International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 15 (2017) pp. 4775-4781
© Research India Publications. http://www.ripublication.com

Breast Cancer Classification using RBF and BPN Neural Networks

Vijayalakshmi S [1] and Priyadarshini J [2]

[1]Research Scholar, SCSE, Vellore Institute of Technology, Chennai Campus, Chennai, India.

[2]Associate Professor, SCSE, Vellore Institute of Technology, Chennai Campus, Chennai, India.

Orcid : 0000-0001-8693-4629

Abstract

This paper explores the possible diagnosis of breast cancer using Radial Basis Function (RBF) for the data set. The use of machine learning and data mining techniques has revolutionized the whole process of breast cancer Diagnosis and Prognosis. Breast Cancer Diagnosis distinguishes benign from malignant breast lumps and Breast Cancer Prognosis predicts when Breast Cancer is likely to recur in patients that have had their cancers excised. We analyze the breast Cancer data available from the Wisconsin Breast Cancer WBC, WDBC from UCI machine learning with the aim of developing accurate prediction models for breast cancer using RBF. Overall, the RBF neural network technique has proved better performance than that of the BPN technique.

Keywords: Classification; Wisconsin breast cancer data, Back Propagation Neural Network, Breast Cancer, Radial Basis Function

INTRODUCTION

Neural Networks are currently a ‘hot’ research area in medicine, particularly in the fields of radiology, urology, cardiology, oncology and etc. Keeping in view of the significant characteristics of Neural Network (NN) and its advantages for implementation of the classification problem, Neural Network technique is highly used in classification of data related to medical field[22]. Owing to their wide range of applicability and their ability to learn complex and nonlinear relationships including noisy or less precise information, Neural Networks (NN) technique is used to solve problems in biomedical engineering. By their nature, Neural Networks are capable of high-speed parallel signal processing in real time. They have an advantage over conventional technologies because they can solve problems that are too complex that do not have any algorithmic solution or for which an algorithmic solution is too complex. Neural Networks are trained by examples instead of rules that are automated. This is one of the major advantages of Neural networks over traditional expert systems.

Cancer refers to the uncontrolled multiplication of a group of cells in particular location of the body. A group of rapidly dividing cells may form a lump, micro calcifications or architectural distortions which are usually referred to as tumors. Breast cancer is any form of malignant tumor which develops from breast cells. Breast cancer is one of most hazardous types of cancer among women in the world. The world health organization’s International Agency for Research on cancer (IARC) estimates that more than 400,000 women expire each year with breast cancer. Today, there is an urgent need in breast cancer control and it is achieved primarily by knowing different risk factors. Secondly, there is need to detect this disease in early stage by knowing different symptoms of this disease, so it can be cured. Breast cancer is mainly of two types: Invasive and Noninvasive. Invasive type is the one in which cancerous cells break through normal breast tissue barriers and spread to other parts of the body. While in non-invasive, cancerous cells remain in a particular location of the breast and do not spread to surrounding tissue, ducts or lobules. Breast analysis techniques have been improved over the last decade. Number of automated classification systems has been developed over last years. Different techniques have varying results. However, there still are issues to be solved: developing new and better techniques. The comparison between different systems helps us to know better system with high performance; this will assist radiologists to take accurate results regarding the disease. Radiologists still produces some variation in reading images. So, there is a need for automatic interpretation of images or automated classification system, and for this purpose classifier is required. Nowadays many techniques are used for classification but Radial Basis Function (RBF) and Back Propagation Neural Network (BPN) shows better results in many instances. This paper gives comparative analysis of RBF and BPN.

Cancer is a general term that refers to cells that grow larger than 2mm in every 3 months and multiply out of control and spreads to other parts of the body. We develop a Artificial Neural Network (ANN) models for breast cancer will be use for all type of the diagnosis and prognosis [1][2] The American Cancer Society has predicted that about 192,370 women in the United States will be diagnosed with invasive breast cancer and 40,170 women will die in 2009 [3].Artificial Neural network systems are made to learn this data by the use of training algorithms that may be specific to the system. Learning involves the extraction of rules or patterns from the historic data.

