Ensemble of Diverse Decorate Classifier with Neural Network

Sweety R. Patel[1] Deepali Koul[2] Sunny Patel[3] Proffesor Anuradha Nawathe[4]
[1,2,3]Research Scholar [4]Associate Professor
[1,2,3,4]Department of Computer Engineering
[1,4]A.V.O.C.E, Sangamner [2]K.I.T.E, Jaipur [3]K.K.W, Nashik

Abstract— Challenging problem in data mining is to classify data sets that suffer from imbalanced Class distributions. In machine learning, the ensemble of classifiers are known to increase the accuracy of single classifiers by combining all of them, but none of these learning techniques alone solve the imbalance problem, to deal with this problem the ensemble algorithms have to be designed specifically. DECORATE is ensemble learning techniques, that directly constructs diverse hypotheses using additional artificially-constructed training examples. ANN is very flexible with respect to missing, incomplete and noisy data and also makes the data to use for dynamic environment. In this paper, we present a method of generating the diversity of the classifier data set from UCI repository by neural networks and learning by DECORATE for optimum accuracy.

Key words: Artificial neural network, Classification, Classifier, DECORATE Ensemble, Diversity, Neural Network Ensembles, UCI Datasets

I. INTRODUCTION

Data Mining, also popularly known as Knowledge discovery in databases (KDD) refer to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. Data mining is being put into use and studied for databases; include object-relational databases, relational databases, object-oriented databases, data warehouses, and transactional databases, semi- structured and unstructured repositories such as the WWW, advanced databases such as multimedia databases, spatial databases, textual databases, time-series databases and flat files. Apply a data-mining algorithm such Anomaly detection , Association rule learning, Clustering, Classification, Regression , Summarization in ways which match with the original goal of the KDD process**.** Classification analysis is the organization of data in given classes. Also called as supervised classification, the classification can use give class labels to order the objects in the data collection? It is a two way process[1] The classification algorithm learns from the training set and builds a model that could be used to either accept or reject credit requests in the future The model is used to classify new instance. For example, after starting a new credit policy, the Our Video Store managers could analyze the customers those who received credits with three possible labels "safe", "risky" and "very risky"[7].

Researchers from many scientific disciplines are designing artificial neural networks to solve a variety of problems in prediction, pattern recognition, optimization, associative memory. A neural network is distributed, parallel information processing structure consisting of processing elements (which can possess a local memory and can carry out localized information processing operations) interconnected via unidirectional signal channels. Each processing element has a only one output connection that branches ("fans out") into as many collateral connections as desired; each carries the same signal - the processing element output signal. The processing output signal can be of any mathematical type desired. The information processing that continuous within each processing element can be defined arbitrarily with the restriction that it must be completely local; that is, it can depend only on the present values of the input signals arriving at the processing element via impinging connections and on the values stored in the processing element's local memory. For example, we typically ask the opinions of several doctors before agreeing to a medical procedure, we read user reviews prior to buying an item (particularly big ticket), we can find future employees by checking their references, etc behaviors vis-à-vis their credit, and label accordingly the Neural networks have emerged as advanced data mining tools in cases where other techniques may not produce satisfactory predictive models. As the term shows, biologically modeling capability inspired by neural networks, but are primarily statistical modeling tools [7]. An ensemble-based system is obtained by combining diverse models, such as experts or classifiers, are strategically combined and generated to solve a particular computational problem. Therefore, such systems are also known as multiple classifier systems, or just ensemble systems. The motivation of ensemble classifier is to produce better results than single models if the classifiers in the ensembles are accurate and diverse. Ensembles are a combination of multiple bases. Final classification results depend on the combined outputs of individual models. Ensembles perform better when base models are unstable classifiers whose output undergoes significant changes in response to small changes in the training data [12].

The difference of individual learners is interpreted as “diversity” in ensemble learning. . It is commonly agreed that the success of ensemble is attributed to diversity the degree of disagreement within an ensemble. It is generally believed that diversity in an ensemble could help to improve the performance of class imbalance learning. So long, no study has been investigated diversity in depth in terms of its dentitions and effects in the context of class imbalance learning. It is not clear whether diversity will have any similar or different impact on the performance of any minority and majority classes. More precisely if Ci(x) is the prediction of the i^{th} classifier for the label of a; C*(x) is the prediction of the whole ensemble, then the diversity of the i-th classifier to example x is given by

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

To compute the diversity of n ensemble of classifier of size n, on training data set of size m, by averaging the above term [13] :

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

This formula 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 relates with base classifier ANN are combined to obtain more precise classification. An inclusive experiment of different methods is done on the several datasets. The experimental results denote that the DECORATE ensemble of ANN enhance the performance of classification.

