EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

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

Research in Computing Science

ISSN: 1870-4069

Vol. 149 No. 6
June 2020

Research in Computing Science

Series Editorial Board

Editors-in-Chief:

Grigori Sidorov, CIC-IPN, Mexico
Gerhard X. Ritter, University of Florida, USA
Jean Serra, Ecole des Mines de Paris, France
Ulises Cortés, UPC, Barcelona, Spain

Associate Editors:

Jesús Angulo, Ecole des Mines de Paris, France
Jihad El-Sana, Ben-Gurion Univ. of the Negev, Israel
Alexander Gelbukh, CIC-IPN, Mexico
Ioannis Kakadiaris, University of Houston, USA
Petros Maragos, Nat. Tech. Univ. of Athens, Greece
Julian Padget, University of Bath, UK
Mateo Valero, UPC, Barcelona, Spain
Olga Kolesnikova, ESCOM-IPN, Mexico
Rafael Guzmán, Univ. of Guanajuato, Mexico
Juan Manuel Torres Moreno, U. of Avignon, France

Editorial Coordination:
Alejandra Ramos Porras

Research in Computing Science, Año 19, Volumen 149, No. 6, junio de 2020, es una publicación mensual, editada por el Instituto Politécnico Nacional, a través del Centro de Investigación en Computación. Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othon de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738, Ciudad de México, Tel. 57 29 60 00, ext. 56571. https://www.rcs.cic.ipn.mx. Editor responsable: Dr. Grigori Sidorov. Reserva de Derechos al Uso Exclusivo del Título No. 04-2019-082310242100-203. ISSN: en trámite, ambos otorgados por el Instituto Politécnico Nacional de Derecho de Autor. Responsable de la última actualización de este número: el Centro de Investigación en Computación, Dr. Grigori Sidorov, Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othon de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738. Fecha de última modificación 01 de junio de 2020.

Las opiniones expresadas por los autores no necesariamente reflejan la postura del editor de la publicación.

Queda estrictamente prohibida la reproducción total o parcial de los contenidos e imágenes de la publicación sin previa autorización del Instituto Politécnico Nacional.

Research in Computing Science, year 19, Volume 149, No. 6, June 2020, is published monthly by the Center for Computing Research of IPN.

The opinions expressed by the authors does not necessarily reflect the editor's posture.

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 Centre for Computing Research of the IPN.

Hybrid Intelligent Systems

Edgar Cossio
Carlos Alberto Ochoa Ortiz Zezzatti
José Alberto Hernández Aguilar
Julio César Ponce Gallegos (eds.)

SEP
SECRETARÍA DE
EDUCACIÓN PÚBLICA

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

Instituto Politécnico Nacional, Centro de Investigación en Computación
México 2020

ISSN: in process

Copyright © Instituto Politécnico Nacional 2020
Formerly ISSNs: 1870-4069, 1665-9899.

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

The editors and the publisher of this journal have made their best effort in preparing this special issue, but make no warranty of any kind, expressed or implied, with regard to the information contained in this volume.

All rights reserved. No part of this publication may be reproduced, stored on a retrieval system or transmitted, in any form or by any means, including electronic, mechanical, photocopying, recording, or otherwise, without prior permission of the Instituto Politécnico Nacional, except for personal or classroom use provided that copies bear the full citation notice provided on the first page of each paper.

Indexed in LATINDEX, DBLP and Periodica

Electronic edition

ISSN 1870-4069

Table of Contents

Page

ISSN 1870-4069

A Multi-Criteria Decision Making for Sustainable Location of Urban Parks

Aida Yarira Reyes Escalante[1], Diego Adiel Sandoval[2], Edith Vera Bustillos[1], Carlos Alberto Ochoa Ortiz Zezzati[1]

[1] Universidad Autónoma de Ciudad Juárez,
Mexico

[2] Instituto Tecnológico de Ciudad Juárez,
Mexico

aida.reyes@uacj.mx, diegoadiel@yahoo.com

Abstract. The purpose and design of urban parks plays a crucial role for sustainable development of cities. Projects for public spaces and the availability of urban parks allow historic preservation, recreation, and a great variety of social, environmental, and economic benefits. Therefore, is important to comply with diverse sustainability indicators toward parks policy based on contemporary needs. Hence, park planners must consider the analysis of accessibility, transportation, contiguity, proximity and connectivity, natural areas, good land, size, influence, and restrictions, among other elements. In this study, we aim to determine a core alternative for planning a suitable location of an urban park by utilizing a multi-criteria decision making methods, MCDM, and through the applications of the Analytic Hierarchy Process, AHP, and Technical for Order of Preference by Similarity to Ideal Solution, TOPSIS. We linked the multi-criterion to the identification of latent locations of green spaces in Juarez Region, Chihuahua, Mexico, by spatial distribution in five sub-regions -northwest, northeast, central, southwest, and southeast (A thru E). The results of this study allowed determining that only one option (D) obtained the top assessment as the optimum alternative to be transformed as urban park, which is located in the northeast area of the landscape assessed. The approach in this study has provided practical ways of managing not only the spatial distribution of urban parks but understanding some holistic criteria for sustainability.

Keywords: Urban parks, AHP methodology, TOPSIS methodology, multi-criteria decision-making, MCDM, sustainable assessment for locations of urban parks.

1 Introduction

The need for sustainable social development options (e.g. communication, green and recreations spaces areas, living areas and environmental protection areas) calls for improved urban development plans. These plans must consider factors like infrastructure, health, protection and security, among others.

pp. 5–20; rec. 2019-08-07; acc. 2019-09-14

Table 1. Original unit of analysis and the applicability to world cities. Source: Morin & Christodoulou [2].

		Applicability to comparison of sustainability among world cities	
		Yes but conditional	No
Original unit of analysis	Global Country		LPI
		CDI	ESI
		DS	HDI
		WF	EVI
		EF	EPI
		GPI	Satellite-based
		ISEW	sustainability
		GS	WI
		EDP(Green GDP)	
	Region or Local City Other		Applications of composite indices to local contexts
		CDI	
		WF	LPI
		Emergy/exergy	

Sustainable cities are considered those where indicators of economic, social and environmental development are constantly monitored and improved. The aim is to maintain balances in all areas, to make urban spaces more environmentally friendly while being optimal for social development and having activities that drive economic development. Cities are of utmost importance for measuring sustainability. The United Nations Human Settlements Program (UN-Habitat) [1] defines an urban agglomeration as the built-up or densely populated area including the suburbs and continuously settled commuter areas, which may be smaller or larger than a metropolitan area.

The conformation of the city is defined both by the inner space that has been carried by the oldest areas as well as by the areas that have been developed around it, areas that have been added not necessarily in planned form, many of them arose from social needs or demographic growth. The components that make up the cities are varied, such as: communication routes, green spaces, equipment, infrastructure, housing areas, shopping areas, surrounded by the most important, all the components are related to each other, either directly or indirectly.

Mori & Christodoulou [2] defines the indexes and indicators to assess the sustainability in the cities. The results shed light about nine indexes, as follow: City Development Index (CDI), Ecological Footprint (EF), Dashboard of Sustainability (DS), Welfare Index (WF), Genuine Progress Indicator (GPI), Index of Sustainable

Economic Welfare (ISEW), Energy/exergy, Environmentally adjusted Domestic Product (EDP), Genuine Saving (GS), see in Table 1.

Based in the results by Morin, et al [2] is remarkably considered the conclusion related to the need of building a unique indicator to measure the sustainability of the city (CSI) where the environmental aspects take up an important role such as biophysical or ecological thresholds, green space, parks, protected areas, and others.

2 Cities and Green Space

Urban planning is perhaps the main ingredient to achieve sustainability in cities. It provides for social, environmental and economic balances. Since the Nineteenth century, there were already concerns about the issues of an adequate urban planning. Oftentimes, cities fail to provide happiness and recreation, as well as other forms of public space use, such as games, opportunities for playing music, entertainment and education, and others, Howe [3].

Additionally, demographic growth has brought increases of air pollution, vehicular traffic, noise, heat, insolation, loss of vegetation, insecurity, and other city life –related problems. The loss of greenspace follows the growth of the population, which, at its time, drives the city into a series of unwanted conditions, some of them already mentioned.

As we have seen, greenspace offers a variety of services in the urban environment. When adequately used and maintained, forests and urban greenspace are laboratories and classrooms of environmental education. They also offer the opportunity of meeting diverse material needs. On the other hand, these spaces give the possibility of satisfying other non-material needs, such as recreation and collective encountering and interaction with other people [4].

Nielsen and Hansen [5] affirm that the style of urban life, with shortage or lacking of greenspace, relates with many mental diseases (e.g. stress, depression and anxiety). Niemelä et al., [6] indicate that greenspace might mitigate the negative consequences of accelerating urbanization. Urban greenspace provides a diverse set of ecosystem services ranging from those derived physical well-being to those related to psychological comfort. Many countries have actively developed a variety of programmers for sustainable urban planning with respect to greenspace.

Urban parks are a special kind of public space. They constitute fundamental elements for managing and defining the equilibrium between built and unbuilt areas in a city. Many spaces, whether structured or unstructured, expansion zones, empty lots, environmental control stripes, may reconvert into an urban park.

Literature regarding evaluation to determine the location of urban parks shows that: in 2018, there were three studies; six in 2017; one during 2016, 2015, 2014 and 2011 respectively. Although we found many of papers addressing the problem of locating an urban park using a variety of techniques: modeling [7, 8, 9, 10, 11]; urban distributions [12, 13]; health [14]; public policies [15, 16]; hot spots [17]; distribution areas [18, 19]; among others. We found that none of the research approached the multi-criteria analysis

Table 2. Elements of a Sustainable Park. Source: Cranz and Boland [21].

Element	**Variables**
Social Goal	Human health; ecological health
Activities	Strolling, hiking, biking, passive & active recreation, bird watching, education, stewardship
Size	Varied, emphasis on corridors
Relation to City	Art-nature continuum; part of larger urban system; model for other
Order	Evolutionary aesthetic
Elements	Native plants, permeable surfaces, ecological restoration green infrastructure, resource self-sufficiency
Promoters	Environmentalists, local communities, volunteers groups, landscape architects
Beneficiaries	Residents, wildlife, cities, planet

methods for the decision making of the location of green areas and in specific public parks.

3 Location Urban Park Methods

According to Cranz [20], the urban park concept has evolved throughout four steps: (a) Pleasure Ground (1850–1900), (b) the Reform Park (1900–1930), (c) the Recreation Facility (1930–1965), and (d) the Open Space System (1965–?).

However, Cranz and Boland [21] suggest a fifth category: Sustainable Park (1990-present). The typology includes both the shifting social purposes that parks served and the corresponding variations in designed form. Each park type evolved to address what were considered to be pressing urban social problems at that time, (see table 2).

The location of the park is the most important decision issue. Jim [22] states that spatial permeation and connectivity of greenspace is desired along new roads (amenity strips on roadsides and medians), amenity parcels in roundabouts, and incidental plots. Also, indicate that within lots, greenspace should be located in the grounds of residential, office, government, institutional and community land uses.

The landscape ecology concepts related to the size, shape and connectivity could be applied with imagination to greenspace planning [23, 24]. Davey, [25] recommended a set of cardinal principles to locate an urban park. He used the nature-reserve design, based on island biogeography theory, namely large size, contiguity, proximity and connectivity, can enhance the quality of green sites.

4 Multi-Criteria Analysis to Define Urban Parks Location

Multiple-criteria decision-making (MCDM) also referent as multiple-criteria decision analysis (MCDA) is a sub-discipline of operations research that explicitly evaluates multiple conflicting criteria in decision making. The MCDA is the process of ranking discrete candidate alternatives and finding the best compromise solution based on the decision maker’s subjective assessments of multiple evaluative criteria [26].

Ting-Yu Chen [27] indicated the MCDA problems becoming increasingly complicated, exact assessments of the choices based on evaluative criteria may be difficult to measure or quantify along the MCDA cycle. Conflicting criteria are typical in evaluating options: cost, customers, tools, equipment, personals, spaces, in all decisions problems criterial can be use the MCDA.

There are different classifications of MCDA problems and methods. A major distinction between MCDA problems is based on whether the solutions are explicitly or implicitly defined. Two options need to follows: Multiple-criteria evaluation problems or Multiple-criteria design problems (multiple objective mathematical programming problems).

Malakooti [28] indicated the MCDM approaches should have the following nine characteristics: Principle-oriented (axiom-based); Convincing; Coherent; Defendable (justifiable); Enlightening (illuminating; informative; supportive); Versatile (allows for the use of different preferential behaviors); Transparent; Systematic; Verifiable (testable and repeatable).

Different approaches to select MCDA methods to solve specific problems have been used to look at the outcomes [29, 30, 31], see Table 3.

5 Problem Setting and Research Objective

A brief review of the land use patterns of Juarez revealed there are many zones available to potential used to urban park [32].

The rapid and unstructured growth of Ciudad Juarez city has prevented an adequate urban planning. According to official data, there are around 4000 parks in the city.

However, currently only two public parks pertaining to the category of urban (that is, a major large-scale urban park) are available: Chamizal Park and Central Park Hermanos Escobar. The first locates in the northern part of the city and the second in the geographical center of the city.

Studies related to the subject matter are limited; we were no able to find evidence of the use of methodologies for the location of an urban park in Ciudad Juárez.

We can state that a methodological system for planning the location of an urban park in of Ciudad Juarez is not available. This, based on the above, the main objective is to determine the location of an urban park in Ciudad Juarez through multi-criteria analysis.

Table 3. MCDA method.

Inputs	**Effort input**	**MCDA method**	**Output**
Utility function	Very HIGH	MAUT	Complete ranking with score
Pairwise comparisons on a ratios scale and interdependencies	I I I	ANP	Complete ranking with score
Pairwise comparisons on an interval scale	I I I	MACBETH	Complete ranking with score
Pairwise comparisons on a ratio scale indifference,	I I	AHP	Complete ranking with score
preference and veto thresholds	I I I	ELECTRE	Partial and complete ranking (pairwise outranking degrees)
	I I I	PROMETHEE	Partial and complete ranking (pairwise preference degree and score
Ideal option and constraints	I I	Goal programming	Feasible solution with deviation score
Ideal and anti-ideal option	I I	TOPSIS	Complete ranking with closeness score
No subjective inputs required	I Very LOW	DEA	Partial ranking with effectiveness score

6 Method

The study relies on an exploratory analysis carried out in 2019 in Juarez, Mexico. We considered five different sectors: North, East, West, South and Central. In addition, this study involves a quantitative design, adopting a multi-criteria decision methodology to determine a basis for planning the location of an urban park.

Through Analytic Hierarchy Process, AHP, and Technical for Order of Preference by Similarity to Ideal Solution, TOPSIS, we evaluated the location's selection.

We utilized the following criteria evaluation: Nature-reserve design: the space is declaring of protection area; Contiguity: the ground is continuous without breaking; Proximity and connectivity: close to the urban area, connections area; Good land: land in good condition to plant vegetation; Parcels: land used to sow; Size: Great or big space; Influence area; Restrictions section: Airplane, military area, unsafe zone, criminal and insecurity, restricted area; Accesses: streets, avenues, high way; Transport accesses: cars, bicycles, motorcycles, public bus, etc.

Furthermore, we included Beta-values, or compliance (or suitability or fitness) value judgments, with the following arbitrary Lickert scale: (1) Worst (no compliance); (2) Very (low compliance; (3) Undesirable compliance; (4) Slightly undesirable compliance; (5) Neutral compliance, and (6) Slightly desirable compliance. We describe the methodology utilized in the multi-criteria analysis: AHP and TOPSIS.

6.1 Analytic Hierarchy Process AHP

The analytic hierarchy process was development by Thomas L. Saaty in the 1977. It is a mathematical structured method [33] and subdivides a complex decision-making problem or planning issue into its components or levels, and arranges these levels into an ascending hierarchic order [34]. In addition, it indicates that the AHP can provide a framework and methodology for the determination of a number of key decisions. The AHP allows its users flexibility in constructing a hierarchy to fit their needs. Also the AHP provides an effective structure for group decision making by imposing a discipline on the group's thought processes [35].

The process AHP is established in different stages, the formulation of the decision problem in a hierarchical structure is the first and main stage. In this stage, the decision maker involved must break down the problem into its relevant components [35]. To make a decision in an organized way to generate priorities we need to decompose the decision into the following steps:

a. Define the problem and determine the kind of knowledge sought.

b. Structure the decision hierarchy from the top with the goal of the decision, then the objectives from a broad perspective, through the intermediate levels (criteria on which subsequent elements depend) to the lowest level (which usually is a set of the alternatives).

c. Construct a set of pairwise comparison matrices. Each element in an upper level is used to compare the elements in the level immediately below with respect to it. The value a_{12} is an approximation of the relative importance of A1 with respect to A2, i.e., $a12 \approx (w1 \,/\, w2)$. This can be generalized and the following:

$$aij \approx \left(wi - wj\right) i, j \ = \ 1, 2, \dots, n \,,$$

$$aii \ = \ 1, i \ = \ 1{,}2, \dots, n,$$

$$Si\ aij = \alpha, \alpha \neq 0, entonces\ aji \ = \frac{1}{\alpha}, i \ = \ 1{,}2, \dots, n \,.$$

If Ai is more important than Aj, then:

$$aij \cong \left(wi - wj\right) > 1. \tag{1}$$

The matrix A must be positive and reciprocal with ones in the main diagonal, and therefore the decision maker only needs to provide the values of the judgments in the upper triangular of the matrix. To fill these values, we use already established scales 1 to 9. The judgments of the criteria are perfectly consistent as long as it is fulfilled that:$aijajk = aik, i, j, k = 1, 2, \ldots, n, lo\ que\ es\ equivalente\ a$: $(wi\ /wj)\ (wj\ /wk) = (wi\ /wk)$.

The eigenvector method produces a natural measure of consistency. Saaty defines the consistency index (CI) as a distance between the λmax and the value of that λmax when the judgments were perfect, ie λmax = n. The CI is defined as follows:

$$CI = \frac{\lambda \max - \mathrm{n}}{(\mathrm{n}-1)}. \tag{2}$$

d. Use the priorities obtained from the comparisons to weigh the priorities in the level immediately below. Do this for every element. Then for each element in the level below add its weighed values and obtain its overall or global priority.

6.2 TOPSIS Method

The Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) is a MCDA tool. It was primarily established by Hwang and Yoon in 1981 for ranking based on resemblance to perfect solution, with advancements done by Yoon in 1987, and Hwang, Lai and Liu in 1993. TOPSIS is a prevalent method suitable for taking a multiple criteria decision for rank ordering by comparison. It is a technique for rank ordering based on closeness to perfect outcomes. The ultimate option is the one that is nearest to the perfect positive outcome and extreme from the negative perfect outcome [36].

This study uses the TOPSIS method. A positive ideal solution maximizes the benefit criteria or attributes and minimizes the cost criteria or attributes, whereas a negative ideal solution maximizes the cost criteria or attributes and minimizes the benefit criteria or attributes. The TOPSIS method is expressed in a succession of six steps as follows:

Step 1: Calculate the normalized decision matrix. The normalized value r $_{ij}$ is calculated as follows:

$$r_{ij} = x_{ij} \sqrt{\sum_{i-1}^{m} x_{ij}^2} \qquad i = 1, 2, \ldots, m\ and\ j = 1, 2, \ldots, n. \tag{3}$$

Step 2: Calculate the weighted normalized decision matrix. The weighted normalized value vij is calculated as follows:

$$Vij = rij\ X\ Wj \qquad i = 1, 2, \ldots, m\ and\ j = 1, 2, \ldots, n. \tag{4}$$

where w_j is the weight of the j^{th} criterion or attribute and $\sum_{j=1}^{n} w_j = 1$.

Step 3: Determine the ideal (A^*) and negative ideal (A^-) solutions:

$$A^* = \{(\max_i v_{ij} \mid j \in C_b), (\min_i v_{ij} \mid j \in C_c)\} = \{v_j^* \mid j = 1, 2, \ldots, m\}, \tag{5}$$

$$A^{-}=\{(\min_{i} v_{ij} \mid j \in C_{b}),(\max_{i} v_{ij} \mid j \in C_{c})\}=\{v_{j}^{-} \mid j=1,2,\ldots,m\}. \tag{6}$$

Step 4: Calculate the separation measures using the m-dimensional Euclidean distance. The separation measures of each alternative from the positive ideal solution and the negative ideal solution, respectively, are as follows:

$$S_{i}^{*}=\sqrt{\sum_{j=1}^{m}(v_{ij}-v_{j}^{*})^{2}, j=1,2,\ldots,m}, \tag{7}$$

$$S_{i}^{-}=\sqrt{\sum_{j=1}^{m}(v_{ij}-v_{j}^{-})^{2}, j=1,2,\ldots,m}. \tag{8}$$

Step 5: Calculate the relative closeness to the ideal solution.

The relative closeness of the alternative A_i with respect to A^* is defined as follows:

$$RC_{i}^{*}=\frac{S_{i}^{-}}{S_{i}^{*}+S_{i}^{-}}, i=1,2,\ldots,m. \tag{9}$$

Step 6: Rank the preference order.

The studies carried out using the two methods of MCDA using AHP and TOPSIS in the analysis of green spaces is increase. The review of the literature allows to visualize how the methods of MCDA are used comparatively to determine the best decisions through AHP and TOPSIS, finding that there is no research where they are used for the decision of Green spaces and urban parks location [37-43].

6.3 Criterial Evaluation Areas

The criterial to perform the evaluation consist in 10 criteria decision:

a. Nature-reserve design: the space is declaring of protection area.
b. Contiguity: the ground is continuous without breaking.
c. Proximity and connectivity: close to the urban area, connections area.
d. Good land: land in good condition to plant vegetation.
e. Parcels: land used to sow.
f. Size: Great or big space.
g. Influence area: 400 mtrs, 1500 mtrs.
h. Restrictions section: Airplane, military area, unsafe zone, criminal and insecurity, restricted area.
i. accesses: streets, avenues, high way.
j. Transport accesses: cars, bicycles, motorcycles, public bus, etc.

The Beta-values, or compliance (or suitability or fitness) value judgments, are made on the following arbitrary licker scale:

1) Worst (no compliance).
2) Very (low compliance).

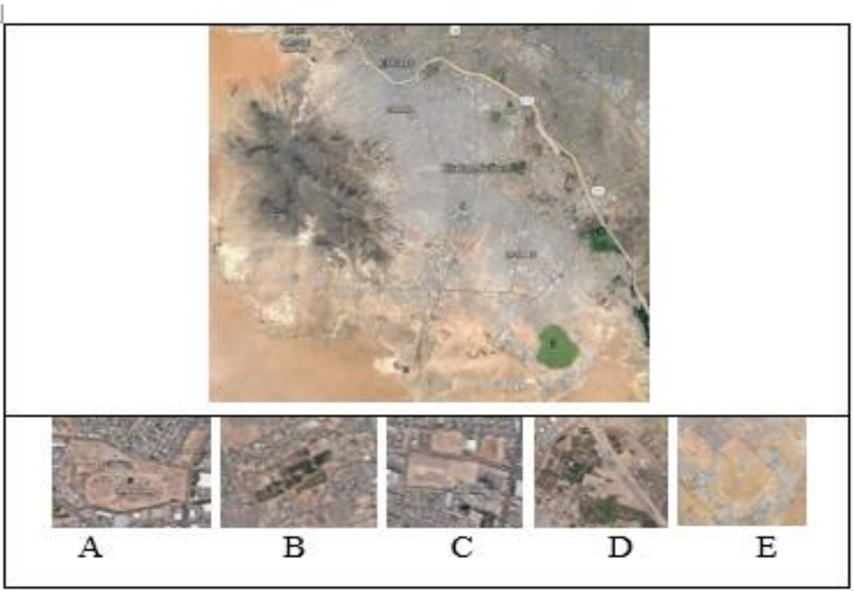


Fig. 1. Decision study sites.

3) Undesirable compliance.

4) Slightly undesirable compliance.

5) Neutral compliance.

6) Slightly desirable compliance.

7) Desirable compliance.

8) Very desirable compliance.

9) Best possible compliance.

7 Results

The results are presented in three moments: (a) map analysis, (b) AHP analysis, and (c) TOPSIS analysis.

(a) The analyses maps determine the site for the location of the park, this analysis was carried out taking as criteria: the location of the space, the free area and access to the site, which is fields uninhabited areas, planting areas, assigned natural areas, airport proximity, restricted or military areas were avoided. We found 5 viable sites for the development of an urban park, see Figure 1. The first evaluation is generated using the beta-values to establish the first relationship, see table 5.

Table 5. Evaluation site using Beta-values.

	A	**B**	**C**	**D**	**E**
Nature-reserve design	1	5	1	9	1
Contiguity	9	9	9	9	7
Proximity and connectivity	9	9	9	7	9
Good land to sow	1	9	1	9	1
Parcels	1	9	1	9	1
Size	5	5	1	9	9
Influence area	9	9	9	5	7
Restrictions section	5	1	1	1	9
Access	9	9	9	9	9
Transport	9	9	9	9	9

Table 6. AHP matrix criteria's example.

Nature-reserve design											
	A	**B**	**C**	**D**	**E**	**Matrix normalized**					**Vector**
A	1	0.2	1	0.11	1	0.05	0.02	0.048	0.05	0.07	0.049
B	5	1	9	0.2	1	0.29	0.1	0.429	0.08	0.076	0.196
C	1	1	1	0.11	1	0.05	0.1	0.048	0.05	0.076	0.065
D	9	7	9	1	9	0.52	0.69	0.429	0.41	0.692	0.549
E	1	1	1	1	1	0.05	0.1	0.048	0.41	0.076	0.138
Total	17	10.2	21	2.42	13						

Table 7. Concentrate Matrix vectors.

	Nature-reserve design	Contiguity	Proximity and connectivity	Good land to sow	Parcels	Size	Influence area	Restrictions	Accesses	Transport
A	0.049	0.20	0.207	0.047	0.047	0.048	0.258	0.287	0.2	0.2
B	0.196	0.20	0.207	0.428	0.428	0.040	0.258	0.059	0.2	0.2
C	0.065	0.20	0.207	0.046	0.047	0.031	0.258	0.059	0.2	0.2
D	0.549	0.20	0.169	0.425	0.428	0.439	0.022	0.059	0.2	0.2
E	0.138	0.16	0.207	0.04	0.047	0.439	0.201	0.534	0.2	0.2

(b) AHP method: According with the stage, the step 1 and 2 was defined in the problem selection. The Step 3. Construct a set of pairwise comparison matrices is presented in the Table 6 the concentrated matrix vector in Table 7. Priority results are presented in Table 8. The best option using AHP is D.

Step 1: Calculate the normalized decision matrix.

	Nature-reserve design	Contiguity	Proximity and connectivity	Good land to sow	Parcels	Size	Influence area	Restrictions section	Accesses	Transport
A	1	9	9	1	1	5	9	5	9	9
B	5	9	9	9	9	5	9	1	9	9
C	1	9	9	1	1	1	9	1	9	9
D	9	9	7	9	9	9	5	1	9	9
E	1	7	9	1	1	9	7	9	9	9

Step 2: Calculate the weighted normalized decision matrix.

	Nature-reserve design	Contiguity	Proximity and	Good land to sow	Parcels	Size	Influence area	Restrictions section	Accesses	Transport
A	0.009174	0.0241	0.024	0.006061	0.00606	0.02	0.028	0.05	0.0222	0.022
B	0.045872	0.0241	0.024	0.054545	0.05455	0.02	0.028	0.01	0.0222	0.022
C	0.009174	0.0241	0.024	0.006061	0.00606	0	0.028	0.01	0.0222	0.022
D	0.082569	0.0241	0.019	0.054545	0.05455	0.04	0.016	0.01	0.0222	0.022
E	0.009174	0.0188	0.024	0.006061	0.00606	0.04	0.022	0.08	0.0222	0.022

Step 3: Determine the ideal and negative ideal solutions.

V+	0.012385	0.0024	0.004	0.008182	0.00273	0	0.003	0.01	0.0011	0.001
V-	0.001376	0.0019	0.003	0.000909	0.0003	0	0.002	0	0.0011	0.001

Step 4: Calculate the separation measures using the m-dimensional Euclidean distance.

Step 5: Calculate the relative closeness to the ideal solution. The relative closeness of the alternative.

Step 6: Rank the preference order.

	Si+	**Si-**	**Pi**	**Rank**
A	0.01403442	0.0044182	0.239433	4
B	0.009364541	0.0080513	0.462296	2
C	0.015746285	0.0015895	0.09169	5
D	0.007490435	0.0085536	0.533132	1
E	0.01344087	0.0083078	0.38199	3

c. TOPSIS method: Using the information in Table 6, the beta-values is using the data to start the evaluation using the TOPSIS method. The best option using TOPSIS is D.

Table 8. Final results.

	Nature-reserve design	Contiguity	Proximity and connectivity	Good land to sow	Parcels	Size	Influence area	Restrictions section	Accesses	Transport	TOTAL
A	0.04	0.208	0.207	0.05	0.047	0.05	0.25	0.28	0.2	0.2	0.157
B	0.19	0.208	0.207	0.43	0.428	0.04	0.25	0.05	0.2	0.2	0.215
C	0.06	0.208	0.207	0.05	0.047	0.03	0.25	0.05	0.2	0.2	0.095
D	0.54	0.208	0.169	0.43	0.428	0.43	0.02	0.05	0.2	0.2	0.275
E	0.13	0.165	0.207	0.05	0.047	0.43	0.20	0.53	0.2	0.2	0.25
Weighted	0.144	0.117	0.097	0.13	0.139	0.04	0.02	0.28	0.0125	0.012	

8 Conclusions

Urbans parks are key elements to pursue sustainability in urban conglomerates. The adequate location of an urban park is paramount for such an endeavor. Parks are public spaces that provides a variety of eco-services. It is important to incorporate sustainable criteria to evaluate their location. Being a determinant factor for urban planning, the location of parks is, therefore, a priority. To determine the best-featured area to locate a park is a task of gliders and researchers interested in these strategies.

This work offers a quantitative multi-criteria framework to determine the best urban alternatives in terms of size, geographical location, access roads, land extension, areas of influence, land types, fundamental elements for decision-making when locating an urban park.

Out of the five identified areas (A-E), evaluated as a potential zone to locate an urban park in the municipality of Juarez, alternative D was optimal according to AHP y TOPSIS methods.

We can conclude that the use of the AHP and TOPSIS methods allow evaluating the location of an urban park using the identified criteria. We conducted from the perspective of the indicated methods and without taking into account other factors like land property, land use, and the eventual existence of public policies that allow knowing if the optimal location has chances of being an urban park.

Both AHP and TOPSIS found option D to be optimal; this zone is located in the Northern area of the city. This area is undergoing an active urban growth, with new constructions and infrastructure projects now in progress. The area was considered in the past decade’s part of the reserve natural area, but in the actually is the most important construction urban area.

The approach that we consolidated in this study, offers a comprehensive and effective decision support method to address a core alternative for planning a suitable location of an urban park. It is recommendable the adoption of this multi-criteria techniques for further decisions regarding the parks policy based on contemporary needs.

9 Recommendation

For future studies it is recommended that land use factors, land ownership, and other relevant restrictions be considered. In addition, from the methodologies, it is relevant to consider alternative scenarios, such as: sensitivity analysis, to strengthen the evaluation criteria.

References

1. UN-Habitat: United Nations Human Settlements Programme: The state of the world's cities report 2006/2007 (2006)
2. Mori, K., Christodoulou, A.: Review of sustainability indices and indicators: towards a new city sustainability index (CSI). Environmental Impact Assessment Review, 32(1), pp. 94–106 (2012)
3. Howe, F.C.: European cities at work (2012)
4. Rivas-Torres, D.: Planeación, espacios verdes y sustentabilidad en el distrito federal. División de Ciencias y Artes para el Diseño Especialización, Maestría y Doctorado en Diseño, Universidad Autónoma de México (2005)
5. Nielsen, T.S., Hansen, K.B.: Do green areas affect health? Results from a Danishsurvey on the use of green areas and health indicators. Health Place, 13(4), pp. 839–850 (2007)
6. Niemelä, J., Saarela, S.R., Söderman, T., Kopperoinen, L., Yli-Pelkonen, V., Väre, S., Kotze, D.J.: Using the ecosystem services approach for better planning and conservation of urban green space: A Finland case study. Biodivers Conserv., 19(11), pp. 3225–3243 (2010)
7. Muge-Unal, C.U.: Evaluating and optimizing urban green spaces for com-pact urban areas: Cukurova district in Adana, Turkey. In: ISPRS International Journal Geo-Information, 7(2), pp. 70 (2018)
8. Conghong Huang, Jun Yang, Peng Jiang: Assessing impacts of urban form on landscape struc-ture of urban green spaces in china using Landsat images based on google earth engine. Remote Sensing, 10(10), pp. 1569 (2018)
9. Niedermann, B., Oehrlein, J., Lautenbach, S., Haunert, J.H.: A network flow model for the analysis of green spaces in urban areas. GIScience, 13, pp. 1–13 (2018)
10. Vallejo, M., Corne, D.W., Vargas, P.A.: Online/offline evolutionary algorithms for dynamic urban green space allocation problems. Journal of Experimental and Theoretical Artificial Intelligence, 29(4), pp. 843–867 (2017)
11. Wicaksono, A., Sarapirome, S.: Urban park area feasibility analysis using fuzzy aggregation of multi-spatial criteria and multi-expert weights. In: International Conference on Digital Information Management (ICDIM´17), pp. 246–251 (2011)
12. Van de Voorde, T.: Spatially explicit urban green indicators for characterizing vegetation cover and public green space proximity: A case study on Brussels. Belgium. In: International Journal of Digital Earth, 10(8), pp. 798–813 (2017)

13. Vallejo, M., Rieser, V., Corne, D.W.: Agent-based modelling for green space allocation in urban areas-factors influencing agent behaviour. In: International Conference on Agents and Artificial Intelligence (ICAART), 1, pp. 257–262 (2015)
14. Conghong Huang, Jun Yang, Hui Lu, Huabing Huang, Le Yu: Green spaces as an indicator of urban health: Evaluating its changes in 28 mega-cities. Remote Sensing, 9(12), pp. 1266 (2017)
15. Zhoulu Yu, Yaohui Wang, Jinsong Deng, Zhangquan Shen, Ke Wang, Jinxia Zhu, Muye Gan: Dynamics of hierarchical urban green space patches and implications for management policy. Sensors, 17(6), pp. 1304 (2017)
16. Banzhaf, E., De la Barrera, F.: Evaluating public green spaces for the quality of life in cities by integrating RS mapping tools and social science techniques. In: Joint Urban Remote Sensing Event (JURSE), pp. 1–4 (2017)
17. Kothencz, G., Kiss, M., Petutschnig, A.: Hot spots for improvements: Where to implement new green spaces?. Smart-World, pp. 1–4 (2017)
18. Tongliga Bao, Xueming Li, Jing Zhang, Yingjia Zhang, Shenzhen Tian: Assessing the distribution of urban green spaces and its anisotropic cooling distance on urban heat island pat-tern in Baotou. In: ISPRS International Journal of Geo-Information, 5(2), pp. 12 (2016)
19. Neuenschwander, N., Hayek, U.W., Grêt-Regamey, A.: Integrating an urban green space typology into procedural 3D visualization for collaborative planning. Computers, Environment and Urban Systems, 48, pp. 99–110 (2014)
20. Galen, C.: The politics of park design: A history of urban parks in America. MIT Press (1982)
21. Galen, C., Boland, M.: Defining the sustainable park: A fifth model for urban parks. Landscape Journal, 23, pp. 102–120 (2004)
22. Jim, C.Y.: Green-space preservation and allocation for sustainable greening of compact cities. Cities, 21, pp. 311–320 (2004)
23. Cook-Edward, A.: Landscape structure indices for assessing urban ecological networks. Landscape and Urban Planning, 58(2), pp. 269–280 (2002)
24. Botequilha-Leita, Ä., Ahern, J.: Applying landscape ecological concepts and metrics in sustainable landscape planning. Landsc Urban Plann, 5, pp. 65–93 (2002)
25. Davey, A.G.: National system planning for protected areas. World Conservation Union, Gland (1998)
26. Wan, S.P., Qin, Y.L., Dong, J.Y.: A hesitant fuzzy mathematical programming method for hybrid multi-criteria group decision making with hesitant fuzzy truth degrees. Knowl-Based Systems, 138, pp. 232–248 (2017)
27. Chen, T.Y.: Remoteness index-based Pythagorean fuzzy VIKOR methods with a generalized distance measure for multiple criteria decision analysis. Information Fusion, 41, pp. 129–150 (2018)
28. Malakooti, B.: Double helix value functions, ordinal/cardinal approach, additive utility functions, multiple criteria, decision paradigm, process, and types (Z Theory I). In: International Journal of Information Technology and Decision Making, 14(6), pp. 1353 (2015)
29. Ishizaka, A., Nemery, P.: Multi-criteria decision analysis: Methods and software. Wiley (2013)
30. Ishizaka, A., Nemery, P.: PROMETHEE. In: Multi-Criteria Decision Analysis: Methods and Software, pp. 135–179 (2013)
31. Guitouni, A., Martel, J.M.M., Bélanger, C.: Managing a decision-making situation in the context of the Canadian airspace protection. Centre de Recherche en Modélisation, Information et Decision (CERMID) (1999)

32. IMIP: Atlas de riesgos naturales y atlas de riesgos antropogénicos. Instituto Municipal de Investigación y Planeación (2016)
33. Saaty, T.L.: The analytic hierarchy process. RWS Publications, 4922 (1990)
34. Saaty, T.L.: Decision making with the analytic hierarchy process. International Journal Services Sciences, 1(1) (2008)
35. Saaty, T.L.: Analytic heirarchy process. Wiley (2014)
36. Xing, W.W.: Distance measure between intuitionistic fuzzy sets. Pattern Recognition Letters, pp. 2063–2069 (2005)
37. Guangdong Tian, Honghao Zhang, MengChu Zhou, Zhiwu Li: AHP, gray correlation, and TOPSIS combined approach to green performance evaluation of de-sign alternatives. IEEE Trans. Systems, Man, and Cybernetics: Systems, 48(7), pp. 1093–1105 (2018)
38. Mohammed, A., Harris, I., Soroka, A., Naim, M.M., Ramjaun, T.: Evaluating green and resilient supplier perfor-mance: AHP-fuzzy topsis decision-making approach. In: Proceedings of the International Conference on Operations Research and Enterprise Systems (ICORES´18), pp. 209–216 (2018)
39. Zyoud, S.H., Fuchs-Hanusch, D.: A bibliometric-based survey on AHP and TOPSIS techniques. Expert Systems with Applications, 78, pp. 158–181 (2017)
40. Campos-Guzman, V., García-Cascales, M.S., Sánchez-Lozano, J.M., Pelta, D.A.: Selection of a fuzzy AHP-TOPSIS electrification system for an isolated rural area in southern México. In: IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–6 (2017)
41. Guoying Yang, Qingling Wang, Jianqing Liu: TOPSIS and AHP model in the application research in the evaluation of coal. In: ICIC, 1, pp. 145–152 (2016)
42. Pazand, K., Hezarkhani, A.: Porphyry Cu potential area selection using the combine AHP - TOPSIS methods: A case study in Siahrud area (NW, Iran). Earth Science Informatics, 8(1), pp. 207–220 (2015)
43. Dammak, F., Baccour, L., Alimi, A.M.: A comparative analysis for multi-attribute decision-making methods: TOPSIS, AHP, VIKOR using intui-tionistic fuzzy sets. . In: IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–5 (2015)

ISSN 1870-4069

Design of a Data Repository for Analysis of Inventory in the Mexican Healthcare System by Means of Data Analytics: Case of Study of a Clinic in the Southeast of Mexico

José Alberto Hernández Aguilar[1], Mariana Ugarte Nava[2],
Julio César Ponce Gallegos[2]

[1] Universidad Autónoma del Estado de Morelos,
Mexico

[2] Universidad Autónoma de Aguascalientes,
Mexico

jose_hernandez@uaem.mx

Abstract. We discussed a data warehouse designed for the health Mexican sector to maintain a proper inventory of the clinics, avoiding as much as possible the misuse of medicines and supplements, for this purpose, we analyzed the characteristics of the main data repositories and their characteristics, later we focus on creating a proposal to support the Mexican health care system oriented on the optimal management of inventories. Our preliminary results indicate that it is possible to maintain adequate inventories of items (medicines and supplements) that are deemed critical from the refill with inventory techniques.

