Detection of Breast Cancer using MLP and RBF Classifiers

C.Meenakshi*
M.Govindarajan**
A.M.Sameeullah***

Abstract

Data mining is the use of algorithms to extract the information and patterns derived by the knowledge discovery from databases. Classification maps data into predefined groups or classes. It is often referred to as supervised learning because the classes are determined before examining the data. The prognosis and diagnosis of cancer has been a challenging research problem for many researchers. The main objective of this proposed work is to compare the performance analysis of various data mining techniques to identify the breast cancer prognosis. This work employs two different kinds of neural network classifiers: the multilayer perceptron (MLP) and the radial basis function (RBF). It demonstrated the classification accuracy of MLP classifier is 79.20 % and radial basis function classifier is 77.78 %. It confirms that the MLP networks produce more specific, accurate results compared to RBF.

Keywords : Data Mining, Classification, Multi Layer Perceptron, Radial Basis Function and Classification Accuracy.

1. Introduction

Data mining methods may be distinguished by either supervised or unsupervised learning methods. In supervised methods, there is a particular pre-specified target variable, and they require a training data set, which is a set of past, heuristic examples in which the values of the target variable are provided. Classification [2] is a very common data mining task. In classification, we need to examine the features of a newly presented object and try to assign it to one of a predefined set of classes. Supervised learning methods are applied to solve classification problems. K-Nearest Neighbor, radial basis function and case based reasoning (CBR) are representative supervised learning methods that can be applied to classification problems. In the process of handling classification tasks, an important issue usually encountered is determining the best performing method for a specific problem. Several studies address the issue. For example, [1] try to find the relationship between the best performing method and data types of input/output variables. However, the common understanding of data mining practitioners and researchers is that there is no one universal best - performing method. That is, different kinds of methods have their own advantages and disadvantages. So, a method can perform best for one specific problem, but given another problem, another method can work better. This situation

*Final Year B.E (CSE) Student, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar, Tamil Nadu, India
**Research Scholar, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar, Tamil Nadu, India
***Professor, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar, Tamil Nadu, India

is called selective superiority [1]. Also, that fact implies that all of the supervised learning methods have their intrinsic limitations to improve classification accuracy.

Classification is one of the primary data mining tasks. The input to a classification system consists of sample instances, called training set, with each instance having several attributes. Attributes can be continuous, coming from an ordered domain, or categorical, coming from an unordered domain. A special class attribute indicates that the label or category to which an example belongs. The goal of classification is to induce a model from the training set, than can be used to predict the class of a new instance. The remainder of this paper is structured as follows: section 2 describes the state of art of work. Section 3 describes classification methods in order to identify the method that are appropriate (in terms of high classification accuracy) for breast Cancer prognosis. Section 4 describes the performance evaluation. Then, section 5 shows the simulated experimental results. Section 6 finally summarizes the major findings and the inferences made out of the findings.

2. State of Art

In this section, the state of the art concerning multilayer perceptron (MLP) and the radial basis function (RBF) classifiers are investigated. The results of this survey will motivate a new approach.

2.1 Related work

Breast Cancer is the leading cause of cancer and the second leading cause of cancer-related deaths of women in the United States [8]. It continues to be the most common malignancy in women. Sharma et al., have investigated that, the early detection of breast cancer is possible by analyzing gene-expression patterns in peripheral blood cells [9]. They identified a set of 37 genes that correctly predicted the diagnostic class in at least 82% of the samples.

Hitherto, the perfect tool to diagnose an early stage of breast tumor has been lacking, except in monitoring it closely throughout later years. However, some reports of early and accurate diction of small breast tumors using thermal analysis (based on increased temperature with angiogenesis) have been reported [10, 11, 12]. Taking the advantage of thermal technology and the existence of advanced neural network systems, this work will attempt in looking forward to evaluate performance of breast cancer data set that may contribute to the medical research and yield clinical benefits.

