Knowledge Discovery in Clinical Databases with Neural Network Evidence Combination

#T Srinivasan [1], Arvind Chandrasekhar [2], Jayesh Seshadri [3], J B Siddharth Jonathan [4]
Department of Computer Science and Engineering,
Sri Venkateswara College of Engineering,
Sriperumbudur, India
[1]tsrini@svce.ac.in, [2]arvindcac@hotmail.com, [3] jayeshs2000@yahoo.co.in, [4]jonathansiddharth@yahoo.co.in

Abstract

Diagnosis of diseases and disorders afflicting mankind has always been a candidate for automation. Numerous attempts made at classification of symptoms and characteristic features of disorders have rarely used neural networks due to the inherent difficulty of training with sufficient data. But, the inherent robustness of neural networks and their adaptability in varying relationships of input and output justifies their use in clinical databases. To overcome the problem of training under conditions of insufficient and incomplete data, we propose to use three different neural network classifiers, each using a different learning function. Consequent combination of their beliefs by Dempster-Shafer evidence combination overcomes weaknesses exhibited by any one classifier to a particular training set. We prove with conclusive evidence that such an approach would provide a significantly higher accuracy in the diagnosis of disorders and diseases.

Keywords—Belief, Dempster-Shafer Theory, Evidence combination, Medical data mining, Neural network, Training, Uncertainty

1. Introduction

The applications of data mining in the field of medicine include diagnosis of diseases, prediction of the effectiveness of surgical procedures, analysis of medical tests and medications, and discovery of relationships among clinical and pathological data. Clinical databases store large amounts of information about patients and their medical conditions. Data mining techniques can be applied on these databases to identify patterns and relationships which can help in studying the symptoms, diagnosis and treatment of diseases [2]. These techniques are especially useful for the prediction or early diagnosis of a disease. In the case of certain diseases like cancer, early diagnosis is very important - it might help save the patient's life. This paper aims to study and apply a formal evidence combination technique for mining medical data for prediction of or screening for a disease. Input data, consisting of feature vectors, is input to three different neural network based classifiers. The classifiers used in this paper are a Back Propagation Network (BPN), Kohenen Learning Network (KN) and a Resilient Back Propagation Network (RProp). Each of the classifiers provides beliefs for each class. These pieces of evidence are then combined to reach a final diagnosis using Dempster's belief combination formula [14]. In this paper, experiments are carried on breast cancer data and thyroid disorder data [15]. The approach proposed has two primary advantages: Robustness across multiple data sets with multiple classifiers and management of uncertainty in the presence of unequal error costs.

In the rest of the paper we first give a brief introduction to the theory of belief functions and evidence. We then describe the three neural network based classifiers under consideration in Section 3. The use of the Dempster-Shafer evidence combination approach in the context of these classifiers is discussed in Section 4. Section 5 describes our experimental evaluation and the results. Section 6 concludes the paper and predicts the direction of future work in this field.

2. Background on Dempster-Shafer Theory

We present here briefly the basis of the Dempster-Shafer theory (DST) or the Mathematical Theory of Evidence (MTE), also sometimes called the theory of probable or evidential reasoning. The DST is usually considered as a generalization of the Bayesian theory of subjective probability to handle uncertain information.

Belief is, very simply, a measure of a trust or confidence that a particular event will occur [9,14]. Let us consider sources of evidence providing various degrees of support for the occurrence of event A. All degrees of support for event A are combined to form a numerical measure of belief that event A occurred. A mathematical function that translates degree of support to belief is known as a Belief Function [1]. Basic belief m(X), which represents the strength or *belief mass* of some evidence for event X provided by the source of information under consideration, has the following properties:

$$\sum m(X)=1 \qquad \text{where } X \in \Omega \qquad (1)$$

$$m(\Phi)=0 \qquad \text{where } \Phi \text{ is empty.} \qquad (2)$$

Here, Ω represents the total event space. Equation 2 indicates that the belief of an empty set is always zero.

