Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 4, April 2013, pg.269 – 277

RESEARCH ARTICLE

A Novel Prediction on Breast Cancer from the Basis of Association rules and Neural Network

S. Palaniappan[1], T. Pushparaj[2]

[1]Assistant Professor, PG Department of Computer Applications, PSNA College of Engineering and Technology, Dindigul – 624 622, Tamil Nadu, India

[2]Assistant Professor, PG Department of Computer Applications, PSNA College of Engineering and Technology, Dindigul – 624 622, Tamil Nadu, India

[1] *spalani@psnacet.edu.in;* [2] *pusht82@yahoo.co.in*

Abstract— The use of machine learning tools in medical diagnosis is increasing gradually. This is mainly because the effectiveness of classification and recognition systems has improved in a great deal to help medical experts in diagnosing diseases. Such a disease is breast cancer, which is a very common type of cancer among woman. As the incidence of this disease has increased significantly in the recent years, machine learning applications to this problem have also took a great attention as well as medical consideration. This paper presents an automatic diagnosis system for predicting breast cancer based on association rules (AR) and neural network (NN). In this study, AR1 and AR2 are used for reducing the dimension of breast cancer dataset and NN is used for intelligent classification. The proposed AR1 + AR2 + NN system performance is compared with NN model. The dimension of input feature space is reduced from nine to four by using AR1 & AR2. In test stage, 3-fold cross validation method was applied to the Wisconsin breast cancer dataset to evaluate the proposed system performances. The correct classification rate of proposed system is 98.4%. This paper demonstrated that the AR1 and AR2 can be used for reducing the dimension of feature space and proposed AR1 + AR2 + NN model can be used to obtain fast automatic diagnostic system for breast cancer.

Key Terms: - Breast cancer diagnosis; Wisconsin Breast Cancer Database (WBCD); Association Rule; Neural Network; Multilayer Perceptron (MLP); feature selection; 3-Fold cross validation

I. INTRODUCTION

BREAST CANCER is considered a major health issue in western countries, and constitutes the most common cancer among women in the world. It is estimated that between 1 in 8 and 1 in 12 women will develop breast cancer during their lifetime. Moreover, in the European Union, as well in the United States, breast cancer remains the leading cause of death for women in their 40s. Breast cancer is a very common and serious cancer for women. There is a considerable increase in the number of breast cancer cases in recent years. It is reported that breast cancer was the second one among the most diagnosed cancers. It is also stated that breast cancer was the most prevalent cancer in the world by the year 2002. Breast cancer outcomes have improved during the last decade with development of more effective diagnostic techniques and improvements in treatment methodologies. A key factor in this trend is the early detection and accurate diagnosis of this disease.

Consequently, breast cancer is an intensely researched area. However, despite many promising leads, the genetic and epigenetic events driving the transformation from normal breast tissue to metastatic breast cancer remain largely unknown. As a consequence, there are no reliable biomarkers to predict the transitions from breast disease to breast cancer to metastatic cancer. The absence of these markers negatively impacts the entire

spectrum of breast disease clinical care, from assessing breast cancer risk to treating breast cancer patients. Hence automated medical diagnostic decision support systems have become an established component of medical technology. The main concept of the medical technology is an inductive engine that learns the decision characteristics of the diseases and can then be used to diagnose future patients with uncertain disease states. In this paper, we investigate a novel approach to automatic breast cancer classification which were trained on the attributes of each record in the Wisconsin breast cancer database, is presented.

