Cumhuriyet Üniversitesi Fen Fakültesi
Fen Bilimleri Dergisi (CFD), Cilt:36, No: 3 Özel Sayı (2015)
ISSN: 1300-1949

Cumhuriyet University Faculty of Science
Science Journal (CSJ), Vol. 36, No: 3 Special Issue (2015)
ISSN: 1300-1949

An Efficient Method for Detection of Masses in Mammogram Images

Javad HADDADNIA[1,*], Omid RAHMANI-SERYASAT[2], Hossein GHAYOUMI-ZADEH[2], Hamidreza RABIEE[2]

[1.] *Associate Professor, Electrical and Computer Engineering Department, Hakim Sabzevari University, Sabzevar, Iran*

[2.] *Department of Electrical and Computer Engineering, Hakim Sabzevari University, Sabzevar, Iran*

Received: 01.02.2015; Accepted: 05.05.2015

Abstract. Breast cancer is one of the most common cancers among women. Mammography is currently the most effective method for early detection of breast cancer. In this paper, a method is proposed for detecting masses in mammogram images. First, based on a specific algorithm, image is segmented and a number of the suspicious regions are obtained. Then, many features are extracted from these regions. To reduce the features, a supervised feature selection method is used. In the final step, a cost-sensitive classifier has been used for classification of the samples. This approach was tested on all images having mass from mini-MIAS data set. Based on the classification results, the percentage of true positive detection rate was 91% false-positive detection was 14% and the area under ROC curve was achieved 96%.

Keywords: Mammogram images, Ranklet features, Co-occurrence matrix, composite classifier, unbalanced data sets, fractal dimension.

1. INTRODUCTION

Breast cancer is considered the most important factor of deaths related to cancers among women. As we know the best prevention method is early detection that leads to improved treatment and reduced loss of human lives. According to national cancer institute of America, in every 3 minutes it is diagnosed one woman is affected with breast cancer and every 13 minutes one woman is died as a result of this disease. Also, it is estimated that 1/8 % of women are affected with breast cancer, and 1/30 % of them die of it. Mammography is currently the most effective method for early and reliable detection of breast cancer [1].

Micro-calcifications and masses are the main symptoms of breast cancer. Mass detection is more difficult than Micro-calcifications because the characteristics of masses may be ambiguous or similar to main tissue of breasts. Masses are often located in areas with dense breast tissue and have smooth borders than Micro-calcifications and on the other hand have many forms. These factors cause mass detection to be considered as a challenging problem.

Various methods have been proposed for detecting masses in mammogram images. For example, Dominguez and colleagues [2] have improved clearness of mammography based on statistical criteria. Then suspicious areas are segmented using multi-level thresh holding. Some features are extracted from each region and at the end; a grading system based on suspicious areas is used. Ozex and colleagues [3] used a template to identify masses. In that system, initially suspicious areas are identified using different thresh holdings, then a template for classification of suspicious area to mass or non-mass is used. Zheng and colleagues [4] combined several artificial intelligence techniques with Discrete Wavelet Transform. Artificial intelligence techniques used include the analysis of fractal dimension, MMRF (Multi-resolution Markov Random Field) and algorithm of Dogs & Rabbits for clustering. At the end, a classifier

*Corresponding author. *Email address*: *jhaddadnia@yahoo.com*

Special Issue: The Second National Conference on Applied Research in Science and Technology based on decision trees is used to classify areas. Dizung and colleagues [5] using Iso-intensity contours gained a number of suspected areas and then extracted co occurrence matrix characteristics from each area and used neural network as a classifier. Bovis and colleagues [6] built 4 co occurrence matrices for each area, and from these 4 co occurrence matrices extracted 70 features, and for classification of areas used RBF and MLP.

Most of the existing techniques in the field of mass detection are combination of image processing techniques and algorithm identification templates, but often they are unaware of the fact that the issue of mass detection is inherently sensitive to the costs. This means that if a mass is wrongly classified as a normal tissue, its risk is much greater than when normal tissue is classified as a mass. Also, another case that negatively affects the classification performance is imbalanced data sets. Since the number of normal regions is much more than abnormal regions, the classification during the training phase will bias to the normal samples.

