Decisions Fusion Strategy: Towards Hybrid Cluster Ensemble

Syed Zahid Hassan [1], Brijesh Verma [2]
[1,2] School of Computing Sciences
Central Queensland University, Queensland, Australia
[1] z.hassan@cqu.edu.au, [2] b.verma@cqu.edu.au

Abstract

Clustering ensembles have renowned as a powerful method for improving both the performance and constancy of unsupervised classification solutions. However, finding a consensus clustering from multiple algorithms is a difficult problem that can be approached from combinatorial or statistical perspectives. We offer a new clustering strategy which is formulated to cluster extracted mammography features into soft clusters using unsupervised learning strategies and 'fuse' the decisions using majority voting and parallel fusion in conjunction with a neural classifier. The idea is to observe associations in the features and fuse the decisions (made by learning algorithms) to find the strong clusters which can make impact on overall classification accuracy. Two novel techniques are proposed for fusion, majority-voting based data fusion, and neural-based fusion. The proposed approaches are tested and evaluated on the benchmark database— digital database for screening mammograms (DDSM). This study compares the performance of the proposed ensemble approach with other fusion approaches for clustering ensembles. Experimental results demonstrate the effectiveness of the proposed method on benchmark dataset.

1. INTRODUCTION

The clustering techniques in data mining are mainly used for data exploration purposes. The clustering algorithms, such SOM, k-Means etc., provide user with the ability to transform raw data into high-level useful knowledge for decision making. More specifically, clustering typically partitions the data based on the attributes similarities (generally using distance function), and group them into a different sets (clusters). Discovery of the 'useful' clusters that make impact on the system performance is the profound research problem which is investigated by the various data mining community [1, 2].

The partitioning of a population of individual into similarity groups has drawn prominent attention in various disciplines, such as biology, medical, finance, marketing etc., and some interesting results have been reported by the bio-informatics researchers [3, 4]. In [5], it is demonstrated that how clustering algorithms can make impact on gene functions discovering and find the sample tissues to identify the causes of cancer.

Despite the existence and successful deployment of various clustering algorithms in a large number of real world problems, yet, there is no single clustering algorithm able to recognize heterogeneous data structure [6]. Generally, all clustering algorithms presume a homogenous clustering criterion over the entire feature space; consequently, all the discovered clusters are similar in properties [6]. The basic limitation associated with every individual clustering algorithm is their inability to identify clusters with different properties. Every clustering algorithm has its own clustering criterion that is relevant to a particular problem domain. They are unable to reveal fundamental distribution of all types of data. On the other hand, it's unfeasible to draw a priori information, which clustering method is appropriate for underlying structure present in the data population [7].

In [6], Martin et al, reported some interesting results on individual clustering algorithm. They monitored the inability of individual clustering algorithm while dealing with data sets which were diverse in nature. Two clustering algorithms: k-Means and single-link algorithms were considered, to find two Spirals and two Globular clusters. It was observed that none of the clustering algorithm was able to discover given three clusters. Interestingly, those globular clusters were successfully detected by the k-Means and a spiral cluster by single-link.

Lately, the need of a combination of diverse clustering algorithms has widely been recognized. The numbers of hybrid clustering endeavours have been initiated all over the globe. Two clustering notions, multi-objective clustering and cluster-ensemble, are extensively reported in the literature [8, 9]. In this research we focus on the latter technique. In cluster-ensemble, a set of classifiers are incorporated by the ensembler, whereby individual classifier's decisions are typically combined by weighted/ unweighted voting to discover new clusters [10].

Application of cluster-ensemble techniques have started to emerge in several application domains, such as medical diagnostics [11], image classifications [12], document clustering [13] etc. Notably, the sstructure of medical data repositories, which consist of complex, large and unlabelled data samples, seems to be a good candidate for unsupervised learning algorithms. The unsupervised learning algorithms such as self-organizing map (SOM), k-Means, k-NN, have been reported in various medical data mining literatures ranges from feature selection, extraction, classification to data visualization. For example, self-organizing map (SOM) is used to identify the clusters in breast cancer diagnosis [14], to predict biopsy outcomes [15] and to model selection of mammography features [16].

To this end, we present a novel combination of data mining algorithms, such as self-organizing map, k-Means, and multi-layer perceptron, in order to utilize the strengths of each individual technique and compensate for each other's weaknesses in data clustering.