II. RELATED WORK

Data classification obtained from known historical data has been one of the most intensively studied subjects in statistics, decision science and computer science. It has been applied in problems of medicine, social science management and engineering. Variable problems such as disease diagnosis, image recognition, and credit evaluation using classification techniques [10]. In medical and other domains, linear programming approaches were efficient and effective methods (Bennett & Mangasarian, 1992; Freed & Glover, 1981; Grinold, 1972; Smith, 1968). Recently, intelligent methods such as NN and support vector machines have been intensively used for classification tasks [13].One of the application areas of analyzing database and pattern recognition is automated diagnostic systems. The aims of these studies are assisting to doctors in making diagnostic decision. Thanks to modern facilities, very large databases can be collect in medicine. These databases need special techniques for analyzing, processing and effective use of them. clustering, etc. (Curt, 1995). Data mining and knowledge discovery in database are an approach to find relationships buried in data [6]. The methodologies consist of data visualization, machine learning and statistical techniques and these can be summarized as classification, prediction, clustering, etc. (Cuet, 1995).Mammography is one of the most used methods to detect the breast cancer [6]. In literature, radiologists show considerable variation in interpreting a mammography (Elmore et al., 1994). As for other clinical diagnosis problems, classification systems have been used for breast cancer diagnosis problem, too. When the studies in the literature related with this classification application are examined, it can be seen that a great variety of methods were used which reached high classification accuracies using the data set. Fine needle aspiration cytology (FNAC) is also widely adopted in the diagnosis of breast cancer. But, the average correct identification rate of FNAC is only 90%. So, it is necessary to develop better identification method to recognize the breast cancer. Statistical techniques and artificial intelligence techniques have been used to predict the breast cancer by several researchers (Kovaler-chuck, Triantaphyllou, Ruiz, & Clayton, 1997; Pendharkar, Rodger, Yaverbaum, Herman, & Benner, 1999). The objective of these identification techniques is to assign patients to either a benign group that does not have breast cancer or a 'malignant' group who has strong evidence of having breast cancer. So, breast cancer diagnostic problems are more general and widely discussed classification problem. (Anderson, 1984; Dillon & Goldstein, 1984; Hand, 1981; Johnson & Wichern, 2002). There are many techniques to predict and classify breast cancer pattern. The authors in [11] proposed several association and classification approaches such as Associations, statistical, mathematical and neural networks for mining breast cancer patterns. The study in [4] proposed a combined neural network and decision trees model for prognosis of breast cancer relapse. In [16] artificial neural network and multivariate adaptive regression splines approach was used to classify the breast cancer pattern. In Dursun Delen, Glenn Walker and Amit Kadam (2004), a comparison of three data mining methods for predicting breast cancer was proposed. In this paper the authors used artificial neural networks and decision trees along with logical regression to predict breast cancer. In [14] a new hybrid method based on fuzzy-artificial immune system and k-nn algorithm was proposed for breast cancer diagnosis. In [17] Wisconsin breast cancer data was classified using multilayer perceptron neural network, combined neural network, probabilistic neural network, recurrent neural network and support vector machine. In [13] isotonic separation technique was used to predict breast cancer. In [19] Support vector machines combined with feature selection was proposed to classify breast cancer diagnosis. Here the method proposed was SVM based combined with feature selection for breast cancer diagnosis on WBCD. In [18] Breast mass classification based on cytological patterns using RBFNN and SVM was proposed to classify breast cancer.

A. Paper Organization

In this paper, an AR1 + AR2 + NN method was proposed to use in breast cancer diagnosis problem. This method consists of two-stages. In the first stage, the input feature vector dimension is reduced by using association rules. This provides elimination of unnecessary data. In the second stage, neural network uses these inputs and classifies the breast cancer data.

III. MATERIAL AND METHODS

A. Wisconsin Breast Cancer Dataset overview

Breast cancer is the most common cancer among women; excluding non-melanoma skin cancers. This cancer affects one in eight women during their lives. It occurs in both men and women, although male breast cancer is rare. Breast cancer is a malignant tumor that has developed from cells of the breast. Although scientists know some of the risk factors (i.e. ageing, genetic risk factors, family history, menstrual periods, not having children, obesity) that increase a woman's chance of developing breast cancer, they do not yet know what causes most breast cancers or exactly how some of these risk factors cause cells to become cancerous. Research is under way to learn more and scientists are making great progress in understanding how certain changes in DNA can cause normal breast cells to become cancerous [17].

