International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:2, 2009

Performance Optimization of Data Mining Application Using Radial Basis Function Classifier

M. Govindarajan, and R. M.Chandrasekaran

***Abstract*—**Text data mining is a process of exploratory data analysis. 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. This paper describes proposed radial basis function Classifier that performs comparative cross-validation for existing radial basis function Classifier. The feasibility and the benefits of the proposed approach are demonstrated by means of data mining problem: direct Marketing. Direct marketing has become an important application field of data mining. Comparative Cross-validation involves estimation of accuracy by either stratified k-fold cross-validation or equivalent repeated random subsampling. While the proposed method may have high bias; its performance (accuracy estimation in our case) may be poor due to high variance. Thus the accuracy with proposed radial basis function Classifier was less than with the existing radial basis function Classifier. However there is smaller the improvement in runtime and larger improvement in precision and recall. In the proposed method Classification accuracy and prediction accuracy are determined where the prediction accuracy is comparatively high.

***Keywords*—**Text Data Mining, Comparative Cross-validation, Radial Basis Function, runtime, accuracy.

I. Introduction

DIRECT marketing has become an important application field for data mining. In direct marketing, companies or organizations try to establish and maintain a direct relationship with their customers in order to target them individually for specific product offers or for fund raising. Large databases of customer and market data are maintained for this purpose. The customers or clients to be targeted in a specific campaign are selected from the database, given different types of information such as demographic information and information on the customer's personal characteristics like profession, age and purchase history.

Classification [4] is one of the primary data mining task. The input to a classification system consists of example tuples, called training set, with each tuple having several attributes. Attributes can be continuous, coming from an ordered domain, or categorical, coming from an unordered domain. A special class attribute indicates the label or category to which an example belongs.

Manuscript received November 25, 2008. This work was supported in part by the first author got Career Award for Young Teachers (CAYT) grant from All India Council for Technical Education, New Delhi.

M.Govindarajan is with the Annamalai University, Annamalai Nagar, Tamil Nadu, India (phone: 91-4144-221946; e-mail: govind_aucse@yahoo.com)

R M.Chandrasekaran is with Annamalai University, Annamalai Nagar, Tamil Nadu, India (phone: 91-4144-238444; e-mail: aurmc@sify.com).

The goal of classification is to induce a model from the training set, than can be used to predict the class of a new tuple. The paper presents radial basis function classifier for direct marketing [2]. The proposed radial basis function classifier is based on comparative cross-validation. The rest of the paper is organized as follows. In section 1 brief introduction about direct marketing and classification methodology is given. Section 2 describes the state of art of work. Section 3 describes proposed radial basis function classifier using comparative cross-validation. The performance evaluation and experimental results are discussed in section 4 and section 5 respectively. The conclusion with summary is in section 6.

II. State of the Art

In this section, the state of the art concerning comparative cross validation of radial basis function algorithm is investigated. The results of this survey will motivate a new approach.

A. Related work

This article focuses on training time [5] and classification accuracy, Precision and recall using comparative cross validation of radial basis function classifier. Comparative Cross validation methods are described in section II. In general, bias and variance principle was described. The problem of training time and classification accuracy for radial basis function classifier is discussed in section III.

The problem of runtime and classification accuracy for neural networks is discussed in [5] [6]. Here, the examples of the combination of RBF and PRBF algorithm are discussed. Altogether investigated direct marketing dataset where comparative cross validation methods are applied to optimize radial basis function. The following steps are carried out to classify the radial basis function [3].

1. Input layer is used to simply input the data.
2. A Gaussian activation function is used at the hidden layer
3. A linear activation function is used at the output layer.

The objective is to have the hidden nodes learn to respond only to a subset of the input, namely, that where the Gaussian function is centered. This is usually accomplished via supervised learning. When RBF functions are used as the activation functions on the hidden layer, the nodes can be sensitive to a subset of the input values.

B. Motivation for a New Approach

Holdout, random subsampling, cross-validation and bootstrap are common techniques for accessing accuracy based on randomly sampled partitions of the given data. The use of such techniques to estimate accuracy increase the overall computation time, yet is useful for model selection. Apart from these techniques in this case, a new technique "comparative cross validation" is proposed which involves accuracy estimation by either stratified k-fold cross-validation or equivalent repeated random subsampling.