3. Classification Techniques

Two data mining methods were used in this study- multilayer perceptron (MLP) and the radial basis function (RBF) classifiers. We chose these two methods based on prior research and relevance to our problem context [3, 4]. Radial basis function neural networks have been widely used for data mining and have also been found to be effective in biomedical engineering model.

3.1 Multilayer Perceptron Classifier

The simple feed forward Neural Network that is actually called a multilayer perceptron. An MLP [5] is a network of perceptrons and used for classifying the height. The neurons are placed in layers with outputs always flowing toward the output layer. If only one layer exists, it is called a perceptron. If multiple layers exist, it is an MLP.

A two-layer neural network capable of calculating XOR. The numbers within the neurons represent each neuron's explicit threshold (which can be factored out so that all neurons have the same threshold, usually 1). This net assumes that if the threshold is not reached, zero (not -1) is output. Note that the bottom layer of inputs is not always considered a real neural network layer [6].

This class of networks consists of multiple layers of computational units, usually interconnected in a feed - forward way. Each neuron in one layer has directed connections to the neurons of the subsequent layer. In many applications the units of these networks apply a sigmoid function as an activation function.

The universal approximation theorem for neural network states that every continuous function that maps intervals of real numbers to some output interval of real numbers can be approximated arbitrarily closely by a multi-layer perceptron [7] with just one hidden layer. This result holds only for restricted classes of activation functions, e.g. for the sigmoid functions. Multi-layer networks use a variety of learning techniques, the most popular being back-propagation. Here the output values are compared with the correct answer to compute the value of some predefined error-function. By various techniques the error is then fed back through the network. Using this information, the algorithm adjusts the weights of each connection in order to reduce the value of the error function by some small amount. After repeating this process for a sufficiently large number of training cycles the network will usually converge to some state where the error of the calculations is small. In this case one says that the network has learned a certain target function. To adjust weights properly one applies a general method for non-linear optimization task that is called gradient descent. For this, the derivative of the error function with respect to the network weights is calculated and the weights are then changed such that the error decreases (thus going downhill on the surface of the error function). For this reason back-propagation can only be applied on networks with differentiable activation functions.

In general the problem of training a network to perform well, even on samples that were not used as training samples, is a quite subtle issue that requires additional techniques. This is especially important for cases where only very limited numbers of training samples are available. The danger is that the network overfits the training data and fails to capture the true statistical process generating the data. Computational learning theory is concerned with training classifiers on a limited amount of data. In the context of neural networks a simple heuristic, called early stopping, often ensures that the network will generalize well to examples not in the training set.

Other typical problems of the back-propagation algorithm are the speed of convergence and the possibility of ending up in a local minimum of the error function. Today there are practical solutions that make back - propagation in multi - layer perceptrons the solution of choice for many machine learning tasks.

3.2 Radial Basis Function Classifier

The RBF design involves deciding on their centers and the sharpness (standard deviation) of their Gaussians. Generally, the canters and SD (standard deviations) are decided first by examining the vectors in the training data. RBF networks are trained in a similar way as MLP. The output layer weights are trained using the delta rule. MLP is the most widely applied neural network technique. RBF have the advantage that one can add extra units with their centers' near parts pf the input, which are difficult to classify. Simple perceptions, MLP, and RBF networks are supervised networks. In an Unsupervised mode, the network adapts purely in response to its inputs. Such networks can learn to pick out structures in their input. One of the most popular models in the unsupervised framework is the self-organizing map (SOM), Radial basis function (RBF) networks combine a number of different concepts from approximation theory, clustering, and neural network theory. A key advantage of RBF networks for practitioners is the clear and understandable interpretation of the functionality of basis functions. Also, fuzzy rules may be extracted from RBF networks for deployment in an expert system. The RBF networks used here may be defined as follows.