The belief function for an event *A* is given by

$$\mathrm{Bel}(A) = \sum m(X) \qquad \text{where } X \subseteq A \text{ and } A \subseteq \Omega \qquad (3)$$

The theory of evidence deals with the evaluation of beliefs from a number of evidences [9] and their combination. For example, consider three sources of evidence P, Q and R. Let the event space be Ω = {A, B}. The measure assigned by evidence P is given by Bel_P (A), Bel_P(B) and Bel_P(uncertainty) . Since the events A and B along with the uncertain component of probability constitute exhaustive events, we have

$$Bel_P(A) + Bel_P(B) + Bel_P(uncertainty)=1 \tag{4a}$$

Similarly we have

$$Bel_Q(A) + Bel_Q(B) + Bel_Q(uncertainty)=1 \tag{4b}$$

$$Bel_R(A) + Bel_R(B) + Bel_R(uncertainty)=1 \tag{4c}$$

A decision can be made based on a combination of these beliefs.

In this paper, we use classifier output to form evidence and a decision such as benign or malignant forms an event, as indicated before. Thus, a possible event space is

Ω = *{benign, malignant}*

In the following sections we will illustrate in brief the operations of the three classifiers under consideration - Back Propagation Network (BPN), Kohenen Network (KN) and Reverse Propagation Network (RN) – and discuss how beliefs for each class and the uncertainty are calculated.

3. THE THREE NEURAL NET BASED CLASSIFIERS

Learning in neural networks proceeds as follows. During the training phase, the network is provided with examples called training sets. Each training set consists of a list of input values and the corresponding output. The network uses these training sets to *learn* the mapping from the input to the output. Each neural network follows a different method of learning.

A. Back Propagation Network (BPN)

The BPN learning process works in small iterative steps: one of the test cases is applied to the network, and the network produces some output based on the current state of it' s synaptic weights (initially, the output will be random). This output is compared to the known-good output, and a mean-squared error signal is calculated. The error value is then propagated backwards through the network, and small changes are made to the weights in each layer. The weight changes are calculated to reduce the error signal for the case in question. The whole process is repeated for each of the test cases. The cycle is repeated until the overall error value drops below some pre-determined threshold and stabilises. At this point the network has learned the given system well enough. The BPN learning network asymptotically approaches the ideal function.

BPN learning depends on two parameters - ξ, learning parameter that specifies the step width of the gradient descent, and $\delta_{\max}$, the maximum difference between a teaching value and an output that is tolerated.

B. Kohenen Learning Network (KN)

A Kohonen neural network is a self organizing mapping technique that allows one to project multidimensional points to two dimensional network. There are two key concepts important in understanding KNs - competitive learning and self-organization.

Competitive learning is simply finding a neuron that is most similar to the input pattern (W stands for the winner neuron). The network modifies this neuron and its neighbour neurons (a competitive learning with self-organization) to be even more similar to it. The input to train the network is a set of uniformly distributed points in each sector (the sector identity is not a part of input), described by x, y and z coordinates. The resultant network is able to organize points according to their sector identity.

For the winning neuron and its physical neighbors, the following training laws are used to modify weights.

$$W_{ij}(t+1) = W_{ij}(t) + \alpha(t).\gamma(t).\lfloor X_i - W_{ij}(t) \rfloor \tag{5a}$$

$$\gamma(t) = \exp\left\{ -\frac{\left(r_{ij} / \sum(t)\right)^2}{2} \right\} \tag{5b}$$

C. Resilient Back Propagation Network (RProp)

The purpose of the RProp training algorithm is to eliminate small gradient changes in the weights and biases the magnitudes of the partial derivatives. Only the sign of the derivative is used to determine the direction of the weight update; the magnitude of the derivative has no effect on the weight update. The size of the weight change is determined by a separate update value. The update value for each weight and bias is increased by a factor δ_{inc} whenever the derivative of the performance function with respect to that weight has the same sign for two successive iterations. The update value is decreased by a factor δ_{dec} whenever the derivative with respect that weight changes sign from the previous iteration. If the derivative is zero, then the update value remains the same. Whenever the weights oscillate, the weight change will be reduced. If the weight continues to change in the same direction for several iterations, then the magnitude of the weight change will be increased.