As per cross validation initial dataset (S) is divided into parts - training [Str] and test [Stst]. Subsequently, k-fold cross validation should divide data [Str] into a secondary training set [(k-1) folds] and a validation set [1 fold]. After training with cross validation, the overall prediction accuracy for Str was always significantly higher than that of Stst. By increasing the size of the Str dataset so that it is more representative of the dataset as a whole (S). That is increasing the number of training vectors, there seem to be getting much more similar training / test accuracy results.

Our goal is to calculate the expectation of the classification accuracy, as given by either Stratified k-fold cross-validation or repeated random subsampling (Jiawei Han, Micheline Kamber 2003). The classification accuracy obtained using Stratified k-fold cross-validation or repeated random subsampling where $|S|T| = N/K_S$

N - Size of S (|S|)
c(x) - The class label associated with x
C - Number of class labels in S
Ni - Number of elements in class i.
Ni = |fx : c(x) = i}|
k - Number of folds in k-fold cross validation (CV).

Let $D = (d_1, d_2 \ldots d_{ks})$ be a partition of S for Stratified k-fold cross-validation

Two accepted techniques for estimating the generalization accuracy are repeated random subsampling and Stratified k-fold cross-validation. In the former is repeated random subsampling, the validation of holdout method in which the holdout method is repeated K times. In this hold out method, S is randomly partitioned in to two independent sets, a training set and test set. Typically two third of data are allocated to training set and the remaining one third is allocated to test set. The training set is used to derive the model, whose accuracy is estimated with test set.

In latter Stratified k-fold cross-validation, the folds are stratified so that the class distribution of the tuples in each fold is approximately the same as that in the initial data.

Repeated random subsampling (T) be the classification accuracy computed by repeated random subsampling with training set T and Stratified k-fold cross-validation (D) be the classification accuracy computed by Stratified k-fold cross-validation with partition D.

Then by definition,

$$CV(D) = 1/K_S \sum_{i=1}^{K_S} \text{Repeated random subsampling } (S/d_i)$$

The expectation is, by substitution and linearity:

$$E[CV] = 1/K_S \sum_{i=1}^{K_S} E[\text{Repeated random subsampling } (S/d_i)]$$

$$= 1/K_S \sum_{i=1}^{K_S} E[E[\text{Repeated random subsampling } (S/d_i) \mid di = d]]$$

By Proposition 6.1 in Ross, 1988 (p.285).

Now:

$$E[CV] = 1/K_S \sum_{i=1}^{K_S} E[\text{Repeated random subsampling } (S/d)]$$

$$= E[\text{Repeated random subsampling } (S/d)]$$

Because E [Repeated random subsampling (S/d)] is independent of i and E[CV] = E [Repeated random subsampling (T)] by a simple correspondence of a test set d and the training set T = S/d.

Let T be the set of permissible training sets. The expectation of the classification accuracy using repeated random subsampling is simply the proportion of possible classified (overall T). The number of possible classification is

$\sum_{T \varepsilon T} |S / T|$, while the total number of correct classification is

$$A = \sum_{T \varepsilon T} \sum_{x \notin T} \text{correct}(x, T)$$

Where the binary function, correct (x,T), returns 1 iff x is correctly labeled by a classifier trained on T.

III. Classification Methods

Classification maps data into predefined groups or classes. It often referred to as supervised learning because the classes are determined before examining data. Classification algorithms require that the classes be defined based on data attribute values. They often describe these classes by looking at the characteristics of data already known to belong to the classes.

A. Existing Radial Basis Function (ERBF)

The RBF networks used here may be defined as follows.

1) RBF networks have three layers of nodes: input layer, hidden layer, and output layer
2) 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 and
3) The activation of each input node (fanout) is equal to its external input where is the th element of the external input

vector (pattern) of the network (denotes the number of the pattern).
4) 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. Any other function which satisfies the conditions derived from theorems of Schoenberg or Micchelli described in [2] may also be used as a basis function. Localized basis functions such as the Gaussian or the inverse multiquadric are usually preferred.
5) 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).

B. Proposed Radial Basis Function (PRBF)

This paper describes proposed radial basis function classifier that performs comparative cross-validation for existing radial basis function classifier. Comparative Cross-validation involves estimation of accuracy by either stratified k-fold cross-validation or equivalent repeated random subsampling.

In *K*-fold cross-validation, the original sample is partitioned into *K* subsamples. Of the *K* subsamples, a single subsample is retained as the validation data for testing the model, and the remaining $K - 1$ subsamples are used as training data. The cross-validation process is then repeated *K* times (the *folds*), with each of the *K* subsamples used exactly once as the validation data. The K results from the folds then can be averaged (or otherwise combined) to produce a single estimation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once. *10*-fold cross-validation is commonly used. Cross validation is still required to prevent over fitting.

