I.J. Intelligent Systems and Applications, **2014, 03, 84-95**
Published Online February 2014 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2014.03.09

A Hybrid RBF-SVM Ensemble Approach for Data Mining Applications

M.Govindarajan
Assistant Professor, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar – 608002, Tamil Nadu, India
E-mail: govind_aucse@yahoo.com

***Abstract*—** One of the major developments in machine learning in the past decade is the ensemble method, which finds highly accurate classifier by combining many moderately accurate component classifiers. This paper addresses using an ensemble of classification methods for data mining applications like intrusion detection, direct marketing, and signature verification. In this research work, new hybrid classification method is proposed for heterogeneous ensemble classifiers using arcing and their performances are analyzed in terms of accuracy. A Classifier ensemble is designed using a Radial Basis Function (RBF) and Support Vector Machine (SVM) as base classifiers. Here, modified training sets are formed by resampling from original training set; classifiers constructed using these training sets and then combined by voting. The proposed RBF-SVM hybrid system is superior to individual approach for intrusion detection, direct marketing, and signature verification in terms of classification accuracy.

***Index Terms*—** Machine learning, Radial Basis Function, Support Vector Machine, Intrusion Detection, Direct Marketing, Signature Verification, Ensemble, Classification Accuracy

I. 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 examples in which the values of the target variable are provided. Classification is a very common data mining task. 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, Michie, Spiegelhalter, and Taylor try to find the relationship between the best performing method and data types of input/output variables. Hybrid models have been suggested to overcome the defects of using a single supervised learning method, such as radial basis function and support vector machine techniques. Hybrid models combine different methods to improve classification accuracy. The term combined model is usually used to refer to a concept similar to a hybrid model. Combined models apply the same algorithm repeatedly through partitioning and weighting of a training data set. Combined models also have been called Ensembles. Ensemble improves classification performance by the combined use of two effects: reduction of errors due to bias and variance.

1.1 Intrusion Detection

Traditional protection techniques such as user authentication, data encryption, avoiding programming errors and firewalls are used as the first line of defense for computer security. If a password is weak and is compromised, user authentication cannot prevent unauthorized use, firewalls are vulnerable to errors in configuration and suspect to ambiguous or undefined security policies (Summers, 1997). They are generally unable to protect against malicious mobile code, insider attacks and unsecured modems. Programming errors cannot be avoided as the complexity of the system and application software is evolving rapidly leaving behind some exploitable weaknesses. Consequently, computer systems are likely to remain unsecured for the foreseeable future. Therefore, intrusion detection is required as an additional wall for protecting systems despite the prevention techniques. Intrusion detection is useful not only in detecting successful intrusions, but also in monitoring attempts to break security, which provides important information for timely countermeasures (Heady et al., 1990; Sundaram, 1996). Intrusion detection is classified into two types: misuse intrusion detection and anomaly intrusion detection.

Several machine-learning paradigms including neural networks (Mukkamala et al.,2003), linear genetic programming (LGP) (Mukkamala et al., 2004a), support vector machines (SVM), Bayesian networks, multivariate adaptive regression splines (MARS) (Mukkamala et al., 2004b) fuzzy inference systems (FISs) (Shah et al.,2004), etc. have been investigated for the design of IDS. In this paper, the performance of decision trees (DT), SVM, hybrid DT–SVM and an ensemble approach is investigated and evaluated. The motivation for using the hybrid approach is to improve the accuracy of the intrusion detection system when

compared to using individual approaches. The primary objective of this paper is ensemble of radial basis function and Support Vector Machine is superior to individual approach for intrusion detection in terms of classification accuracy.

1.2 Direct Marketing

Direct marketing (Sara Madeira Joao M.Sousa, 2000) has become an important application field for data mining. In direct marketing (C. L. Bauer, 1998) 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.

The customers of a company are regarded as valuable business resources in competitive markets, leading to efforts to systematically prolong and exploit existing customer relations. Consequently, the strategies and techniques of customer relationship management (CRM) have received increasing attention in management science. CRM features data mining as a technique to gain knowledge about customer behaviour and preferences.

Data mining problems in the CRM domain, such as response optimization to distinguish between customers who will react to a mailing campaign or not, churn prediction, in the form of classifying customers for churn probability, cross-selling, or up-selling are routinely modeled as classification tasks, predicting a discrete, of- ten binary feature using empirical, customer centered data of past sales, amount of purchases, demographic or psychographic information etc. Customer retention has a significant impact on firm profitability. Gupta et al find that a 1% improvement in retention can increase firm value by 5% (Gupta, Sunil, et al., 2004). Churn refers to the tendency for customers to defect or cease business with a company. Marketers interested in maximizing lifetime value realize that customer retention is a key to increasing long-run firm profitability. A focus on customer retention implies that firms need to understand the determinants of customer defection (churn) and are able to predict those customers who are at risk of defection at a particular point in time.

Response modeling is usually formulated as a binary classification problem. The customers are divided into two classes, respondents and non-respondents. A classifier is constructed to predict whether a given customer will respond or not. From a modeling point of view, however, several difficulties arise (Shin, H. J., & Cho, S, 2006 and Zahavi, J., & Levin, N, 1997). One of the most noticeable is a severe class imbalance resulting from a low response rate: typically less than 5% of customers are respondents (Gonul, F. et al., 2000). A typical binary classifier will result in lopsided outputs to the non-respondent class (Kubat, et al., 1997). In other words, the classifier will predict most or even all customers not to respond. Although the classification accuracy may be very high since a majority of customers are in fact non-respondents. In this work, a model which identifies a subset of customers is constructed that includes as many respondents and as few non-respondents as possible Various classification methods have been used for response modeling such as statistical and machine learning methods. Recently, SVMs have drawn much attention and a few researchers have implemented them for response modeling (Shin, H. J., et al., 2006 and Yu, E., et al., 2006).