Under RProp learning (typically simulated on non-recursive networks), the learning depends on two parameters - δ_0, the starting value for all δ s, and $\delta_{\max}$, the upper limit for the update values δ .

4. EVIDENCE COMBINATION

A. Uncertainty Evaluation

Here, we see how to evaluate uncertainty for each classifier. We use the class differentiation quality as our uncertainty measure [2]. The idea behind this perspective is that the closer the values of beliefs for K classes to each other, the more uncertain the classifier is about its decision. As the beliefs start spreading apart uncertainty starts decreasing.

Let uncertainty be denoted as H(U) [2]. If there are K possible classifications, the distance between the belief values and the value 1/K are evaluated. If all the classes have the same distance then the ambiguity involved in the classification is the highest. If one class shows maximum possible distance then the ambiguity involved is the least. Generalizing from this, a measure of uncertainty can be computed.

$$H(U) = 1 - \left(\frac{K}{K-1} \right) \sum_{i=1}^{K} \left(m(i) - \frac{1}{K} \right)^2 \tag{6}$$

We use this measure to compute uncertainty as H(U) and then normalize the belief values $Bel(i) = \alpha.m(i)$ so that

$$\sum_{i=1}^{K} Bel(i) + Bel(\theta) = \sum_{i=1}^{K} \alpha.m(i) + \beta.H(U) \tag{7}$$

But $$\sum_{i=1}^{K} Bel(i) + Bel(\theta) = 1 \tag{8a}$$

And $$\sum_{l=1}^{K} m(i) = 1 \tag{8b}$$

Thus, we obtain

$$\alpha = 1 - \beta.H(U) \tag{9}$$

The Dempster-Shafer Theory (DST or Dempster's Rule) of evidence combination deals with these beliefs.

B. Combining classifiers using DST

Dempster's Rule assumes that observations are independent and have a non-empty set intersection [14]. Any two beliefs Bel_1 and Bel_2 with elements A_i and B_i respectively may be combined into a new belief function using Dempster's rule of combination [3]. Let the combined belief mass be assigned to C_k, where C is a set of all subsets produced by $A \cap B$. The mathematical representation of the rule is

$$Bel(C_k) = \frac{\sum_{A_i \cap B_i = C_k ; C_k \neq \phi} Bel(A_i) \times Bel(B_i)}{1 - \sum_{A_j \cap B_j = \phi} Bel(A_j) \times Bel(B_j)} \tag{10}$$

In this paper, we perform pair-wise combination of classifiers. We first combine, for instance, beliefs of Back Propagation Network (BPN) classifier and Kohenen Network (KN) classifier. In the second step, we combine the output of the first step (BK) with the evidence from the Resilient Back Propagation Network (RProp) classifier.

Let us assume that the BPN classifier provides beliefs Bel_BPN(A) and Bel_BPN(B), where Bel_BPN is the belief provided by BPN and A and B are the two classes (positive and negative prediction) under consideration.

Similarly for KN classifier beliefs are given as Bel_KN(A) and Bel_KN(B). Uncertainties for the two classifiers are U_BPN and U_KN respectively.

Naturally,

U_BPN =1 - Bel_BPN(A) - Bel_BPN(B) (11a)
U_KN =1 - Bel_KN(A) - Bel_KN(B) (11b)

Bel(A) is a belief mass given to class A, say benign class of the disease under consideration. It is evaluated by multiplying benign belief masses of BPN and KN, assuming independent evidence sources. Added to this is the product of uncertainty in KN and benign belief of BPN, belief for benign of KN and uncertainty of BPN. To obtain the combined belief for the hypothesis, all these basic beliefs are summed.