The rest of this paper is organized as follows: Section 2 reviews previous work on various classification techniques used for cancer data set. Section 3 describes Artificial Neural Network technique based on Radial Basis Function. Section 4 presents the proposed system experimental analysis Finally, we conclude this paper in Section 5.

LITERATURE REVIEW

In the literature, there are many studies done on cancer detection and/or data mining. [3], used data mining for the diagnosis of ovarian cancer. For the analysis, serum proteomics that distinguish the serum ovarian cancer cases from non-cancer ones are used. An SVM (Support Vector Machine) based method is applied and statistical testing and GA (Genetic Algorithms) based methods are used for feature selection. [4] aimed to propose a new 3-D microwave approach based on SVM classifier whose output is transformed to a posteriori probability of tumor presence. Gene expression data sets for ovarian, prostate and lung cancers are analyzed in another paper [5].

An integrated gene search algorithm (preprocessing: GA and correlation based heuristics, making predictions/ data mining: decision tree and SVM algorithms) for genetic expression data analysis is proposed. In [6] the clinical and imaging diagnostic rules of peripheral lung cancer by data mining techniques that are Association Rules (AR) of knowledge discovery process and Rough Set (RS)reduction algorithm and Genetic Algorithm (GA) of generic data analysis tool (ROSETTA) are extracted[7], deals with complementary learning fuzzy neural network (CLFNN) for the diagnosis of ovarian cancer. CLFNN-micro-array, CLFNN-blood test, CLFNN-proteomics demonstrates good sensitivity and specificity. So, it is shown that CLFNN outperforms most of the conventional methods in ovarian cancer diagnosis [8], applies the classification technology to construct an optimum cerebra vascular disease predictive model. Classification algorithms used are decision tree, Bayesian classifier, and back propagation neural network. The objective of [9] is to develop an original method to extract sets of relevant molecular bio markers (gene sequences) that can be used for class prediction and as a prognostic and predictive tool. With the help of the analysis of DNA microarrays, molecular biomarkers are generated and this analysis is based on a specific data mining technique: Sequential Pattern Discovery.

The performance of data classification by integrating artificial neural networks with multivariate adaptive regression splines (MARS) approach is explored for mining breast cancer pattern[10].This approach is based on firstly to use MARS in modeling the classification problem, then obtained significant variables are used as input variables of designed neural networks model. A comparison of three data mining techniques artificial neural networks, decision trees, and logistic regression is realized in a study to predict the survivability of breast cancer [11]. Accuracy rates are found as 93.6%, 91.2%, and 89.2%respectively. Many aspects of possible relationships among DNA viruses and breast tumors are considered [12]. Feasible clusters in DNA virus combinations that depend on the observed probability of breast cancer, fibro adenoma and normal mammary tissue are created in this study and viral prerequisites for breast carcinogenesis and the protective are determined. Obtaining bioinformatics about breast tumor and DNA viruses, and building an accurate diagnosis model for breast cancer and fibro adenoma are aimed [13].

A hybrid SVM-based strategy with feature selection to render a diagnosis between the breast cancer and fibro adenoma and to find important risk factor for breast cancer is constructed. DNA viruses, HSV-1, EBV, CMV, HPV and HHV-8 are evaluated. There is also another study related to breast cancer. Breast cancer pattern is mined using discrete particle swarm optimization and statistical method [14]. Besides, to detect breast cancer, association rules (AR) and neural network (NN) are used this time [15]. AR is used to reduce the dimension of the database and NN is used for intelligent classification. In Menendeza et al. (2010), a Self-Organizing Map (SOM) based clustering algorithm for preprocessing of samples from a breast cancer screening program is introduced. Prediction of the recurrence of breast cancer is investigated [3]. The accuracy of Cox Regression and SVM algorithms are compared and it is shown that a parallelism of adequate treatment and follow-up by recurrence prediction prevent the recurrence of breast cancer. In this study, different from the studies stated above, breast cancer is tried to be predicted whether as a benign or malignant case through seven different algorithms which have not been tried for breast cancer yet in the literature and a performance analysis is aimed to be performed.

