Consider a triangular element as shown in Fig. 2, The potential V_{e1}, V_{e2} and V_{e3} at nodes 1, 2 and 3 respectively, are obtained using equation (4)

$$\begin{bmatrix} V_{e1} \\ V_{e2} \\ V_{e3} \end{bmatrix} = \begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} \tag{6}$$

And the coefficients a, b and c can be determined from equation (6)

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix}^{-1} \begin{bmatrix} V_{e1} \\ V_{e2} \\ V_{e3} \end{bmatrix} \tag{7}$$

Equation (4) can be rewritten as

$$V_e = \begin{bmatrix} 1 & x & y \end{bmatrix} \frac{1}{2A} \begin{bmatrix} (x_2y_3 - x_3y_2) & (x_3y_1 - x_1y_3) & (x_1y_2 - x_2y_1) \\ (y_2 - y_3) & (y_3 - y_1) & (y_1 - y_2) \\ (x_3 - x_2) & (x_1 - x_3) & (x_2 - x_1) \end{bmatrix} \begin{bmatrix} V_{e1} \\ V_{e2} \\ V_{e3} \end{bmatrix} \tag{8}$$

where A is the area of the element e. Then (8) can be expressed as

$$V_e = \sum_{i=1}^{3} N_i(x, y) V_{ei} \tag{9}$$

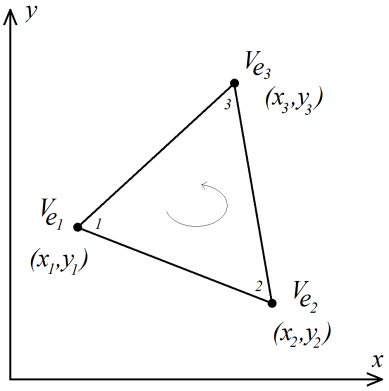


Fig. 2. A typical triangular element. The local numbering could be as 1-2-3.

where N_i represent linear interpolating functions. They are called *element shape functions*.

Now, the energy per unit length associated with the element e is given by [10,9]:

$$W_e = \frac{\varepsilon}{2} \int \mid \mathbf{E} \mid^2 dS = \frac{\varepsilon}{2} \int \mid \boldsymbol{\nabla} V_e \mid^2 dS \tag{10}$$

assuming a two dimensional solution region free of charge ($\rho_S = 0$). Now if it is applied the gradient to (9) this is

$$\boldsymbol{\nabla} V_e = \sum_{i=1}^{3} V_{ei} \boldsymbol{\nabla} N_i \tag{11}$$

and substituting (11) into (10)

$$W_e = \frac{\varepsilon}{2} \sum_{i=1}^{3} \sum_{j=1}^{3} V_{ei} \left[\int \nabla N_i \cdot \nabla N_j dS \right] \tag{12}$$

If is defined the term in brackets as

$$C_{ij}^{(e)} = \int \boldsymbol{\nabla} N_i \cdot \boldsymbol{\nabla} N_j dS \tag{13}$$

equation (12) becomes

$$W_e = \frac{\varepsilon}{2} \left[V_e \right]^T \left[C^{(e)} \right] \left[V_e \right] \tag{14}$$

Asembling all elements Finally is needed to assemble each element in the solution region, for this, the energy associated with the assemblage of all elements is required

$$W = \sum_{e=1}^{N} W_e = \frac{1}{2}\varepsilon[V]^T[C][V] \tag{15}$$

with

$$[V] = \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_n \end{bmatrix} \tag{16}$$

N is the number of elements, n is the number of nodes and $[C]$ is the *global coefficient matrix*, that is the assemblage of each individual element coefficient matrices.

$$[C] = \begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1\,41} \\ C_{21} & C_{22} & \cdots & C_{2\,41} \\ \vdots & \vdots & \ddots & \vdots \\ C_{41\,1} & C_{41\,2} & \cdots & C_{41\,41} \end{bmatrix} \tag{17}$$

Here C_{ij} is obtained with the fact that the potential distribution must be continuous across interelement boundaries. The size of C depends on the amount of involved nodes and C_{ij} is the coupling between nodes i and j. Some properties of the C_{ij} matrix are:

- Is symmetric ($C_{ij} = C_{ji}$) as the element coefficient matrix
- Since $C_{ij} = 0$ if does not exist coupling between nodes i and j, for a large number of elements $[C]$ becomes sparse and banded.
- It is singular.