Thus,

Bel_comb(A) = Bel_BPN(A) x Bel_KN(A)
+ U_BPN x Bel_KN(A)
+ Bel_BPN(A) x U_KN (12)

Three terms exist in the above equation, because each term containing an uncertain term needs to be considered as *potential support* for any hypothesis.

5. Experimental Evaluation

An experimental evaluation was carried out on two datasets – one on breast cancer and the other on thyroid disorders [15]. The breast cancer data has a total of 700 instances, each consisting of 9 attributes. All the attributes take values between one and ten. The classes are benign and malignant and they are denoted as 0 and 1 respectively. The thyroid disorders dataset consists of 215 records, each consisting of 5 attributes, with three possible classes in output – normal activity, hypothyroid activity and hyperthyroid activity – denoted as 0, 1 and 2 respectively.

A. Example 1 – Breast cancer

The breast cancer data has a total of 700 instances, each consisting of 9 attributes. All the attributes take values between one and ten. The classes are benign and malignant and they are denoted as 0 and 1 respectively. 458 records belong to class benign and 242 records belong to class malignant. 450 datasets were used to train the network and 250 datasets (in five groups of 50 each) were used as test sets. In the five test sets, 148 were class benign and 102 were class malignant.

Table 1 shows the test results of breast cancer dataset in the form of a confusion matrix. Confusion matrices are listed for the three individual neural net based classifiers as well as the combination. The class denoted by 2 corresponds to indecision. This classification is extremely important in contexts where the cost of a false or erroneous classification is very high. In this case, cancer diagnosis, for instance, misdiagnosing the cancer or its symptoms as benign has a very high cost. In such circumstances, it may be preferable to refer the decision to a higher authority or expert rather than risk a potentially false decision.

TABLE 1: CLASSIFICATION OF BREAST CANCER

BPN				KN			
	0	1	2		0	1	2
0	145	1	2	0	131	8	9
1	1	88	13	1	4	86	12
RProp				B+K+R Combination			
	0	1	2		0	1	2
0	136	1	11	0	147	0	1
1	2	95	5	1	0	99	3

As shown in this table, BPN classifier shows the maximum accuracy in classification of records belonging to class 0 (benign). RProp classifier shows maximum accuracy in classification of records belonging to class 1 (malignant). As evident from the result set, the combination classifier is the most accurate overall. Fig. 1, 2 and 3 show the structure of the BPN, KN and Rprop respectively. The following table compares the accuracy of the methods:

TABLE 2: ACCURACY - BREAST CANCER DATA

Test	BPN (%)	KN (%)	RProp (%)	B+K+R (%)
1	94.0	86.00	90.00	96.00
2	96.0	96.00	90.00	100.00
3	92.0	84.00	94.00	98.00
4	94.0	88.00	94.00	98.00
5	90.0	80.00	92.00	98.00
Overall	93.2	86.80	92.00	98.40

In Table 2 above for the breast cancer data set, the overall accuracy of BPN classifier is 93.2%, 86.8% for KN and 92% for RProp. The overall accuracy of the combination classifier is 98.40%, which is the best overall accuracy by far.

B. *Example 2 – Thyroid disorders*

The thyroid disorders dataset consists of 215 records, each consisting of 5 attributes. The input data can be classified into three classes – normal activity, hypothyroid activity and hyperthyroid activity denoted as 0, 1 and 2 respectively. 138 records, consisting of 96 normal cases, 22 hypothyroid cases and 20 hyperthyroid cases, were used for training. 77 datasets (54 normal, 15 hypothyroid and 8 hyperthyroid) were used for testing the network.

Table 3 shows the test results of this dataset in the form of a confusion matrix. The class denoted by 3 corresponds to indecision in this case.