Recently, hybrid data mining approaches have gained much popularity; however, a few studies have been proposed to examine the performance of hybrid data mining techniques for response modeling (Maryam et al., 2013).

1.3 Signature Verification

Optical Character Recognition (OCR) is a branch of pattern recognition, and also a branch of computer vision. OCR has been extensively researched for more than four decades. With the advent of digital computers, many researchers and engineers have been engaged in this interesting topic. It is not only a newly developing topic due to many potential applications, such as bank check processing, postal mail sorting, automatic reading of tax forms and various handwritten and printed materials, but it is also a benchmark for testing and verifying new pattern recognition theories and algorithms. In recent years, many new classifiers and feature extraction algorithms have been proposed and tested on various OCR databases and these techniques have been used in wide applications. Numerous scientific papers and inventions in OCR have been reported in the literature. It can be said that OCR is one of the most important and active research fields in pattern recognition. Today, OCR research is addressing a diversified number of sophisticated problems. Important research in OCR includes degraded (heavy noise) omni font text recognition, and analysis/recognition of complex documents (including texts, images, charts, tables and video documents). Handwritten numeral recognition, (as there are varieties of handwriting styles depending on an applicant's age, gender, education, ethnic background, etc., as well as the writer's mood while writing), is a relatively difficult research field in OCR.

In the area of character recognition, the concept of combining multiple classifiers is proposed as a new direction for the development of highly reliable character recognition systems (C.Y.Suen et al., 1990) and some preliminary results have indicated that the

combination of several complementary classifiers will improve the performance of individual classifiers (C.Y.Suen et al., 1990 and T.K.Ho et al., 1990). The primary objective of this paper is ensemble of radial basis function and Support Vector Machine is superior to individual approach for recognizing totally unconstrained handwritten numerals in terms of classification accuracy.

The rest of this paper is organized as follows: Section 2 describes the related work. Section 3 presents hybrid intelligent system and Section 4 explains the performance evaluation measures. Section 5 focuses on the experimental results and discussion. Finally, results are summarized and concluded in section 6.

II. Related Work

2.1 Intrusion Detection

The Internet and online procedures is an essential tool of our daily life today. They have been used as an important component of business operation (T. Shon and J. Moon, 2007). Therefore, network security needs to be carefully concerned to provide secure information channels. Intrusion detection (ID) is a major research problem in network security, where the concept of ID was proposed by Anderson in 1980 (J.P. Anderson, 1980). ID is based on the assumption that the behavior of intruders is different from a legal user (W. Stallings, 2006). The goal of intrusion detection systems (IDS) is to identify unusual access or attacks to secure internal networks (C. Tsai , et al., 2009) Network-based IDS is a valuable tool for the defense-in-depth of computer networks. It looks for known or potential malicious activities in network traffic and raises an alarm whenever a suspicious activity is detected. In general, IDSs can be divided into two techniques: misuse detection and anomaly detection (E. Biermann et al.,2001; T. Verwoerd, et al., 2002)

Misuse intrusion detection (signature-based detection) uses well-defined patterns of the malicious activity to identify intrusions (K. Ilgun et al., 1995; D. Marchette, 1999) However, it may not be able to alert the system administrator in case of a new attack. Anomaly detection attempts to model normal behavior profile. It identifies malicious traffic based on the deviations from the normal patterns, where the normal patterns are constructed from the statistical measures of the system features (S. Mukkamala, et al., 2002). The anomaly detection techniques have the advantage of detecting unknown attacks over the misuse detection technique (E. Lundin and E. Jonsson, 2002). Several machine learning techniques including neural networks, fuzzy logic (S. Wu and W. Banzhaf, 2010), support vector machines (SVM) (S. Mukkamala, et al., 2002; S. Wu and W. Banzhaf, 2010) have been studied for the design of IDS. In particular, these techniques are developed as classifiers, which are used to classify whether the incoming network traffics are normal or an attack. This paper focuses on the Support Vector Machine (SVM) and Radial Basis Function (RBF) among various machine learning algorithms.

The most significant reason for the choice of SVM is because it can be used for either supervised or unsupervised learning. Another positive aspect of SVM is that it is useful for finding a global minimum of the actual risk using structural risk minimization, since it can generalize well with kernel tricks even in high-dimensional spaces under little training sample conditions.

In Ghosh and Schwartzbard (1999), it is shown how neural networks can be employed for the anomaly and misuse detection. The works present an application of neural network to learn previous behavior since it can be utilized to detection of the future intrusions against systems. Experimental results indicate that neural networks are ‘‘suited to perform intrusion state of art detection and can generalize from previously observed behavior’’ according to the authors.

2.2 Direct Marketing

Direct marketing aims at obtaining and maintaining direct relations between suppliers and buyers within one or more product/market combinations. In marketing, there are two main different approaches to communication: mass marketing and direct marketing (Ling, C. X., & Li, C, 1998). Mass marketing uses mass media such as print, radio and television to the public without discrimination. While direct marketing involves the identification of customers having potential market value by studying the customers’ characteristics and the needs (the past or the future) and selects certain customers to promote. Direct marketing becomes increasingly popular because of the increased competition and the cost problem. It is an important area of applications for data mining, data warehousing, statistical pattern recognition, and artificial intelligence. In direct marketing, models (profiles) are generated to select potential customers (from the client database) for a given product by analyzing data from similar campaigns, or by organizing test mail campaigns (Setnes, M., & Kaymak, U, 2001). Various classifiers have been employed such as logistic regression, neural networks and support vector machine.