Solving the Resulting Equations From variational calculus, it is known that Laplace's equation is satisfied when the total energy in the solution region is minimum. We require that the partial derivatives of W with respect to each nodal value of the potential be zero. Using the Band Matrix Method together with Gauss-Jordan method it is possible to minimize the total energy in the solutions region (derivatives of W respect each nodal value of the potential be zero). Finally the vector solution of equation (16) can be found and a potential distribution with equation (9) may be obtained. Because of a separation of fixed and free nodes in the Band Matrix method, free element coefficient matrix is not singular and it is possible to inverse for finding free node potentials. Once analytically analyzed the problem, an algorithm for solving the problem can be designed to implement into the STM32 microcontroller. Fig. 3 shows the solution algorithm to solve Laplace's equation in a two dimensions region with Dirichlet boundary conditions.

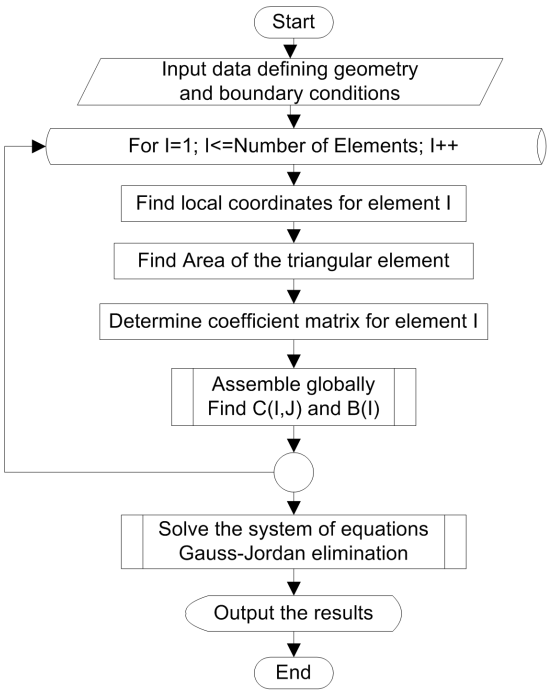


Fig. 3. Algorithm to solve Laplace's equation.

1.3 Implementation on the STM32F Discovery

Laplace's equation solution by the FEM is typically made using a PC, with solution algorithms implemented in compilers of different high-level languages or using specialized commercial software. But some specific applications require calculations programmed into embedded systems. Examples are: Electrostatics, Electrical Impedance Tomography, temperature measurements, image analysis an so on. A big advantage to solve a differential equation in an embedded system is the reduction of space in the final implementation. Another advantage is that the system performs only a particular task [11], in this case the solution of an elliptic differential equation. Power consumption and portability are other advantages.

In this work, a microcontroller (MCU) of the family STM32 32-bit ARM Cortex was used. The STM32 family of 32-bit Flash microcontrollers are based on the ARM©Cortex©-M processor. It offers a 32-bit product range that combines very high performance, real-time capabilities, digital signal processing, and low-power, low-voltage operation, while maintaining full integration and ease of development [12]. Such computational characteristics are sufficient to implement the algorithm shown in Fig. 3.

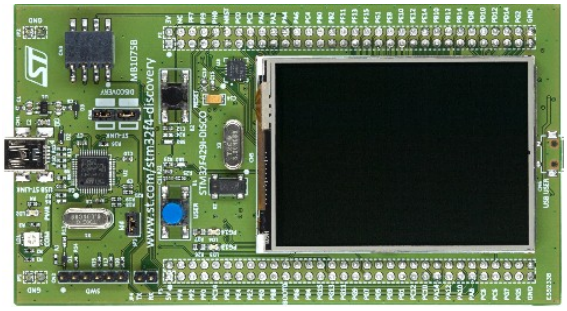


Fig. 4. Embedded system STM32F429IDiscovery where solution was loaded [13]

Fig. 4 shows a target with MCU STM32F429ZIT6 and a LCD to display numeric results. This kind of board is the embedded system used to implement the solution.

