SDCA: System to detect cancerous abnormalities

Eddy Sánchez de la Cruz[1], Homero Alpuín-Jiménez[1], Humberto de Jesús Ochoa Domínguez[2], and Pilar Pozos-Parra[1]

[1] Universidad Juárez Autónoma de Tabasco. Cunduacán, Tabasco, México
eddsacx@gmail.com, {homero.alpuin,pilar.pozos}@dais.ujat.mx
[2] Universidad Autónoma de Ciudad Juárez.Ciudad Juárez, Chihuahua, México
hochoa@uacj.mx

Abstract. In this article we present SDCA, which is a system to detect cancerous abnormalities in digital mammograms. The SDCA try to give at radiologist a second opinion in the analysis of a digital mammogram to increase the reliability of detecting breast cancer. SDCA is a semi-automation of KDD process (*Knowledge Discovery in Databases*). The KDD process is a method that uses strategies of Artificial Intelligence (AI) to extract patterns of behavior in databases with large volumes of information. Two SDCA characteristics outstanding are 1) the implementation of *Mejía filtering method* in the data cleansing module, and 2) the implementation of *Decorate strategy Classification* in the classification module.
The results shows that SDCA get 95% of detections classified correctly. SDCA was developed using Matlab GUIDE, and tests were done with the database (DB) of digital mammographic MIAS.

Key words: SDCA, KDD process, detect cancerous abnormalities, digital mammograms, classification.

1 Introduction

For the radiologist, mammograms are highly useful to identify abnormalities carcinogenic potential. The difficulty arises when the radiologist's review does not guarantee the detection of cancerous abnormalities. Therefore, this research serves as a support in the detection of abnormal regions. Therefore, the area of interest for this research is the analysis of medical images using the KDD process.
This research, then, serves to give the radiologist a second opinion on the detection of abnormal and thus increase the reliability of diagnosis.
SDCA is an extension of work previously presented in [11] and [10]. This article describes the segmentation, filtering and classification modules.
The rest of the paper is divided as follows: Section 2 briefly describes the KDD process for this research. Section 3 describes the filter module. Section 4 describes the segmentation module. Section 5 describes the classification module.

Section shows experimental results and finally in section ends with a conclusion and future work.

2 KDD process

The following describes the KDD process steps for this research:

– Selection: Given a set of different digital mammography DB, the most representative was selected respect its use in other research. MIAS (see Table 1) is a reduced version of the original, and for a long time, been used to test research in [9], [1], [3], [8], [7], and also strongly recommended by the University of South Florida [5].

Name	Description
MIAS	Small database that contains the same images as the original version, but reduced to a size of 1024 x 1024 pixels, available in `http://www.wiau.man.ac.uk/services/MIAS/MIASmini.html`

Table 1. DB MIAS

– Data Preparation: The images are filtered using the *Mejía filtering method*, to reduce noise.

– Data Transformation: Segmented manually, the area of interest by the radiologist, normalizes each segmented area to have the same dimensions and finally gives the frequency histogram of gray levels of each segmented area, and stored for create a testing base.

– Data Mining: new samples are obtained and applied Decorate classification strategy to classify the abnormality, if it exists.

– Patterns evaluation: The patterns obtained are evaluated by the radiologist.

3 Filtering module

This module was integrated into the prototype, because with this filtering method is obtained excellent results in the enhancement of abnormalities. This work was presented in [7]. The method works by using the Transform Contourlet Nonsubsampled (NSCT) and the Prewitt filter. The method is based on the classical approach used in the processing methods for image processing.

For a more detailed explanation see [7].

4 Segmentation module

In this module, the radiologist manually segmented the area he considers abnormalities. The result is shown in Fig. 2.

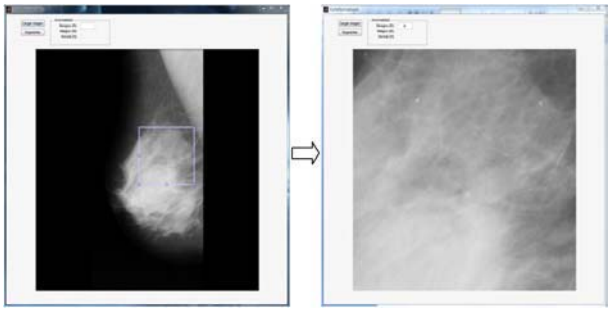

Fig. 1. Segmentation module

Then the area of interest is normalized so that all images have the same dimensions.

After being normalized interest area, gives the frequency histogram of gray levels, ranging from 0 to 255 (see Fig. 3).

