Ensembles of Classification Methods for Data Mining Applications

M. Govindarajan
Assistant Professor, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar – 608002, Tamil Nadu, India.
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. In this research work, new ensemble classification methods are proposed using classifiers in both homogeneous ensemble classifiers using bagging and heterogeneous ensemble classifiers using arcing classifier and their performances are analyzed in terms of accuracy. A Classifier ensemble is designed using Radial Basis Function (RBF) and Support Vector Machine (SVM) as base classifiers. The feasibility and the benefits of the proposed approaches are demonstrated by the means of real and benchmark data sets of data mining applications like intrusion detection, direct marketing and signature verification. The main originality of the proposed approach is based on three main parts: preprocessing phase, classification phase and combining phase. A wide range of comparative experiments are conducted for real and benchmark data sets of direct marketing. The accuracy of base classifiers is compared with homogeneous and heterogeneous models for data mining problem. The proposed ensemble methods provide significant improvement of accuracy compared to individual Classifiers and also heterogeneous models exhibit better results than homogeneous models for real and benchmark data sets of data mining applications.

Index Terms — Data Mining, Ensemble, Intrusion Detection, Direct Marketing, Signature Verification, Radial Basis Function, Support Vector Machine, Accuracy.

1. Introduction

Data mining methods may be distinguished by either supervised or unsupervised learning methods. One of the most active areas of research in supervised learning has been to study methods for constructing good ensembles of classifiers. It has been observed that when certain classifiers are ensemble, the performance of the individual classifiers.

Recently, advances in knowledge extraction techniques have made it possible to transform various kinds of raw data into high level knowledge. However, the classification results of these techniques are affected by the limitations associated with individual techniques. Hence, hybrid approach is widely recognized by the data mining research community. 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 (Haykin, 1999).

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. 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

In general, businesses worldwide use mass marketing as their marketing strategy for offering and promoting a new product or service to their customers. The idea of mass marketing is to broadcast a single communication message to all customers so that maximum exposure is ensured. However; since this approach neglects the difference among customers it has several drawbacks. In fact a single product offering cannot fully satisfy different needs of all customers in a market and unsatisfied customers with unsatisfied needs expose businesses to challenges by competitors who are able to identify and fulfill the diverse needs of their customers more accurately. Thus in today’s world where mass marketing has become less effective, businesses choose other approaches such as direct marketing as their main marketing strategy (Hossein Javaheri 2007).

Direct marketing is concerned with identifying which customers are more likely to respond to specific promotional offers. A response model predicts the probability that a customer is responsive/non-responsive to an offer for a product or service. A response modeling is usually the first type of target modeling that a business develops as its marketing strategy. If no marketing promotion has been done in the past, a response model can make the marketing campaign more efficient and might bring in more profit to the company by reducing mail expenses and absorbing more customers (Parr Rud 2001). Response model can be formulated in to a binary classification problem in which customers are divided in to two groups of respondents and non-respondents. Typically historical purchase data is used to model customer response. In direct marketing a desirable response model should contain more respondents and fewer non-respondents (Shin 2006).

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.

This paper proposes new ensemble classification methods to improve the classification accuracy. The main purpose of this paper is to apply homogeneous and heterogeneous ensemble classifiers for real and benchmark dataset of data mining applications to improve classification accuracy. Organization of this paper is as follows. Section 2 describes the related work. Section 3 presents proposed methodology 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.

2. 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

Various data mining techniques have been used to model customer response to catalogue advertising. Traditionally statistical methods such as discriminant analysis, least squares and logistic regression have been applied to response modeling.

Given the interest in this domain, there are several works that use DM to improve bank marketing campaigns (Ling and Li, 1998) (Hu, 2005) (Li et al, 2010). In particular, often these works use a classification DM approach, where the goal is to build a predictive model that can label a data item into one of several predefined classes (e.g. “yes”, “no”). Several DM algorithms can be used for classifying marketing contacts, each one with its own capabilities and possibilities. Examples of popular DM techniques are: Naive Bayes (NB) (Zhang, 2004), Decision Trees (DT) (Aptéa and Weiss, 1997) and Support Vector Machines (SVM) (Cortes and Vapnik, 1995).

