International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-3 Number-1 Issue-9 March-2013

An Ensemble Method based on Particle of Swarm for the Reduction of Noise, Outlier and Core Point

Satish Dehariya[1], Divakar Singh[2]

Abstract

The majority voting and accurate prediction of classification algorithm in data mining are challenging task for data classification. For the improvement of data classification used different classifier along with another classifier in a manner of ensemble process. Ensemble process increase the classification ratio of classification algorithm, now such par diagram of classification algorithm is called ensemble classifier. Ensemble learning is a technique to improve the performance and accuracy of classification and predication of machine learning algorithm. Many researchers proposed a model for ensemble classifier for merging a different classification algorithm, but the performance of ensemble algorithm suffered from problem of outlier, noise and core point problem of data from features selection process. In this paper we combined core, outlier and noise data (COB) for features selection process for ensemble model. The process of best feature selection with appropriate classifier used particle of swarm optimization. Empirical results with UCI data set prediction on Ecoil and glass dataset indicate that the proposed COB model optimization algorithm can help to improve accuracy and classification.

Keywords

Classification, Ensemble, PSO

1. Introduction

Ensemble classification technique plays a vital role in data mining classification of data. The performance of individual classifier is not better in concern of accuracy and majority voting. The ensemble method started in last decade of machine learning research repository. When instances with known label are given the learning is called supervised learning and if instances are unlabeled the learning is called unsupervised learning.

Satish Dehariya, Department of Computer Science & Engineering, BUIT, Bhopal, India.
Divakar Singh, Department of Computer Science & Engineering, BUIT, Bhopal.

But unsupervised learning provides useful classes of items which is called clusters. Clusters are groups of similar types of objects. These groups are formed with classification methods [3]. These classifications are done by classifiers. But when an object need to be classified into predefined group or class on the basis of number of observed attributes related to that object a classification problem is occurred. Another type of learning is reinforcement learning where information's are provide by the environment in the form of scalar reinforcement signal which constitutes a measurement of system operation i.e. how well system is operating. Meta-learning uses set of attributes called meta-attributes to represent the characteristics of learning tasks [4, 5]. So it is not a good way to utilize one method or algorithm to solve a particular problem because every algorithm has strength with some limitations. So the best idea is use strengths of one method over the limitations of another algorithm. So techniques of applying algorithms in such way are called ensemble of classifiers. COB (core, outlier, and boundary) method quantitatively measures the accuracies of majority voting ensembles for binary classification.[8]. Good ensemble methods are that in which each individual classifiers are accurate and diverse. But ensemble methods are combination of predictions made by a set of individual classifiers. Accurate classifier is meant to be produce accurate prediction than the random classifier and diverse classifier is meant to be produce prediction independently. For experimental purpose of COB three different ensemble methods bagging, random forests, and a randomized ensemble, two different numbers of individual classifiers and three different machine learning algorithms decision trees, k-nearest neighbors, and support vector machines are used. The COB model results that the accuracy of ensemble method is worse with the present of nonempty core subset than the accuracy of the binomial method. The COB model is an enhancement to the binomial model with addition of two subsets core and outlier. The majority votes are decomposed into three terms an average individual accuracy, good diversity and bad diversity [9.10]. Diversity can be defined as a consequence of two decisions (1) the choice of error function and (2) the choice of combiner function in the design of ensemble problem in the machine

learning. While illustrating the majority votes two special case pattern of success and pattern of failure were introduced in this paper. Probability distributions over all possible combinations of correct/incorrect votes are defined to improve the individual accuracy *p*. Each combination where exactly (T+1)/2votes are correct, appears with probability α. In this pattern no votes are wasted [11, 12]. The above section discuss introduction of stream data classification. In section 2 we discuss related work for ensemble classification. In section 3 discuss proposed approach for Majority Voting. In section 4 discuss experimental result analysis and finally discuss conclusion of our paper in section 5.