In this section some of the related prior work on data mining methods for breast cancer diagnosis is discussed. Song et al. [16] presented a new approach for automatic breast cancer diagnosis based on artificial intelligence technology. They focused to obtain a hybrid system for diagnosing new breast cancer cases in collaboration between Genetic Algorithm (GA) and Fuzzy Neural Network. They also showed that inputs reduction (features selections) can be used for many other problems which have high complexity and strong non-linearity with huge data to be analyzed. Arulampalam and Bouzerdoum [18] proposed a method for diagnosing breast cancer and called Shunting Inhibitory Artificial Neural Networks (SIANNs). SIANN is a neural network stimulated by human biological networks in which the neurons interact among each other's via a nonlinear mechanism called shunting inhibition. The feed forward SIANNs have been applied to several medical diagnosis problems and the results were more favorable than those obtained using Multilayer Perceptions (MLPs). In addition, a reduction in the number of inputs was investigated. Setiono[17] proposed a method to extract classification rules from trained neural networks and discussed its application to breast cancer diagnosis. He also explained how the pre-processing of data set can improve the accuracy of the neural network and the accuracy of the rules because some rules may be extracted from human experience, and may be erroneous.

The data pre-processing involves the selection of significant attributes and the elimination of records with missed attribute values from Wisconsin Breast Cancer Diagnosis dataset. The rules generated by Setiono's method were more brief and accurate than those generated by other methods mentioned in the literature. Meesad and Yen [18] proposed a hybrid Intelligent System (HIS) which integrates the Incremental Learning Fuzzy Network (ILFN) with the linguistic knowledge representations. The linguistic rules have been determined based on knowledge embedded in the trained ILFN or been extracted from real experts. In addition, the method also utilized Genetic Algorithm (GA) to reduce the number of the linguistic rules that sustain high accuracy and consistency. After the system being completely constructed, it can incrementally learn new information in both numerical and linguistic forms. The proposed method has been evaluated using Wisconsin Breast Cancer Dataset (WBC) data set. The results have shown that the proposed HIS perform better than some well-known methods.

RADIAL BASIS FUNCTION

Artificial Neural Network (ANN) is a branch of computational intelligence that employs a variety of optimization tool to "learn" from past experiences and use that prior training to classify new data, identify new patterns or predict. In this work neural network models have used for the diagnosis and prediction of breast cancer. Based on the tests reports our diagnostic system will predict either Benign or Malignant. The several methods like Back propagation network (BPA), Radial Basis Function (RBF), Learning Vector Quantization (LVQ) and Competitive learning (CL) to reduce the sample-per-feature ratio and investigated multiple machine learning methods to find an optimal classifier.

All the models were simulated using a variety of parameters. Each of the models was tested using many combinations of parameters in independent experiments. After a number of experimentations and parameter alterations using theoretical understandings and best practices, the optimal classifier were identified. Experimental results of breast cancer diagnostic system using neural network models with the experimentally identified best parameter values are shown in table for each model. The optimized network parameter has found like number of hidden layer, hidden layer neurons and learning rate. Performances for testing in terms of True positive rate (TPR), True negative rate (TNR), accuracy (AC), false negative rate (FNR) and false positive rate (FPR) are calculated. All above terminologies are calculated for all models and selected best classifier with the help of performance comparison. Breast Cancer Database The Wisconsin breast cancer diagnosis (WBCD) database is the result of the efforts made at the University of Wisconsin Hospital for accurately diagnosing breast masses based solely on an FNA test. The purpose of the data set is to classify a tumor as either benign or malignant based on cell descriptions gathered by microscopic examination. The data set contains 10 attributes like Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Clump Thickness, Single Epithelial Cell Size Bare Nuclei, Bland Chromatin, Normal Nucleoli and Mitoses plus the class distribution Benign and Malignant. 699 patients database are contained which 458 are benign examples and 241 are malignant examples.16 attributes values are missed therefore we deleted all from database and 444 are used for training which contains 260 benign and 184 malignant cases. Similarly 239 used for testing which contains 184 benign and 55 malignant cases [19].