2 Results

The development of the source code has four stages:

– First, the input of the number of elements, the number of nodes, the number of fixed nodes, the prescribed values of the potentials at the free nodes, the

x and y coordinates, element-node relationship and prescribed potentials at fixed nodes.
- Second, to find the element coefficient matrix $[C^{(e)}]$ for each element and the global coefficient matrix $[C]$.
- Third, to solve the inverse of the gobal matrix by Gauss-Jordan elimination technique.
- Finally, outputting the result of the computation on a screen or display.

Numerical results are shown by an electrostatic application example.

2.1 Validation with an electrostatic example

The algorithm that solve Laplace's equation was compiled in a free version of Keil uVision 5 and loaded into a STM32F429IDiscovery board. For test the embedded system, it was used a triangle region example from [10] that consists in a two-dimensional problem as shown in Fig. 5. The triangular region is divided into 25 triangular elements with a total of 21 nodes and some Dirichlet boundary conditions are taken as: $100V$ in nodes 11, 15, 18 and 20; $50V$ in nodes 6 and 21; $0V$ in nodes 1, 2, 3, 4, 5, 6 and same $0V$ in nodes 7, 12, 16, 19.

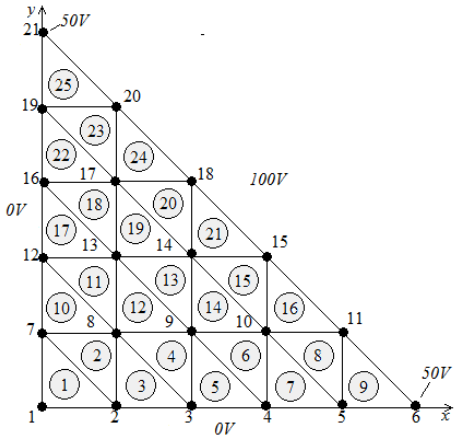


Fig. 5. Solution region divided into 25 elements

The numerical solution is computed in each node and displayed into STM32F Discovery board screen as is shown in Fig. 6. It is possible to observe a numerical comparison with results of the same example from [10] in Table 1.

An important observation is that Discovery board can process operations with matrices of size up 5 x 5 without problems; with operations between matrix

Fig. 6. STM32F running an example

Node No.	Reference Potential	Computed Potential	Error
1	0.000	0.0000	0.0000
2	0.000	0.0000	0.0000
3	0.000	0.0000	0.0000
4	0.000	0.0000	0.0000
5	0.000	0.0000	0.0000
6	50.000	50.0000	0.0000
7	0.000	0.0000	0.0000
8	18.182	18.1818	0.0002
9	36.364	36.3636	0.0004
10	59.091	59.0909	0.0001
11	100.000	100.0000	0.0000
12	0.000	0.0000	0.0000
13	36.364	36.3636	0.0004
14	68.182	68.1818	0.0002
15	100.000	100.0000	0.0000
16	0.000	0.0000	0.0000
17	59.091	59.0909	0.0001
18	100.000	100.0000	0.0000
19	0.000	0.0000	0.0000
20	100.000	100.0000	0.0000
21	50.000	50.0000	0.0000

Table 1. Comparison between reported potential and computed potential.

sizes of 90 x 90 can be observed that there are some processing problems, and for sizes for more than 100 x 100 it is needed to add auxiliary variables when we compile in Keil free version compiler. The auxiliary variable can be a simple double type or a double vector type. The syntax is:

```
for(...){
...
auxiliary[i]=C[i]
C[i] = auxiliary[i] + value[i]*matrix[i][j];
...
}
```

instead of

```
for(...){
...
C[i] = C[i] + value[i]*matrix[i][j];
...
}
```

3 Conclusions

It is possible to solve Laplace's equation using a microcontroller of the family STM32 and avoiding the typical use of a PC. Both the global matrix compute and the Gauss-Jordan elimination are supported and solved very fast. To solve large vector and matrix operations such as addition in the Keil compiler, it is needed an auxiliary variable to compute correctly operations of recursive type $variable[i] = variable[i] + function(i)$; using this form: $auxiliary[i] = variable[i]$; and $variable[i] = auxiliary[i] + function(i)$;.