2. Related Work

In this section we discuss method for ensemble classifier for improving majority voting of classifier and improved the accuracy of classification technique. Xueyi Wang entitled "A New Model for Measuring the Accuracies of Majority Voting Ensembles "a new model called COB (core, outlier, and boundary) which quantitatively measures the accuracies of majority voting ensembles for binary classification [1]. Good ensemble methods are that in which each individual classifiers are accurate and diverse. But ensemble methods are combination of predictions made by a set of individual classifiers. Accurate classifier is meant to be produce accurate prediction than the random classifier and diverse classifier is meant to be produce prediction independently. For experimental purpose of COB three different ensemble methods bagging, random forests, and a randomized ensemble, two different numbers of individual classifiers and three different machine learning algorithms decision trees, k-nearest neighbors, and support vector machines are used. The COB model results that the accuracy of ensemble method is worse with the present of nonempty core subset than the accuracy of the binomial method. The COB model is an enhancement to the binomial model with addition of two subsets core and outlier. PSOvin Brown, and Ludmila I. Kuncheva entitled ""Good" and "Bad" Diversity in Majority Vote Ensembles" accuracy is not straight forward with the desired diversity in classifier ensembles is proposed [2]. The majority votes are decomposed into three terms an average individual accuracy, good diversity and bad diversity. Diversity can be defined as consequences of two decisions (1) the choice of error function and (2) the choice of combiner function in the design of ensemble problem in the machine learning. While illustrating the majority votes two special case pattern of success and pattern of failure were introduced in this paper. A probability distribution over all possible combinations of correct/incorrect votes is defined to improve the individual accuracy *p*. Each combination where exactly (T+1)/2votes are correct, appears with probability α. In this pattern no votes are wasted. Tao, Dacheng, Tang, Xiaoou, Li, Xuelong, Wu and Xindong entitled "Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval" a new asymmetric bagging and random subspace mechanism is designed [3]. Relevance feedback schemes based on support vector machines (SVM) have been widely used in content-based image retrieval (CBIR). However, the performance of SVM-based relevance feedback is often poor when the number of labeled positive feedback samples is small. This is mainly due to three reasons: 1) an SVM classifier is unstable on a small-sized training set; 2) SVM's optimal hyper plane may be biased when the positive feedback samples are much less than the nePSOtive feedback samples, and 3) over fitting happens because the number of feature dimensions is much higher than the size of the training set. The proposed method addressed all these three Bagging can substantially improve clustering accuracy and yields information on the accuracy of cluster assignments for individual observations. In addition, bagged clustering procedures are more robust to the variable selection scheme, i.e. their ac-curacy is less sensitive to the number and type of variables used in the clustering. Improving and assessing the accuracy of a given clustering procedure using a resampling method is known as bagging. In supervised learning bagging is used to generate and aggrePSOte multiple clustering's. In this paper two new sampling methods BagClust1 and BagClust2 are proposed to improve and assess the accuracy of a given clustering procedure. In BagClust1 the clustering procedure is repeatedly applied to each bootstrap sample and the final partition is obtained by plurality voting. The BagClust2 method forms a new dissimilarity matrix by recording for each pair of observations the proportion of time they were clustered together in the bootstrap clusters. Nikunj C. Oza and KaPSOn Tumer entitled "Classifier Ensembles: Select Real-World Applications" classifier ensembles and ensemble applications are presented [6]. Ensuring that the particular classification algorithm matches the properties of the data is crucial in providing results that meet the needs of the particular application domain. One way in which the impact of this algorithm/application match can be alleviated is by using ensembles of classifiers, where a variety of

classifiers are pooled before a final classification decision is made. Classifier ensembles provide an extra degree of freedom in the classical bias/variance tradeoff, allowing solutions that would be difficult to reach with only a single classifier. Many learning algorithms generate a single classifier that can be used to make predictions for new examples. The way in which multiple classifiers are combined are simple averaging, weighed averaging, stacking, bagging and boosting. Robert E. Banfield, Lawrence O. Hall, Kevin W. Bowyer and W.P. Kegelmeyer entitled "A Comparison of Decision Tree Ensemble Creation Techniques" Randomization-Based technique for creating an ensemble of classifiers is proposed [7]. BAGGING is one of the older, simpler, and better known techniques for creating an ensemble of classifiers. Bagging creates an ensemble of classifiers by sampling with replacement from the set of training data to create new training sets called "bags". A number of other randomization-based ensemble techniques boosting, random subspaces, random forests, and randomized C4.5 have been introduced. In bagging, only a subset of examples typically appears in the bag which will be used in training the classifier. Out-of-bag error provides an estimate of the true error by testing on those examples which did not appear in the training set. Authors have developed an algorithm which appears to provide a reasonable solution to the problem of deciding when enough classifiers have been created for an ensemble. It works by first smoothing the out-of-bag error graph with a sliding window in order to reduce the variance. Leo Breiman entitled "Bagging Predictors" a method for generating multiple versions of a predictor and using these to get an aggregated predictor is proposed [8].

