EVALUATION OF NEURAL NETWORK BASED CLASSIFICATION SYSTEMS FOR CLINICAL CANCER DATA CLASSIFICATION

K. Mumtaz
Vivekanandha Institute of Information and Management Studies, Tiruchengode, India

S.A.Sheriff
Asan Memorial College of Arts and Science, Chennai, India

Dr. K. Duraiswamy
KS Rangasamy College of Technology, Tiruchengode, India

Abstract

The two important tasks of Data mining are clustering and classification. Breast cancer represents the second leading cause of cancer deaths in women today. Cancer cells are to be classified as either malignant or benign. This paper aims at the study of suitable machine learning technique for multidimensional, clinical cancer data classification. In this paper, three kinds of neural network based classification systems are evaluated for the proposed cancer data classification problem. The evaluated models are: 1.Adaptive Resonance Theory Based Neural Network (ART), 2.Self Organizing Map Based Neural Network (SOM) and 3.Back Propagation Neural network (BPN).

Keywords: Breast cancer diagnosis, Data mining, ANN, BPN, ART, FART, SOM.

I. Introduction

Knowledge discovery in databases (KDD) is defined as the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data [1, 2]. Some people treat data mining as a synonym for KDD. Data mining is an interdisciplinary field with a general goal of predicting outcomes and uncovering relationships in data [3, 4]. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Recent progress in data mining research has led to the developments of numerous efficient methods to mine interesting patterns and knowledge from large databases.

One of the major challenges in medical domain is the extraction of comprehensible knowledge from medical diagnosis data. The use of machine learning tools in medical diagnosis is increasing gradually. This is mainly because of the effectiveness of classification and recognition systems to help medical experts in diagnosing diseases.

Classification is also described as supervised learning [5]. It is a method of categorizing or assigning class labels to a pattern set under the supervision of a teacher. Decision trees and neural networks are the most commonly used tools for pattern classification. Here a training data set of records is accompanied by class labels. New data can be classified based on the training set by generating descriptions of the classes. In addition to the training set, there is also a test data set which is used to determine the effectiveness of a classification. In principle, the popular neural network can be trained to recognize the data directly.

The back-propagation neural network in particular has proven successful in creating useful models from large masses of complex data. Because of its pattern recognition nature it has proven robust with respect to missing data and other data irregularities.

Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters [22].

In this paper, three neural network based classification models are evaluated for their suitability for clinical cancer data classification. The objective of classification is to determine whether the outcome (class) would be 'Benign' or 'Malignant'.

II. Machine Learning

Learning methods

Learning methods in neural networks can be broadly classified into three types namely supervised learning, unsupervised learning and reinforcement learning.

a. Supervised learning

In this, every input pattern that is used to train the network is associated with an output pattern. A teacher is assumed to be present during the learning process, when a comparison is made between the network's computed output and the correct expected output to determine the error.

b. Unsupervised learning

In this learning method, the target output is not presented to the network. It is as if there is no teacher to present the desired patterns and hence, the system learns of its own by discovering and adapting to structural features in the input parameters.

c. Reinforcement learning

In this method, a teacher though available, does not present the expected answer but only indicates if the computed output is correct or incorrect. A reward is given for a correct answer computed and a penalty for a wrong answer [6].

Learning algorithms

The most basic method of training a neural network is trial and error. If the network isn't behaving the way it should, change the weighting of a random link by a random amount. If the accuracy of the network declines, undo the change and make a different one. It takes time, but the trial and error method does produce results [17].

III. The Evaluated Models

Inspired by the structure of the brain, a neural network consists of a set of highly interconnected entities, called nodes or units. Each unit is designed to mimic its biological counterpart, the neuron. Each accepts a weighted set of inputs and responds with an output. Application areas of neural networks include system identification and control (vehicle control, process control), game playing and decision making (chess, racing), pattern recognition(radar systems, face identification, object recognition), sequence recognition(gesture, speech, handwritten text recognition), medical diagnosis, financial applications, data mining, visualization and e-mail spam filtering [16]. Neural networks are the most popular and widely used Data Mining techniques.

In this section we describe the three neural network models namely Adaptive Resonance Theory Based Neural Network (ART), Self Organizing Map Based Neural Network (SOM) and Back Propagation Neural network (BPN) which are under evaluation.

A. The BPN

The feedforward, back-propagation architecture was developed in the early 1970's by several independent sources (Werbor; Parker; Rumelhart, Hinton and Williams). Currently, the back-propagation architecture is the most popular, effective, and easy-to-learn model for complex, multi-layered networks. The typical back-propagation network has an input layer, an output layer, and at least one hidden layer. There is no theoretical limit on the number of hidden layers but typically there are just one or two.