TABLE 3: CLASSIFICATION OF THYROID DISORDER

BPN					KN				
	0	1	2	3		0	1	2	3
0	33	2	8	11	0	46	2	2	4
1	0	12	0	3	1	1	11	1	2
2	0	0	2	6	2	0	0	7	1
RPN					B+K+R Combination				
	0	1	2	3		0	1	2	3
0	53	0	0	1	0	52	0	0	2
1	7	3	0	5	1	1	8	0	6
2	1	0	6	1	2	0	0	7	1

As shown in this table, RProp classifier shows the maximum accuracy in classification of records belonging to class 0. BPN classifier shows maximum accuracy in classification of records belonging to class 1, and KN for those of class 2. As evident from the result set, the combination classifier has the greatest overall accuracy. In particular, it classifies far fewer cases wrongly than the other schemes. When it does not classify a case correctly, it classifies it as uncertain in most

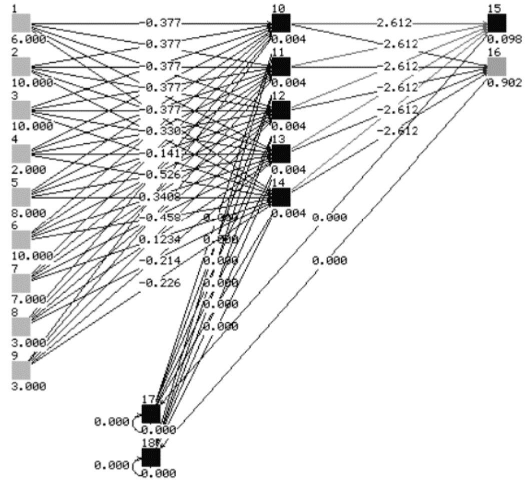

Fig. 1: Back Propagation Network for the Breast Cancer example

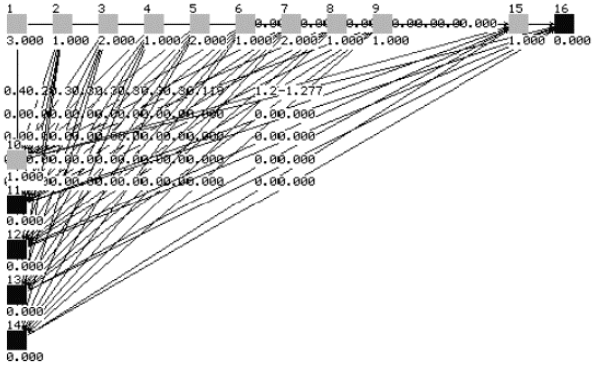

Fig. 2: Kohonen Learning Network for the Breast Cancer example

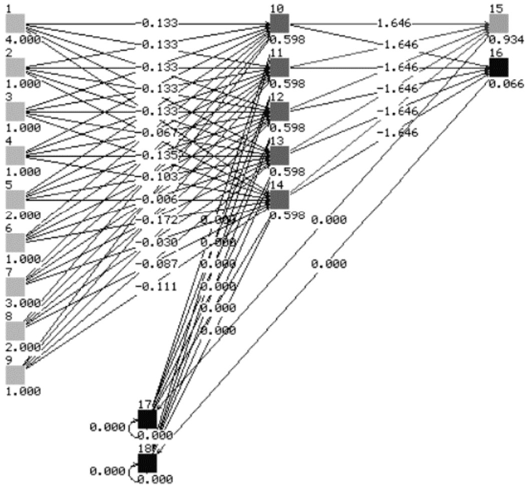

Fig. 3: Resilient Back Propagation Network for the Breast Cancer example

cases. This is, of course, far safer than a misclassification. Fig. 4, 5 and 6 show the structure of the BPN, KN and Rprop respectively.