If there are many more positive instances than negative instances in a dataset, there is a chance that a given fold may not contain any negative instances. To ensure that this does not happen, stratified *K*-fold cross-validation is used where each fold contains roughly the same proportion of class labels as in the original set of samples.

In general, stratified 10-fold cross validation is recommended for estimating accuracy (even if computation power allows using more folds) due to its relatively low bias and variance.

The Repeated random sub-sampling validation method randomly splits the dataset into training and validation data. For each such split, the classifier is retrained with the training data and validated on the remaining data. The results from each split can then be averaged. The advantage of this method (over k-fold cross validation) is that the proportion of the training/validation split is not dependent on the number of iterations (folds). The disadvantage of this method is that some observations may never be selected in the validation subsample, whereas others may be selected more than once. In other words, validation subsets may overlap.

IV. Performance Measures

This section a detailed performance evaluation of proposed radial basis function classifier.

A. 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 error and error rate for existing radial basis function classifier is one. So classification accuracy is high (99.76 %). But error and error rate for proposed radial basis function classifier is 328 (uncorrected classes) and 0.7866 respectively for 417 instances. So the classification accuracy for proposed radial basis function classifier is low (21.34 %).

B. Prediction Accuracy

Prediction can be viewed as a type of classification. The prediction accuracy is given by:

Accuracy = 1 – error rate

Where the test error (rate), or generalization error, is the average loss over the test set. Thus, the following error rate is obtained which can be expressed in term of Mean Squared Error (MSE). The error rate (MSE) for existing radial basis function classifier is 0.0003. So Prediction accuracy is 99.96 %. But error rate (MSE) for proposed radial basis function classifier is 0.1041. So the Prediction accuracy is 89.58 %.

C. Runtime

The serial runtime of a program is the time elapsed between the beginning and the end of its execution on a sequential computer.

D. Bias-Variance Tradeoff

The bias of an estimator is the difference between the expected value of the estimator and the actual value. The degree to which numerical data tend to spread is called the dispersion, or variance of the data

The bias-variance tradeoff principle can be stated as follows:

- Dataset with too few parameters are inaccurate because of a large bias (not enough flexibility).
- Dataset with too many parameters are inaccurate because of a large variance (too much sensitivity to the sample).

For a given dataset, the comparative cross-validation exhibits high bias and variance due to large numbers of uncorrected class and there by large of errors respectively. Hence we obtained low accuracy.

E. Precision and Recall

Precision and Recall are two widely used measures for evaluating the quality of results in domains such as Information Retrieval and statistical classification [12]. Precision can be seen as a measure of exactness or fidelity, whereas Recall is a measure of completeness.

In a statistical classification [15] task, the Precision for a class is the number of true positives (i.e. the number of items correctly labeled as belonging to the class) divided by the total number of elements labeled as belonging to the class (i.e. the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class). Recall in this context is defined as the number of true positives divided by the total number of elements that actually belong to the class (i.e. the sum of true positives and false negatives, which are items which were not labeled as belonging to that class but should have been).

V. Experimental Results

In this section, the properties and advantages of this approach are demonstrated by means of directing marketing data set and also the performance of PRBF is evaluated. The performance of classification algorithm is usually examined by evaluating the accuracy of the classification.

TABLE I
PROPERTIES OF DATA SETS

Dataset Factor of	INSTANCES	Attribues
Direct Marketing	417	256

The Performance of classification is examined 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.

Classification accuracy is usually calculated by determining the percentage of tuples 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. The Performance of classification is examined 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 II
RUN TIME (SECONDS)

Existing radial basis function (ERBF)	Proposed radial basis function (PRBF)	Faster By
12.84	10.94	1.90

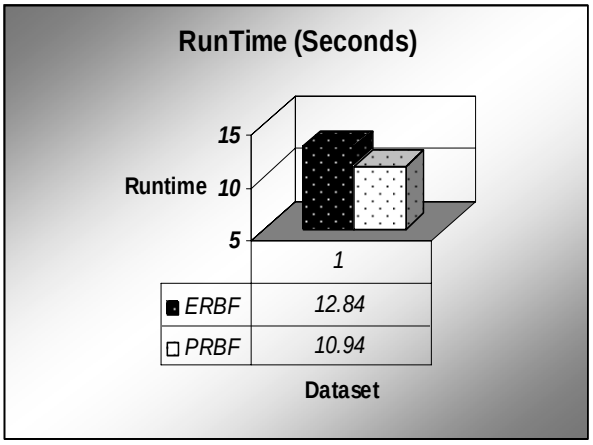

Fig. 1 Run Time (Seconds)

