Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 5, May 2013, pg.134 – 138

RESEARCH ARTICLE

Decorate Ensemble of Artificial Neural Networks with High Diversity for Classification

Mittal C. Patel[1], Prof. Mahesh Panchal[2], Himani P. Bhavsar[3]
[1]PG Student, India
[2]Head of the Computer Engineering, India
[3]PG Student, India

[1] Me.mittal87@gmail.com; [2] mkhpanchal@gmail.com; [3] himani411@gmail.com

Abstract— An important data mining task is classification which is used to predict target categories of data instances. DECORATE is one of the most popular ensemble learning techniques, and can use strong learner to build diverse committees in a straightforward strategy. Artificial Neural networks (ANN) are very flexible with respect to incomplete, missing and noisy data and also makes the data to use for dynamic environment. ANN is dependent on how best is the configuration of the net in terms of number of weights, neurons and layers. In this paper, DECORATE with ANN as a base classifier is used to classify data from UCI repository. An experiment is conducted on the public datasets, and the analysis results show that the DECORATE ensemble of ANN improves the performance of classification obviously.

Key Terms: - Artificial neural network; Classification; Classifier; DECORATE Ensemble; Diversity; Neural Network Ensembles; UCI Datasets.

I. INTRODUCTION

Data Mining is the process to analyze the data from all different views and finally forming them into meaningful information. Data Mining is the central step in the KDD process that performs the different tasks like classification, clustering, summarization, regression, task analysis.

Now days, an active research topic is classification in the data mining. The purpose of classification is used to predict the target category of instance. Classification is a data mining (machine learning) technique used to predict group membership for data instances. For example, you may wish to use classification to predict whether the weather on a particular day will be “sunny”, “rainy” or “cloudy”. Popular classification techniques include decision trees and neural networks.

Classification is two-way process: [1]

1) Model Construction
   - ➔ Predicts categorical class labels (discrete or nominal)
2) Model Usage
   - ➔ Classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data.

Classification can be applied in many fields such as bankruptcy prediction [1], text processing [2], biomedical data analysis [3], image processing [4], fault diagnosis [5], etc. Recently, many popular techniques in the fields of data mining and machine learning have been used as classification methods, such as support vector machines (SVM) [6], k-nearest neighbors (kNN) [7], naive Bayes [8], CART [9], C4.5 [10], Rochhio algorithm [11], etc.

An **artificial neural network** (ANN), often just called a "neural network" (NN), is a mathematical model or computational model based on biological neural networks, in other words, is an emulation of biological neural

system. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. [12]. Thus, ANN is powerful non-linear statistical data modeling tools and usually used to model complex relationships between inputs and outputs or to find patterns in data [13].

Ensemble learning is a kind of techniques that train a set of component classifiers and then, combine their predictions to classify new instances, and it has become one of the most explored topics within data mining and machine learning communities. The motivation of ensemble classifier is that improve the performance of single one bye combing the outputs of several individual predictors. Research work has shown that to combine a set of simple classifiers may result in better performance in comparison to any single sophisticated classifier[14].

Diversity is the disagreement of an ensemble member with the ensemble's prediction as a measure of diversity. More precisely if Ci(x) is the prediction of the i-th classifier for the label of x; C*(x) is the prediction of the entire ensemble, then the diversity of the i-th classifier on example x is given by

$$d_i(x) = \begin{cases} 0: \text{ if } C_i(x) = C^*(x) \\ 1: \text{otherwise} \end{cases}$$

To compute the diversity of an ensemble of size n, on a training set of size m, now, average the above term [15]:

$$\frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} d_i(x_j)$$

This measure estimates the probability that a classifier in an ensemble will disagree with the prediction of the ensemble as a whole [15].

In this paper, DECORATE ensemble with base classifier ANN are combined to obtain more accurate classification. A comprehensive experiment of different methods is conducted on the several public datasets. The experimental results indicate that the DECORATE ensemble of ANN improves the performance of classification obviously.