Keywords: Data warehouse, healthcare system, inventory, two-bin system.

1 Introduction

The companies that are suppliers of entities in the area of health care need to maximize the benefits of their delivery operations of inputs. For their part, the health sector entities, both public and private, face several challenges in the management of material resources, as the vast majority has major shortcomings with regard to the visibility and updating of the inventory of inputs, processes of patient care, in addition to that have complex structures of payment to suppliers, among other problems. To cope with these challenges, this obliges them to give special importance to logistic systems, especially where the clinician has to manage inventories. The steps that need to be done manually make the processes more prone to errors, slow, and therefore, expensive.

An approach to solve this situation is supported by Information Technology (IT), where databases are key elements in applications that manage supplies, and there are lot of applications related to the health sector.

Table 1. Features identified for treatment of patients based on [3, 4].

<<Treatment>>
1. Type of case
2. Type of department
3. Patient Type
4. Type of partial payment
5. Number of days of medication
6. Gender of the doctor
7. Cost of the medication
8. Cost of diagnosis
9. Cost of the review
10. Cost of medication administration.

In the real world, you can find very complete database systems, as the American Health Care System, the health system in Singapore and in the case of Latin America excels that of Brazil. According to the literature [1] and the reality observed, data bases of the health sector are usually very large and contain many tables and information, for example, the Brazilian Health Sector system requires 380 tables to operate. In addition to this, the information can be difficult to access, and in several Mexican institutions, information derived from day-to-day operations is not digitized or even stored in electronic media without any normalization process. According to [2] that is why we will see this research from the point of view of data analytics.

2 Related Work

For the design of the data repository literature was reviewed with the aim to identify the most frequent operations of health institutions, for this purpose were reviewed several works, among whose was considered those that discuss the treatment for patients, the operation of the clinical laboratory, the emergency area, the management of the inventory of medicine, and the use of drugs and supplies.

Treatment for patients. In [3], several data model for getting clinical effectiveness are discussed and how *the treatment* given to a patient over a period of time is modeled with a table. This requires a folio number, the department, the patient type, and type of payment, the number of days of medication, the gender of the doctor, the cost of the medicine, the cost of diagnosis, the cost of revision and the cost of medication administration, see Table 1.

It may be inferred from the above table that there is a need for additional tables with the data in the event, the patient, the Doctor, Medicines and billing.

These elements are fundamental for the control of the costs of patient care. It should be noted that these fields are part of the database of Medicare in the United States of America, which is considered as big data [5].

Operation of the clinical laboratory. Another fundamental element in the operation of health institutions corresponds to the operation of the clinical laboratory. In [6] is proposed a star schema, at the center of this star is the table of the results of the clinical laboratory with primary key (IDResult), which connects with other entities from foreign keys (Fk- Foreign key), in which the patient ID (IDPatient), the id of the Doctor (IDPhys), for the control of the test the Id of the test (IDTest), to locate the test in the IDTime) and time (for the control of facilities (IDLocation). The entities involved are patient, doctor, laboratory test, time, location, all of which are linked to the output table of the Clinical Laboratory.

Operation of the emergency area. The operation of the emergency room mix some of the entities described above in a massive form [7], namely laboratory testing and treatment however includes the entities: radiological examination, Evolution of the exam, Episode, monitoring and diagnosis, all of them including the temporality (timestamp) on each one.

Inventory of medicine. The inventory of the medicine and particularly high-value of the drug was modeled by [8] cited in [9], it proposes a very simple form for the control of stocks and of the possible deviations in the same:

Amount stolen = (Stocks at the start of the day - Stocks at the start of the next day) - Medicine Prescription Assortment (1).

Therefore, if at the start of the day had 100 pieces of a medicine, and the next day there are only 90 pieces, but only took 7 pieces, there would be a shortfall of 3 pieces:

$$\text{Stolen Inventory} = (90) - 100 - 7 = 10 - 7 = 3.$$

It follows the detail of the recipe, which can be used to calculate the cost of medicines. It should be noted that it would be required to associate this information with the Doctor and Patient. To identify who issued the treatment to and to who was ad ministered.

The scheme described in Table 2 can be adapted for the management of supplies to health institutions, for example: syringes, gauze, etc. as well as equipment.

Considering the above discussed, the repository design is shown in figure 1 as entity relationship diagram. This schema concentrates the management of supplies of drugs, equipment, and others. It shows the entities involved and events that can occur in the day-to-day operations of a health institution, which is modeled on what has been called IDEvent in the Events table. In each event is given appointment the doctor, patient, and if necessary the use of laboratory tests, as well as the results, that trigger in treatment, which may require one or more prescriptions which in turn impacts the inventory level.

Table 2. Entities required for inventory control of medicines, based on [9].

<<Medicine>>
IdMedicine
Name
Presentation
Existence
Price
Time Stamp
<<Prescription>>
IdPresciption
IdDoctor
IdPatient
Date
IdMedicine(s)

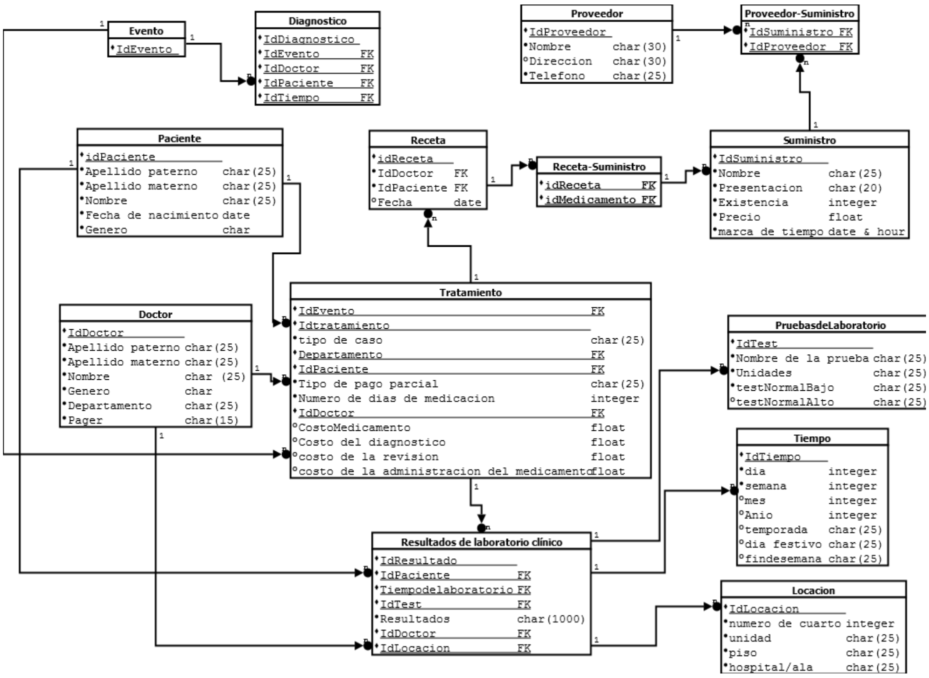


Fig. 1. Proposed Data Repository.

It should be noted that in this design particular care in handling of the spaces, of the costs of services, as well as of the times in which the processes are carried out. We will focus in the inventory management.

According to [10], the inventory is defined as "all the money that the system invests in the purchase of things the system aims to sell", if we apply this concept to the health sector and what we modify slightly would be something like: "All the money the health sector invests in the purchase of medicines and supplies that the system uses to deliver health quality services.”

In this sense, of the inventory conceived as a resource indicating that [11]: "The inventory is any stored resource that serves to satisfy any need now or in the future", these authors indicant that all organizations have some system of Planning and control of inventory, in this way a bank has methods to control your inventory of cash, a hospital has methods for controlling the bank of blood and other important inputs.

3 Methodology

For this investigation, the method of study case will be used to analyze the inventory of a health clinic in the southeast of Mexico using data analytics, the case study method is a methodological tool that can be applied in any area of knowledge and allows quantitative and qualitative methodologies to be used [12].

The purpose of all models and techniques used on inventories is to determine in a rational way how much and when to order [11]. For this research we used the model of inventory for cost reduction in the real world described by [11] and adapted from [13], which consists of the following stages:

1. Definition of the problem.
2. Development of the model.
3. Data collection.
4. Development of a solution.
5. Evidence of the solution.
6. Analysis of the results.
7. Implementation of the results.

Definition of the Problem. How to reduce missing medicines and supplies for a health care clinic in the southeastern of Mexico using data analytics.

Development of the Model. The figure 2 shows the proposed model. The first stage determines what to purchase and store for the inventory. In the second stage is carried out the demand forecast for inventory, and the third stage is controlled the process. A feedback loop allows adjusting the plan and the projection from the experience and observation [11]. For a better control is proposed the use of a Two-bin Kan-Ban system (see Figure 2).

Data Collection. For proof of the use of medicines and supplies, information was provided from operations of a health clinic of the Mexican Southeast, with information from the period from January to September 2016. The fill format is displayed in the following figure.

As can be seen in the previous figure, the inventory control is done manually; these control sheets were captured and preprocessed in Excel. Highlighting the Elements Date, Area, Quantity, Material requested, Stocks, Name, Existence and Observations (see figure 4).

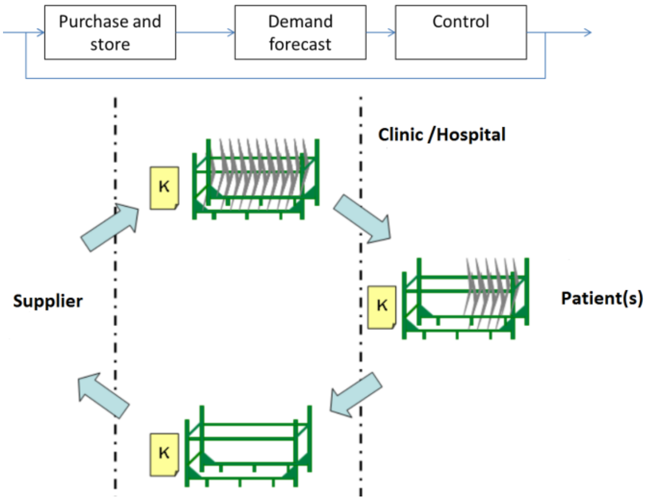
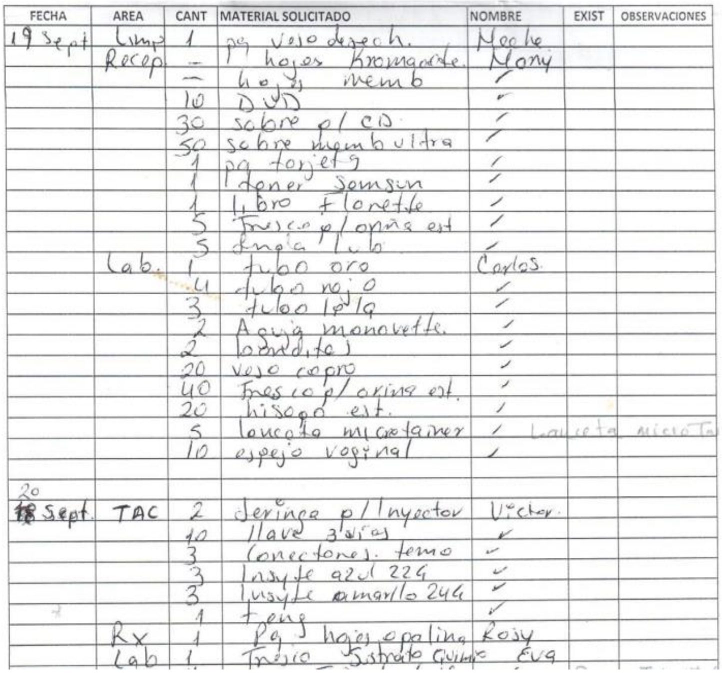


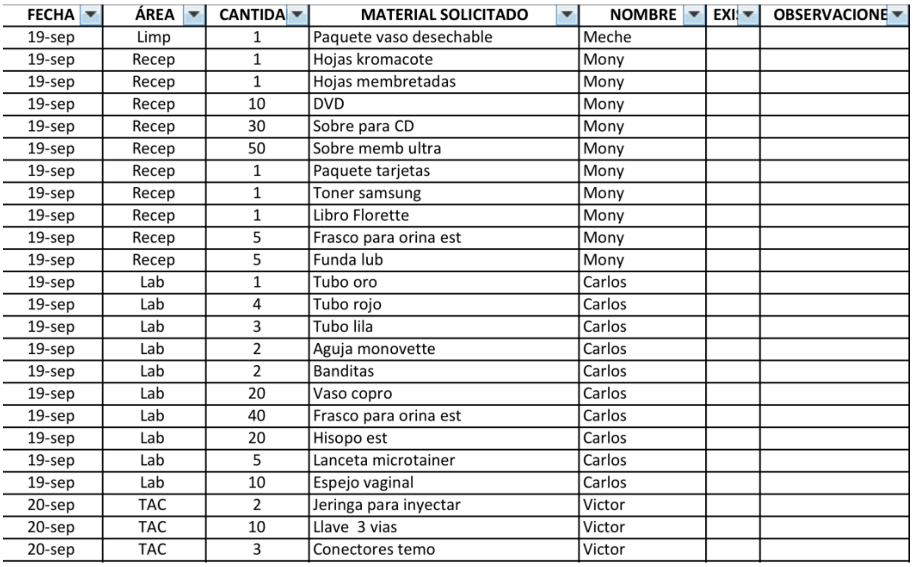
Fig. 2. System of planning and inventory control [11].



FECHA	AREA	CANT	MATERIAL SOLICITADO	NOMBRE	EXIST	OBSERVACIONES
19 Sept	Limp	1	pq vaso desech.	Meche		
	Recep	–	hojas Kromacote.	Mony		
		–	hojas memb	✓		
		10	DVD	✓		
		30	sobre p/ CD	✓		
		50	sobre memb ultra	✓		
		1	pq tarjetas	✓		
		1	toner Samsung	✓		
		1	libro florette	✓		
		5	Frasco p/ orina est	✓		
		5	Angela tub	✓		
	Lab.	1	tubo oro	Carlos.		
		4	tubo rojo	✓		
		3	tubo lila	✓		
		2	Aguja monovette.	✓		
		2	banditas	✓		
		20	vaso copro	✓		
		40	Frasco p/ orina est.	✓		
		20	hisopo est.	✓		
		5	lanceta microtainer	✓	Lanceta	microTa
		10	espejo vaginal	✓		
20 ~~18~~ Sept.	TAC	2	Jeringa p/ Inyector	Victor.		
		10	llave 3 vias	✓		
		3	Conectores. temo	✓		
		3	Insyte azul 224	✓		
		3	Insyte amarillo 244	✓		
		1	Teng	✓		
	Rx	1	Pq hojas opalina	Rosy		
	Lab	1	Frasco Sustrato Quimio	Eva		

Fig. 3. Detail of Control Sheets of medicines and supplies in a clinic in the health sector.

As can be seen, the previous table database can be used to calculate the deviations in the management of the inventory as it has the field inventories. However, this information is not provided.



FECHA	ÁREA	CANTIDA	MATERIAL SOLICITADO	NOMBRE	EXI	OBSERVACIONE
19-sep	Limp	1	Paquete vaso desechable	Meche		
19-sep	Recep	1	Hojas kromacote	Mony		
19-sep	Recep	1	Hojas membretadas	Mony		
19-sep	Recep	10	DVD	Mony		
19-sep	Recep	30	Sobre para CD	Mony		
19-sep	Recep	50	Sobre memb ultra	Mony		
19-sep	Recep	1	Paquete tarjetas	Mony		
19-sep	Recep	1	Toner samsung	Mony		
19-sep	Recep	1	Libro Florette	Mony		
19-sep	Recep	5	Frasco para orina est	Mony		
19-sep	Recep	5	Funda lub	Mony		
19-sep	Lab	1	Tubo oro	Carlos		
19-sep	Lab	4	Tubo rojo	Carlos		
19-sep	Lab	3	Tubo lila	Carlos		
19-sep	Lab	2	Aguja monovette	Carlos		
19-sep	Lab	2	Banditas	Carlos		
19-sep	Lab	20	Vaso copro	Carlos		
19-sep	Lab	40	Frasco para orina est	Carlos		
19-sep	Lab	20	Hisopo est	Carlos		
19-sep	Lab	5	Lanceta microtainer	Carlos		
19-sep	Lab	10	Espejo vaginal	Carlos		
20-sep	TAC	2	Jeringa para inyectar	Victor		
20-sep	TAC	10	Llave 3 vias	Victor		
20-sep	TAC	3	Conectores temo	Victor		

Fig. 4. Detail of Control Sheets of medicines and supplies in southeast clinic (own source).

Table 3. Measures of the Theft in Public Hospitals in Venezuela [9] based on surveys to the Staff.

Concept	**Percentage of Staff that indicates that the theft occurs**	**Stolen percentage (%)**
Surgical Supplements	67.0	10.1
Medications	64.4	13.4
Computer	50.1.	5.7
Food	42.3	12.2
Other	28.2	3.4

Statistical Analysis. In [14] was used univariate logistic regression analysis to identify the association between supply change management and the audit variables for a review of POC (Point of Care) diagnostics in 100 Clinics in Ghana, based on this analysis they identified similar problematics to those experienced in the Clinic of southeast of Mexico like high clinic attendance, poor documentation of inventory level, and poor monitoring of monthly consumption level. For this study we used two variables: the quantity of material requested and the date to model the usage of medicines and supplies during a period of time, by using of time series which works in linear data to make forecasting [15, 16].

Development of a solution. According to [17] cited in [9] the level of missing in public hospitals round the 10% on average, varies according to the type of article, as well as surgical supplements have a 10.1% of missing, medicines the 13.4%, the team 5.7%, the food 12.2% and other 3.4%, see table 3.

3.1 Algorithm for the Optimization of Critical Items (Supplies and Medicines)

By what was previously described, in order to optimize the items critical to the operation of health institutions intends to use the methodology developed by [18], which consists of 4 stages:

1. Identify the critical items through an ABC classification, which considers factors such as costs and inventory turns, so that they selected the items they need a policy of inventory.
2. Analysis of the behavior of the demand of the items to identify the existence of patterns. At this stage, we identified the demand of the item in order to identify the most appropriate form of their prognosis, and thus be able to estimate their behavior.
3. The parameters of the inventory policy according to the results of the previous stages.
4. Inventory policies evaluation. For health institutions is proposed the implementation of a policy of continuous review (s, Q).Where s represents reordering point and Q represents a product request of a fixed size Q. Each time you reach the states, system generates an order with the amount Q. This policy is also known as "Two-bin system", this system consists of two Kan-Ban systems in tandem and has been shown to give excellent results in public health institutions [19] and exemplified at St Clair Hospital in https://www.youtube.com/watch?v=yjSwwPF5BUU.

This could be achieved according to the following formulas:

$$EOQ = \sqrt{\frac{2AD}{v\,r}}. \tag{2}$$

The Equation 2 represents the optimal amount of order, expressed in units, where:

Fixed cost of placing an order in monetary value.
D- Annual Demand,
v- Value,
r- ratio,
v*r - Inventory Management Cost as a percentage of the value of the product, in annual percentage rate:

$$ROP = \mathrm{d*t}. \tag{3}$$

The Equation 3. Allows you to calculate the refill point, where:

D is the demand in units/day,
T is the waiting time known (days).

Evidence of the solution. Next, we show the performance of the proposed policies by means of simulation in worksheet of the selected products.

In the database of the Clinic under study, we identified critical items that need a policy of inventory, some examples are: sealed envelopes, letterhead sheet, urine bottle,

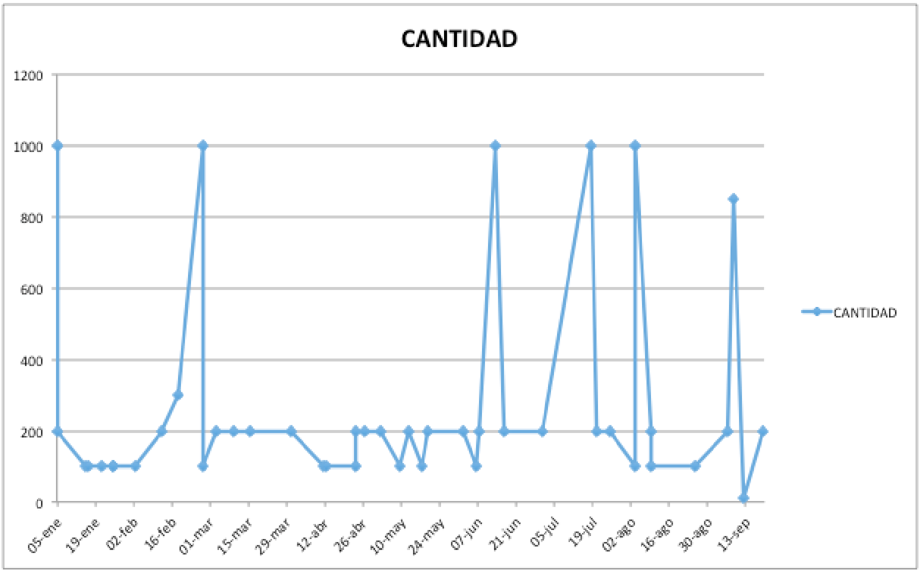


Fig. 4. Identification of critical items in the Clinic of Chiapas, the figure shows the consumption of sealed envelopes.

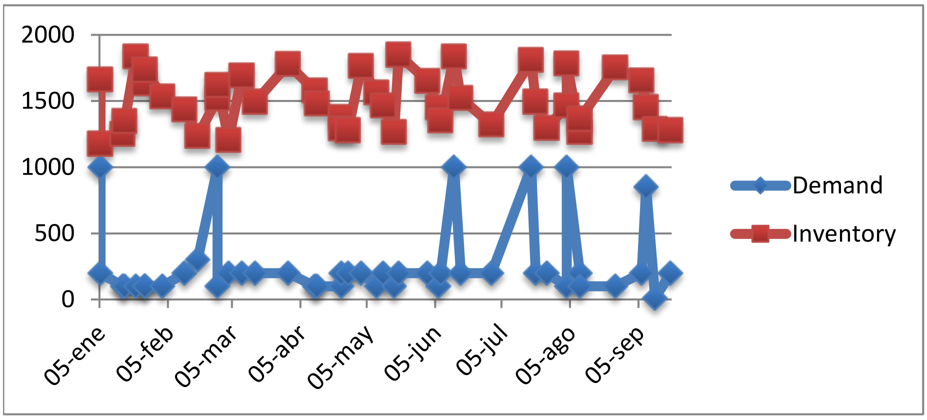


Fig. 5. Simulation of item on letterhead sheet stationery.

copro cup, yellow tips, speculum, tongue depressors, among others. Analysis and implementation of the results will be discussed in next section.

4 Results and Discussion

We proceeded to analyze the behavior of the demand. Table 4 shows the actual demand (January to September) of some critical items.

1. The parameters passed to inventory policy. To do this we tested a value of 100, a cost of inventory management 5% annual and a "fixed" refill (ROP) of 10%. With

Table 4. Identification of critical items.

No. Article	Critical items (high consumption)	Units	Cost	Unit cost*
1	On the letterhead stationery	11760	3.5	41160
2	Letterhead Sheet	21728	1.5	32592
3	Copro cup	1630	4	6520
4	urine bottle	4949	4.3	21280.7
5	Yellow Tips	5899	0.4	2359.6
6	Speculum	341	5	1705
7	Tongue depressors	1886	1	1886

Table 5. Policy parameters of inventory, resources (own and mercadolibre.com).

No. Artic	Critical items (high consume)	Refill point S	Quantity Q
1	On the letterhead stationery	1176	686
2	Letterhead Sheet	2172.8	932
3	Glass Copro	163	255
4	For urine bottle	494.9	445
5	Yellow Tips	589.9	486
6	Speculum	34.1	117
7	Tongue depressors	188.6	275

these parameters are calculated the optimum amount of order for critical items under discussion.

2. The parameters passed to inventory policy. To do this we tested a value of 100, a cost of inventory management 5% annual and a "fixed" refill (ROP) of 10%. With these parameters are calculated the optimum amount of order for critical items under discussion.

3. We evaluated the performance of the proposed policies, in the following figure simulation for the item on letterhead stationery, considering the refill point 1176 and an order quantity Q of 686 units.

As can be seen in the simulation, the proposed model ensures that the item is in stock at all times during the operations of the institution of health, that is why it is vital to analyze the behavior of each item and its peak demand periods in time.

5 Conclusions and Future Work

A Data repository to analyze inventory in the Mexican Healthcare system was designed, we proposed the use of data analytics due the high quantity and potential of information managed in this type of systems.

We proposed a two-bin Kan-ban system to control inventory and reduce missing supplies and medicines for a health care clinic in the southeastern of Mexico. We demonstrate is possible to use single data analytics to achieve that goal.

This work is be not means completed, our future work will focus on the use of data analytics mentioned in [20] for Mexican health care institutions.

References

1. Schneeweiss, S., Avorn, J.: A review of uses of health care utilization databases for epidemiologic research on therapeutics. Journal of Clinical Epidemiology, 58, pp. 323–337 (2005)
2. Raghupathi, W., Raghupathi, V.: Big data analytics in healthcare: Promise and potential. Health Information Science and Systems, 2 (2014)
3. Kahn, M.G., Batson, D., Schilling, L.M.: Data model considerations for clinical effectiveness researchers. MedCare, 50 (2012)
4. Deshpande, A.M., Brandt, C., Nadkarni, P.M.: Temporal query of attribute-value patient data: Utilizing the constraints of clinical studies. International Journal of Medical Informatics, 70, pp. 59–77 (2003)
5. Medeiros, B.C., Satram-Hoang, S., Hurst, D., Hoang, K.Q., Momin, F., Reyes, C.: Big data analysis of treatment patterns and outcomes among elderly acute myeloid leukemia patients in the United States. Annals of Hematology, 94, pp. 1127–1138 (2015)
6. Lyman, J.A., Scully, K., Harrison, J.H.: The development of health care data warehouses to support data mining. Clinics in Laboratory Medicine, 28(1), pp. 55–71 (2008)
7. Lin, W.T., Wu, Y.C., Zheng, J.S., Chen, M.Y.: Analysis by data mining in the emergency medicine triage database at a Taiwanese regional hospital. Expert Systems with Applications, 38, pp. 11078–11084 (2011)
8. La Forgia, G., Cross, H.: Health financing and management in belize: An assessment for policymakers. A Compendium of Technical Notes, Cost Recovery, 2 (2001).
9. Di Tella, R., Savedoff, W.D.: Diagnosis corruption: Fraud in latin america's public hospitals. Harvard Business School, Faculty & Research (2001)
10. Goldratt, E.M., Gibler, N.A.: El sindrome del pajar: Cómo extraer información del océano de datos. Ediciones Castillo (1999)
11. Render, B., Stair, R.M., Hanna, M.E., Hale, T.S.: Quantitative analysis for management. Pearson (2012)

12. Cristina, P., Carazo, M.: El método de estudio de caso Estrategia metodológica de la investigación científica. Pensam Gestión, pp. 165–193 (2006)
13. Lee, H.L., Billington, C.: The evolution of supply-chain-management models and practice at hewlett-packard. Interfaces (Providence), 25, pp. 42–63 (1995)
14. Kuupiel, D., Tlou, B., Bawontuo, V., Drain, P.K., Mashamba-Thompson, T.P.: Poor supply chain management and stock-outs of point-of-care diagnostic tests in Upper East Region's primary healthcare clinics, Ghana. PLoS One, 14 (2019)
15. Olson, D.L., Wu, D.: Predictive data mining models. Springer (2020)
16. Madrigal-Espinoza, S.D.: Modelos de regresión para el pronóstico de series temporales con estacionalidad creciente. Computación y Sistemas, 18, pp. 821–831 (2015)
17. Jaén, M.H., Paravisini, D.: Diseño institucional, estructura de incentivos y corrupción en hospitales públicos en Venezuela. RePEc (1999)
18. Aguirre-Lasprilla, S.: Parametrización y evaluación de política de inventario (s,Q) en hospitales: Un caso de estudio en la ciudad de Barranquilla. Prospectiva, 13, pp. 99 (2015)
19. Graban, M.: Lean hospitals: Improving quality, patient safety, and employee engagement. CRC Press (2018)
20. Raghupathi, W., Raghupathi, V.: Big data analytics in healthcare: Promise and potential. Health Information Science and Systems (2014)

ISSN 1870-4069

Image Creation Using Generative Adversarial NetWorks (GAN) Applied to the Fashion Industry

Alejandro Acosta, Alberto Ochoa Zezzatti, Gustavo Delgado,
José Mejía

Universidad Cuauhtemoc,
Maestria en Big Data
Mexico

alejandro.acosta@uabc.edu.mx

Abstract. The object manipulation of an image has become very popular lately. It is common to see applications about facial recognition or any other type of objects [2, 4, 19]. Technology allows people to achieve these tasks fast and in a really powerful way. In the fashion industry, there exists a lot of images due to photographers that can be taken and be analyzed to detect features and metadata. Deep learning has proven to be an excellent approach to the image recognition field by using convolutional neural networks (CNN). In this work, will be presented some image recognition and manipulation applied to the fashion industry by using CNN in order to create new objects from existing ones. Giving objects a certain level of realism on contours, boundaries, textures, colors, and shapes from the original "clothing" object.

Keywords: Deep learning, convolutional neural networks, fashion, clothing, image detection, image manipulation, machine learning, computer vision, generative adversarial networks.

1 Introduction

In the fashion industry, clothing is one of the most important subjects around it. The parades, photography, designers, stylists and much more, work to show to the world how amazing the new clothes are so people may look for that new skirt or new blouse in the stores the very next day to purchase. Technology makes everything better, and this includes the fashion industry. Image manipulation has become popular these days that is applied for practically every aspect of life.

Fashion designing is a particular area in the industry. It is known that this industry relies on designers and their work when designing the next big thing on clothes to present. In addition, it is known that graphic designers help, creating clothes on the computer from what the fashion designers pass to them.

pp. 33–44; rec. 2019-08-09; acc. 2019-09-14

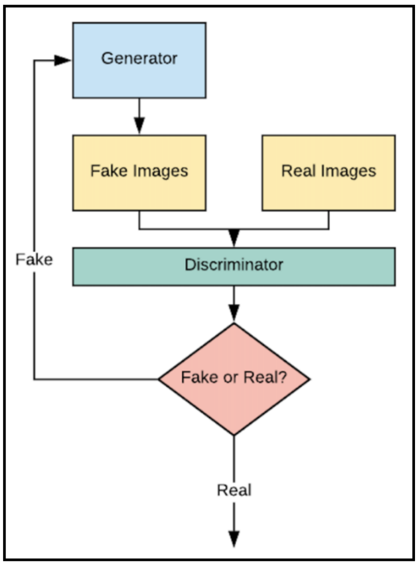


Fig. 1. Flowchart about GANs process.

An outstanding way to improve this process is to create new fashion designs based on a series of previous works created by fashion designers.

This will allow them to create new clothes that they could not even think about it or never occurs to them before [1].

Deep learning algorithms that learn from previous works and create new things are becoming powerful around the internet that demonstrate that it is possible for computers to create new things. Recent research shows that there is an approach that allows deep learning to create new things. The subject is about Generative Adversarial Networks (GAN). This kind of deep learning gives a new generative approach to image processing tasks in computer vision [5, 6].

[7] GANs are a kind of convolutional neural networks (CNN) presented to the world around 2014. These networks combine concepts about machine learning and Game Theory.

GAN model is also commonly known by DCGAN (Deep Convolutional Generative Adversarial Networks) could be understood a game between two agents: a Generator and a Discriminator. El objective of the Generator (G agent) is to generate the best information that it can from the real dataset (training step). On the other hand, the Discriminator (D agent) is in charge of classifying the real data and the fake data generated by the G. At the same time, the G is getting better on data creation since it is learning from the output that the D is discarding (classifying). The goal of this is that G can cheat D by generating realistic information that D can’t distinguish from fake or real information. Both algorithms help each other to get better and improve by themselves [9].

The information that the GAN algorithms will process and will be talked about in this paper are going to be images. Apparel images in specific. There is another characteristic around GANs and it is that are designed to reach a Nash equilibrium at which each player cannot reduce their cost without changing the other players parameters. This makes the algorithms to be prepared so they can create realistic images from any random noise data [10, 17]. The Ds are based on a particular type of algorithm named as Backpropagation. That is what allows to D to be recursive and iterative so it can learn from previous trainings and testing phases. This series of steps also will let the Discriminator to be refined and it is going to be better at distinguishing between fake and real images [11].

2 Related Work

As mentioned already. GAN algorithms are used around the internet for generation data and it is important to highlight those other related works and papers that have been a game-changer on the application that offers. Jun-Yan Zhu, and other colleagues of him, in their paper from 2017 titled “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks” [8], introduced their very famous CycleGan that provides an interesting way to reconstruct images from pair of data that is an input and an output. Quoting the paper mentioned above: “Image-to-image translation is a class of vision and graphics problems where the goal is to learn the mapping between an input image and an output image using a training set of aligned image pairs.”

Also, the paper mentions how this algorithm works in order to accomplish the goal they are proposing. Quoting the paper again, “We present an approach for learning to translate an image from a source domain X to a target domain Y in the absence of paired examples. Our goal is to learn a mapping:

$$G: X \rightarrow Y, \tag{1}$$

such that the distribution of images from G(X) is indistinguishable from the distribution Y using an adversarial loss. Because this mapping is highly under-constrained, we couple it with an inverse mapping:

$$F: Y \rightarrow X, \tag{2}$$

and introduce a cycle consistency loss to enforce $F(G(X)) \approx X$ (and vice versa). Qualitative results are presented on several tasks where paired training data does not exist, including collection style transfer, object transfiguration, season transfer, photo enhancement, etc. Quantitative comparisons against several prior methods demonstrate the superiority of our approach”. This is one of the many works published that are so impressive in terms of impact and the application that can be applied.

The very first time that the GAN name appears, was because of Ian Godfellow [13] in 2014. He is the creator of this type of networks and every work related to GANs are thanks to Godfellow.

There are GANs that are based on Bayesian multivariable probabilistic models, where uses a set of random variables provided a graph. This way a generator can create images from not exactly a vector but from a cloud of data (distribution) [15].

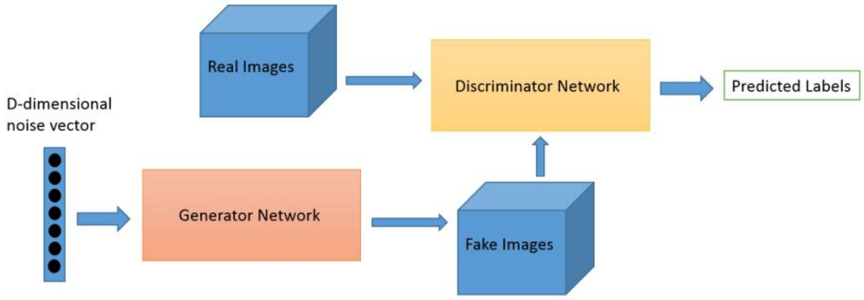


Fig. 2. Data Flow (source: https://www.oreilly.com/ideas/deep-convolutional-generative-adversarial-networks-with-tensorflow).

3 Proposed Approach

Creates a deep learning algorithm that will learn from fashion images and is going to be able to analyze and create new objects by recognizing patterns on the base images using Generative Adversarial Networks.

To address the approach to generate new objects given a dataset that will help to train GAN algorithms, GAN must be defined and describe how these networks work.

[6] GAN generates data distributions through the adversarial process. GAN gradually improves the quality of generations by the adversarial training process. GAN exhibits the following advantages: 1) GAN belongs to the type of non-parametric production-based modeling methods, which does not require prior approximate distributions of training data. 2) GAN works on the whole image and takes less time to generate samples by directly using global information.

What makes GAN so outstanding is its special structures. GAN is a deep adversarial framework consisting of a generative network named generator and a discriminatory network called discriminator.

The generator captures the data distributions, which wish to pass through the test of the discriminator, and the discriminator estimates the probability whether the sample is from true distributions.

By using GAN, it is required to ensure that the two agents are interacting at the same time, Generators and Discriminators. That is what is very interesting about this matter. A Generator is “The Artist”, a neural network that is trying to create pictures of clothes that look real. Discrimination is “The Critic”, a network that is examining clothing pictures to determine which ones look real or fake.

Showing a different diagram but this time as GAN data flow, it can be represented as the following fig. 2.

The discriminator it is consider a classification model or networks that uses unsupervised classification to determine whether the input image is real or not [16, 18, 19].

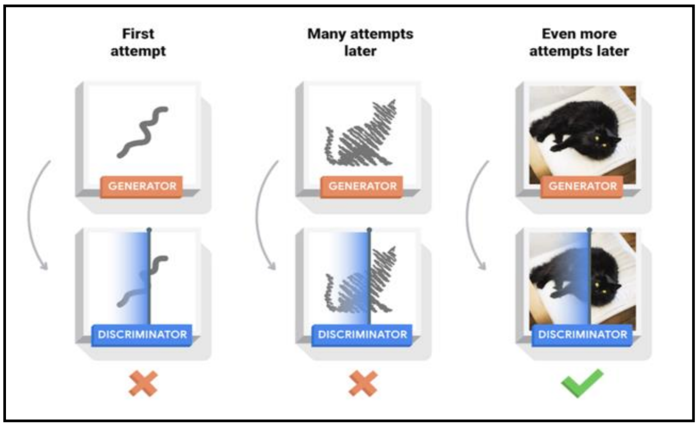


Fig. 3. Illustrative Generator and Discriminator process.

