Efficient Clustering for Cluster based Boosting

Yogesh D. Ghait
ME Computer Engineering
Dr. D. Y. Patil College of Engineering
Ambi, Talegaon-Dabhade – India

ABSTRACT

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. Boosting is the iterative process which aims to improve the predictive accuracy of the learning algorithms. Clustering with boosting improves quality of mining process. When supervised algorithms applied on training data for learning, there may be possibility of biased learning which affects the accuracy of prediction result. Boosting provides the solution for this.It generates subsequent classifiers by learning incorrect predicted examples by previous classifier. Boosting process possesses some limitations. Different approaches introduced to overcome the problems in boosting such as overfitting and troublesome area problem to improve performance and quality of the result.Cluster based boosting address limitations in boosting for supervised learning systems. In literature Cluster based boosting [6] is used to address limitations in boosting for supervised learning systems. In paper [6], k-means is used as a clustering algorithm. Encapsulation of another clustering method with CBB may result into increase in the performance. In our proposed work we used fuzzy c means (FCM), Expectation Minimization and Hierarchical algorithm with CBB and compared the results.

Keywords

Clustering, classification, boosting ,decision tree

1. INTRODUCTION

Supervised learning algorithm may not learn training data correctly and completely. Possibility of incorrect and incomplete learning causes the prediction accuracy degradation. To overcome this issue one approach is improve the accuracy of the supervised learning algorithm iteratively is boosting. Boosting generates subsequent classifiers by learning incorrect predicted examples by previous classifier. All generated classifiers then used for classification of the test data.Adaboost is the conventional boosting algorithm, in this paper Adaboost is said as Boosting.

2. ADVANTAGES OF THE BOOSTING

Most common problem of the supervised learning algorithm is over-fitting. In over-fitting, classifier learning process starts memorizing the training data instead of learning. This happens due to high complexity of the data. If classifier memorized the training data then its prediction accuracy will be low when classifiers tested on non-training data. Literature theoretically proves that boosting is over-fitting resistant [8]. Evaluation on the different real datasets proves that boosting yields higher predictive accuracy than using single classifier [9].

3. LIMITATIONS OF THE BOOSTING

Boosting has some limitations on certain types of the data.

3.1 Boosting can’t handle Noisy label data. In this data training data is wrongly labeled. Noisy training data results into wrong learning and lowers the prediction accuracy. For example consider the scenario that first function fails to predict the instance due to noisy data learning. Boosting considers that first function learned incorrectly and focuses on mis-classified example to learn new function to classify this example correctly assuming that provided labels (which are wrong) are correct. Here boosting is learning noisy data and generates the new function by learning such examples, which will degrade the prediction accuracy.

3.2 Troublesome area- Consider in area A1 in training data F1 and F2 are relevant features. Relevant features are the features where class label of the instance depends on the values of these features. Consider another area A2 in training data with relevant features F2, F3, F4. Suppose supervised learning algorithm learns F2, F2, F3 as relevant feature. When the learned function classifies the instances which belongs to A1 area then it may predict wrong results because in area A1, F3 and F4 are the irrelevant features. In this scenario A1 is troublesome area; due to such troublesome areas boosting can’t depend on previous function to decide whether the instances are classified correctly or wrongly.

4. REASON OF THE LIMITATIONS

Boosting considersonly incorrectly classified instances for subsequent function learning. These instances holds complex structure therefore they are not classified by first function. When such examples used for learning resulting functions are complex. As from above limitations it is clear that it is difficult to depend on the first function to decide the correctness of the instances due to noisy label and troublesome areas. At the same time, the training process for these subsequent functions tends to ignore problematic training data on which the initial function predicted the correct label.

5. RELATED WORK

[1]Boosting algorithms focuses on inaccurate classified instances for subsequent function learning.With a growing size of classifiers boosting usually does not overfitthe training data [1]. Schapire et al. attempted to explain this in terms of the margins the classifier achieves on training examples. Margin is a quantity that measured as the confidence in theprediction of the combined classifier. This paper focuses on Breiman’s arc-gv algorithm for maximizing margins. Further it explains why boosting is resistant to overfitting and how it refines the decision boundary for accurate predictions.

[2]For classification problem, boosting proves to be efficient technique and provides better results. Boosting a simple clustering algorithm provides improved and robust multi-clustering solutionwhich improves quality of the partitioning.

To provide consistent partitioning of datasetwhich is required for clustering,this paper proposes a new clustering methodology the boost-clustering algorithm. The new algorithm is a multi-clustering methodbased on general principles of boosting. Proposed approach used basic clustering algorithm in iterative way and aggregation of the multiple clustering results by using weighted voting. Experiments show that this approach is better and efficient.

[3]When complexity of data increases, classifiers start memorizing the data instead of learning which results in low prediction accuracy. This is a problem of overfitting. Noise data influence boostingleads to overfitting. As described in [3] when bayesian error is not zerothat is for overlapping classes, standard boosting algorithms are not appropriate thus results in overfitting. This can be avoided by removing confusing samples misclassified by Bayesian classifier. This paper proposes an algorithm which removes confusing samples. Results show that removing confusing samples helps boostingtoreduce generalization error and avoid overfiiting.