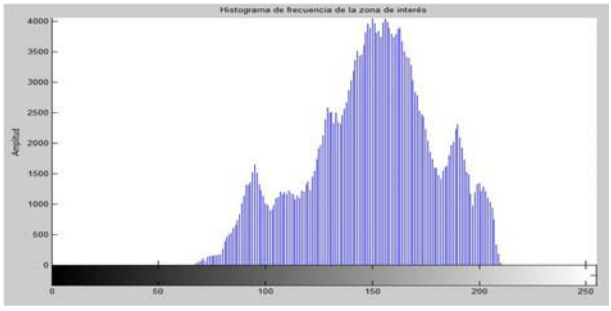

Fig. 2. Frequency histogram of gray levels

This histogram is stored in a xlsx file called *Histograma.xlsx* for purpose of building the testing base.

Then the radiologist choose a tag if the detected abnormality is benign (B), malignant (M) or normal (N).

Finally for this module, open the file *Histograma.xlsx* and saved in CSV format (comma delimited) for later use.

5 Classification module

In this module selects the *Histograma.csv* file to implement the classification of the abnormality:

The classification is activated by pressing the **Detección** button, which indicates the call to Decorate strategy, which uses, in this case, the algorithm LADTree base, and this, in turn, is based on the algorithm LoogitBoost 1.1.

The basic algorithm LoogitBoost:

Algorithm 1.1: LogitBoost algorithm

Result: **If** $p\,(1|a)$>0.5 predict the first class **Else** the second

begin
  for $j = 1$ *until* t **do**
    for $a\,[i]$ **do**
      Assign the target value for the regression to $z\,[i] = (y\,[i] - p\,(1\,|a\,[i]))\,(p\,(1\,|a\,[i])\Lambda\,(1 - p\,(1\,|a\,[i]))$ Assign the weight of the instance to $[i] = p\,(1\,|a\,[i])\Lambda\,(1 - p\,(1\,|a\,[i]))$
      Fit a regression model f_j at data with class values $z\,[i]$ and weights $w\,[i]$
end

Finally, the result is released (see Fig. 6) for analysis by the radiologist.

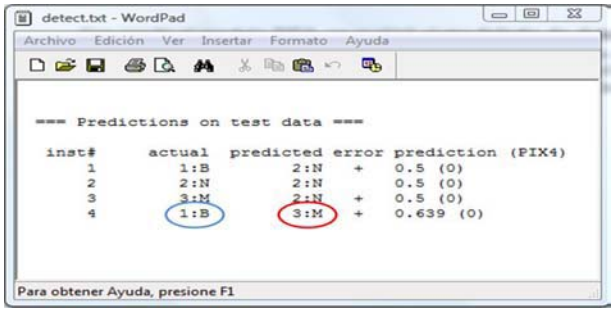

Fig. 3. Classification results

This table shows four predictions made by SDCA, the first three samples are needed for the proper operation of the system, however, the new sample, which interests the radiologist is the fourth, and in this new sample the radiologist predicts the area of interest was the first type, i.e., benign (1: B) (blue circle), but the system indicates that normality can be either three, or malignant (3: M) (red circle). This aid increase the reliability, as is now required to take another mammogram from another angle to confirm whether the abnormality is type 3: M.

6 Results

Classification Algorithms We tested each classification algorithm for each strategy, the strategies are: bayesian algorithms (*bayes*) classification functions (*functions*), algorithms to generate rules (*rules*), meta classifiers (*meta*), lazy algorithms (*lazy*), algorithms generation of decision trees (*trees*) and miscellaneous algorithms (*misc*).

We used the 322 digitized mammograms of MIAS BD (Table 1) for the first test and even to choose the algorithm that performed better. As shown in the table below, the best result was obtained with the strategy Decorate. This strategy was 95% of instances correctly classified.

Strategy	**Algorithm**	**CCI***	**% of CCI***
Bayes	NaiveBayes	13	65%
Functions	Logistic	14	70%
Rules	PART	12	60%
Meta	Decorate	19	95%
Lazy	IB1	12	60%
Trees	RandomForest	17	85%
Misc	HyperPipes	10	50%

* CCI ← Correctly Classified Instances

Experiments and analysis For testing we used a data sample of eighty, divided into four sets of twenty instances each. These results have been presented previously [10] and [11].

In the first dataset is obtained 95%; in the second dataset is obtained, similarly, 95%; in the third dataset is obtained 90% and finally, in the fourth dataset is obtained 100% of instances correctly classified. These results fluctuate between 90% and 100%, giving an average of 95% of instances correctly classified. This shows that SDAC is very reliable to use in the classification of cancerous abnormalities.

