ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Development of Hybrid Ensemble Approach for Automobile Data

M.Govindarajan*, A.Mishra

* Assistant Professor, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar – 608002, Tamil Nadu, India.

Professor, Department of Mechanical Engineering, Indira Gandhi Institute of Technology, Sarang, Odisha, India

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 automobile data like Auto Imports and Car Evaluation Databases. In this research work, new hybrid classification method is proposed using classifiers in a heterogeneous environment using arcing classifier 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 Auto Imports and Car Evaluation Databases in terms of classification accuracy.

Keywords: Machine learning, Radial Basis Function, Support Vector Machine, Ensemble, Classification Accuracy.

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 [10] 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 goal of ensemble learning methods is to construct a collection (an ensemble) of individual classifiers that are diverse and yet accurate. If this can be achieved, then highly accurate classification decisions can be obtained by voting the decisions of the individual classifiers in the ensemble.

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.

Related work

Data mining tasks like clustering, association rule mining, sequence pattern mining, and classification are used in many applications. Some of the widely used data mining algorithms in classification include Support vector machines and neural networks.

Support vector machines (SVMs) are relatively new techniques that have rapidly gained popularity because of the excellent results N. Cristianini, et al. [2] have achieved in a wide variety of machine learning problems, and solid theoretical underpinnings in statistical learning theory.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

On the other hand, Artificial Neural Networks (ANN) as a classifier algorithm are also widely-used in data mining for performing classification in a number of applications. D. Delen et al., [3] uses ANN and compares its performance against decision trees mining algorithm to develop a prediction models for breast cancer. J. A. Marchant and C. M. Onyango [9] performs a comparison between ANN and Support Vector Machine (SVM) for Drug/Nondrug Classification.

T. Ho [6]; J. Kittler [8] show the ensemble technique, which combines the outputs of several base classification models to form an integrated output, has become an effective classification method for many domains.

Freund and Schapire [4] [5] propose an algorithm the basis of which is to adaptively resample and combine (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.

In this research work, proposes a new hybrid method for mechanical problem. A new architecture based on coupling classification methods (RBF and SVM) using arcing classifier adapted to mechanical problem is defined in order to get better results.

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.

Hybrid RBF-SVM System

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

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.

Existing Classification Methods

Radial basis Function Neural Network

Oliver Buchtala, et al., [11] designed RBF that 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.

Support Vector Machine

Vapnik, V [12] presented support vector machine, 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).

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

Given a set of N linearly separable training examples $S = \{x_i \in R^n | i = 1,2,\ldots,N\}$, where each example belongs to one of the two classes, represented by $y_i \in \{\div 1, -1\}$, 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:

$$f(\mathrm{X}) = sign(W.X + b) \quad (1)$$

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

$$f(x) = sign\left(\sum_{i=1}^{N} a_i y_i (x_i . x) + b\right) \quad (2)$$

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.

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 mechanical problem is defined in order to get better results.

Proposed RBF-SVM Hybrid System

According to Breiman. L, [1], Given a set D, of d tuples, arcing 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, D_i, by sampling D with replacement. Same example from given dataset *D* may occur more than once in the training dataset D_i.
3. Use D_i to derive a model, M_i
4. Classify each example d in training data D_i and initialized the weight, W_i for the model, M_i, based on the accuracies of percentage of correctly classified example in training data D_i.
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;

Performance evaluation measures

Cross Validation Technique

Jiawei Han and Micheline Kamber [7] found that Cross-validation 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,

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

the folds are selected so that the mean response value is approximately equal in all the folds.

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.

Experimental results and discussion

Auto Imports Database Description

This data set consists of three types of entities: (a) the specification of an auto in terms of various characteristics, (b) its assigned insurance risk rating, (c) its normalized losses in use as compared to other cars. The second rating corresponds to the degree to which the auto is more risky than its price indicates. Cars are initially assigned a risk factor symbol associated with its price. Then, if it is more risky (or less), this symbol is adjusted by moving it up (or down) the scale. Actuarians call this process "symboling". A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.

Table 1. Properties of Auto Imports Database

Data Set Characteristics:	Multivariate	Number of Instances:	205
Attribute Characteristics:	Categorical, Integer, Real	Number of Attributes:	26
Associated Tasks:	Regression	Missing Values	Yes

It contains the following attributes:

1. symboling: -3, -2, -1, 0, 1, 2, 3.
2. normalized-losses: continuous from 65 to 256.
3. make: alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, Volvo.
4. fuel-type: diesel, gas.
5. aspiration: std, turbo.
6. num-of-doors: four, two.
7. body-style: hardtop, wagon, sedan, hatchback, convertible.
8. drive-wheels: 4wd, fwd, rwd.
9. engine-location: front, rear.
10. wheel-base: continuous from 86.6 120.9.
11. length: continuous from 141.1 to 208.1.
12. width: continuous from 60.3 to 72.3.
13. height: continuous from 47.8 to 59.8.
14. curb-weight: continuous from 1488 to 4066.
15. engine-type: dohc, dohcv, l, ohc, ohcf, ohcv, \ rotor.
16. num-of-cylinders: eight, five, four, six, three, twelve, two.
17. engine-size: continuous from 61 to 326.
18. fuel-system: 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
19. bore: continuous from 2.54 to 3.94.
20. stroke: continuous from 2.07 to 4.17.
21. compression-ratio: continuous from 7 to 23.
22. horsepower: continuous from 48 to 288.
23. peak-rpm: continuous from 4150 to 6600.
24. city-mpg: continuous from 13 to 49.
25. highway-mpg: continuous from 16 to 54.
26. price: continuous from 5118 to 45400.