TABLE 4: ACCURACY – THYROID DISORDER DATA

Test	BPN (%)	KN (%)	RProp (%)	B+K+R (%)
1	66.67	93.33	80.00	86.67
2	66.67	93.33	80.00	86.67
3	73.33	80.00	73.33	86.67
4	60.00	66.67	73.33	93.33
5	41.18	70.59	82.35	76.47
Overall	61.03	83.12	80.52	87.01

Table 4 compares the overall accuracy of the approaches. As can be seen accuracies of BPN, KN and RProp classifiers are a poor 61.03%, 83.12% and 80.52% respectively. The combination classifier, however, has an overall accuracy is 87.01%. The superiority of the combination classifier is even more evident as the size of the dataset under consideration increases.

Fig. 7 shows the percentage accuracy of each classifier for the two examples considered in this section. Clearly, the combination classifier is the most accurate and reliable.

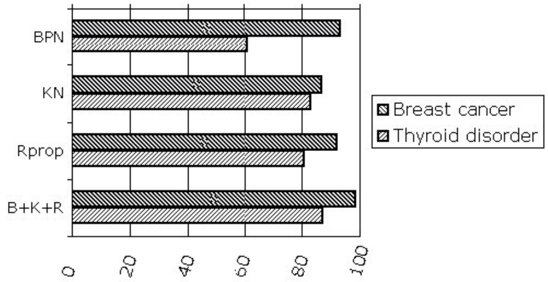

Fig. 7: Percentage accuracy of each classifier

6. CONCLUSION

We have described a method for classifying medical data by combining multiple classifiers. We have demonstrated the combination of evidences from three different classifiers using the Dempster-Shafer theory. Class differentiation quality is used for the computation of uncertainties. The combination approach has shown the best classification accuracy across two domains: breast cancer classification (benign, malignant) and thyroid disorder classification (normal, hypothyroid, hyperthyroid). The combination approach remained robust in the presence of fairly different classifier performances. This approach to classification is attractive for medical applications because of its ability to handle varying classifier performances robustly and the ability to classify samples as uncertain in the presence of classifier uncertainty. Future work could include comparisons with other classification mechanisms like Bayesian logic,