Aristides Gionis et al have shown that the numbers of clusters discovered by their algorithms seem to be very reasonable choices: for the Votes dataset most people vote according to the official position of their political parties, so having two clusters is natural; for the Mushrooms dataset, notice that both ROCK and LIMBO achieve much better. (Aristides Gionis et al., 2005). Many aspects of churn have been modeled in the literature. First, whether churn is hidden or observable influence the overall approach to modeling. In some industries, customer defection is not directly observed, as customers do not explicitly terminate a relationship,

but can become inactive. In other industries, however, the defection decision is observable as customers cease their relationship via actively terminating their contract with the firm (Fader, P. S., et al., 2004).

The modeling approach could also depend critically on the relative importance placed on explanation/interpretation *vis a vis* prediction. Models that are better at explanation may not necessarily be better at prediction. The empirical literature in marketing has traditionally favored parametric models (such as logistic or probit regression or parametric hazard specifications and zero-inflated poisson models) that are easy to interpret. Similar to the previous discussion on acquisition, churn is a rare event that may require new approaches from data mining, machine learning and non-parametric statistics that emphasize predictive ability (Hastie, T., et al., 2001). These include projection-pursuit models, jump diffusion models, neural network models, tree structured models, spline-based models such as Generalized Additive Models (GAM), and Multivariate Adaptive Regression Splines (MARS), and more recently approaches such as support vector machines and boosting (Lemmens, et al., 2003).

Tang applied feed forward neural network to maximize performance at desired mailing depth in direct marketing in cellular phone industry. He showed that neural networks show more balance outcome than statistical models such as logistic regression and least squares regression, in terms of potential revenue and churn likelihood of a customer (Tang, Z, 2011).

2.3 Signature Verification

In the past several decades, a wide variety of approaches have been proposed to attempt to achieve the recognition system of handwritten numerals. These approaches generally fall into two categories: statistical method and syntactic method (C. Y. Suen, et al., 1992). First category includes techniques such as template matching, measurements of density of points, moments, characteristic loci, and mathematical transforms. In the second category, efforts are aimed at capturing the essential shape features of numerals, generally from their skeletons or contours. Such features include loops, endpoints, junctions, arcs, concavities and convexities, and strokes.

Suen et al.,(1992) proposed four experts for the recognition of handwritten digits. In expert one, the skeleton of a character pattern was decomposed into branches. The pattern was then classified according to the features extracted from these branches. In expert two, a fast algorithm based on decision trees was used to process the more easily recognizable samples, and a relaxation process was applied to those samples that could not be uniquely classified in the first phase. In expert three, statistical data on the frequency of occurrence of features during training were stored in a database. This database was used to deduce the identification of an unknown sample. In expert four, structural features were extracted from the contours of the digits. A tree classifier was used for classification. The resulting multiple-expert system proved that the consensus of these methods tended to compensate for individual weakness, while preserving individual strengths. The high recognition rates were reported and compared favorably with the best performance in the field.

The utilization of the Support Vector Machine (SVM) classifier has gained immense popularity in the past years (C. J. C. Burges., et al., 1997 and U. Krebel, 1999). SVM is a discriminative classifier based on Vapnik's structural risk minimization principle. It can be implemented on flexible decision boundaries in high dimensional feature spaces. Generally, SVM solves a binary (two-class) classification problem, and multi-class classification is accomplished by combining multiple binary SVMs. Good results on handwritten numeral recognition by using SVMs can be found in Dong, et al.'s paper.

Artificial Neural Networks (ANN), due to its useful properties such as: highly parallel mechanism, excellent fault tolerance, adaptation, and self-learning, have become increasingly developed and successfully used in character recognition (A. Amin, et al., 1996 and J. Cai, et al., 1995). The key power provided by such networks is that they admit fairly simple algorithms where the form of nonlinearity that can be learned from the training data. The models are thus extremely powerful, have nice theoretical properties, and apply well to a vast array of real-world applications.

2.4 Arcing Classifier

Freund and Schapire (1995,1996) propose an algorithm the basis of which is to **a**daptively **r**esample and **c**ombine (hence the acronym--arcing) so that the weights in the resampling are increased for those cases most often misclassified and the combining is done by weighted voting.

Previous work has demonstrated that arcing classifiers is very effective for RBF-SVM hybrid system. (M.Govindarajan et al., 2012).

Xu et al. (1992) proposed four combining classifier approaches according to the levels of information available from the various classifiers. The experimental results showed that the performance of individual classifiers could be improved significantly. Huang and Suen (1993, 1995) proposed the Behavior-Knowledge Space method in order to combine multiple classifiers for providing abstract level information for the recognition of handwritten numerals. Lam and Suen (1995) studied the performance of combination methods that were variations of the majority vote. A Bayesian formulation and a weighted majority vote (with weights obtained through a genetic algorithm) were

implemented, and the combined performances of seven classifiers on a large set of handwritten numerals were analyzed.

III. Hybrid Intelligent System

This section shows the proposed RBF-SVM hybrid system which involves Radial Basis Function (RBF) and Support Vector Machine (SVM) as base classifiers.

3.1 RBF-SVM Hybrid System

The proposed hybrid intelligent system is composed of three main phases; pre-processing phase, classification phase and Combining Phase.

3.1.1 Dataset Pre-processing

Before performing any classification method the data has to be pre-processed. In the data pre-processing stage it has been observed that the datasets consist of many missing value attributes. By eliminating the missing attribute records may lead to misclassification because the dropped records may contain some useful pattern for Classification. The dataset is pre-processed by removing missing values using supervised filters.

3.1.2 Existing Classification Methods

a) Radial basis Function Neural Network

The RBF (Oliver Buchtala, et al., 2005) design involves deciding on their centers and the sharpness (standard deviation) of their Gaussians. Generally, the centres 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. 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 (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).
- ✓ 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.

b) Support Vector Machine

The support vector machine (SVM) is a recently developed technique for multi dimensional function approximation. The objective of support vector machines is to determine a classifier or regression function which minimizes the empirical risk (that is the training set error) and the confidence interval (which corresponds to the generalization or test set error) (Vapnik, V, 1998).