3. Proposed Method for Classification

In this paper we proposed an optimized ensemble classifier for the reduction of noise, core point and outlier in individual classifier for performing an ensemble classifier. For the process of optimization PSO are used. PSO is heuristic function and the nature of heuristic function is gets optimal result. In the process of ensemble of individual classifier combined with selection of feature vector of data. The multiple support vector machines combined with feature vector and spread of data in form of noise, core and outliers are calculated with binomial distribution function. The combined data of noise core and outlier passes through simple PSO and form a new class of COB and improved the voting ratio of ensemble classifier. For the input of PSO create a sub set of COB features set. We randomly assigned population of PSO according to selection of COB feature set. We define COB on the variable id which matrix contain the COB upper and lower value set. For the selection of COB population used velocity function given by

$$F(xi) = \frac{f(xi)}{\sum_{i=0}^{n} F(xi)} \quad \text{.....(1)}$$

Where f (xi) is the velocity of individual xi and F (xi) is the velocity of that individual Cob selected. Here in the process of PSO goal of pbsest and gbest are set. For the process of weighted velocity we fixed the value of variable velocity is p=0.07. And finally gets the optimized set of classifier. And finally gets the optimized set of classifier. Proposed model of our process shown in figure.

Figure 1: Proposed Model

Steps for algorithm

Input data set s number of classifier M

1. For d=1 to n
2. Rd=random sample from feature set
3. Md=M(Rd)

Calculate COB with binomial distribution with Ensemble optimization with Particle of swarm algorithm uses 2 objectives, i.e. minimizing error functions within each classifier (equation 1) and maximizing the ensemble value between classifier (equation 2). The calculations used as follows.

$$\sigma_t^2 = \frac{1}{N}\sum_{j=1}^{N} \sigma_{ij}^2 \quad \text{……1)}$$

Where i = 1, 2... K; K is the number of classifier

$$\sigma^2 \sum_{i=1}^{K} \sigma_i^2 \quad \text{……2)}$$

Whereas:

σ_i^2 : Error in $i -^{th}$ classifier

n_i : number of data in banyak data pada $i-^{th}$ classifier
x_{ij} : Data in i-classifier, $j-^{th}$ variabe
z_{ij} : i-classifier average in j-variable
V : Number of variable

The first function is to minimize error average in classifier which formulated as follow:

$$V(w) = \frac{1}{k}\sum_{i=1}^{k} \sigma^2 i \quad , i = 1,2 \dots \dots k \qquad 3)$$

Whereas:
V (w) : Error in classifier
k : Number of classifier

The second function is to maximize inter-classifier error which formulated as follow:

$$V(b) = \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{v}(z_{ij} - \bar{z}_j) \qquad 4)$$

V (b) : Error in classifier
z_{ij} : I-classifier average in variable
$\bar{z}$: Grand mean of $j-^{th}$ variable

A. Velocity Function with Ranking Approach

Velocity function is calculated by using pareto ranking approach, where each individual datum is evaluated by the overall population based on the non-domination concept. Next ranking approach is done by equations 5.

$$r_2(x,t) = 1 + nq(x,t) \dots \dots \dots \dots \dots \dots \dots \dots \dots .5)$$

Whereas:
$r_2(x,t)$: $x-^{th}$ Completion rank in $t-^{th}$ iteration
$nq(x,t)$: Solution number which dominate x completion in $t-^{th}$ iteration

4. Find upper and lower COB value and along with difference set weight parameter
5. Generate random population of COB matrix
6. Check velocity constraints
7. Apply random velocity v=0.07
8. Ensemble output
9. Exit

4. Experimental Result

For the process of experimental result analysis of proposed algorithm we collected 6 datasets from the UCI Machine Learning Repository. The datasets have item sizes vary from 102 to 104 and feature sizes from 4 to 102. A few datasets have missing values and we replaced them with null values. The nominal data types are changed to integers and are numbered starting from 1 based on the order of the appearance. For those dataset with multiple classes, we use class 1 as the positive class and all other classes as the negative class. We used a 10-fold cross validation for each experiment. For the total of 10 rounds of cross validation for each dataset in each experiment, we recorded the mean of the average accuracy of individual classifiers. Our all process performs in matlab 7.8.

Table 1: show that experimental value of result for five classifier on number of classifier is 10 and 10 and also show that empirical calculation of MAE, MRE and accuracy of ensemble classifier for liver data set.

Number of Classifier	Classifier	Mae	Mre	Accuracy	Execution Time
10	DT	5.45	12.39	82.85	12
20		3.725	11.76	82.85	13.58
10	KNN	6.00	14.56	87.95	13.99
20		5.0	16.93	8.95	13.51
10	SVM	4.5	12.6	92.95	13.79
20		3.5	14.81	92.95	14.29
10	SVM-COB	4.0	17.9	93.95	14.24
20		3.0	13.81	93.95	14.05
10	SVM-COB-PSO	3.89	11.6	96.35	14.71
20		2.89	12.18	96.35	14.74

Table 2: shows that experimental value of result for five classifier on number of classifier is 10 and 10 and also show that empirical calculation of MAE, MRE and accuracy of ensemble classifier for glass data set.