Architecture of BPN

The BPN, also called multi-layer feed-forward neural network or multi-layer perceptron, is very popular and is most widely used. The BPN is based on the supervised procedure, i.e. the network constructs a model based on examples of data with known outputs. The architecture of the BPN is a layered feedforward neural network, in which the non-linear elements (neurons) are arranged in successive layers, and the information flows unidirectionally, from input layer to output layer, through the hidden layer(s) [13].

A three layered feedforward neural network consisting of one input layer, one hidden layer and one output layer is shown below.

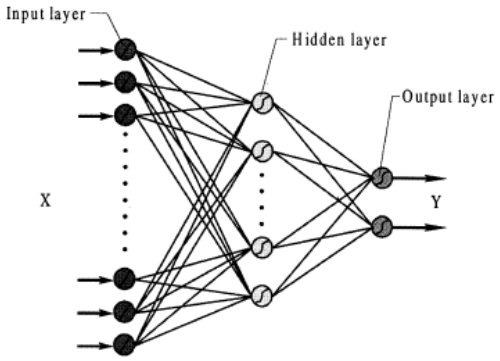


Fig.1 Feedforward Neural Network

Training of BPN

The training process normally uses some variant of the Delta Rule, which starts with the calculated difference between the actual outputs and the desired outputs. Using this error, connection weights are increased in proportion to the error times a scaling factor for global accuracy. Training inputs are applied to the input layer of the network, and desired outputs are compared at the output layer. During the learning process, a forward sweep is made through the network, and the output of each element is computed layer by layer. The difference between the output of the final layer and the desired output is back-propagated to the previous layer(s), usually modified by the derivative of the transfer function, and the connection weights are normally adjusted using the Delta Rule. This process proceeds for the previous layer(s) until the input layer is reached.

B.The SOM

A self-organizing map (SOM) or self-organizing feature map (SOFM) is a neural network approach that uses competitive unsupervised learning. Learning is based on the concept that the behavior of a node should impact only those nodes and arcs near it. Weights are initially assigned randomly and adjusted during the learning process to produce better results. During this learning process, hidden features or patterns in the data are uncovered and the weights are adjusted accordingly. The model was first described by the Finnish professor Teuvo Kohonen and is thus sometimes referred to as a Kohonen map.

The self-organizing map is a single layer feedforward network where the output syntaxes are arranged in low dimensional (usually 2D or 3D) grid. Each input is connected to all output neurons. There is a weight vector attached to every neuron with the same dimensionality as the input vectors. The goal of the learning in the self-organizing map is to associate different parts of the SOM lattice to respond similarly to certain input patterns.

A two dimensional Kohonen Self Organizing Feature Map network is shown in the figure below.

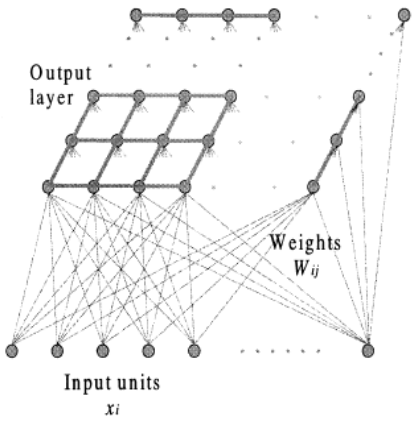


Fig.2 The SOFM Network

Training of SOM

When a training sample is given to the network, its Euclidean distance to all weight vectors is computed. The neuron with weight vector most similar to the input is called the Best Matching Unit (BMU). The weights of the BMU and neurons close to it in the SOM lattice are adjusted towards the input vector. The magnitude of the change decreases with time and is smaller for neurons physically far away from the BMU. The update formula for a neuron with weight vector $W_v(t)$ is $W_v(t + 1) = W_v(t) + \Theta(v, t)\alpha(t)(D(t) - W_v(t))$, where $\alpha(t)$ is a monotonically decreasing learning coefficient and $D(t)$ is the input vector. The neighborhood function $\Theta(v, t)$ depends on the lattice distance between the BMU and neuron v. In the simplest form it is one for all neurons close enough to BMU and zero for others, but a Gaussian function is a common choice, too. Regardless of the functional form, the neighborhood function shrinks with time [17]. At the beginning when the neighborhood is broad, the self-organizing takes place on the global scale. When the neighborhood has shrunk to just a couple of neurons the weights are converging to local estimates. This process is repeated for each input vector for a number of cycles.

