A PERSPECTIVE OF DATA MINING METHOD BASED ON DRBF NEURAL NETWORKS

[1] S. SAI SATYANARAYANA REDDY, [2]P.ASHOK REDDY, [3]V.KRISHNA REDDY

[1]Professor, Department of Computer Science & Engineering,Mylavaram,LBRCE,A.P.India-521230
[2]Assistant Professor, Department of Master of Computer Appliations,Mylavaram,LBRCE,A.P.India.
[3]Associate Professor,Department of Computer Science &Engineering,Mylavaram,LBRCE,A.P.India
E-mail: [1]saisn90@gmail.com , [2] ashokreddimca@gmail.com

ABSTRACT

Recently there has been significant advances in the use of wavelet network methods in various data mining processes, With the extensively application of many databases and sharp development of Internet, The capacity of utilizing information technology to fabricate and collect data has improved greatly. It is an imperative problem to mine useful information or knowledge from large databases or data warehouses. Therefore, data mining technology is urbanized rapidly to meet the need. But data mining often faces so much data which is raucous, disorder and nonlinear. Providentially, ANN is suitable to solve the before-mentioned problems of DM because ANN has such merits as good vigor, flexibility, parallel-disposal, distributing-memory and high tolerating error. This paper gives a detailed discussion about the relevance of ANN method used in DM based on the analysis of all kinds of data mining technology, and especially lays stress on the categorization Data Mining based on RBF neural networks. Pattern classification is an important part of the RBF neural network function. Under on-line environment, the training dataset is variable, so the batch learning algorithm which will generate plenty of surplus retraining has a lower efficiency. an suitable metric for imbalanced data is applied as a filtering technique in the context of Nearest Neighbor rule, to improve the classification accuracy in RBF and MLP neural networks This paper deduces an incremental learning algorithm from the gradient descend algorithm to improve the blockage. ILA can adaptively adjust parameters of RBF networks driven by minimizing the error cost, without any surplus retraining. Using the method projected in this paper, an on-line cataloging system was constructed to resolve the IRIS classification problem.

Key words: *Data Dredging, Data Warehousing Concepts, PNN/GRNN networks ,Intrusion Detection, Information Security, and Data Networks Security, schooling algorithm Neural Networks, RBF Neural Networks.*

I. INTRODUCTION

Definition of Data Mining

Data mining is a procedure of extraction of information and knowledge that are hided in data, unknown by people and potentially useful from a large quantity of data with multiple characteristics that is uncompleted, containing noise, fuzzy and random. As a kind of cross-discipline field that synchronizes multiple disciplines including database technology, artificial intelligence, neural networks, statistics, knowledge acquirement and information extraction, nowadays data mining has becomes one of the most front research direction in the international realms of information-based decision making. Analyzing and comprehending data from different aspects, people use data mining methods to dig out useful knowledge and hidden information of prediction from a large amount of data that are stored in database and data warehouse. The methods include association rules, classification knowledge, clustering analysis, tendency and deviation analysis as well as similarity analysis. By finding valuable information from the analysis results, people can use the information to guide their business actions and administration actions, or assist their scientific researches. All of these provide new opportunities and challenges to the development of all kinds of fields related to data processing.

II. TECHNIQUES IN DATA MINING

The machine learning and the statistics are two main technical approaches of data mining. The machine learning, as a broad subfield of artificial intelligence, is concerned with the development of algorithms and techniques that allow computers to learn ability to achieve the tasks of identifying, inducing, classification, predication etc. Artificial Neural Network (ANN) and Decision Tree are the most widely applied methods in those fields. ANN is an information processing pattern that is stimulated by the way biological nervous systems, such as the brain, process information. It is collected of a large number of highly interconnected special consideration elements working in unison to solve detailed problems. The most typical neural networks are the BP neural network, the Hopfield neural networks and the adaptive neural networks. As the other technical support of data mining, the statistics offers the most fundamental theory of data mining techniques based on the precise mathematical approach. In recent years, agglomerative methods are widely applied in data mining. They can take each record as a class at the beginning. Then new classes should continuously agglomerated used K-mean algorithms until only one class could be obtained. They can solve many problems such as pattern recognition or data classification

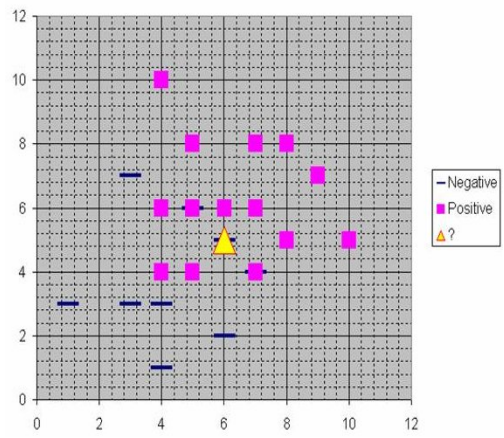
RBF Neural Networks:

A Radial Basis Function neural network has an input layer, a hidden layer and an output layer. The neurons in the secreted layer contain Gaussian transfer functions whose outputs are inversely proportional to the expanse from the center of the neuron.