One of the things that is often overlooked, is feature selection phase. Especially when the number of extracted features is high, the selection process is necessary because presence of irrelevant features causes declines in performance of classification. Another limitation of the proposed methods in articles is that most of them are used for detection of just one special group of masses (like Speculated mass or Circumscribed mass) and cannot be generalized to different masses in terms of type and size. In this paper, first each mammogram is segmented using special techniques and some preliminary suspicious area is achieved. This method is effective in detecting all kinds of masses and allows the radiologist to determine in what range (in respect of size) the existing are identified. Then, to reduce suspicious areas that cannot have mass, a lot of features were extracted from each region. Using a supervised feature selection method, a proper subset of the extracted features was chose. In the following, sampling method was used to eliminate bias created on normal samples. At the end a combined classifier method was used that is sensitive to cost and uses a cost matrix for applying the considered cost. In the following of Section 2 a method is proposed for finding suspicious areas. In section 3, the features extracted from the suspicious area are introduced and their selection method is examined. Section 4 will describe the classification system designed in this paper. In Section 5, testing methods and the results obtained are summarized and conclusions are presented in Section 6.

2. FINDING SUSPICIOUS AREAS

First, to reduce the volume of computations, the image size is reduced from 1024 x 1024 to 512 x 512, then to remove impulsive noises on the image of the mammogram that may drastically affect the next steps, the median filter of size 3*3 is used. After this pre-processing, mammograms should be segmented. The purpose of this image segmenting is to obtain some suspicious areas or ROI (Region of interest) within which a mass may be hidden. In general, some ROI may be detected on image, all of which are indicative of a mass. In other words, the ROI mapping - to -mass is not a one - to - one mapping. At this point, method of splitting into half was used to find the roots of inspiration, and a number of suspicious areas are provided to mass. The algorithm steps are as follows:

1-The image is divided by a number of overlay cells with same sizes. (Grading image into cells with dimensions S*S).

2-In each cell, the pixel with the highest gray level is found. Its Coordination is shown with the index and its value with *m*.

3-*F* indicates the beginning of the range in which the answer is sought and *R* is also the end of range. Initially, *F* is assumed 0, and *R* is assumed as *m*.

4-*Th = (F+R)/2* is considered and an iso-intensity contour with value of *Th* is applied to the image.

5-The region within contour that includes index is considered. If the surface area exceeds a limit (*Area_Max*) or circularity of the region is greater than maximum (*Circ_Max*), then *F=Th* and we go back to step 4. Otherwise, if the surface area is less than a minimum (*Area_Min*), then put *R=Th* and we return to step 4. If the two above conditions are not maintained then this area is determined as a primary suspicious areas, and we go to next cell and Step 2 is followed. Another condition for stop of process on a cell is that the difference between two successive *Th* is less than a limit (Δ).

In this study, *Area_Max* is considered as equal 8000 and *Area_Min* equal to 155 because there is no mass whose area is greater than 8000 (ie. a mass with a radius of 50) and smaller than 155 (a mass with a radius of 7). It should be noted that these values are obtained about images of 512 x 512. One of advantages of this method is that it provides degree of freedom for radiologist, so that radiologist himself can determine value of these parameters and specify dimensions in which he is going to discover system of existing masses. *Circ_Max* is determined, practically and through successive experiments, as 7. Also, in this paper Δ is considered as 4 gray levels and s as 32. The higher the S, the number of suspicious areas is reduced and the number of false positive samples (FP; the samples that were incorrectly identified by the system as mass) is reduced, but also the rate of (TP) (really true examples) and is decreased and vice versa. So, there should be balance between S, FP and TP. Through Consecutive tests it is found that the best value for S is equal to 32. Figure 1 shows an example of the result of the implementing the mentioned segmentation algorithm.

Now it's time to give label to the regions obtained from the segmentation stage (TP or FP). The Label related to each area is determined according to the following procedure: First, a Bounding Box is drawn in a way that the considered area is completely included.

Suppose that (*LX, LY*) indicates the length and width of the rectangle, and (*Xcog, Ycog*) represents the center of gravity of the considered area, and {(Xb, Yb), Rb} represents the center and radius of the mass after biopsied. If

$$|Xb\text{-}Xcog| < \max(Rb, LX/2) \text{ and } |Yb\text{-}Ycog| < \max(Rb, LY/2),$$

then a region is labeled as TP, otherwise it will be a FP.

The Condition used for labeling an area as TP is a strong requirement and this requirement has been used in numerous papers [5.19, 20]. For example, in Figure 1, after applying the segmentation algorithm, four suspicious area were obtained among which just the area that is marked with red arrows labeled as TP and the other 3 area show areas that have been wrongly represented as mass, hence these 3 regions are labeled as FP. We intend to reduce this number using machine learning techniques.