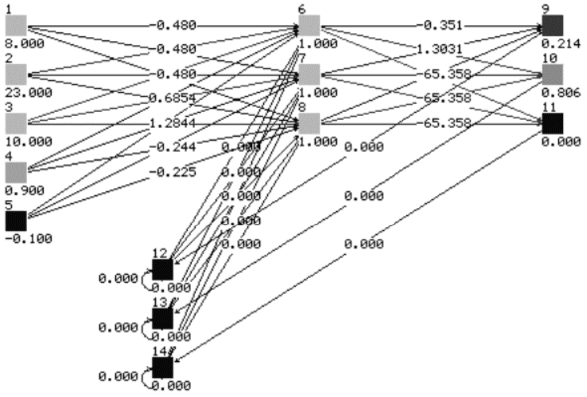

Fig. 4: Back Propagation Network for the Thyroid disorder example

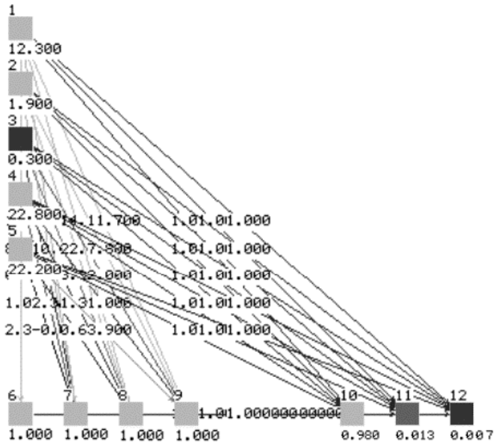

Fig. 5: Kohonen Learning Network for the Thyroid disorder example

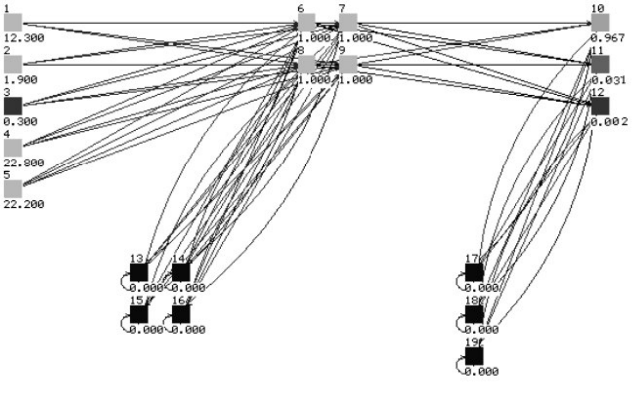

Fig. 6: Resilient Back Propagation Network for the Thyroid disorder example

fuzzy logic, transferable belief model and so on. In addition, research can be conducted to identify the optimum combination of classifiers giving the most robust and reliable performance. With adaptations, the technique used in this paper can be applied to classifiers other than the neural network and compared with our approach.

References

[1] Philippe Smets, Rudolf Kruse, "The Transferable Belief Model for Belief Representation: Uncertainty Management in Information Systems" 1996: 343-368

[2] Liu Rujie, Yuan Baozong, "A D-S Based Multi-Channel Information Fusion Method Using Classifier's Uncertainty Measurement" Proceedings Of ICSP2000, pp. 1297-1300.

[3] André Ayoun, Philippe Smets, "Data association in multi-target detection using the transferable belief model." International Journal of Intelligent Systems 16(10): 1167-1182 (2001).

[4] J.C. Prather, D.F, Lobach, L.K. Goodwin, J.W. Hales, M.L. Hage and W.E. Hammond, "Medical data mining: knowledge discovery in a clinical data warehouse." Proceedings/AMIA Annual Fall Symposium 1997

[5] O. R. Zaïane, M. L. Antonie, A. Coman, "Mammography Classification By an Association Rulebased Classifier", Proceedings of MDM/KDD 2002: 62-69.

[6] Lotfi A. Zadeh, "Syllogistic Reasoning as a Basis for Combination of Evidence in Expert Systems," IJCAI 1985: 417-419.

[7] Robin R. Murphy, "Dempster-Shafer Theory for Sensor Fusion in Autonomous Mobile Robots" IEEE Transactions on Robotics and Automation, 14:2, 197-206,1998.

[8] M. L. Antonie, O. R. Zaïane, A. Coman, "Application of Data Mining Techniques for Medical Image Classification". MDM/KDD 2001: 94-101

[9] M. Q. Ji, M. M. Marefat and P. J. Lever, "An Evidential Approach for Recognizing Shape Features", Proceedings of IEEE AIA, 1995.

[10] Y. A. Aslandogan, G. A. Mahajani, S. Taylor, "Evidence combination in medical data mining", Proceedings of ITCC 2004.

[11] O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

[12] W. H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193.

[13] M. Brameier, W. Banzhaf, "A Comparison of Linear Genetic Programming and Neural Networks in Medical Data Mining" IEEE Transactions on Evolutionary Computation, 2000.

[14] G. Shafer, "A Mathematical Theory of Evidence," Princeton University Press, 1976.

[15] http://www.ics.uci.edu/~mlearn/MLRepository.html

[16] P.C. Voukydis, "A neural network system for detection of life-threatening arrhythmias, based on Kohonen networks", Computers in Cardiology, 10-13 Sept. 1995 Pages 165 – 167

[17] P. A. Mastorocostas, "Resilient back propagation learning algorithm for recurrent fuzzy neural networks", Electronics Letters, Volume: 40, Issue: 1, 8 Jan. 2004 Pages: 57 – 58

[18] K. H. Talukder, "Performance analysis of back-propagation learning algorithm for handwritten digit classification", Multi Topic Conference, 2002. Abstracts. INMIC 2002. Dec. 27-28, 2002

[19] Li Pan; Hong Zheng; S. Nahavandi, "The application of rough set and Kohonen network to feature selection for object extraction", 2003 International Conference on Machine Learning and Cybernetics, Volume: 2, 2-5 Nov. 2003