During the mapping process a new input vector may quickly be given a location on the map, it is automatically classified or categorized. There will be one single winning neuron: the neuron whose weight vector lies closest to the input vector. (This can be simply determined by calculating the Euclidean distance between input vector and weight vector.)

SOM Algorithm

1. Each node's weights are initialized.
2. A vector is chosen at random from the set of training data and presented to the lattice.
3. Every node is examined to calculate which one's weights are most like the input vector. The winning node is commonly known as the Best Matching Unit (BMU).
4. The radius of the neighborhood of the BMU is now calculated.
5. Each neighboring node's (the nodes found in step 4) weights are adjusted to make them more like the input vector. The closer a node is to the BMU; the more its weights get altered.
6. Repeat step 2 for N iterations [13].

B. The ART

The basic ART system is an unsupervised learning model. It typically consists of a comparison field and a recognition field composed of neurons, a vigilance parameter, and a reset module. Higher vigilance produces highly detailed memories (many, fine-grained categories), while lower vigilance results in more general memories (fewer, more-general categories). The following figure shows the architecture of ART.

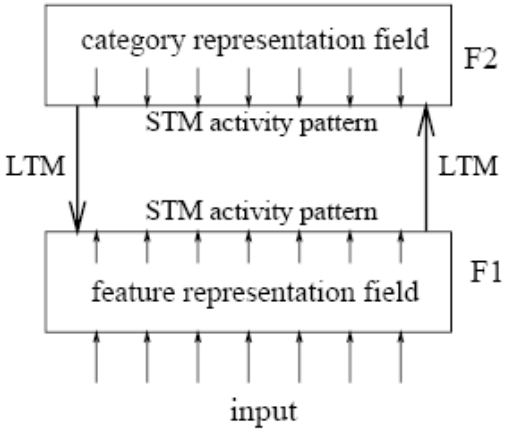


Fig.3 The ART Architecture

The system consists of two layers F1 and F2 which are connected to each other via the LTM. The input pattern is received at F1, whereas classification takes place in F2. The input is not directly classified. First a characterization takes place by means of extracting features, giving rise to activation in the feature representation field. The expectation, residing in the LTM connections translates the input pattern to a categorization in the category representation field. The classification is compared to the expectation of the network, which resides in the LTM weights from F2 to F1. If there is a match, the expectations are strengthened, otherwise the classification is rejected.

Training of ART

There are two basic methods of training ART-based neural networks: slow and fast. In the slow learning method, the degree of training of the recognition neuron's weights towards the input vector is calculated to continuous values with differential equations and is thus dependent on the length of time the input vector is presented. With fast learning, algebraic equations are used to calculate degree of weight adjustments to be made, and binary values are used [16].

The comparison field takes an input vector (a one-dimensional array of values) and transfers it to its best match in the recognition field. Its best match

is the single neuron whose set of weights (weight vector) most closely matches the input vector. Each recognition field neuron outputs a negative signal (proportional to that neuron's quality of match to the input vector) to each of the other recognition field neurons and inhibits their output accordingly. In this way the recognition field exhibits lateral inhibition, allowing each neuron in it to represent a category to which input vectors are classified. After the input vector is classified, the reset module compares the strength of the recognition match to the vigilance parameter. If the vigilance threshold is met, training commences. Otherwise, if the match level does not meet the vigilance parameter, the firing recognition neuron is inhibited until a new input vector is applied; training commences only upon completion of a search procedure. In the search procedure, recognition neurons are disabled one by one by the reset function until the vigilance parameter is satisfied by a recognition match. If no committed recognition neuron's match meets the vigilance threshold, then an uncommitted neuron is committed and adjusted towards matching the input vector.

Types of ART

ART 1 is the simplest variety of ART networks, accepting only binary inputs [8].

ART 2 extends network capabilities to support continuous inputs [9].

ART 2-A is a streamlined form of ART-2 with a drastically accelerated runtime, and with qualitative results being only rarely inferior to the full ART-2 implementation [10].

ART 3 builds on ART-2 by simulating rudimentary neurotransmitter regulation of synaptic activity [11].

Fuzzy ART implements fuzzy logic into ART's pattern recognition, thus enhancing generalizability. An optional (and very useful) feature of fuzzy ART is complement coding, a means of incorporating the absence of features into pattern classifications, which goes a long way towards preventing inefficient and unnecessary category proliferation [12].