The rest of this paper is structured as follows:

Section 2 introduces the DECORATE Ensemble of ANN in detail with method to generate Artificial data. Section 3 gives comprehensive evaluation of its effectiveness by applying it to several public datasets and comparing the results obtained with other methods. Finally, section 4 presents conclusions.

II. NEW DECORATE ENSEMBLE OF ANN

New extended DECORATE algorithm from original DECORATE algorithm will work according to section 2.2. Section 2.1 will give overview of Ensemble Learning.

2.1 Ensemble Learning

Ensemble learning is to use a set of classifier to learn partial solutions for a given problem, and then integrate these solutions by using some strategies to construct a final solution to the original problem. Recently, ensemble learning is one of the most popular fields in the data mining and machine learning communities, and has been applied successfully in many real applications such as microarray data classification [21], text classification [22] and Spam email classification [23]. As one of the most popular ensemble learning techniques, DECORATE (Diverse Ensemble Creation by Oppositional Relabeling of Artificial Training Examples) is a simple meta-learner that can use any strong learner as a base classifier to build diverse committees in a fairly straightforward strategy [24]. The motivation of DECORATE is based on the fact that to combine the outputs of multiple classifiers is only useful if they disagree on some inputs [25]. Decorate is designed to use additional artificially generated training data, and add different randomly constructed instances to the training set in order to generate highly diverse ensembles. In this paper, ANN algorithm and DECORATE ensemble technique are combined to obtain more accurate classification.

2.2 Proposed Algorithm

Algorithm includes that first of all take the original dataset, do the partition of it. Generate the training dataset and testing dataset according to partition. Create different training datasets and testing dataset for the given iteration. Then apply ANN on training iterations and generate output according to testing dataset. Finally, calculate the error and accuracy for it (previously counted error).Generate artificial data according to the procedure and add it to the original dataset, then again apply ANN on it and take the result and add it to the

ensemble classifiers if its error is less than e1.Repeat this procedure until specified no of iterations or desired no of ensemble size achieves.

Finally, combine all results by applying Average or Voted method.

2.2.1 Artificial Data generation method

For each tuple from the given dataset, check whether the attributes of given tuple are matched with other tuples? If it is, then list out classes of those similar tuples. From the list of classes take minimum occurred class for that tuple and it to the artificial data generation list. But, if it is not then takes controversial class of that tuple and add it to the artificial data generation list.

III. RESULTS AND DISCUSSIONS

3.1 Dataset

An experimental evaluation of the DECORATE ensemble of ANN is presented here. Four dataset from the UCI machine learning dataset repository, i.e., the iris dataset, the Image Segment dataset, the Breast-cancer dataset and the Glass Identification are used in experiment.

3.2 Performance Measures

In the experiment, Accuracy metric is adopted to analyze the performance of classification. As shown in Table 1, four cases are considered as the result of classifier to the instance [16].

Class C		Result of Classifier	
		belong	Not belong
Real Classification	Belong	TP	FN
	Not belong	FP	TN

Table 1: Cases of the classification for one class

TP (True Positive): the number of instances correctly classified to that class.
TN (True Negative): the number of instances correctly rejected from that class.
FP (False Positive): the number of instances incorrectly rejected from that class.
FN (False Negative): the number of instances incorrectly classified to that class.
The Accuracy indicates the proportion of correctly classified instances and can be defined as follow:
Accuracy = (TP + TN) / (TP + TN + FP + FN)

3.3 Experimental Result Analysis

To measure the effectiveness of the DECORATE ensemble of ANN, RBF network, Random Forest and Naïve base are implemented as benchmarks in the experiments. Fig 1 shows the classification results of all base classifiers.