4 Materialization

The first part of the materialization of the work was to find a language and framework that allows handling GANs. Tensorflow and Python do the job excellently. There is an existing repository that is already built to generate images with the GAN approach. The repository belongs to Greg Surma, Github username is "gsurma" and the repository name is "image generator". This repository includes a full implementation of a GAN algorithm with Tensorflow and Python already. This is an existing work that was taken to be modified as needed for this project since the repository is open for exploring. Secondly, the dataset was replaced. Instead of using the dataset on the code repository, it was changed to use a collection of 500 t-shirts images. The project was adjusted for slowdown the epochs because the bigger the epochs amount are indicated, the bigger the computer power and resources will be needed. It is a good idea to run not too many epochs if you are on a computer without the necessary resources. The estimation is that 10 epochs can take up to 4 hours to complete. However, if there are just a few epochs, the results are not going to be as expected. Next, the training begins by using the t-shirt dataset. It is necessary and very important to crop all the images at the same size. For our purposes, all the images were resized to 200 x 200px.

The proposal is to use GAN algorithm based on Tensorflow framework that provides tools to accomplish these tasks. The main framework is based on the Python language that is also very much used for the community as mentioned above.

Looking at the big picture, the process is basically generating an image by the Generator and being analyzed by the Discriminator (Fig. 3). While the discriminator rejects the image, the generator will create a new one based on the rejection made. This is how the generator will learn. Once the discriminator validates and tags the generated image as valid, the image will be considered as realistic.

Each time that the discriminator rejects, the generator will start over again in a next iteration generating a new one to be analyzed. Each iteration, loop or repetition in this

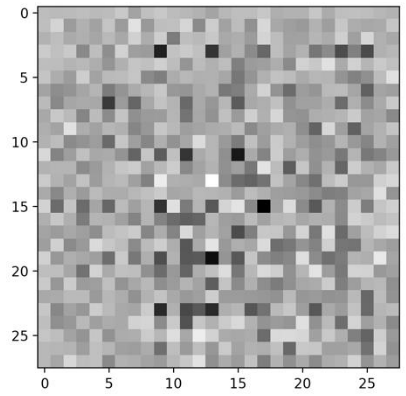


Fig. 4. Generated image by the Generator algorithm.

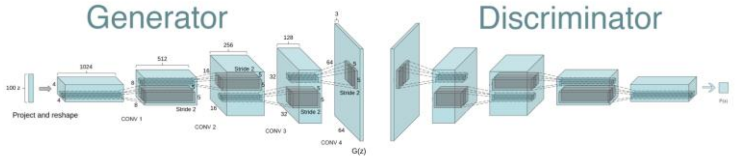


Fig. 5. An example architecture for generator and discriminator networks that uses convolutional layers to process visual information.

paper will be called as an "Epoch". The first part is to create the generator using python code and Tensorflow as primary deep learning framework. The generator uses tf.keras.layers.Conv2DTranspose (up sampling) layers function to produce an image from a seed. It starts with a dense layer that takes the input, then up samples several times until you reach the desired image size of whatever is specified. (Fig. 4).

The following part is for the discriminator to take the image generated from the random noise and determine if is real or fake. The discriminator is a Convolutional Neural Network (CNN-based) image classifier. The discriminator model is the one that will be trained in order to distinguish between real or fake images. To train the model, it was taken a specific in DataWorldTeam a Web site with many adequate datasets with several fashion images. Analyzing and explaining further the Generator and the Discriminator algorithms at the working phase, both can be explained with more details.

The Generator, as mentioned, it takes random noise as input and samples the output in order to fool the Discriminator that it's the real image. Once the Generator's output goes through the Discriminator, it is known that the Discriminator verdict whether it thinks that it was a real image or a fake one. This information has to be used to feed the Generator and perform backpropagation here. If the Discriminator identifies the Generator's output as real, it means that the Generator did a good job and it should be given a good feedback. However, if the Discriminator recognized that it was given a fake image, it means that the Generator failed and it should be given a negative feedback.

The following diagram shows how the generator and discriminator are built from the inside Fig 5.

In Fig. 5, it is explained how DCGAN are used for simple modeling. A 100 dimensional uniform distribution Z is projected to a small spatial extent convolutional representation with many feature maps. A series of four fractionally-strided convolutions (in some recent papers, these are wrongly called deconvolutions) then convert this high level representation into a 64 × 64-pixel image. Notably, no fully connected or pooling layers are used [12]. For the Discriminator, it gets both real images and fake ones and tries to tell whether they are legit or not. The system designers or people implementing the algorithm know whether they came from a dataset (reals) or from a generator (fakes). This information can be used to label them accordingly and perform a classic backpropagation allowing the Discriminator to learn over time and get better in distinguishing images.

When the Discriminator correctly classifies fake and real images, then it can be considered that it has a positive feedback on loss gradient. If it fails at classifying them, it can be said that has negative feedback on it. This process allows the algorithm to learn and get better at every turn. Quoting a paragraph of [12] for visualizing of the Discriminator features, “Previous work has demonstrated that supervised training of CNNs on large image datasets results in very powerful learned features (Zeiler & Fergus, 2014). Additionally, supervised CNNs trained on scene classification learn object detectors (Oquab et al., 2014).

We demonstrate that an unsupervised DCGAN trained on a large image dataset can also learn a hierarchy of features that are interesting. Using guided backpropagation as proposed by (Springenberg et al., 2014), we show in Fig. 5 (figure on that paper) that the features learnt by the discriminator activate on typical parts of a bedroom, like beds and windows. For comparison, in the same figure, we give a baseline for randomly initialized features that are not activated on anything that is semantically relevant or interesting.” This is very important to understand the work that the Discriminator does, because is not just detect if an image is real or not, but to learn that there are some common features in the dataset that needs to be considered as a MUST when the generator presents a generated image and it tries to pass on.

To make these previous paragraphs better understandable too, here is a definition of Backpropagation that has become a very important part of GANs. It is an algorithm used to efficiently train Artificial Neural Networks (ANN). Its main feature is that it is a Recursive algorithm that go back and learn to get better down the road until it reaches the goal that it was implemented for. In order for the Generator and Discriminator to learn, it is necessary to implement Loss Functions to introduce the backpropagation idea into this. A Loss Function is used for measuring the discrepancy between the two outputs. For the developed work on this paper, it is also important to consider a very important thing before doing the training of the algorithms. It is the Optimizers. The idea of this is to adopt optimizers in order to balance the Generator and Discriminator learning rates. Optimizers allow the algorithms to balance its rate for learning. This is

Table 1. Orthogonal array.

Variable Value								
A	B	C	D	E	F	G	H	Color
H	H	H	H	H	H	H	L	1
H	H	H	H	H	H	L	H	2
H	H	H	H	H	L	H	H	3
H	H	H	H	L	H	H	H	3

important since this approach is working with both at the same time and will not be a good idea if only one algorithm is actually learning and the other not.

4.1 Training

Now, for the training of the algorithms it is going to be set with 1 Epoch to test the functionality of the Generator and Discriminator. Results may vary based on the epochs that are specified. For the first epoch, as can be expected, the generated image will be discarded for it is a blurry image of pure noise data. As aside note, it is important to continually verify the optimizer and loss functions at each Epoch so it can balance both algorithms the better possible.

In order to be able similar, the most efficient arrangement of clothes in a dresser, we developed an atmosphere able to store the data of each one of the representing these clothes, this with the purpose of distributing of an optimal form to each one of the evaluated clothes. One of the most interesting characteristics observed in this experiment was the diversity of the cultural patterns established by each clothe with respect to their symbolic capital. The scenes structured associated with the agents cannot be reproduced in general, since they only represent a little while dice in the space and time of the different clothes.

These represent a unique form and innovating of adaptive behavior which solves a computational problem that it does not try to clustering the clothes only with a factor associated with his external appearance (attributes of each clothe), trying to solve a computational problem that involves a complex change between the existing representations.

The generated configurations can be metaphorically related to the knowledge of the behavior of a potential customer with respect to an optimization problem (to select culturally specify similar clothes, without being of the same kind [3]). The main experiment consisted of detailing each one of the clothes on a collection, with 500 agents, and one condition of unemployment of 50 époques, this allowed us to generate the best selection of each kind of clothes and their possible location in a Dresser, which was obtained after comparing the different cultural and social similarities from each clothe, and to evaluate with Multiple Matching Model each one of them [10].

The developed tool classified each one of the clothes pertaining to each kind, with different wardrobe for clothes that included identity and for clothes only with cultural identity, this permit identifies changes in the time respect at other clothes. The design of the experiment consists in an orthogonal array test, with the interactions between the

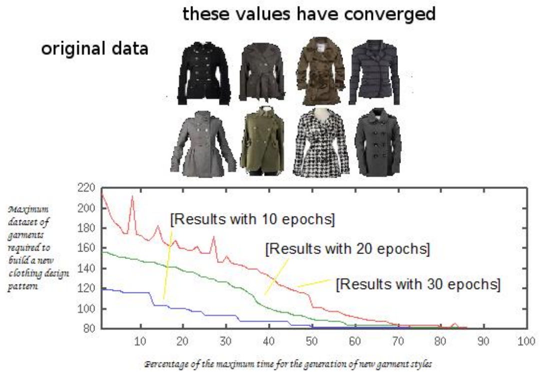


Fig. 6. Representation of our GAN to design new models of clothes.

Table 2. Comparison between epochs.

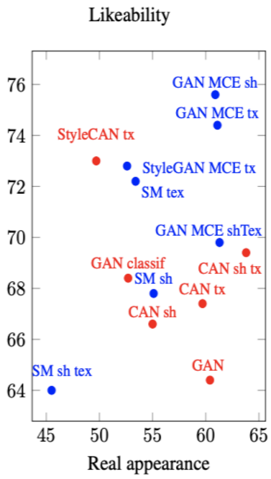
	1 Epochs	10 Epochs	50 Epochs	100 Epochs
Image Size	128	128	128	128
Noise Size	100	100	100	100
Batch Size	64	64	64	64

variables: emotional control, ability to fight, intelligence, agility, force, resistance, social leadership, and speed. These variables are studied in a range of color (1 to 64). The orthogonal array is L-N (2**8), in other words, 8 factors in N executions, N is defined by the combination of possible values of the 8 variables the possible range of color (see Table 1).

According to the results obtained for the time of accommodation in the transport must be performed to accommodate the time of unloading in the following design. On the basis of the results obtained in the experiment, the similarity average between clothes, it was found that the average is 4.0625, this means that you have to improve in the management of the use of the combination colors and standardize according to the weather.

Other factors that affect are the turn as well as the use of accessories, such as: bracelets, necklaces, pins and charms, an implementation in order to resolve this problem is the use of fashion visualization.

As can be seen to hold this type of materials is necessary to use merchandise direct and indirect. The merchandise indirect secure ground, it gives resistance to lateral and longitudinal movement of the clothes. While direct ties anchored to manikins in the dresser. To perform this process, it is necessary to count with the correct dimensions and the distance from the street and the height of these, cables or chains, taking into account the material of these materials.



Method/Human Method	over- all	shape nov.	shape comp.	tex. nov.	tex. comp.	real fake
DCGAN MCE shape	**3.78**	**3.58**	**3.57**	**3.64**	**3.57**	**60.9**
DCGAN MCE tex	**3.72**	**3.57**	3.52	**3.61**	**3.58**	**61.1**
StyleCAN tex	3.65	3.37	3.31	3.44	3.21	49.7
StackGAN	3.62	3.45	3.38	3.43	3.33	51.9
StyleGAN MCE tex	3.61	3.38	3.29	3.50	3.37	53.4
DCGAN SM tx	3.60	3.43	3.42	3.46	3.41	52.6
StackGAN MCE tex	3.59	3.36	3.31	3.44	3.28	55.9
StyleGAN	3.59	3.28	3.21	3.27	3.15	47.2
StackGAN CAN sh	3.51	3.56	**3.56**	3.58	3.40	50.7
DCGAN MCE shTex	3.49	3.40	3.24	3.40	3.31	**61.3**
StackGAN CAN tex	3.48	**3.57**	3.54	3.55	3.50	48.4
DCGAN CAN shTex	3.47	3.28	3.18	3.33	3.16	**63.8**
StackGAN MCE sh	3.45	3.27	3.16	3.28	3.12	60.4
StackGAN CAN shTex	3.42	3.37	3.32	3.44	3.32	49.5
DCGAN classif	3.42	3.32	3.32	3.37	3.29	52.7
DCGAN SM sh	3.39	3.27	3.12	3.30	3.23	55.1
DCGAN CAN tex	3.37	3.23	3.12	3.35	3.09	59.7
DCGAN CAN sh	3.33	3.28	3.16	3.27	3.12	55.0
StackGAN MCE shtex	3.30	3.25	3.28	3.31	3.27	41.5
DCGAN	3.22	2.95	2.78	3.24	2.83	60.4
DCGAN SM shTex	3.20	3.10	3.00	3.13	3.1	45.5

Fig. 7. Analysis of various studies around different types of GAN.

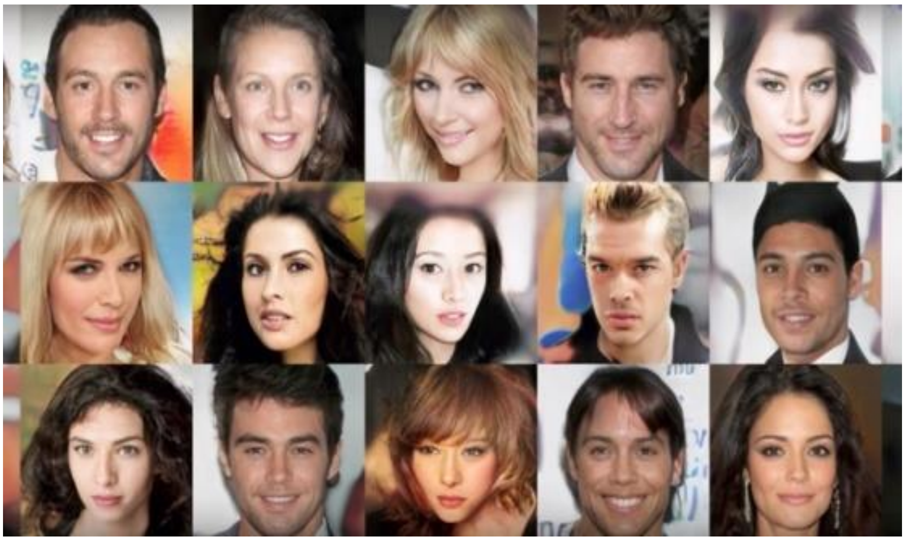


Fig. 8. Representation of Tall Models built with a GAN using 1000 diverse faces on a dataset of DataWorldTeam.

4.2 Results

All the very first Epoch results are expected to be images with no sense at all since the generator is basically constructing "something" that it doesn't have any reference on either what it is building is right or not. Until the discriminator tell, it is obviously in the next steps, see Figure 6.

4.3 Comparison between Epochs

Here it can be visualized a comparison rates between the different Epochs and the adjustments on every specs around the optimizer and loss functions. We show in Table 2 these features.

4.4 Analysis of GANs

In the paper "Design Inspiration from Generative Networks" [14], the authors conducted a study where were mixed 2 different generations. 500 total images picked randomly from 5 best models and 300 real down-sampled images from a RTW dataset. They asked if the images were real or generated to about 45 participants who rated 20 images each in average. They obtain 20% of the generations thought to be real, and 21.5% of the original dataset images were considered to be generated.

In this study, it is very notable that DCGAN networks are very reliable to deliver good results than the rest of GANs.

5 Conclusion

The generative process offered by GANs method has the advantage that uses a simple approximation of sampling from an unknown distribution, such as the current presented by the Generator from any noise seed data. Unfortunately, GAN requires a lot of computing power to run a good number of epochs to ensure generating very good output images. That means the algorithm will not reach its optimal level because it requires many epochs that used in this work.

Another disadvantage is that both algorithms, Generative and Discrimination are not synchronized due to the initial conditions. Discriminator has a better start since it has a standpoint because it can compare with the actual dataset. Therefore, thinking in game theory, discrimination is wining.

Following this logic, the Generator requires more time to create better copies of the fake images to reach the point when it can beat the Discriminator [7]. In future research it is possible to make diverse tall models with different aspects to modelling clothes in a specific society as Guam, Nepal or Timor-Leste with average size of 1.57 to female or 1.64 to male, as is shown in figure 8.

References

1. Liu, Z., Yan, S., Luo, P., Wang, X., Tang, X.: Fashion landmark detection in the wild. Lecture Notes in Computer Science, pp. 229–245 (2016)
2. Lao, B., Jagadeesh, K.: Convolutional neural networks for fashion classification and object detection. Semantic Scholar (2015)
3. Yu, W., Liang, X., Gong, K., Jiang, C., Xiao, N., Lin, L., Yat-Sen, S.: University, DarkMatter AI research. Layout-Graph Reasoning for Fashion Landmark Detection (2019)

4. Cetinic, E., Lipic, T., Grgic, S.: A deep learning perspective on beauty, sentiment and remembrance of art. In: Institute of Electrical and Electronics Engineers (IEEE), pp. 1–1 (2019)
5. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Visual Geometry Group, Department of Engineering Science, University of Oxford (2015)
6. Yangjie Cao, Li-Li Jia, Yong-Xia Chen, Nan Lin, Cong Yang, Bo Zhang, Zhi Liu, Xue-Xiang Li, Honghua Dai: Recent advances of generative adversarial networks in computer vision. In: Institute of Electrical and Electronics Engineers (IEEE), 7, pp. 14985–15006 (2019)
7. Atenas, F., Sanhueza, F., Valenzuela, C.: Redes neuronales adversarias convolucionales para generación de imágenes (2017)
8. Jun-Yan, Z., Park, T., Isola, P.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks efros. Berkeley AI Research (BAIR) laboratory, UC Berkeley (2015)
9. Li, F.F., Johnson, J., Yeung, S.: CS231n: Convolutional neural networks for visual recognition (2017)
10. Chernov, A.V.: On some approaches to find Nash equilibrium in concave games. Automation and Remote Control, 80(5), pp. 964–988 (2019)
11. Kanezaki, A.: Unsupervised image segmentation by backpropagation. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 760–769 (2018)
12. Radford, A., Metz, L.: Unsupervised representation learning with deep convolutional generative adversarial networks (2016)
13. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Annual Conference on Neural Information Processing Systems (NIPS) (2014)
14. Sbai, O., Elhoseiny, M., Bordes, A., LeCun, Y., Couprie, C.: Design inspiration from generative networks. Cornell University (2018)
15. Reyes, M.A., Sánchez, L.A., Mote, R.E.: Una propuesta genética y bayesiana para resolver problemas de clasificación en aplicaciones médicas. Research in Computing Science (2016)
16. Figueredo-Avila, G.A.: Clasificación de la manzana royal gala usando visión artificial y redes neuronales artificiales. Research in Computing Science 114(1), pp. 23−32 (2016)
17. Ortiz-Rangel, E., Mejía-Lavalle, M., Sossa-Azuela, H.: Using pulsed neural networks to improve filtering of images contaminated with gaussian noise. Computacion y Sistemas, 21(2), pp. 381−395 (2016)
18. Zecua, E., Caballero, I., Martínez-Carranza, J., Reyes, C.A.: Clasificación de estímulos visuales para control de drones. Research in Computing Science 114(1), pp. 201−212 (2016)
19. Mejía-Lavalle, M., Lux, M., Pérez, C., Martínez, A.: Digital images text detection as strategy to improve content-based image retrieval. Research in Computing Science (2016)

ISSN 1870-4069

Biomimetric Drone for Controlling Bird Pests and Optimizing Citriculture

Antonio Romero[1], Eddy Sánchez De la Cruz[1],
Alberto Ochoa[2], Alberto Hernández[3]

[1] Instituto Tecnológico Superior de Misantla,
Mexico

[2] Universidad Autónoma de Ciudad Juárez,
Mexico

[3] Universidad Autónoma de del Estado de Morelos,
Mexico

{192t0033, esanchezd}@misantla.tecnm.mx,
jose_hernandez@uaem.mx,alberto.ochoa@uacj.mx

Abstract. The use of a biometrically inspired intelligent drone that is capable of reducing the effects caused by bird attacks on citrus harvest, in order to determine and reduce the size of the flocks and optimize the citrus. Currently, a relevant aspect in farming systems around the world, is the effect of poultry pests, which reduce agricultural productivity significantly annually, therefore, through the use of drones, provide visual recognition to prevent indirect effects that this type of pests generates, through the incorporation of the DERT algorithm, where a random sample of n points is generated in the solution space, in order to evaluate and order them according to their value, in addition, the initial solutions they are generated randomly in the feasible region of the problem, and the search phase is carried out by means of some local search procedure once an initial solution has been generated, they determine whether or not the improvement method is applied to it.

Keywords: Smart drone, biomimetric, citriculture, visual recognition.

1 Introduction

The main problems that farmers face are plant diseases and pests that affect the crops, which causes significant losses and threatens the health food safety for the population, as well as economic problems for the farmer. Currently the technique used to detect attacks of birds in crops is a traditional method of control in the fields, the direct ocular inspection, which consists of personally supervising the plantation in a visual way, this method is slow and is not applicable to large areas of ground.

However, there are others, which provide efficiency, but at a higher cost, from space the detection of attacks in crops is suitable for viewing from altitude. That is why a biometrically inspired unmanned aircraft can be optimal to monitor a specific crop, in this way, drone technology offers a solution to the problems or difficulties that currently exist for a farmer, since they allow to expand inspections more extensive and complete, with their help the work on the agricultural sector is optimized, in addition, the manual work is reduced to have this collaboration, because it benefits the activities of the field because the drone becomes the eyes of the farmer and facilitates the task of the people with the help of the cameras capture images that show if the crop is being attacked by a pest, in the same way, show the state of the plantation to know the probabilities and the risks of this.

On the other hand, there is a variety of techniques to repel pests, there is a set of equipment and tactics that cause a state of alert in poultry species, which interrupt or prevent access to their feeding area. Through the use of sound, visual or mobile elements in areas where the damage is greater, these methods must be used in a precise manner so that their effect can be evident.

1.1 Drones and Aereal Photography

A creative and sustainable solution to this problem has come hand in hand with photography, this gives another perspective of analysis for different fields of research such as cartography, forestry, aquaculture, archeology, agriculture, livestock, among others, since integrating different types of cameras allows to obtain information to produce systems that help in different areas. Part of the trends of aerial photographs is that they are used as a means of information, which are processed to obtain useful information for various analyzes. And a fundamental part is to have the information regarding the catches to facilitate their analysis. The interest in smart vehicles is growing as a result of its large field of applications. Drones are becoming increasingly popular due to the characteristics they present, as well as their numerous applications in surveillance, inspection, search and rescue, among other uses. In addition, they have the advantage of being able to be implemented in both the private and public sectors. Among the applications with drones are the following.

Surveillance: in Mexico, the government of Puebla, carries out surveillance with elements of the police supported by drones, using them to monitor marches, rallies and high-risk circumstances, according to [1]:

- Supervision and control of land: they provide support in areas of difficult access or irregular, drones move with great ease, as they are in fields of agriculture, fish farms, etc. [2].

- Stress detection in plants: The spectral signature can reveal if individual plants are thriving or if they are subject to stress due to drought, nutritional deficiencies or are under attack by insects or viruses [3].

Fig. 1. DJI Drone used in pest detection.

- Fruit counting: they are able to digitize the key characteristics of the plantation, and the high resolution images of the drone clearly show the individual coconut trees, allowing them to make a visual count of the total number of palms [4].
- Filming: currently the entertainment industry equips unmanned vehicles with high resolution cameras to obtain high quality aerial photographs and videos.

In contrast, the Kowat company studies the fear of birds against their predators and builds drones that mimic their appearance and behavior, so they can scare them efficiently. The devices also have speakers that reproduce the sound of predators and an autopilot. The latest models of the company incorporate sensors that allow them to dodge obstacles and pursue objectives in automatic flight mode. These improvements optimize flight control and allow different flight patterns to be programmed that reproduce the behavior of predators more efficiently, once the species of birds that attack crops or fish farms have been analyzed, the company selects a type of drone and defines an area of takeoff and landing. Subsequently, the hours in which the drone should fly are programmed depending on the behavior of the birds and the flight patterns are programmed. The drone will take off and land automatically at the scheduled times. The effectiveness of the drone when frightening birds is checked periodically to make adjustments in the timing and flight pattern, it is necessary to have a specialized technician to put the drone to work and make periodic checks of its operation. Once the initial adjustments are made, the technology operates automatically and it is not necessary to have specialized labor [5] (see Fig. 1).

1.2 Drone Technology for Combating Pests

Inbright, a Mexican company, developed a drone that gives relevant information in the field of agriculture such as soil temperature, the number of plants, detect if there are pests in crops, through images obtained by the drone, it is possible to obtain information such as how large the plant is, the distance between plants, the average areas where the terrain is drier and where more humid.

Fig. 2. Fruit Aggressor Bird.

If there is a pest, identify the progress of the pest based on the damage it is doing to the plant. The unmanned aircraft "fixed wing" has thermal and infrared cameras that give the values to farmers so they can make decisions based on them.

The company offers two types of cameras for the fixed-wing drone. The first one is thermal and it is the one that obtains the data of the temperature of the soil together with the plants, and it can be used to determine the culture strategy. The second is infrared, and serves to gather information about soil moisture and diseases or pests. Subsequently, a software built by the company evaluates the images obtained by the cameras positioned in the unmanned aerial vehicle and generates the specifications [6].

2 Control Avian Pests Using Biomimetric Drones in Citrus Harvest

Currently, there are mechanisms to address the problems caused by insects and rodents in crops, including methods for the protection of growing areas have experienced significant progress, allowing to combat to a large extent the harmful effects that these pests cause them. However, there are still difficulties to achieve the same effectiveness in the control of other types of pests that affect the food industry, bird pests. In addition, the economic aspect and the climatic effects provide support in the increase of crop production, it is certain that humans will not be the only ones that will be at the end of the consumption chain. There will also be birds, who, in a discretionary way, produce considerable damage to the agricultural sector, therefore, it is necessary to use vision systems that greatly facilitate the tasks to be developed, which, together with other systems, can make both the control of the crops and the measures are carried out autonomously. [7] (see Fig. 2).

According to, [8] for the pest control of birds there is not a single method suitable for all cases depends on each particular situation, the species to be controlled, the number of specimens and their attachment to the place to be protected.

Table 1. Veracruz, the largest citrus producer nationwide.

State of the republic	**Production percentage**
Veracruz	63%
San Luis Potosi	13%
Michoacán	13%
Tamaulipas	11%

Therefore, before choosing any of the available methods, it is advisable to consult an expert that indicates which could be effective in each specific case, with the aim of using time and money in those that really have a chance of success. In this regard, it is important to note that many of the wild birds are carriers of pathogenic microorganisms, especially bacteria such as Salmonella, Campylobacter and Escherichia coli. On the other hand, it should be mentioned that the use of technology in the agricultural sector is still minimal, because the issue is unknown and causes uncertainty, as well as doubts in farmers who use traditional methods in their hectares, causing the impossibility of evaluating and controlling the state of their crops in a precise way and, therefore, implies that these are more prone to infestations of pests and diseases, besides that the cost to mitigate these problems is very high. Given the above, in the city of Misantla, there has been little use for Information and Communication Technologies for pest control of birds, therefore, in this project, it is proposed to implement drone technology to help solve the problems that the farming society of the region presents.

2.1 Proposed Methodology

A relevant aspect of the research was to determine the crop that required the implementation of drone technology to optimize the processes and thereby generate improvement in the agricultural sector as according to agricultural statistics conducted in 2017 by SEDARPA and the state government of Veracruz, the city of Misantla and its surrounding communities, are the main producers of citrus fruits, that is, producers of orange, lemon, among others, at the state level. (see Table 1). Once the harvest is defined, define what problems this presents, where the main problems are contamination and total loss of the product due to avian pests.

In this region of Veracruz, there are different types of birds associated with damage and consumption of this type of crop in particular, which is why the visual reconnaissance of the unmanned aircraft, allows specifying which areas of this crop is being affected by pests, as farmers in the area now resort to the traditional way of monitoring their crops, which is not as efficient as the use of drone technology, since it allows covering larger areas in less time.

Regarding the production of the sown area, the following table contains an estimate of the production generated by the state (see Table 2). That is why, the issue to be addressed is citrus, one of the most productive techniques in the state of Veracruz, being one of the largest producers.

Table 2. Veracruz state production.

Sown fruit	Production percentage
Sweet orange	68%
Mexican lemon	21%
Persian limon	6%
Grapefruits, tangerines and tangerines	5%

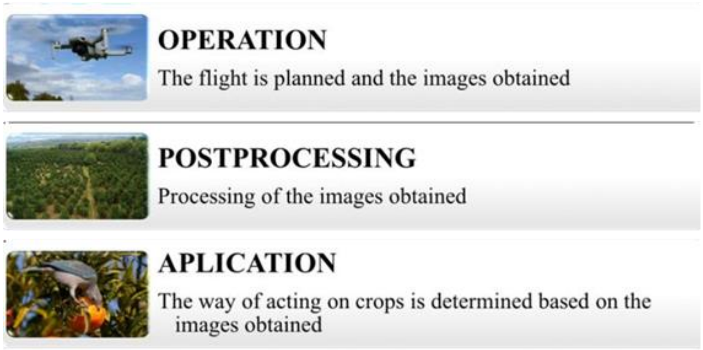

OPERATION
The flight is planned and the images obtained
POSTPROCESSING
Processing of the images obtained
APLICATION
The way of acting on crops is determined based on the images obtained

Fig. 3. Project methodology.

For the processing and obtaining of the images, as well as their processing, it was very important to follow the following methodology. (see Fig. 3).

2.2 Luminescence System in Response to Pest Attack

The use of light as a bird repulsion system is an optimal procedure for places where there are residents near the place where attacks of flocks of birds usually occur, due to the use of light rays directed in a specific direction and that do not affect the comfort of residents of the area.

Based on reports of bird encounters. The use of laser-based and light-based bird repulsion systems is known to have been implemented 30 years ago, however, it did not receive much attention and in the same way it did not receive much research in this regard.

To understand how a bird can behave before a reflex created by activating some kind of light, it is necessary to study the behavior in these situations in birds [9].

In addition, birds have a fourth type of pigment that allows them to be more sensitive to wavelengths (350nm-700nm) that are less than those of humans. This allows them to perceive a range of colors that we know as ultraviolet (UV) rays (see Fig. 4).

In a recent study conducted by researchers at Purdle University, the psychological-visual behavior of birds and their response to lights of different wavelengths was discussed. The method for the investigation was the use of multiple specimens of birds wavelengths (350nm-700nm) that are less than those of humans. This allows them to perceive a range of colors that we know as ultraviolet (UV) rays. (see Fig. 4), of the

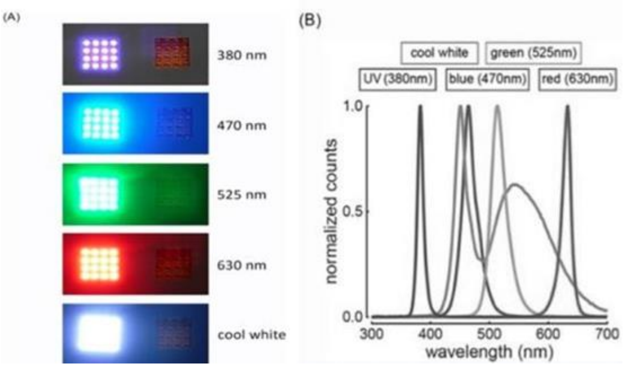


Fig. 4. Light treatments.

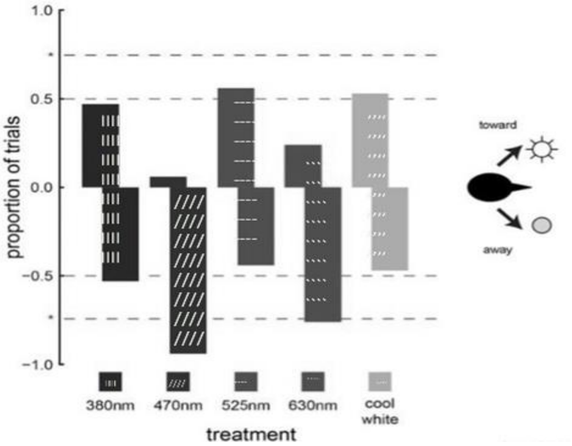


Fig. 5. Cowbird responses to LED lights.

species (Brown head cowbirds). The results obtained are that the birds have a psychological response to the lights of high chromatic contrast and are indifferent to the different wavelengths that they also compared. Avian collisions with man-made objects and vehicles (for example, buildings, automobiles, airplanes, power lines) have recently increased.

Lights have been proposed to alert the birds and minimize the chances of collisions, but it is a challenge to choose lights that are tuned to the avian eye and that can also lead to avoid them given the differences between human and avian vision.

A test of choice is proposed to address this problem by first identifying the wavelengths of light that over stimulate the retina using species-specific perceptive models and then evaluating the avoidance / attraction responses of the brown-headed cowbirds at these lights during the day using a behavior. The implementation of this mechanism has the purpose of provoking a reaction in flocks of birds of alert before an unknown event, it is planned to produce a light source modulated with a suitable wavelength so that the visual psychology of the birds can be used [10].

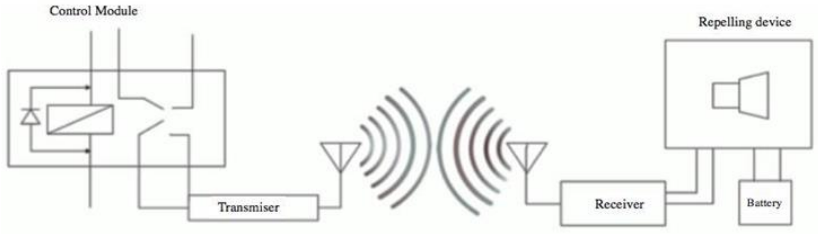


Fig. 6. Wireless remote control of a repelling device.

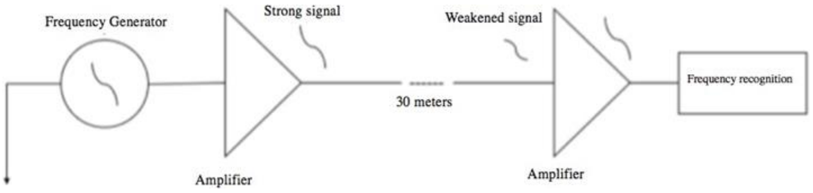


Fig. 7. Communication interface based on frequency signal transmission.

2.3 Methodology of Light

The requirements that the bird-based bird repulsion system must fulfill must be the following:

- Present a minimum weight for the drone.
- Own a better use of energy.
- Show greater flight autonomy.
- Have a practical size for the use that is going to be given, taking into account that the unmanned aircraft may be surrounded by different obstacles.
- The system must need a DC power supply so that it can be powered by the drone's battery.
- It must not harm the welfare of the local fauna or alter the order of it.
- The system must have variables in its method of repelling birds by light, to avoid the normalization of the birds before said system.

The type of light that is required to use must be based on the sensitivity of the species to be treated, it is directly related if the type of bird is nocturnal or diurnal. As the light mechanism will be used at night, you should be careful with the type of light source you choose. Since in the case of using a high consumption light source, there is a possibility that the drone does not have a good autonomy in relation to the energy consumption compared to the charge in the battery, see Fig. 5.

2.4 Sound Mode in a UAV for the Control of Pests in Citrus Harvest

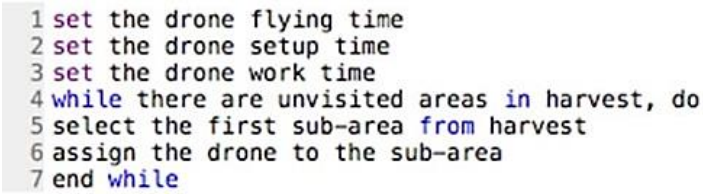
The system has as a reference the instinct of the bird to survive, fleeing from the imminent dangers that nature has and, for this occasion, simulates sounds to make them believe that they are in danger. The sounds are essential for the survival of birds because it is a way of communicating with other birds, among the aspects that influence the emission and capture of sounds are: reproduction, protection of the territory, alert and communication with other birds. [11] With this model to scare the birds is to generate a false danger to them, but without causing any kind of permanent damage. It is expected that this system will work both day and night, since as it is known this type of pests attacks during the day and night in search of food generating losses.

On the other hand, the main objective of the sound implementation is to considerably reduce the damage caused by the birds to the benefit of the citrus producers. This system can work in two different ways, which are presented below:

- Through ultrasonic sounds: they are those that do not belong to the audible spectrum for the human being and whose audible frequency is between 20Hz-20kHz, that is, they are transmitted at such a high frequency that it is not perceptible for humans, but for animals. [12] For this, it could be used between 1 and 5 KHz, since it is the average in which the audible spectrum of the birds is found. The main advantage of implementing it is that they are effective because they generate a disorientation in the birds that makes them go to another place where they consider it a safe place.
- And through audible spectrum: which consist inemitting audible sounds for both human beings and birds, being able to be more effective because they would emit familiar sounds that scare away birds, like the sounds of natural predators or recreations of sounds of dangerous situations for them.

A disadvantage exposed by the previous methods, is the adaptation process presented by the birds, that is, after a certain time they will be able to realize that the dangerous situation is a simulation, for this, a way to solve the deficiencies of these method is to modify the emitted sound, interspersing it between thunderous sounds, these can be sounds like that of a shotgun that has been shot, or simply a sound strong enough to make birds fear, also between sounds of natural predators, since it is possible to simulate sounds of animals that tend to attack a specific type of bird, in order to generate a state of alert in these. So, a speaker of small size, but with the ability to play high frequency audios can play the goal of repelling sound, because these are light and require more power to operate.

The model proposed to avoid sound is divided into 2 proposals, one through a wired controller and another through a wireless controller for the sound systems that will be implemented in the drone to cover an area of one hectare. For this, it is planned to model a drone with some of the sound systems mentioned above; In order to facilitate its use, it is planned to use those that are of audible spectrum. (see Fig. 6 and 7).



```
set the drone flying time
set the drone setup time
set the drone work time
while there are unvisited areas in harvest, do
select the first sub-area from harvest
assign the drone to the sub-area
end while
```

Fig. 8. Algorithm Operation.

It is proposed to assemble all the components inside a control card and also everything about the drone, trying to be the best so as not to cause any flight inconvenience, so that this works in the most effective way, a list of audios will be placed in a SD memory with the objective of recreating the stimuli that cause fear in each type of bird in specific.

It is possible to control many parameters of the drone so that it adapts to what is sought, from adjusting the speed of rotation to marking a maximum height, the connection between smartphone / tablet and the drone is made through WiFi. The drone creates a WiFi network to which you have to connect. So you get a greater range to take the drone farther or higher, or both. When talking about very long distances, in the tests, which have been carried out, it has reached more than 30 meters of distance and height of 10 meters, this due to the circumstances of the climate and the terrain, for this reason it has been impossible because the wind was blowing and more height, more wind. At the beginning of using it, there were certain response problems. The signal arrived very little, but in a couple of takeoffs and landings, the drone improved its flight response. In the cases of response problems, it should be modified that this, external mode was enabled.

While the use of drones in research and other practical applications is still taking its first steps, the initial testing of the drone has already proven its usefulness. The drones can be used to perform analyzes in large areas and difficult to access, in a relatively short time and with high precision. For farmers, the drone images will be useful to help detect potential losses in their crops in advance, giving them enough time to respond to bird flock attacks.