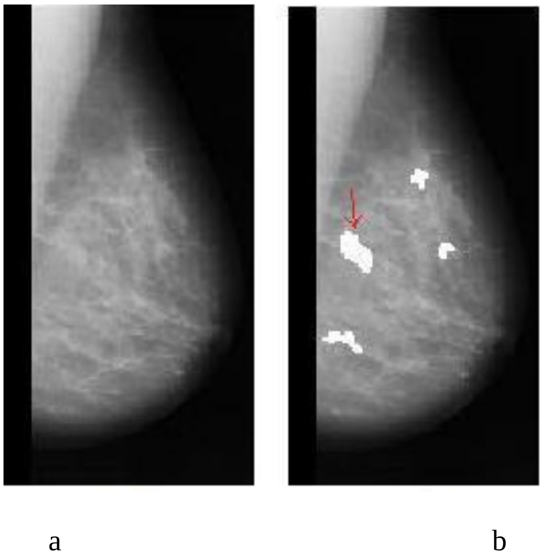

Figure 1. The result of segmentation algorithm on the image mdb091. (a) The main image. (b) The specious area obtained.

3. EXTRACTION AND SELECTION OF FEATURE

In this step, several features are extracted for each region. First, characteristic of intensity [7] (including the mean of gray levels, standard deviation of gray levels, Kurtosis, Skewness and entropy) are extracted. Then, fractal dimension of each area is added to the feature vector of that region based on the approach proposed in [8]. Now, Ranklet characteristics [9], which make up the largest group of features extracted from an area, are investigated. If a Ranklet transfer is applied on the image I, *Nri* non parametric Ranklet image is produced. In fact, *Nri* is obtained by multiplying *Nr* (number of resolutions that are assumed in the analysis) by No (number of directions being examined). In the proposed method, the number of directions is assumed as 3 (horizontal, vertical and diagonal). To extract features, first Bounding Box related to each ROI is changed to16 x 16 Dimensions. Then, resolutions [14, 8, 4, 2} are applied and horizontal, vertical and diagonal Ranklet images related to each resolution is gained; 12 images is obtained for every 16 x 16 image. Each of Ranklet coefficients of these images are added to the general vector as a feature. One of the valuable features of this transfer is not changing in linear / non-linear monotone transformations of gray levels. For each of the 12 resulting image, two other features are extracted too. These two features are called convergence mean and variance of code. Thus, for every 12 Ranklet images, two other features are added to the total feature vector. Then, 3 Convergence means and 3 code variances for each resolution is gained (one for horizontal images, one for vertical and one for diametrical image). Now, two other features are extracted for each resolution and they are average convergence mean and variance mean of code in 3 directions. Thus, 8 = 2 * 4 feature is added to the general feature vector that these 8 features, in addition to being invariant against leaner/ non linear changes of gray level, are invariant with respect to rotation. As the last group of the extracted features, the features resultants from co-occurrence matrix are reviewed. First, co-occurrence matrix is made in directions 0, 45, 90 and 135 ° with d = (dx, dy) = (1,1). Then, Haralick features [10] and the features available in [11] (total 22 features) are extracted from each of these matrices. Therefore, 88 = 4 * 22 feature have been obtained until now. In addition to these 88 features, mean and standard deviation for each of the 22 features in 4 mentioned directions are connected to the vector. It should be noted that the latter properties are invariant against rotation.

In most of machine learning algorithms, the complexity of algorithm depends on the number of features (dimensions) as well as the number of samples (N). To reduce computational complexity, required memory, increase resistance to noise and outlier points and also to increase the classification efficiency (by reducing the irrelevant features), dimension is decrease. After extracting all the mentioned features, a total of 1392 features is achieved. Now, we want to reduce the number of features and choose a suitable subset of them. In this paper, the criterion that was used in CFS (Correlation-based Feature Subset Selection) method, [12] has been used as a measure of eligibility of a subset. According to the eligibility criteria, appropriate subset is one whose members are highly correlated with the class labels and on the other hand are not dependent to each other. According to this principle, irrelevant and redundant features would be removed. For choosing the best subset, the first best strategy has been used. This strategy is a greedy method of achieving a nearly optimal solution.