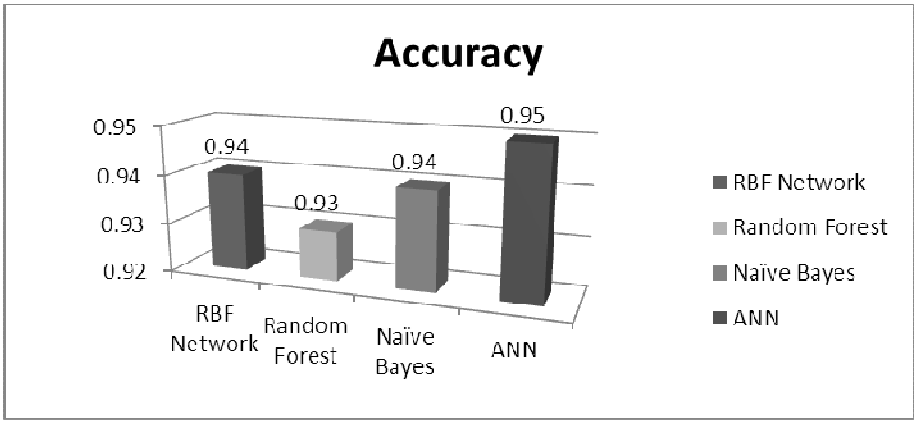

Figure 1: ANN with Other Classifiers

In figure 1, comparison of base classifiers is given by its accuracy. When RBF Network is applied on different dataset, it gives 94% accurate result which is 1% less than result of ANN. When Random Forest is applied on different dataset, it gives 93% accurate result which is 2% less than result of ANN. Same as RBF network and Random Forest, When Random Forest is applied on different dataset, it gives 93% accurate result which is 2% less than result of ANN.

For the DECORATE ensemble of ANN, cross-validation approach is used to evaluate classification performance. Suppose 10 cross validation is used for each dataset, then each dataset is split into 10 parts and the algorithm is run once for each parts. Nine parts are grouped

Together to form the train dataset for training and the remaining tenth is used as test dataset for test. The training-test procedure is conducted ten times and the average of the ten performances is used as final result. Fig 2 shows result of Average combining method and Fig 3 shows result of Voting combining method on New DECORATE algorithm. Fig 2 and 3 gives the difference of accuracy of original DECORATES and extended New DECORATE algorithm.

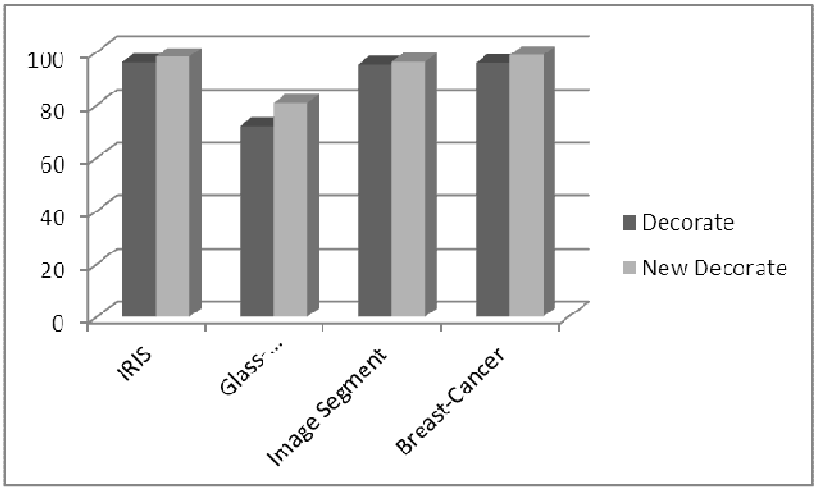

Figure 2: UCI Repository Datasets with comparison of DECORATE and New DECORATE algorithm (Using Voted Method)