Neural Networks have also been used in response modeling. Bounds and Ross showed that neural networks could improve the response rate from 2% up to 95% (Bounds 1997). Viaene et al have also used neural networks to select input variables in response modeling (Viaene, Baensens et al. 2001). 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 2011). Bentz and Merunkay also showed that neural networks did better than multinomial logistic regression (Bentz 2000).

To overcome the neural networks limitations, Shin and Cho applied Support Vector Machine (SVM) to response modeling. In their study, they introduced practical difficulties such as large training data and class imbalance problem when applying SVM to response modeling. They proposed a neighborhood property based pattern selection algorithm (NPPS) that reduces the training set without accuracy loss. For the other remaining problem they employed different misclassification costs to different class errors in the objective function (Shin 2006).

Although SVM is applied to a wide variety of application domains, there have been only a couple of SVM application reports in response modeling. Cheung, Kwok, Law, and Tsui (2003) used SVM for content-based recommender systems. The system is definitely a form of direct marketing that has emerged by virtue of recent advances in the World Wide Web, e-business, and on-line companies. They compared Naive Bayes, C4.5 and 1-nearest neighbor rule with SVM. The SVM yielded the best results among them. More specific, SVM application to response modeling was attempted by Viaene et al. (2001b).

Performance comparison of the methods has been one of the controversial issues in direct marketing domain. Suh, Noh, and Suh (1999) and Zahavi and Levin (1997a, 1997b) found that neural network did not outperform other statistical methods. They suggested combining the neural network response model and the statistical method. On the other hand, Bentz and Merunkay (2000) reported that neural networks outperformed multinomial logistic regression. Potharst, Kaymak, and Pijls (2001) applied neural networks to direct mailing campaigns of a large Dutch charity organization. According to their results, the performance of neural networks surpassed that of CHAID or logistic regression.

Ha, Cho, and MacLachlan (2005) proposed a response model using bagging neural networks. The experiments over a publicly available DMEF4 dataset showed that bagging neural networks give more improved and stabilized prediction accuracies than single neural networks and logistic regression.
Much of the previous work on ensembles of classifier models (Breiman, L. 2001) has focused on homogeneous ensemble classifiers – i.e., collections of classifier models of a single type. This work also focuses on heterogeneous ensemble classifiers, where the collection of classifiers are not of the same type. Note that such classifier models are also referred to as hybrid ensemble classifiers.

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 Daneshmandi et.al, 2013). A hybrid approach is built by combining two or more data mining techniques. A hybrid approach is commonly used to maximize the accuracy of a classifier. Coenen et al. proposed a hybrid approach with C5, a decision tree algorithm and case based reasoning (CBR). In their study, first cases were classified by means of C5 algorithm and then the classified cases were ranked by a CBR similarity measure. This way they succeeded to improve the rank of the classified cases (Coenen 2000). Chiu also proposed a CBR system based on Genetic Algorithm to classify potential customers in insurance direct marketing. The proposed GA approach determines the fittest weighting values to improve the case identification accuracy. The created model showed better learning and testing performance (Chiu 2002).

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.

RenataF. P. Neves et al (2011) have proposed SVM based offline handwritten digit recognition. Authors claim that SVM outperforms the Multilayer perceptron classifier. Experiment is carried out on NIST SD19 standard dataset. Advantage of MLP is that it is able to segment non-linearly separable classes. However, MLP can easily fall into a region of local minimum, where the training will stop assuming it has achieved an optimal point in the error surface. Another hindrance is defining the best network architecture to solve the problem, considering the number of layers and the number of perceptron in each hidden layer. Because of these disadvantages, a digit recognizer using the MLP structure may not produce the desired low error rate.