RBF networks are similar to K-Means clustering and PNN/GRNN networks. The main difference is that PNN/GRNN networks have one neuron for each point in the training file, whereas RBF networks have a variable number of neurons that is usually much less than the number of training points. For problems with small to medium size training sets, PNN/GRNN networks are usually more accurate than RBF networks, but PNN/GRNN networks are impractical for large training sets.

How RBF networks work

Although the implementation is very different, RBF neural networks are conceptually similar to K-Nearest Neighbor (k-NN) models. The basic idea is that a predicted target value of an item is likely to be about the same as other items that have close values of the predictor variables. Consider this figure:



Assume that each case in the training set has two predictor variables, *x* and *y*. The cases are plotted using their *x, y* coordinates as shown in the figure. Also assume that the target variable has two categories, positive which is denoted by a square and negative which is denoted by a dash. Now, suppose we are trying to predict the value of a new case represented by the triangle with predictor values *x*=6, *y*=5.1. Should we predict the target as positive or negative?

Notice that the triangle is position almost exactly on top of a dash representing a negative value. But that dash is in a fairly unusual position compared to the other dashes which are clustered below the squares and left of center. So it could be that the underlying negative value is an odd case.

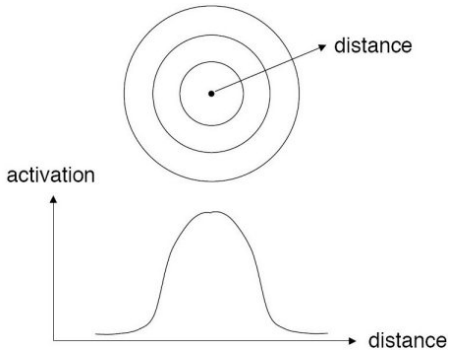
The nearest neighbor classification performed for this example depends on how many neighboring points are considered. If 1-NN is used and only the closest point is considered, then clearly the new point should be classified as negative since it is on top of a known negative point. On the other hand, if 9-NN classification is used and the closest 9 points are considered,

then the effect of the surrounding 8 positive points may overbalance the close negative point.

An RBF network positions one or more RBF neurons in the space described by the predictor variables (x,y in this example). This space has as many dimensions as there are predictor variables. The Euclidean distance is computed from the point being evaluated (e.g., the triangle in this figure) to the center of each neuron, and a radial basis function (RBF) is applied to the distance to compute the weight (influence) for each neuron. The radial basis function is so named because the radius distance is the argument to the function.

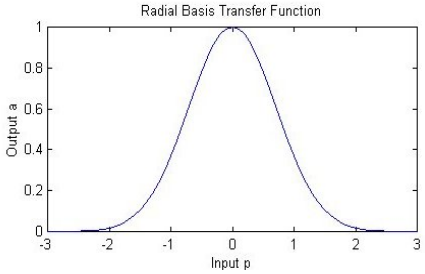
$$Weight = \text{RBF}(distance)$$

The further a neuron is from the point being evaluated, the less influence it has.

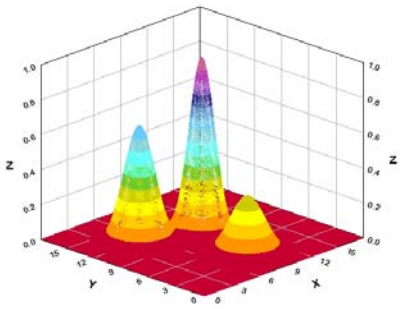


Radial Basis Function

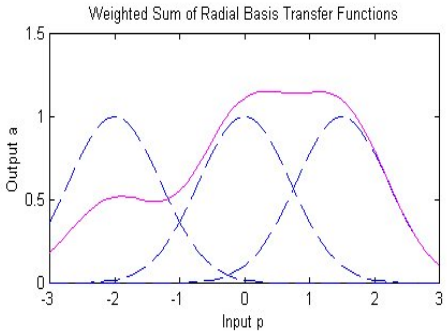
Different types of radial basis functions could be used, but the most common is the Gaussian function:



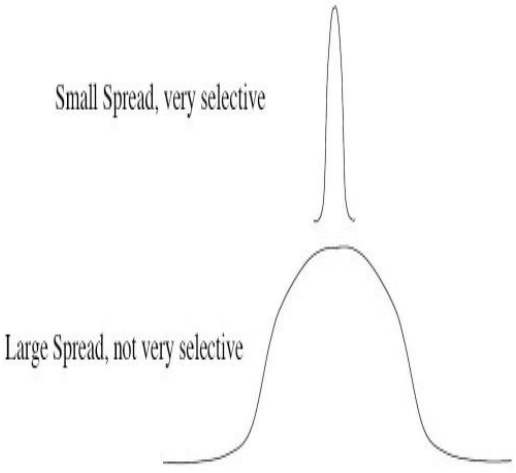
If there is more than one predictor variable, then the RBF function has as many dimensions as there are variables. The following picture illustrates three neurons in a space with two predictor variables, *X* and *Y*. *Z* is the value coming out of the RBF functions:



The best predicted value for the new point is found by summing the output values of the RBF functions multiplied by weights computed for each neuron.

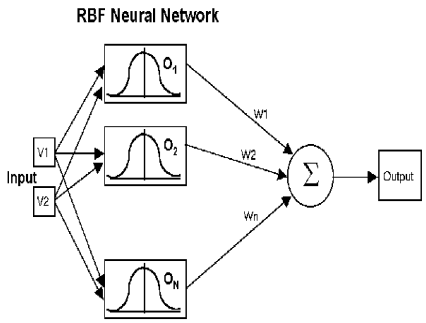


The radial basis function for a neuron has a center and a radius (also called a spread). The radius may be different for each neuron, and, in RBF networks generated by DTREG, the radius may be different in each dimension.



With larger spread, neurons at a distance from a point have a greater influence.

RBF Network Architecture



RBF networks have three layers:

1. Input layer – There is one neuron in the input layer for every predictor variable. In the case of definite variables, N-1 neurons are used where N is the number of categories. The input neurons (or processing before the input layer) standardizes the group of the values by subtracting the median and dividing by the inter quartile range. The input neurons then feed the values to every of the neurons in the hidden layer.
2. Hidden layer – This layer has a variable number of neurons (the optimal number is determined by the training process). Every neuron consists of a radial basis function centered on a point with as many dimensions as there are predictor variables. The spread (radius) of the RBF function may be dissimilar for each dimension. The centers and spreads are determined by the training procedure. When presented by means of the x vector of input values from the input layer, a hidden neuron computes the Euclidean distance of the test case from the neuron's center point and then applies the RBF kernel function to this distance using the spread values. The resulting value is passed to the summation layer.
3. Summation layer – The value coming out of a neuron in the hidden layer is multiplied by a weight associated with the neuron (W1, W2, ...,Wn in this figure) and passed to the abstract which adds up the weighted values and presents this sum as the output of the network. Not shown in this figure is a bias value of 1.0 that is multiplied by a weight W0 and fed into the summation layer. For categorization problems, there is one output (and a separate set of weights and summation unit) for each target category. The value output for a type is the probability that the case being evaluated has that category.

Training RBF Networks

The following parameters are determined by the working out process:

1. The quantity of neurons in the concealed layer.
2. The coordinates of the center of each hidden-layer RBF role.
3. The radius (spread) of each RBF function in each aspect.
4. The weights applied to the RBF function outputs as they are passed to the abstract layer.

Various methods have been used to guide RBF networks. One approach first uses K-means clustering to locate cluster centers which are then used as the centers for the RBF functions. Conversely, K-means clustering is a computationally intensive procedure, and it often does not generate the optimal number of centers. Another approach is to use a break subset of the training points as the centers.

DTREG uses a schooling algorithm developed by Sheng Chen, Xia Hong and Chris J. Harris. This algorithm uses an evolutionary approach to determine the finest center points and spreads for each neuron. It also determines when to stop adding neurons to the network by monitoring the estimated leave-one-out (LOO) error and terminating when the LOO error beings to increase due to over fitting.

The computation of the optimal weights flanked by the neurons in the hidden layer and the summation layer is done using ridge regression. An iterative method developed by Mark Orr (Orr, 1966) is used to compute the optimal regularization Lambda parameter that

minimizes generalized cross-validation (GCV) error.

Radial Basis Function Networks

After the FF networks, the radial basis function (RBF) network comprises one of the mainly used network models.

Fallowing Figure illustrates an RBF network with inputs $x_1,...,x_n$ and output $\hat{y}$. The arrows in the figure indicate parameters in the network. The RBF network consists of individual hidden layer of basis functions, or neurons. At the input of every neuron, the distance between the neuron center and the input vector is calculated. The output of the neuron is formed by applying the basis function to this distance. The RBF network output is produced by a weighted sum of the neuron outputs and the unity bias shown.