Given a set of N linearly separable training examples $S = \left\{ x_i \in R^n \,|\, i = 1, 2, ..., N \right\}$, where each example belongs to one of the two classes, represented by $y_i \in \left\{ \div 1, -1 \right\}$, the SVM learning method seeks the optimal hyperplane **w . x** +b = 0, as the decision surface, which separates the positive and negative examples with the largest margins. The decision function for classifying linearly separable data is:

$$f(X) = \text{sign}\left(W.X + b\right) \tag{3.1}$$

Where **w** and b are found from the training set by solving a constrained quadratic optimization problem. The final decision function is

$$f(x) = sign\left(\sum_{i=1}^{N} a_i y_i (x_i..x) + b \right) \tag{3.2}$$

The function depends on the training examples for which a_i s is non-zero. These examples are called support vectors. Often the number of support vectors is only a small fraction of the original data set. The basic SVM formulation can be extended to the non linear case by using the nonlinear kernels that maps the input space to a high dimensional feature space. In this high dimensional feature space, linear classification can be performed. The SVM classifier has become very popular due to its high performances in practical applications such as text classification and pattern recognition.

The support vector regression differs from SVM used in classification problem by introducing an alternative loss function that is modified to include a distance measure. Moreover, the parameters that control the regression quality are the cost of error C, the width of tube ε and the mapping function ϕ. In this research work, the values for polynomial degree will be in the range of 0 to 5. In this work, best kernel to make the prediction is polynomial kernel with epsilon = 1.0E-12, parameter d=4 and parameter c=1.0.

A hybrid scheme based on coupling two base classifiers using arcing classifier adapted to data mining problem is defined in order to get better results. The main originality of proposed approach relies on associating two techniques: extracting more information bits via specific linguistic techniques, space reduction mechanisms, and moreover a arcing classifier to aggregate the best classification results.

c) Proposed RBF-SVM Hybrid System

Given a set D, of d tuples, arcing (Breiman. L, 1996) works as follows; For iteration i (i =1, 2,…..k), a training set, D_i, of d tuples is sampled with replacement from the original set of tuples, D. some of the examples from the dataset *D* will occur more than once in the training dataset D_i. The examples that did not make it into the training dataset end up forming the test dataset. Then a classifier model, M_i, is learned for each training examples *d* from training dataset D_i. A classifier model, M_i, is learned for each training set, D_i. To classify an unknown tuple, X, each classifier, M_i, returns its class prediction, which counts as one vote. The hybrid classifier (RBF-SVM), M^*, counts the votes and assigns the class with the most votes to X.

Algorithm: Hybrid RBF-SVM using Arcing Classifier

Input:

- *D, a set of d tuples.*
- *k = 2, the number of models in the ensemble.*
- *Base Classifiers (Radial Basis Function, Support Vector Machine)*

Output: Hybrid RBF-SVM model, M^*.

Procedure:

1. *For i = 1 to k do // Create k models*
2. *Create a new training dataset, Di, by sampling D with replacement. Same example from given dataset D may occur more than once in the training dataset Di.*
3. *Use Di to derive a model, Mi*
4. *Classify each example d in training data Di and initialized the weight, Wi for the model, Mi, based on the accuracies of percentage of correctly classified example in training data Di.*
5. *endfor*

To use the hybrid model on a tuple, X:

1. if classification then
2. let each of the k models classify X and return the majority vote;
3. if prediction then
4. let each of the k models predict a value for X and return the average predicted value;

IV. Performance Evaluation Measures

4.1 Cross Validation Technique

Cross-validation (Jiawei Han and Micheline Kamber, 2003) sometimes called rotation estimation, is a technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. 10-fold cross validation is commonly used. In stratified K-fold cross-validation, the folds are selected so that the mean response value is approximately equal in all the folds.

4.2 Criteria for Evaluation

The primary metric for evaluating classifier performance is classification Accuracy: the percentage of test samples that are correctly classified. The accuracy of a classifier refers to the ability of a given classifier to correctly predict the label of new or previously unseen data (i.e. tuples without class label information). Similarly, the accuracy of a predictor refers to how well a given predictor can guess the value of the predicted attribute for new or previously unseen data.

V. Experimental Results and Discussion

5.1 Dataset Description

5.1.1 Intrusion Detection

a) Real Dataset Description

The Acer07 dataset, being released for the first time is a real world data set collected from one of the sensors in Acer eDC (Acer e-Enabling Data Center). The data used for evaluation is the inside packets from August 31, 2007 to September 7, 2007.

b) Benchmark Dataset Description

The data used in classification is NSL-KDD, which is a new dataset for the evaluation of researches in network intrusion detection system. NSL-KDD consists of selected records of the complete KDD'99 dataset (Ira Cohen, et al., 2007). NSL-KDD dataset solve the issues of KDD'99 benchmark [KDD'99 dataset]. Each NSL-KDD connection record contains 41 features (e.g., protocol type, service, and ag) and is labeled as either normal or an attack, with one specific attack type.

5.1.2 Direct Marketing

a) Real Dataset Description

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be (or not) subscribed. The classification goal is to predict if the client will subscribe a term deposit (variable y).

b) Benchmark Dataset Description

The data includes all collective agreements reached in the business and personal services sector for locals with at least 500 members (teachers, nurses, university staff, police, etc) in Canada in 87 and first quarter of 88. Data was used to test 2 tier approach with learning from positive and negative examples.

