Cancer Detection using the *KDD* Process

Eddy Sánchez de la Cruz, Homero Alpuín Jiménez, and Pilar Pozos Parra

Universidad Juárez Autónoma de Tabasco,
División Académica de Informática y Sistemas,
Carretera Cunduacán-Jalpa Km. 1,
Cunduacán, Tabasco, Mexico
eddsac@hotmail.com, {homero.alpuin,pilar.pozos}@dais.ujat.mx

Abstract. Due to increasing deaths from breast cancer, it becomes necessary to incorporate computer technologies to support medical diagnosis. Despite this technological effort, the level of breast cancer detection is keeping low. This article presents the KDD process (*Knowledge Discovery in Databases*) as an alternative to obtain a trustable detection in medical imaging. Two important strategies that are used in this research are: *the Mejia's method*, which is used for data cleaning, and *LADTree algorithm*, which is used for detection. The implementation of the *KDD process* shows satisfactory results.

Key words: Breast Cancer, Mammograms, KDD Process, Detection.

1 Introduction

Data analysis tries to find trends or changes in data behavior. These trends and variations will be called, patterns. If the patterns are useful and relevant for the domain, then it is called knowledge.

At the beginning, the data analysis was performed manually using statistical techniques. Currently, this type of analysis is not feasible because of the large amount of information that can contain a modern database (DB) and for several data formats such as tables (Relational databases), sequences [1], graphs [12], images [8], and audio. In order to solve these problems the Data Mining arises (DM). DM is an automatic or semiautomatic process that seeks to discover hidden patterns in a dataset that are potentially useful for users of the DB [13], [1]. The main idea is to enhance and extract the information for interpretation and analysis by humans [10]. Extracting information from images is a complex process that uses different strategies to tackle the problem such as Artificial Neural Networks (ANN) [6], Bayesian networks [3], fuzzy logic [15], case based reasoning [3], bio-inspired algorithms [5] and hybrids of these.

In this article, the KDD process is used to analyze mammograms. For a specialist, the images are valuable to detect abnormalities; however, the specialist's analysis may be subject to errors, i.e. not guaranteed to find all the anomalies. Therefore, this research is use as a support in the doctor's decision to detect

Advances in Soft Computing Algorithms
Research in Computing Science 49, 2010, pp. 109-117

breast cancer. The rest of the article is divided as follows: Section 2 briefly explains the KDD process in this research. Section 3 discusses the first two phases of the KDD process. Section 4 explains the data transformation, data mining, and the pattern evaluation obtained, and then we test with a data set which shows a good detection of breast cancer. Finally, we conclude with a discussion for further work.

2 KDD Process

KDD is an interactive and iterative process which involves many steps and includes a lot of decisions that must be taken by the developer [9].

The KDD process for this research, where the steps are as follow:

- **Selection:** Given a set of different database takes the most representative.
- **Preparing data:** Data are filtered to reduce noise.
- **Data transformation:** The images are converted to table, taking the histogram of each one.
- **Data mining:** Applying an algorithm to obtain a pattern of behavior.
- **Assessment patterns:** Patterns are evaluated using a test database to determine whether breast cancer detection is appropriate.

3 Medical Images Preparation

Medical images analysis is a relevant issue for providing support to medical diagnosis. Some work has been done to detect abnormalities in mammographies using algorithms that implement different strategies [6], [3], [15], [3], [5] . Knowing that a mammography is an array of pixels, we want to find the pattern of the pixels that determines the existence of an anomaly, (see figure 1). The following sections will describe the first two phases of the KDD process: Selecting the database and preparing data.

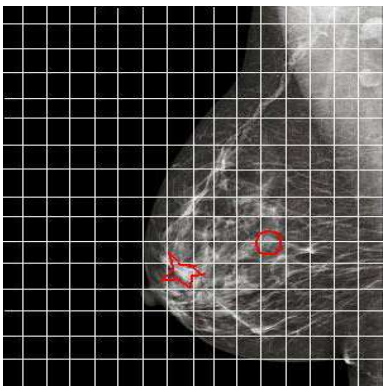

Fig. 1. Areas abnormalities.

3.1 Selecting the Database

We performed an investigation to obtain a public domain mammography's database and MIAS, which is a reduced database, was selected as it was used to test research in [7], [10] and which the South Florida University recommends in [4]:

Table 1. Database MIAS mammograms.

Name	**Description**
MIAS	Database created by the *Mammographic Image Analysis Society*, United Kingdom. Source: http://www.wiau.man.ac.uk/services/MIAS/MIASweb.html There is a reduced version that contains the same images as the original version, but reduced to a size of 1024 x 1024 pixels, available at http://peipa.essex.ac.uk/info/mias.html

3.2 Preparing Data

The image often contains information that is irrelevant for a given application. In such cases, the image must be filtered for removing all the irrelevant information.