Muhammad et al (2012) have discussed hybrid feature extraction in their work. SVM is used as a classifier. Authors have combined structural, statistical and correlation functions to derive hybrid features. In first step, elementary stroke location is identified with the help of chosen elementary shape. To make it more robust, certain structural / statistical features are added in it. The added structural / statistical features are based on projections, profiles, invariant moments, endpoints and junction points. This enhanced, powerful combination of features results in a 157-variable feature vector for each character. It includes 100 correlation features and 57 structural/statistical features. Correlation features are based on Pearson’s correlation coefficient.

Shubhangi et al, (2009) have extract similar correlation function based features for Chinese hand-printed character recognition. Classification is done based on minimum distance decision rule. While proposed method perform final classification based on support vector machine (SVM).

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.
Malayalam is a language spoken by millions of people in the state of Kerala and the union territories of Lakshadweep and Pondicherry in India. It is written mostly in clockwise direction and consists of loops and curves. Neural network based approach is discussed in (Amritha Sampath et al, 2012) for Malayalam language. In pre processing step, noise is removed by applying threshold (number of pixels in rectangular bounding box).

Postal address recognition system for Arabic language is proposed by M.Charfi et al. (2012) Writing translates style of writing, Mood and personality of the writer, which makes it difficult to characterize. From scanned envelop, printed boarder and stamp logo are suppressed. Address is located and using histogram method, lines, words and characters are segmented. Temporal order of strokes can be helpful for robust recognition. In literature, way of temporal order reconstruction is proposed. End stroke point, Branching point and Crossing point are detected from city name. Elliptical model is applied on preprocessed digit or character and matching process is applied.

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.

2.4 Bagging Classifier

Breiman (1996c) showed that bagging is effective on “unstable” learning algorithms where small changes in the training set result in large changes in predictions. Breiman (1996c) claimed that neural networks and decision trees are example of unstable learning algorithms.

The boosting literature (Schapire, Freund, Bartlett, & Lee, 1997) has recently suggested (based on a few data sets with decision trees) that it is possible to further reduce the test-set error even after ten members have been added to an ensemble (and they note that this result also applies to bagging).

2.5 Arcing Classifier

Freund and Schapire (1995,1996) proposed an algorithm the basis of which is to adaptively resample and combine (hence the acronym--arcning) 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). A hybrid model can improve the performance of basic classifier (Tsai 2009).

In this paper, a hybrid direct marketing system is proposed using radial basis function and support vector machine and the effectiveness of the proposed bagged RBF, bagged SVM and RBF-SVM hybrid system is evaluated by conducting several experiments on real and benchmark datasets of data mining applications. The performance of the proposed bagged RBF, bagged SVM and RBF-SVM hybrid classifiers are examined in comparison with standalone RBF and standalone SVM classifier and also heterogeneous models exhibits better results than homogeneous models for real and benchmark data sets of data mining applications.

3 Proposed Methodology

3.1 Preprocessing for real and benchmark Datasets

Before performing any classification method the data has to be preprocessed. In the data preprocessing 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 preprocessed by removing missing values using supervised filters.

3.2 Existing Classification Methods

3.2.1 Radial Basis Function Neural Network

Radial basis function (RBF) networks (Oliver Buchtala et al, 2005) combine a number of different concepts from approximation theory, clustering, and neural network theory. A key advantage of RBF networks for practitioners is the clear and understandable interpretation of the functionality of basis functions. Also, fuzzy rules may be extracted from RBF networks for deployment in an expert system.

The RBF networks used here may be defined as follows.

1. RBF networks have three layers of nodes: input layer \( u_i \), hidden layer \( u_h \) and output layer \( u_o \).
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 \( w_{ij} \) is associated with the connection between nodes \( i \) and \( j \).
3. The activation of each input node (fanout) \( i \in u_i \) is equal to its external input.

\[
def \quad a_i(k) = x_i(k) \quad (3.1)
\]
where \( x_i(k) \) is the element of the external input vector of the network \( X(k) \) of the network \( k = 1, 2, \ldots \) (\( k \) denotes the number of the pattern).