The network consist of consists of N input nodes, K hidden nodes and M output nodes. Let o_{pm} and o_{pk} be the output of output node m and hidden node k from input pattern p, respectively. Assume ω_{km} is the network weight for hidden node k and output node m, and ω_{nk} is the network weight for input node n and hidden node k. Also, let x_{pn} be the input value in input node n for input pattern p, and t_{pm} be the target output value in the output node m for input pattern p. Note that the symbol Δ represents the difference between the current and the new value in the next iteration. The standard algorithm for BPN is as follow:

Step 1: Initialization: Initialize all weights and refer to them as current weight $\omega_{km}(0)$ and ω_{nk}. Set the learning rate μ and the momentum factor α to small positive values (e.g. 0.1). Set the error threshold E and the iteration number i = 0.

Step 2: Forward Pass: Select the input pattern $x_p = \{ x_{p1}, \ldots\ldots\ldots\ldots, x_{pn}\}$ from the training set and compute o_{pm} (i) and o_{pk} .

Use the desired target $t_p = \{ t_{p1},y, t_{pm} \}$ associated with $x_p$to compute the sum of the squared system error, E(i), for all input patterns. If E(i) <= E, then the algorithm iscompleted and the convergence is achieved; otherwise, go to step 3 (Backward Pass).

Step 3: Backward Pass: Compute the changes of the weights for the next iteration $\Delta\omega_{km}(i+1)$ and $\Delta\omega_{nk}(i+1)$.

The training and testing strategy is probably the most important issue in designing a BPN. The disadvantage of this BPN is care must be taken in choosing the appropriate size of the training set. On one hand, if the training set is too small, will converge easily, but it will not predict with acceptable accuracy because the network is unable to learn the underlying patterns sufficiently well with such a limited amount of data. On the other hand, if the training set is too large, it will be more difficult for the network to converge. In fact, sometimes it fails to converge at all. Moreover, even if the network can converge, it may have learned some historical patterns that may no longer be effective because the conditions have already changed too drastically.

A function set F and a terminal instruction set T used for generating a hierarchical RBF model are described as

$$S=F \cup T = \{ +_2 ,+_3 \ldots .+_N \} \cup \{x_1 ,.., x_n\} \qquad (1)$$

where +i(i = 2; 3; .. ;N) denote non-leaf nodes instructions and taking i arguments. $x_1,x_2,..,x_n$are leaf nodes instructions and taking no other arguments. The output of a non-leaf node is calculated as a RBF neural network model. From this point of view, the instruction +i is also called a basis function operator with i inputs. In general, the basis function networks can be represented as:

$$\sum_{i=1}^{m} w_i \Psi_i(x,\theta) \qquad (2)$$

where x belongs to $R^n$is input vector $\Psi_i(x,\theta)$is the parameter vector used in the basis functions.

In the creation process of HiRBF tree , if a nonterminal instruction, +i(i = 2; 3; 4; .. ;N) is selected, i real values are randomly generated and used for representing the connection strength between the node +i and its children. In addition, two adjustable parameters a_i and b_i are randomly created as Gaussian radial basis function parameters. The overall output of HiRBF tree can be computed from left to right by depth first method, recursively as shown in fig 1.

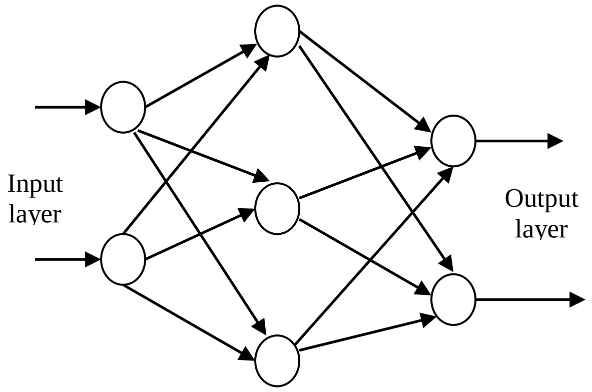

Figure 1: RBF neural network

The following the general steps that is used for processing. The general learning procedure for constructing the HiRBF network can be described as:

Step 1: Create an initial population randomly (HiRBF trees and its corresponding parameters);

Step 2: Structure optimization is achieved.