The used data source is Wisconsin Breast Cancer Dataset (WBCD) taken from the University of California at Irvine (UCI) Machine Learning Repository. This data set is taken from fine needle aspirates from human breast tissue was commonly used among researchers who use machine learning (ML) methods for breast cancer classification. In this study, the WBCD was used and analyzed. They have been collected by Dr. William H. Wolberg (1989–1991) at the University of Wisconsin–Madison Hospitals. There are 699 records in this database. Each record in the database has nine attributes. The nine attributes detailed in Table 1 are graded on an interval scale from a normal state of 1–10, with 10 being the most abnormal state. In this database, 241 (65.5%) records are malignant and 458 (34.5%) records are benign.

B. Association rules (Feature Extraction layer)

In order to see how AR can be used in breast cancer data with NN, first of all it is needed to define AR. AR find interesting associations and/or relationships among large set of data items. AR shows attribute value conditions that occur frequently together in a given dataset. They allow capturing all possible rules that explain the presence of some attributes according to the presence of other attributes.

1. Apriori Algorithm

The Apriori algorithm is a state of the art algorithm most of the association rule algorithms are somewhat variations of this algorithm [2]. The Apriori algorithm works iteratively. It first finds the set of large 1-item sets, and then set of 2- itemsets, and so on. The number of scan over the transaction database is as many as the length of the maximal item set. Apriori is based on the following fact: The simple but powerful observation leads to the generation of a smaller candidate set using the set of large item sets found in the previous iteration. Thus, AR aims at discovering the patterns of co-occurrence of attributes in a database. For instance, an association rule in a supermarket basket data may be in 10% of transactions, 85% of the people buying milk also buy yoghurt in that transaction.

One of the most popular data mining approaches is to find frequent itemsets from a transaction dataset and derive association rules. Finding frequent itemsets (itemsets with frequency larger than or equal to a user specified minimum support) is not trivial because of its combinatorial explosion. Once frequent itemsets are obtained, it is straightforward to generate association rules with confidence larger than or equal to a user specified minimum confidence.

The Apriori first scans the transaction databases D in order to count the support of each item i in I, and determines the set of large 1- itemsets. Then, iteration is performed for each of the computation of the set of 2-itemsets, 3-itemsets, and so on.

It then iterates on the following three steps and extracts all the frequent itemsets.

1. Generate *Ck*+1, candidates of frequent itemsets of size *k* +1, from the frequent itemsets of size *k*.
2. Scan the database and calculate the support of each candidate of frequent itemsets.
3. Add those itemsets that satisfies the minimum support requirement to *Fk*+1.

The Apriori algorithm is as follows:

Ck: Candidate itemset of size k
Lk : frequent itemset of size k
L1 = {frequent items};
for (*k* = 1; *Lk* !=∅; *k*++) **do begin**
  Ck+1 = candidates generated from *Lk*;
  for each transaction *t* in database do
    increment the count of all candidates in *Ck+1*
      that are contained in *t*
  Lk+1 = candidates in *Ck+1* with min_support

end
return ∪*k Lk*;

After the candidates are generated, their counts must be computed in order to determine which of them are large. This counting step is really important in the efficiency of the algorithm, because the set of the candidate itemsets may be possibly large.

Neural networks (classifier layer)

C. Neural networks (NNs) are biologically inspired and mimic the human brain. They are occurring neurons. These neurons are connected each other with connection links. These links have weights. They multiplied with transmitted signal in network. The output of each neuron is determined by using an activation function such as sigmoid and step. Usually nonlinear activation functions are used. NN's are trained by experience, when applied an unknown input to the network it can generalize from past experiences and product a new result [5] [7] [8].