TABLE III
CLASSIFICATION ACCURACY

Existing radial basis function (ERBF)	Proposed radial basis function (PRBF)	Reduced By
99.76 %	21.34 %	78.42 %

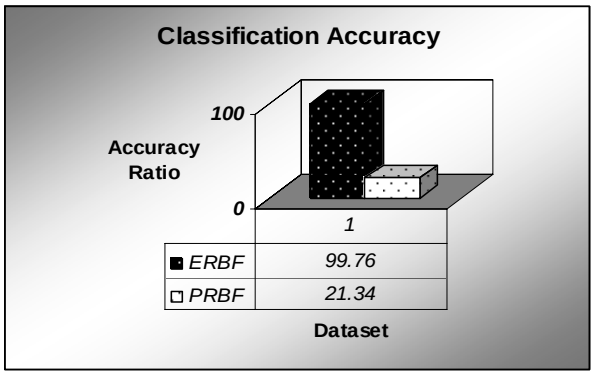

Fig. 2 Classification Accuracy

TABLE IV
PREDICTION ACCURACY

Existing radial basis function (ERBF)	Proposed radial basis function (PRBF)	Reduced By
99.96 %	89.58 %	10.38 %

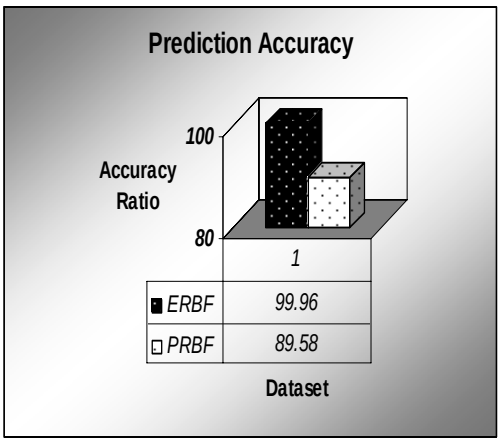

Fig. 3 Prediction Accuracy

TABLE V
PRECISION AND RECALL

Algorithms	Precision	Recall
Existing radial basis function (ERBF)	0.99	0.99
Proposed radial basis function (PRBF)	12.8	16.4

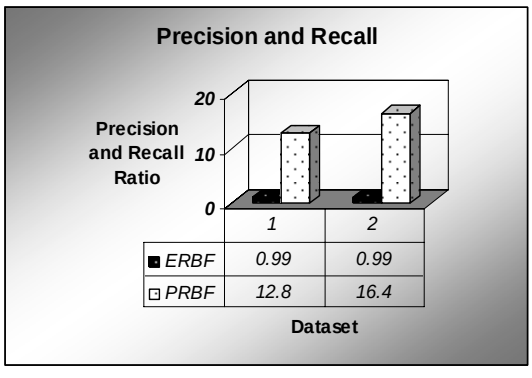

Fig. 4 Precision and Recall

VI. CONCLUSION

Classification is an important problem in data mining. In this work one text mining classifier is developed using radial basis function classifier to measure the training time, Classification accuracy, precision and recall for direct marketing data set. First, utilized our developed text mining algorithm, including text mining techniques based on classification of data upon one data set. After that, existing radial basis function classifier is employed to deal with measure training time, Classification accuracy, precision and recall for direct marketing data set. Experimental results show that the proposed method may have high bias; its performance (accuracy estimation in our case) may be poor due to high variance. Thus the accuracy with proposed radial basis function classifier was less than with the existing radial basis function classifier. However there is smaller the improvement in runtime and larger improvement in precision and recall. In the proposed method classification accuracy and prediction accuracy are determined where the prediction accuracy is comparatively high. This algorithm is independent of specify data sets so that many ideas and solutions can be transferred to other classifier paradigms.

ACKNOWLEDGMENT

Authors gratefully acknowledge the authorities of Annamalai University for the facilities offered and encouragement to carry out this work. This part of work is supported in part by the first author got Career Award for Young Teachers (CAYT) grant from All India Council for Technical Education, New Delhi. They would also like to thank the reviewer's for their valuable remarks.