4. Each hidden node (neuron) \( j \in \mathbb{H} \) determines the Euclidean distance between “its own” weight vector and the activations of the input nodes, i.e., the external input vector.

\[
s_j(k) = \|W_j - X(k)\| \tag{3.2}
\]

The distance \( s_j(k) \) is used as an input of a radial basis function in order to determine the activation \( a_j(k) \) of node \( j \). Here, Gaussian functions are employed.

\[
a_j(k) = e^{-s_j(k)^2 / r_j^2} \tag{3.3}
\]

The parameter \( r_j \) of node \( j \) is the radius of the basis function; the vector \( W_j \) is its center.

Localized basis functions such as the Gaussian or the inverse multiquadric are usually preferred.

5. Each output node (neuron) \( l \in \mathbb{O}_0 \) computes its activation as a weighted sum:

\[
a[l] = \sum_j [\eta[l] a_j(k)] \sum_i [\eta[l] a_i(k)] + \eta[B] \tag{3.4}
\]

The external output vector of the network, \( y(k) \), consists of the activations of output nodes, i.e.,

\[
y_i(k) = a_i(k). \tag{3.5}
\]

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.

### 3.2.2 Support Vector Machine

Support vector machines (Cherkassky et al., 1998; Burges, 1998) are powerful tools for data classification. Classification is achieved by a linear or nonlinear separating surface in the input space of the dataset. The separating surface depends only on a subset of the original data. This subset of data, which is all that is needed to generate the separating surface, constitutes the set of support vectors. In this study, a method is given for selecting as small a set of support vectors as possible which completely determines a separating plane classifier. In nonlinear classification problems, SVM tries to place a linear boundary between two different classes and adjust it in such a way that the margin is maximized (Vanajakshi and Rilett, 2004). Moreover, in the case of linearly separable data, the method is to find the most suitable one among the hyperplanes that minimize the training error. After that, the boundary is adjusted such that the distance between the boundary and the nearest data points in each class is maximal.

In a binary classification problem, its data points are given as:

\[
D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\}, \quad y = \{-1, 1\}.
\]

where

- \( y \) is a binary value representing the two classes and,
- \( x \) is the input vector.

As mentioned above, there are numbers of hyperplanes that can separate these two sets of data and the problem is to find the hyperplane with the largest margin. Suppose that all training data satisfy the following constraints:

\[
w \cdot x + b \geq +1 \quad \text{for} \quad y = +1
\]

\[
w \cdot x + b \leq -1 \quad \text{for} \quad y = -1
\]

where

- \( w \) is the boundary vector
- \( x \) is the input vector
- \( b \) is the scalar threshold (bias).

Therefore, the decision function that can classify the data is:

\[
f(y) = \text{sgn}((w \cdot x) + b)
\]

Thus, the separating hyperplane must satisfy the following constraints:

\[
y_i((w \cdot x_i) + b) \geq 1
\]

where \( l \) is the number of training sets.

The optimal hyperplane is the unique one that not only separates the data without error but also maximizes the margin. It means that it should maximize the distance between closest vectors in both classes to the hyperplane. Therefore the hyperplane that optimally separate the data into two classes can be shown to be the one that minimize the functional:

\[
\frac{1}{2} \|w\|^2 - C \sum_i \xi_i
\]

where \( C \) is a parameter that controls the trade-off between maximizing the margin and minimizing the classification error, \( \xi_i \) is the slack variable.
Algorithm: RBF and SVM ensemble classifiers using bagging

Input:
- D, a set of d tuples.
- k = 1, the number of models in the ensemble.
- Base Classifiers (Radial Basis Function, Support Vector Machine).

Output: Bagged RBF and SVM, M^*

Method:
1. for i = 1 to k do // create k models.
2. Create a bootstrap sample, Di, by sampling D with replacement, from the given training data set D repeatedly. Each example in the given training set D may appear repeated times or not at all in any particular replicate training data set 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 bagged RBF and SVM models 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;

3.4 Heterogeneous Ensemble Classifiers using Arcing

3.4.1 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, Di, 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 Di. The examples that did not make it into the training dataset end up forming the test dataset. Then a classifier model, Mi, is learned for each training examples d from training dataset Di. A classifier model, Mi, is learned for each training set, Di. To classify an unknown tuple, X, each classifier, Mi, 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;
The basic idea in Arcing is like bagging, but some of the original tuples of D may not be included in Di, where as others may occur more than once.