4. CLASSIFICATION OF SAMPLES

Data set created in the previous section is an unbalanced database is because the number of FP samples is much more than TP. Classifiers often have trouble in encountering unbalanced data sets. Also, another point that should be considered is that issue of mass detection is inherently sensitive to cost. To solve the cost problem, a cost matrix is defined as is seen in Table 1, and the matrix is involved in classification of the samples.

Table 1. Cost matrix for a two-class problem.

Predicted class			
True class		mass	Non-mass
	mass	**0**	***C1***
	Non-mass	***C2***	**0**

Using cost matrix is one of the strategies for coping with unbalanced data sets. Another strategy is using sampling methods. In this paper, a sampling method called SMOTE [13] has been used. In this method, k ones of the nearest neighbors to a sample are selected from smaller class. Then, using these neighbors, depending on the sampling rate, some fabricated samples are made by modeling the prototype and added to the set.

In this paper, a meta- classifier called Meta-Cost [14] is used to create a cost-sensitive classifier. In fact, meta- classifier transfers the basic classifier that is placed in it,based on a cost matrix such as Table 1. The basic classifier which we have placed in Meta-Cost, is a classifier that combines 3 Ensemble classifier called Bagging [15], AdaBoost. M1 [16] and rotation forest [17] through mechanism of a majority vote. In other words, a combination of Ensemble classifiera classifiers is made. Naive Bayes has been used as a basic classifier in AdaBoost and decision trees as basic classifier in Bagging and rotation forest. In summary, in this step, first using sampling, the data set is balanced, then a classifier is made that is cost-sensitive and is composed of a combination of three combined classifiers.

5. EXPERIMENTS AND RESULTS

All mammograms used in this paper are mammograms available in database mini-MIAS [18] and have mass. Among 322 images available in the mini-MIAS, 55 images contain mass and in these 55 images there are 58 masses. After the segmentation step, 57 TP and 209 FP were obtained from these 55 images. These 57 TP represent 53 masses (as previously mentioned, the ROI-mapping- to -mass is not a one - to - one mapping). Therefore, sensitivity of system after segmentation and before classification is equal 0.913 = 58/53. Also, there is false positive (FP) the sample as mean of 3.8 on each image (3.8 = 55/209). In the next steps it will tried to reduce the number of FPs greatly using machine learning techniques, without much loss of sensitivity. In investigating a specific technique, we do 10-fold cross validation for 10 times and then, efficiency criteria of the classifier in these 10 implementations, are reported based on the mean.

First, we'll examine the differentiation power of each group of features individually and then the impact of these features are examined together (Table 2).

As can be seen in Table 2, due to the space complexity, using all features for training the proposed classifier (even with devoting the maximum memory that the operating system assigns to a process) is impractical. That's why we had to test impact of mechanism of choosing feature on the largest group of features; Ranklet features (1254 features) can be tested. Due to irrelevant features, efficiency of classifier on these attributes is extremely low. But after applying the mechanism of proposed feature selection, only 47 features were selected as suitable. Along with this reduction of dimension, the complexities of time and space are also reduced greatly. On the other hand, with the removal of a lot of irrelevant features in this group we were able to dramatically improve the classifier efficiency criteria. Next, from the features extracted for a sample (1392 features), only 49 features were selected by feature selection mechanism and classification and used for training classifier. It is seen that (last row in Table 2) adding features that have high TP rate (eg, fractal dimension) increase the classifier efficiency (especially TP Rate) is. Low rate of TP classifier in Table 2 is due to the unbalanced nature of the dataset.

Table 2. Results of Classification using different groups of characteristics and impact of feature selection on classification efficiency.

	True positive rate	False positive rate	The area under ROC curve
Intensity features	0.2336	0.0205	0.6499
Features of co-occurrence matrix	0.2792	0.0471	0.6892
Fractal dimension feature	0.3632	0.0695	0.6530
Ranklet features	0.0237	0.008	0.6397
All featurres	Not implantable		
Feature selection by CFS+Ranklet feature	0.1965	0.0301	0.7539
Feature selection by CFS+all extracted features	0.3525	0.0302	0.7801

Now impact of the cost matrix on increase of TP rate is examined (Figure 2). As it was expected, with increase of C1/C2 ratio, TP rate increases too. The question that arises is how long is this ratio increased? When this ratio is excessively increased, the classifier is biased towards too much positive samples and this bias will reduce the classifier efficiency. As a result, C1/C2 ratio increase until mutations occur at a rate of TP. Otherwise, excessive increase of this ratio, will not make much improvement in the rate of TP, but FP rate will increase dramatically. Therefore, radiologist should be careful enough in setting this parameter.