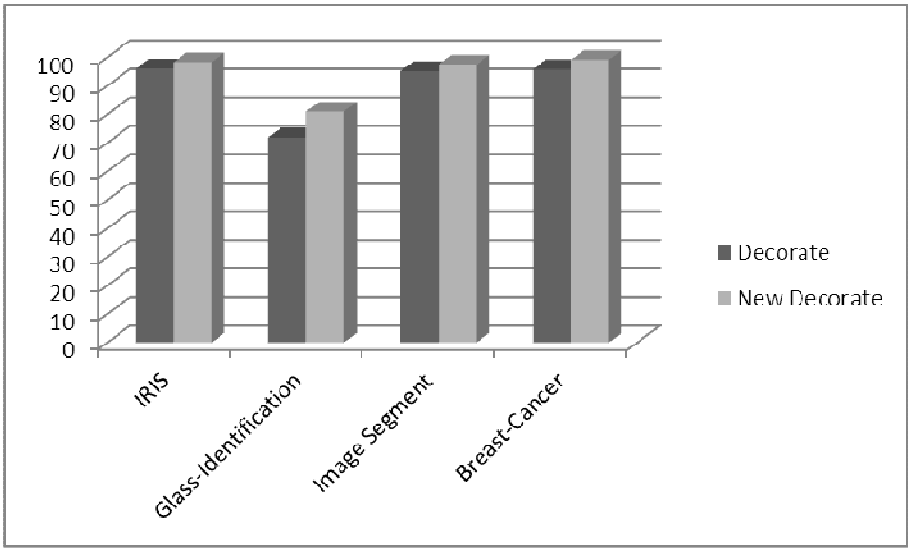

Figure 3: UCI Repository Datasets with comparison of DECORATE and New DECORATE algorithm (using Voted method)

IV. CONCLUSION

ANN is a powerful technique and an ensemble of accurate and diverse neural networks was found capable of providing better results than a single neural network. If two classifiers produce different errors on new input data then both classifiers are considered to be "diverse". Diversity in an ensemble of neural networks can be handled by manipulating input data or output data. DECORATE is a classifier combination technique to construct a set of diverse base classifiers using additional artificially generated training instances. The output can be achieved by average and voted combination strategies. In this paper, The DECORATE ensemble and ANN method are combined for classification, and the approach has been tested by public datasets from the UCI

Machine Learning Repository. The experimental results says that the DECORATE ensemble of ANN method achieves obvious improvement of classification performance.

REFERENCES

[1] Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques (2nd edition), The University of Illinois at Urbana-Champaign, Morgan Kaufmann, 2006.

[2] A. X. Sun, Y. Liu, E. -P. Lim. Web classification of conceptual entities using co-training, Expert Systems with Applications, 38 (2011), 14367 – 14375.

[3] J. -H. Hong, S. -B. Cho. Gene boosting for cancer classification based on gene expression profiles,Pattern Recognition, 42 (2009), 1761 – 1767.

[4] A. J. Ishak, A. Hussain, M. M. Mustafa. Weed image classification using Gabor wavelet and gradient field distribution, Computers and Electronics in Agriculture, 66 (2009), 53 – 61.

[5] Z. Q. Geng, Q. X. Zhu. Rough set-based heuristic hybrid recognizer and its application in fault diagnosis, Expert Systems with Applications, 36 (2009), 2711 – 2718.

[6] V. Vapnik. The nature of statistical learning theory, Springer, New York, 1995.

[7] T. Cover, P. Hart. Nearest neighbor pattern classification, IEEE Transactions on Information Theory, 13 (1), (1967), 21 – 27.

[8] P. Domingos, M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss,Mach Learn, 29 (1997), 103 – 130.

[9] L. Breiman, et al. Classification and regression trees, Wadsworth, Belmont, 1984.

[10] J. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Matteo, CA, 1993.

[11] T. Joachims. A probabilistic analysis of the Rochhio algorithm with TFIDF for text categorization,in: Proceedings of the 14th international conference on machine learning, 1997, pp. 143 – 151.

[12] DR. YASHPAL SINGH, ALOK SINGH CHAUHAN, "Neural Networks in Data Mining", JATIT, PP 37 – 42, 2009.

[13] http://en.wikipedia.org/wiki/Artificial neural network, 2012.

[14] T. G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization, Machine Learning, 40 (2000), 139 – 158.

[15] Lei SHI, Haiping SI, Hongbo QIAO, Xinming MA, "DECORATE Ensemble of Artificial Neural Network for Classification", Journal of Computational Information Systems 8: 8503–8509, Aug-2012.

[16] Y. Yang, X. Liu. A re-examination of text categorization methods, in: Proceedings of the 22th annual international ACM SIGIR conference on research and development in the information retrieval, 1999, pp. 42 – 49. Intelligence, volume 12, PP 993–1001, October 1990.