TABLE I
Wisconsin breast cancer data description of attributes

Wisconsin breast cancer data description of attributes				
Attribute number	**Attribute description**	**Values of attributes**	**Mean**	**Standard deviation**
1	Clump Thickness	1–10	4.42	2.82
2	Uniformity Of Cell Size	1–10	3.13	3.05
3	Uniformity Of Cell Shape	1–10	3.2	2.97
4	Marginal Adhesion	1–10	2.8	2.86
5	Single Epithelial Cell Size	1–10	3.21	2.21
6	Bare Nuclei	1–10	3.46	3.64
7	Bland Chromatin	1–10	3.43	2.44
8	Normal Nucleoli	1–10	2.87	3.05
9	Mitoses	1–10	1.59	1.71
N = 699 Observations, 241 Malignant and 458 Benign.				

NNs models have been used for pattern matching, nonlinear system modeling, communications, electrical and electronics industry, energy production, chemical industry, medical applications, data mining and control because of their parallel processing capabilities. So that they are also known as connectionist models or parallel distributed processing models, while traditional computers, because of their architecture, NN are inefficient at these tasks, especially pattern-matching tasks. Neural networks essentially comprise three pieces: the architecture or model; the learning algorithm; and the activation functions [10]. Neural networks are programmed or "trained" to store, recognize, and associatively retrieve patterns or database entries. It is precisely these two abilities (pattern recognition and function estimation) which make artificial neural networks (ANN) so prevalent a utility in data mining.

The first piece is the neural network's pattern of connections between neurons, the architecture or model. The second part is the method which determines the weights on the connections, the training or learning algorithm. These weights represent the information being processed by the neural network. The third component is he function which determines each neuron's output signal, the activation function.

When designing a NN model a number of considerations must be taken into account. First of all the suitable structure of the NN model must be chosen, after this the activation function and the activation values need to be determined. The number of layers and the number of units in each layer must be chosen. Generally desired model consist of a number of layers. The most general model assumes complete interconnections between all units. These connections can be bidirectional or unidirectional.

The strength of the connection between an input and a neuron is noted by the value of the weight. Negative weight values reflect inhibitory connections, while positive values designate excitatory connections [8]. The next two components model the actual activity within the neuron cell. An adder sums up all the inputs modified by their respective weights. This activity is referred to as linear combination. Finally, an activation

function controls the amplitude of the output of the neuron. An acceptable range of output is usually between 0 and 1, or -1 and 1.

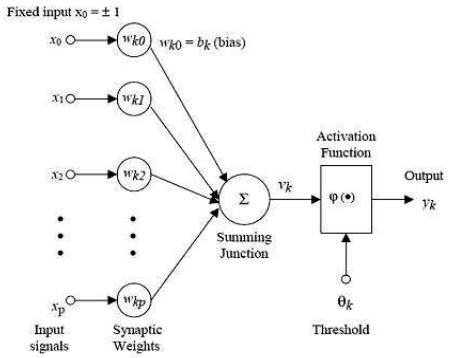
Feature extraction is the key for pattern recognition so that it is arguably the most important component of designing the intelligent system based on pattern recognition since even the best classifier will perform poorly if the features are not chosen well.

A feature extractor should reduce the pattern vector (i.e., the original waveform) to a lower dimension, which contains most of the useful information from the original vector [16]. Fig.2 shows the proposed automatic detection system block diagram. It consists of two parts: (a) feature extraction and reduction with AR (b) classification with NN.

D. AR Layer

AR is a method to find the associations and/or relationships among items in large databases. So, we can use it to detect relations among inputs of any system and later eliminate some unnecessary inputs. We propose two different techniques to eliminate inputs. These are named as AR1 and AR2, respectively.

Fig.1 shows a neural network has to be configured such that the application of a set of inputs produces (either 'direct' or via a relaxation process) the desired set of outputs. The network is trained by providing it with input and matching output patterns. These input-output pairs can be provided by an external teacher, or by the system which contains the neural network. A crucial aspect of carrying out learning and prediction analysis with a neural network system is to split the database into two independent sets: the training set (80% of the dataset), which is used to train the neural network, and the test set (20% of the dataset) to validate its predictive performance.