Step 3: If a better structure is found, then go to step 4, otherwise go to step 2.

Step 4: Parameter optimization is achieved. In this stage, the architecture of HiRBF model is fixed, and it is the best tree developed during the end of run of the structure search.

Step 5: If the maximum number of local search is reached, or no better parameter vector is found for a significantly long time then go to step 6; otherwise go to step 4.

Step 6: If satisfactory solution is found, then the algorithm is stopped; otherwise go to step 2.

Training methodologies are based on a set of input–output training pairs (*xk*; *yk*) (*k* = 1, 2, . . .,*K*). However, in contrast to the usual neural network training procedures that are completed in one phase, most standard training procedures for RBF networks consist of two distinct phases:

- Calculation of the hidden layer parameters;
- Determination of the connection weights between the hidden layer and the output layer.

The algorithm also assures that for any input data in the training set, there is at least one selected hidden node that is close enough according to a distance criterion.

EXPERIMENTAL ANALYSIS

The focus of this system is to evaluate the effectiveness of the proposed approach in diagnosing breast cancer. The Wisconsin Breast Cancer datasets from the UCI Machine Learning Repository is used [19], to distinguish malignant (cancerous) from benign (non-cancerous) samples. A brief description of these datasets is presented in table 1. Each dataset consists of some classification patterns or instances with a set of numerical features or attributes.

Table I: Breast Cancer Data Set

Dataset	No. of Attributes	No. of Instances
Wisconsin Breast Cancer(Original)	11	699
Wisconsin Diagnosis Breast cancer (WDBC)	32	569
Wisconsin Prognosis Breast Cancer(WPBC)	34	198

The data used in this study are provided by the UC Irvine machine learning repository located in breast-cancer-Wisconsin sub-directory, filenames root: breast-cancer-Wisconsin having 699 instances, 2 classes (malignant and benign), and 9 integer-valued attributes. We removed the 16 instances with missing values from the dataset to construct a new dataset with 683 instances. Class distribution: Benign: 458 (65.5%) Malignant: 241 (34.5%). In this study, the classification accuracy of the existing BPN technique and proposed Radial Basic Function analyzed. Construction, learning and test are different phases used in classification problems modeled by neural networks. Three layers exist in a network of back propagation, including an input layer with nine parameters, a hidden layer containing five neurons and an output layer containing a single neuron. The value of the neuron of output layer indicates if the entry corresponds to a cancer case or not. The weights of the network connection are set randomly in the learning phase. The input parameters are normalized between 0 and 1. Table 2 shows the experimental results of cancer dataset using BPN and RBF network. The goal is to have high accuracy, besides high precision and recall metrics. The result for the various dataset are given in table II and shown in fig 2-4 respectively.