REFERENCES

[1] Oliver Buchtala, Manual Klimek and Bernhard Sick, Member, IEEE " Evolutionary Optimization of Radial Basis Function Classifier for Data Mining Applications", IEEE Transactions on systems,man,andcybernets,vol.35,No.5, October,2005
[2] Blake, C., & Merz, C. (1998). UCI repository of machine learning databases. http://www.ics.uci.edu/˜mlearn/MLRepository.html.
[3] C. L. Bauer. A direct mail customer purchase model. Journal of Direct Marketing, 2:16–24, 1988.
[4] Dietterich, T. (1998). Approximate statistical tests for comparing supervised classification learning algorithms.Neural Computation, 10, 1895–1923.
[5] Friedman, J., Bentley, J., &Finkel, R. (1977). An algorithm for finding best matches in logarithmic expected time. ACM Transactions on Mathematical Software, 3, 209–226.
[6] Jiawei Han, Micheline Kamber " Data Mining – Concepts and Techniques" Elsevier, 2003, pages 359 to 365.
[7] N. Jovanovic, V. Milutinovic, and Z. Obradovic, Member, IEEE, "Foundations of Predictive Data Mining" (2002)
[8] J. M. Sousa, U. Kaymak, and S. Madeira. A comparative study of fuzzy target selection methods in direct marketing. In Proceedings of the 11th IEEE International Conference on Fuzzy Systems, Hawaii, USA, May 2002.
[9] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of International Joint Conference on Artificial Intelligence (pp. 1137–1143).
[10] Margaret H.Dunham, "Data Mining- Introductory and Advanced Topics" Pearson Education, 2003, page 112.
[11] Mitchell, T. (1997). Machine learning. New York: McGraw-Hill.
[12] Naohiro lshiil, Eisuke suchiya, Yongguangao and Nobuhiko yamaguchi, "Combining Classification Improvements by Ensemble Processing"

Proceedings of the 2005 Third ACIS Int'l Conference on Software Engineering Research, Management and Applications (SERA'05) 0-7695-2297-1/05 $20.00 © 2005 IEEE
[13] Ross, S. (1988). A first course in probability. New York: Macmillan.
[14] Sara Madeira Joao M.Sousa, "Comparison of target selection methods in direct Marketing" Technical University of Lisbon, Institution Superior T'echicio, Dept.Mechanical Eng./IDMEC, 1049-001 Lisbon, Portugal (2002).
[15] Vapnik, V. (1998). Statistical learning theory. New York: Wiley.

M. Govindarajan received the B.E and M.E and Pursuing Doctoral Degree in Computer Science and Engineering from Annamalai University, Tamil Nadu, India in 2001 and 2005 and 2006 respectively. He is currently a lecturer (Senior Scale) at the Department of Computer Science and Engineering, Annamalai University, Tamil Nadu, India. He has presented and published more than 20 papers in Conferences and Journals. His current Research Interests include Data Mining and its applications, Algorithms, Text Mining. He was the recipient of the Achievement Award for the field and to the Conference Bio-Engineering, Computer science, Knowledge Mining (2006), Prague, Czech Republic and All India Council for Technical Education "Career Award for Young Teachers (2006), New Delhi, India. He is Life Member of Computer Society of India, Indian Society for Technical Education and Session Member of Indian Science Congress Association, Associate member of Institute of Engineers.

R. M.Chandrasekaran received the B.E Degree in Electrical and Electronics Engineering from Maduari Kamaraj University in 1982 and the MBA (Systems) in 1995 from Annamalai University, M.E in Computer Science and Engineering from Anna University and PhD Degree in Computer Science and Engineering from Annamalai University, Tamil Nadu, India in 1998 and 2006 respectively. He is currently working as a Founder Registrar of Anna University, Tirchirappalli and formerly Professor at the Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar, Tamil Nadu, India. From 1999 to 2001 he worked as a software consultant in Etiam, Inc, California, USA. He has conducted Workshops and Conferences in the Areas of Multimedia, Business Intelligence and Analysis of algorithms, Data Mining. He has presented and published more than 32 papers in conferences and journals and is the author of the book Numerical Methods with C++ Program (PHI, 2005). His Research interests include Data Mining, Algorithms, Networks, Software Engineering, Network Security, and Text Mining. He is Life member of Computer Society of India, Indian Society for Technical Education, Institute of Engineers and Indian Science Congress Association.