ARTMAP, also known as Predictive ART, combines two slightly modified ART-1 or ART-2 units into a supervised learning structure where the first unit takes the input data and the second unit takes the correct output data, and then used to make the minimum possible adjustment of the vigilance parameter in the first unit in order to make the correct classification [12].

Fuzzy ARTMAP is merely ARTMAP using fuzzy ART units, resulting in a corresponding increase in efficacy [13].

IV. Evaluation and Results

"Wisconsin Breast Cancer Database" is the database used in this research to study the performance of the classification algorithms under evaluation.

Breast Cancer Dataset

Breast cancer dataset (Wisconsin Breast Cancer Database) is obtained from the UCI online machine learning repository at http://www.ics.uci.edu/~mlearn/MLRepository.html.

Metrics Used For Evaluation

In order to measure the performance of a clustering and classification system, a suitable metric is needed. The algorithms under consideration were evaluated using the measures namely Run Time and Rand Index.

a. Total Run Time

We calculated the total run time as the sum of time required for training and the time required for testing. Here we compare the CPU times only. Since the time taken for training is the very much higher and the time required for testing the network with same number of records is very very insignificant, we just mentioned the time taken for training.

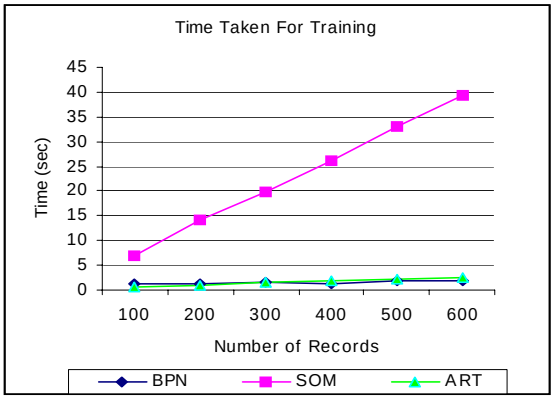
The Performance in Terms of Speed

The following table and graph shows the performance of the BPN, SOM and ART neural network based classification systems in terms of the total run time measured in terms of speed. From the table and the graph we can know that the time taken for training SOM based neural network was very high while comparing it with the other two. The other two methods (BPN and ART) consumed almost equal time for training. But the time taken for training BPN was little bit lower than ART.

Table 1. Performance in Terms of Speed

Sl.No	No. of Records	Time Taken for Training and Testing		
		BPN Error:0.01 Epochs : 100	SOM Error:0.01 Epochs : 100	ART Vigilance:0.3 Epochs : 100
1	100	1.16	7.05	0.61
2	200	1.33	14.21	1.03
3	300	1.46	19.98	1.63
4	400	1.31	26.19	1.99
5	500	1.74	32.98	2.27
6	600	1.77	39.43	2.61

Fig.4 The Time Study Chart



b. Rand index or Rand measure

The Rand index or Rand measure is a commonly used technique for the measure of similarity between two data clusters. This measure was found by W. M. Rand.

Given a set of n objects S = {O1... On} and two data clusters of S which we want to compare: X = {x1... xR} and Y = {y1, ..., yS} where the different partitions of X and Y are disjoint and their union is equal to S; we can compute the following values:

a is the number of elements in S that are in the same partition in X and in the same partition in Y,

b is the number of elements in S that are not in the same partition in X and not in the same partition in Y,

c is the number of elements in S that are in the same partition in X and not in the same partition in Y,

d is the number of elements in S that are not in the same partition in X but are in the same partition in Y.

Intuitively, one can think of a + b as the number of agreements between X and Y and c + d the number of disagreements between X and Y.

The Rand index, R has a value between 0 and 1 with 0 indicating that the two data clusters do not agree on any pair of points and 1 indicating that the data clusters are exactly the same.

The Performance in Terms of Accuracy

The following graph and table show the performance of BPN, SOM and ART in terms of accuracy measured using Rand index. It is noted that Rand Index was high in the case of BPN based neural network. The other two methods (SOM and ART) produced almost equal result. But the performance in the case of SOM was little bit better than the ART).