This paper is divided into five sections. Section 2 discusses our proposed ensemble methodology. Section 3 explains the database used for experiment purpose. Section 4 presents the preliminary experiments and results, evaluates the performance of the proposed approach by performing quantitative analysis and discussion on the results achieved. The conclusions and future directions are presented in section 5.

2. PROPOSED METHODOLOGY

We propose our data clustering strategy by forming the unique cluster ensemble that utilizes the strengths of various unsupervised and supervised clustering algorithms, such as self organising map, k-Means, multi-layer perceptron etc, to enhance medical system' decisions-making capability. These algorithms vary in their methods of search and representation to ensure diversity in the errors of the learned models.

There are two types of hybrid strategies proposed and investigated in this paper: majority-voting based data fusion and neural-based data fusion, as depicted in Figure1.

In former approach, we created an ensembler of unsupervised clustering algorithms, whereby the decisions of each individual algorithm are combined by using a simple majority-voting scheme. Notably, the decisions are combined on the tested data samples. In this majority-voting scheme, each algorithm assigns the confidence level to its generated clusters based on the maximum cases one cluster contains. More specifically, the strong clusters are obtained by calculating the both maximum values clusters contain in decision matrix and by measuring the accuracy of each individual algorithm that contributes in decision-making.

In latter approach, neural-based data fusion, MLP is incorporated with unsupervised ensembler as a classifier. The MLP classifier learns with the soft clusters, generated by the clustering algorithms, and classifies them into appropriate classes, i-e benign and malignant. MLP's result can later be explored for further investigation and decision-making.

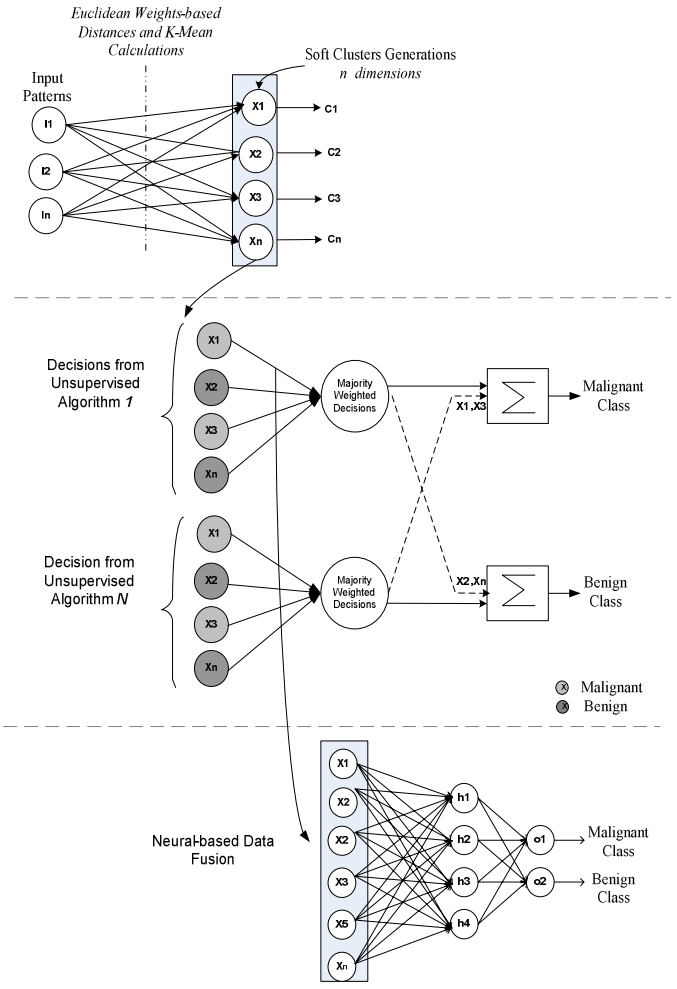

Figure1. Proposed Data Clustering Strategy for Decision Making

To explicate our methodology further, we formed an unsupervised clustering-ensembler by using self-organising map and k-Means clustering criterions. The proposed ensembler calculates the Euclidean distance and k-Means values in the input cases and generates various soft clusters. For instance,

We have input patterns: $IP = [I_1, I_2, I_3, ... I_n]$ and;

Weight Values: $W = [W_1, W_2, W_3, ... W_n]$

We calculate the Euclidean distances to find the winning units in SOM map:

$$ED = (I_1 - W_1)^n + (I_2 - W_2)^n ... + (I_n - W_n)^n$$

The output unit who has the least Euclidean distance is considered as winning or image unit for the input case.

$$ED = \| IP - W \|$$

Note: to generate number of soft clusters, we design the output units almost twice the dimension of input features spaces. Functionally, SOM consisted of 16 neurons partitioned in a single layer in a 2-D grid of 4 x 4 neurons.