All images have some amount of noise, which can be cause by the camera or the means of the signal transmission. Noise usually is manifest as isolated pixels that take a different gray level of its neighbors.

Filters are designed to reduce noise that can occur as a result of an image capture process, scanning or transmission [14]. This stage of KDD process uses the Mejía's filter proposed in [10]. This method is based on Non-subsampled Contourlet Transform (NSCT) and the Prewitt filter. The filter scheme is shown in figure 2.

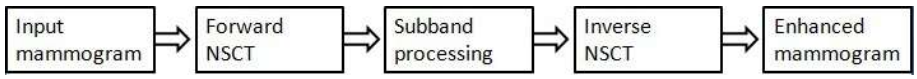

Fig. 2. Block diagram of the transform methods for images processing.

Describing in [10], this filtering process is efficiently used to enhancement a type of cancer, called micro-calcifications, in mammograms images. Figure 3 shows an image and the image after the filter. In the filtered image, we can see the enhancement of the brighter pixels which reveals the existence of micro-calcifications.

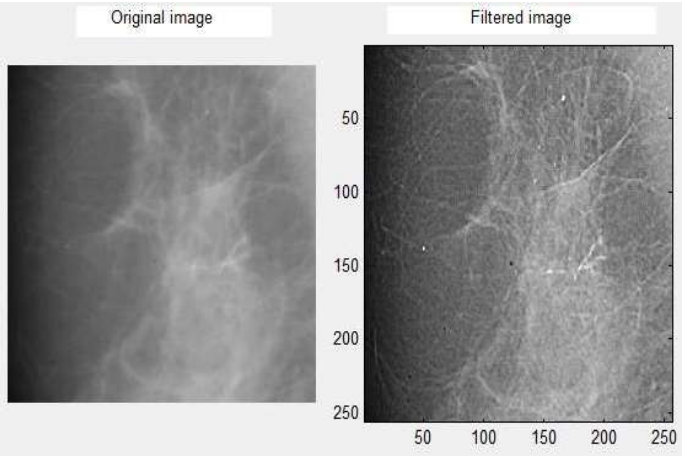

Fig. 3. NSCT Filter.

4 Processing and Analyzing Medical Images

In this section we explain the last three phases of the KDD process: Data transformation, Data mining and Patterns assessment.

4.1 Data Transformation

At this third phase of the KDD process, we obtained the histogram of each filtered mammography. The histograms describe frequency distribution of gray levels. The X axis represents the gray scale ranging from 0 to 255, and the Y axis represents the number of pixels that correspond to each gray level, an example of this type of histogram is shown in the figure 4.

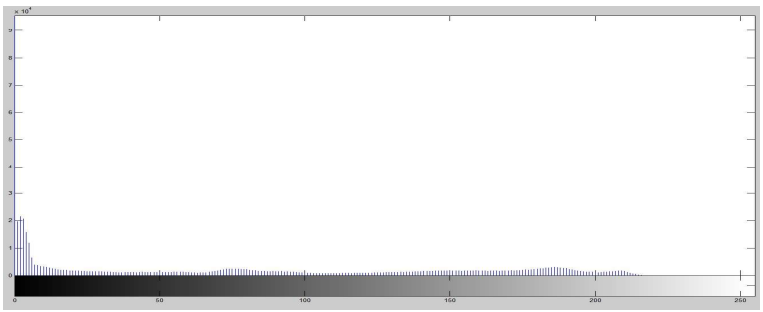

Fig. 4. Mammography's Histogram.

We take 100 instances (histograms) to build the training base, 50 with cancer and 50 without cancer. Of these 100 instances 13 are benign (B), 37 are normal (N) and 50 are malignant (M). Finally, in the data transformation, we build a file with the training base in ARFF format. An ARFF (= Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files have two sections. The first section is the *Header* information, which is followed the *Data* information [2]. The Header of the ARFF file contains the name of the relation, a list of the attributes (the columns in the data), and their types [2], an example of this type of file is shown below:

```
% 1. Title: Mammograms Database EDDSAC
% 2. Sources:
% (a) Creator: Sanchez, Eddy
% (b) email: eddsacx@gmail.com
% (c) Date: June, 2010
%
@RELATION ima_training

@attribute Severidad {B,N,M} % B = Benign, N = Normal
 and M = Malignant
@attribute Pix0 numeric

⋮

@attribute Pix255 numeric
```

The Data of the ARFF file looks like the following:

```
@data
%-------------------------------------------------------------
% This histogram corresponds to an image
% which presents a benign abnormality (B).
%-------------------------------------------------------------
B,520417,26413,16747,14412,11561,9346,10517,
7801,5836,5303,4872,4381,4115,3718,3575,
3153,2767,2609,2424,2350,2235,2126,2226, ...
```

4.2 Data Mining

To implement the pattern detection we use the software WEKA (*Waikato Environment for Knowledge Analysis*) v. 3.6.1. It is a collection of machine learning algorithms for data mining tasks. It contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes [11]. Weka is open source software issued under the GNU General Public License [11].