The second part of this paper can be shaped as follows:

Section 2, instigate the DECORATE Ensemble of ANN in point factor with method to generate Artificial data.

Section 3 instigates inclusive evaluation of its efficacy by applying it to many datasets and equalizing the results derived with other methods. Lastly, section 4 exhibit conclusions.

We explore the, inducement behind the development of NEW DECORATE ENSEMBLE OF ANN

II. LITERATURE SURVEY

A. Ensemble neural Network

In order to solve the problem of classification and prediction, an ensemble of accurate and diverse neural networks was found capable of providing better results than a single neural network. [9]

Each network has different weight, which decreases in generalization error and in variance of single classifier. Experiments with Iris data, breast cancer data set, and diabetes data set from the UCI machine learning repository showed NNEs produce better performance as compared to other ensemble methods. [3]

Combining multiple evolved ANNs has been actively researched recently. The main idea of neural networks ensemble is that a population of ANNs contains more information than any single ANN in the population. Such information can be used to improve generalization performance and reliability. Generally multiple ANNs in the last generation are combined to construct an ensemble that has better generalization performance provided that the last generation individuals complement each other in the generalization.[14]

B. Diversity

Therefore, both diversity and uncertainty can be used to support decision making in the ensemble selection process. [10]. Ensemble methods usually improve the prediction performance over a single classifier. It is also well known that the performance of an ensemble is related to both the accuracy and diversity of its base learners. Krogh and Vedels by have shown that ensemble errors depend on both average error of the base models and diversity.

We conclude that introducing additional diversity among the models of multi-label ensembles allows for maximization of their performance. To maximize the effect of combining multiple ANNs, a method for large diversity of neural networks in evolution should be used. [13]

In ensemble normal process, outputs of all networks is combined. There may be a situation where outputs from multiple networks to be different. At that time Dietterich suggested that if two classifiers produce different errors on new input data then both classifiers are considered to be "diverse". [9]

It is well-known that ensemble performance relies heavily on sufficient diversity among the base classifiers. With this point, the blueprint used to balance base classifier accuracy and diversity must be considered a base component of any ensemble algorithm.[14]

C. Decorate Ensemble

DECORATE (the most popular method) is a classifier combination technique to construct a set of diverse base classifiers using additional artificially generated training instances. The predictions from the base classifiers are then integrated into one to output the final results by using some Strategies. The DECORATE algorithm is used for diverse ensemble method.

The DECORATE can also be effectively used for the following:

- Active learning, accurate model is required to learn to reduce the number of training examples;
- Squeezing unlabeled data to ameliorate accuracy in a semi- supervised learning;
- Combining both semi-supervised and active learning for outstrip results;
- procuring improved class membership probability evaluates, to accelerate in cost sensitive decision making;
- Minimizing the error of regression methods;
- Enhancing the accuracy of relational learners.

III. NEW DECORATE ENSEMBLE OF ANN

New advancement in DECORATE algorithm from original DECORATE algorithm will work according to given following section.

A. Ensemble Learning

An Ensemble based systems can be used in problem domains other than improving the generalization performance of a classified as Incremental learning, Error correcting output codes, Feature selection. It is obtained by combining diverse models. Therefore, such systems are also referring to as multiple classifier systems. Example, for diagnosis of a neurological disorder, a neurologist may use the electroencephalogram (one-dimensional time series data), magnetic resonance imaging MRI, functional MRI, or positron emission tomography PET scan images (two- dimensional spatial data), the amount of certain chemicals in the cerebrospinal fluid along with the subjects demographics such as age, gender, education level of the subject, etc. (scalar and or categorical values).These heterogeneous features cannot be used all together to train a single classifier (and even if they could - by converting all features into a vector of scalar values such training is unlikely to be successful)[10]. In such cases, an ensemble of classifiers can be used, where a separate classifier is trained on each of the feature sets independently.

The decisions made by each classifier can then be combined by any of the combination rules. Decorate is a recently introduced ensemble method that constructs diverse committees using Artificial data. For missing data, DECORATE is the most robust has been applied successfully in many real applications such as Spam email classification [8], text classification [8], micro array data

classification [4]. DECORATE (Diverse Ensemble Creation by Oppositional Relabeling of Artificial Training Examples) is one of the most popular ensemble learning techniques, DECORATE is a meta-learner for building diverse ensembles of classifiers by using specially constructed artificial training examples. Inclusive experiments have determined that this technique is consistently more precise than the Bagging, Base classifier and Random Forests. Decorate also obtains higher accuracy than Boosting on small training sets, and reach analogous performance on wide-ranging training sets. In this paper, ANN as base classifier and DECORATE ensemble technique are combined to obtain more accurate classification.