4. Performance Evaluation Measures
4.1 Cross Validation Technique
Cross-validation (Jaewi 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 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.

5. Experimental Results and Discussion
5.1 Intrusion Detection
5.1.1 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.

5.1.2 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.

5.2 Direct Marketing
5.2.1 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).

5.2.2 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.

5.3 Signature Verification
5.3.1 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.

5.3.2 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.

5.4 Experiments and Analysis
5.4.1 Intrusion Detection
In this section, new ensemble classification methods are proposed using classifiers in both homogeneous ensemble classifiers using bagging and heterogeneous ensemble classifiers using arcing classifier and their performances are analyzed in terms of accuracy.
5.4.1.1 Homogeneous Ensemble Classifiers using Bagging

The Acer07 and NSL-KDD datasets are taken to evaluate the proposed Bagged RBF and bagged SVM classifiers.

a) Proposed Bagged RBF and Bagged SVM

<table>
<thead>
<tr>
<th>Real Dataset</th>
<th>Classifiers</th>
<th>Classification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acer07 dataset</td>
<td>RBF</td>
<td>99.53%</td>
</tr>
<tr>
<td></td>
<td>Proposed Bagged RBF</td>
<td>99.86%</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>99.80%</td>
</tr>
<tr>
<td></td>
<td>Proposed Bagged SVM</td>
<td>99.93%</td>
</tr>
</tbody>
</table>

Table 1. The Performance of Base and Proposed Bagged Classifiers for Real Dataset

Figure 1. Classification Accuracy of Base and Proposed Bagged Classifiers Using Real dataset

In this research work, new ensemble classification methods are proposed using classifiers in homogeneous ensemble classifiers using bagging and their performances are analyzed in terms of accuracy. Here, the base classifiers are constructed using radial basis function and Support Vector Machine. 10-fold cross validation (Kohavi, R, 1995) technique is applied to the base classifiers and evaluated Classification accuracy. Bagging is performed with radial basis function classifier and support vector machine to obtain a very good classification performance. Table 1 and Table 2 show classification performance for real and benchmark datasets of intrusion detection using existing and proposed bagged radial basis function neural network and support vector machine. The analysis of results shows that the proposed bagged radial basis function and bagged support vector machine classifiers are shown to be superior to individual approaches for real and benchmark datasets of intrusion detection problem in terms of classification accuracy. According to Fig. 1 and 2 proposed combined models show significantly larger improvement of Classification accuracy than the base classifiers. This means that the combined methods are more accurate than the individual methods in the field of intrusion detection.

5.4.1.2 Heterogeneous Ensemble Classifiers Using Arcing

The Acer07 and NSL-KDD datasets are taken to evaluate the proposed hybrid RBF-SVM classifiers.

a) Proposed Hybrid RBF-SVM System

<table>
<thead>
<tr>
<th>Real Dataset</th>
<th>Classifiers</th>
<th>Classification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acer07 dataset</td>
<td>RBF</td>
<td>99.40%</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>99.60%</td>
</tr>
<tr>
<td></td>
<td>Proposed Hybrid RBF-SVM</td>
<td>99.90%</td>
</tr>
</tbody>
</table>

Table 3. The Performance of Base and Proposed Hybrid RBF-SVM Classifiers for Real Dataset

Figure 2. Classification Accuracy of Base and Proposed Bagged Classifiers Using Benchmark Dataset

Figure 3. Classification Accuracy of Base and Proposed Hybrid RBF-SVM Classifiers Using Real dataset
In this research work, new hybrid classification methods are proposed using classifiers in heterogeneous ensemble classifiers using arcing classifier and their performances are analyzed in terms of accuracy. 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 3 and 4. According to Fig. 3 and 4, the proposed hybrid models show significantly larger improvement of classification accuracy than the base classifiers and the results are found to be statistically significant.