We construed and assigned the random reference input vectors (neuron weights) to each partition. For each input, the Euclidean distance between the input and each neuron was calculated. The reference vector with minimum distance is identified. After the most similar case is determined, all the neighbourhood neurons, connected with the same link, adjust their weight with respect to the reference vector to form a group in two dimensional grids. The whole process is repeated several times, decreasing the amount of learning rate to increase the reference vector, until the convergence is achieved.

In k-Means criterion, we randomly partitioned the input data into k-cluster centers along with its all closest features. With each input feature, it calculates the mean point of each feature and constructs a new partition by associating data-entities to one of the k clusters. Cluster features are moved iteratively between k clusters and intra-and-inter-cluster similarity. Distances are measured at each move. Features remained in the same cluster if they were closer to it otherwise move into new cluster. The centers for each cluster are recalculated after every move. The convergence achieved when moving object increased intra-cluster distances and decreases inter-cluster dissimilarity. The whole k-Means process can be represented as:

$$V = \sum_{i=1}^{k} \sum_{x \in s} |x - \mu_i|$$

Where V is the vector space with k clusters, S_i, $i = 1,2,...,k$ and μ_i is the centroid or mean point of all the points $x \in s$.

The clustering decisions made by SOM and K-Means can be demonstrated by using decision matrix:

$$\text{SOM and K-Means} = \begin{matrix} C1 & C2 & C3 & C4 & Cn \end{matrix} \begin{bmatrix} X_1 & X_2 & X_3 & X_4 & X_N \\ Y_1 & Y_2 & Y_3 & Y_4 & Y_N \end{bmatrix}$$

C1-Cn represents output clusters generated by SOM/k-Means algorithm, whereby *X1-Xn* and *Y1-Yn* represent individual algorithm's decision for each case. For instance, *X1-Xn* are the clusters that represent cases which are classified as a class malignant, whereas *Y1-Yn* clusters represent cases which belong to class benign.

To find the strong clusters in each decision matrix formed by individual clustering algorithm; we calculate the maximum number of cases that one cluster contains, check which class this cluster belong to and then compare it with the other clusters. For example, the maximum values can be found by using simple comparison rules,

If X1 > Y1 the case belong to class malignant otherwise benign

Y1 > X1 the case belong to class benign otherwise malignant

After finding the maximum values, we combine and fuse all the strong clusters to generate the final cluster. This cluster exhibits particular class. The idea of this majority based fusion is to take benefits from those clusters which were misclassified by different clustering algorithms.

Alternatively, we also investigated the neural based data fusion. For this reason, we designed a multi-layer perceptron neural network, which was fully connected, layered, feed-forward architecture. After finding the strong clusters (with maximum cases), we feed those cases into MLP for further classification.

3. DATA REPOSITORY

Dataset of digital mammograms is used in this research and it is taken from Digital Database for Screening Mammography (DDSM) established by University of South Florida. The underlying structure of this database is heterogeneous in nature, contains categorical, textual and images data.The DDSM database contains approximately 2,500 case studies, whereby each study includes two images of each breast, along with some associated patient information (age at time of study, breast density rating, subtlety rating for abnormalities, keyword description of abnormalities) and image information (scanner, spatial resolution etc). The database contains a mixture of normal, benign, benign without call-back and cancer volumes selected and digitized. Images containing suspicious areas have associated pixel-level information about the locations and types of suspicious regions.

For evaluation purpose, we considered data set that consists of six features (measurements) from 200 mammograms cases: 100 benign and 100 malignant. The 100 mammograms were used for training purposes and 100 for testing. Typically, features presented in our data sets includes: Patients Age, Density, Shape, Margin, Assessment Rank and Subtlety. The input data contain raw data as well as extracted features which are used as an input to the clustering algorithms. The input data are normalized between 0-1.

4. EXPERIMENTS RESULTS AND DISCUSSION

The proposed approach is tested on a benchmark database in order to evaluate the system performance and accuracy. The experimental results are presented below in Tables 1 and 2.