2.

Fig. 1. Artificial neuron model

TABLE II
MLP architecture and training parameters

architecture	
The number of layers	3
The number of neuron on the layers	Input: 4, 8, 9 Hidden: 11 Output: 1
The initial weights and biases	Random
Activation functions	 Tangent-sigmoid Tangent-sigmoid Linear
Training parameters	
Learning rule	 Back-propagation
Sum-squared error	0.01

E. Applications

1) AR1 Layer

The AR1 technique uses all input parameters and their all records to find relations among the input parameters. If we find rules that have enough support value and high confidence value, than we can eliminate some inputs thanks to these rules. In the AR form (X=>Y), item set also depend on X item set. Thus, we can eliminate all items in Y item set. So, these are not necessary to use in NN inputs.

2) AR2 Layer

Especially, we can use AR2 with classification problems. AR2 uses all input parameters but not all their records. We find only large item sets for every class. All items in these large item sets are most important items to classification. Thus, we can only use these items to classify all data. If an item of large item set of any class is large in other classes and it has different value, this item must be used as NN inputs.

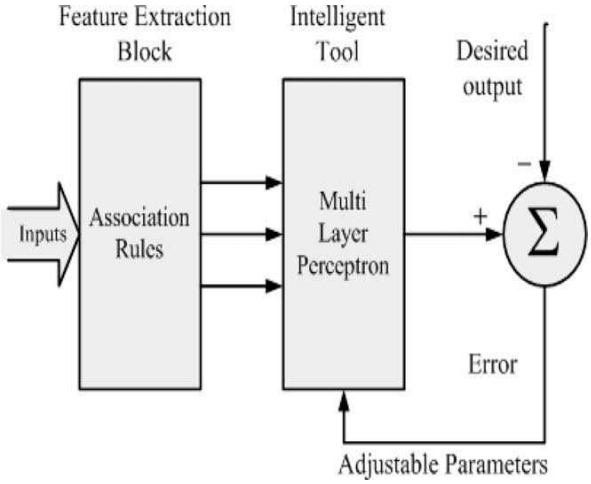


Fig. 2. The block diagram of the automatic detection system

In this study, we used AR1 and AR2 to reduce the number of NN inputs for breast cancer detection problem. We eliminated only one input parameter of NN By using AR1 technique. Because, one of the rule is

Input: 1-3-8-9)2
Value: 1-1-1-1)1 confidence is 100%.

According to this rule; if the value of 1st, 3rd, 8th and 9th input parameters are 1, the value of 2nd input parameter is 1. Then it says that 2nd input already depend on others. So we did not use 2nd input parameter in NN input. Wisconsin breast cancer database has two classes. These are benign and malignant classes. Using AR2, we found large itemsets of benign and malignant classes given as follows:

Input: 2-8-9
Value: 1-1-1 (large itemsets for benign class)
Input: 6
Value: 10 (large item for malignant class)

According to this large itemset, we can say that 2nd, 8th and 9th input parameters already can define benign class and 6th input parameter can define malignant class. These parameters are the most important parameters for breast cancer detection problems. So, we only used these inputs in NN.

3) NN Layer

Multi-layer perceptron (MLP): the intelligent classification is realized in this layer by using features, which are obtained from AR layer. The Multi-Layer Perceptron (MLP), also called a feed-forward network, involves estimated weights between the inputs and a hidden layer where the hidden layer has a nonlinear activation function [8]. The training parameters and the structure of the MLP used in this study are shown in Table II. These were selected for the best performance, after several experiments.