Using cost matrix has another major reason and it is the fact that issue of identification of a mass is cost sensitive and that is why using cost matrix is unavoidable. SMOTE technique has another parameter called sampling rate and by setting this parameter we try to balance the considered data set and to reduce the bias created on larger class samples and thus to increase TP rate. In this paper, value of this parameter is chosen as 200%. Applying this criterion, TP rate is obtained as 0.8463 and FP rate as 0.1156 and the area under the ROC curve as 0.938. The obtained results confirm the efficiency of the technique to improve the classification efficiency. According to the results, the proposed cost matrix and sampling method is used as combination, so that could benefit advantages of both methods. The results of combination of these two methods are seen in Figure 2.

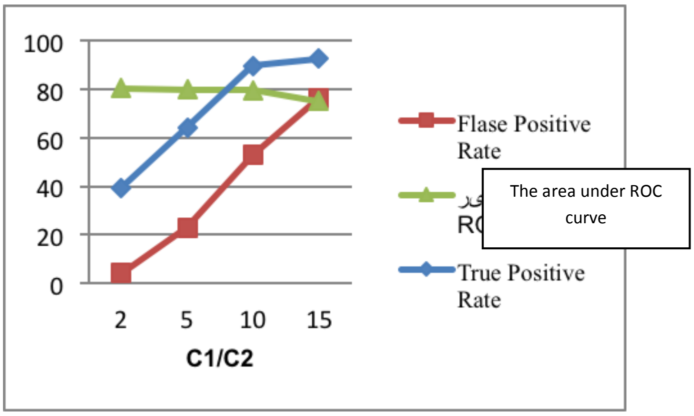

Figure 2. Impact of the cost matrix on criteria of classifier's efficiency.

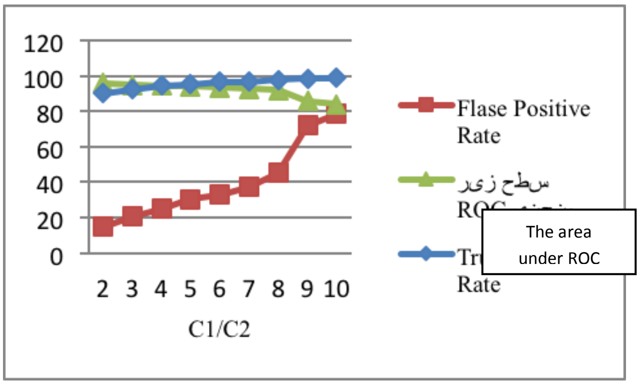

Figure 3. Combination of sampling with rate of 200% and different cost matrices.

It is also observed in Figure 3 that when the C1/C2 ratio excesses a limit, then little improvement occurs in the rate of TP and FP mutants considerably. How to balance the TP rate and FP rate seems to depend on the radiologist's opinion. The radiologist can impose his opinion by setting the C1/C2 ratio of the cost matrix, this means giving degree of freedom to the radiologist. In this section for stating the best result, the results are examined only from the perspective of machine learning. In Figure 3, the best value of the area under ROC curve has occurred when C1/C2 ratio is 2. Applying this ratio we obtained Sensitivity = 0.91 and FP rate of 0.14 in classification stage. According to the results of the classification step, Sensitivity of entire system will be approximately 83% (0.832 = 58/53 * 0.91) and the number of false positives samples per each image is on average equal to 0.53 (0.53 = 55/209 * 0.14). In table, 3 proposed methods in this paper is compared with methods that have been tested on the mini-MIAS data set.

Table 3. Comparison of the efficiency of the proposed method with other methods.

	total Sensitivity	False Positive per Image	The area under ROC curve in classification
Bovis [6]	0.54	1.281	0.74
Dizung[5]	0.85	3.47	0.815
Zhang[4]	97.3	3.92	-
Dominguez[2]	0.8	2.3	-
Ozex [3]	0.81	0.33	-
The proposed method	0.83	0.53	0.96

It should be noted that in [4] the number of masses in the mentioned data set is 37 ; that proposed method in that article could detect 36 samples, and obtain sensitivity = 36/37 = 97.3 while the number of in this data set is equal to 58 mass. Also in [3] all mammograms containing masses have not been tested and results have been reported on 41 images containing mass.