From the comparative results shown in Table 1, it is observed that our proposed cluster ensemble approach provides better results than the stand alone individual technique. It is also noticed that the proposed approach outperforms all individual approaches in all main output categories (see Table 1): classification accuracy, misclassification accuracy and error rates. Out of the total of 100 digital mammogram cases of the test dataset, SOM made 12% misclassifications; K-Means made 16% misclassifications, and interestingly MLP also misclassified 12% cases. The proposed majority-voting based approach made 7 misclassifications. Markedly, the proposed neural-based data fusion does not show any errors and give 100% performance. The reason to achieve maximum performance from MLP-based data fusion could be the small amount of data sets. However, this interesting result is under investigation. The classification accuracies achieved by SOM,

k-Means, MLP and proposed approaches are 88%, 84%, 88%, and 93% and 100% respectively.

TABLE 1
RESULTS SHOWING THE IMPROVEMENT IN CLASSIFICATION ACCURACIES

Algorithms	Classification Error [%]	Root Mean Square Error	Classification Accuracy [%]
SOM	12	0.2777	88
K-Means	16	0.2433	84
MLP	12	0.2666	88
Proposed (Majority Weighted-Voting Scheme)	7	N/A	93
Proposed (Neural - based Fusion)	0	0	100

The experiments were also performed comparing the accuracies of algorithms by individual class: benign and malignant. For each class, the ROC analysis attributes, such as TP rate, FP rate, and F-measure, are measured with particular algorithm as shown in Table 2. It is noticeable that the attributes frequency measures for both classes benign and malignant are quite high with the proposed ensemble approach.

TABLE 2
DETAILED ACCURACY BY CLASSES: TP = TRUE POSITIVE RATE; FP = FALSE POSITIVE RATE AND F-MEASURE= FREQUENCY MEASURE OVER CLASS ACCURACY

	Classes	TP Rate	FP Rate	F-Measure
Individual Algorithm	Benign (SOM)	0.81	0.04	0.87
	Malignant (SOM)	0.96	0.24	0.88
	Benign (k-Means)	0.94	0.14	0.84
	Malignant (k-Means)	0.86	0.06	0.83
	Benign (MLP)	0.82	0.06	0.87
	Malignant (MLP)	0.94	0.18	0.88
Proposed Majority-Voting Approach	Benign	NA	NA	93
	Malignant	NA	NA	93
Proposed Neural-based Fusion Approach	Benign	0	0	100
	Malignant	0	0	100

We created a confusion matrix to evaluate individual classifier performance by displaying the correct and incorrect pattern classifications. Typical Confusion Matrix can be represented as:

Confusion Matrix

a	b	<-------- Classified as
x1	x2	a = Malignant
y1	y2	b = Benign

Where row (x1 and x2) represents the actual patterns and column (x1 and y1) represents the classified patterns for class a (Malignant). The difference between the actual patterns and the classified patterns can be used to determine the performance of a classifier.

To explicate it further, we draw the Confusion Matrix for each classifier to evaluate how many patterns in a given class are classified correctly/incorrectly.

SOM Confusion Matrix

a	b	<-------- Classified as
48	2	a = Malignant
10	40	b = Benign

This SOM classifier successfully classified 88 cases out of 100 cases presented. The row values (48, 2) are the actual cases for the class malignant, and row values (10, 40) represent the actual class benign. However, the classified outputs are represented by column a (48, 10) and column b (2, 40). The comparison of these rows and columns, between actual pattern and classified patterns, can provide interesting insights. For instance: for the malignant class accuracy, we notice that the original malignant patterns were (48, 2) and the classifier indicates (48, 10). Thus, it classified 48% cases correctly as a malignant class and misclassified 2 cases. It is also noticeable that those two patients will be given clear when they were supposed to be treated like a cancer patients. Similarly, for the benign class accuracy, the actual cases are (10, 40) and whereas the classifier indicates (2, 40). The 40% cases were classified correctly as a class benign and 10% cases were misclassified. In this scenario, those 10 patients who are not the victim of cancers will be treated like a cancer

patient despite it being the opposite scenario. However, the overall outcome is much more favourable: 48% classified correctly as a malignant class and 40% classified correctly as a benign class.

K-Means Confusion Matrix

a	b	<-------- Classified as
38	11	a = Malignant
5	46	b = Benign

By applying the above-mentioned confusion matrix method on the K-Means classifiers, the 38% cases were classified correctly as a class malignant (11 cases were misclassified) and 46% cases classified correctly as a class benign (misclassified 5 cases), overall achieved 84% classification accuracy.