The experimental results show that proposed hybrid RBF-SVM is superior to individual approaches for intrusion detection problem in terms of classification accuracy.

### 5.4.2 Direct Marketing

In this section, new ensemble classification methods are proposed using classifiers in both homogeneous ensemble classifiers using bagging and heterogeneous ensemble classifiers using arcing classifier and their performances are analyzed in terms of accuracy.

#### 5.4.2.1 Homogeneous Ensemble Classifiers Using Bagging

The bank marketing and labor relations datasets are taken to evaluate the proposed Bagged RBF and bagged SVM classifiers.

#### a) Proposed Bagged RBF and Bagged SVM

<table>
<thead>
<tr>
<th>Benchmark Dataset</th>
<th>Classifiers</th>
<th>Classification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSL-KDD dataset</td>
<td>RBF</td>
<td>84.74%</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>91.81%</td>
</tr>
<tr>
<td></td>
<td>Proposed Hybrid RBF-SVM</td>
<td>98.46%</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Real Dataset</th>
<th>Classifiers</th>
<th>Classification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bank Marketing</td>
<td>RBF</td>
<td>71.16 %</td>
</tr>
<tr>
<td></td>
<td>Proposed Bagged RBF</td>
<td>76.16 %</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>69.00 %</td>
</tr>
<tr>
<td></td>
<td>Proposed Bagged SVM</td>
<td>73.33 %</td>
</tr>
</tbody>
</table>

Figure 5. Classification Accuracy of Base and Proposed Bagged Classifiers Using Real dataset

<table>
<thead>
<tr>
<th>Benchmark Dataset</th>
<th>Classifiers</th>
<th>Classification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labor Relations</td>
<td>RBF</td>
<td>94.73 %</td>
</tr>
<tr>
<td></td>
<td>Proposed Bagged RBF</td>
<td>96.34 %</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>89.47 %</td>
</tr>
<tr>
<td></td>
<td>Proposed Bagged SVM</td>
<td>96.49 %</td>
</tr>
</tbody>
</table>

Figure 6. Classification Accuracy of Base and Proposed Bagged Classifiers Using Benchmark Dataset

<table>
<thead>
<tr>
<th>Benchmark Dataset</th>
<th>Classifiers</th>
<th>Classification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RBF</td>
<td>84.74%</td>
</tr>
<tr>
<td></td>
<td>Proposed Bagged RBF</td>
<td>91.81%</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>98.46%</td>
</tr>
</tbody>
</table>

TABLE 4. THE PERFORMANCE OF BASE AND PROPOSED HYBRID RBF-SVM CLASSIFIER FOR BENCHMARK DATASET

<table>
<thead>
<tr>
<th>Benchmark Dataset</th>
<th>Classifiers</th>
<th>Classification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real Dataset</td>
<td>Classifiers</td>
<td>Classification Accuracy</td>
</tr>
<tr>
<td></td>
<td>RBF</td>
<td>71.16 %</td>
</tr>
<tr>
<td></td>
<td>Proposed Bagged RBF</td>
<td>76.16 %</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>69.00 %</td>
</tr>
<tr>
<td></td>
<td>Proposed Bagged SVM</td>
<td>73.33 %</td>
</tr>
</tbody>
</table>

TABLE 5. THE PERFORMANCE OF BASE AND PROPOSED BAGGED CLASSIFIERS FOR REAL DATASET

<table>
<thead>
<tr>
<th>Benchmark Dataset</th>
<th>Classifiers</th>
<th>Classification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labor Relations</td>
<td>RBF</td>
<td>94.73 %</td>
</tr>
<tr>
<td></td>
<td>Proposed Bagged RBF</td>
<td>96.34 %</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>89.47 %</td>
</tr>
<tr>
<td></td>
<td>Proposed Bagged SVM</td>
<td>96.49 %</td>
</tr>
</tbody>
</table>