The processing of images and the analysis of the data, will provide an added value of the technology of the drones, it is very important the quality of the data obtained by the drones, and this has been great, the implementation of the drone was obtained with great precision in land uses, allowing the estimation of the area for each use.

Currently, see they are using several different techniques in imaging systems to improve the accuracy of camera pixels. For technology-oriented multispectral imaging systems, the red-green-blue (RGB) and near infrared (NIR) scanners were developed specifically for space-based scans and subsequently for aerial images. Applications with sequential traditional methodologies are restricted to domains where extended recording time, namely microscopy, remote sensing and biomedical imaging. Several approaches to instant multispectral images have recently been developed. Some of these snapshot procedures use fiber optics to reformat a two-dimensional image into a one-

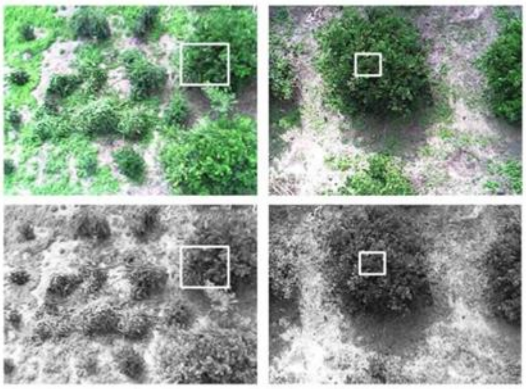


Fig. 9. Processing results.

dimensional matrix and then use a conventional one-dimensional image spectrometer to obtain spectral information.

The algorithm begins by generating the areas to be processed, from now on, at selected altitudes provided by the user. Then, it proceeds iteratively for each area from the lower to the upper altitude, dividing it to obtain simply connected subareas, in a number equal to that of the UAVs used. This objective is evaluated by a random closing search given the large number of possible closing connections.

With the increase in advanced technology, the multispectral imaging standard has been used today in various applications, for surveillance, security and defense purposes. Traditional spectral imaging methods involve sequential time scanning of a spatial or spectral dimension combined with instant images of the other two dimensions. The applications of these traditional time sequential techniques are restricted to arenas where an extended recording time is acceptable, as in microscopy, remote sensing and biomedical images.

2.5 DERT Algorithm

The DERT algorithm is based on the list programming approach, where a list represents an ordered sequence of values, in which the same value can occur more than once, for this, if the same value is repeated several times, each occurrence is considered a different element.

The interest in this approach is to explore strategies with less computational complexity to be later applied to prevent and combat bird pests in citrus crops through the use of drones, therefore, the DERT algorithm is basically composed of two main phases: prioritization phase, in which a priority rank assignment is established for each of the drones and the geographical sub-area assignment phase, where each drone is assigned to that geographic sub-area that optimizes a predefined cost function . The DERT algorithm is shown in Figure 5.

The algorithm contemplates that a drone can be assigned to several commissions, but can only do one at a time. A commission consists of leaving the initial base in an area assigned to perform a job. In this case, the work done by an unmanned aircraft in a particular area is to monitor.

3 Results

For the image data set, all noise was eliminated to maintain only the areas of interest For the image data set, all noise was eliminated to keep only the areas of interest necessary, that is, the background was removed to be able to visualize only the affected areas, that is the orange tree, in the same way for images of crops with pests, in addition to applying grayscale.

Finally, each image was homologous to the same size and, therefore, the correct processing was applied.

One of the most relevant aspects of this research is that if it detects the type of pest in time, farmers will benefit from the reduction of losses as well as fruit in the harvest season, as well as money and fruit copies.

After having processed the images, it is notable to appreciate in Figure 8, the presence of bird specimens in citrus crops, which leads to their classification, in order to perform a bird classification and how much damage each species generates. Without neglecting the classification of crops as to the type of citrus that is being affected.

4 Conclusions and Future Research

To conclude the present investigation, the increase of the productivity of the citrus fruits in the fruit fields represents an increase of the profits in the economic system, which will cause an improvement in the economy of the citriculture in the state of Veracruz since with the technology implemented and the use of censors and sounds, it will be possible to reduce pest attacks of birds, and at the same time, optimize the process of monitoring farmers, reducing time and labor.

Also, this research demonstrates how the implementation of drones with intelligent systems and mechanisms sensitive to different stimuli in the environment, means a great innovation in the productive sector and now, it is a reality to say that the agricultural sector will also innovate with the use of this technology, which will benefit not only productivity and stability in the agricultural sector, but will benefit the economy of farmers and agricultural workers.

To the extent that drones are capable of providing a useful service, there will be an evolution with its use in agriculture. For the service to be useful, it is necessary that the information provided by the drone was not previously known or was not detectable or that this information was much more expensive to collect, depending on the area or type of terrain in which it is implemented.

The autonomy of drones, often limited to a few minutes of flight, although on the agricultural sector are planning unmanned aircraft with autonomies that are close to the time, with adjusted acquisition costs. Drones are one of the tools with the greatest future prospects in fields such as the prevention and protection of natural resources and precision agriculture.

Currently, the use of drones has become increasingly common in the area of agriculture. These unmanned aerial vehicles are capable of providing accurate information on the production of crops; some models, even, allow to increase the efficiency in terms of fumigation, fertilization and pest control as we have already seen.

With the use of unmanned spacecraft, it is possible to fly over crops quickly and capture useful information for those who manage production, and with just one drone it is possible to monitor hectares accurately, which allows us to evaluate, in general terms, the ground conditions. This includes aspects such as hydration, temperature, rate of growth, premature disease localization, avian pest attack, etc.

On the other hand, drones are not new, since for some decades they have been implemented in military reconnaissance and search tasks, their implementation in the field is not recent, drones are part of the growing trend in precision agriculture and Although this technology is not new, it needs to be implemented in different sectors.

References

1. El Universal: Drones in charge of security in Puebla (2013)
2. Greenwood, F.D.: The challenge of comparing crop imagery over space and time. ICT Update, 82, pp. 2–4 (2016)
3. Allen, W.: Drones detect crop stresses more effectively. ICT Update, 82, pp.10–11 (2016)
4. Reynolds, E.: Drones for agriculture. ICT Update, 82, pp. 20–21 (2016)
5. Wang, Z., Lucas, A., Wong, K.C., Charmitoff, G.: Biomimetic drones to scare birds. In: 17th Australian International Aerospace CongressAt: Melbourne (2017)
6. Redacción NBSP: The Mexican drone that can detect pests in the fields (2016)
7. Samuelle, J.: Pigeons are a threat to agriculture (2014)
8. González, O.: Bird pests in the food industry. Manufacturing.net (2009)
9. Blackwel, B., Bernhardt, G., Cepek, J., Dolbeer, R.: Lasers as non- lethal avian repellents: potential applications in the airport environment. In: Federal Aviation Administration Technology Transfer Conference (2002)
10. Goller, B., Blackwell, B., De Vault, T., Baumhardt, P., Juricic, E.: Assessing bird avoidance of high-contrast lights using a choice test approach: Implications for reducing human-induced avian mortality. PeerJ, 6(10), pp. e5404 (2018)
11. Contreras, A., Garcia, J., Tejeda, A.: Birds as a pest, controls and management (2003)
12. Guerrero, F.: Controlled audio emitter system designed to scare away intrusive birds. The Professional School of Electronic Engineering Retrieved (2016)

ISSN 1870-4069

Mobile Application for Recognition of Arachnids' Bites: An Approach based on Ambient Intelligence

Ivette Mendoza[1], Eddy Sánchez[1], Alberto Ochoa[2], Alberto Hernández[3]

[1] Instituto Tecnológico Superior de Misantla,
Mexico

[2] Universidad Autónoma de Ciudad Juárez,
Mexico

[3] Universidad Autónoma del Estado de Morelos,
Mexico

{192t0032, esanchezd}@misantla.tecnm.mx,
alberto.ochoa@uacj.mx, jose_hernandez@uaem.mx

Abstract. In various regions of the world, the bite or sting of a spider is crucial, especially in children. This problem can produce in some cases, benign symptoms where they do not need medical treatment, others, usually present complications due to scratching and with this cause a superinfection in the lesion. These accidents require treatments to prevent death, amputations or permanent sequelae. Likewise, the impact of sequelae and deaths represent a high burden on health, society and economy. For this problem, currently, a recognition system has not been developed to help address this problem. On the other hand, the objective of this paper is that, with Ambient Intelligence, it is allowed through telemetry and knowledge management systems can be supported in a remote diagnosis that affects the improvement of the diagnosis made and its potential long-term use for a substantial improvement of the patient, generating recommendations, through a mobile application that is connected to a center of specialties, near the City of Misantla.

Keywords: Ambient intelligence, bite spider, arachnid poison, telemetry, remote diagnosis, intelligent system.

1 Introduction

In Mexico, there are approximately 50 families comprising 5,579 arachnid species, but of these only a dozen of the families Theriidae and Loxoscelidae, with the genera

Latrodectus and Loxosceles cause problems for humans [1]. There are 100 species of violinist spiders worldwide, and the most toxic are found on the American continent, especially in South America and are called Loxosceles reclusa, Laeta, Boneti. It is commonly known as a violinist, brown, brown spider, among others. Its habitat is dark temperate, moist and poorly ventilated places, such as sheds, wooden warehouses and warehouses [2]. Spiders are invertebrates of the arachnid family, such as ticks and scorpions. They are characterized by being carnivorous (even cannibals) and hunting characters, since they only eat live prey, that is, insects that hunt, since they do not eat the dead ones. All spiders have small poisonous glands, with more or less poison to kill their prey [3].

The spiders, in order to kill their prey, inject the venom through their chelicerae, two appendices with a powerful musculature and a sharp distal nail that nail the prey. The venom comes from a poisonous gland located in the cephalic zone, which drains directly to the chelicerae through ducts [4]. The spiders of the genus Loxosceles are characterized by being not very aggressive, since they only bite when provoked. They are nocturnal and hide in dark places where they wait for the prey to jump over, since they make little use of the spider web.

1.1 Spider Bite

Most bites by spiders are resolved without complications, when a timely intervention is made on them. However, in local loxosis by recluse spider, the necrosis of the skin can be serious and requires an exhaustive management of the wound, until surgical treatments and depending on the affected member. Therefore, it is impossible to know what kind of spider the aggressor has been, at the moment when it bites. As well as identifying the place to which they should turn, that is, a center of specialty where they attend this type of emergency.

Ambient intelligence is a multidisciplinary approach that aims to improve the way environments and people interact with each other. Make the places where they live and work are more beneficial for people. The main peculiarity of the technological development of digital environments is that they can produce a reinforcement and even an amplification effect of the cognitive abilities of people. In this sense, we can speak of the existence of a specific type of intelligence: ambient intelligence [5]. And, everything that requires a clinical and diagnostic model falls within the ambient intelligence. Since spider bites are analyzed with a medical database.

It is known that spiders respond to a wide variety of ambient conditions and can be indicators of plant associations and habitat disturbances [7]. Therefore, the objective of this paper is to develop a mobile application that recognize arachnid bites to optimize medical diagnostic processes in children and generate recommendations to channel the patient to a specialty center closest, in case of presenting this problem.

In addition, we need to reduce the risk of death in infants, which has been reflected in the center of Veracruz, Mexico, caused by arachnid.

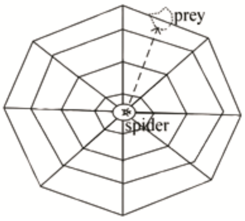


Fig. 1. Spider prey sketch.

Spiders have been a source of intrigue and fear for humans for centuries, and there are numerous myths about the medical effects of spiders. Many people believe that the bites of several species of spiders cause necrotic ulceration, despite the evidence that most suspected cases of necrotic arachnoidism are caused by something other than a spider bite. Latrodectism and loxoscelism are the most important clinical syndromes resulting from the spider bite. Latrodectism results from the bites of widow spiders (Latrodectus spp) and causes local, regional or generalized pain associated with non-specific symptoms and autonomic effects. Loxoscelism is caused by Loxosceles spp, and the cutaneous form manifests as pain and erythema that can develop into a necrotic ulcer.

Systemic loxoscelism is characterized by intravascular hemolysis and occasional renal failure. Antivenoms are an important treatment for spider poisoning, but they have been less successful than those for snake poisoning, with concerns about their effectiveness for both latrodectism and loxoscelism [8]. The bite of an arachnid can have great complications, such as necrosis. This is the degeneration of a tissue by the death of its cells. This mortality is produced by the action of a harmful agent that generates an irreparable injury. On the other hand, we find the one that is produced as a result of a series of changes in the cell that is the one that, by itself, decides that it has to die. This is a main element in the investigation. Therefore, it is essential that when a person presents this problem of arachnid bite, it is urgently addressed since otherwise it could result in deat in it, if not treated immediately.

1.1 Spider Web Definitions

Definitions of the spider web. In nature [9], the spider is based on the great network for the life of its prey. In most cases, the cobweb structure is approximated as a potential center of symmetry (see Fig. 1).

The cobweb, consists mainly of three parts, namely the center point, the radius and the spiral silk thread. The radiation that comes out of the center of the silk is called convergence (ray), used as a spider's web, the silk spiral thread of the ray is called chord (spiral). Through the web, spiders can feel the vibration to determine if there are prey and their net position. Each ray of the web will transmit the vibrations to the center of the web, so the spiders will usually remain in the center waiting for information.

Differences between dangerous and non-dangerous spider webs:

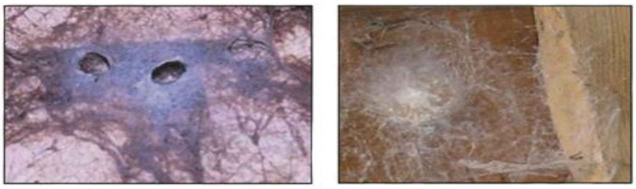


Fig. 2. Loxosceles fabrics. Irregular, cottony, compact, with apparent disorganization.

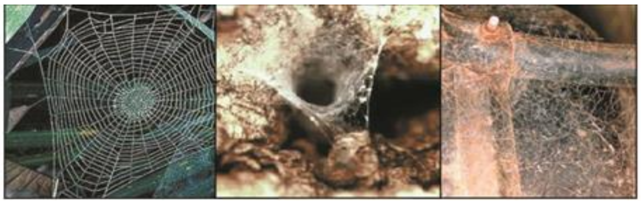


Fig. 3. In general, they are regular, organized, symmetrical and demarcated.

a) Dangerous spider webs, Fig. 2.
b) Non-dangerous spider webs, Fig. 3.

Spider Morphology

As well as insects, spiders and other arachnids have two main parts in their body. The anterior part of your body is the cephalothorax or prosoma. The dorsal part of the cephalothorax is called the carapace, the ventral part is known as the sternum. The back of the body is called abdomen or opisthosoma. The body is divided into 2 regions or labels: the prosoma (previous) and the opisthosoma (posterior to), joined by a structure called pedicel, which can be a marked constriction in some arachnids (union caulogaster, for example, spiders and amblipigids), or a wide union of both labels (hologaster union, for example, scorpions and opiliones).

The prosoma has 6 pairs of appendages: chelicerae (2-3 segments), pedipalps (maximum 6 segments) and 4 pairs of ambulatory legs (7 segments that can be subdivided or merged depending on order). The dorsal part of the prosoma is usually covered by a shield or carapace, which lacks antennas and where the ocelli are found, which can be from scratch in the case of troglobian species or that inhabit caves, 2, 4, 6 and even 8 ocelos, always in even numbers. The opisthosoma may or may not be segmented and generally lacks segmented appendices (or are highly modified); In addition, in this region are the reproductive and respiratory structures [10].

2 Analysis of Related Works

Experts from the Arachnology Laboratory of the Center for Parasitological and Vector Studies [11], created a digital system that allows users to take pictures with their mobile devices and identify the different species of arachnids and scorpions, and in turn, identify arachnid species that pose some kind of risk to human health as well as harmless ones. Sergio Rodriguez Gil, one of the creators of the application that goes by

Table 1. Degree of poisoning from arachnid bite.

Degree of intoxication	Clinical picture
Grade I / mild	Pain in injury site, in lumbosacral region, abdomen. Asthenia, adynamia, diaphoresis, sialorrhea, hyperreflexia.
Grade II / moderate	Accentuated dyspnea, epiphora, headache, spasm, contracture or muscle rigidity, priapism.
Grade III / severe	Accentuated miosis, mydriasis, trismus, heart rhythm disorders, bronchospasm.

name: Is it a spider or a scorpion? This is a tool created in Argentina, which serves the community so that, on the one hand, it has enough information about these animals and can differentiate which are dangerous and, on the other, so that it is possible to report their presence more quickly.

This innovation is now available to download in different digital stores for devices with Android operating system. In addition, from the immediate service received by the community, the data collected by the application is configured as a great information base for future scientific work: "What users send allows mapping the zones of appearance and seasonal behaviors, and even giving a notion about the reactions generated by the presence of these species, the erroneous or accurate preconceptions that circulate about them in society" [12]. Its operation is very simple. The first thing the user must answer is whether the animal in question stung it or not. If the answer is affirmative, a notification arrives as a matter of urgency and the first recommendation is that you approach the closest attention center so as not to delay the attention.

Another spider identification assistant made for mobile devices, is called "Spidentify", created in Australia. This application was created by Minibeast Wildlife who reveals the mysteries of one of the most feared animal groups in Australia and places a lot of detailed information on 250 arachnid species. This tool identifies detailed information about each spider, including an instantly accessible bite danger classification. Useful pop-up windows explain technical words in a simple language. Users in this application can browse the field guide by category, including habitats and families. Search the place to show any species in the application, search terms in the glossary or find closely related species. The complementary content explains the spider's anatomy, dispels myths, answers common questions and more [13].

3 Recognition of Arachnid Bites

3.1 Epidemiology, Diagnosis and Prevention

The epidemiology of a spider bite depends on the interaction between the spider and humans, the ecology of the spider and the ambient. The distribution of spiders of medical importance is the most important factor in the identification of where clinically important arachnidism occurs throughout the world and is analyzed for each of the spider groups [14].

The diagnosis of spider bite [15], is usually clinical, and the defined bites must be based on a clear history of a spider that bites the person and is then identified.

3.2 Treatment

Antivenoms are an important therapeutic intervention for poisoning syndromes, and antivenoms exist for many groups of spiders [5]. However, antivenoms have been less successful in the treatment of arachnidism than those of snake or scorpion poisoning. The use of antivenom is based on clinical experience, which has led to discrepancies in the proportion of patients treated. For example, in Brazil, antivenom is rarely used to treat Phoneutria poisoning despite substantial and distressing effects, but it is widely used to treat Loxosceles poisoning, although in theory it is unlikely to be effective [6].

Although Loxosceles antivenom is used in South America and is an effective invitator, the delay in clinical presentation and the irreversible nature of skin necrosis mean that this laboratory finding does not translate into effective treatment [16]. In contrast, antivenom for tissue poisoning in the form of a funnel fabric is highly effective and can save lives because early and inverse neurotoxic effects can be administered [5].

3.3 Methodological Proposal for the Recognition of an Arachnid Bite

According to the tests, the use of a mobile application is feasible, to identify the bite of a spider and to know which was the aggressor, using photographs in real time for the cases. This investigation tries to recommend the patient with said problem, to the place where it should be resorted in case of bite of a poisonous spider, to avoid the amputations or even the death of said spider in the case of some child.

Since, it is impossible to know what kind of spider the aggressor has been, when he bites and the idea of not knowing what species has been attacked in a person, is one of the main problems for this work. That is why with AmI, a series of tools to solve this problem arise, which in this case are technology, information, image processing, among others.

3.4 Design of the Mobile Application

The georeferencing of the location to identify the place that should be attended to avoid a tragedy, determining the correct center of specialties to treat the problem based on the arachnid bite, using a mobile application, which makes possible a decisive aspect where it is recommended catalyze the patient to the right place, at the right time.

Using a georeferencing model, it is possible to determine the precise location of the different places where medical care is carried out, so that, through the telemetry and knowledge management systems, a remote diagnosis for babies is admitted. they suffer the bite of a spider and that affects the improvement of the diagnosis made, as well as its potential long-term use for a substantial improvement of the patient (see Fig. 5).

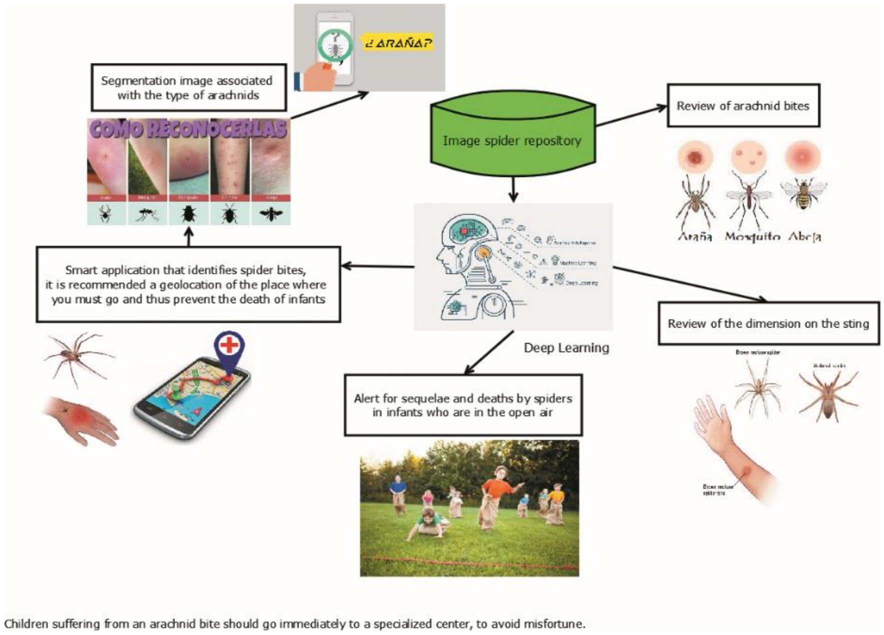

¿ARAÑA?
Segmentation image associated with the type of arachnids
COMO RECONOCERLAS
Image spider repository
Review of arachnid bites
Araña
Mosquito
Abeja
Smart application that identifies spider bites, it is recommended a geolocation of the place where you must go and thus prevent the death of infants
Deep Learning
Review of the dimension on the sting
Alert for sequelae and deaths by spiders in infants who are in the open air
Children suffering from an arachnid bite should go immediately to a specialized center, to avoid misfortune.

Fig. 4. Proposal of the model, associated with the investigation**.**

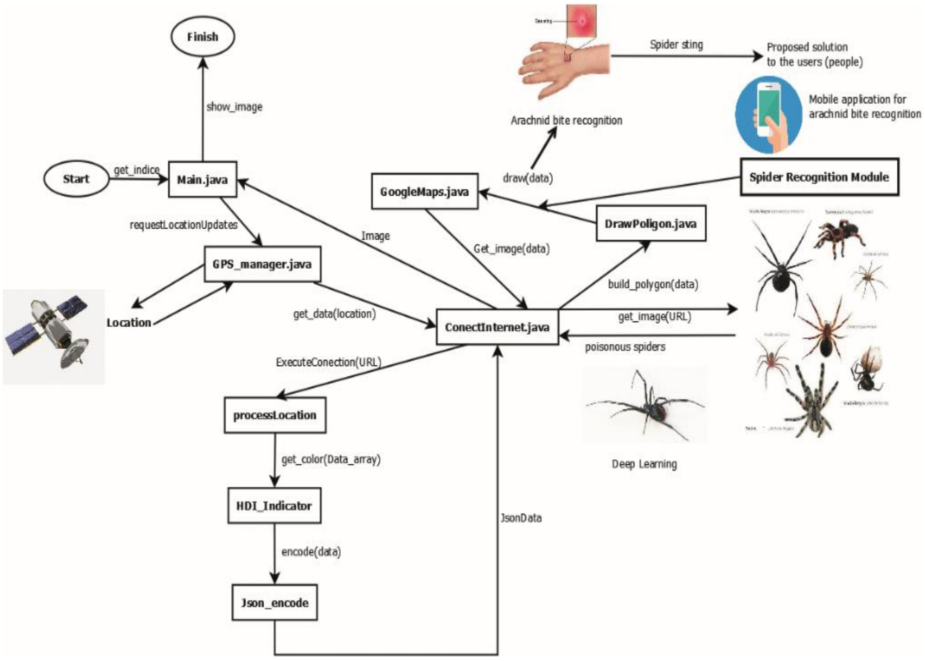


Fig. 5. Modular description of the components for the development of the mobile application.

Fig. 6. Identify the place to turn to in case of arachnid bite.

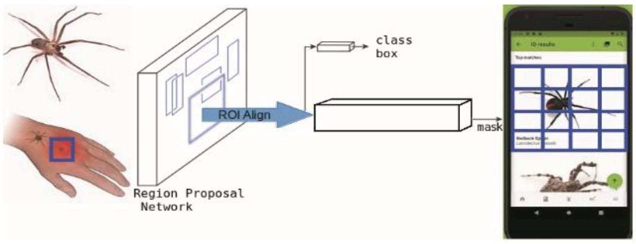


Fig. 7. Simplified diagram of the R-CNN mask network used for classification / segmentation of images. In the diagram, the network generates three types of output, the box where the arachnid is located, a binary mask that delimits the arachnid and the type (genre) of arachnids found.

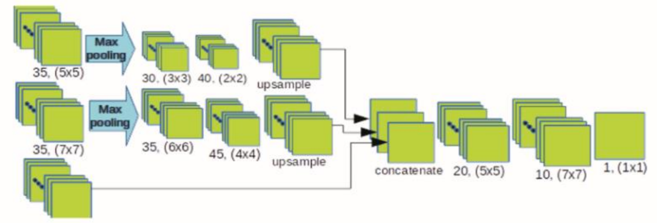


Fig. 8. The proposed architecture for the mask module. It consists of three channels of convolutional layers that are concatenated to have a good feature map with information at different scales.

3.5 Considerations of the Problems and their Impact on Society

According to [10] in the world there are about 5 million accidents per year of poisoned animals, of which between 50 and 75% require treatment to prevent death, amputations or permanent sequelae. The impact of the sequelae and deaths represent a high burden on health, social and economic.

Most of these problems caused by an arachnid bite are:

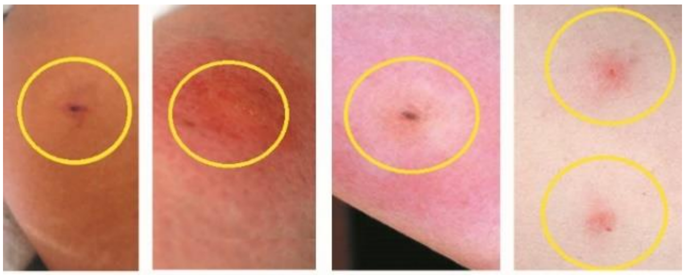


Fig. 8. Sample images of the input data. As can be seen, the bites can be similar and it is difficult to identify the type of arachnid that has been the aggressor.

- They produce similar benign symptoms.
- Some do not need medical treatment.
- Others, usually present complications, due to scratching, since there may be superinfection in the lesion.
- It is also possible to trigger serious reactions that can even compromise the life or amputation of a limb.
- It is impossible to know what type of spider the aggressor has been, at the time it bites.
- Identification of the place to be used, a center of specialties where they attend this type of emergency.

3.6 Spider Recognition Module

This module for the mobile application is shown in Fig. 7, which has an exchange with the image recognition of spiders and the rest of the data. To recognize the arachnid bite and identify it, this work requires a dataset that contains images of different types of arachnid species, as well as the different types of spider webs. The arachnid identification stage consists of a deep learning architecture. An input image of the sting and, subsequently, it is segmented into several groups which are then classified as the recognition of the arachnid type. The architecture is based on the regional convolutional mask neural network (RCNN) that extends RCNN by adding an object mask to the existing branch for recognition of the bounding box.

Mask R-CNN is an architecture of deep learning to solve the segmentation of instances, it uses the module for the classification of R-CNN fast, but with an additional module to create masks. The R-CNN mask adopts an identical first stage of the region proposal network, however, it adds a binary mask for each stage to produce a binary mask from the input image. In this architecture, each module of the network has its own assigned loss, which allows the network to generate masks for each class without competition between classes. In Fig. 7, the architecture of the network is shown.

In this paper, a new architecture was proposed for the mask module in the R-CNN mask, since the network will only focus on the recognition of arachnids, simplifying the module so that it is easier to train and have less computational load.

The proposed architecture consists of three convolutional channels, where each channel aims to select different feature sizes: a large channel with 11x11 size cores, a medium size channel with 4x4 size output filter, and finally, a small size channel with 2x2 output filters. The outputs of the different channels are concatenated in a feature map that contains information at different scales of the input image, keep in mind that, for concatenation, the output of the medium and small channels is sampled so that the output of all three channels have the same size. Subsequently, the output is fed to a pair of convolution layers to recover the size of the original image and, finally, a convolution layer of filter size 1x1 to have a level image of gray with the masks of the spider regions. Each mask in the output image has a gray level that indicates the class of the region enclosing the mask [17]. And, the details of the described architecture (see Fig. 8).

In Fig. 8, several images used to train the network are shown. To increase the transmission speed to the server and reduce the computational cost, the input images were reduced to a fixed size of 128x128 pixels. This initial data set was processed with data augmentation techniques [18], to have a final data set of images; for this purpose, each image of the initial data set was processed with translations, rotations and scale changes. In addition, each image was manually segmented into the background and the different types of arachnids. The network became a medium error loss function, the implementation was programmed using the keras framework [19], on a computer with an Intel Core i5-4210U 2.40 GHz processor and a Nvidia GeForce GTX 840M graphics card (8GB of RAM).

In addition to this, it is convenient to indicate that the image recognition module is still being improved for its correct implementation in the cell phone, but due to the compact form of the architecture of this research, it is considered that it is very viable that it can be implemented. For any type of mobile with Android operating system.

To know the operation of the proposed model, the estimation of the bites by various types of arachnids is used, after that, the comparisons between the bites to determine by means of the type of spider web what the aggressor really was and based on this, recommend the patient a nearby specialty center.

These images were taken from a database, which has scientific names that are part of the following catalog of taxonomic authorities: CONABIO (Comp.). 2012. Catalog of taxonomic authorities of arachnids (Arachnida: Arthropoda) of Mexico. SNIB-CONABIO database. Mexico.

That includes information of the BK006 project. And, they were evaluated by the intelligent application through their representatives in the deep learning model. For the estimation of the matrix of the selection of identification of one type of arachnid, 30 executions of the experiment were carried out under the same conditions, and, for the group of finalist proposals, a design of experiments according to the attributes of each proposal to obtain a better estimate of the final classification.

An atmosphere capable of storing the data of each one of the stings in the people was developed, including the type of arachnid species that represents the sting, this, with the purpose of evaluating when taking an image from the application in real time and saying which it has been the arachnid bitten on the person. One of the most important characteristics observed in this work was the identification of the spider bite, not being able to distinguish which has been of the different existing species in the

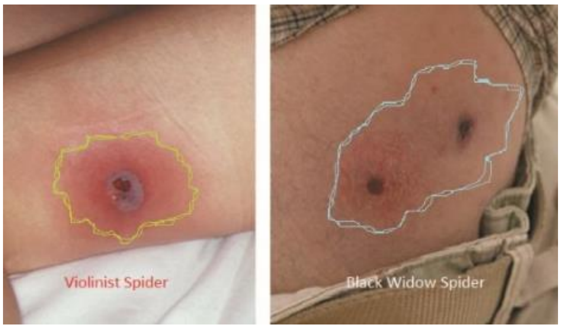


Fig. 9. Results of the proposed network segmentation.

database, since some spider bites are similar to another type of species and that's where the confusion exists.

Structured scenes with agents cannot be reproduced in general, and only a moment is represented in the space and time of different societies. This is a unique and innovative form of adaptive behavior that solves a computational problem that does not try to group societies only with a factor associated with their external society, the treatment of a computational problem that involves a complex change between existing relationships. they can relate metaphorically to the knowledge of community behavior with respect to an optimization problem to culturally select similar societies and their arachnids, without being the same quadrant associated with similar ones.

The main experiment was to detail each one of the 5 different types of arachnids and their bites, this allowed to generate the best selection of each quadrant and its possible location in a Diorama associated with the arachnids, which was obtained after comparing the different similarities between the stings that you are leaving after attacking, and evaluate with the multiple coincidence model of each of them. The tool developed classified each of the spiders belonging to each quadrant. The design of the experiment consists of an orthogonal matrix test, with the interactions between the variables: emotional control, fighting ability, intelligence, agility, strength, resistance, social leadership and speed. These variables are studied in a range of emotions associated with the arachnid; another variable is the color of the spiders represented by numbers (0 to 256).

Metrics for the evaluation of segmentation are as follows.To measure network performance, we used the precision for the recognition module in R-CNN and for the segmentation mask stage using the Union Intersection (IoU), which is a common metric [20], used to evaluate the accuracy of the segmentation. This metric calculates the total number of pixels at the intersection between the set of predicted pixels (A_{pred}) and truth pixels in the ground (A_{GT}) for each class, and is divided by the number of pixels in its joint, as shown in equation 1:

$$IoU \frac{A_{pred} \cap A_{GT}}{A_{pred} \cup A_{GT}}. \tag{1}$$

3.7 Fragment of Code

Advancing and traversing the bounding boxes and the names tagged for each arachnid and drawing them in the output image for viewing purposes. Example:

Algorithm 1. Example of a program when performing image classification, object detection, semantic segmentation and instance segmentation. Performing the segmentation of instances with Mask R-CNN.

```
1 # loop over the recognized spider
2 for ((top, right, bottom, left), name) in zip (boxes, names):
3           # draw the predicted spider name on the image
4           cv2.rectangle(image, (left, top), (right, bottom), (0, 255, 0), 2
5           y = top - 15 if top - 15 > 15 else top + 15
6           cv2.putText(image, name, (left, y), cv2.FONT_HERSHEY_SIMPLEX,
7 0.75, (0, 255, 0), 2)
8 # show the output image
9 cv2.imshow("Image", image)
  cv2.waitKey(0)
```

In this code fragment, the for cycle begins to traverse the bounding boxes of the different types of detected arachnids and the predicted names. To create an iterable object that can easily be traversed values, we call zip (tables, names), which results in tuples from which you can extract the coordinates and the name of the box.

The coordinates of the box are used to draw a green rectangle in the line of code 4. The coordinates are also used to calculate where the text should be drawn for the name of the arachnid (line 5), followed by placing the text of the name in the image (lines 6). If the spider bounding box is at the top of the image, we must move the text below the top of the box (handled on line 5), otherwise the text would be cut. Finally, proceed to show the image until a key is pressed (lines 8 and 9).

4 Results

For the image data set, all noise was eliminated to maintain only the areas of interest necessary, that is, the background was removed to only appreciate either the spider, its spider web and the spider bite. Finally, each image was homologous to the same size and, therefore, correct processing was applied.

One of the most relevant aspects of this research is that if it detects the type of arachnid in time, the patient can be cataloged to a specialty center where it is attended to avoid death. And taking into account that the death of infants can be reduced, this proposed mobile application for Android operating systems arises. The output of the segmentation on the arachnid bites and the combination with the sorting output is shown (see Fig. 10).

The prediction of future events is a difficult task to perform, since it requires extensive multivariate analysis, and it is also impossible to do it in several subjects [20]. There are several methods that have been used as an auxiliary tool for the construction of estimation models. In this case, in the review of the literature, it has been detected that there are not enough antecedents in the area. In this work, the use of ubiquitous calculus, image processing and deep inclinations combine to predict behavior in an evaluation of the bite by a child arachnid from ambient intelligence.

The approach proposes a model that includes two main characteristics: the identification of the arachnid bite and to recognize the type of arachnids by their characteristics, in this case identify which, without margin of error by means of its web. The model incorporates basic information on the various types of arachnids from an arachnid repository, which was obtained from CONABIO.

5 Conclusions and Future Research

In Mexico, currently, they have not done a job like that is presented above. In addition, the great contribution of this article are the techniques of image pre-processing and machine learning for the case study and incorporating the three data sets into one.

The animals are characterized by possessing venom glands and specialized structures for inoculant tales. In Mexico, there is a large number of poisonous species and it is located in two biogeographical zones. When having an accident with any of these animals, you should go to the doctor, to avoid serious consequences or death. According to the experiment, it is confirmed that it is possible to implement a technological platform for the recognition of arachnid bites in real time, using a mobile device that captures the part of the bite. This application is able to identify which was the arachnid attacked in a person and based on it, act immediately avoiding a misfortune in the affected party.

It is recommended that the intelligent tool be adapted to the people who are with this problem and thus be catalyzed in an efficient way as far as the displacement, to a nearby center of specialties. It is expected that the mobile application will be developed not only for Android operating systems, but also for iOS devices and, be cross-platform.

References

1. Mexican Biodiversity: How many species are there 2018 (2019)
2. Francke, O.F.: Biodiversity arthropoda (Chelicerata: Arachnida ex Acari) en Mexico. Revista Mexicana de Biodiversidad, 85, pp. 408–418 (2014)
3. Rastogi, S., Liberles, D.A.: Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evolutionary Biology, 5 (2005)
4. Harris, T.: Spider's venom how things work (2019)
5. Isbister, G.K., Page, C.B., Buckley, N.A., Fatovich, D.M., Pascu, O., MacDonald, S.P.J., Calver, L.A., Brown, S.G.A.: Randomized controlled trial of intravenous antivenom versus placebo for latrodectism: The second redback antivenom evaluation (RAVE-II) study. In: Annals of Emergency Medicine, 64(6), pp. 620–628 (2014)
6. Isbister, G.K., Brown, S.G.A., Miller, M., Tankel, A., Macdonald, E., Stokes, B., Ellis, R., Nagree, Y., Wilkes, G.J., James, R., Short, A., Holdgate, A.: A randomised controlled trial of intramuscular vs. intravenous antivenom for latrodectism - The RAVE study. In: QJM:

An International Journal of Medicine, 101(7), pp. 557–565 (2008)

7. Gutiérrez, L., Jiménez-Jiménez, M.L.: Spiders of wetlands southern Baja California (2004)
8. Álvarez, F.H.B., Cervantes, M.M., Fernández, A.A.: Latrodectism in a pediatric patient. Mediciego, 25(1), pp. 72–78 (2019)
9. Chen, H., Chau, M., Zeng, D.: CI spider: a tool for competitive intelligence on the Web. Decision Support Systems, 34(1), pp. 1–17 (2002)
10. Cushing, P.E.: Spiders (Arachnida: Araneae). Encyclopedia of Entomology (2019)
11. CEPAVE: Launch of digital application to identify spiders and scorpions. National Council of Scientific and Technical Research (CONICET – UNLP) (2017)
12. Guisade: Launch of digital application to identify spiders and scorpions. National Council of Scientific and Technical Research (CONICET) (2017)
13. Wegner, G.S.: Spider identification made simple. The Chemical Company (2019)
14. Isbister, G.K.: Necrotic arachnidism: The mythology of a modern plague. The Lancet, 364, pp. 549–553 (2004)
15. Stoecker, G.J.: Diagnosis of loxoscelism in a child confirmed with an enzyme-linked immunosorbent assay and noninvasive tissue sampling. Journal of the American Academy of Dermatology, pp. 55, pp. 888–90 (2006)
16. Pauli, I.M.J.: Analysis of therapeutic benefits of antivenin at different time intervals after experimental envenomation in rabbits by venom of the brown spider (Loxosceles intermedia). Toxicon, 53, pp. 660–671 (2009)
17. Rosebrock, A.: Mask R-CNN with OpenCV. Deep Learning, Semantic Segmentation (2018)
18. Ronneberger, O., Fischer, P., Brox, T.: U-Net: redes convolucionales para la segmentación de imágenes biomédicas. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI) (2015)
19. Papandreou, g., Chen, L.Ch., Murphy, K., Yuille, A.L.: Weakly- and semi-supervised learning of a dcnn for semantic image segmentation. In: Computer Vision and Pattern Recognition (2015)
20. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Computer Vision and Pattern Recognition, pp. 2980–2988 (2017)

ISSN 1870-4069

Recent Advances in Biomedical Image Segmentation Using Neural Networks

Cecilia Irene Loeza Mejía[1], Balzhoyt Roldán Ortega[1], R.R. Biswal[2],
Gregorio Fernández Lambert[1], D. Reyes González[1]

[1] Tecnológico Nacional de México,
Departamento de Posgrado e Investigación,
Mexico

[2] Escuela de Ingeniería y Ciencias,
Tecnológico de Monterrey, Zapopan, Jalisco,
Mexico

rroshanb@tec.mx

Abstract. In recent years, the analysis and processing of biomedical images had considerable relevance, as it has proven to be an effective way of obtaining information regarding human body in a less invasive way and thus helps in extracting the characteristics that could potentially represent a disease. Different aspects of segmentation algorithms and features have been largely studied in the last decade relating to various areas. However, there is not a single method or solution because of the variation in the property of images, medical imaging techniques and modalities, variability and noise for each object of interest. This work presents a comparison of different methods including deep learning for segmentation of multimodal biomedical images. In addition, the application of U-Net architecture for the lung region segmentation of chest computed tomography in the Lung TIME dataset was evaluated.