MLP Confusion Matrix

a	b	<-------- Classified as
47	3	a = Malignant
9	41	b = Benign

Individual MLP classified 47% and 41% cases correctly as a class malignant and benign respectively, with the ratio of 3 misclassified cases of a class malignant and 9 cases for a class benign, overall computed 88% accuracy.

From the decision-making perspective, it's also noticeable that by fusing the outputs of all clustering algorithms, based on majority-voting and neural-based data fusion methods we obtained final clusters which are adequately/accurately classified, such as demonstrating 93% and 100% accuracies respectively.

A. Comparative Analysis with Relevant Approaches

Performing comparative analysis with other existing approaches for digital mammograms classification is a challenging task. Most of the proposed approaches found in the literature are generally tested on digital mammograms but adopted different databases or same databases with different features compare to what we have tested with our approach. However, there are some approaches which are quite relevant to our work have demonstrated some interesting facts. In [17], Mahmoud proposed the approach for the classification of tumors (masses) in mammograms using two segments approach. In the first stage, he extracted mammography features by using a combination of morphological operations and a region growing technique. In the second phase, segmented regions are classified as normal, benign, or malignant tissues based on different measurements (shape, intensity variation, spread pattern etc). Experiments were performed on mammogram images of the MIAS database and 82.9% classification accuracy was claimed, as show in Figure2. Anna in [18], investigated the texture properties of the tissue surrounding microcalcification and their contribution towards breast cancer diagnosis.

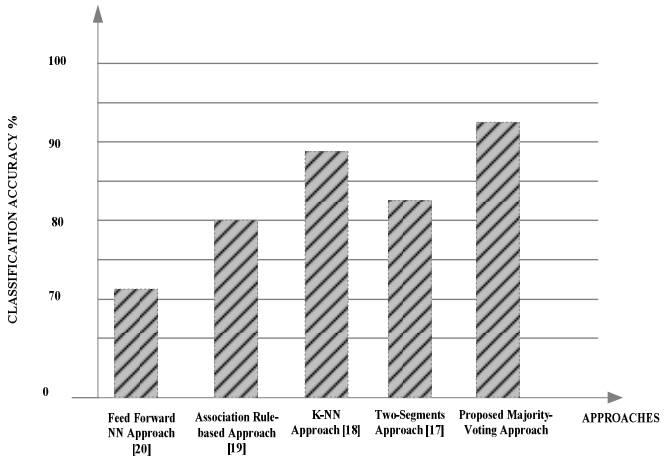

Figure2. Comparison of Proposed Approach with the Existing Approaches

Anna [18], used K-NN approach to discriminate benign and malignant classes in digital mammograms. The Digital Database for Screening Mammography (DDSM) was used, which consisted of 100 mammographics images. The overall classification accuracy demonstrated was 89%, as shown in Figure2. Osmar in [19], deployed a association rule-based classifier for mammography classification and managed to attain over 80% in accuracy. In [20], Keir proposed bootstrap aggregation (bagging) technique to extract features and used feed forward neural network to classify the mammography images, obtained by DDSM. The overall classification accuracy reported on four-classes problem was 71.4%, as shown in Figure2.

5. CONCLUSIONS

The experiments results show that proposed cluster ensemble approach is useful for the analysis of clinical parameters and their combinations for the cancer diagnosis. We demonstrated that with our proposed decision fusions techniques; majority-voting and MLP based neural fusion, the accuracy of the rules used to generate overall diagnosis for the cancer disease was improved. Individual SOM and MLP classifiers achieved accuracies of 88 %, whereby k-Means accuracy was 84 %. The majority-voting and neural-based data fusion schemes reported accuracy of 93 % and 100% respectively, which are very promising. The proposed approach is also able to visualize the data which helps in interpretation of the results.

In future we plan to test our proposed approach on other benchmark databases to evaluate its performance. More experiments are still in progress with different hybrid combinations.

ACKNOWLEDGEMENT

This work is supported by ARC ISSNIP Postgraduate Grant.

REFERENCES

[1] George, D. and Derek, A. (2004), "Linkens: Adaptive systems and hybrid computational intelligence in medicine". Artificial Intelligence in Medicine 32(3): pp. 151-155.

[2] Goonatilake, S. and Khebbal, S. (1995), Intelligent Hybrid Systems. John Wiley and Sons, Chichester.

[3] Narayanan, E. K. A. (2005). Intelligent Bioinformatics: The Application of Artificial Intelligence Techniques to Bioinformatics Problems. John Wiley & Sons.