TABLE 6. THE PERFORMANCE OF BASE AND PROPOSED BAGGED CLASSIFIERS FOR BENCHMARK DATASET
ensemble classifiers using bagging and their performances are analyzed in terms of accuracy. Here, the base classifiers are constructed using radial basis function and Support Vector Machine. 10-fold cross validation (Kohavi, R, 1995) technique is applied to the base classifiers and evaluated Classification accuracy. Bagging is performed with radial basis function classifier and support vector machine to obtain a very good classification performance. Table 5 and 6 show classification performance for real and benchmark datasets of direct marketing using existing and proposed bagged radial basis function neural network and support vector machine. The analysis of results shows that the proposed bagged radial basis function and bagged support vector machine classifiers are shown to be superior to individual approaches for real and benchmark datasets of direct marketing problem in terms of classification accuracy. According to Fig. 5 and 6 proposed combined models show significantly larger improvement of Classification accuracy than the base classifiers and the results are found to be statistically significant. This means that the combined methods are more accurate than the individual methods in the field of direct marketing.

5.4.2.2 Heterogeneous Ensemble Classifiers Using Arcing

The bank marketing and labor relations datasets are taken to evaluate the proposed hybrid RBF-SVM classifiers.

a) Proposed Hybrid RBF-SVM System

<table>
<thead>
<tr>
<th>Benchmark Dataset</th>
<th>Classifiers</th>
<th>Classification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labor Relations Dataset</td>
<td>RBF</td>
<td>94.73 %</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>89.47 %</td>
</tr>
<tr>
<td></td>
<td>Proposed Hybrid RBF-SVM</td>
<td>98.24 %</td>
</tr>
</tbody>
</table>

TABLE 8. THE PERFORMANCE OF BASE AND PROPOSED HYBRID RBF-SVM CLASSIFIER FOR BENCHMARK DATASET

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Classifiers</th>
<th>Classification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bank Marketing dataset</td>
<td>RBF</td>
<td>71.16 %</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>69.00 %</td>
</tr>
<tr>
<td></td>
<td>Proposed Hybrid RBF-SVM</td>
<td>88.33 %</td>
</tr>
</tbody>
</table>

Figure 8. Classification Accuracy of Base and Proposed Hybrid RBF-SVM Classifiers Using Benchmark Dataset

In this research work, new hybrid classification methods are proposed using classifiers in heterogeneous ensemble classifiers using arcing classifier and their performances are analyzed in terms of accuracy. The dataset 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 7 and 8. According to Fig. 7 and 8, the proposed hybrid models show significantly larger improvement of classification accuracy than the base classifiers and the results are found to be statistically significant. The experimental results show that proposed hybrid RBF-SVM is superior to individual approaches for direct marketing problem in terms of classification accuracy.

5.4.3 Signature Verification

In this section, new ensemble classification methods are proposed using classifiers in both homogeneous ensemble classifiers using bagging and heterogeneous ensemble classifiers using arcing classifier and their performances are analyzed in terms of accuracy.
5.4.3.1 Homogeneous Ensemble Classifiers Using Bagging

The NIST and U.S. Zip code datasets are taken to evaluate the proposed Bagged RBF and bagged SVM classifiers.

a) Proposed Bagged RBF and Bagged SVM

TABLE 9. THE PERFORMANCE OF BASE AND PROPOSED BAGGED CLASSIFIERS FOR REAL DATASET

<table>
<thead>
<tr>
<th>Real Dataset</th>
<th>Classifiers</th>
<th>Classification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIST dataset</td>
<td>RBF</td>
<td>76.5 %</td>
</tr>
<tr>
<td></td>
<td>Proposed Bagged RBF</td>
<td>91.8 %</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>89.2 %</td>
</tr>
<tr>
<td></td>
<td>Proposed Bagged SVM</td>
<td>98.0 %</td>
</tr>
</tbody>
</table>

Figure 9. Classification Accuracy of Base and Proposed Bagged Classifiers Using Real dataset