Evaluación. En los últimos años se han presentado aportaciones para ayudar al radiólogo en el diagnóstico de cáncer de mama. Las pruebas en estos trabajos se han realizado con diferentes BDs.
En la tabla 2, vemos que [6] obtuvo 95.35%, sin embargo, la BD nos dice que estos resultados son propios para la comunidad española y, además, el tamaño de la muestra es pequeño. [8] obtuvo 91% utilizando la misma BD que se usa en esta investigación, sin embargo, el tamaño de la muestra es muy pequeño. En [2] se obtuvo 73% utilizando una BD diferente, resultados realmente bajos y, además, no se menciona el tamaño de la muestra. Finalmente, en esta invetigación, se obtiene 95%, resultados satisfactorios, teniendo en cuenta que la BD utilizada goza de amplia aprobación por la comunidad científica, y que el tamaño de la muestra es considerable. Por lo que se concluye que SDAC es altamente fiable para apoyar el diagnóstico del radiólogo.

Método	**año**	**BD utilizada**	**Tamaño muestra**	**% de ICC***
SDAC	2011	MIAS	80	95%
[2]	2008	DDSM[5]	s/n	73%
[8]	2008	MIAS	30	91%
[6]	2005	Hospital Puerta de Hierro de Madrid	43	95.35%

* ICC ← Instancias Correctamente Clasificadas

Table 2. Comparación de resultados obtenidos con otros trabajos previamente reportados en la literatura

7 Conclusión y trabajos futuros

En este trabajo, presentamos SDCA para detectar anormalilades cancerígenas en mastografías digitales. SDCA es una semi-automaización del proceso KDD. Los resultados obtenidos muestran que SDCA aumenta la fiabilidad en la detección de anormalidades cancerígenas, dando al radiólogo una segunda opinión sobre la revisión de la mastografía.

Como trabajo futuro se propone aumentar el número de datos de pruebas, para verificar que se mantenga el promedio de fiabilidad alrededor de 95%. Además, teniendo en cuenta los buenos resultados aqui obtenidos e inspirados en [4], quienes primero trabajaron con mastografías y luego con Pap Smear Microscopic Image, se pretende migrar el proceso KDD para detectar cáncer cervicoúterino.

References

1. M. Antonie, O. Zaiane, and A. Coman. Application of data mining techniques for medical image classification. *In Proc. Of Second Intl. Workshop on Multimedia Data Minino (MDM/KDD2001) in conjunction with Seventh ACM SIGKDD*, pages 94–101, 2001.
2. Enrique Calot, Hernán Merlino, and Paola Britos Ramón García-Martínez. Clasificación de tumores en mamografías mediante uso combinado de rbp y filtros sobel. 2008.
3. Ahmed Farag and Samia Mashali. Dct based features for the detection of microcalcifications in digital mammograms. 2004. Univ of Texas at El Paso. IEEE.
4. Francisco Gallegos-Funes, Margarita Gómez-Mayorga, José Lopez-Bonilla, and Rene Cruz-Santiago. Rank m-type radial basis function (rmrbf) neural network for pap smear microscopic image classification. *C. Roy Keys Inc. http://redshift.vif.com*, 16:4, 2009.
5. M. Heath, K. Bowyer, D. Kopans, R. Moore, and W. Kegelmeyer. The digital database for screening mammography (ddsm). *Medical Physics Publishing. ISBN: 1-930524-00-5*, pages 212–218, 2001. University of South Florida Digital Mammography Home Page `http://marathon.csee.usf.edu/Mammography/Database.html`.
6. José Avelino Manzano Lizcano, Laura Moyano Pérez, and Carmen Sánchez Ávila. Sistema para la detección automatizada de microcalcificaciones en mamografía digitalizada utilizando la transformada contourlet. *Congreso de Métodos Numéricos en Ingeniería. ISBN: 978-607-7557-71-5*, pages 2–11, 2005.
7. 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.
8. Lorena Vargas Quintero, Leiner Barba Jiménez, Cesar Torres, and Lorenzo Mattos. Transformada wavelet y técnicas de filtrado no lineal aplicadas a la detección de microcalcificaciones en mamografías digitales. *Memorias. XIII Simposio de Tratamiento de Señales, Imágenes y Visión Artificial STSIVA. ISSN 978-958-8477-00-8*, II:23–26, 2008.
9. R. Rangayyan, N. El-Faramawy, Leo Desautels, and O. Alim. Measures of acutance and shape for classification of breast tumors. *IEEE Transactions on Medical Imaging*, 16:799, 1997.
10. Eddy Sánchez, Pilar Pozos-Parra, and Homero Alpuín-Jiménez. Cancer detection using the kdd process. *Advances in Soft Computing Algorithms. ISSN: 1870-4079*, 49:109–117, 2010.
11. Eddy Sánchez, Pilar Pozos-Parra, and Homero Alpuín-Jiménez. Detección de cáncer de mama usando el proceso kdd en mastografías digitales. *Avances en Informática y Sistemas Computacionales. ISBN: 978-607-7557-71-5*, V:40–51, 2010.