Number of Classifier	Classifier	Mae	Mre	Accuracy	Execution Time
10	DT	5.45	12.39	82.85	12
20		3.725	11.76	82.85	13.58

10	KNN	6.00	14.56	87.95	13.99
20		5.0	16.93	8.95	13.51
10	SVM	4.5	12.6	92.95	13.79
20		3.5	14.81	92.95	14.29
10	SVM-COB	4.0	17.9	93.95	14.24
20		3.0	13.81	93.95	14.05
10	SVM-COB-PSO	3.89	11.6	96.35	14.71
20		2.89	12.18	96.35	14.74

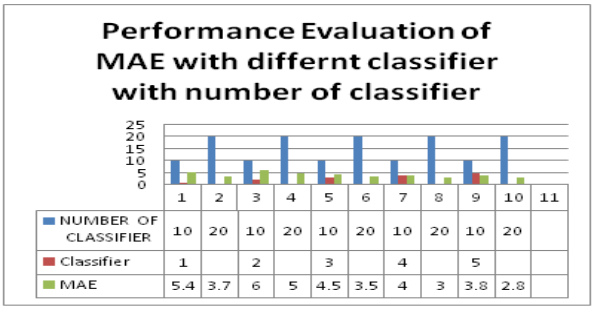

Figure 2: shows that comparative result analysis of all five classifier in terms of number of classifier and mean absolute error rate. The convention of classifier is 1,2,3,4,5 as DT,KNN,SVM,SVM-COB,SVM-COB-PSO.

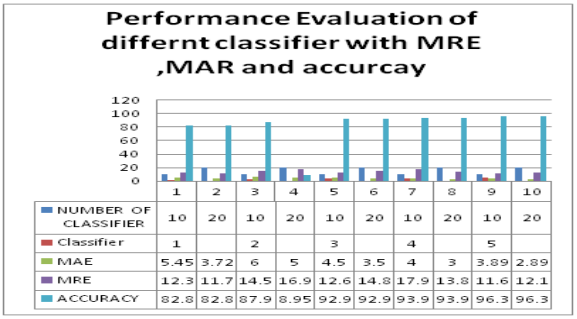

Figure 3: shows that comparative result analysis of all five classifier in terms of number of classifier, mean absolute error rate, relative error and accuracy. The convention of classifier is 1,2,3,4,5 as DT,KNN,SVM,SVM-COB,SVM-COB-PSO.

5. Conclusion and Future Work

In this paper we proposed an optimized ensemble method based on PSO. Our method combined Noise, core point and outlier of unclassified data of ensemble classifier. These three are combined together and form Cob model, these COB model passes through PSO and reduces the unclassified data improve the majority voting of classifier. Our experimental result shows better in compression of old and traditional ensemble classifier. Our experimental task performs in UCI data set such as glass, wine, liver and etc. The model is stable under different machine learning algorithms, dataset sizes, or feature sizes.

References

[1] Xueyi Wang "A New Model for Measuring the Accuracies of" in IEEE World Congress on Computational Intelligence, 2012.
[2] Brown, Gavin, and Ludmila I. Kuncheva. ""Good" and "bad" diversity in majority vote ensembles." In Multiple Classifier Systems, pp. 124-133. Springer Berlin Heidelberg, 2010.
[3] Tao, Dacheng, Tang, Xiaoou, Li, Xuelong, Wu and Xindong "Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval" in IEEE Transactions, 2006.
[4] Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew and Alex Ksikes "Ensemble Selection from Libraries of Models" 21st International Conference on Machine Learning, 2004.
[5] Sandrine Dudoit and Jane Fridlyand "Bagging to improve the accuracy of a clustering procedure" in IEEE Transcation, 2002.
[6] Nikunj C. Oza and KaPSOn Tumer "Classifier Ensembles: Select Real-World Applications" in Elsevier, 2007.
[7] Robert E. Banfield, Lawrence O. Hall, Kevin W. Bowyer and W.P. Kegelmeyer "A Comparison of Decision Tree Ensemble Creation Techniques" in IEEE Transactions, 2007.
[8] Leo Breiman "Bagging Predictors" in Kluwer Academic Publishers, 2006.
[9] Dietterich, Thomas G. "Ensemble methods in machine learning." In Multiple classifier systems, pp. 1-15. Springer Berlin Heidelberg, 2000.
[10] S. B. Kotsiantis "Supervised Machine Learning: A Review of Classification Techniques" in Informatica 30, 2007.
[11] Thomas G. Dietterich "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization" in Kluwer Academic Publishers, 1999.
[12] Guoqiang Peter Zhang entitled "Neural Networks for Classification: A Survey" in IEEE Transactions, 2000.