TABLE 10. THE PERFORMANCE OF BASE AND PROPOSED HYBRID RBF-SVM CLASSIFIERS FOR REAL DATASET

<table>
<thead>
<tr>
<th>Real Dataset</th>
<th>Classifiers</th>
<th>Classification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIST dataset</td>
<td>RBF</td>
<td>76.5 %</td>
</tr>
<tr>
<td></td>
<td>Proposed Hybrid RBF</td>
<td>99.3 %</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>93.98 %</td>
</tr>
<tr>
<td></td>
<td>Proposed Hybrid RBF-SVM</td>
<td>95.45 %</td>
</tr>
</tbody>
</table>

Figure 10. Classification Accuracy of Base and Proposed Bagged Classifiers Using Benchmark Dataset

In this research work, new ensemble classification methods are proposed using classifiers in homogeneous ensemble classifiers using bagging and their performances are analyzed in terms of accuracy. Here, the base classifiers are constructed using radial basis function and Support Vector Machine. 10-fold cross validation (Kohavi, R, 1995) technique is applied to the base classifiers and evaluated Classification accuracy. Bagging is performed with radial basis function classifier and support vector machine to obtain a very good classification performance. Table 9 and 10 show classification performance for real and benchmark datasets of recognizing totally unconstrained handwritten numerals using existing and proposed bagged radial basis function neural network and support vector machine. The analysis of results shows that the proposed bagged radial basis function and bagged support vector machine classifiers are shown to be superior to individual approaches for real and benchmark datasets of handwriting recognition problem in terms of classification accuracy. According to Fig. 9 and 10 proposed combined models show significantly larger improvement of Classification accuracy than the base classifiers. This means that the combined methods are more accurate than the individual methods in the field of handwriting recognition.

5.4.3.2 Heterogeneous Ensemble Classifiers Using Arcing

The NIST and U.S. Zip code datasets are taken to evaluate the proposed hybrid RBF-SVM classifiers.

a) Proposed Hybrid RBF-SVM System

TABLE 11. THE PERFORMANCE OF BASE AND PROPOSED HYBRID RBF-SVM CLASSIFIERS FOR REAL DATASET

<table>
<thead>
<tr>
<th>Real Dataset</th>
<th>Classifiers</th>
<th>Classification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIST dataset</td>
<td>RBF</td>
<td>76.5 %</td>
</tr>
<tr>
<td></td>
<td>Proposed Hybrid RBF-SVM</td>
<td>99.3 %</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>89.2 %</td>
</tr>
</tbody>
</table>

Figure 11. Classification Accuracy of Base and Proposed Hybrid RBF-SVM Classifiers Using Real Dataset
TABLE 12. THE PERFORMANCE OF BASE AND PROPOSED HYBRID RBF-SVM CLASSIFIER FOR BENCHMARK DATASET

<table>
<thead>
<tr>
<th>Benchmark Dataset</th>
<th>Classifiers</th>
<th>Classification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>U.S. Zip code dataset</td>
<td>RBF</td>
<td>86.46 %</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>93.98 %</td>
</tr>
<tr>
<td></td>
<td>Proposed Hybrid RBF-SVM</td>
<td>99.13 %</td>
</tr>
</tbody>
</table>

The experiment results lead to the following observations.

- SVM exhibits better performance than RBF in the important respects of accuracy.
- The proposed bagged methods are shown to be significantly higher improvement of classification accuracy than the base classifiers.
- The hybrid RBF-SVM shows higher percentage of classification accuracy than the base classifiers.
- The χ² statistic is determined for all the 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 χ² significance table, it is found that this value is significant with a degree of freedom of 1. In general, the result of χ² statistic analysis shows that the proposed classifiers are significant at p < 0.05 than the existing classifiers.
- The accuracy of base classifiers is compared with homogeneous and heterogeneous models for data mining problems and heterogeneous models exhibit better results than homogeneous models for real and benchmark data sets of data mining applications.
- The data mining applications could be detected with high accuracy for homogeneous and heterogeneous models.

The future research will be directed towards developing more accurate base classifiers particularly for the data mining applications.

Acknowledgements

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

References


[38] Jiawei Han, Micheline Kamber. Data Mining – Concepts and Techniques, Elsevier Publications, 2003.
Ensembles of Classification Methods for Data Mining Applications

21