Car Evaluation Database Description

The dataset is obtained from UCI Machine Learning Repository, which is supplied by the University of California. The car evaluation database was originally derived from a simple hierarchical decision model. The model evaluates cars according to the following concept structure:

CAR - Car acceptability
PRICE - Overall price
Buying - Buying price
Maint - Price of maintenance
TECH - Technical characteristics
COMFORT - Level of comfort
Doors - Number of doors
Persons - Capacity in terms of passengers
Lug_boot - The size of luggage boot
Safety - Estimated safety of the car

Table 2. Properties of Car Evaluation Database

Data Set Characteristics:	Multivariate	Number of Instances:	1728
Attribute Characteristics:	Categorical	Number of Attributes:	6
Associated Tasks:	Classification	Missing Values	No

PRICE, TECH, and COMFORT are three immediate concepts. Every concept is related to its lower level descendants by a set of examples. The car evaluation

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

database contains examples with the structural information removed, i.e., directly relates CAR to six input attributes: buying, maint, doors, persons, lug_boot, and safety. There are 1,728 instances that completely cover the attribute space with 6 attributes (no missing attribute values) as follows:

• buying: v-high, high, med, low
• maint: v-high, high, med, low
• doors: 2, 3, 4, 5-more
• persons: 2, 4, more
• lug_boot: small, med, big
• safety low, med, high

The class distribution, which is the number of instances per class is shown in Table 3

Table 3. Class distribution

Class Name	Number of instance per class	Percentage (%)
Unaac	1210	70.023
Acc	384	22.222
Good	69	3.993
Vgood	65	3.762

Experiments and Analysis

Auto Imports Database

The auto imports database is taken to evaluate the proposed bagged RBF and SVM for automobile prediction system.

Table 4.The Performance of Existing and Proposed Hybrid RBF-SVM Classifier for Auto Imports Database

Dataset	Classifiers	Classification Accuracy
Auto Imports Database	RBF	61.95 %
	SVM	71.21 %
	Proposed Hybrid RBF-SVM	97.56 %

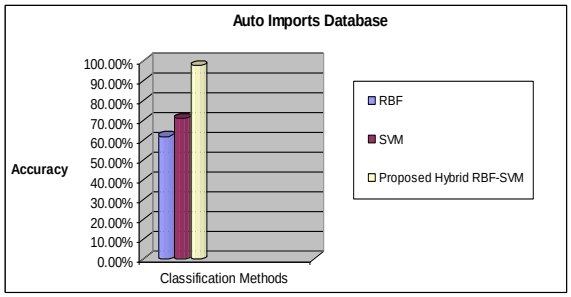

Figure 1. Classification Accuracy of Base and Proposed Hybrid RBF-SVM Classifiers Using Auto Imports Database

Car Evaluation Database

The car evaluation database is taken to evaluate the proposed bagged SVM and RBF for car marketing prediction system.

Table 5: The Performance of Existing and Proposed Hybrid RBF-SVM Classifier for Car Evaluation Database

Dataset	Classifiers	Classification Accuracy
Car Evaluation Database	RBF	88.25 %
	SVM	93.75 %
	Proposed Hybrid RBF-SVM	98.66 %

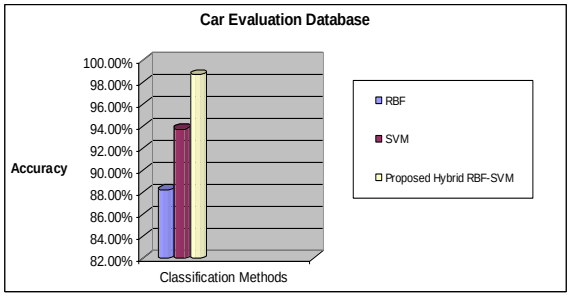

Figure 2. Classification Accuracy of Base and Proposed Hybrid RBF-SVM Classifiers Using Car Evaluation Database

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 4 and 5. According to Table 4 and 5, 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.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

The proposed ensemble of RBF and SVM is shown to be superior to individual approaches for automobile data like Auto Imports and Car Evaluation Databases in terms of Classification accuracy.

Conclusion

In this research, some new techniques have been investigated for automobile data 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 Auto Imports Database and Car Evaluation Database.
- 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 automobile data.

Acknowledgements

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

References

[1] Breiman. L, “Bias, Variance, and Arcing Classifiers”, Technical Report 460, Department of Statistics, University of California, Berkeley, CA, 1996.
[2] N. Cristianini, B. Schoelkopf, "Support vector machines and kernel methods, the new generation of learning machines", Artificial Intelligence Magazine, 23(3), 2002, pp. 31–41.
[3] D. Delen, G. Walker, and A. Kadam, “Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods”, Artificial Intelligence in Medicine, Elsevier, 2004, pp. 121-130.
[4] 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.
[5] 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.
[6] T. Ho, J. Hull, S. Srihari, “Decision combination in multiple classifier systems”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 16, 1994, pp. 66–75.
[7] Jiawei Han , Micheline Kamber, “Data Mining – Concepts and Techniques”, Elsevier Publications, 2003.
[8] J. Kittler, “Combining classifiers: a theoretical framework”, Pattern Analysis and Applications, 1, 1998, pp.18–27.
[9] J. A. Marchant and C. M. Onyango, “Comparison of a Bayesian Classifier with a Multilayer Feed-Forward Neural Network using the Example of Plant/Weed/Soil Discrimination”, Computers and Electronics in Agriculture, Elsevier, 39, 2003, pp. 3-22.
[10] Michie, D., Spiegelhalter, D. J., & Taylor, C, “Machine learning, Neural and statistical classification”, Ellis Horwood, 1994.
[11] 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.
[12] Vapnik, V, “Statistical learning theory”, New York, John Wiley & Sons, 1998.