5.1.3 Signature Verification

a) Real Dataset Description

The dataset used to train and test the systems described in this paper was constructed from NIST's Special Database 3 and Special Database 1 which contain binary images of handwritten digits. NIST originally designated SD-3 as their training set and SD-1 as their test set. However, SD-3 is much cleaner and easier to recognize than SD-1. The reason for this can be found on the fact that SD-3 was collected among Census Bureau employees, while SD-1 was collected among high-school students. Drawing sensible conclusions from learning experiments requires that the result be independent of the choice of training set and test among the complete set of samples. Therefore it was necessary to build a new database by mixing NIST's datasets.

b) Benchmark Dataset Description

The data used in classification is 10 % U.S. Zip code, which consists of selected records of the complete U.S. Zip code database. The database used to train and test the hybrid system consists of 4253 segmented numerals digitized from handwritten zip codes that appeared on U.S. mail passing through the Buffalo, NY post office. The digits were written by many different people, using a great variety of sizes, writing styles, and instruments, with widely varying amounts of care.

5.2 Experiments and Analysis

5.2.1 Intrusion Detection

a) Real Dataset

The Acer07dataset is taken to evaluate the proposed hybrid RBF-SVM for intrusion detection system.

Table 1: The Performance of Existing and Proposed Hybrid RBF-SVM Classifier for real world dataset

Real Dataset	Classifiers	Classification Accuracy
Acer07dataset	RBF	99.40%
	SVM	99.60%
	Proposed Hybrid RBF-SVM	99.90%

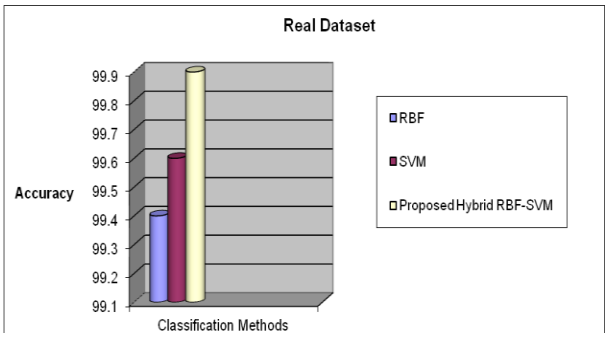

Fig. 1: Classification Accuracy of Base and Proposed Hybrid RBF-SVM Classifiers Using Real Dataset

b) Benchmark Dataset

The NSL- KDD dataset is taken to evaluate the proposed hybrid RBF-SVM for intrusion detection system.

Table 2: The Performance of Existing and Proposed Hybrid RBF-SVM Classifier for bench mark dataset

Benchmark Dataset	Classifiers	Classification Accuracy
NSL- KDD dataset	RBF	84.74%
	SVM	91.81%
	Proposed Hybrid RBF-SVM	98.46%

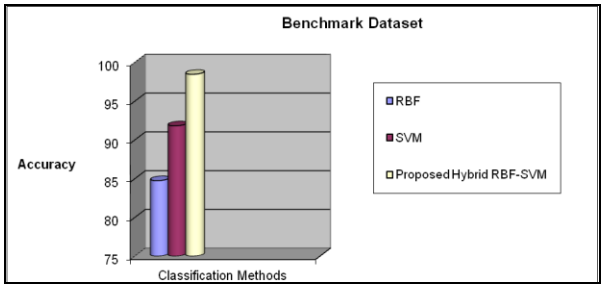

Fig. 2: Classification Accuracy of Base and Proposed Hybrid RBF-SVM Classifiers Using Benchmark Dataset

5.2.2 Direct Marketing

In this section, new ensemble classification method is proposed using arcing classifier and its performance is analyzed in terms of accuracy.

a) Real Dataset

Table 3: The Performance of Existing and Proposed Hybrid RBF-SVM Classifier for real world dataset

Real Dataset	Classifiers	Classification Accuracy
Bank Marketing dataset	RBF	71.16%
	SVM	69.00%
	Proposed Hybrid RBF-SVM	88.33%

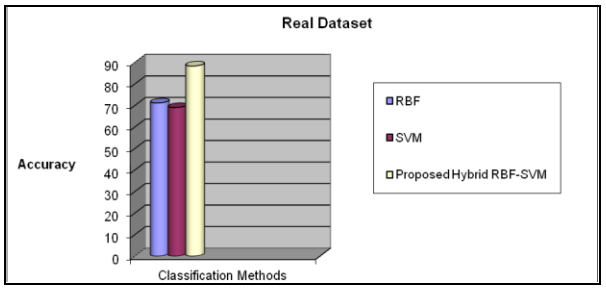

Fig. 3: Classification Accuracy of Base and Proposed Hybrid RBF-SVM Classifiers Using Real Dataset

b) Benchmark Dataset

The bank marketing dataset is taken to evaluate the proposed hybrid RBF-SVM classifier.

Table 4: The Performance of Existing and Proposed Hybrid RBF-SVM Classifier for bench mark dataset

Benchmark Dataset	Classifiers	Classification Accuracy
Labor dataset	RBF	94.73%
	SVM	89.47%
	Proposed Hybrid RBF-SVM	98.24%

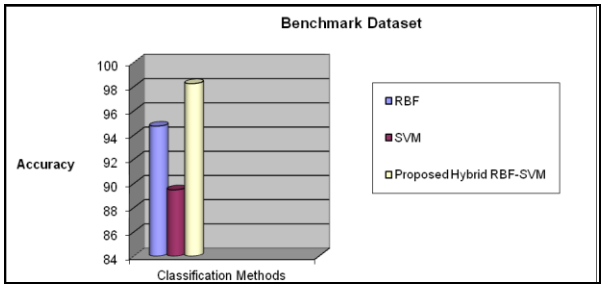

Fig. 4: Classification Accuracy of Base and Proposed Hybrid RBF-SVM Classifiers Using Benchmark Dataset

5.2.3 Signature Verification

a) Real Dataset

The NIST dataset are taken to evaluate the proposed hybrid RBF-SVM for handwriting recognition system.