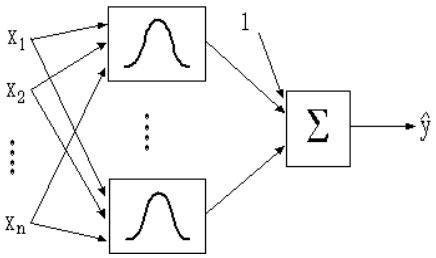


Figure: An RBF network with one output.

The RBF network in above Figure is often complemented with a linear part. This corresponds to supplementary direct connections from the inputs to the output neuron. Mathematically, the RBF network, excluding a linear part, produces an output given by

$$\hat{Y}(\theta) = \sum_{i}^{n} g(\theta,X) = \sum_{i=1}^{nb} W_i^2 \, e\text{-}\lambda_i^2(X\text{-}W1/i)^2$$

$$+W^2_{nb+1}+\chi_1 X_1+\ldots\chi_n X_n$$

Here *nb* is the number of neurons, each containing a basis function. The parameters of the RBF network consist of the positions of the basis functions w_i^1, the inverse of the width of the basis functions λ_i, the weights in output sum w_i^2, and the parameters of the linear part χ_1 ,...,χ_n. In many cases of function approximation, it is advantageous to have the additional linear part but it can be excluded by using the options.

The parameters are often lumped together in a common variable θ to make the notation compact. Then you can use the generic description *g(θ,x)* of the neural network model, where *g* is the network function and *x* is the input to the network.
Also, RBF networks may be multi-output as illustrated in fallowing figure.

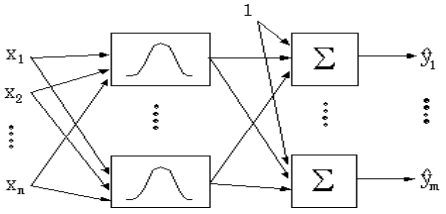


Figure: A multi-output RBF network.

FF networks and RBF networks can be used to solve a common set of problems. The integral commands provided by the package and the associated options are very similar. Problems where these networks are useful include:

- Function approximation.
- Classification.
- Modeling of self-motivated systems and time series.

III. CONCLUSION:

Current research in data mining mainly focuses on the discovery algorithm and visualization techniques. There is growing awareness that, in practice, it is easy to discover a huge number of patterns in a database where most of these patterns are actually obvious, redundant, and useless or uninteresting to the user. To prevent the user from being beleaguered by a large number of uninteresting patterns, techniques are desired to identify only the useful and interesting patterns and present them to the user.

Neural networks are suitable in data-rich environments and are typically used for extracting embedded knowledge in the form of rules, quantitative evaluation of these rules, clustering, Self-organization, classification and regression. They have an advantage, over other types of machine learning algorithms.

IV FUTURE SCOPE:

A new time series classification method based on time series motifs and dynamic radial basis function networks has been presented. The new algorithm has been applied to online signature verification. In our future work, we want to evaluate the proposed method with skilled forgeries as well. Furthermore, we are integrating the presented algorithm into a software framework for online signature verification and evaluate its performance in an ensemble of different classification methods. We also will apply the proposed algorithm to other time series classification tasks.

REFERENCES:

[1]. C. M. Bishop, Neural Networks for Pattern.

[2]. Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons

[3]. Data Mining Techniques by Michael Berry & Gordon Linoff / Paperback / 2004

[4]. Principles Of Data Mining by Hand, Mannila, and Smyth

[5]. Data Mining Algorithms by Graham Williams

[6]. Fausett L., Fundamentals of Neural
[7]. Networks, Prentice-Hall.

[8]. Gurney K., An Introduction to Neural Networks, UCL Press

[9]. Haykin S., Neural Networks , 2nd Edition, Prentice Hall,

[10]. .Radial Basis Function Networks 2(Hardcover-Mar2001)by Hewlett (Author), L.C.Jain (Author), Robert J. Hewlett

[11]. 10. J. Moody and C. J. Darken, "Fast learning in networks of locally tuned processing units," Neural Computation, 1, 281-294 (1989).

[12]. T. Poggio and F. Girosi, "Networks for approximation and learning," Proc.

[13]. Martin D. Buhmann Radial Basis Fu nctions: Theory and Implementations.

[14]. Yee, Paul V. and Haykin, Simon Regularized Radial Basis Function Networks: Theory and Applications. John Wiley.

[15]. John R. Davies, Stephen V. Coggeshall, Roger D Johns, and Daniel Schutzer, "Intelligent Security Systems," in Freedman, 15. Roy S., Flein, Robert A., and Lederman, Jess, Editors Artificial Intelligence in the Capital Markets. Chicago: Irwin.

[16]. S. Chen, C. F. N. Cowan, and P. M. Grant, "Orthogonal Least Squares Learning Algorithm for Radial Basis Function Networks", IEEE Transactions on Neural Networks, Vol 2, No 2.

****