Table II: Comparison of Existing and Proposed System

Technique	Accuracy	Precision	Recall
BPN	90.42	0.81	0.74
RBF	98.26	0.90	0.97

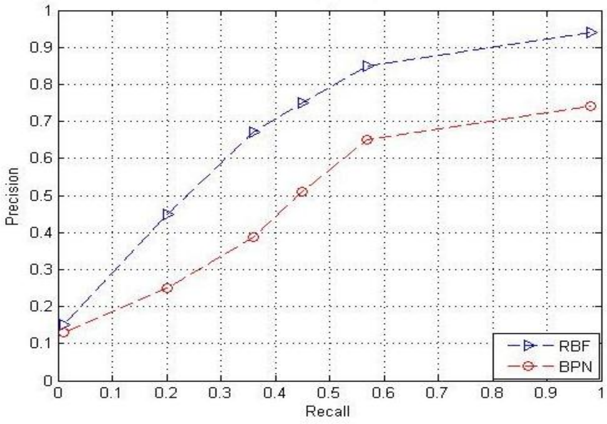

Figure 2: Comparative Analysis for WBCO

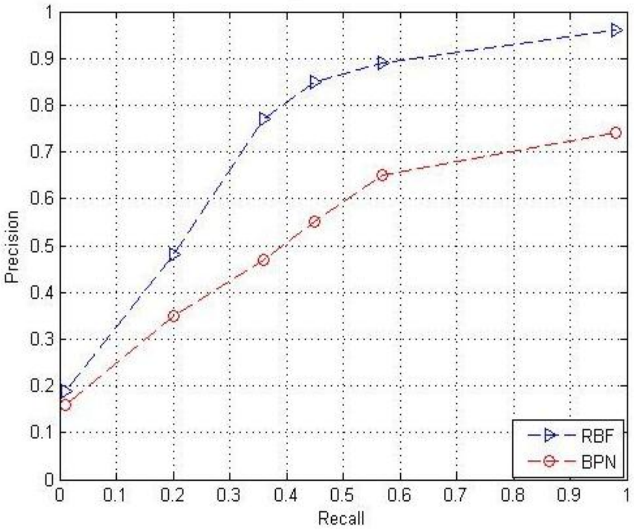

Figure 3: Comparative Analysis for WDBC

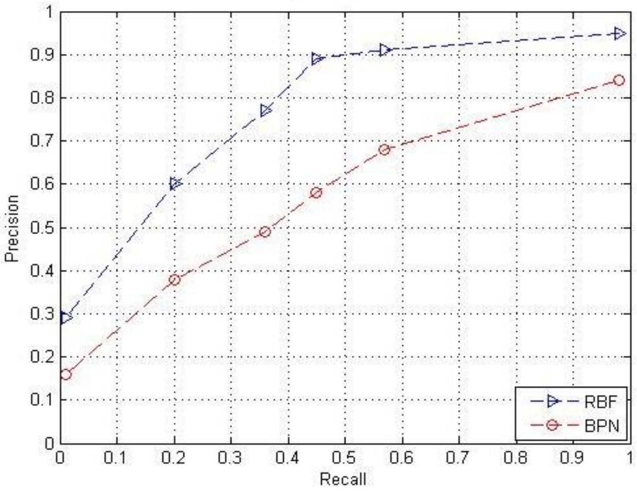

Figure 4: Comparative Analysis for WPBC

We have adopted approaches to build models to study neuronal breast cancer classification. We presented the results by modeling the attributes by neural networks. These results show the effectiveness of one technique over another and effectiveness of the extracted parameters. We can note that the RBF network, on the classification, performs better than the BPN network. Finally we note that the model produced by the supervised technique RBF can be considered a successful model for the detection and classification of various breast cancers. The study in this paper for the automatic classification of breast cancer based on RBF neural network is a new study that affects breast cancer development. Future studies will be realized using unsupervised learning technique for our problem of classification, as well as other techniques in data mining domains in order to improve the accuracy of classification.

CONCLUSION

In this paper, a new approach is developed to study the breast cancer classification based on neural network technique. The objective of this study is to create an effective tool for building neural models to help us making a proper classification of various classes of breast cancer. The interest in neural networks is justified by their own properties: learning ability, generalization and reminiscence. In this system the implementation of RBF neural to classify the breast cancer data set has been proposed. This paper has also outlined, discussed and resolved the issues, related to breast cancer dataset using the artificial neural network technique. Unlike the existing BPN technique the proposed approach gives better performance and accuracy. This study clearly shows that the preliminary results are promising for the application of the data mining methods. The analysis does not include records with missing data; future work will include the missing data which also increase the performance of the system.

REFERENCES

[1] Jiankun Hu, Member, IEEE, and Athanasios V. Vasilakos ,“Energy Big Data Analytics and Security: Challenges and Opportunities”, IEEE Transactions on Smart Grid, Vol. 7, No. 5, September 2016

[2] Ge Song; Justine Rochas; Lea El Beze; Fabrice Huet; Frédéric Magoulès, “K Nearest Neighbour Joins for Big Data on MapReduce: A Theoretical and Experimental Analysis”, IEEE Transactions on Knowledge and Data Engineering, Vol. 28, No. 9, 2016.

[3] Kim K.S., W. Kim, K.Y. Na, J.M. Park, J.Y. Kim, K.Y. Lee, J.E. Lee, S.W. Kim, R.W. Park, and Y.S. Jung (2010). “New recurrence prediction model for breast cancer by data mining”, p. 136.