Keywords: Segmentation, biomedical image, neural networks.

1 Introduction

Biomedical images are especially important and have contributed to the progress of medicine, allowing us to see the human body in a less invasive way, helping to provide an efficient diagnosis. The information regarding internal structures, molecular composition and interaction of human body is crucial to evaluate the changes that occur over time for personalized treatment.

There are different kinds of physical processes that allow biomedical images to be generated [1]. These processes are mainly based on (i) ultrasound backscattering (ii) X-ray transmission: radiography, computed tomography (CT), (iii) Gamma ray emission from radioisotopes: Positron emission tomography (PET) and (iv) Spin precession in

pp. 73–82; rec. 2019-08-13; acc. 2019-09-14

magnetic fields: magnetic resonance imaging (MRI). In addition, there are multiple imaging modalities which can be (i) two dimensional (M x M pixels) or (ii) three-dimensional (M x M x M voxels).

Medical image processing allows the analysis of a large amount of information quickly and efficiently, with the aim of searching for certain characteristics and identifying a certain group of diagnoses. Segmentation is often the first stage in pattern recognition systems [2] and it is an exceedingly difficult problem in any image [3]. Segmentation is an important tool in medical image processing because it directly affects the accuracy and reliability of the diagnosis results [4].

Moreover, it is used for feature extraction, quantification of measurements, and allows an image to be transformed to a more meaningful form, as it obtains region of interest (ROI) with similar features in an orderly manner and groups them into a class [6]. ROI possesses a group of pixels defined in different forms such as circle, ellipse, polygon or irregular shapes [7]. Segmentation is used, for example: for the detection of organs and to distinguish pathological tissue from normal tissue [2, 8]. In this article, segmentation in the spinal cord, spine, prostate, breast, lung, arteries and brain was reviewed. In addition, U-Net has been implemented for segmentation of the pulmonary region.

This paper is organized as follows: in the second section a review of segmentation techniques based on neural networks in the multimodal biomedical imaging (obtained from computed tomography scan, magnetic resonance imaging, ultrasound or collecting images over time [6]) is presented. The third section shows the materials and methods that were used in segmentation of chest computed tomography. The fourth section shows the results obtained in the segmentation of the pulmonary region and the comparison with the results of different neural networks and tools for segmentation. While, the last section a summary of segmentation in medical image processing and areas of improvement in neural networks.

2 Background

2.1 Techniques before Segmentation

Data augmentation. It allows modifying samples of the dataset to create more training images, which reduces the possibility of over-fitting [11] and improves the generalization of machine learning models. Zhuang et al. [4] used shift (vertical and horizontal), shear transformation and flipping about the horizontal plane on the training dataset. In contrast, Almajalid et al. [5] used rotation (±90◦ in each time) and elastic deformations [26].

Dabiri et al. [11] used random rotation, horizontal and vertical flip transformations. Instead Gros et al. [16] included shifting (±10 voxels in each direction), flipping, rotation (±20◦ in each direction), and elastic deformations [26]. In contrast Van et al. [25] applied random amounts of translation, rotation, scaling, shearing, elastic warping, cropping, blurring, contrast alterations, noise addition, and mirroring in the anterior-posterior axis.

Instead Li et al. [17] applied rotation, horizontal flipping, zoom, height/width shift. Moreover Rat et al. [18] extracted randomly displaced image patches up to 3 voxels along each dimension of the image domain.

Pre-processing. At this stage, generally, the image noise is eliminated, and the contrast is improved. In biomedical images, morphological filters can be applied. To preprocess computed tomography (CT) images, the authors use morphological image processing, region-based image processing, and contrast adjustment [9], instead Dabiri et al. [11] converted DICOM images to grayscale PNG images and standardized their pixel value lie in the range of [0,1]. On the other hand, to preprocess magnetic resonance imaging (MRI) Ilhan et al. [20] applied morphological operations and pixel subtraction operations, in contrast, Gros et al. [16] used CNN to detect and crop parts of the image. Moreover, Lu et al. used [31] classical z-score normalization algorithm, Sheela et al. [32] applied median filter. To preprocess Yang et al. [33] applied N4 bias field correction algorithm and intensity normalization. Instead, to preprocess ultrasound images, Almajalid et al. [5] used Speckle Reducing Anisotropic Diffusion and histogram equalization; in contrast Nithya et al. [21] used the median filter.

2.2 Segmentation

The segmentation of biological images typically consists of partitioning an image into multiple regions of interest representing anatomical objects, for an efficient analysis and visualization. The different technical challenges and difficulties that are encountered during image segmentation procedures include, but not limited to, heterogeneous pixel intensities, noisy boundaries, and non-regular shapes with high variability.

There are several approaches to segmentation using machine learning, which can be supervised and unsupervised. The segmentation techniques are region-based, edge-based and multispectral or multimodal, which is based on integration of information from several images.

In addition, there are hybrid approaches, which result from the combination of individual procedures [3]. The Table 1 shows the neural networks techniques used in multimodal biomedical images.

The type of technique used may varies according to the type of biomedical image. Vania et al. [9] used in CT CNN and FCN, instead Liu et al. [10] used CDP-ResNet, alternately Dabiri et al. [11] Liu et al. [27] and Huang et al. [28] used DNN, instead Lessman et al. [14] used FCN.

On the other hand in MRI images, Lessman et al. [14] used FCN, instead Bui et al. [15] implement 3D-SkipDenseSeg, instead Gros et al. [16] applied CNN, by contrast Li et al. [17] used MMAN, Rak et al. [18] implement a combination of CNN and graph cuts and Tong et al. [30] proposed RIANet.

Moreover, in ultrasound images, Zhuang et al. [4] implement RDAU-NET, Almajalid et al. [5] used U-Net, Karimi et al. [22] and Xu et al. [23] applied CNN, instead Yang et al. [29] DPU-Net.

Table 1. Segmentation strategies in Biomedical Images.

Abbreviation	Technique	Aspect of interest
CNN	Convolutional Neural Network	It requires a minimum preprocessing and extract features from image pixel [5].
CDP-ResNet	Cascaded Dual-Pathway Residual Network	It extracts features by cascading two dual-pathway residual networks [10].
DNN	Deep Neural Network	It is more accurate when is trained with augmented data [27].
DPU-Net	Dual Path U-Net	It has more convolutional layers and lower depth for each convolutional layer [29].
FCN	Fully Convolutional Network	It can multitask concurrently, and It can be used in complex segmentation tasks [14].
MMAN	Multimodality Aggregation Network	It provides good performance in brain segmentation [17].
RDAU-NET	Residual-Dilated-Attention-Gate-UNet	By increasing the layers, the network will have greater learning capacity [4].
RIANet	Recurrent Interleaved Attention Network	It has achieved competitive segmentation results with fewer parameters [30].
U-Net	Convolutional Networks for Biomedical Image Segmentation	It is a novel framework of CNN designed for precise and fast segmentation of images [13].
3D-SkipDenseSeg	3D Fully Convolutional, Skip-Connected DenseNet	It has highly accurate results particularly in applications with data-sparse [15].

3 Materials and Methods

3.1 Dataset Lung TIME

The computed tomography scans that were used correspond to the publicly available Lung TIME [24] dataset, which contains two subsets: TIME1 and TIME2. TIME1 includes 148 CT of adolescent patients and TIME2 contains 9 CT of adult patients from the Faculty Hospital in Motol, Prague, Czech Republic.

Each CT contains several images (slices) with a resolution of 512x512 pixels. In this work, 2003 slices were used and for each slice, binarization was performed for a constant threshold of -350 Hounsfield Units, proposed by Pulagam et al. [35] to separate the pulmonary region from CT. As preprocessing the images were rescaled in different resolutions: 32x32, 64x64 and 128x128 pixels. **¡Error! No se encuentra el origen de la referencia.** shows an example of a slice and its binarization in different

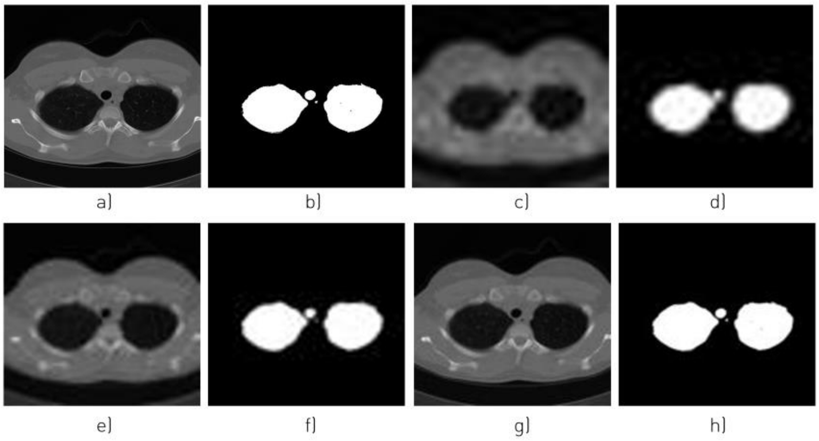


Fig. 1. Binarization and CT slice scaling a) original slice 512x512, b) binarized 512x512, c) slice 32x32, d) binarized 64x64, e) slice 64x64, f) slice 128x128, g) binarized 128x128.

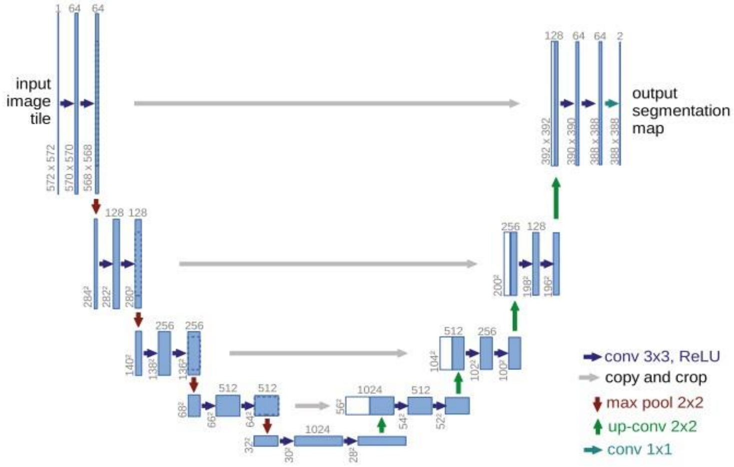


Fig. 2. The structure of U-Net [13].

resolutions of the image. Each of the 2003 slices and their respective binary image were used to train and evaluate the U-Net network.

3.2 U-Net

U-net is a convolutional neural network that allows to capture context and to achieve precise localization for precise and fast segmentation of images [13]. In this work, we applied U-Net using the Adam optimizer with a learning rate of 0.0001 and the binary cross-entropy loss function.

Table 2. Results of lung region segmentation using U-Net.

Input image	Epochs	Train ACC	Test ACC	Time (hours)
32x32	5	60.71%	60.14%	0.35
64x64	5	74.94%	74.28%	1.32
128x128	5	74.96%	74.26%	4.11
32x32	10	60.72%	60.65%	1.23
64x64	10	74.89%	74.84%	3.59
128x128	10	74.97%	74.50%	7.30

Table 3. Tools used in biomedical image segmentation.

Biomedical Image	Tools	Used in
CT	Python and Tensorflow	[9]
	PyTorch framework	[14]
MRI	Caffe framework	[15]
	Keras and Tensorflow	[16, 18]
	PyTorch framework	[14]
Ultrasound	MATLAB	[21]
	MATLAB and Dev C++	[34]
	Tensorflow	[22]
	Python using Keras and Tensorflow	[4,5]

Figure 2 shows the U-Net architecture. To implement U-Net, Tensorflow 2.0 and Python 3.7 were used. The laptop that was used has the following characteristics: Windows 10 operating system, Intel Core i3-5015U CPU @ 2.10 GHz (4 CPUs) and 6GB RAM.

Accuracy was calculated with different input image sizes (32x32, 64x64 and 128x128 pixels), using 70% of the slices for training and the remaining for testing.

4 Results and Discussion

Table 2 depicts the image input size, epochs number, the accuracy of the training and test, and the execution time for the training in hours by applying a U-Net architecture for the segmentation of the pulmonary region in Lung TIME dataset.

Table 4. Biomedical image segmentation using Neurals Networks.

Technique	Image	Segmentation applied in	Results	Limitations/Challenges
CNN	MRI	Spinal cord and intra medullary lesions [16]	95% DICE in spinal cord 60% DICE in intramedullary multiple sclerosis	Is sensitive to the quality of the detection module.
	US	Prostate [22]	93.9 ± 3.5% DICE	12 hours of training.
		Breast [23]	85.1% JS	Is prone to errors in boundary of glandular tissues.
CNN & FCN	CT	Spine [9]	94% DICE, 97% SE, 99% SP	13 hours of training.
CDP-ResNet	CT	Lung nodules [10]	81.58% DICE	Is a semi-automatic segmentation method.
DNN	CT	Skeletal muscle at the L3 and T4 levels [11]	96.34 ± 2.77% JS 98.11 ± 1.47% DICE 98.15 ± 1.63% SE, 99.81% SP	Their performance depends on the ground truth labels that are provided.
DPU-Net	US	Arterial walls [29]	HD over 40 MHz dataset 87% JS in lumen and media HD over 20 MHz dataset 90% JS in lumen and media	A fixed-sized kernel cannot be a universal solution.
FCN	CT	Vertebra [14]	94.9 ± 2.1% DICE	Limitations in maximum number of filters per layer.
	MRI	Vertebral body [14]	94.4 ± 3.3% DICE	
MMAN	MRI	Brain [17]	DICE was 86.40% in GM, 89.70% in WM and 84.86% in CF	Its performance varies in different image modes.
RDAU-NET	US	Breast lesions [4]	DICE, SE and SP above 80%	Image resolution affects segmentation.
RIANet	MRI	Anatomical structures of the heart [30	DICE 94.2% in left ventricular 92.3% in right ventricular and 91% in myocardium	Segmentation results of right ventricular is slightly worse.
U-Net	US	Breast [5]	DICE 82.5%	The detected contour can include a large area of non-tumor region, such as shadows.
3D- SkipDenseSeg	MRI	Brain [15]	DICE 90.37 ± 1.38% in WM, 92.27 ± 0.81% in GM and 95.79 ± 0.54% in CF	Sensitivity in computing the distances to the surfaces caused by the low contrast tissues between different classes.
U-Net	CT	Lung region	ACC above 60%	Image resolution affects segmentation accuracy and time of training.

It is observed that the accuracy (ACC) improves with increasing the input size of the slice and the number of epochs, however, the computational complexity increases.

Table 3 shows the tools used for biomedical image segmentation while table 4 shows a summary of the state-of-the-art techniques used in the segmentation of multimodal biomedical imaging specifically CT, MRI and ultrasound (US), using neural networks.

In addition, the limitations or challenges in each segmentation technique, have been presented. The different metrics that were used are Dice similarity coefficient (DICE), Jaccard Similarity (JS), Sensitivity (SE), Specificity (SP) and Accuracy (ACC). In the case of brain segmentation, the following symbology is used: Gray Matter (GM), White Matter (WM) and Cerebrospinal Fluid (CF).

Segmentation was applied in different human body structures such as spinal cord, spine, prostate, chest, lung, arteries and brain. As can be seen, variants of CNN networks [9, 14, 16] and deep learning [11] have performed better compared to other neural network architectures as CDP-ResNet [10] and MMAN [17].

However, CNN and deep learning require a lot of computing power in and the use of the same architecture does not generate the same accuracy in different types of biomedical images. Also, the areas of opportunity in segmentation of biomedical image using neural networks is to reduce computation time during testing and decrease errors in the partitioning of the boundary of tissues.

4 Conclusion

The segmentation of multimodal biomedical imaging is a great challenge and requires constant improvement in accuracy and performance, as it uses multiple images with noise, similar features and irregular shapes, which directly affects the tasks of classifying body structures and diseases prediction and prognosis. In addition, the increase in the dimensions of the images significantly increases the computational complexity of the algorithms.

As a result, various tools, methods and approaches are used in biomedical image segmentation. The metrics that the is generally used to evaluate segmentation performance is DICE score. Neural networks architectures have shown excellent results in the segmentation of multimodal biomedical imaging of CT, MRI and ultrasound.

There are many areas of improvement in neural networks, such as the reduction of training time, the search for kernel that can be used as universal solutions, the improvement of segmentation in the border tissues and the obtaining of results with less image training.

References

1. Toennies, K.: Guide to Medical Image Analysis. Springer (2017)
2. Dougherty, G.: Digital image processing for medical applications. Cambridge (2009)
3. Deserno, T.: Biomedical image processing. Springer (2011)
4. Zhuang, Z., Li, N., Joseph-Raj, A.N., Vijayalakshmi, G.V. Mahesh, Qiu, S.: An RDAU-NET model for lesion segmentation in breast ultrasound images. Plos One, 14(8) (2019)

5. Almajalid, R., Shan, J., Du, Y., Zhang, M.: Development of a deep-learning-based method for breast ultrasound image segmentation. In: 17th IEEE International Conference on Machine Learning and Applications (2018)
6. Bronzino, J.: Handbook of medical imaging. Academic Press (2000)
7. Koprowski, R.: Medical and biological image analysis. IntechOpen (2018)
8. González, D., Villuendas, Y., Argüelles, A.: Experimental comparison of bioinspired segmentation algorithms applied to segmentation of digital mammographies. Research in Computing Science, 138, pp. 109–116 (2017)
9. Vania, M., Mureja, D., Lee, D.: Automatic spine segmentation from CT images using convolutional neural network via redundant generation of class labels. Journal of Computational Design and Engineering, 6, pp. 224–232 (2019)
10. Liu, H., Cao, H., Song, E., Jin, R., Jin, Y., Hung, Ch.: A cascaded dual-pathway residual network for lung nodule segmentation in CT images. Physica Medica, 63, pp. 112–121 (2019)
11. Dabiri, S., Popuri, K., Cespedes-Feliciano, E.M., Caan, B.J., Baracos, V.E., Faisal-Beg, M.: Muscle segmentation in axial computed tomography (CT) images at the lumbar (L3) and thoracic (T4) levels for body composition analysis. Computerized Medical Imaging and Graphics, 75, pp. 47–55 (2019)
12. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
13. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention, pp. 234–241 (2015)
14. Lessmann, N., van Ginneken, B., de Jong, P.A., Išguma, I.: Iterative fully convolutional neural networks for automatic vertebra segmentation and identification. Medical Image Analysis, 53, pp. 142–155 (2019)
15. Bui, T., Shin, J., Moon, T.: Skip-connected 3D DenseNet for volumetric infant brain MRI segmentation. Physica Medica, 54 (2019)
16. Gros, C., De Leener, B., Badji, A., Maranzano, J., Eden, D., Dupont, S.M., Talbott, J., Zhuoquiong, R., Liu, Y., Granberg, T., Ouellette, R., et al.: Automatic segmentation of the spinal cord and intramedullary multiplesclerosis lesions with convolutional neural networks. Neuroimage, 184, pp. 901–915 (2019)
17. Li, J., Yu, Z.L., Gu, Z., Liu, H., Li, Y.: MMAN: Multi-modality aggregation network for brain segmentation from MR images. Neurocomputing, 358, pp. 10–19 (2019)
18. Rak, M., Steffen, J., Meyer, A., Hansen, C., Tönnies, K.D.: Combining convolutional neural networks and star convex cuts for fast whole spine vertebra segmentation in MRI. Computer Methods and Programs in Biomedicine, 177, pp. 47–56 (2019)
19. Rak, M., Tönnies, K.: A learning-free approach to whole spine vertebra localization in MRI. In: Medical Image Computing and Computer-Assisted Intervention, pp. 283–290 (2016)
20. Ilhan, U., Ilhan, A.: Brain tumor segmentation based on a new threshold approach. Procedia Computer Science, 120, pp. 580–587 (2018)
21. Nithya, A., Appathurai, A., Venkatadri, N., Ramji, D.R., Anna-Palagand, C.: Kidney disease detection and segmentation using artificial neural network and multi-kernel k-means clustering for ultrasound images. Measurement, 149 (2019)
22. Karimi, D., Zeng, Q., Mathur, P., Avinash, A., Mahdavi, S., Spadinger, I., Abolmaesumi, P., Salcudean, S.E.: Accurate and robust deep learning-based segmentation of the prostate clinical target volume in ultrasound images. Medical Image Analysis, 57, pp. 186–195 (2019)

23. Xu, Y., Wang, Y., Yuan, J., Cheng, Q., Wang, X., Carson, P.L.: Medical breast ultrasound image segmentation by machine learning. Ultrasonics, 91, pp. 1–9 (2019)
24. Dolejší, M., Kybic, J.: The lung time—annotated lung nodule dataset and nodule detection framework. In: Proceedings of SPIE, 7260 (2009)
25. van Sloun, R.J.G., Wildeboer, R.R., Mannaerts, C.K., Postema, A.W., Gayet, M., Harrie, Beerlage, H.P., Salomon, G., Wijkstra, H., Mischi, M.: Deep learning for real-time, automatic, and scanner-adapted prostate (zone) segmentation of transrectal ultrasound, for example, magnetic resonance imaging–transrectal ultrasound fusion prostate biopsy. European Urology Focus (2019)
26. Simard, P., Steinkraus, D., Platt, J.: Best practices for convolutional neural networks applied to visual document analysis. In: Seventh International Conference on Document Analysis and Recognition (2003)
27. Liu, C., Gardner, S., Wen, N., Elshaikh, M.A., Siddiqui, F., Movsas, B., Chetty, I.J.: Automatic segmentation of the prostate on CT images using deep neural networks (DNN). International Journal of Radiation Oncology Biology Physics, 104(4), pp. 924–932 (2019)
28. Huang, X., Sun, W., Tseng, T.L., Li, Ch., Qian, W.: Fast and fully-automated detection and segmentation of pulmonary nodules in thoracic ct scans using deep convolutional neural networks. Computerized Medical Imaging and Graphics, 74, pp. 25–36 (2019)
29. Yang, J., Faraji, M., Basu, A.: Robust segmentation of arterial walls in intravascular ultrasound images using dual path U-Net. Ultrasonics, 96, pp. 24–33 (2019)
30. Tong, Q., Li, C., Si, W., Liao, X., Tong, Y., Yuan, Z., Heng, P.A.: RIANet: Recurrent interleaved attention network for cardiac MRI segmentation. Computers in Biology and Medicine (2019)
31. Lu, G., Zhou, L.: Localization of prostatic tumor's infection based on normalized mutual information MRI image segmentation. Journal of Infection and Public Health (2019)
32. Sheela, A., Suganthi, G.: Automatic brain tumor segmentation from MRI using greedy snake model and fuzzy C-Means optimization. Journal of King Saud University - Computer and Information Sciences (2019)
33. Yang, T., Song, J., Li, L.: A deep learning model integrating SK-TPCNN and random forests for brain tumor segmentation in MRI. Biocybernetics and Biomedical Engineering, 39, pp. 613–623 (2019)
34. Panigrahi, L., Verma, K., Singh, B.K.: Ultrasound image segmentation using a novel multi-scale gaussian kernel fuzzy clustering and multi-scale vector field convolution. Expert Systems with Applications, 115, pp. 486–498 (2019)
35. Pulagam, A., Rao, V., Inampudi, R.: Automated pulmonary lung nodule detection using an optimal manifold statistical based feature descriptor and SVM classifier. Biomedical and Pharmacology Journal, 10(3), pp. 1311–1324 (2017)

ISSN 1870-4069

Disease Prediction Applying Machine Learning: Case Study of Breast Cancer and Diabetes

Edgar Gonzalo Cossio Franco[1], María de los Ángeles Núñez Herrera[2],
Keuri Adilene Machain Tarula[3], Carlos Daniel Robles Ontiveros[4],
Javier Agustín Ramírez Martínez[3], Marco Julio Franco Mora[5],
Manuel Iván Estrada Chávez[5]

[1] Instituto de Información Estadística y Geográfica de Jalisco,
Mexico

[2] Instituto Tecnológico de la Piedad,
Mexico

3 Instituto Tecnológico del Sur de Nayarit,
Mexico

[4] Instituto Politécnico Nacional,
Mexico

[5] Instituto Tecnológico Superior de Ciudad Hidalgo,
Mexico

Abstract. Breast cancer and diabetes are part of the leading causes of death in the world [1]. According to the World Health Organization, Cancer ranks second while diabetes ranked seventh in 2016 [2]. Through the application of artificial intelligence (AI) it is possible to train a machine to predict future scenarios in terms of determining a positive or negative diagnosis of cancer or diabetes based on historical data. The objective of this research is to implement AI through Machine Learning (ML) for predictive purposes regarding breast cancer and diabetes and thereby diagnose in time. For the present study, linear regression and the J48 algorithm were used.

Keywords: Cancer, diabetes, artificial intelligence, machine learning, linear regression, j48 algorithm.

1 Introduction

Machine learning is an artificial intelligence technique that combines a set of algorithms with the purpose of training machines. This training allows them to learn and based on it discover patterns that allow you to predict scenarios based on historical data.

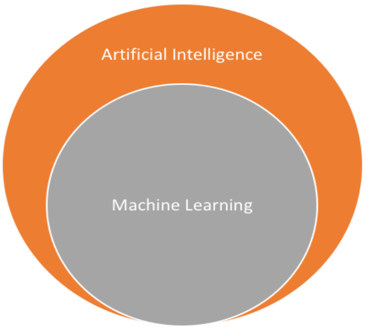


Fig. 1. Machine Learning and artificial intelligence.

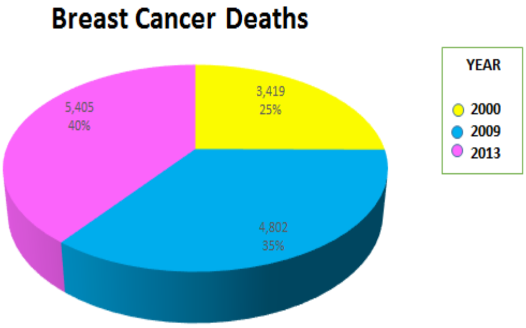


Fig. 2. Increase in deaths with breast cancer.

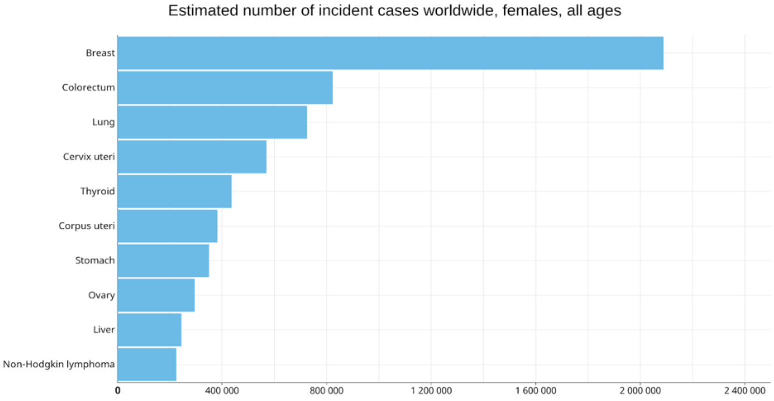


Fig. 3. Deaths worldwide according to type of cancer [5].

Among the machine learning algorithms it is possible to find: decision trees, Naive Bayes, logistic regression, K-means algorithm, linear discriminant analysis, support Vector Machines, Isotonic Separation, Random Forests, Neural Networks, Genetic Algorithms, among others [3].

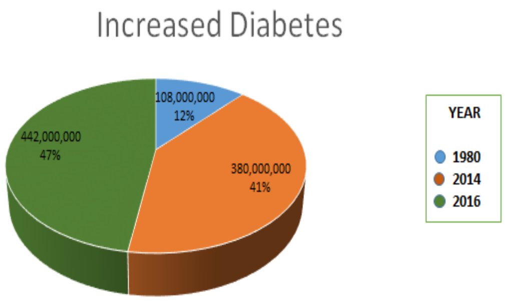


Fig. 4. Increase in deaths due to diabetes worldwide [7].

In recent years, machine learning has been applied with the objective of predicting scenarios in the health, banking, financial, educational and in general all those fields where there are large volumes of data (Big data) [4]. In figure 1 it is possible to see the location of machine learning with respect to artificial intelligence.

In the field of health, artificial intelligence, particularly machine learning, plays an important role because every day unstructured data is generated that allow the identification of patterns through grouping, analysis, segmentation, disease processing and prediction; In the case of the present investigation, we work with two particular cases: cancer and diabetes.

Cancer ranks second worldwide worldwide while diabetes ranks number seven in 2016, according to WHO data [1, 2]. Under this scenario it is important to work on strategies that contribute to the early prevention strategy for possible eradication.

1.1 Problem

When a cancer is detected after its onset, most of the time there is little to do. In figure 2 it is possible to see deaths worldwide according to the types of cancer.

Regarding diabetes, WHO reported that the disease was ranked 7 worldwide in terms of deaths, which represents 422 million people, corresponding to one in 11 people [6].

The reality is that it is an alarming fact and demands that the strategy be taken more seriously. The cause of diabetes occurs when there is a high level of blood sugar and the pancreas is not able to produce insulin or there is but in minimal amounts [7]. The increase in deaths from diabetes was 62 million in just two years, as shown in figure 3.

1.2 Related Work

To carry out this research, work was consulted that has to do with the prediction of cancer and diabetes through machine learning.

In [8] an investigation was carried out with different Machine Learning methods, the research was carried out with an integrated trend of mixed data, such as clinical and genomic, ensuring that the application of such data in machine learning methods could improve the accuracy of Cancer sustainability, recurrence and survival prediction.

Data
result
T²

Fig. 5. General process.

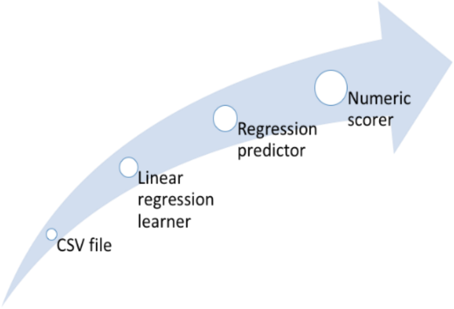
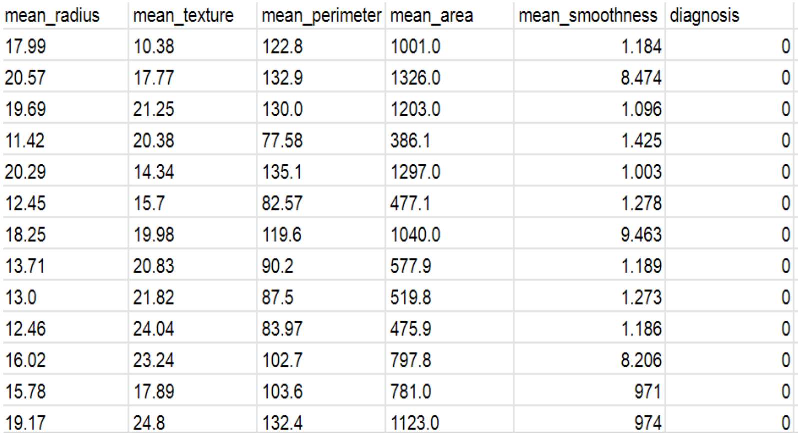


Fig. 6. Dataset treatment.

Table 1. Variable cancer.



mean_radius	mean_texture	mean_perimeter	mean_area	mean_smoothness	diagnosis
17.99	10.38	122.8	1001.0	1.184	0
20.57	17.77	132.9	1326.0	8.474	0
19.69	21.25	130.0	1203.0	1.096	0
11.42	20.38	77.58	386.1	1.425	0
20.29	14.34	135.1	1297.0	1.003	0
12.45	15.7	82.57	477.1	1.278	0
18.25	19.98	119.6	1040.0	9.463	0
13.71	20.83	90.2	577.9	1.189	0
13.0	21.82	87.5	519.8	1.273	0
12.46	24.04	83.97	475.9	1.186	0
16.02	23.24	102.7	797.8	8.206	0
15.78	17.89	103.6	781.0	971	0
19.17	24.8	132.4	1123.0	974	0

On the other hand, in [9], applying different supervised and unsupervised learning methods, the regression is highlighted as the most accurate method for a more accurate result using the Gradient Descent algorithm, which consists in that the cost function is reduced when the model adjusts its parameters. Another method used is neural networks, which are inspired by the biological neuronal system, although they do not work in the same way. This method receives data in a layer called input layer. The data will be purchased and the model automatically identifies the characteristics of the data and labels or names them. This is the first model he selects for his research for the discrimination of malignant or benign tumors among patients with breast cancer, built with many hidden layers to better generalize the data.

The research of [10] applies methods of support vector machines and generalized discriminant analysis for classification of information with an accuracy ranging from 78.21 to 82.05 percent in the prediction of diabetes disease.

In [11], an algorithm based on neural networks is applied for the detection of breast cancer, inspired by social search algorithms of animals which offers promising results even above the particle accumulation algorithm. The k-means method is applied with modifications in [12], as well as the Support Vector Machine method is applied for the prediction of diabetes with an accuracy of 99.64%.

2 Methodology

For the present investigation, we worked with the methodology shown in Figure 4 where the process of obtaining, analyzing, training and results of the data set is described. The linear regression method and decision trees were used.

2.1 Process

As part of the process that guided the present investigation, data exploration was necessary, for which the dataset of Pima indigenous women of at least 21 years of age is used, in the case of diabetes. In the case of cancer, the data were obtained from the Hospitals of the University of Wisconsin, Madison, from Dr. William H. Wolberg. The data set is from the National Institute of Diabetes and Digestive and Kidney Diseases.

In this process, the data were obtained for analysis and processing. In the case of diabetes, the variables are composed as:

Diabetes: Number of pregnancies, Age, Pedigree, Plasma, Blood Pressure, Insulin in the body, Body mass, Skin thickness. In the case of cancer, the following variables are available.

2.1.1 Data

Cancer: Radius of the tumor, Texture of the tumor, Perimeter of the tumor, Area of the tumor and softness of the tumor.

Both datasets were processed as shown in Figure 5 where four strategic moments are identified: the dataset reader (CSV file), linear regression learner, regression predictor and numeric scorer.

2.1.1.1 CSV

Cancer: In the cancer dataset the variables to contribute are very specific since the data obtained are from tumors extracted from the body of people. Table 1 shows an extract of the variables.

Mean Radius: it is the measure of the radius of the extracted tumor that is given in millimeters, the radius can be large or small and it is not determined that it is benign or malignant until it is in conjunction with the other variables.

Mean Texture: It is also known as the average grayscale on radiographs or tomographs performed on the tumor. Grayscale depends a lot on the texture of the tumor.

Mean Perimeter: this measure is the perimeter of the tumor, as it is well known that the tumors are neither round nor oval at all but it is a deformed fat pellet therefore when it has a perimeter greater than 2 cm it is said to be accurate malignant, but that does not always turn out that way until they are related to the other variables and have a more accurate prognosis.

Mean Area: this variable is the most important in a medical analysis and diagnosis of the tumors in conjunction with the perimeter and texture of the tumor, since if its area exceeds 5 cm or the size of a lemon it is probably that you have cancer. In most cases, where the area is 5 cm it is not a single tumor anymore, if they are not several together as a bunch of grapes, diagnosed as advanced cancer where despite the treatment there is no longer a cure.
Mean Smoothness: Average variation in radio lengths, as mentioned above, the tumors have an irregular shape, which takes an average of variations between the different possible radii of these.

Diabetes: The diabetes dataset shows variables that have to do with the factors that can determine it. Table 2 shows an extract.

Preg: pregnancy is the number of pregnancies that women have had regardless of whether or not they had abortions. The number of pregnancies in every woman is important because they create new substances during and after pregnancy, which if they reproduce to a greater extent or do not reproduce cause serious side effects on health.

Plas: is the plasma glucose concentration at 2 hours in an oral glucose tolerance test. When the plasma glucose is above 11.1 mmol/dl it is diabetes.

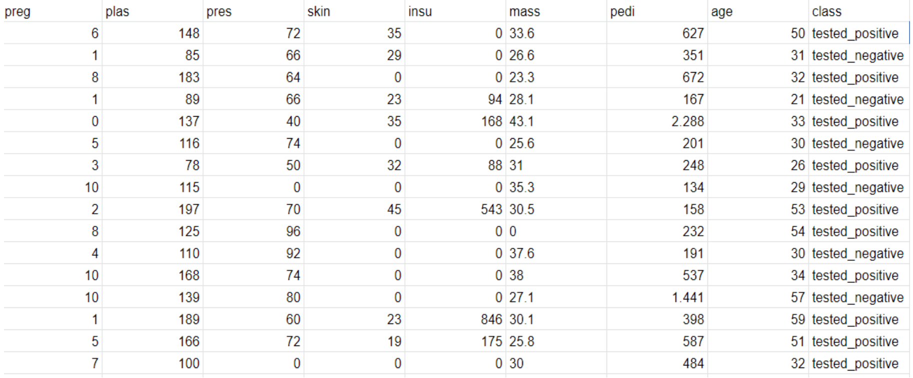
Pres: diastolic blood pressure (mm Hg) commonly called high blood pressure or hypertensive people who suffer from it. High blood pressure shows a very high prevalence in type 2 diabetes mellitus and is a risk factor for the development of cardiovascular complications. Strict control of blood pressure to figures less than 130/80 mm Hg reduces cardiovascular and renal morbidity and mortality to a greater extent than the control of other complications.