- RBF networks have three layers of nodes: input layer, hidden layer , and output layer .
- Feed-forward connections exist between input and hidden layers, between input and output layers (shortcut connections), and between hidden and output layers. Additionally, there are connections between a bias node and each output node. A scalar weight is associated with the connection between nodes.
- The activation of each input node (fan out) is equal to its external input where is the element of the external input vector (pattern) of the network (denotes the number of the pattern).
- Each hidden node (neuron) determines the Euclidean distance between "its own" weight vector and the activations of the input nodes, i.e., the external input vector. The distance is used as an input of a radial basis function in order to determine the activation of node. Here, Gaussian functions are employed. The parameter of node is the radius of the basis function; the vector is its center.
- Each output node (neuron) computes its activation as a weighted sum The external output vector of the network, consists of the activations of output nodes, i.e.,. The activation of a hidden node is high if the current input vector of the network is "similar" (depending on the value of the radius) to the center of its basis function. The center of a basis function can, therefore, be regarded as a prototype of a hyper spherical cluster in the input space of the network. The radius of the cluster is given by the value of the radius parameter. In the literature, some variants of this network structure can be found, some of which do not contain shortcut connections or bias neurons. Parameters (centers, radii, and weights) of the RBF networks must be determined by means of a set of training patterns with a target vector and (supervised training). For a given input the network is expected to produce an external output.

A radial basis function (RBF) is a real-valued function whose value depends only on the distance from the origin. They are used in function approximation, time series prediction, and control. In artificial neural networks radial basis functions are utilized as activation functions.

4. Performance Evaluation

This section a detailed performance evaluation of Multilayer Perceptron and radial basis function networks.

4.1 Classification Accuracy

The primary metric for evaluating classifier performance is classification Accuracy - the percentage of test samples that are correctly classified.

Natural performance measure for classification problems:

- Success: instance's class is predicted correctly
- Error: instance's class is predicted incorrectly
- Error rate: proportion of errors made over the whole set of instances
- Accuracy: proportion of correctly classified instances over the whole set of instances
  Accuracy = 1 – error rate

The classification accuracy of 79.20 % and 77.78 % for multilayer perceptron and the radial basis function, respectively.

5. Experimental Results

In this section we demonstrated the properties and advantages of our approach by means of breast cancer data set and also we present the performance of Multilayer Perceptron and radial basis function networks. The performance of classification algorithms is usually examined by evaluating the accuracy of the classification. Classification accuracy is usually calculated determining the percentage of instances placed in the correct class. This ignores the fact that there also may be a cost associated with an incorrect assignment to the wrong class. This perhaps should also be determined. We examine the Performance of classification much as is done with information retrieval systems. With only two classes, there are four possible outcomes with the classification. The upper left and lower right quadrants are correct actions. The remaining two quadrants are incorrect actions.

Table : 1. Properties of Data Set

Dataset	Instances	Attributes
Breast cancer	726	10

Table: 2. Classification Accuracy

Data Mining Techniques	Classification Accuracy
Multilayer Perceptron (MLP)	79.20 %
Radial Basis Function (RBF)	77.78 %

6. Conclusion

Classification is an important problem in data mining. In this work we developed two different kinds of neural network classifiers: the multilayer perceptron (MLP) and the radial basis function (RBF) classifier to measure the classification accuracy for Breast Cancer data set. The decision-making was