[4]Adaptive boosting algorithm is important and popular among the classification methodologies. With low noisy data overfitting rarely present with Adaboost. High Noisy data affects the Adaboost which causes overfitting problem. This paper studies Adaboost and proposed two regularization schemes from the viewpoint of mathematical programming to improve the robustness of AdaBoostagainst noisy data.Forcontrolling the distribution skewness in the learning process to preventthe outerlier samples from spoiling decision boundaries, paper introduced a penalty scheme mechanism.By using two convex penalty functions, two soft margin concepts i.e. two new Adaboost algorithms are defined.

[5]Ensemble classifiers are used to increase predictive accuracy with respect to the base classifier. First base classifiers are generated and combined to generate ensemble classifiers and this is achieved through boosting.Boosting improves the performance and predictive accuracy of learning algorithms in machine learning. Boosting process combinesweak classifiers to produce strong classifiers.Paper[5] Contains comprehensive evolution and evaluation of Boosting on various criteria (parameters) with Bagging. It shows that Boosting has superior prediction capabilities than bagging as classifies the samples more correctly.

[6]Paper proposed a novel CBB (Cluster Based Boosting) approach that partitions the training data into clusters and these clusters contains highly similar member data, and then integrates these clusters directly into the boosting process. This paper applies selective boosting strategy on each cluster based on previous function accuracy on member data and additional structure provided by the cluster.Paper appliesmethod of clustering the training data to improve subsequent functionsand helps boosting process with high prediction accuracy.Selective boosting approach uses high learning rate, low learning or no boosting strategy on each cluster.Proposed scheme addresses two specific problems, one is filtering subsequent functions when data has noisy label and troublesome area; second is overfitting in subsequent functions.

6. VARIOUS CLUSTERING ALGORITHMS[7]

Clustering is the task of partitioning the data set into highly similar group where member in each group is highly similar within the group (in some sense or other) and differs from other group members. Clustering is vital task in machine learning. Different clustering methods are available in literature.

The COWEB algorithm introduced for clustering objects in an object-attribute data set. The COBWEB algorithm generates hierarchical clustering, where clustersare described probabilistically. For the tree construction, COWEB uses a heuristic measure called category utility. On the basis of this heuristic measure splitting and merging of classes can be done which allows COWEB to do bidirectional search while K-means is unidirectional. COWEB has some limitations. It is expensive to update and store the clusters because of probabilitydistribution representation of clusters. Also it is complex in terms of time and space for skewed data input as classification tree is not height balanced.

DBSCAN is the density based clustering algorithm which finds the number of clusters starting from the estimated density distribution of corresponding nodes. As contrast to k-means DBSCAN does not need to know the number of clusters in the data initially. It can find clusters completely surrounded by (but not connected to) a different cluster.

Algorithm uses Euclidean distance measure which is useless in case of high dimensionality.

Farthest First algorithm variant of K means. This algorithm places each cluster center in turn at the point furthest from the existing cluster centers. This point must lie within the data area. Algorithm helps to boost up the clustering process. Farthest-point based heuristic method is fast and suitable for large-scale data mining applications.

K-Means clustering algorithm:In data mining, *k*-means clustering aims to partition *n* observationsinto *k* clusters where each observation belongs to thecluster with the nearest mean and it is a simple unsupervised learning algorithm. With a large number of variables, K-Means may be computationally faster and produce tighter clusters than above mentioned techniques.But for different initial partitions values of K affect outcome. K may be difficult to predict because of fixed number of clusters and it does not work well with non-globular clusters.

[9] Among the clustering methods which are fuzzy, fuzzy c means (FCM) is renowned method because it is advantageous and avoids ambiguity. It also has ability to maintain more information as compare to any other hard clustering methods. It has various applications in areas such as image clustering and segmentation, pattern recognition etc.

[10] Expectation Maximization (EM) is used for finding the maximum likelihood estimates of parameters and it is an iterative method. EM algorithm has two steps first is expectation stage related to unknown variables by using estimation of parameters and other is maximization step provides new estimation of the parameters.

[11] Hierarchical clustering is a popular tool for data analysis and it used to construct the binary trees that integrates similar group of points. It is able to provide better summary of data by its construction and it imposes hierarchical structure on the data. Hierarchical clustering works in two ways first is Agglomerative which is bottom-up approach and another is Divisive which is top-down approach.

In literature, various boosting methodologies are available; paper [6] is addresses limitations we discussed above in boosting for supervised learning systems. We observed thatfor clustering k-means is used and there is scope to find and combine best suitable algorithm with CBB. For better performance encapsulation of another clustering method with CBB can be used. In our proposed work we used fuzzy c means (FCM), Expectation Minimization and Hierarchical algorithm with CBB and compared the results.