Table 5: The Performance of Existing and Proposed Hybrid RBF-SVM Classifier for real world dataset

Real Dataset	Classifiers	Classification Accuracy
NIST dataset	RBF	76.50%
	SVM	89.20%
	Proposed Hybrid RBF-SVM	99.30%

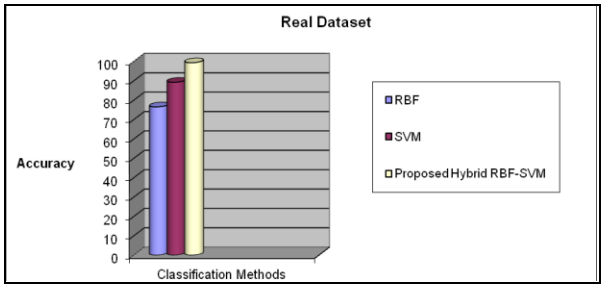

Fig. 5: Classification Accuracy of Base and Proposed Hybrid RBF-SVM Classifiers Using Real Dataset

b) Benchmark Dataset

The U.S. Zip code dataset are taken to evaluate the proposed hybrid RBF-SVM for handwriting recognition system.

Table 6: The Performance of Existing and Proposed Hybrid RBF-SVM Classifier for bench mark dataset

Benchmark Dataset	Classifiers	Classification Accuracy
U.S. Zip code dataset	RBF	86.46%
	SVM	93.98%
	Proposed Hybrid RBF-SVM	99.13%

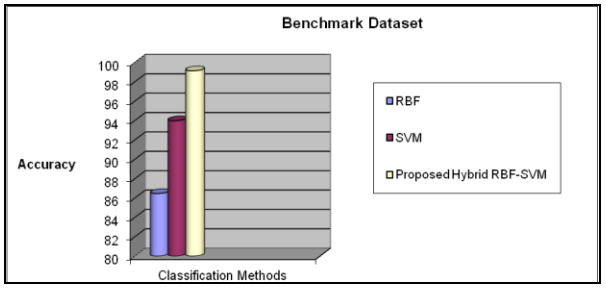

Fig. 6: Classification Accuracy of Base and Proposed Hybrid RBF-SVM Classifiers Using Benchmark Dataset

The data set described in section 5 is being used to test the performance of base classifiers and hybrid classifier. Classification accuracy was evaluated using 10-fold cross validation. In the proposed approach, first the base classifiers RBF and SVM are constructed individually to obtain a very good generalization performance. Secondly, the ensemble of RBF and SVM is designed. In the ensemble approach, the final output is decided as follows: base classifier's output is given a weight (0–1 scale) depending on the generalization performance as given in Table 1 to 6. According to Table 1 to 6, the proposed hybrid model shows significantly larger improvement of classification accuracy than the base classifiers and the results are found to be statistically significant.

The x^2 statistic x2 is determined for all the above approaches and their critical value is found to be less than 0.455. Hence corresponding probability is $p < 0.5$. This is smaller than the conventionally accepted significance level of 0.05 or 5%. Thus examining a x^2 significance table, it is found that this value is significant with a degree of freedom of 1. In general, the result of x2 statistic analysis shows that the proposed classifiers are significant at $p < 0.05$ than the existing classifiers.

The proposed ensemble of RBF and SVM is shown to be superior to individual approaches for data mining applications like intrusion detection, direct marketing, and signature verification in terms of Classification accuracy

VI. Conclusion

In this research, some new techniques have been investigated for data mining applications and evaluated their performance based on classification accuracy. RBF and SVM have been explored as hybrid models. Next a hybrid RBF-SVM model and RBF, SVM models as base classifiers are designed. Finally, hybrid systems are proposed to make optimum use of the best performances delivered by the individual base classifiers and the hybrid approach. The hybrid RBF-SVM shows higher percentage of classification accuracy than the base classifiers and enhances the testing time due to data dimensions reduction.

The experiment results lead to the following observations.

- SVM exhibits better performance than RBF in the important respects of accuracy for Intrusion detection and Signature Verification problems.
- RBF exhibits better performance than SVM in the important respects of accuracy for direct marketing problem.
- Comparison between the individual classifier and the combination classifier: it is clear that the combination classifiers show the significant improvement over the single classifiers for data mining applications.

Acknowledgment

Author gratefully acknowledges the authorities of Annamalai University for the facilities offered and encouragement to carry out this work.

References

[1] Aristides Gionis and Heikki Mannila and Panayiotis Tsaparas. Clustering Aggregation,ICDE, 2005.

[2] A. Amin, H. B. Al-Sadoun, and S. Fischer. Hand-printed Arabic Character Recognition System Using An Artificial Network, Pattern Recognition, 29(4), 1996, pp. 663-675.

[3] J.P. Anderson, Computer security threat monitoring and surveillance, Technical Report, James P. Anderson Co., Fort Washington, PA, 1980.

[4] C. L. Bauer. A direct mail customer purchase model, Journal of Direct Marketing, 2, 1998, pp.16–24.

[5] E. Biermann, E. Cloete and L.M. Venter. A comparison of intrusion detection Systems, Computer and Security, 20, 2001,pp.676-683.

[6] Breiman. L. Bias, Variance, and Arcing Classifiers, Technical Report 460, Department of Statistics, University of California, Berkeley, CA, 1996.

[7] C. J. C. Burges and B. Scholkopf. Improving the Accuracy and Speed of Support vector Learning Machine, Advanced in Neural Information Processing Systems 9, MIT Press, Cambridge, MA, 1997, pp. 375-381.