Skin: Thickness of the triceps skin fold (mm).

Insu: 2-hour serum insulin (mu U / ml) is a hormone that takes glucose from the blood and transports it into the body's cells where it is used as energy. Diabetes occurs when the pancreas does not produce enough insulin or when the body does not use insulin properly.

Mass: Body mass index (weight in kg / (height in m) ^ 2). Diabetes favors the appearance of muscular atrophy. Given this scenario, researchers have discovered that

Table 2. Diabetes variables.



preg	plas	pres	skin	insu	mass	pedi	age	class
6	148	72	35	0	33.6	627	50	tested_positive
1	85	66	29	0	26.6	351	31	tested_negative
8	183	64	0	0	23.3	672	32	tested_positive
1	89	66	23	94	28.1	167	21	tested_negative
0	137	40	35	168	43.1	2.288	33	tested_positive
5	116	74	0	0	25.6	201	30	tested_negative
3	78	50	32	88	31	248	26	tested_positive
10	115	0	0	0	35.3	134	29	tested_negative
2	197	70	45	543	30.5	158	53	tested_positive
8	125	96	0	0	0	232	54	tested_positive
4	110	92	0	0	37.6	191	30	tested_negative
10	168	74	0	0	38	537	34	tested_positive
10	139	80	0	0	27.1	1.441	57	tested_negative
1	189	60	23	846	30.1	398	59	tested_positive
5	166	72	19	175	25.8	587	51	tested_positive
7	100	0	0	0	30	484	32	tested_positive

an increase in blood sugar levels triggers the decrease in muscle mass. In addition, they have observed that the abundance of the transcription factor KLF15 increased in the skeletal muscle of diabetic mice.

Pedi: Pedigree function of diabetes (genetics). There is a complex genetic transmission of diabetes. A predisposition to the disease is inherited to which different environmental triggers are added. This is true for both diabetes 2 and 1, although more marked in the first.

Age: Age of the people with whom the research was conducted in the indigenous Pima group.

2.1.1.2 Linear Regression Learner

Method used to train the machine with the dataset from the set of independent variables to a dependent. Perform a multivariate linear regression.

2.1.1.3 Regression Predictor

It is the node that predicts the response using the regression model. This node must be connected to CSV reader and the Linear Regression learner. It is executed as long as the test data contains the columns related to each other. The node adds at the end of the dataset a column to the input table that contains the prediction for each row; depending on the case of the dependent variable.

2.1.1.4 Numeric Scorer

This node statistically calculates the values of the numerical columns called r¡ and the predicted values known as p¡. The calculated variables are as follows:

R²: It is a statistics used in the context of statistical models whose main purpose is the prediction of future results. In short, it can be defined as the precision in which the prediction is made according to the dependent and independent variable:

$$1\text{-SS res/SS tot}= 1-\sum(p\ i-r\ i)^2 / \sum (r\ i-1/n *\sum r\ i)^2 .$$

Error Absolute Medium: measurement of difference between two continuous variables:

$$(1/n * \sum \mid p\ i\text{-}r\ i).$$

Error Quadratic Medium: sum of the squares of the waste. It is a measure of the difference between the data and an estimation model:

$$(1/n * \sum (p\ i\text{-}r\ i)^2).$$

Error Root Mean Square: As used in the differences between the values predicted by a model or an estimate and the observed values. Represents the square root of the second sample of the differences between the predicted values and the observed values:

$$(\mathrm{sqrt}\ (1/n * \sum(p\ i\text{-}r\ i)^2)).$$

Difference with middle sign: It shows that it summarizes how well a set of estimates equals the quantities, which they must estimate. It is a statistics that is used to evaluate an estimation procedure:

$$(1/n * \sum (p\ i\text{-}r\ i)).$$

2.1.2 T2 (Training and Test)

Before applying the Machine Learning method, a Data Science is performed that is responsible for cleaning the dataset to make a more accurate prediction and the nodes can be configured correctly.

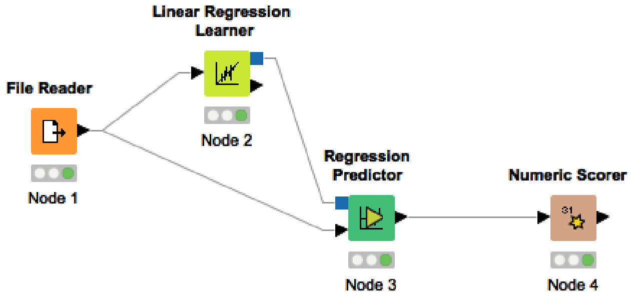


Fig. 6. Model for prediction.

Linear Regression

The Linear Regression method is the relationship between independent or explanatory variables and dependent or response variables. Which will allow obtaining a prediction of the dependent variable or response based on the given values of the independent variable.

The linear regression is expressed by the following expression:

$$Y=Mx + n.$$

J48 Algorithm

It is an induction algorithm, which generates a rule or tree structure from subsets or case windows extracted from the training dataset. Its form of processing is based on generating a structure of rules and assesses its goodness using criteria that measure the accuracy in the classification of cases using two main criteria to direct the process:

1. Calculates the value of the information provided by a candidate rule or branch of the tree, with a routine called info.

2. Calculate the overall improvement that a branch or rule provides using a routine called gain or benefit.

3 Results

The results obtained in the research were satisfactory for the topics discussed since they are the main causes of deaths worldwide in women, where statistics increase every day also spreading in men and children.

The predictions and classification of data with the Linear Regression, Decision Tree j48 methods have the following precision:

- Linear Regression:
- Cancer, R^2=0.93,
- Diabetes, R^2=0.95.

4 Conclusion

Through the application of machine learning, such as linear regression and J48 algorithm properly trained and tested, it was possible to establish a reliable model of prediction of breast cancer and diabetes in .93 and .95 percent, respectively, which is encouraging as it opens the possibility of application and testing with other methods that make the prediction stronger.

5 Future Work

It is contemplated to continue working with different tools offered by artificial intelligence such as:

Neural networks of single layer and multilayer for the prediction and prevention of high impact issues in society; criminal incidence and causes of maternal death during pregnancy.

The possibility of georeferencing the criminal incidence through heat maps is contemplated. Another objective is to create dashboards that show the real-time information of a fact linked to the dataset.

References

1. Who: Cancer key facts, from WHO (2019)
2. Who: Cancer key facts, from WHO (2019)
3. Abreu, P.H., Santos, M.S., Abreu, M.H., Andrade, B., Silva, D.C.: Predicting breast cancer recurrence using machine learning techniques: A systematic review. ACM Computing Surveys, 49(3), pp. 1–40 (2016)
4. Lugo-Reyes, S., Maldonado-Colín, G., Murata, C.: Artificial intelligence to assist the clinical diagnosis in medicine. Alergia México Magazine, 61(2), pp. 110–120 (2014)
5. Who: International Agency of Research on Cancer: Who (2019)
6. Who: International Agency of Research on Cancer: Who (2016)
7. Who: International Agency of Research on Cancer: Who (2014)
8. Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., Photoiadis, D.I.: Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal (2015)
9. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I.: Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal (2016)
10. Polat, K., Güneş, S., Arslan, A.: A cascade learning system for classification of diabetes disease: generalized discriminant analysis and least square support vector machine. Expert Systems with Applications (2006)
11. He, S., Wu, Q.H, Saunders, J.R.: Breast cancer diagnosis using an artificial neural network trained by group search optimizer. Transactions of the Institute of Measurement and Control, 31(6), pp. 517–531 (2009)
12. Afzali, S., Yildiz, O.: An effective sample preparation method for diabetes prediction. International Arab Journal of Information Technology (IAJIT), 15(6), pp. 968–973 (2018)

ISSN 1870-4069

Engagement and Reasons for Selecting University Studies Using the Clustering Technique

Erika Yunuen Morales Mateos[1], María Arely López Garrido[1],
José Alberto Hernández Aguilar[2], Carlos Alberto Ochoa Ortiz[3],
Arturo Corona Ferreira[1], Oscar Alberto González González[1]

[1] Universidad Juárez Autónoma de Tabasco Cunduacán,
Mexico

[2] Universidad Autónoma del Estado de Morelos, Cuernavaca,
Mexico

[3] Universidad Autónoma de Ciudad Juárez,
Mexico

erika.morales@ujat.mx, arely.lopez@ujat.mx,
jose_hernandez@uaem.mx, alberto.ochoa@uacj.mx,
arturo.corona@ujat.mx,oscar.gonzalez@ujat.mx

Abstract. The objective of this work is to discover groupings of university students, characterized by their student engagement, motivation for career selection and academic performance, this through the application of a descriptive data mining technique called low cluster The k-means algorithm. The study was carried out in a university in southern Mexico with students belonging to careers in technology. The UWES-S instrument and a series of questions related to the reasons for choosing university studies were used to measure student engagement. A descriptive data analysis and a heat map were performed, which graphically represents the variables involved and their relationship, then the techniques necessary for the development of clustering were applied. In order to carry out these analyzes, the R language was used. As results, two clusters were found, cluster 1 is identified by students with a high level of student engagement, and the main reason for selecting the career is due to the possibility of work It offers, followed by the academic quality of the career, on the other hand, cluster 2 is characterized by a medium student engagement also by the possibility of work it offers followed by the aptitudes for studies.

Keywords: Engagement, career choice, heat map, cluster, k-means.

1 Introduction

University students can present different behavior in front of their university careers, different levels of student engagement, which is probably related to other variables such as career selection reasons, to know if the career they study is the one they wanted to study. Students make the decision to select their career according to a series of parameters, such as if they have aptitudes for it, for the academic quality of the career, and even for the job opportunities offered by the career, among others.

These are elements that are interesting to know, as universities are concerned with offering quality programs to their students and terminal efficiency is an important factor, since the objective is for students to finish their careers in accordance with the guidelines established in the programs of studies. Therefore, in this study it is proposed to analyze student engagement, motivation for career selection and academic performance, in students of technology careers of a university in southern Mexico, so that groups of students with characteristics are identified similar among members of a group and different from other groups.

The identification of groups can be done through clustering, which is a data mining technique that organizes information into segments, has a great capacity for prediction as new data can be classified into existing groups, thanks to this It knows that they share common characteristics and behaviors, and in addition to each identified group, other techniques based on different algorithms can be applied, so that the data can be studied in a more efficient way [1].

In general terms, research has been carried out that measures student engagement with statistical techniques, offering interesting results when related to other variables, among these works are the following:

A work developed by Vizoso and Arias [2] determined the relationship between engagement, burnout, academic performance and career choice priority, in a group of students from the University of León from different careers, resulting in students who are pursuing the career as a first option have higher levels of engagement, that is, high levels of vigor, dedication and absorption, as well as greater academic performance, than those who are not studying the preferred career. As for the burnout that is a state of exhaustion and cynicism, both groups of students feel equally fatigued, it is manifested that it may be due to the academic stress of studying a university degree.

Carrasco and Martínez [3] conducted an investigation on the level of engagement and its implication with academic performance with university students of health sciences, found with respect to the level of engagement, of the 100% evaluated sample He found that 82.0% of university students are at the high level and 18.0% at the low level, which generally means that all students have adequate levels of engagement, although no relationship was found. Significantly with academic performance in general, the analyzes by faculties reported that there is a significant relationship in two faculties.

In a research carried out at the University of Malaga in Spain, on perceived stress, burnout and engagement in university students, the results obtained show that students have medium-low levels in cynicism, means in exhaustion, inadequacy, vigor, absorption and stress perceived and medium-high in dedication.

In general, as students perceive more stress, they also show greater exhaustion, cynicism and inadequacy and less vigor, dedication and absorption [4].

2 Proposed Materials and Methods

2.1 Description of the Data

This paper shows mainly results on student engagement and the reason for the selection of university studies that have students in technology careers, from a population of 141 students of a computer science faculty at a university in southern Mexico, who Willingly agreed to answer a questionnaire that includes the UWES-S and additional elements related to the investigation, the period of application of the survey was February-August 2015. The sample selected was not probabilistic, directed and for convenience [5].

2.2 Utretch Work Engagement Scale for Students (UWES-S)

Salanova and Schaufeli [6] define engagement as "a positive psychological state characterized by high levels of energy and vigor, dedication and enthusiasm for work, as well as total absorption and concentration in work activity". Student engagement characterized by vigor, dedication and absorption is measured through the Utretch Work Engagement Scale for Students (UWES-S), which is composed of 17 items, measured through a Likert scale with values from zero to six .

Vigor is related to high levels of energy, effort, not fatigue easily, and persistence. Dedication refers to the meaning of studies, to feeling enthusiastic, proud, and inspired. Absorption is feeling happily immersed in their studies and presenting difficulty separating from them, feeling that time passes quickly and forgets everything around [7].

2.3 Circumstances of Choice of Studies

Four variables were used to know the reasons why university students selected their careers, later these variables were related to student engagement. These four variables of the choice of studies are: the educational value, academic quality of the career, aptitudes for studies and the possibility of work. For this, four questions that measure these variables were taken as a basis and are integrated in an answer, where 0 means nothing, 2 equals very little, 3 means little, 4 equals half-mind, 5 corresponds to enough, and 6 means a lot [8].

2.4 Analysis of the Data

A descriptive analysis was applied to know the characteristics of the study population, such as quantities and percentages of the variables gender, age, career and social stratum.

Subsequently, the minimum, maximum, average and standard deviation values of the variables engagement, vigor, dedication and absorption were calculated, as well as the variables related to the reason for choosing university studies.

A heat map is presented that shows through different colors and all the intensity of the relationship between variables, from a light tone to a high ratio, through yellow and orange colors to intense red that indicates a minimum relationship [9].

To know and describe characteristics of groups in the population, the clustering technique was applied with the variables involved in the study. With this technique, the data is grouped so that those belonging to the same group have similarities to each other and those belonging to different groups show significant differences [1]. The algorithm used to carry out the clustering was the k-means, where the number of clusters to be created must be specified in advance.

To determine the number of clusters that should be presented, the elbow method was used, where in a graph the point of the elbow is the place where there is a significant change and is the number of clusters that should be generated.

The R language was used to perform the different analyzes presented. R is a powerful software in the implementation of complex graphics and analysis, since it has a wide variety of packages for these tasks [9, 10].

Clustering

Clustering is a descriptive task that consists in obtaining "natural" groups from a data set. The data is grouped based on the principle of maximizing the similarity between the elements of a group by minimizing the similarity between the different groups. Clustering is related to summarization, where each group is considered as a summary of the elements that form them in order to accurately describe the data [11].

An algorithm widely used to group by partitioning is the k-means thanks to its simplicity, it is described in Algorithm 1. To use it you must specify the number of clusters that are going to be generated, it is the parameter k, for which you randomly select k elements, which represent the center or average of each cluster. Subsequently, each of the instances (example) is assigned to the center of the nearest cluster according to the Euclidean distance that separates it from it.

For each cluster the centroid of all its instances is calculated. These cen-troids are taken as the new centers of the clusters. The entire process is repeated with the new cluster centers. The iteration continues until the assignment of the same instances to the same clusters is repeated, since the central points have stabilized and will remain in variables after each iteration [1].

Algorithm 1. K-means [1].

```
Algorithm 1: K-means
Choose k examples that act as seeds (k number of clusters).
For each example, add example to the most similar class.
Calculate the centroid of each class, which become the new seeds.
If you do not reach a convergence criterion (for example, two iterations do not change the
classifications of the examples), go back to step 2.
```

Table 1. Characteristics of the population study sample.

Variables	**Values**	**N**	**%**
Gender	Women (1)	54	38
	Men (2)	87	62
Age	18-19	35	25
	20-21	69	49
	22-23	37	26
Career	LIA (1)	51	36
	LSC (2)	71	50
	LTI (3)	12	9
	LT (4)	7	5
Social stratum	Low-Low (1)	17	12
	Low-High (2)	43	30
	Medium-Low (3)	68	47
	Medium-High (4)	11	7
	High-Low (5)	5	3
	High-High (6)	1	1

Table 2. Descriptive statistics of the population study sample.

Variables	**Mean**	**Standard Deviation**	**Minimum**	**Maximum**
Engagement	4.32	0.94	1.60	6.00
Vigor	4.06	0.99	1.50	6.00
Dedication	4.77	0.98	1.80	6.00
Absorption	4.13	1.05	1.20	6.00
Formative value	4.22	1.24	1.00	6.00
Academic quality	4.23	1.52	1.00	6.00
Aptitudes	4.37	1.09	1.00	6.00
Possibility work	4.67	1.11	1.00	6.00
Average	8.15	0.69	6.70	9.80
Selected career	2.45	1.33	0.00	4.00

3 Results

Table 1 shows a sample of 141 students, 87 men and 54 women, the age ranges from 18-19, 20-21, 22-23, where the values of 35, 69 and 37 correspondingly, the population of students with ages in the range of 20-21 being greater. The courses considered are four, where the largest population is in LSC (Bachelor in Computer Systems) with 71, followed by LIA (Bachelor in Administrative Computer Science) with 51, LTI (Bachelor in Information Technology) with 12 and LT (Licensed in Telematics) with 7. The social stratum of the students is mostly medium-low with 68 students.

A descriptive analysis was carried out that shows the mean, standard deviation, minimum and maximum, for the variables presented in Table 2. The student

Table 3. Proposed scale to measure the UWES-S (Adapted from Schaufeli and Bakker [7]).

Valores	**Engagement**
Very low	score < 2.20
Low	2.20 ≤ score < 3.30
Medium	3.30 ≤ score < 4.70
High	4.70 ≤ score ≤ 6.00

engagement, presents a minimum of 1.60 and a maximum of 6.00, so as an average of 4.32. Of the dimensions of the student engagement, the dedication presents the highest average value with 4.77, minimum of 1.80, maximum of 6.00, followed by absorption with an average of 4.13, minimum of 1.20, maximum of 6.00, finally the vigor with a average of 4.06, minimum of 1.50, maximum of 6.00,

Regarding the career choice by, the training value, academic quality, aptitudes and work possibility presents in the average the values of 4.22, 4.23, 4.37, 4.67 correspondingly, as well as a minimum of 1.00 and a maximum of 6.00. As for the average, it has values of minimum 6.70, maximum of 9.80 and an average of 8.15. Another variable is the selected race where the average is 2.45, the minimum of 0 and the maximum of 4.00.

Table 3 shows an equivalence table for the categorical values according to their score, with established values the observations of the cases can be evaluated. Schaufeli and Bakker present more techniques to obtain the level of engagement to the UWES-S [7].

Using the proposed scale, it can be interpreted in a general way that the student engagement is 4.32, so that the students feel regularly engaged.

3.1 Heat Map of the Correlations

The correlation is a descriptive task that analyzes the percentage of similarity between the values of two numerical variables. The mathematical model with a correlation coefficient r is used, which takes values between 1 and -1.

The strongly correlated variables have a coefficient of 1 or -1 (positively or negatively), on the contrary if the value is 0 there is no correlation. These analyzes allow studying relationships between cause-effect attributes [1].

To find the possible relationships between variables, a heat map was used, since visualization is one of the first forms of data inspection.

The variables in the data set are shown in the heat map of Figure 1, the level of correlation is observed depending on the intensity of the color, the most intense color indicates a correlation closer to the value 0 and the lighter color indicates a correlation closest to 1.

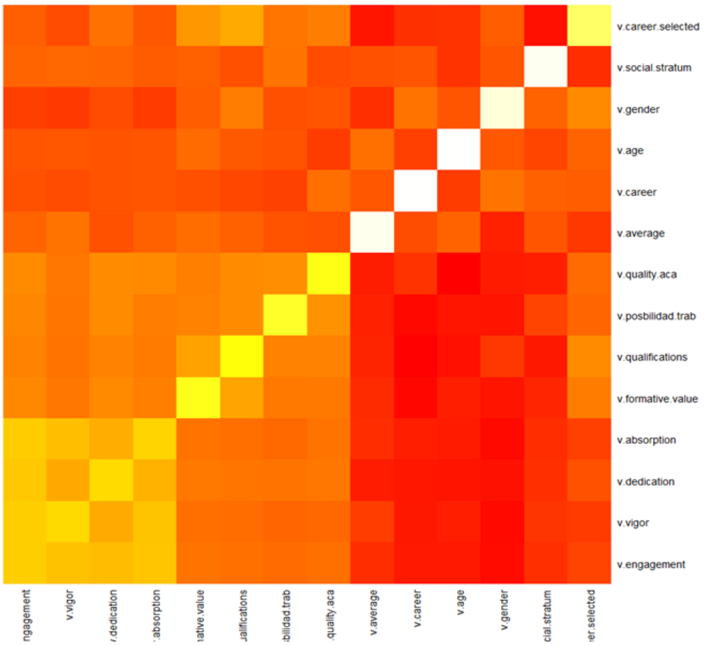


Fig. 1. Heat map of the correlations between attributes.

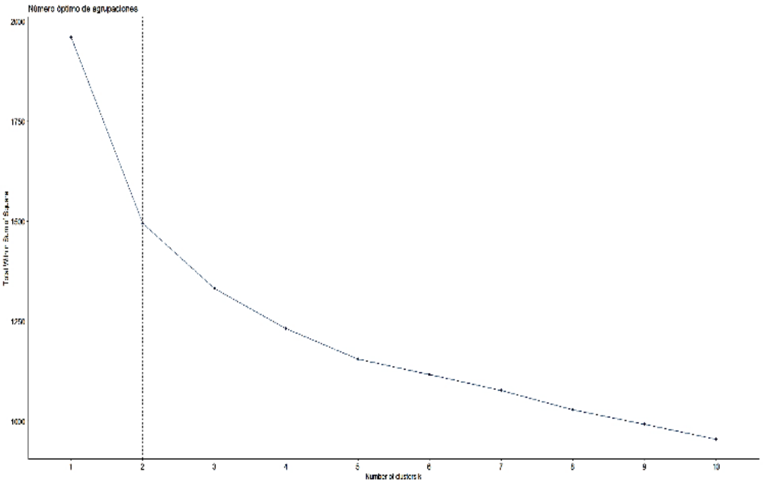


Fig. 1. Elbow technique to determine the optimal number of clusters.

Figure 1 shows that the most related variables are those found in the lower left with lighter colors, student engagement, vigor, dedication, absorption, formative value, academic quality, skills and possibility. work; on the contrary, the less related are found in the lower right, a box with darker shades (red) is observed.

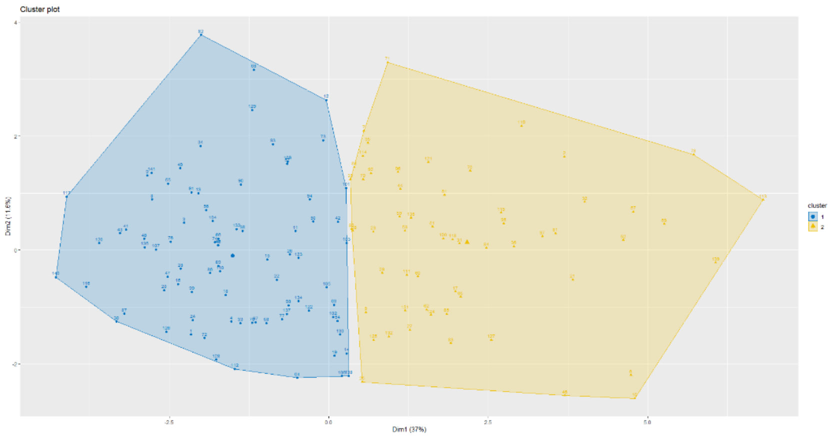


Fig. 2. Cluster graph obtained with the k-means algorithm.

3.2 Clustering

Clustering is a technique that allows you to find common properties that group a set of data. It seeks to maximize the distance of the instances to the grouping to which they do not belong, minimizing the distance of the own [1].

To carry out this analysis, the optimal number of clusters was first determined, the elbow technique that determines the moment when increasing the number of clusters does not imply a substantial measure in a quality measure was applied. Figure 2 shows that according to the elbow technique the optimal number of clusters is two.

Subsequently, the clustering of two groups was created, using the k-means algorithm, Table 4 shows the results,

Cluster 1 corresponds to the student engagement with a high level with 4.88, as well as its dimensions of vigor, dedication and absorption, with high values, 4.59, 5.34, 4.73 respectively; As for the variables that determine the reason for choosing studies, he leads the possibility of work offered by the career with 5.15, followed by the academic quality of the career with 4.91, very close to the aptitudes for studies with 4.90, subsequently by the formative value of the career with 4.81. Additionally, they present approximately an academic average of 8.21, they are from the LSC career, their age is 21, it is made up of men and women, the social stratum is medium-low and the career they study was the one they wanted to study.

Cluster 2 corresponds to the student engagement with a medium level with 3.51, as well as its dimensions of vigor, dedication and absorption, with average values, 3.30, 3.96 and 3.28 respectively; For the variables that determine the reason for the choice of studies, it leads the possibility of work offered by the career with 3.98, followed by the aptitudes for the studies with 3.62, later by the formative value of the career with 3.37 and the academic quality of the race with 3.25. As for the additional variables, this group has an academic average of 8.10, they are mainly from the LSC and LIA, their

Table 4. Characteristics of the obtained cluster.

Variables	Cluster 1	Cluster 2
v.engagement	4.88	3.51
v.vigor	4.59	3.30
v.dedication	5.34	3.96
v.absorption	4.73	3.28
v.formative.value	4.81	3.37
v.qualifications	4.90	3.62
v.posbilidad.trab	5.15	3.98
v.quality.aca	4.91	3.25
v.average	8.21	8.10
v.career	1.89	1.72
v.age	20.79	20.75
v.gender	0.63	0.58
v.social.stratum	2.71	2.39
v.career.selected	2.95	1.74

age is 21 years, it is made up of men and women, the social stratum is low-high and the Career they studied was the one they wanted to study to a small extent.

Figure 3 shows the grouping obtained with the k-means algorithm, it is observed that the resulting clusters are not overlapping and all instances were classified.

4 Conclusions

In this work, an analysis has been developed that involves variables related to studies, that is, student engagement, knowing the reasons for career selection, knowing if the career they study is what they wanted, academic performance and other personal identification variables. The R language was used for the implementation of data analysis and representation techniques. Descriptive statistics indicated that for the general population it has a engagement of 4.32 considered medium, dedication of 4.77, followed by absorption with 4.13, and vigor 4.06, an average of academic performance of 8.15 and the selected career where the average is 2.45 That means medium-mind.

In the heat map, the relationship between variables was identified, the ones that present the most relationship are the dimensions of student engagement with those of reasons for the selection of university studies. Finally, a data mining technique called clustering was applied using the k-means algorithm, where two groups were identified, group 1 presents high characteristics of student engagement unlike group 2 that presents a medium engagement; both groups present as a reason for choosing studies the possibility of work offered by the career, at different levels, group 1 quite, group 2 medium; In addition, group 1 selected their career fairly and group 2 a little, both groups have an average of about 8.

In these results in group 2 despite the average level of engagement, since it was not the career selected at a high level, maintains a favorable academic average. It is proposed to continue studying both identified groups, with other techniques to learn more about population behavior, as well as to replicate this study in other careers to know if there are differences given the nature of the same.

References

1. García, J., Molina, J.M., Berlanga, A., Patricio, M.A., Bustamante, A.L., Padilla, W.R.: Ciencia de datos, técnicas analíticas y aprendizaje estadístico. Alfaomega (2019)
2. Vizoso, C.M., Arias, O.: Engagement, burnout y rendimiento académico en estudiantes universitarios y su relación con la prioridad en la elección de la carrera. Revista de Psicología y Educación, 11(1), pp. 45–60 (2016)
3. Carrasco, M.A., Martínez, C.: Nivel de engagement y su implicancia en el rendimiento académico en estudiantes universitarios de ciencias de la salud de la Unheval-Huänuco. Revista Boletín Redipe, 8(2), pp. 131–139 (2019)
4. Vallejo, M., Aja, J., Plaza, J.J.: Estrés percibido en estudiantes universitarios: influencia del burnout y del engagement académico. International Journal of Educational Research and Innovation (IJERI), 9, pp. 220–236 (2018)
5. Hernández, R., Fernández, C., Baptista, M.: Metodología de la investigación. McGraw-Hill Interamericana (2010)
6. Salanova, M., Schaufeli, W.B.: El Engagement de los empleados un reto emergente para la dirección de recursos humanos. Estudios Financieros, (261), pp. 109–138 (2004)
7. Schaufeli, W.B., Bakker, A.: Utrecht work engagement scale (UWES). Escala de Engagement en el trabajo de Utrecht, Ocuppational Health Psychology Unit: Utrech University (2003)
8. Artunduaga, M.: Cuestionario sobre rendimiento académico y deserción en la universidad. Universidad Complutense de Madrid Facultad de Ciencias de la Educación Departamento de Métodos de Investigación y Diagnóstico en Educación (MIDE) (2005)
9. Guisande, C., Vaamonde, A.: Gráficos estadísticos y mapas con R. Ediciones Díaz de Santos (2013)
10. Development Core Team: A language and environment for statistical computing. R Foundation for Statistical Computing (2019)
11. Hernández, J., Ramírez, M.J., Ferri, C.: Introducción a la minería de datos. Editorial Pearson Educación (2004)

ISSN 1870-4069

Information Reliability in Estimate Process: A Novel Blockchain Model

Inés Borunda, Iván Pérez, Erwin Martínez, Alberto Ochoa Zezzatti

Universidad Autónoma de Ciudad Juárez,
Mexico

{al187052, emartine, ivan.perez,
alberto.ochoa}@uacj.mx

Abstract. This study is based on the implementation of the Blockchain in the process of the "recuse for quote" process that feeds the system of estimates for the quotation of new products, or changes to existing ones, used by automotive companies These companies develop close working and business relationships with their suppliers, that's why is extremely valuable to have accurate information, since it depends on it to make strategic decisions to improve the quality, reduce costs and minimize delivery times. The proposed data network model (blockchain) manages the exchange of information between the different departments involved in an estimated procedure, so that all parties can synchronize the reliable information, avoiding using modified information, which creates confidence in the decisions making. Decision-making with inaccurate information implies the loss of trust and competitiveness with suppliers and customers. The implementation the blockchain helps to detect any type of manipulation or alteration in the information in a timely manner and improve the integrity of the infrastructure in the information security.

Keywords: Blockchain, information security, estimated procedure and decision making.

1 Introduction

One of the problems faced by manufacturing companies is to have the reliability of their information systems, since this, they feed the system to be able to provide an estimated quote for new products. Considering that the estimates require certainty in the unit costs of production, the value of direct labor, the components in the list of materials and the indirect charges that may occur soon. This implies that the estimated costs must indicate what it can cost to produce a new product, providing the analysis tools necessary for decision making. However, it is not uncommon to find discrepancies between the information provided by the suppliers and the information in the estimate systems.

This risk has been previously identified [1] who points out that because the information is asymmetric the problem of fraud can occur between the subjects of the company. Therefore, the supervision and restriction corresponding to the security systems is required to guarantee the reliability of the data, through the different departments involved in the request for quote process. Prioritizing policies according to management guidelines, in order to guarantee the reliability of information and assertiveness in decision making, the risk of generating estimates with inaccurate information can result in the loss of business, credibility as a company and sustainability, in the future.

Therefore, it is proposed to apply blockchain to this key process in manufacturing organizations. By 2020, it is estimated that 60% of the leading manufacturers will depend on digital platforms, which will be responsible for supporting the functions that are responsible for 30% of their revenues. By 2021, the range of new technologies will be integrated into the manufacturing sector, with 20% of the leading manufacturers depending on some combination of artificial intelligence, internet of things; cognitive systems and blockchain [2].

That is why achieving transparency requires accurate and secure data collection in the storage of these, a difficult task that is currently entrusted to third parties through centralized information repositories [3].

Therefore, it is extremely important to consider that sectors such as: health, insurance, government and supply chain management are likely to be transformed by the *blockchain* [4, 5] discussed the value of the *blockchain* application in the smart contract. [6] proposed *blockchain* applications to influence laws and regulations. [7] built the data of the shared security network system.

However, *blockchain* has not been detected in the handling of customer-supplier information regarding the "*recuse for quotes*" used to quote estimates. This model provides the security tools based on cryptography using mathematical formulas incorporated in the hash function that allows the information received to not be modified by any department. It is worth mentioning that each department involved in the "recuse for quote" process can add information regarding their position, without modifying the information of the other departments.

2 Theoretical Framework

2.1 BlockChain

Blockchain technology is relatively recent, so it is frequently related to bitcoin. However, the *blockchain* and bitcoin are different. Because the information in the *blockchain* is encrypted through mathematical formulas, as well as non-symmetric encryption algorithms that guarantee the security of data in the transaction and technology of economic models, obtaining the principle of cryptography, reconciling the parties involved without the need for an intermediary third party [8]. What makes *bitcoin* a successful application in the *blockchain*. In essence, the *blockchain* is a database distributed, decentralized, secure and reliable similar to an account book (distributed accounting), which records all transactions digitally permanently with the ease of being able to achieve its trajectory [9].

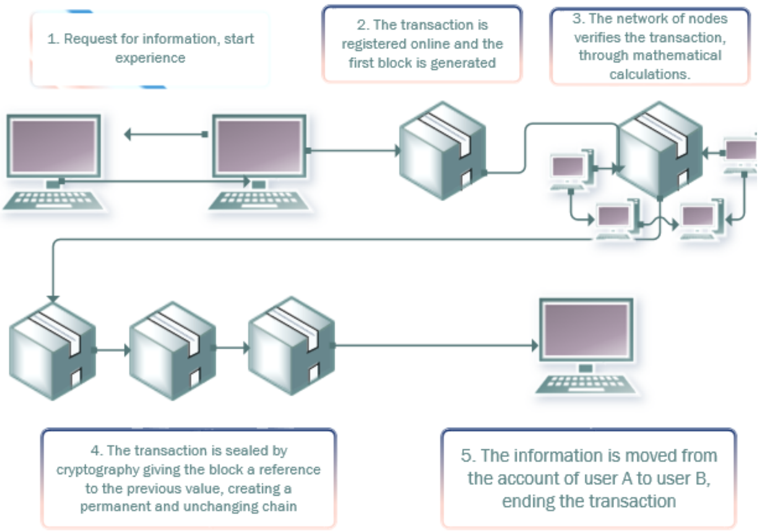

1. Request for information, start experience
2. The transaction is registered online and the first block is generated
3. The network of nodes verifies the transaction, through mathematical calculations.
4. The transaction is sealed by cryptography giving the block a reference to the previous value, creating a permanent and unchanging chain
5. The information is moved from the account of user A to user B, ending the transaction

Fig. 1. Information transfer by using blockchain.

When the transaction is sent from A to B, as you can see on Fig. 1, a private key is granted to A and B "the receiver" is decrypted with the sender's public key [10] giving the opportunity that the individual can prove his property in anonymity [11].

In this way, the information is certain that it will not be lost unless all nodes are destroyed or more than 51% of the nodes of the entire network are compromised. Same, that elaborates a unique and unrepeatable hash in each block that must coincide with the hash of the previous block thus guaranteeing the reliability of the transferred information.

The loss of any node does not affect the operation of the entire system, due to its distribution design. Being the collective participation a decisive factor that guarantees transparency, to ensure the correctness and security of the transaction, reassuring that human intervention has no effect on the system.

Since the accounts remain synchronized in real time, and data that has been validated and registered cannot be manipulated [12]. It should be mentioned that the node must be certified to join the consensus process [13].

2.2 Hash

Involving the hash function that transforms an input of arbitrary size into an output of fixed length of n bits, being unidirectional the hash cannot be altered. Therefore, with this technology we can solve the problem of fraud among the personnel working in the company by providing an information base so that the analysis in the decision-making process is more precise for each part of the process [14]. Which leads to a decision making in an assertive, timely and correct manner. Corroborating the benefits incorporated by using the hash algorithm and cryptography in data transfer.

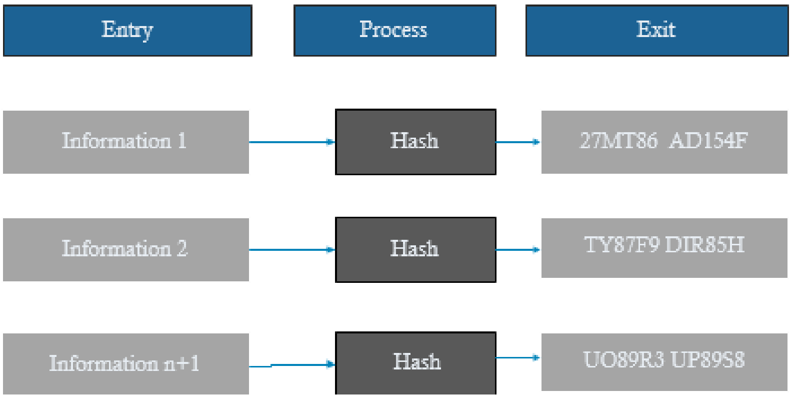


Fig. 2. Hash Function.

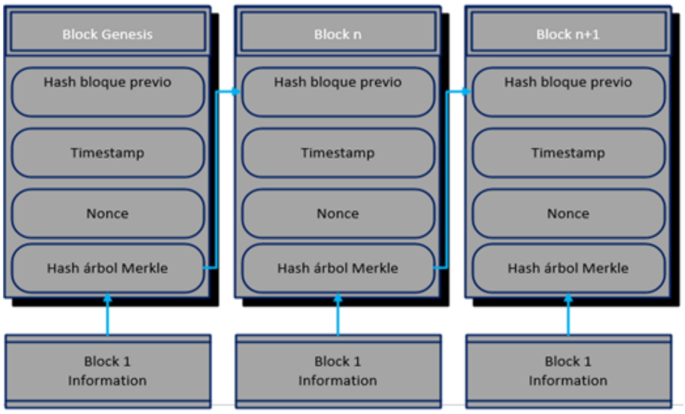


Fig. 3. Blockchain block structure.

The first block is known as the genesis block; it is shown in Figure 2. A block consists of a header and a body [15]. The body of the block contains the list of transactions [16]. The block header contains several fields, mainly the Block Version that indicates the set of rules that must be followed for validation, a hash of the previous block header, a timestamp, one hash root of the Merkle tree represents the hash value of all transactions in the block [17] as we can see in Figure 2.

The nonce is a 32-bit field that increases until the equation is solved [18]. In addition to be a distributed master book, blockchain is also defined by three key concepts: consensus, smart contract and cryptography [19]. When making a transaction, smart contracts are invoked to execute the term of a contract / procedure on each node in the network [20].

In blockchain, the block structure is made up of block header and block body, andthe hash values of the business data are gradually divided into pairs and form the Merkle tree structure.