Weka contains different classification strategies, such as: bayesian algorithms (*bayes*), classification functions (*functions*), lazies algorithms (*lazy*), committee

algorithms *meta*), multi-instance learn algorithms (*mi*), rules generation algorithms (*rules*), trees generation decision algorithms (*trees*), and miscellany algorithms *misc*). We tried with the classifiers of each strategy and the best result was obtained with the *Decorate classifier*. Decorate classifier is a meta-learner, classified within committee strategies. Decorate classifier can use different base classifiers. In this research we use as base classifier the multiclass decision tree: LADTree. This base classifier uses the algorithm LogitBoost, the algorithm is as follows:

Algorithm 1.1: Algorithm LogitBoost.

Result: **If** $p(1|a) > 0.5$ predict the first class, if not the second

```
1 begin
2   for j = 1 until t do
3     for a[i] do
4       Assign the objective value for the regression to
        z[i] = (y[i] − p(1|a[i]))/(p(1|a[i]) ∗ (1 − p(1|a[i])))
5       Assign the weight of the instance to
        [i] = p(1|a[i]) ∗ (1 − p(1|a[i])
6     Fit a regression model f_j at datas with class values z[i] and weight
      w[i]
7
8 end
```

4.3 Patterns Assessment

For testing, we used four dataset of 20 instances each one. The following tables show the results:

inst#	actual	predicted	error		inst#	actual	predicted	error
1	1:B	1:B			11	2:N	2:N	
2	1:B	1:B			12	3:M	3:M	
3	2:N	2:N			13	2:N	2:N	
4	2:N	2:N			14	1:B	1:B	
5	3:M	3:M			15	2:N	2:N	
6	2:N	2:N			16	3:M	3:M	
7	2:N	2:N			17	3:M	3:M	
8	1:B	1:B			18	2:N	2:N	
9	2:N	2:N			19	1:B	1:B	
10	3:M	3:M			20	2:N	1:B	+
Correctly Classified 19 <−− 95%								
Incorrectly Classified 1 <−− 5%								

In the first dataset, of 20 instances 5 are benign (B), 10 normal (N) and 5 malignant (M). We obtained 95 % correctly classified instances, i.e. 19 of 20.

inst#	actual	predicted	error		inst#	actual	predicted	error
1	2:N	2:N			11	3:M	3:M	
2	3:M	3:M			12	2:N	2:N	
3	3:M	3:M			13	2:N	2:N	
4	2:N	2:N			14	1:B	1:B	
5	1:B	1:B			15	2:N	2:N	
6	2:N	1:B	+		16	3:M	3:M	
7	1:B	1:B			17	2:N	2:N	
8	1:B	1:B			18	3:M	3:M	
9	2:N	2:N			19	1:B	1:B	
10	2:N	2:N			20	1:B	1:B	

Correctly Classified 19 <-- 95%
Incorrectly Classified 1 <-- 5%

In the second dataset, of 20 instances 6 are benign (B), 9 normal (N) and 5 malignant (M). We obtained 95 % correctly classified instances, i.e. 19 of 20.

inst#	actual	predicted	error		inst#	actual	predicted	error
1	1:B	2:N	+		11	3:M	3:M	
2	2:N	2:N			12	3:M	3:M	
3	3:M	3:M			13	2:N	2:N	
4	2:N	2:N			14	1:B	1:B	
5	2:N	2:N			15	1:B	1:B	
6	2:N	2:N			16	3:M	3:M	
7	3:M	3:M			17	2:N	2:N	
8	2:N	2:N			18	3:M	3:M	
9	1:B	3:M	+		19	1:B	1:B	
10	2:N	2:N			20	1:B	1:B	

Correctly Classified 18 <-- 90%
Incorrectly Classified 2 <-- 10%

In the third dataset, of 20 instances 6 are benign (B), 8 normal (N) and 6 malignant (M). We obtained 90 % correctly classified instances, i.e. 18 of 20.

inst#	actual	predicted	error		inst#	actual	predicted	error
1	1:B	1:B			11	2:N	2:N	
2	2:N	2:N			12	2:N	2:N	
3	3:M	3:M			13	3:M	3:M	
4	3:M	3:M			14	2:N	2:N	
5	3:M	3:M			15	2:N	2:N	
6	2:N	2:N			16	2:N	2:N	
7	2:N	2:N			17	2:N	2:N	
8	3:M	3:M			18	2:N	2:N	
9	2:N	2:N			19	2:N	2:N	
10	2:N	2:N			20	1:B	1:B	

Correctly Classified 20 <-- 100%
Incorrectly Classified 0 <-- 0%

Finally, in the fourth dataset, of 20 instances 2 are benign (B), 13 normal (N) and 5 malignant (M). We obtained 100 % correctly classified instances, i.e. 20 of 20.