6. CONCLUSION

In this paper, a method is proposed for detecting masses in mammogram images. This method can be used to detect all types of mass with any size. Giving different degrees of freedom to a radiologist in detecting masses has been proposed as one of the most obvious features of the method. Since mass detection is inherently a cost sensitive issue, an efficient classifier cost-sensitive was used; radiologist can set the cost matrix parameter achieve the desired results. This method was tested on mammograms from mini-MIAS data set that have mass and achieved very good results, especially in the classification step. Classifier mechanisms proposed in this paper can be used for classification of masses problems (in terms of being benign or malignant). This issue will be addressed in future works.

REFERENCES

[1] De Oliveira, M., Braz, J., Cardoso, P., Gattass, M., " *Detection of Masses in Digital Mammograms using K-means and Support Vector Machine"*, Electronic Letters on Computer Vision and Image Analysis, vol. 8, No 2, pp. 39-50, 2009.

[2] Dominguez, A., Nandi, A., *"Detection of masses in mammograms via statistically based enhancement, multilevel-thresholding segmentation, and region selection*", journal of Computerized Medical Imaging and Graphics, Vol. 32, No. 4, pp. 304-315, 2008.

[3] Ozekes S., Osman O., Camurcu A.Y., "*Mammographic Mass Detection Using A Mass Template*", Korean J Radiol, Vol. 6, No. 3, p.221-228, 2005.

[4] Zheng L., Chan A.K., "*An Artificial Algorithm for Tumor Detection in Screening Mammograms*", IEEE Transactions on Medical Imaging, Vol. 20, No. 7, p. 559-567, 2001.

[5] Dzung V., Nguyen D.T., Dzung T, Pham V.T., "*An Automated Method to Segment and Classify Masses in Mammograms*", International Journal of Electrical and Computer Engineering, Vol. 8, No. 4, 2009

[6] Bovis K., Singh S., Fieldsend J, Pinder Ch., "*Identification of masses in digital mammograms With MLP and RBF Nets*", in Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks Com, Vol 1, pp. 342-347, 2000.

[7] Christoyanni I., Dermatas E., Kokkinakis G, "*Fast detection of masses in computer-aided mammography*", IEEE Signal Process. Mag., Vol. 17, No. 1, pp. 54-64, 2000.

[8] Yang S., Wang C., Chung Y., Hsu G., Lee S., Chung P., Chang C., "*A computer aided system for mass detection and classification in digitized mammograms*", Biomedical Engineering-Application-Basis & Communications, Vol. 17, No. 5, pp. 215-228, 2005.

[9] Massotti M., Campanini R., "*Texture classification using invariant ranklet features*", Pattern Recognition Letters, Vol. 29, No. 14, pp. 1980-1986, 2008.

[10] Haralick R.M., Shanmugan K., Dinstein I., "*Textural Features for Image Classification*", IEEE Transactions on Systems,Man and Cybernetics, Vol. 3, No. 6, pp. 610-621, 1973.

[11] Soh L., Tsatsoulis C., "*Texture Analysis of SAR Sea Ice Imagery Using Gray Level Co-occurrence Matrices*", IEEE Transaction on Geoscience and Remote Sensing, Vol. 37, No. 2, pp. 780-795, 1999

[12] Hall M.A., "Correlation-based Feature Subset Selection for Machine Learning", Ph.D. Thesis, University of Waikato, Hamilton, New Zealand 1999.

[13] Chawla N.V., "*Synthetic Minority Over-sampling Technique*", Journal of Artificial Intelligence Research, Vol. 16, pp. 321-357, 2002.

[14] Domingos P., "*MetaCost: A general method for making classifiers cost-sensitive*", Fifth International Conference on Knowledge Discovery and Data Mining, pp. 155-164, 1999.

[15] Breiman L., "*Bagging predictors*", Machine Learning, Vol. 24, No. 2, pp. 123-140, 1996.

[16] Freund Y., Schapire R.E., "*Experiments with a new boosting algorithm*", Thirteenth International Conference on Machine Learning, San Francisco, pp. 148-156, 1996.

[17] Rodriguez J.J, Kuncheva L., Alonso C., "*Rotation Forest: A new classifier ensemble method*". IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 28, No. 10, pp. 1619-1630, 2006.