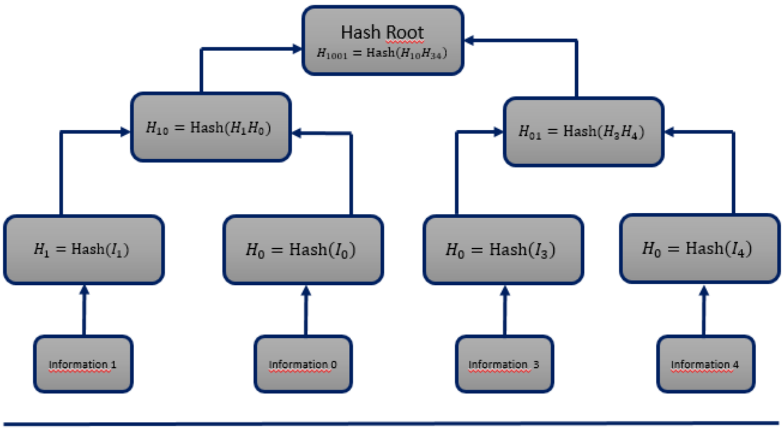


Fig. 4. Merkle Hash Tree.

The leaf nodes are the initial hash values of the business data, and the body of the Merkle tree is stored in the body of the block, and the root of the tree is stored in the header of the block. Commercial data is a mapping relationship built with the leaf nodes of the Merkle Tree, and commercial data can be stored in the body of the block [21].

In a centralized system, the administrator could be bribed, and the entire system could be subject to manipulation and falsification of information [22]. With the structure of the blockchain this risk is avoided.

3 Proposal of the Model

For a large production company, the endogenous risk of its supply chain can be divided into two parts: the credit risk caused by the asymmetry of information between the companies within the supply chain and the risk caused by incomplete information [23]. Therefore, the main objective of the model is to create certainty in the information that reaches the company and is transmitted by the different departments involved in the process of "recuse for quote". Being an improvement plan for assertive decision making based on transparency in information, and offering managers relevant information for decision making and providing information to third parties about the value of the company or organization.

Figure 5 shows the interaction of departments A, B and C, where A sends information to B, B cannot modify it, but can add new information and transfer it to C, where in turn C can add information, but not modify the previous one, since it is protected by the Hash cryptology.

Creating a reliable ecosystem between suppliers and their customers [24]. This is achieved through a policy that focuses on the transparency of the chain to ensure the traceability of the product, where accurate data collection and secure storage of them is required. Suppliers and the company must be synchronized by the use of the Etherium

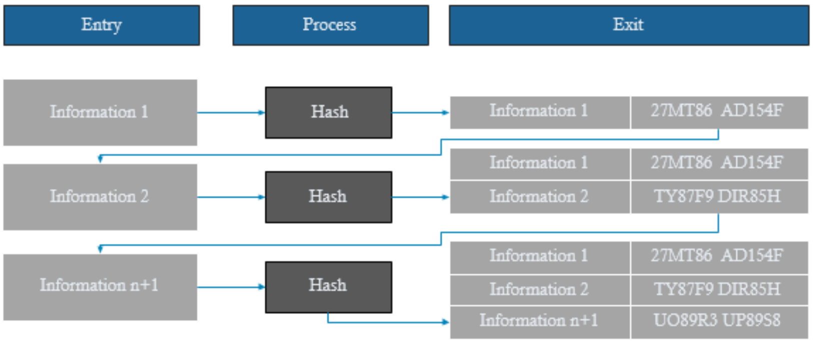


Fig. 5. Incorporation of several processes in the information network chain, using blockchain.

system, in which it will be used to transfer the information to the different departments involved.

3.1 Model Description

a) When accessing the information to the system, the provider protects its information through a hash, which will be kept during the reception of each department, so that no department can modify the information received.

b) The private key is protected by a cryptographic coprocessor.

c) The departments involved that require access to the information and add new content, will be able to do so since this is where the innovation arises to the model, because with the previous model the information cannot suffer any alteration in its trajectory.

d) Data is collected and transferred by each department involved in the estimation process, where they will be safe and reliable before being stored in an immutable and decentralized database.

e) In the transition between blocks, the next block will copy the previous one having the previous information content blocked, however, you can add new block information, quickly generating virtual data according to the recovery requirements and with the content of the last block as input .

f) The blocks are chained to each other by the immediately preceding block. In this way, any modification that was intended to be introduced at some point in the chain would affect the hash and only the hash of the last block would have to be verified to detect where the anomaly happened.

g) The hash is created by mathematical formulas, which make it unrepeatable.

h) Each block may have the information of the previous block without the right to write about the part of the information, however, it will have its own space to add information from its department, which will go to the next block, where the process is repeated , until culminating, providing a system of authentication and storage of information.

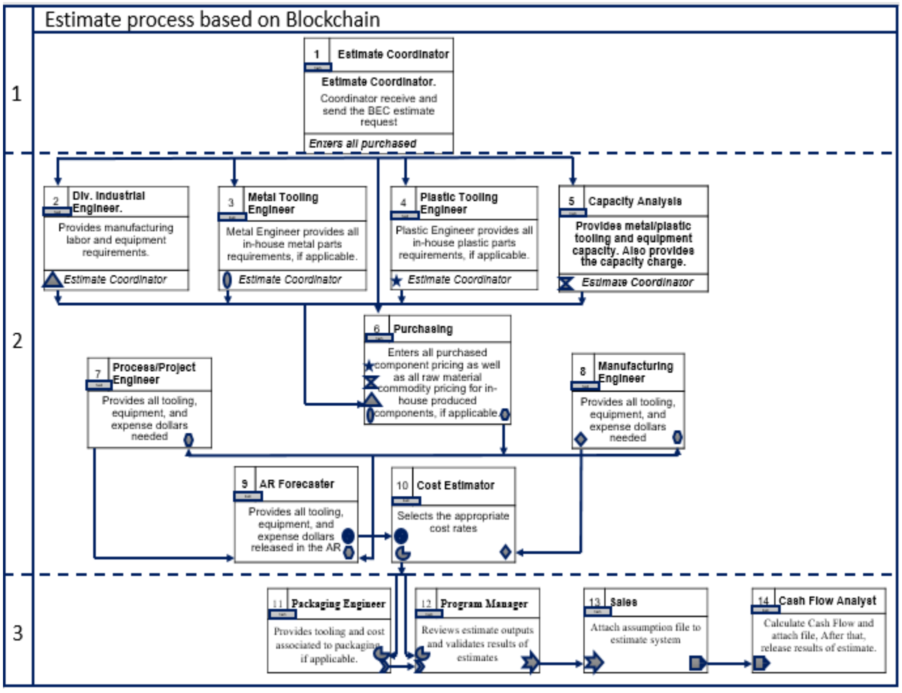

Estimate process based on Blockchain
1
2
3
1 Estimate Coordinator
Estimate Coordinator.
Coordinator receive and send the BEC estimate request
Enters all purchased
2 Div. Industrial Engineer.
Provides manufacturing labor and equipment requirements.
Estimate Coordinator
3 Metal Tooling Engineer
Metal Engineer provides all in-house metal parts requirements, if applicable.
Estimate Coordinator
4 Plastic Tooling Engineer
Plastic Engineer provides all in-house plastic parts requirements, if applicable.
Estimate Coordinator
5 Capacity Analysis
Provides metal/plastic tooling and equipment capacity. Also provides the capacity charge.
Estimate Coordinator
6 Purchasing
Enters all purchased component pricing as well as all raw material commodity pricing for in-house produced components, if applicable.
7 Process/Project Engineer
Provides all tooling, equipment, and expense dollars needed
8 Manufacturing Engineer
Provides all tooling, equipment, and expense dollars needed
9 AR Forecaster
Provides all tooling, equipment, and expense dollars released in the AR
10 Cost Estimator
Selects the appropriate cost rates
11 Packaging Engineer
Provides tooling and cost associated to packaging if applicable.
12 Program Manager
Reviews estimate outputs and validates results of estimates
13 Sales
Attach assumption file to estimate system
14 Cash Flow Analyst
Calculate Cash Flow and attach file, After that, release results of estimate.

Fig. 6. Estimate process based on Blockchain.

i) What guarantees the construction of the information storage mechanism is to improve the storage efficiency and privacy protection of this. The large chain of information network of the company based on blockchain depends on a private industrial network and the Internet, as well as the staff of the different departments affiliated with the blockchain network with prior authorization, avoiding interference to the users system that are not relevant by reducing the risk of malicious users falsifying information in the system together.

Applying the model in the manufacturing industry we can appreciate the structure presented in Figure 6, which shows the relationship that exists between the different departments involved in the estimation process.

As we can see, companies are structured by different specialized departments, which are involved in the transfer and use of information. Its objectives are to optimize its production processes, increase sales and achieve profitable and sustainable growth.

As Figure 6 shows, the initial information reaches a different department, and can be altered by them. The image shows how hash cryptography is implemented since the information reaches the company, along the way, the information can be easily tracked through the code and verifies where the modification was made.

The difficulty can be changed by increasing or reducing the number of zeros in front of the hash of the block header on which the hash function is then applied. An additional figure is shown from each department, which represents the cryptograph, where each hash must match the next block. With which the correlation of businesses of different

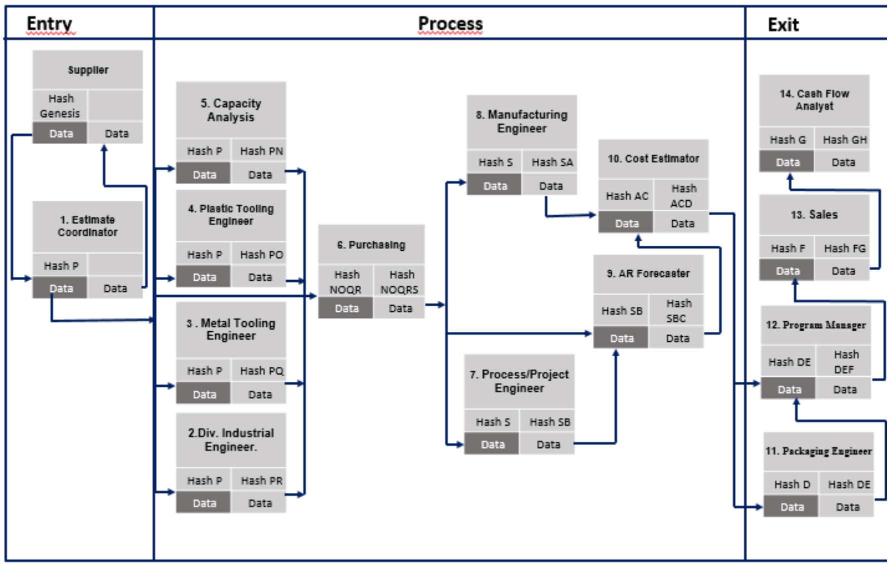


Fig. 7. Model for the Estimate process based on Blockchain.

subjects in ascending and descending sense can be detected constructing their relations of interaction of information.

Demarcating the fellows that do not have commercial interaction relationships or any connection, so that they cannot understand the information of other fellows, whose relationship of information interaction of a section [23].

4 Simulation Methodology

A centralized management system represents a threat to data integrity, availability and resilience, because the system is subject to corruption, fraud and manipulation [25]. Therefore, the distributed information technology Blockchain can help maintain transparency and reliability in the information flows in the different departments, generating a reliable system, under the following work scheme where the data received is unalterable. Due any modification in a specific block will invalidate all subsequent blocks [26].

Considering that the communication between the front end and the blockchain is done through the HTTP server on the representative state transfer (REST) API, using JSON to encode and decode requests and responses [27]. The model proposed in Figure 7 is shown, which is composed of different departments involved in an estimation process. In the first instance, department 1 asks the suppliers for the information, which is not encrypted by hash, since the suppliers must generate the information.

Once such information returns to department 1, it will be encrypted by hash P, which will be transferred to departments 2, 3, 4, 5 and 6, where the new hash will be the union of some digits of the previous plus the new remaining for 2PR, 3PQ, 4PO and 5PN who

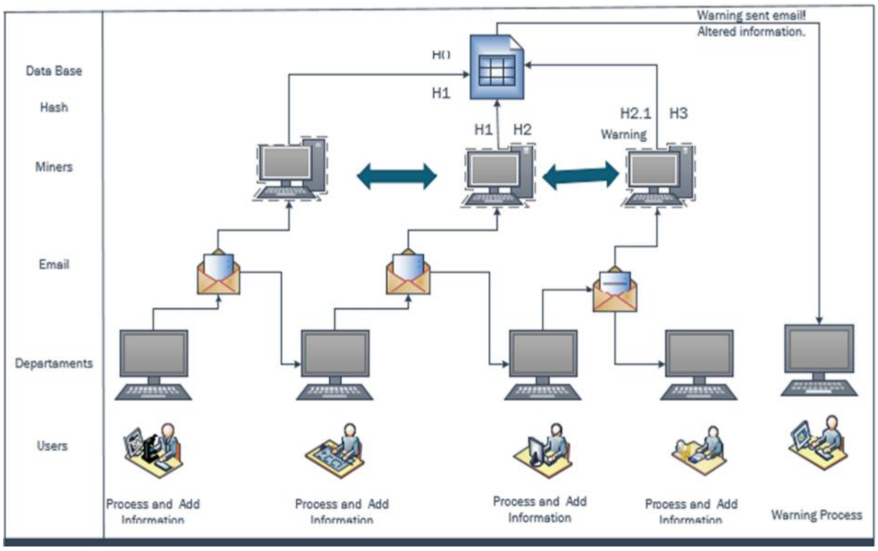


Fig. 8. Model for the Estimate process based on Blockchain.

cannot modify the information received in "P", but having enabled the part of the system that corresponds to add their data, "R", "Q", "O" and "N "Respectively, they will also send their encrypted information, in addition to the encrypted information that they had previously received, in this case to department 6 who has the hash S, to which" N, O, Q, R "is added. Department 6 transfers the "S" to departments 7, 8 and 9, which generate their own hash to encrypt their information and send it so on until the information is transmitted by all the departments involved.

The model allows establishing a virtual security environment with the incorporation of the different departments, through a set of mathematical equations that guarantee the integrity of the information in the decision-making process in the company or a specific area of it.

The promise he makes us implies a future in which no one has absolute power in the network, and no one can lie about past or present events [28], providing full certainty in decision making.

The proposal of the model focuses on the use of Ethereum, as a blockchain-based encryption model, however this has a limited capacity in handling a large amount of data, which would generate the need to investigate about the BigchainDB for manipulation of large amounts of information in less time.

The evaluation of Hyperledger's legation that has a better performance. Within the investigation it was also found that terms of security attack, Ethereum and Parity are vulnerable exposing the system to double spending attack [29]. However, Ethereum is still considered the most advantageous in terms of scalability, reliability and maturity of the system, which allows a large number of participants [30].

5 Conclusions and Future Research

Having this clear how vulnerable the manufacturing sector is to fraud, the implementation of *blockchain* technology would generate an immutable tamper-proof record for the departments involved. Being the objective of these functions to be able to detect if a message has been modified or not, to verify the integrity of the message. Blockchain reduces fraud, errors and delays identified in the supply chain ecosystem, increases trust between the customer and the provider in data management, *blockchain* was integrated, to ensure availability, accuracy and accessibility of data for all the chain, improving business decisions and providing insight into all system vulnerabilities [29].

This leads to an efficient implementation of the consensus protocol to improve the growth of the economy, ensuring the correct functioning of the *blockchain* and avoiding any malfunction of the *blockchain* architecture [31]. Having a correct manipulation of the use due to the information, through the different departments will contribute reliability of this. It is worth mentioning that a disadvantage in that *blockchain* has a weight restriction in the information and transfer of it.

However, to avoid the low performance of the *blockchain*, dual storage architecture can be implemented to handle a large amount of data. The communication protocol based on data rate, communication range, energy consumption, and cost should also be considered.

6 Future Research

The implementation of the *blockchain*-based computer security model is a long process of adaptation, modification and awareness of the different departments and therefore personal in charge of what requires a culture of collaboration. Modifying the organizational culture based on previously established processes, generates a huge challenge in the industry, and more if some processes have been flawed over the years. Therefore, it is essential to have the participation of both management levels and key personnel of the different departments.

What makes this project an opportunity for improvement, ensuring reliable information for assertive decision-making based on achieving the company's objectives. The next process is to test the model regarding the veracity, speed and performance of the system using blockchain in the company, simulating the processes see the Figure 8 to evaluate the model.

References

1. Fu, Y., Zhu, J.: Big production enterprise supply chain endogenous risk management based on blockchain. IEEE Access, 7, pp. 15310–15319 (2019)
2. Mire, S.: Blockchain for manufacturing (2018)

3. Abeyratne, S., Monfared, R.: Blockchain ready manufacturing supply chain using distributed ledger. International Journal of Research in Engineering and Technology (2016)
4. Kshetri, N.: Blockchain’s roles in meeting key supply chain management objectives. International Journal of Information Management (2018)
5. Meijer, D., Carlo, R.W.: The UK and blockchain technology: A balanced approach. Journal Payments Strategy Systems, 9 (2016)
6. Yang, D., Pan, Z.D.: The blockchain bring finance and law optimization. China Finance (2016)
7. Wang, J., Lingchao, G., Aiqiang, D., Shaoyong, G., Hui, Ch., Xin, W.: Block chain based data security sharing network architecture research. Journal of Computer Research and Development (2016)
8. Nakamoto, S.: Bitcoin: a peer-to-peer electronic cash system (2016)
9. Daming, L., Zhiming, C., Lianbing, D., Xiang, Y., Harry, H.: Wang information security model of block chain based on intrusion sensing in the IoT environment. Cluster Computing, 22(Z1), pp. 1−18 (2018)
10. Ping, Z., Yu, D., Bin, L.: White paper on China's blockchain technology and application development. (2016)
11. Zheng, Z., Xie, S., Dai, H., Wang, H.: Blockchain challenges and opportunities: A survey. Researchgate (2016)
12. Yonggui Zhu, Jianming Zhu: Big production enterprise supply chain endogenous risk management based on blockchain. IEEE Access, 7, pp. 15310-15319 (2019)
13. Azzia, R., Kilany, R., Chamouna, M.S.: The power of a blockchain-based supply chain. Computers & Industrial Engineering, 135, pp. 582−592 (2019)
14. Pierro, M.: What is the blockchain? Computing in science and engineering (2017)
15. Zheng, Z., Xie, S., Dai, H., Wang, H.: Blockchain challenges and opportunities: A survey. Researchgate (2016)
16. Gupta, M.: Blockchain for dummies. 2nd IBM limited edition. IBM (2018)
17. Christidis, K., Devetsikiotis, M.: Blockchains and smart contracts for the internet of things. IEEE Access (2016)
18. Yonggui, Z.: Jianming big production enterprise supply chain endogenous risk management based on blockchain. (2019)
19. T. Anh, D., Ji, W., Gang, C., Rui, L., Beng-Chin, O., Kian-Lee, T.: Blockbench: A framework for analyzing private blockchains. In: Proceedings of the ACM international conference on management of data (2017)
20. Yonggui Fu, Jianming Zhu: Big production enterprise supply chain endogenous risk management based on blockchain. IEEE Access, 7 (2019)
21. Bocek, T., Stiller, B.: Smart contracts-blockchains in the wings. Digital marketplaces unleashed, pp. 169–184, Springer (2018)
22. Moreno, J., Serrano, M.A., Fernández-Medina, E.: Main issues in big data security. Future Internet, 8(3), pp. 44 (2016)
23. Pincheira, C., Salek-Ali, M., Vecchio, M., M., Giaffreda, R.: Blockchain-based traceability in agri-food supply chain management: A practical implementation. In: IoT vertical and Topical Summit on Agriculture-Tuscany (2018)
24. Sankar, L.S., Sindhu, M., Sethumadhavan, M.: Survey of consensus protocols on blockchain applications. In: Advanced Computing and Communication Systems (2017)

ISSN 1870-4069

N-gramas de sílabas vs. n-gramas de caracteres para la tarea de atribución de autoría en un corpus multi-tema

Héctor Javier Hernández[1], Hiram Calvo[1], Eduardo López[2],
Juan Pablo Posadas-Durán[2],
Ilya Markov[3], Grigori Sidorov[1]

[1] Instituto Politécnico Nacional,
Centro de investigación en Computación,
México
[2] Instituto Politécnico Nacional,
Escuela Superior de Ingeniería Mecánica y Eléctrica Unidad Zacatenco,
México
[3] University of Amsterdam,
Holanda

hhernandezs1203@alumno.ipn.mx,
hcalvo@cic.ipn.mx,lopezmareseduardo@gmail.com,
jposadasd@ipn.mx, imarkov@uva.nl, sidorov@cic.ipn.mx

Resumen. Los caracteres de n-gramas como marcadores de estilo han demostrado un buen desempeño para la tarea de atribución de autoría [5]; sin embargo, es difícil dar una interpretación específica de la información que estas características están capturando. En este trabajo se plantea la comparación de n-gramas de caracteres y n-gramas de sílabas como marcadores de estilo para comprobar qué tanto pueden ayudar estas últimas a identificar correctamente la autoría. Partimos de la hipótesis de que los n-gramas de caracteres capturan cierta información morfológica de las palabras, y que por ello el uso de sílabas podría ser equivalente. Realizamos experimentos con diversas características, y encontramos que el desempeño de las sílabas está cercano al que presentan los n-gramas de caracteres, si bien los caracteres permiten realizar la tarea con un mejor desempeño.

Palabras clave: N-gramas de caracteres, n-gramas de sílabas, atribución de autoría.

N-grams of Syllables vs. N-grams of Characters for the Authorship Attribution Task in a Multi-topic Corpus

Abstract. N-gram characters as style markers have yielded a good performance for the task of authorship attribution [5]; however, it is difficult to give a specific interpretation of the information that these characteristics are capturing. In this work, the comparison of character n-grams and syllables n-grams are proposed as style markers to verify to which extent the latter help to correctly identify authorship. We start from the hypothesis that character n-grams capture certain morphological information from words, and that, because of this, the use of syllables could be equivalent. We conducted experiments with various characteristics, and found that the performance of the syllables is to that of character n-grams for this task, although characters allowed to perform the authorship attribution task with better performance.

Keywords: Character n-grams, syllable n-grams, authorship attribution.

1. Introducción

La atribución de autoría es la línea de investigación que se dedica al problema de identificar el autor de un documento dentro de un conjunto de posibles autores. En esta área, una de las características que ha demostrado obtener mejores resultados han sido los n-gramas de caracteres [1].

A pesar de que los n-gramas de caracteres no corresponden a una característica propia de la lingüística, su eficacia para resolver el problema de atribución se puede explicar en el hecho de que ciertos n-gramas de caracteres corresponden a elementos como sílabas, prefijos o sufijos. Es así que el propósito de esta investigación es determinar si el buen desempeño de los n-gramas se debe a que capturan esta información morfológica. Para ello, realizaremos una comparación entre ambos tipos de características.

Otra estrategia usada frecuentemente en la solución de problemas relacionados con el Procesamiento de Lenguaje Natural es la denominada frecuencia de palabras que consiste en obtener la frecuencia de ocurrencia de cada palabra en un documento. La estrategia presenta la desventaja de que puede ser que las palabras no siempre estén escritas de la misma manera [1], por lo cual las inflexiones que corresponden a una palabra se consideran tan diferentes como dos palabras no relacionadas entre sí.

En este trabajo se hace una evaluación de los n-gramas y de las sílabas como marcadores de estilo en un problema de atribución de autoría usando un corpus en español que incluye diferentes tópicos; además se propone y evalúa un método basado en el uso de las siguientes características: n-gramas de caracteres filtrados y sílabas. En la sección 2 presentamos algunos trabajos relacionados con esta propuesta, así como la base de la clasificación de n-gramas de caracteres; en la sección 3 presentamos nuestra propuesta; en la sección 4 presentamos el corpus construido; en la sección 5 presentamos algunos experimentos; en la sección 6 nuestros resultados, y finalmente en la sección 7 nuestras conclusiones.

2. Trabajos relacionados

Los n-gramas de caracteres han demostrado ser exitosos en la tarea de atribución de autoría aunque, como se explica en [1], los n-gramas son probados bajo condiciones controladas, es decir, en donde los textos de un corpus pertenecen a un mismo tema o dominio; sin embargo, en un escenario más real, los textos pueden tener autores que escriban de uno o más temas.

Tabla 1. Categorías de sílabas.

Categoría	Subcategoría	Definición
Afijo	Prefijo	Primera sílaba de una palabra.
Afijo	Sufijo	Última sílaba de una palabra.
Palabra	Palabra completa	Palabras de una sola sílaba
Palabra	Palabra al centro	Las sílabas que se encuentren al centro de una palabra
Palabra	Multi palabra	Contiene la última sílaba de una palabra y la primera de la siguiente, siempre y cuando las dos palabras tengan más de una sílaba.
Puntuación	Puntuación inicial	Bigrama compuesto por: signo de puntuación + sílaba
Puntuación	Puntuación central	Trigrama compuesto por: sílaba + signo de puntuación + sílaba.
Puntuación	Puntuación final	Bigrama compuesto por: sílaba + signo de puntuación

En trabajos previos, los n-gramas de caracteres son tratados de la misma manera, sin importar la estructura o la posible información morfológica/sintáctica que estos n-gramas nos puedan proporcionar. No es hasta que en [2] se propone una clasificación de n-gramas de caracteres, y el resultado de esta clasificación muestra que hay n-gramas en específico que parecen aportar más información y por tanto mejorar el desempeño de la clasificación. En este trabajo usaremos la misma clasificación, con la diferencia de reemplazar los n-gramas de caracteres por n-gramas de sílabas.

3. Sílabas como marcadores de estilo

Se propone usar las sílabas como marcadores de estilo, además de esto se propone seguir la clasificación de n-gramas de caracteres propuesta en [2], pero aplicada a los n-gramas de sílabas.

3.1. Extracción de sílabas

La división silábica *per se* es un problema complejo en cualquier idioma, en este trabajo se limita a implementar soluciones existentes a este problema. Para poder obtener las sílabas de una palabra empleamos el siguiente proceso para obtener la división silábica de las palabras, como resultado de una búsqueda, obtuvimos un diccionario de división silábica del sitio web [4], el cual fue la primera fuente de consulta para obtener una división correcta, si la palabra no se encuentra en la

Tabla 2. Documentos que componen el corpus CCAP-s1.

Autor	**Cine**	**Comida**	**Fotografía**	**Arte**	**Estilo de vida**	**Total**
Alejandro Arroyo C.	52	16	16	9	34	149
Alejandro López	51	16	17	12	54	181
Andrea Méndez B.	26	5	21	44	33	147
Daniela Fernandez	6	4	9	5	14	51
Iván Montejo	14	7	4	8	12	58
Mafer Fernández L.	9	2	7	12	15	57
Total por categoría	**158**	**50**	**74**	**90**	**162**	**643**

colección, entonces se empleamos una librería para python llamada *pyphen* [3] la cual está bajo la licencia GPL 2.0+/LGPL 2.1+/MPL 1.1 tri-license.

4. Corpus

Para el desarrollo de este proyecto, se necesita usar un corpus que refleja el uso del lenguaje en español en escenarios reales. Una de las características importantes que este recurso debe cumplir es que debe de ser multi-tema; es decir, los diferentes autores deben de tener documentos de diferentes temas; por ejemplo, el autor uno debe tener documentos de las categorías "A","B", "C" y "D", de igual forma el resto de los autores que integran el corpus. Esta característica presento un reto complicado ya que en un primer momento se intentó obtener el corpus de periódicos nacionales y de revistas, sin embargo, surgió el problema de que la gran mayoría de los autores (periodistas) de estos medios, sólo se centraban en una categoría, por lo que estos medios de comunicación no pudieron ser utilizados. Como resultado de la continuación de la búsqueda, se determinó utilizar el sitio web llamado "Cultura Colectiva", el cual es una plataforma de comunicación de contenido original en español; y contiene publicaciones de un gran número de autores y variedad de temas como: cine, comida, arte, diseño, historia, entre otras. La importancia de un corpus multi-tema estriba en que, al realizar los experimentos en un corpus de este tipo, representa un escenario más realista, ya que, en situaciones reales, los factores como número de autores, número de documentos por autor, y el tema de cada documento puede variar, y el objetivo de este trabajo es demostrar que se puede resolver esta tarea a través de un análisis de información morfológica, en un entorno más complejo y desbalanceado. El corpus recopilado (CCAP) consta de 4,134 documentos. Estos documentos pertenecen a una colección de 11 categorías diferentes, y 21 autores diferentes, para este trabajo se tomó la decisión de acotar este corpus a solo 643 textos, de 6 autores, y 6 temas diferentes (CCAP-s1). El detalle de la distribución de dichos textos en CCAP-s1 puede apreciarse en la Tabla 2. Ambos corpus pueden encontrarse en http://idic.likufanele.com/~ccap.

5. Experimentos

En este trabajo se propone un escenario más real al entrenar SVM usando un conjunto de textos pertenecientes a una categoría y haciendo pruebas con textos de otra categoría, en este trabajo se han probado diferentes combinaciones de características, pero solo nos enfocaremos en las que presentaron mejores resultados.

5.1. Desambiguación de los n-gramas de caracteres y de sílabas

Dentro de un conjunto de n-gramas de caracteres existe la posibilidad de que un n-grama pertenezca a más de una categoría [2], por ejemplo el bigrama *de*, se puede considerar como un n-grama que bien podría estar dentro de la categoría de *n-grama como palabra completa*, sin embargo, este bigrama también lo podemos encontrar al inicio, al centro o al final de una palabra (*destello*, *clandestino*, *jade*), es por esto que en este trabajo se emplea el concepto de n-gramas *etiquetados* y *sin etiquetas*; con esto pretendemos desambiguar el sentido de un n-grama, y pretendemos identificar los n-gramas que puedan brindar información más relevante. Así mismo aplicamos esta misma idea para el uso de sílabas como marcadores de estilos, una silba podría pertenecer a más de una categoría.

5.2. Evaluación de los métodos propuestos

Para la evaluación del método propuesto, utilizamos la precisión de la predicción que el modelo hace; Los experimentos se realizaron entrenando máquinas de vectores soporte usando WEKA con su configuración por defecto.

Para todos los experimentos usamos n-gramas de caracteres con n=3, la razón de esta decisión está basada en los resultados de que se reportan en trabajos previos [2].

5.3. Configuración para el corpus cultura colectiva

La configuración para los experimentos con el corpus de cultura colectiva consistió en tomar los documentos de una categoría y hacer pruebas con otra, cabe aclarar que los conjuntos de entrenamiento y de pruebas no están estrechamente correlacionados, ya que el lenguaje como el contexto será diferente. Al tratarse de 6 diferentes categorías, se realizaron 30 combinaciones diferentes, de entrenamiento-prueba, de estas combinaciones se presentan la que dio un mejor resultado.

6. Resultados

El corpus en español que representa un escenario mucho más interesante y complejo, el esquema que se siguió, en esencia es el mismo, se tiene que dividir el corpus en una sección para el entrenamiento y otra para las pruebas, más sin embargo, lo interesante de estos experimentos estriba en que, los conjuntos de entrenamiento y pruebas no estarán estrechamente relacionados, es decir, el uso del lenguaje en dichos conjuntos será distinto, ya que ambos pertenecen a diferentes temas, lo interesante es poder observar, aquellas características que prevalecen y otorgan información sobre el estilo

Tabla 3. N-gramas vs. sílabas (P=Porcentaje de precisión).

Tipo de Característica	**Características sin etiquetar**		**Características etiquetados**		**Combinación de afijos + puntuación**	
	P	característica s	P	características	P	características
n-gramas	56.0	4,913.7	56.3	5734.7	56.5	30,047
sílabas	54.2	5,801.7	54.8	6419.8	52.9	1821.3

que una persona tiene al escribir, sin importar el tema del que esté hablando. Para estos experimentos al tratarse de 6 categorías se realizaron las 30 posibles combinaciones de entrenamiento-prueba, es decir, arte vs. cine, arte vs. comida, arte vs. diseño, arte vs. estilo, arte vs. fotografía, y viceversa, para finalmente obtener el promedio de éstas. Para los experimentos con n-gramas de caracteres empleamos el siguiente set de características: (*Prefijos, Sufijos, Espacio + prefijo, Espacio + sufijo*) y para los experimentos con sílabas empleamos el siguiente set de características (*Prefijos, Sufijos, Multi palabra, Toda la palabra, Palabra al centro*). A continuación, se muestran los resultados obtenidos para este corpus.

7. Conclusiones y trabajo a futuro

Los n-gramas basados en caracteres han demostrado tener un buen desempeño para la tarea de atribución de autoría, y tal parece que parte de este éxito se debe a que estos n-gramas están capturando información morfológica de las palabras lo cual se ha podido mostrar parcialmente mediante los experimentos presentados en este trabajo, sin embargo, este trabajo presenta varias áreas de oportunidad, como mejorar la manera en la que se obtienen las sílabas, ya que no podemos obtener una división silábica que sea cien por ciento precisa, y al no tener la división, perdemos información.

Como trabajo futuro se plantean las siguientes acciones:

- Robustecimiento de recursos léxicos para el idioma español,
- Plantear un algoritmo o metodología más eficiente para la división silábica,
- Probar diferentes algoritmos de aprendizaje automático para resolver esta tarea.

Referencias

1. Stamatatos, E.: On the robustness of authorship attribution based on character n-gram features. Journal of Law and Policy, 21, pp. 421–439 (2013)
2. Sapkota, U., Bethard, S., Montes, M., Solorio, T.: Not all character N-grams are created equal: A study in authorship attribution. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 93–102 (2015)
3. Kozea, A.: Community project. Pyphen de A Kozea Community Project Sitio https://pyphen.org/ (2008)

4. OSLIN: Diccionario de división silábica. OSLIN https://web.archive.org/web/20160203184525/es.oslin.org/syllables.php. (2015)
5. Koppel, M., Schler, J., Argamon, S.: Computational methods in authorship attribution. JASIST, 60, pp. 9–26 (2009)

ISSN 1870-4069

Análisis e implicaciones de la detección del humor en textos

Victor Manuel Palma-Preciado, Grigori Sidorov, Alexander Gelbukh

Instituto Politécnico Nacional, Centro de Investigación en Computación,
México

{victorpapre, gr1965}@gmail.com,
{gelbukh}@gelbukh.com

Resumen. El presente trabajo tiene como objetivo analizar el fenómeno de la detección de humor en textos y aquellas implicaciones que conlleva, como puede ser el preparar los datos para tener un análisis correcto del set tomado, así mismo algunas metodologías actuales, todo esto tomando como base una selección de trabajos del periodo 2018-2019 los cuales hablan de la detección de humor, creación de conjunto de dato y talleres de detección de humor. Los trabajos fueron seleccionados dado el interés y los resultados que obtuvieron los autores, ya que entre ellos hay algunos con un enfoque más vanguardista. Como resultado, se obtuvo un análisis de los trabajos elegidos. Se puede observar que los métodos para detectar diferentes tipos de humor, aun teniendo la misma tarea por resolver, pueden ser variados y arrojar resultados competitivos uno frente a otros.

Palabras clave: Humorismo, detección de humor, conjuntos de datos, bromas.

Analysis and Implications of Humor Detection in Texts

Abstract. The paper aims to analyze a part of the phenomena of humor detection in texts and its implications, such as preparing the data to have a correct analysis of the set taken, as well as current methodologies. The analysis is based on a selection of works from the period 2016-2019, which talk about humor detection, creation of data sets and humor detection workshops. The works were selected given the interest and the results obtained by the authors, since among them there are some with a more avant-garde approach. As a result, an analysis of the chosen works was obtained. It can be seen that the methods for detecting different types of humor, even having the same task to be solved, can be varied and yield competitive results against each other.

Keywords: Humorism, humor identification, datasets, jokes.

1. Introducción

Muchas veces es difícil como humano identificar el humor, ya que a menudo el humor es subjetivo, dado que nos permite expresar un rango amplio de emociones, normalmente es difícil para las personas identificarlo, ahora extrapolando la misma tarea, pero aun nivel computacional, se puede llegar a decir que es difícil identificar diferentes tipos humor y esto tiene cierto sentido, ya que el contexto nos dirá si por ejemplo, es una frase irónica o sarcástica, también de la misma forma si una frase intenta demostrar doble sentido o solo es una frase normal sin esa intención.

Todos esto problemas están presentes a la hora de identificar las intenciones tras de una frase y más aún cuando se tienen variantes de la lengua tomando por ejemplo los países de habla hispana donde ciertas palabras pueden tener un significado diferente o de connotación sexual que podría ser mal interpretado. Por lo tanto, se busca un modelo que nos permita identificar de manera positiva, si algo es gracioso o no y también que tan gracioso puede llegar a ser. Dado que siempre es bueno tener una noción general del humor y sus características, esta vez se analizarán textos concernientes a la creación de corpus y la detección del humor, por lo tanto, de este trabajo se puede esperar un análisis breve de ello.

La metodología usada se centró en compilar información relevante para la tarea de detección de humor en texto, en general se tomaron artículos y secciones de revistas así mismo algunos trabajos de Workshops donde el objetivo fuera la detección de humor. Seleccionado para el periodo 2016-2019, aunque solo se utilizó un trabajo del 2016 se deicidio considerar de estar forma el periodo de esta forma.

2. El concepto del humor

El humor tiene un rol importante en la sociedad, en este sentido [2], hablan de que el humor es una actividad comunicativa altamente inteligente, que puede provocar risas o asombro, pero sin la menor duda se puede decir que es un fenómeno sociológico. Si las máquinas llegarán a entender el idioma hasta cierto punto, les sería más fácil predecir la intención humana, lo que traería consigo una mejora en la capacidad que existe en la interacción humano-máquina. Sin embargo, [12], plantean la existencia de un problema en la detección del humor para el procesamiento del lenguaje natural [12]. La detección del humor a pesar de ser un fenómeno completamente conocido en la lingüística, aún presenta cierta carencia en las tareas computacionales [5,10].

Por lo tanto, el entender el humor se vuelve una tarea complicada al tener diferentes usos en la lengua, entre ellos los de tipo figurativo, metáforas, ironía y sarcasmo que son aspectos de la misma comunidad en la que vivimos y, por lo tanto, forman parte del acervo cultural escrito [3] Estos conceptos se utilizan de diferentes formas ingeniosas para expresar distintas ideas y esto pasa con mayor frecuencia en las redes sociales, en donde las frases o sus construcciones son más difíciles de entender, la creatividad y la forma en que se construye el humor permite visualizar las características personales de quien lo construye, por lo tanto, es importante crear nuevos métodos y recursos para analizar el humor en los textos [9].

3. Líneas base

Las líneas base (*baselines*) se definen como la base de comparación y se utilizan para evaluar el desempeño a través del tiempo. En el lenguaje natural cada autor toma consideraciones diferentes para sus líneas base, algunas de las líneas base utilizadas como el caso de [1], tomaron dos consideraciones: en la primera, se consideró que la línea base fuera bajo una posibilidad del 50% de que el *tweet* se clasificará como humor o no y la segunda tuvo en cuenta la revisión del corpus, en el que la mayoría de los tweets humorísticos iniciaban con un guion, lo cual le daba precisión, pero con una exhaustividad (recall) menor, con una dificulta al identificar entre diferentes tipos de humor, como limitante de su modelo.