7. PROPOSED WORK

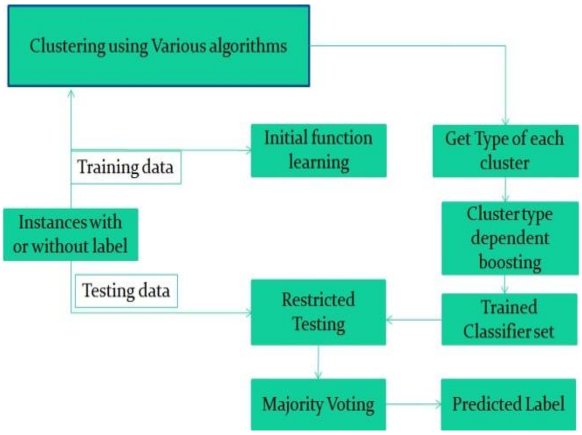
Clustering using Various algorithms
Initial function learning
Get Type of each cluster
Training data
Instances with or without label
Cluster type dependent boosting
Testing data
Restricted Testing
Trained Classifier set
Majority Voting
Predicted Label

Fig 1: Proposed System

The Cluster Based Boosting solution is based on unsupervised clustering that tries to decompose or partition the training data into clusters where the member instances in a cluster are similar to each other and as different as possible from members in other clusters. The training data is broken into the cluster. During this process, CBB computes the BIC (Bayesian Information Criterion) for the set of clusters. Second, CBB chooses the set of clusters with the lowest BIC. Third, CBB learns the initial function using all the training data.

After clustering by using one of the techniques among fuzzy c means, expectation maximization or hierarchical algorithm, CBB performs selective boosting based on the cluster type. Clusters are categorized based on following four categories.

- Heterogeneous struggling (HES): The cluster contains members with different labels and previous functions struggle to predict the correct labels. Since such a cluster generally contains troublesome training data and previous functions have been struggling, CBB uses boosting with a high learning rate (high-eta boosting) on this type. Learning subsequent functions focusing on incorrect members until accuracy improves.
- Heterogeneous prospering (HEP): The cluster contains members with different labels, but previous functions are still able to predict the correct label. CBB uses boosting with a low learning rate (low-eta boosting) on this type—learning fewer subsequent functions focusing on incorrect members.
- Homogenous struggling (HOS): The cluster contains members with a single label, but the previous functions struggle to predict the correct labels. Since this type is easy for a function to predict (simply by predicting the majority label), CBB learns a single, subsequent function on all members without boosting on incorrect members.
- Homogenous prospering. The cluster contains members with predominately a single label and the previous functions already predict the correct label for most of the members. CBB does not learn any subsequent functions on this type.to prevent those functions from learning the label noise.

After this process we get the subsequent functions i.e. trained set of classifiers. There are two different ways that these subsequent functions can be used: restricted and unrestricted. Restricted counts the subsequent functions learned on the cluster to which the new instance would be assigned and disregards votes from other clusters. It is more consistent with the selective boosting on each cluster.

Then a weighted vote of these subsequent functions is calculated and assigned to each of them which will be further used to predict the labels for a new instance.

8. CONCLUSION

In this paper, we discussed various boosting problem and proposed solutions and also described some clustering techniques. Use of boosting is advantageous for more accurate results in machine learning. Cluster based boosting approach addresses limitations in boosting on supervised learning algorithms.In order to performance enhancement in our work,weintegrate the boosting methodology with fuzzy c means (FCM), Expectation Minimization and Hierarchical algorithm which are different available clustering techniques and analyzeditthe outputted results.

9. REFERENCES

[1] L. Reyzin and R. Schapire, “How boosting the margin can also boost classifier complexity,” in Proc. Int. Conf. Mach. Learn., 2006, pp. 753–760.

[2] D. Frossyniotis, A. Likas, and A. Stafylopatis, “A clustering method based on boosting,” Pattern Recog. Lett., vol. 25, pp. 641–654, 2004.

[3] A. Vezhnevets and O. Barinova, “Avoiding boosting overfitting by removing confusing samples,” in Proc. Eur. Conf. Mach. Learn., 2007, pp. 430–441.

[4] Y. Sun, J. Li, and W. Hager, “Two new regularized adaboostalgorithms,”in Proc. Int. Conf. Mach. Learn. Appl., 2004, pp. 41–48.

[5] A. Ganatra and Y. Kosta, “Comprehensive evolution and evaluation of boosting,” Int. J. Comput.Theory Eng., vol.2, pp. 931–936, 2010.

[6] Cluster-Based Boosting,”L. Dee Miller and Leen-KiatSoh, Member, IEEE , 2015

[7] “Boosting: Foundations and Algorithms,” Rob Schapire.

[8] W. Gao and Z-H. Zhou, “On the doubt about margin explanation of boosting,” Artif.Intell., vol. 203, pp. 1–18, Oct. 2013

[9] Yinghua Lu, Tinghuai Ma1, ChanghongYin2 ,Xiaoyu Xie2 , Wei Tian and ShuiMingZhong, “Implementation of the Fuzzy C-Means Clustering Algorithm in Meteorological Data” 2013

[10] Expectation Maximization Algorithm, IEE signal Processing, 2006.

[11] Ryan Tibhsirani,”Hierarchical Clustering,” 2013.