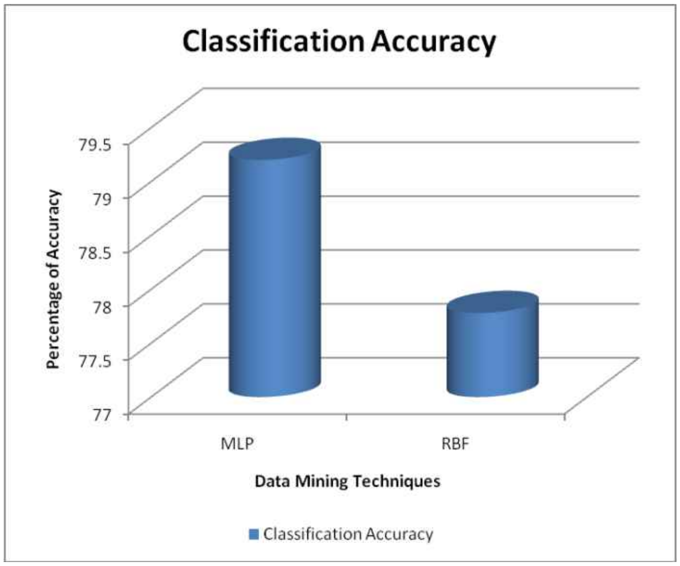

Figure: 1. Classification Accuracy

accomplished using two Classifiers (MLP and RBF), with breast cancer data set. The accuracy depends on several factors, such as the size and quality of the training set and also parameters chosen to represent the input. The results presented in Table 2 show the classifiers are effective in classification accuracy. Having said this, the MLP networks produce more specific, accurate results compared to RBF. In future the approaches will be used for exploring systems for large data sets and real time data sets.

7. Acknowledgements

Authors gratefully acknowledge the authorities of Annamalai University for the facilities offered and encouragement to carry out this work. They would also like to thank the reviewer's for their valuable remarks.

8. References

1. *Michie, D., Spiegelhalter, D. J., & Taylor, C. (1994). Machine learning. Neural and statistical classification. Ellis Horwood.*
2. *Margaret H.Dunham, "Data Mining - Introductory and Advanced Topics" Pearson Education, 2003, pp. 106-112.*
3. *Shital Shah, Adrew kusiak (2007), "Cancer gene search With data-mining and genetic algorithms", Computers in Biology and Medicine, Volume 37, issue 2, pp. 251-261, Elsevier.*
4. *Linua Li, Hong Tang, Zuobao Wu, Jianli Gong, Michael Gruidl, Jun Zou, Melvyn Tockman, Robert A.Clark, (2004), "Data Mining techniques for Cancer detection using serum proteomic profiling", Artificial Intelligence in Medicine, 32(2), pp 71-83, Elsevier.*
5. *Jiawei Han, Micheline Kamber "Data Mining – Concepts and Techniques" Elsevier, 2003, pp 360 to 365.*
6. *Bharath, Ramachandran. Neural Network Computing. McGraw-Hill, Inc., New York, 1994. pp. 4-43.*
7. *Margaret H.Dunham, "Data Mining- Introductory and Advanced Topics" Pearson Education, 2003, pp. 112-114*
8. *A.Jemal, T.Murray, E.Ward, et al. "Cancer statistics 2005, CA :Cancer J.Clin. 55 (2005). pp. 10-30.*
9. *P.Sharma, N.S.Sahni, T.Robert, P.Skaane, U.Petter, B.Hege, J.Marianne, K.Lena, M.Cecilie, S.Pradeep, Z.Alia, A.Jarle, S.Torill, A.A.Lars, S.Ellen, B.Anne-Lise, L.Anders, "Early detection of breast cancer based on gene-expression Patterns in peripheral blood cells, Breast Cancer Res. 7 (2005) R634-R644*
10. *M.Gautherine, C.Gross, "Contribution of infrared thermography to early diagnosis, pretheraputic and post-irradiation follow-up of breast carcinomas", Med Mundi. 21(1976) pp. 135-149.*
11. *J.R.Harris, M.E.Lippman, U.Vernoesi, W.Willet, "Breast cancer (Part-I), N.Engl.J.Med. 327 (1992) pp. 319-328.*
12. *E.Y.K.Ng, Y.Chen, L.N.Ung, "Computerised breast thermography: study of image segmentation and temperature cyclic variations, Int.J.Med.Eng. Technol. 25 (2001) pp. 12-16.*