[8] J. Cai, M. Ahmadi, and M. Shridhar. Recognition of Handwritten Numerals with Multiple Feature and Multi-stage Classifier, Pattern Recognition, 28(2), 1995, pp. 153-160.

[9] Fader, P. S., B. G. S. Hardie, and K. L. Lee. Counting Your Customers' the Easy Way: An

Alternative to the Pareto/NBD Model, Working Paper, Wharton Marketing Department, 2004.

[10] Freund, Y. and Schapire, R. A decision-theoretic generalization of on-line learning and an application to boosting. In proceedings of the Second European Conference on Computational Learning Theory, 1995, pp. 23-37.

[11] Freund, Y. and Schapire, R. Experiments with a new boosting algorithm, In Proceedings of the Thirteenth International Conference on Machine Learning, 1996, pp.148-156 Bari, Italy.

[12] Ghosh AK, Schwartzbard A. A study in using neural networks for anomaly and misuse detection. In: The proceeding on the 8th USENIX security symposium, <http://citeseer.ist.psu.edu/context/1170861/0>; 1999, [accessed August 2006].

[13] Gonul, F. F., Kim, B. D., & Shi, M. Mailing smarter to catalog customers. Journal of Interactive Marketing, 14(2), 2000, pp.2–16.

[14] M.Govindarajan, RM.Chandrasekaran. Intrusion Detection using an Ensemble of Classification Methods, In Proceedings of International Conference on Machine Learning and Data Analysis, San Francisco, U.S.A, 2012, pp.459-464.

[15] Gupta, Sunil, Donald R. Lehmann, and Jennifer Ames Stuart. Valuing Customers," Journal of Marketing Research, 41(1), 2004, pp.7–18.

[16] Hastie, T., R. Tibshirani, and J. Friedman. The Elements of Statistical Learning: Data Mining, Inference and Prediction. New York: Springer-Verlag, 2001

[17] Heady R, Luger G, Maccabe A, Servilla M. The architecture of a network level intrusion detection system. Technical Report, Department of Computer Science, University of New Mexico, 1990.

[18] T.K.Ho, J.J.Hull, and S.N.Srihari, Combination of Structural Classifiers, in Proc. IAPR Workshop Syntatic and Structural Pattern Recog., 1990, pp.123-137.

[19] Y. S. Huang and C. Y. Suen. An Optimal Method of Combining Multiple Classifiers for Unconstrained Handwritten Numeral Recognition, Proceedings of 3rd International Workshop on Frontiers in Handwriting Recognition, 1993.

[20] Y. S. Huang and C. Y. Suen. A Method of Combining Experts for the Recognition of Unconstrained Handwritten Numerals, IEEE Transactions on PAMI, 17(1), 1995, pp.90-94.

[21] K. Ilgun, R.A. Kemmerer and P.A. Porras. State transition analysis:A rule-based intrusion detection approach, IEEE Trans. Software Eng, 21, 1995, pp.181-199.

[22] Ira Cohen, Qi Tian, Xiang Sean Zhou and Thoms S.Huang. Feature Selection Using Principal Feature Analysis, In Proceedings of the 15th international conference on Multimedia, Augsburg, Germany, September, 2007, pp.25-29.

[23] Jiawei Han , Micheline Kamber. Data Mining – Concepts and Techniques, Elsevier Publications, 2003

[24] KDD'99 dataset, http://kdd.ics.uci.edu/databases, Irvine, CA, USA, 2010.

[25] U. Krebel. Pairwise Classification and Support Vector Machines, Advances in Kernel Methods: Support Vector Learning, MIT Press, Cambridge, MA, 1999, pp.255-268.

[26] Kubat, M., Holte, R., & Matwin, S. Learning when negative examples abound. Lecture Notes in Artificial Intelligence (LNAI 1224), 1997, pp.146–153, Prague, The Czech Republic

[27] L. Lam and C. Y. Suen. Optimal Combinations of Pattern Classifiers, Pattern Recognition Letters, 16(9), 1995, pp.945-954.

[28] Lemmens, Aurélie and Christophe Croux. Bagging and Boosting Classification Trees to Predict Churn, Working Paper, Teradata center, 2003.

[29] Ling, C. X., & Li, C. Data mining for direct marketing: Problems and solutions, Proceedings of the KDD98, 1998, pp.73–79.

[30] E. Lundin and E. Jonsson. Anomaly-based intrusion detection: privacy concerns and other problems", Computer Networks,34, 2002,pp.623-640.

[31] Maryam Daneshmandi, Marzieh Ahmadzadeh. A Hybrid Data Mining Model to Improve Customer Response Modeling in Direct Marketing, Indian Journal of Computer Science and Engineering, 3(6), 2013,pp.844-855.

[32] D. Marchette. A statistical method for profiling network traffic". In proceedings of the First USENIX Workshop on Intrusion Detection and Network Monitoring (Santa Clara), CA, 1999,pp.119-128.

[33] Michie, D., Spiegelhalter, D. J., & Taylor, C. Machine learning. Neural and statistical classification, Ellis Horwood, 1994.

[34] Mukkamala S, Sung AH, Abraham A. Intrusion detection using ensemble of soft computing paradigms, third international conference on intelligent systems design and applications, intelligent systems design and applications, advances in soft computing. Germany: Springer; 2003, pp.239–48.

[35] Mukkamala S, Sung AH, Abraham A. Modeling intrusion detection systems using linear genetic

programming approach, The 17th international conference on industrial & engineering applications of artificial intelligence and expert systems, innovations in applied artificial intelligence. In: Robert O., Chunsheng Y., Moonis A., editors. Lecture Notes in Computer Science, vol. 3029. Germany: Springer; 2004a, pp.633–42.