We have four incorrectly classified instances (see table 2). 2 instances are benign (B) and 2 are normal (N). Instances malignant (M) indicate that breast cancer does exist. In table 2 we can see that none malignant instance is incorrectly classified, which implicates that the classificator performs a correct diagnoses of all malignant instances as shown in table 3.

Given above results we are motivated to use this classificator with another mastography databases in order to verify the correct performance of our classificator.

Table 2. Instances incorrectly classified.

dataset	inst#	actual	predicted	error
1	20	2:N	1:B	+
2	6	2:N	1:B	+
3	1	1:B	2:N	+
3	9	1:B	3:M	+

Table 3. Classification of M instances.

dataset	inst#	actual	predicted	err		dataset	inst#	actual	predicted	err
1	5	3:M	3:M			3	7	3:M	3:M	
1	10	3:M	3:M			3	11	3:M	3:M	
1	12	3:M	3:M			3	12	3:M	3:M	
1	16	3:M	3:M			3	16	3:M	3:M	
1	17	3:M	3:M			3	18	3:M	3:M	
2	2	3:M	3:M			4	3	3:M	3:M	
2	3	3:M	3:M			4	4	3:M	3:M	
2	11	3:M	3:M			4	5	3:M	3:M	
2	16	3:M	3:M			4	8	3:M	3:M	
2	18	3:M	3:M			4	13	3:M	3:M	
3	3	3:M	3:M							

5 Conclusions and Future Work

In this research we combined the *Mejía's filtering method* and the *Decorate classificator* in the *KDD process* to detect cancer in digital mammograms; the results show a satisfactory detection of breast cancer using the MIAS database.

However, as future work we want to get another mammogram databases to repeat the test and corroborate that this approach provides a good breast cancer detection in general cases.

References

1. Sanghamitra Bandyopadhyay, Ujjwal Maulik, Lawrence B. Holder, and Diane J. Cook. Advanced methods for knowledge discovery from complex data. 2005. Springer.
2. Remco R. Bouckaert, Eibe Frank, Mark Hall, Richard Kirkby, Peter Reutemann, Alex Seewald, and David Scuse. Weka manual for version 3-6-0. 2008. University of Waikato, Hamilton, New Zealand.
3. Ernesto Coto. Methods for medical image segmentation. 2003. Springer-Verlag.
4. DDSM. Ddsm: Digital database for screening mammography. 2009. University of South Florida Digital Mammography Home Page http://marathon.csee.usf.edu/Mammography/Database.html.
5. Eddy Sánchez de la Cruz and Jorge P. Torres. Sia in the diagnosis of cancer. 2009. National Congress of Computers and Systems (CONAIS). Villahermosa, Tabasco, México.
6. N. D. Duque, J. C. Chavarro, and R. Moreno. Seguridad inteligente. 2007. Science Et Technica , Vol. 13, No. 35.
7. Ahmed Farag and Samia Mashali. Dct based features for the detection of micro-calcifications in digital mammograms. 2004. Univ of Texas at El Paso. IEEE.
8. Morales Gonzalez and Aurora B. Image feature extraction of bone marrow cells for the classification of acute leukemias. Master's thesis, 2006. National Institute of Astrophysics, Optics and Electronics.
9. José Molina López and Jesús García Herrero. *TECHNICAL ANALYSIS OF DATA - PRACTICAL APPLICATIONS USING MICROSOFT EXCEL And WEKA.* Universidad Carlos III from Madrid, 2004.
10. José M. Mejía Mu noz. The nonsubsampled contourlet transform for enhancement of microcalcifications in digital mammograms. 2009. 8th Mexican International Conference on Artificial Intelligence MICAI-2009. Guanajuato, México.
11. WEKA University of Waikato. Weka 3: Data mining software in java. 2010. http://www.cs.waikato.ac.nz/ ml/weka/.
12. Iván Olmos, Jesús González, and Mauricio Osorio. Mining common patterns on graphs. 2005. International Conference on Computational Intelligence and Security, Lecture Notes in Artificial Intelligence, Vol. 3802, 41-48, Springer Verlag.
13. Abraham Silberschatz, Henry Korth, and S. Sudarshan. *Database Fundamentals.* McGrawHill, 4a Ed., 2002.
14. Universidad-Jaén. Noise reduction in a digital image. 2006. Electronic Engineering Department. Area Systems and Automation Engineering. Pg. 2-7.
15. César Cardona Valencia. Evaluation of algorithms based on fuzzy logic applied to the preprocessing and edge detection in digital images, 2004.