Each neuron has an activation function to determine the output. There are many kind of activation function. Usually nonlinear activation functions such as sigmoid, step are used. NN's are trained by experience, when

applied an unknown input to the network it can generalize from past experiences and product a new result [5] [8]. The weight of an input is a number which when multiplied with the input gives the weighted input. These weighted inputs are then added together and if they exceed a pre-set threshold value, the neuron fires. In any other case the neuron does not fire.
The output of the neuron net is given by Equation:

$$y(t+1)=a\left(\sum_{j=1}^{m} W_{ij}X_j(t)-\theta_i\right) and f_i=\underline{=}net_i=\sum_{j=1}^{m} W_{ij}X_j(t)-\theta_i \quad (1)$$

where, X = (X1,X2 ,. . . ,Xm) represent the m input applied to the neuron, Wi represent the weights for input Xi, hi is a bias value, a(.) is activation function.

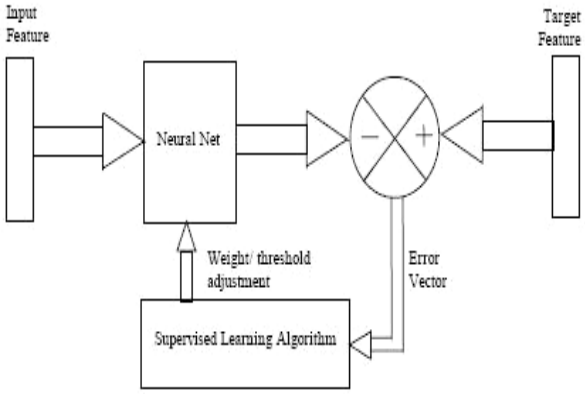


Fig. 3. Neural network training set model

A multi-layer perception with Back Propagation was employed for structuring model and the network was divided into input, hidden and output layers. In the input layer, numbers of elements were restricted by specifications of Software. Therefore, categorical variables were converted to dummy variables, and variables that had continuous values were set down as continuous variables. The sigmoid function was employed, and the weight was decided randomly in the connections for the network.

4) Implementation

The above mentioned two stages (two layers) along with their algorithms and a diagnostic tool is created using java language. The Apriori and MLP algorithms are implemented using java by which the numbers of attributes are reduced from nine to eight then to four and the reduced inputs are given to NN layer for breast cancer classification. The same was implemented in weka (A Data mining tool) and results are compared.

TABLE III
Performance comparison for breast cancer detection using NN, AR1 + NN, AR2 + NN, and ar1+ar2+NN

The classifier	The epochs	Correct classified	Miss classified	Correct rate (%)
NN (9, 11, 1)	61	216	11	95.2
AR1 + NN (8,11,1)	44	216	6	97.4
AR2 + NN (4,11,1)	33	217	10	95.6
AR1+AR2+NN (8,4,11,1)	22	219	4	98.4

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

This study was performed using Wisconsin breast cancer database with 9 attributes and 699 records. In test stage, 3-fold cross validation method was applied and average values were calculated. The performance comparison and correct classification rates are tabulated in Table III. As shown in Table III, the best

classification performance was obtained with AR1 + NN with eight inputs and its correct classification rate is 97.4%. The correct classification rate of NN with 9 inputs is 95.2%. The correct classification rate of AR2+NN is 95.6% and the correct classification rate was obtained with AR1+AR2+NN is 98.4%. So, we can use AR1+AR2 for reducing input parameters at minimum number and NN for best classification performance.

V. CONCLUSION AND FUTURE WORK

With the improvements in expert systems and clinical diagnostic tools, the effects of these innovations are entering to more application domains day by day and medical field is one of them. Decision making in medical field can be a trouble sometimes. Classification systems that are used in medical decision making provide medical data to be examined in shorter time and more detailed. According to the statistical data for breast cancer in the world, this disease is among the most prevalent cancer types. In the same time, this cancer type is also among the most curable ones if it can be diagnosed early. So, AR1+AR2 are used for reducing the dimension of breast cancer database and NN is used for intelligent classification. The proposed AR1+AR2+NN system performance is compared with NN model. The dimension of input feature space is reduced from nine to four by using AR. In test stage, 3-fold cross validation method was applied to the Wisconsin breast cancer database to evaluate the proposed system performances. The correct classification rate of proposed system is 98.4% for four inputs and 95.6% for eight inputs. This research demonstrated that the AR1+AR2+NN approach can be used for efficient automatic breast cancer diagnostic systems and the same model can be used to obtain efficient automatic diagnostic systems for other diseases.