[36] Mukkamala S, Sung AH, Abraham A, Ramos V. (2004b), Intrusion detection systems using adaptive regression splines. In: Seruca I, Filipe J, Hammoudi S, Cordeiro J, editors. Proceedings of the 6th international conference on enterprise information systems, ICEIS'04, vol. 3, Portugal, 2004b,pp.26–33

[37] S. Mukkamala, G. Janoski and A.Sung. Intrusion detection: support vector machines and neural networks, In proceedings of the IEEE International Joint Conference on Neural Networks (ANNIE), St. Louis, MO, 2002, pp.1702-1707.

[38] Oliver Buchtala, Manuel Klimek, and Bernhard Sick, Member, IEEE. Evolutionary Optimization of Radial Basis Function Classifiers for Data Mining Applications, IEEE Transactions on systems, man, and cybernetics—part b: cybernetics, 35(5), 2005.

[39] Sara Madeira Joao M.Sousa. Comparison of target selection methods in direct Marketing, Technical University of Lisbon, Institution Superior Technician, Dept. Mechanical Eng./IDMEC, 1049-001 Lisbon, Portugal, 2000.

[40] Setnes, M., & Kaymak, U. Fuzzy modeling of client preference from large data sets: an application to target selection in direct marketing. IEEE Transactions on Fuzzy Systems, 9(1), 2001, pp.153–163.

[41] Shah K, Dave N, Chavan S, Mukherjee S, Abraham A, Sanyal S. Adaptive neuro-fuzzy intrusion detection system, IEEE International Conference on Information Technology: Coding and Computing (ITCC'04), 1, USA: IEEE Computer Society; 2004, pp.70–74.

[42] Shin, H. J., & Cho, S. Response modeling with support vector machines. Expert Systems with Applications, 30(4), 2006, pp.746–760.

[43] T. Shon and J. Moon. A hybrid machine learning approach to network anomaly detection, Information Sciences, 177, 2007, pp.3799-3821.

[44] C.Y.Suen, C.Nadal, T.A.Mai, R.Legault, and L.Lam, Recognition of totally unconstrained handwritten numerals based on the concept of multiple experts, Frontiers in Handwriting Recognition , C.Y.Suen, Ed., IN Proc.Int.Workshop on Frontiers in Handwriting Recognition, Montreal, Canada, Apr. 2-3, 1990, pp.131-143.

[45] C. Y. Suen, C. Nadal, R. Legault, T. A. Mai, and L. Lam. Computer recognition of unconstrained handwritten numerals, Proc. IEEE, 80, 1992, pp.1162–1180.

[46] Summers RC. Secure computing: threats and safeguards. New York: McGraw-Hill, 1997.

[47] Sundaram A. An introduction to intrusion detection. ACM Cross Roads; 2(4), 1996.

[48] W. Stallings. Cryptography and network security principles and practices, USA: Prentice Hall, 2006

[49] Tang, Z. Improving Direct Marketing Profitability with Neural Networks. International Journal of Computer Applications, 29(5), 2011,pp.13-18.

[50] C. Tsai , Y. Hsu, C. Lin and W. Lin. Intrusion detection by machine learning: A review, Expert Systems with Applications, 36, 2009,pp.11994-12000.

[51] Vapnik, V. Statistical learning theory, New York, John Wiley & Sons, 1998.

[52] T. Verwoerd and R. Hunt. Intrusion detection techniques and approaches, Computer Communications, 25, 2002,pp.1356-1365.

[53] S. Wu and W. Banzhaf. The use of computational intelligence in intrusion detection systems: A review, Applied Soft Computing, 10, 2010, pp.1-35.

[54] L. Xu, A. Krzyzak, and C. Y. Suen. Methods of Combining Multiple Classifiers and Their Applications to Handwritten Recognition, IEEE Transactions on Systems, Man, Cybernetics, 22(3), 1992, pp.418-435.

[55] Yu, E., & Cho, S. Constructing response model using ensemble based on feature subset selection. Expert Systems with Applications, 30(2), 2006, pp.352–360.

[56] Zahavi, J., & Levin, N. Issues and problems in applying neural computing to target marketing. Journal of Direct Marketing, 11(4), 1997, pp.63–75.

Author's Profiles

M.Govindarajan received the B.E and M.E and Ph.D Degree in Computer Science and Engineering from Annamalai University, Tamil Nadu, India in 2001 and 2005 and 2010 respectively. He did his post-doctoral research in the Department of Computing, Faculty of Engineering and Physical Sciences, University of Surrey, Guildford, Surrey, United Kingdom in 2011 and pursuing Doctor of Science at Utkal University, orissa, India. He is currently an Assistant Professor at the

Department of Computer Science and Engineering, Annamalai University, Tamil Nadu, India. He has presented and published more than 75 papers at Conferences and Journals and also received best paper awards. He has delivered invited talks at various national and international conferences. His current Research Interests include Data Mining and its applications, Web Mining, Text Mining, and Sentiment 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, Career Award for Young Teachers (2006), All India Council for Technical Education, New Delhi, India and Young Scientist International Travel Award (2012), Department of Science and Technology, Government of India New Delhi. He is Young Scientists awardee under Fast Track Scheme (2013), Department of Science and Technology, Government of India, New Delhi and also granted Young Scientist Fellowship (2013)**,** Tamil Nadu State Council for Science and Technology, Government of Tamil Nadu, Chennai. He has visited countries like Czech Republic, Austria, Thailand, United Kingdom, Malaysia, U.S.A, and Singapore. He is an active Member of various professional bodies and Editorial Board Member of various conferences and journals.

How to cite this paper: M.Govindarajan,"A Hybrid RBF-SVM Ensemble Approach for Data Mining Applications", International Journal of Intelligent Systems and Applications(IJISA), vol.6, no.3, pp.84-95, 2014. DOI: 10.5815/ijisa.2014.03.09