Algorithm includes ANN need to be trained in the learning process and then the synaptic connections that exist between the neurons are adjusted appropriately. First of all take the training set T, consisting of N instances, ensemble size, maximum number of iterations to build an ensemble, factor to determine number of artificial instances to generate to get ensemble classifier. Do the partition of original dataset. Provide the given training set T as the input of base learner ANN to obtain a classier according to partition. Generate uncommon training datasets and testing dataset for the given iteration. Then apply ANN base learner on training dataset T iterations and generate outputs ensemble classifier according to testing dataset. Finally, Compute ensemble error and accuracy for it (previously computed error).Generate artificial data based on the distribution of training data and add it to the original dataset T, then again apply base learner ANN on T to obtain a new classier and take the result and add it to the ensemble classifiers if its error is less than previously occurred error. Reiterate this procedure until given no of iterations or desired no of ensemble classifier size achieves.

– Algorithm for DECORATE ensemble of ANN

Input: T, training set consisting of N instances;

ANN, base learner;

C_{size}, desired ensemble size;

Imax, maximum number of iterations to build an ensemble;

R_{size}, a used to determine number of artificial data to generate. AFdata, Artificial data set

Output: ensemble classifier

(1) Let i = 1 and trials = 1
(2) Provide the given training set T as the input of base learner
ANN to obtain a classifier C_i
(3) Initialize ensemble C* = C_i
(4) Compute ensemble error as e= (1/N) * ($\Sigma^{N}_{i=1}$ I (C*(xi≠yi)).
(5) While i < C_{size} and trials <I_{max}
(6) Generate Rsize × N training instances R based on the
Distribution of training data
For Each tuple X in C*
{
Find tuples with unique attributes and Label data in R with probability of class labels inversely proportional to those predicted by C*
AFdata = AFdataU {tuple with new class};
}
(7) If (size of AF_{data}> Rsize)
{
AFdata = Random $AF_{data}$of SizeRsize;
}
(8) T = T∪R
(9) Apply base learner ANN to obtain new classifier C
(10) C = C∪C′
(11) T = T − R, remove the artificial data
(12) Compute training error e′ of C* as in step 4
(13) If e′ ≤ e, let i = i + 1 and set e= e′; Otherwise, remove C′
From the ensemble set C, i.e. C* = C* − C′
(14) Trials= trials + 1
(15) End While

Finally, incorporate all results by assigning Average or Voted
Method to obtain more accurate classification

– *Artificial Data generation method*

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

IV. 4Results And Discussion

A. Dataset

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

B. Performance Measure

In the experiment, to analyze the performance of classification Accuracy metric is adopted. As shown in Table 1, as the result of classifier to the instance [10] four cases are considered. Note that the datasets vary in the numbers of classes, training examples, numeric and nominal attributes thus providing a diverse tested [10]

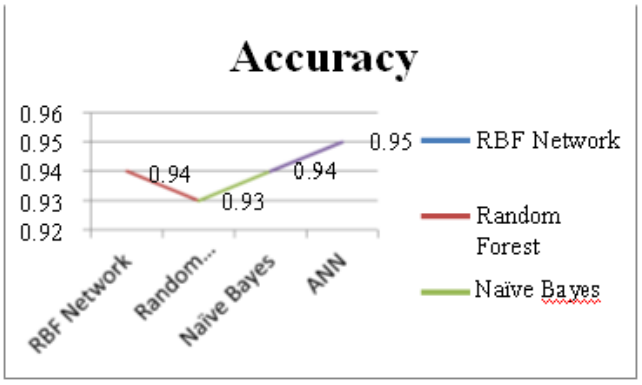

Fig. 1: ANN with Other Classifiers

In above figure 1, comparison of base classifiers is given by its accuracy. When Random Forest is applied on different dataset, it gives 93% accurate result which is 2% less than that of ANN. When RBF Network is applied on different dataset, it gives 94% accurate result which is 1% less than that of ANN. Same as RBF network and Random Forest, When it is applied on various dataset, it delivers 93% correc0074 result which is For the DECORATE ensemble of base classifier ANN, to evaluate classification performance cross-validation approach is used. Suppose for each data set 10 cross validation are used, then every data set is split into equal 10 parts after that algorithm will be executed once for each parts. Other nine parts are grouped together to form training data set and remaining 10 is