Acknowledgments The authors would like to express their gratitude to Instituto Politécnico Nacional for the support given in SIP 20144683 and SIP 20151350 projects. Thanks to CONACYT for scholarship (CVU/Scholarship holder) 350051/237150.

References

1. Rokhlin, V.: Rapid solution of integral equations of classical potential theory. Journal of Computational Physics. 60, 187–207 (1985).
2. Wendland, W.L., Stephan, E., Hsiao, G.C., Meister, E.: Mathematical Methods in the Applied Sciences. 1, 265–321 (1979).
3. Komla, D., Pascal, O.: A finite volume method for the Laplace equation on almost arbitrary two-dimensional grids. ESAIM: Mathematical Modelling and Numerical Analysis. 39, 1203-1249 (2005).

4. Liu, C.S.: A Meshless Regularized Integral Equation Method for Laplace Equation in Arbitrary Interior or Exterior Plane Domains. CMES. 19, 99–109 (2007).
5. Johansen, H. , Colella, P.: A Cartesian Grid Embedded Boundary Method for Poissons Equation on Irregular Domains. Journal of Computational Physics. 147, 60–85 (1998).
6. Jomaa, Z., Macaskill, C.: The embedded finite difference method for the Poisson equation in a domain with an irregular boundary and Dirichlet boundary conditions. Journal of Computational Physics. 202, 488–506 (2005).
7. Burden, R.L., Faires, J.D.: Numerical Analysis. Cengage Learning, (2011).
8. Boyce, W.E., DiPrima, R.C.: Elementary Differential Equations and Boundary Value Problems. Wiley, (2013).
9. Griffiths, D.J.: Introduction to electrodynamics. Prentice Hall, (1999).
10. Sadiku, M.O.: Elements of Electromagnetics. Oxford, (2004).
11. Heath, S.: Embedded Systems Design. Newnes, (2003).
12. STMicroelectronics, `http://www.st.com/web/en/catalog/mmc/FM141/SC1169`, 05-05-2015.
13. STMicroelectronics, `http://www.st.com/web/catalog/tools/FM116/SC959/SS1532/LN1848/PF259090?s_searchtype=partnumber`, 01-08-2015

Reviewing Committee

Magdalena Marciano Melchor
Hind Taud
Itzamá López Yáñez
Miguel G. Villarreal Cervantes
Oscar Camacho Nieto
Edgar Roberto Ramos Silvestre
Hildeberto Jardón Kojakhmetov
William De La Cruz De Los Santos
Jamal Toutouh El Alamin
John Cortes Romero
Alexander Gelbukh
Mateo Valero Cortés
Mónica Isela Acuautla Meneses
Najib Tounsi
Roberto Sepúlveda Lima
Víctor Manuel Ramírez Rivera
Martha Dunia Delgado Dapena
Cornelio Yáñez Márquez
Cuauhtémoc López Martín
Yenny Villuendas Rey
Alfredo Víctor Mantilla Caeiros
Edgar Omar López Caudana
Gerardo Abel Laguna Sánchez
Hiram Calvo Castro
José Fermi Guerrero Castellanos
Julián Patiño Ortiz
Pedro Sánchez Santana
Ollin Peñaloza Mejía
Stephane Couturier
Néstor Velasco Bermeo
Mario Aldape Pérez
Luis Octavio López Leyva
Alejandro Rosete Suárez
Giovanni Guzmán Lugo
Roberto Zagal Flores
David Ortega Pacheco
Gerardo Abel Laguna Sánchez
Elena Acevedo Mosqueda
Rolando Flores Carapia
Miguel Patiño Ortiz
Antonio Hernández Zavala
Benjamín Luna Benoso

Impreso en los Talleres Gráficos
de la Dirección de Publicaciones
del Instituto Politécnico Nacional
Tresguerras 27, Centro Histórico, México, D.F.
octubre de 2015
Printing 500 / Edición 500 ejemplares

ISSN: 1870-4069
http://rcs.cic.ipn.mx

RCS
Research in Computing Science