En el experimento de [9], se utilizó una escala para cada tweet del uno al cinco, en el cual, uno era el no gracioso y cinco como el más gracioso entre ese rango de números. Para medir los resultados se utilizó *Root Mean Squared Error* (RMSE). La *baseline* tomada fue justo en la mitad con tres puntos sobre el texto y su *baseline* sobre los datos de 1.14 RMSE. Todo esto bajo un corpus de 12,000 *tweets* en español, aunque se tuvieron diferentes variaciones del español esto no fue un impedimento para un post procesamiento de tweets consideradas de baja calidad, con las bases anteriores las tareas desarrolladas en el taller de *HAHA task* [1].

Mientras que las baseline usados por [11], utilizaron *Random Forest* con Word2Vec (W2V) adicionado de *Human Centric Feature* (HCF), optando por un *drop rate* de 0.5 y para testear el desempeño con factores F se considera el tamaño del filtro y HN.

Por otro lado, [6] decidieron utilizar teorías ampliamente reconocidas para el modelado del humor. Las líneas base que establecieron, se basaron en el trabajo hecho por [11], presentaron las características de sus modelos de la siguiente forma: estructura de incongruencia, la ambigüedad, el efecto de la interpersonalidad en la que las emociones juegan un papel, en la subjetividad que aumenta la posibilidad de que se genere el humor, el estilo fonético y como parte complementaria se utilizó *KNN Features*, construida bajo el indicador de similitud semántica.

Pero de la misma forma que [6] y también [2] utilizaron las baselines de [11], con otras ponderaciones para CNN, *Human Theroy driven Features* (HTF), *Human Centric Features* (HCF), para lo cual, los autores re implementaron el método propuesto por [11] más el uso de *KNN features*, Word2Vec similar a KNN características y por último HCF más Word2Vec.

4. Construcción de corpus

Se puede definir como un corpus a cualquier colección de piezas lingüísticas ordenadas de acuerdo a un criterio lingüístico para servir de muestra del lenguaje que se quiere estudiar. Se observa que los autores pueden desarrollar sus corpus [12] y por otro lado, existen aquellos que utilizan corpus ya marcados [1, 2, 9]

Al respecto, [12] crearon un conjunto de datos sobre el humor compuesto de 9123 chistes compilados de forma manual en el idioma chino. Sus anotaciones no solo contenían bromas, también incluían anotaciones hechas sobre el humor lingüístico, no solo aquellos que tuvieran una cierta cantidad de humor, también las palabras que desencadenaban el humor, sus relaciones, características y categorías humorísticas. Estos autores aseveran que el rápido crecimiento de las redes sociales, trae un número significativo de texto y dentro de estos los humorísticos. En la mayoría el humor se genera con dos conceptos incongruentes entre sí y examinándolo en un marco semántico, para saber si efectivamente ambos están semánticamente desconectados y justamente validar la declaración anterior, se forma un contraste, que puede generar ambigüedad.

Por otra parte, [12] hablan de la importancia de los corpus, ya que en palabras sucintas un corpus es indispensable para el análisis de humor y determinara también la calidad de su detección automática. Por lo tanto, en lo que se basa la construcción de este grupo de datos, es en la premisa de los sets utilizados, para que expresen de forma adecuada cómo surge el humor. En general los datos recopilados fueron de twitter, en un esquema de anotaciones con palabras clave, que justamente hacían al texto gracioso, por la forma, la relación de los caracteres, el escenario, la categoría humorística y su grado humorístico.

[12] indican que trabajos previos no tenían en cuenta ciertas características en la detección del humor y su calidad. Las cuales hacen que su trabajo obtenga un esquema más completo, una de las principales diferencias obtenidas por estos autores fue la de anotar no solo lo que era humorístico sino también lo que causaba el humor.

El modelo desarrollado por los autores cuenta con las siguientes características de anotación:

JokeModel = (Relationship, Scene, Category, HumorLevel, Keyword, DataSource).

Se explica que cada tópico dentro del modelo maneja una conexión general, en el caso de la relación entre los actores puede ser el caso de doctor-paciente, amantes, subordinados, ya que obtener la relación ayuda a esclarecer el contexto de la broma, lo mismo pasara con la escena en la que se desarrolla. En el caso de la categoría y el nivel humorístico no hay consenso sobre qué se debe de usar, por lo tanto, los autores decidieron desarrollar sus propias categorías de humor y una anotación categórica de cinco niveles en los grados del humor.

Los autores determinan que las palabras claves resultan en uno de los desafíos más grandes, ya que muchas consideraciones tienen que ser tomadas, el leer el texto entero y establecer un significado general del texto, así como cada palabra del texto, para establecer su significado en el contexto que se utiliza, a su vez de estas mismas palabras determinan si su significado es de incongruencia, conflicto, ambigüedad, sorpresa o que desarrollen una emoción fuerte que haga al texto ser gracioso en el contexto dado, también decidir si el contexto en el que fue usada la palabra puede ser entendido.

El proceso que se tomó en cuenta para las anotaciones de los chistes se dio por validación cruzada, en el que solo si los anotadores estaban de acuerdo a la observación se consideraba completa. En general la fuerte necesidad de grupos de datos trae consigo nuevas formas de crearlos, nuevos parámetros que usar, así mismo trae consigo nuevas formas de interpretar los datos, como dicen los autores, es intrigante saber que la palabra con más alta frecuencia en ciertos chistes está sumamente ligada con la escena en la que son puestos, determinando su locación.

Por otro lado, [2] utilizaron corpus ya marcados, establecieron que para evaluar de manera correcta ocuparían un conjunto de datos que consistiría en muestras humorísticas (positivas) y no humorísticas (negativas).

Por lo tanto, su conjuntode datos consistirá de cuatro partes: *Puns of the Day* por [11], 16000 *One Liners* [7], *Short Jokes dataset* y *PTT jokes*, los cuales cuenta con bromas variadas, con diferentes tamaños y lenguajes.

En cambio, [9] en el preprocesamiento de sus datos, iniciaron con la limpieza de los símbolos de numeral que aparecen con mucha frecuencia en tweets, así como de emoticons y *URL*, además de palabras usadas comúnmente en el ámbito de las redes sociales como los RT (*retweet*) y FAV (favoritos), estos últimos reemplazados por comodines que expresan el significado general de las palabras.

Se puede observar cierta tendencia a utilizar corpus con datos recabados de redes sociales o corpus con una amplia trayectoria de uso, se sobrentiende que no siempre será las mismas marcaciones para el corpus en el caso de que el modelo difiriera, ya que en raras ocasiones ocurrirá que se presente un corpus que esté exactamente marcado para el modelo que se vaya a utilizar. Se deben tomar ciertas consideraciones para ajustar la información, de tal forma que sea útil el marcaje y si esto no fuera suficiente, existe la opción de unir diferentes corpus para obtener uno que represente mejor nuestros datos.

5. Metodología para la detección del humor

La metodología para la detección del humor varía según el autor y la actividad que quiera desarrollar, pero en un enfoque general es utiliza *Deep learning*, aunque algunos autores llegan a utilizar machine learning de la forma de *Tree-CRF*.

Al respecto, [1] separan la información y evitan el sesgo de la mala escritura en twitter, este sesgo debe ser manejado de diferentes formas, entre ellas, mediante el análisis de sentimientos en forma de *Word-Vector* en árboles de *Condition Random Fields (Tree-CRF)* como clasificador de los datos, a su vez utiliza MalParser [8].

En el análisis de humor basado en anotaciones hechas por el humano (IberEval Workshop, 2018), en el cual las tareas requeridas fueron las siguientes: clasificar si un texto es humorístico o no y también predecir qué tan gracioso es. A partir de lo anterior, [9] propuso el uso de *Atention-based Recurrent Neural Network* (ARNN), donde la capa de cuidado o de atención ayude a calcular cada termino encaminado a encontrar las clases de humor.

Es importante recalcar que dichas actividades recaen sobre la información obtenida de Twitter, el cual, se volvió una plataforma popular para la obtención de contenido espontáneo creado por los usuarios.

También se nombran otros métodos para el reconocimiento del humor basados en aprendizaje supervisado, dentro de Deep Neuronal Network se encuentra *Long Short Term Memory* (LSTM) y también su variante bidireccional (Bi-LSTM). [9] hablaron del uso de Redes Neuronales Recurrentes (RNN) para la obtención de características para datos secuenciales. Estos autores utilizaron un ARNN, pero con la variante bidireccional de LSTM para generar el contexto del vector, lo cual, será pasado directamente a otra red LSTM para detectar si es humor o no, los autores describieron que dicho método no fue explorado en el campo de las arquitecturas basadas en ARNN para el reconocimiento del humor en español hasta ese momento.

[6] utilizó un modelo que tenía en tres características principales: complejidad métrica, humor en el texto y la expresión humorística del texto.

Además, midieron de diferente forma la complejidad de las oraciones, las cuales fueron cuantificadas como características respecto al número de sustantivos en frases, números de verbos, frases proposicionales, conjunciones subordinadas y otras características. Propusieron dentro de la estructura estadística, sus reglas de producción y por último las relaciones de dependencia, lo cual, indicó las relaciones entre palabras.

De igual forma, intentaron aplicar una combinación de POS *tag* y relación dependencia, obteniendo resultados poco favorables en comparación a solo utilizar relación de dependencia.

Como parte de la metodología que elaboraron [2], quienes utilizaron *Convolutional neural network* (CNN) diseñadas para extraer las características locales en dimensiones grandes de datos, entre ellas imágenes o conversaciones, quienes se apoyaron en una estructura basada en la tarea de clasificación de texto [4], lo primero fue tokenizar las sentencias de entrada de la forma word-vector con dimensiones D a una matriz de dos dimensiones, se utilizaron vectores GloVe, los cuales, fueron entrenados en token 6B, 400k vocabulario de Wikipedia 2014 más Gigaword 5 en la capa de *embedding*. También se utilizaron diferentes tamaños de filtros con un rango de tres a 20, para cada filtro de tamaño, aplicaron de 100 a 200 filtros, explotando el uso de *max pooling* y aplanando la salida, lo que conllevó a un vector aplanado 1D de dimensiones N en la salida predicha, se tomó en cuenta que se mejora el rendimiento bajo ciertas consideraciones, decidieron no enfatizar tanto la profundidad de la red, ya que esto traía consigo un incremento en la dificultad, tanto como sea profunda la red.

6. Análisis de modelos para detectar el humor

Como fue analizado, los enfoques con *Deep Learning* tienen resultados interesantes que permiten entender que este enfoque puede ser el adecuado para años venideros, sin olvidar los enfoques clásicos con los cuales comparar línea base.

Se exploraron características lingüísticas sobre el estilo por parte de [2], por ejemplo, número de palabras, número de caracteres usados, el énfasis por repetición y también el número de entre comillas. En cambio, para las características de estructura y contenido se usaron diferentes vocabularios como lo son de estilo topológico, animal, obsceno, sexual, además se tomó en cuenta la ambigüedad lexica (*Synsets*)(ADDESE). A su vez, se usaron algunas de las características afectivas, los sentimientos y el uso de emoticonos, para saber sí representaban una emoción negativa o positiva (*Emoticon Sentiment*).

Los datos superaron la línea base propuestos de 50% humor o no humor, se obtuvo un F_1 en su primera corrida de 0.7851, en la tercera y última de 0.7702. En general se puede observar que bajo el experimento desarrollado para las tareas propuestas los autores obtuvieron el mejor rendimiento en precisión (accuracy) en la primera fase frente a las demás, así mismo lograron una precisión y puntaje de F_1 más altos entre los contendientes, no de la misma forma en la exhaustividad (recall), donde tuvieron valores bajos.

Por otro lado, [6] adhirieron las características sintácticas estructurales en las líneas base propuestas, obtuvieron una mejoría notable en el rendimiento tanto del puntaje en precisión (accuracy) con un 7.9%, para el de puntaje de F_1 en un 7.2%, se puede apreciar que las características sintácticas pueden obtener mejores resultados con algunas propiedades del humor. Lo cual, significa que su modelo ayudo a identificar el humor y a explicar cómo estas estructuran sintácticas se relacionaron con el fenómeno lingüístico del humor. Postularon que, al identificar humor, los textos humorísticos utilizan palabras simples, pero estructuras sintácticas más complejas. Y a su vez, los textos tienden a parecerse más a conversaciones, no solo eso también este tipo de textos son más vívidos y específicos, tienden a explicar que el texto deja al lector imaginar la situación y que ciertas palabras son utilizadas para realzar la situación que plantea.

Por otro lado, en la experimentación de [2] quienes decidieron optar por las líneas base [11], utilizaron *Random Forest* con Word2Vec (W2V) adicionado de Human *Centric Feature* (HCF), optaron por una taza de caída del 0.5 y para testear el desempeño con factores F y HN. El valor de F significa el incremento del tamaño del filtro, mientras que HN indica las capas del *Highway* para entrenar la red profunda, en este caso se eligieron tres capas ya que provee de mayor estabilidad y precisión (accuracy)en los pasos de entrenamiento. Se obtuvo un incremento en la puntuación F_1 de 0.859 a 0.903 con el modelo planteado usando CNN.

7. Conclusiones y trabajos a futuro

De los modelos analizados se aprecia que, en cuanto a la generación de corpus, la mejor fue la de [12] ya que su planteamiento de creación abarca características novedosas de marcado. En cambio, en la tarea de detección de humor [2,6] lograron modelos que mejoran las características de precisión (accuracy) y el puntaje de F_1. Los procesos con *Deep Learning* obtienen los resultados más favorables en la detección del humor, sin dejar pasar los enfoques clásicos.

En trabajos futuros se pretende extender para obtener un estado del arte y no un análisis pequeño como el que se puede observar en este artículo.

Agradecimientos. Agradecemos a CONACYT, SNI, IPN (SIP, COFAA), apoyo de proyectos SIP 20200859 y 20200797 y Conacyt A1-S-47854.

Referencias

1. Castro, S., Chiruzzo, L., Rosa, A.: Overview of the HAHA Task: Humor Analysis based on Human Annotation at IberEval (2018)
2. Chen, P., Soo, V.: Humor recognition using deep learning. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2 (2018)
3. Ghosh, A., Veale, T.: Fracking sarcasm using neural network (2016)
4. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, pp. 1746–1751 (2014)
5. Kreuz, R., Glucksberg, S.: How to be sarcastic: The echoic reminder theory of verbal irony. Journal of Experimental Psychology (1989)
6. Liu, L., Zhang, D., Song, W.: Modeling sentiment association in discourse for humor recognition. Association of Computational Linguistics (2018)
7. Mihalcea, R., Strapparava, C.: Making computers laugh: Investigations in automatic humor recognition. In: Conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 531–538 (2005)
8. Nakagawa, T., Inui, K., Kurohashi, S.: Dependency tree-based sentiment classification using CRFs with hidden variables. In: NAACL-HLT, pp. 786–794 (2010)
9. Ortega-Bueno, R., Muñiz-Cuza, C.E., Medina-Pagola, J.E., Rosso, P.: UO-UPV: Deep linguistic humor detection in Spanish social media (2018)

10. Utsumi, A.: Verbal irony as implicit display of ironic environment: Distinguishing ironic utterances from nonirony 1. Journal of Pragmatics (2000)
11. Yang, D., Lavie, A., Dyer, C., Hovy, E.: Humor recognition and humor anchor extraction. In: Conference on Empirical Methods in Natural Language Processing, pp 2367–2376 (2015)
12. Zhang, D., Zhang, H., Liu, X., Lin, H., Xia, F.: Telling the whole story: a manually annotated Chinese dataset for the analysis of humor in jokes (2019)

ISSN 1870-4069

Generalidades para funciones de correlación entre distribuciones de probabilidad

Maria Elena Ensastegui-Ortega, Ildar Batyrshin, Alexander Gelbukh

Instituto Politécnico Nacional,
Centro de Investigación en Computación,
México

{elena.ensastegui, batyr1}@gmail.com,
gelbukh@gelbukh.com

Resumen. El artículo presenta funciones de correlación basadas en similitudes y disimilitud entre distribuciones de probabilidad. La medidas de similitud y distancia que son usadas comparan dos distribuciones de probabilidad o tuplas de números reales. Se exploran funciones de distribución de probabilidad, medidas de distancias/similitud que cumplen con las propiedades de función de similitud o disimilitud. A partir de estas nuevas funciones de similitud o disimilitud se construyen nuevas funciones de correlación. También se da un bosquejo general de lo que son las funciones de similitud, disimilitud y correlación.

Palabras clave: Funciones de similitud o disimilitud, funciones complementarias, funciones de correlación.

Generalities for Correlation Functions between Probability Distributions

Abstract. The paper presents correlation functions based on similarity and dissimilarity between probability distributions. The similarity and distance measures that are used compare two probability distributions or tuples of real numbers. Probability distribution functions, measures of distance or similarity that meet the similarity or dissimilarity function properties are explored. From these new similarity or dissimilarity functions, new correlation functions are constructed. A general outline of what the similarity, dissimilarity, and correlation functions are is also given.

Keywords: Similarity or dissimilarity functions, complementary functions, correlation functions.

pp. 131–140; rec. 2019-04-19; acc. 2020-05-01

1. Introducción

Las aplicaciones a los algoritmos de Machine Learning y Minería de Datos han llevado al ser humano a un crecimiento acelerado en muchas áreas tanto del conocimiento como de la vida diaria, esto se debe a que los seres humanos cada día pueden usar herramientas que cumplan con tareas automatizadas dando a los seres humanos la opción de optimizar su tiempo.

Las medidas de similitud y correlación se utilizan en la recuperación de información, clasificación de datos, aprendizaje automático, análisis de relaciones y toma de decisiones en ecología, lingüística computacional, procesamiento de imágenes y señales, análisis de datos financieros, bioinformática y ciencias sociales.

Los algoritmos de aprendizaje no supervisado hacen uso de medidas de similitud para poder agrupar conjuntos de datos, estos algoritmos trabajan tanto con datos nominales como numéricos, estos datos se mide que tanto se relacionan entre sí, con medidas de similitud.

En este articulo, la segunda sección, se explican las propiedades de funciones de similitud y disimilitud, la tercera es sección una breve explicación sobre distribuciones de probabilidad, en la cuarta sección se exploran medidas de similitud como: Bhattacharyya, Coseno, Czekanowski, Ruzicka, Jaccard y Dice que también son funciones de distribuciones de probabilidad [2] se puede ver que cumplen que son funciones $S : \Omega \times \Omega \rightarrow [0; 1]$ s Ω si para todo x, y en Ω cumplen con las propiedades de simetría y reflexiva por tanto son funciones de similitud [1], a partir de estas se construyen nuevas funciones de correlación.

En la sección cinco se exploran medidas de distancias como: Sorensen, Wave Hedges, Soergel, Tanimoto, Jaccard y Dice, que también son funciones de distribuciones de probabilidad [2] se puede ver que cumplen que son funciones $S : \Omega \times \Omega \rightarrow [0; 1]$ s Ω si para todo x, y en Ω cumplen con las propiedades de simetría e irreflexiva por tanto son funciones de disimilitud [1] a partir de estas se construyen nuevas funciones de correlación.

En la sección seis, se muestra que algunas de estas funciones de similitud y disimilitud son complementarias. En la sección siete, se da un breve resumen de las funciones de correlación creadas. Por último en la sección siete se discuten las concluciones y trabajo a futuro.

2. Funciones de similitud, disimilitud y de correlacion

Los datos de los cuales podremos extraer conocimiento pueden venir de un conjunto diferente del vacío, sea Ω tal conjunto que se denominará un dominio universal o un conjunto subyacente en la definición de similitud. Ya que podemos considerar a Ω como un dominio específico para un tipo de datos considerado: el conjunto de todas las n-tuplas binarias, el conjunto de todos los vectores con valores en los reales de longitud n, el conjunto de imágenes u objetos considerados en algún problema, etc[1].

Una función $S : \Omega \times \Omega \rightarrow [0; 1]$ se llama una función de similitud en Ω si para todo x, y en Ω cumple las siguientes propiedades:

- Simetría: $S(x, y) = S(y, x)$ [5],
- Reflexiva: $S(x, x)$=1.

Una función D: $\Omega \times \Omega \to [0; 1]$ es una función de disimilitud en Ω si para todo x, y en Ω esta cumple las siguientes propiedades:

- Simetría: $D(x, y) = D(y, x)$,
- Irreflexiva: $D(x, x) = 0$.

Podemos decir que estas funciones son complementarias si para todo x, y en Ω se cumple que: $S(x, y) + D(x, y) = 1$.

Para funciones complementarias de similitud o disimilitud tenemos que:

$$S(x, y) = 1 - D(x, y), D(x, y) = 1 - S(x, y). \tag{1}$$

Una función A: $\Omega \times \Omega \to [0; 1]$ es una función de correlación en Ω si para todo x, y en Ω esta cumple las siguientes propiedades [5]:

- Simetría: $A(x, y) = A(y, x)$,
- Reflexiva: $A(x, x) = 1$,
- Negativa:$A(x, y) < 0$ para algún x, y en Ω.

Dichas funciones de correlación se denominarán funciones de correlación débiles si no satisfacen la propiedad de relación inversa la cual nos dice:

Proposición 1. Suponga que S y D son funciones de similitud y disimilitud en Ω de tal manera que para algunos x, y en Ω se cumple: $S(x, y) < D(x, y)$ y, entonces la función definida para todo (x, y) en Ω esta dada por:

$$A(x, y) = S(x, y) - D(x, y), \tag{2}$$

es una función de correlación. Si S y D son complementarios, entonces la función A será una función de correlación si para algún x, y en Ω se cumple: $S(x, y) < ,5$.

La fórmula obtenida para A tiene una interpretación razonable: la correlación entre x e y es positiva si la similitud entre ellas es mayor que la disimilitud, y la correlación es negativa en caso contrario.

Si las funciones de similitud S y disimilitud D son complementarias, la función de correlación A definida por (1) se llama complementaria a S y D. Las funciones complementarias S, D y A se designarán como (S, D, A) y se denominarán tripleta de correlación . De la definición de las funciones complementarias disimilitud, similitud y de (2) se deduce que las funciones de similitud, disimilitud y correlación de la tripleta de correlación (S,D,A) pueden obtenerse una de otra para todo (x, y) en Ω como sigue [4]:

$$S(x, y) = 1 - D(x, y), D(x, y) = 1 - S(x, y), \tag{3}$$

$$A(x, y) = 2S(x, y) - 1, S(x, y) = \frac{1}{2}A(x, y) + 1. \tag{4}$$

$$A(x, y) = 1 - 2D(x, y), D(x, y) = \frac{1}{2}(1 - A(x, y)). \tag{5}$$

3. Distribuciones de probabilidad

Un espacio probabilístico formalmente es una tripleta (Ω, F, P) con (Ω, F) un espacio métrico y P una medida de probabilidad. Supongamos que existe $A \in F$ entonces podemos decir que $P(A)$ se le llama probabilidad de A [3].

Consideramos que P es una función que asigna valores en el intervalo (0,1) y cumple las siguientes propiedades: Sea w una variable aleatoria en el conjunto F [3]:

1. $P(w) \geq 0$,
2. $\sum_{w \in \Omega} p(w) = 1$.

Sea X un conjunto finito numerable y consideramos a la variable aleatoria w, decimos que $p_w(x) = P(C_x)$ con $\{C_x = y : w(y) = x\}$, es una distribución de probabilidad de la variable w.

En términos simples una función de distribución de probabilidad asigna evento que ocurre sobre la variable aleatoria la probabilidad de que dicho evento ocurra [6].

4. Construcción de funciones correlación con funciones de similitud

Sea $x = x_1, \ldots, x_n$ una distribución de probabilidad finita con $x_i \geq 0$ para todo $i = 1, \ldots, n$ y $\sum_{i=1}^{n} x_i = 1$. Sea $y = y_1, \ldots, y_n$ una distribución de probabilidad finita con $y_i \geq 0$ para todo $i = 1, \ldots, n$ y $\sum_{i=1}^{n} y_i = 1$. Definimos al coeficiente Bhattacharyya como sigue [5]:

$$S(x,y) = \sum_{i=1}^{n} \sqrt{x_i y_i}. \tag{6}$$

Podemos mostrar que la anterior similitud cumple las propiedades de simetría y reflexiva .

Ya que esta medida de similitud cumple con las propiedades para ser una funciń de similitud podemos concluir que es una función de similitud, por tanto podemos hacer uso de (4) para construir una función de correlación débil como sigue:

$$A(x,y) = 2\sum_{i=1}^{n} \sqrt{x_i y_i} - 1. \tag{7}$$

La simititud coseno esta dada por la siguiente formula. Consideramos n-tuplas $x = (x_1, ..., x_n)$ y $y = (y_1, ..., y_n)$ con valores en los reales [4].

$$cos(x,y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}}, \tag{8}$$

donde $x_i, y_i \geq 0$ $i = 1, ..., n$.

Es fácil notar que la similitud coseno cumple con las propiedades de simetría y reflexiva. Ya que es simétrica, reflexiva, es una función de similitud para construir una función de correlación débil consideramos la formula (4).

Entonces tenemos que la función de correlación débil de la función similitud coseno esta dada por:

$$A(x,y) = \frac{2\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} - 1. \tag{9}$$

La similitud Czekanowski compara dos distribuciones de probabilidad [2]. Sean P_i y Q_i dos distribuciones de probabilidad la similitud Czekanowski esta dada por:

$$S_{Cze} = \frac{2\sum_{i=1}^{d} min(P_i, Q_i)}{\sum_{i=1}^{d}(P_i + Q_i)}. \tag{10}$$

Se puede mostrar que la medida de similitud cumple con $S_{Cze} : \Omega \times \Omega \to [0;1]$, ademas de que es reflexiva y simétrica, entonces la similitud Czekanowski es una función de similitud, por las propiedades que cumplen las funciones de similitud por tanto podemos usar la ecuación (4) para construir una función de correlación débil para la función de similitud Czekanowski como sigue:

$$A_{Cze} = \frac{4\sum_{i=1}^{d} min(P_i, Q_i)}{\sum_{i=1}^{d}(P_i + Q_i)} - 1. \tag{11}$$

La similitud Ruzicka compara dos distribuciones de probabilidad [2]. Sean P_i y Q_i dos distribuciones de probabilidad la similitud Ruzicka esta dada por:

$$S_{Ruz} = \frac{\sum_{i=1}^{d} min(P_i, Q_i)}{\sum_{i=1}^{d} max(P_i, Q_i)}. \tag{12}$$

Se puede mostrar que la medida de similitud cumple con $S_{Cze} : \Omega \times \Omega \to [0;1]$, y cumple con las propiedades de simetría y reflexiva, por tanto podemos usar la ecuación (4) para construir una función de correlación débil para la similitud Ruzicka, como sigue:

$$A_{Ruz} = \frac{2\sum_{i=1}^{d} min(P_i, Q_i)}{\sum_{i=1}^{d} max(P_i, Q_i)} - 1. \tag{13}$$

La similitud Jaccard compara dos distribuciones de probabilidad [2]. Sean P_i y Q_i dos distribuciones de probabilidad la similitud Jaccard esta dada por:

$$S_{Jac} = \frac{\sum_{i=1}^{d} P_i Q_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2 - \sum_{i=1}^{d} P_i Q_i}. \tag{14}$$

Se puede mostrar que la medida de similitud cumple con $S_{Jac} : \Omega\times\Omega \to [0;1]$, y cumple con las propiedades de simetría y reflexiva, por tanto podemos usar la ecuación (4) para construir una función de correlación débil para la similitud Jaccard, como sigue:

$$A_{Jac} = \frac{2\sum_{i=1}^{d} P_i Q_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2 - \sum_{i=1}^{d} P_i Q_i} - 1. \tag{15}$$

La similitud Dice compara dos distribuciones de probabilidad [2]. Sean P_i y Q_i dos distribuciones de probabilidad la similitud Dice esta dada por:

$$S_{Dice} = \frac{2\sum_{i=1}^{d} P_i Q_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2}. \tag{16}$$

Se puede mostrar que la similitud Dice tiene su dominio en el conjunto (0,1), y cumple con las propiedades de simetría y reflexiva, por tanto podemos usar la ecuación (3) para construir una función de correlación débil para la similitud Dice, como sigue:

$$A_{Dice} = \frac{4\sum_{i=1}^{d} P_i Q_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2} - 1. \tag{17}$$

5. Construcción de funciones correlación con funciones de disimilitud

La distancia Sorensen [2] puede ser usada para comparar dos distribuciones de probabilidad. Sean P_i y Q_i distribuciones de probabilidad, la distancia Sorencen esta dada por:

$$D_{sor} = \frac{\sum_{i=1}^{d} |P_i - Q_i|}{\sum_{i=1}^{d} P_i + Q_i}. \tag{18}$$

Se puede mostrar que esta distancia $D_{sor} : \Omega \times \Omega \rightarrow [0;1]$, ademas de que es irreflexiva, y simétrica, entonces la distancia Sorencen es una función de disimilitud,por las propiedades que cumplen las funciones de disimilitud por tanto podemos usar la ecuación (5) para construir una función de correlación débil para la función de disimilitud Sorencen como sigue:

$$A_{sor} = 1 - \frac{4\sum_{i=1}^{d} |P_i - Q_i|}{\sum_{i=1}^{d} (P_i + Q_i)}. \tag{19}$$

La distancia Wave Hedges compara dos distribuciones de probabilidad [2]. Sean P_i y Q_i dos distribuciones de probabilidad, entonces la distancia Wave Hedges esta dada por:

$$d_{WH} = \sum_{i=1}^{d} \frac{1 - min(P_i, Q_i)}{max(P_i, Q_i)}. \tag{20}$$

Se puede mostrar que la distancia cumple con $d_{WH} : \Omega \times \Omega \rightarrow [0;1])$ y cumple con las propiedades de simetría e irreflexiva, entonces podemos decir que es una función de disimilitud, por tanto podemos usar la ecuación (5) para construir una función de correlación débil para la disimilitud Wave Hedges, como sigue:

$$A_{WH} = 1 - \sum_{i=1}^{d} \frac{2(1 - min(P_i, Q_i)}{max(P_i, Q_i))}. \tag{21}$$

La distancia Soergel compara dos distribuciones de probabilidad [2]. Sean P_i y Q_i dos distribuciones de probabilidad, la distancia Soergel esta dada por:

$$d_{Sg} = \frac{\sum_{i=1}^{d} |P_i - Q_i|}{\sum_{i=1}^{d} max(P_i, Q_i)}. \tag{22}$$

Se puede mostrar que esta distancia cumple con $D_{Sg} : \Omega \times \Omega \rightarrow [0; 1]$, ademas de que es irreflexiva, y simétrica, entonces la distancia Soergel es una función de disimilitud,por las propiedades que cumplen las funciones de disimilitud por tanto podemos usar la ecuación (5) para construir una función de correlación débil para la función de disimilitud Soergel como sigue:

$$A_{Sg} = 1 - \frac{2\sum_{i=1}^{d} |P_i - Q_i|}{\sum_{i=1}^{d} max(P_i, Q_i)}. \tag{23}$$

La distancia Tanimoto compara dos distribuciones de probabilidad [2]. Sean P_i y Q_i dos distribuciones de probabilidad, la distancia Tanimoto esta dada por:

$$D_{Tani} = \frac{\sum_{i=1}^{d} max(P_i, Q_i) - min(P_i, Q_i)}{\sum_{i=1}^{d} max(P_i, Q_i)}. \tag{24}$$

Se puede mostrar que esta distancia $D_{Tani} : \Omega \times \Omega \rightarrow [0; 1]$, ademas de que es irreflexiva, y simétrica, entonces la distancia Tanimoto es una función de disimilitud,por las propiedades que cumplen las funciones de disimilitud por tanto podemos usar la ecuación (5) para construir una función de correlación débil para la función de disimilitud Tanimoto como sigue:

$$A_{Tani} = 1 - \frac{2\sum_{i=1}^{d} max(P_i, Q_i) - min(P_i, Q_i)}{\sum_{i=1}^{d} max(P_i, Q_i)}. \tag{25}$$

La distancia Jaccard compara dos distribuciones de probabilidad [2]. Sean P_i y Q_i dos distribuciones de probabilidad la similitud Jaccard esta dada por:

$$D_{Jac} = \frac{\sum_{i=1}^{d} (P_i - Q_i)^2}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2 - \sum_{i=1}^{d} P_i Q_i}. \tag{26}$$

Se puede mostrar que la distancia cumple con $D_{Jac} : \Omega \times \Omega \rightarrow [0; 1]$, y cumple con las propiedades de simetría e irreflexiva, por tanto podemos usar la ecuación (5) para construir una función de correlación débil para la función de disimilitud Jaccard, como sigue:

$$A_{Jac} = 1 - \frac{2\sum_{i=1}^{d} (P_i - Q_i)^2}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2 - \sum_{i=1}^{d} P_i Q_i}. \tag{27}$$

La distancia Dice compara dos distribuciones de probabilidad [2]. Sean P_i y Q_i dos distribuciones de probabilidad la similitud Dice esta dada por:

$$D_{Dice} = \frac{\sum_{i=1}^{d} (P_i - Q_i)^2}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2}. \tag{28}$$

Se puede mostrar que la distancia cumple con $D_{Dice} : \Omega \times \Omega \to [0;1]$, y cumple con las propiedades de simetría e irreflexiva, por tanto podemos usar la ecuación (5) para construir una función de correlación débil para la función de disimilitud Dice, como sigue:

$$A_{Dice} = 1 - \frac{\sum_{i=1}^{d} (P_i - Q_i)^2}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2}. \tag{29}$$

6. Funciones complementarias

Podemos notar que la función de disimilitud Sorensen puede expresarse de la siguiente manera [2]:

$$D_{sor} = 1 - S_{Cze} = \frac{\sum_{i=1}^{d} |P_i - Q_i|}{\sum_{i=1}^{d} P_i + Q_i}. \tag{30}$$

Lo que nos dice que por la ecuación (3) que la función de similitud Czekanowski y la función de disimilitud Sorensen son complementarias.

Se puede observar que, $S_{Jac} + D_{Jac} = 1$ [2] podemos concluir que son funciones complementarias por la ecuación (1). Sabemos que $S_{Dice} + D_{Dice} = 1$ [2] podemos concluir que son funciones complementarias por la ecuación (1).

Tabla 1. Funciones de similitud.

N	Funcion de Similitud	Funcion de Correlacion
(6)	$S(x,y) = \sum_{i=1}^{n} \sqrt{x_i y_i}$	$A(x,y) = 2\sum_{i=1}^{n} \sqrt{x_i y_i} - 1$
(8)	$cos(x,y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}}$	$A_{Cos}(x,y) = \frac{2\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} - 1$
(10)	$S_{Cze} = \frac{2\sum_{i=1}^{d} min(P_i,Q_i)}{\sum_{i=1}^{d} (P_i+Q_i)}$	$A_{Cze} = \frac{4\sum_{i=1}^{d} min(P_i,Q_i)}{\sum_{i=1}^{d} (P_i+Q_i)} - 1$
(12)	$S_{Ruz} = \frac{\sum_{i=1}^{d} min(P_i,Q_i)}{\sum_{i=1}^{d} max(P_i,Q_i)}$	$A_{Ruz} = \frac{2\sum_{i=1}^{d} min(P_i,Q_i)}{\sum_{i=1}^{d} max(P_i,Q_i)} - 1$
(14)	$S_J = \frac{\sum_{i=1}^{d} P_i Q_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2 - \sum_{i=1}^{d} P_i Q_i}$	$A_J = \frac{2\sum_{i=1}^{d} P_i Q_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2 - \sum_{i=1}^{d} P_i Q_i} - 1$
(16)	$S_{Dice} = \frac{2\sum_{i=1}^{d} P_i Q_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2}$	$A_{Dice} = \frac{4\sum_{i=1}^{d} P_i Q_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2} - 1$

Tabla 2. Funciones de disimilitud.

N	Funcion de Disimilitud	Funcion de Correlacion
(18)	$D_{sor} = \frac{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert}{\sum_{i=1}^{d} P_i + Q_i}$	$A_{sor} = 1 - \frac{4 \sum_{i=1}^{d} \lvert P_i - Q_i \rvert}{\sum_{i=1}^{d} (P_i + Q_i)}$
(20)	$d_{WH} = \sum_{i=1}^{d} \frac{1 - min(P_i, Q_i)}{max(P_i, Q_i)}$	$A_{WH} = 1 - \sum_{i=1}^{d} \frac{2(1 - min(P_i, Q_i)}{max(P_i, Q_i))}$
(22)	$d_{Sg} = \frac{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert}{\sum_{i=1}^{d} max(P_i, Q_i)}$	$A_{Sg} = 1 - \frac{2 \sum_{i=1}^{d} \lvert P_i - Q_i \rvert}{\sum_{i=1}^{d} max(P_i, Q_i)}$
(24)	$D_{Ta} = \frac{\sum_{i=1}^{d} max(P_i, Q_i) - min(P_i, Q_i)}{\sum_{i=1}^{d} max(P_i, Q_i)}$	$A_{Ta} = 1 - \frac{2 \sum_{i=1}^{d} max(P_i, Q_i) - min(P_i, Q_i)}{\sum_{i=1}^{d} max(P_i, Q_i)}$
(26)	$D_J = \frac{\sum_{i=1}^{d} (P_i - Q_i)^2}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2 - \sum_{i=1}^{d} P_i Q_i}$	$A_J = 1 - \frac{2 \sum_{i=1}^{d} (P_i - Q_i)^2}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2 - \sum_{i=1}^{d} P_i Q_i}$
(28)	$D_{Dice} = \frac{\sum_{i=1}^{d} (P_i - Q_i)^2}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2}$	$A_{Dice} = 1 - \frac{\sum_{i=1}^{d} (P_i - Q_i)^2}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2}$

7. Resultados

Por último en la Tabla 1 muestra las funciones de correlación creadas a partir de las funciones de similitud, mientras que la Tabla 2 muestra las funciones de correlación creadas a partir de las funciones de disimilitud.

8. Conclusiones y trabajo a futuro

En efecto se crearon nuevas funciones de correlación que cumplen con la propiedad de correlación, de medidas de similitud y distancias de distribuciones de probabilidad, que cumplían con las características para ser funciones de similitud o en su caso funciones de disimilitud por tanto era posible construir su función de correlación, también pudimos mostrar que algunas de ellas eran funciones de similitud complementarias.

Como trabajo a futuro se planea crear nuevas funciones de correlación que cumplan con la propiedad fuerte de correlación. Ademas de encontrar conjuntos de datos para los cuales se pueden aplicar y medir su precisión.

Referencias

1. Batyrshin, I.: Data Science: Similarity, Dissimilarity and Correlation Functions. (2019)
2. Sung-Hyuk Cha: Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. (2007)
3. Koralov, L., Sinai, Ya.: Theory of Probability and Random Processes. (2007)
4. Batyrshin, I.: Towards a general theory of similarity and association measures: similarity, dissimilarity and correlation functions. Journal of Intelligent and Fuzzy Systems, vol. 36, no. 4, pp. 2977–3004 (2019)
5. Batyrshin, I.: Constructing correlation coefficients from similarity and dissimilarity functions. Acta Polytechnica Hungarica, 16(10), pp. 191–204 (2019)

6. Díaz Mata, A.: Estadística aplicada a la administración y economía. México, McGraw Hill (2013)

Electronic edition
Available online: http://www.rcs.cic.ipn.mx

ISSN: 1870-4069
http://rcs.cic.ipn.mx

Centro de Investigación
en Computación