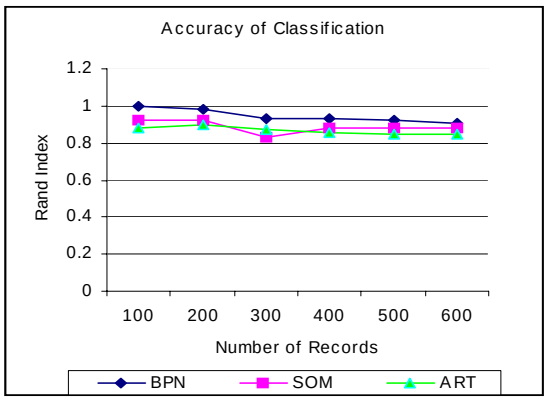


Fig.5 The Time Classification Performance Chart

Table 2. Performance in Terms of Classification Accuracy

Sl.No	No. of Records	Rand Index of Calculated and True Class Labels		
		BPN Error:0.01 Epochs : 100	SOM Error:0.01 Epochs : 100	ART Vigilance:0.3 Epochs : 100
1	100	1.0	0.92	0.88
2	200	0.98	0.92	0.90
3	300	0.93	0.83	0.87
4	400	0.93	0.88	0.86
5	500	0.92	0.88	0.85
6	600	0.91	0.88	0.85

Further, if the number of records increases, then the classification accuracy gradually decreases in all the cases. The classical BPN based neural network performed very well in all the cases and produced significantly good results than the other two methods.

V. Conclusion

The Performance of the three kinds of neural network based classification algorithm was tested with "Wisconsin Breast Cancer Database ". Several tests were made on the system and overall significant results were achieved. The classical BPN based classifier is very good in performance when compared to SOM and ART.

Since the number of dimensions in this cancer database is very low, we did not use any feature extraction or feature selection technique in this research. Instead, we directly used the cancer data for training. If we adopt any suitable feature extraction and feature selection technique, then we may expect better accuracy even while classifying higher number of records.

The present classification system may fail to produce accurate classification results for high dimensional data. So in such cases, feature extraction or feature selection technique or dimensionality reduction will be a necessary one. Future works may address these issues.

Acknowledgment

The first author extends her thanks to Prof. Dr.M.Karunanithi, Chairman & Secretary, Vivekanandha Educational Institutions, for his constant encouragement and support throughout the research work.

References

[1] U.M. Fayyad, G.Piatetsky-Shapiro, P.Smyth, and R. Uthurusamy, eds., "Advances in Knowledge Discovery and Data Mining", Menlo Park, CA: AAAI/MIT Press, 1996.

[2] K. J. CIOS, W. Pedrycz, and R. Swiniarski, "Data Mining Methods for Knowledge Discovery", Dordrecht: Kluwer, 1998.

[3] J Han and M. Kamber, "Data Mining: Concepts and Techniques", San Diego: Academic Press, 2001.

[4] D. Hand, H.Mannila, and P.Smyth, "Principles of Data Mining", London: MIT Press, 2001.

[5] J.T. Tou and R.C. Gonzalez, "Pattern Recognition Principles", London: Addison-Wesley, 1974

[6] S.Rajasekaran and G.A. Vijayalakshmi Pai, "Neural Networks, Fuzzy Logic and Genetic Algorithms Synthesis and Applications", PHI, 2007.

[7] Holsheimer, M., Kersten, M., Mannila, H., Toivonen, H. "A Perspective on Databases and Data Mining", Proceedings KDD '95.

[8] Carpenter, G.A. & Grossberg, S. (2003), Adaptive Resonance Theory, In M.A. Arbib (Ed.), The Handbook of Brain Theory and Neural Networks, Second Edition. Cambridge, MA: MIT Press

[9] Grossberg, S. (1987), Competitive learning: From interactive activation to adaptive resonance, Cognitive Science (Publication).

[10] Carpenter, G.A. & Grossberg, S. (1987), ART 2: Self-organization of stable category recognition codes for analog input patterns, Applied Optics.

[11] Carpenter, G.A., Grossberg, S., & Rosen, D.B. (1991), ART 2-A: An adaptive resonance algorithm for rapid category learning and recognition, Neural Networks (Publication).

[12] Carpenter, G.A. & Grossberg, S. (1990), ART 3: Hierarchical search using chemical transmitters in self-organizing pattern recognition architectures, Neural Networks (Publication)

[13] http://gemi.mpl.ird.fr/PDF/Lek.EM.1999.pdf

[14] http://www.ai-junkie.com/ann/som/som2.html

[15] http://www.learnartificialneuralnetworks.com

[16]http://en.wikipedia.org/wiki/Adaptive_resonance_theory

[17]http://www.virtualvetures.ca/~neil/neural/neuron-d.html

[18] ftp://ftp.fas.sfu.ca/pub/cs/han/kdd

[19] http://www.data-miner.com/

[20] http://www.kd1.com/

[21]http://www.mathworks.com/access/helpdesk/help/

[22] http://home.dei.polimi.it/matteucc/Clustering/