Used as test data set for testing. For a final result training test procedure is applied ten times and average of ten performances is taken. Fig 2 shows result of Average combining method and Fig 3 shows result of Voting combining method on New DECORATE algorithm. Both Figure 2 and 3 gives the dissimilarity of accuracy of original DECORATES and extended New DECORATE algorithm

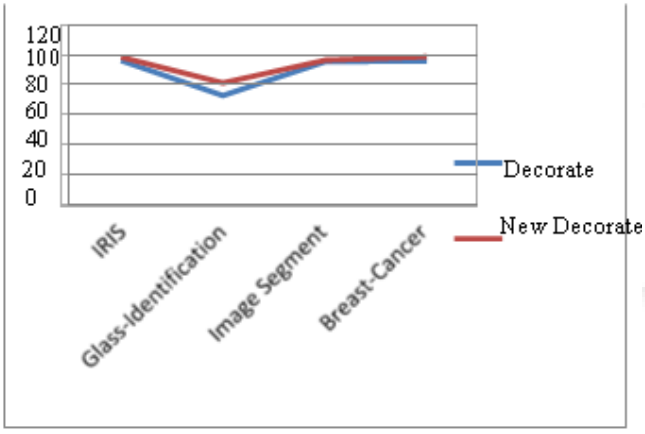

Fig. 2: Repository Datasets(UCI) with comparison of DECORATE and New DECORATE algorithm (Using Voted Method)

V. Conclusions And Future Scope

As an emulation of biological neural system, ANN is a rigorous technique and it has been applied successfully in various fields. DECORATE is a classier combination technique to construct a set of diverse base classier using additional artificially generated training instances. The predictions from the base classier are then integrated into one to output final results by using some strategies. So, The DECORATE ensemble and ANN method are combined for classification, and the approach will be tested by public datasets from the UCI Machine Learning depository. The experimental results indicate that the DECORATE ensemble of new ANN method achieves self-evident improvement of classification performance.

References

[1] P. Domingos, M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss,MachLearn, 29 (1997), 103 – 130
[2] A. J. Ishak, A. Hussain, M. M. Mustafa. Weed image classification using Gabor wavelet and gradient field Computers, 66 (2009), 53 – 61.
[3] T. Jochims A probabilistic analysis of the Rochhio algorithm with TFIDF for text: Proceedings of the international conference on machine learning in datamining, 1997, pp. 143 – 151
[4] Haiping SI, Lei SHI, Xinming MA Hongbo QIAO, "DECORATE Artificial Neural Network for classification",Journal of computer information System 8:8503-8509,August -2012
[5] Y. Liu, E. -P. Lim. A. X. Sun, Web classification of conceptual entities using co- training, Systems, 38(2011),14367-14375.
[6] Y. Yang, X. Liu. A re-examination of text methods, in: Proceedings of the 22th annual international ACM SIGIR conference on research and development in the information retrieval, 1999, pp. 42. Intelligence, PP 993–1001, Oct 1990
[7] Micheline Kamber and Jiawei Han and, Data Mining: concepts and technique, The University of Urbana-Champaign, Morgan Kaufmann,2006.
[8] S. B. Cho J.H.Hong. Gene boosting for cancer classification based on gene expression profiles,Pattern Recognition algorithm , 42 (2009),1761-1767
[9] Q. X. Zhu, Z. Q. Geng,. ,. Rough set-based heuristic hybrid recognizer ,its application in fault diagnosis, Expert Systems with Applications, 36 (2008-2009), 2711 – 2718
[10] L. Breiman, et al. Classification and regression trees, Wadsworth, Belmont,19984
[11] Vapnik. The nature of statistical and learning theory, Springer, New York(USA),1995
[12] T. G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees boosting,bagging, and randomization, DataMining, 40 (2000-2001),139-158
[13] P. Hart ,T. Cover,. Nearest neighbor pattern classification, IEEE Transactions on Information Theory, 13 (vol 1), (1967), 21-27
[14] ALOK CHAUHAN ,DR. YASHPAL S., , "Neural Networks in Data Mining technique ",JATIT, PP37–42,2009
[15] X. Liu ,Y. Yang. A re-examination of the text categorization of methods, in: Proceedings of the 22 annual international ACM SIGIR conference on R and Dt in the information retrieval system, 1999, pp. 42 – 49. Intelligence, vol 12, PP 993–1001,Oct 1990
[16] J. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Matteo, CA,1993