[4] Wang, J. T. L., Zaki, M. J., Toivonen, H. T. T. & Shasha, D. E. (Eds.) (2003). Data Mining in Bioinformatics. Advanced Information and Knowledge Processing. Springer publishers.

[5] Costa, I.G., Carvalho, F.A.T., & Souto, M. C. P. (2004). "Comparative Analysis of Clustering Methods for Gene Expression Time Course Data". Genetics and Molecular Biology, 27 (4), 623-631.

[6] Martin H. C. Law, Alexander P. Topchy, Anil K. Jain, (2004). "Multiobjective Data Clustering". IEEE computer society conference on computer vision and pattern recognition. Vol 3, pp. 424-430.

[7] Law, M., Topchy, A. & Jain, A. K. (2004). "Multiobjective Data Clustering". In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition: Vol. 2 (pp. 424-430).

[8] Handl, J. & Knowles, J. (2007). "An Evolutionary Approach To Multiobjective Clustering". IEEE Transactions On Evolutionary Computation, 11 (1), 56-76.

[9] Boulis, C. & Ostendorf, M. (2004). "Combining Multiple Clustering Systems". In J. Boulicaut, F. Esposito, F. Giannotti, D. Pedreschi (Eds.), 8th European conference on Principles and Practice of Knowledge Discovery in Databases, Lecture Notes in Computer Science, 3202, 63-74.

[10] Evgenia Dimitriadou, Andreas Weingessel, and Kurt Hornik. (1999). "Voting in Clustering and Finding the Number of Clusters". In H. Bothe, E. Oja, E. Massad, and C. Haefke, editors, Proceedings of the International Symposium on Advances in Intelligent Data Analysis (AIDA 99). pp 291-296. ICSC Academic Press, 1999.

[11] Greene, D., Tsymbal, A., Bolshakova, N., & Cunningham, P. (2004). "Ensemble Clustering in Medical Diagnostics". In Proceedings of the 17th IEEE Symposium on Computer-Based Medical Systems (pp. 576- 581). IEEE Computer Society.

[12] Lourenco, A. & Fred, A. (2005). "Ensemble Methods in the Clustering of String Patterns". In Proceedings of the Seventh IEEE Workshops on Application of Computer Vision: Vol. 1 (pp. 143- 148). IEEE Computer Society, Washington, DC.

[13] Greene, D. & Cunningham, P. (2006). "Efficient Ensemble Methods for Document Clustering". Department of Computer Science, Trinity College Dublin. (Tech. Rep. TCD-CS-2006-48)

[14] Chen, D., Chang, R. F., and Huang, Y. L, (2000), "Breast Cancer Diagnosis Using Self-organizing Map for Sonography". Ultrasound in Medicine Biology vol 26, pp:405–11.

[15] West, D., and West, V. (2000), "Model Selection for a Medical Diagnostic Decision Support System: A Breast Cancer Detection Case". Artificial Intelligence in Medical. Vol 20, pp:183–204.

[16] Pattaraintakorn, P., Cercone, N., and Naruedomkul, K. (2005), "Hybrid Intelligent Systems: Selecting Attributes For Soft-Computing Analysis". In 29th Annual International Computer Software and Applications Conference (COMPSAC). Volume 1, Issue 26, pp: 319 – 325.

[17] Mahmoud R. Hejazi, Yo-Sung Ho (2005). "Automated Detection of Tumors in Mammograms Using Two Segments for Classification". PCM 2005. Vol 1, pp: 910-921

[18] Anna Karahaliou , Ioannis Boniatis, Spyros Skiadopoulos, Philippos Sakellaropoulos, Eleni Likakai, George Panayiotakis, Lena Costaridou (2006). "A Texture Analysis Approach For Characterizing Microcalcifications On Mammograms" . In the international special topic conference on Information technology in bio medicine, pp: 251-257.

[19] Osmar R. Zaļane, Maria-Luiza Antonie, Alexandru Coman, (2002), "Mammography Classification by an Association Rule-Based Classifier". Third International ACM SIGKDD Workshop on Multimedia Data Mining (MDM/KDD'2002) in conjunction with Eighth ACM SIGKDD, pp. 62-69, Edmonton, Alberta, Canada, 17-19 July 2002

[20] Keir Bovis and Sameer Singh (2002). "Classification of Mammographic Breast Density Using a Combined Classifier Paradigm". In 4th International Workshop on Digital Mammography, pp: 177-180.