[4] Kerheta, A., M. Raffettob, A. Bonia, and A. Massa (2006). “An SVM-based approach to microwave breast cancer detection”, Engineering Applications of Artificial Intelligence, No. 19, pp. 807-818.

[5] Shah, S., and A. Kusiak (2007). “Cancer gene search with data mining and genetic algorithms”, Computers in Biology and Medicine. No. 37, pp. 251-261.

[6] Qiang, Y., Y. Guo, X. Li, Q. Wanga, H. Chenc, and D. Cuicc (2007). “The diagnostic rules of peripheral lung cancer preliminary study on data mining techniques”, Journal of Nanjing Medical University. No. 21(3), pp. 190-195.

[7] Tan, T. Z., C. Queka, G. S. Ng, and K. Razvi (2008). “Ovarian cancer diagnosis with complementary learning fuzzy neural network”, Artificial Intelligence in Medicine. No. 43, pp. 207-222,

[8] Yeh, D.Y., C.H. Cheng, and Y.W. Chen (2011). “A predictive model for cerebrovascular disease using data mining”, Expert Systems with Applications.

[9] Fabregue, M., S. Bringay, P. Poncelet, M. Teisseire, and B. Orsetti (2011). “Mining microarray data to predict the histological grade of a breast cancer”, Journal of Biomedical Informatics. No. 44, pp. 12-16.

[10] Choua, S.M., T.S. Leeb, Y. E. Shaoc, and I.F. Chen (2004). “Mining breast cancer pattern using artificial neural networks and multivariate adaptive regression splines”, Expert Systems withApplications, No. 27, pp. 133-142.

[11] Delen, D., G. Walker, and A. Kadam (2005). “Predicting breast cancer survivability: a comparison of three data mining methods”, Artificial Intelligence in Medicine, No. 34, pp. 113-127.

[12] Liao H.C., and J.H. Tsai (2007). “Data mining for DNA viruses with breast cancer, fibroadenoma, and normal mammary tissue”, Applied Mathematics and Computation, No. 188, pp. 989-1000.

[13] Huang, C., H.C. Liao, and M.C. Chen (2008). “Prediction model building and future selection with support vector machines in breast cancer diagnosis”. Expert Systems with Applications, No. 34, pp. 578-587.

[14] Yeh, W.C., W.W. Chang, and Y. Y. Chung , “A new hybrid approach for mining breast cancer pattern using discrete partical swarm optimization and statistical method”, Expert Systems with Applications. No. 36, pp. 8204-8211, 2009

[15] Karabatak, M., and M.C. Ince , “An expert system for detection of breast cancer based on association rules and neural network”, Expert Systems with Applications, No. 36, pp. 3465-3469, 2009.

[16] Song, H., et al., New methodology of computer aided diagnostic system on breast cancer, *in Proceedings of the Second international conference on Advances in Neural Networks*-Volume Part III. 2005, Springer-Verlag: Chongqing, China. pp. 780-789.

[17] Setiono, R., Generating Concise and Accurate Classification Rules for Breast Cancer Diagnosis. *Artificial Intelligence in Medicine*, 2000. 18(3): pp. 205-219.

[18] Meesad, P. and G. Yen, Combined numerical and linguistic knowledge representation and its application to medical diagnosis. *Component and Systems Diagnostics, Prognostics, and Health Management* II, 2003. 4733: pp. 98-109.

[19] Frank, A. & Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA:University of California, School of Information and Computer Science.

[20] L. Bocchi, G. Coppini, J. Nori, and G. Valli, “Detection of single and clustered microcalcifications in mammograms using fractals models and neural networks,” *Med. Eng. Phys.*, vol. 26, no. 4, pp. 303–312, 2004.

[21] J. Moody and C. Darken, “Fast learning in networks of locally-tuned processing units,” *Neural Comput.*, vol. 1, pp. 281–294, 1989.

[22] Vijayalakshmi.A, Priyadarshini.J, “An Analysis of Particle Swarm Optimization Technique for Breast Cancer Dataset”, International Journal of Control Theory and Applications, pp 297-308, Volume: 9, Issue: 3, 2016