REFERENCES

[1] Agrawal, R., & Srikant, R. "Fast algorithms for mining association rules in large databases," In Proceedings of the 20th international conference on very large databases, pp. 487–499, 1994.
[2] Agrawal, R., Imielinski T., & Swami, A "Mining association rules between sets of items in large databases," In Proceedings of the ACM SIGMOD international conference on management of data, 1993.
[3] Anderson, T. W. An introduction to multivariate statistical analysis. New York: Wiley, (1984).
[4] Aragonés, M. J., Ruiz, A. G., Jiménez, R., Pérez, M., & Conejo, E. A. "A combined neural network and decision trees model for prognosis of breast cancer relapse," Artificial Intelligence in Medicine, 27, 45–63, (2003).
[5] Bishop, C. M "Neural networks for pattern recognition," Oxford: Clarendon Press, (1996).
[6] Choua, S.M., Leeb, T.S., Shaoc, Y. E., & Chenb, I.F "Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines," Expert Systems with Applications, 27, 133–142, (2004).
[7] Hanbay, D., Turkoglu, _I., & Demir, Y "An expert system based on wavelet decomposition and neural network for modeling Chua's circuit," Expert Systems with Applications, doi:10.1016/j.eswa.2007.03.002, (2007)..
[8] Haykin, S. "Neural networks, a comprehensive foundation," New York: Macmillan College Publishing Company Inc, (1994).
[9] Kovalerchuck, B., Triantaphyllou, E., Ruiz, J. F., & Clayton, J. "Fuzzy logic in computer-aided breast-cancer diagnosis: Analysis of lobulation," Artificial Intelligence in Medicine, 11, 75–85, (1997).
[10] Michie, D., Spiegelhalter, D. J., & Tayor, C. C "Machine learning, neural and statistical classification," London: Ellis Horwood, (1994).
[11] Pendharkar, P. C., Rodger, J. A., Yaverbaum, G. J., Herman, N., & Benner, M "Associations statistical, mathematical and neural approaches for mining breast cancer patterns," Expert Systems with Applications, 17,223–232, (1999).
[12] Rushing, J. A., Ranganath, H. S., Hinke, T. H., & Graves, S. J "Image segmentation using association rule features,", IEEE Transactions on Image Processing, 11, 558–566, (2002).
[13] Ryua, Y. U., Chandrasekaranb, R., & Jacobc, V. S. "Breast cancer prediction using the isotonic separation technique," European Journal of Operational Research, 181, 842–854, (2007).
[14] Sahan, S., Polat, K., Kodaz, H., & Günes, S. "A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis,", Computers in Biology and Medicine, 37, 415–423, (2007).
[15] Smith, F. W. "Pattern classifier design by linear programming," IEEE Transactions on Computers C-17(4), 367–372, (1968).
[16] Türkoglu, I., Arslan, A., & Ilkay, E. "An intelligent system for diagnosis of the heart valve diseases with wavelet packet neural networks," Computers in Biology and Medicine, 33, 319–331, (2003).

[17] Übeyli, E. D. "Implementing automated diagnostic systems for breast cancer detection," Expert Systems with Applications, 33, 1054–1062, (2007).
[18] R.S.Subhashini, V.Ramalingam, S.Palanivel "Breast mass classification based on cytological patterns using RBFNN and SVM," Expert Systems with Applications, (2009).
[19] Mehmet Faith Akay "Support vector machines combined with feature selection was proposed to classify breast cancer diagnosis," Expert Systems with Applications, (2009).

