ISSN XXXX XXXX © 2017 IJESC

Research Article Volume 7 Issue No.8

Clustering and Boosting in Data Mining

E.Suriyapriya[1], M.Praveena, M.Phil[2]
M.Sc Student[1], Assistant Professor[2]
Department of Computer Science
Dr.SNS Rajalakshmi College of Arts & Science, Coimbatore, India

Abstract:
In this paper, a novel approach for Developing Cluster and booster on the basis of Data Mining is discussed. Clustering with boosting improves quality of mining process. Boosting is the iterative process which aims to improve the predictive accuracy of the learning algorithms. Our objective is to use a feature selection technique which will remove irrelevant features and redundancies from the available dataset. 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. Use of boosting in many applications proved its effectiveness. Cluster based boosting is used to address limitations in boosting for supervised learning systems.

Keywords: Data preprocessing, Clustering, Boosting.

I. INTRODUCTION

The data mining can be divided by their learning process or representation of extracted knowledge. Support vector machine neighbor classifiers, and Probability based classifiers. Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters. 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. Help users to understand the natural grouping or structure in a data set. In general, It can be divided into two categories. First, each member of the ensemble is generated independent of the other models, usually with different initial conditions or parameters. The second category could be considered as the clustering counterpart of the boosting-based methods in classification domain Used either as a stand-alone tool to get insight into data distribution or as a preprocessing step for other algorithms. Boosting means, once learning process is completed or classifier is learned, boosting generates subsequent classifiers by learning incorrect predicted examples by previous classifier. Boosting generates subsequent classifiers by learning incorrect predicted examples by previous classifier. All generated classifiers then used for classification of the test data here any member of the ensemble of classifiers are trained sequentially to compensate the short comings of the previously trained models, usually using the notion of sample weights.

II. DATA PREPROCESSING

Data preprocessing is often unused but it is very important in data mining process. The phrase "scrap In, Scrap Out" is particularly applicable to data mining and machine earning. Since the number of the partitions is fixed for the base models, we first count the number of shared samples between clusters of both the first model and the next one in the ensemble. Analyzing data that has not been thoroughly screened for such problems can produce distorted results. Thus, delegation and quality of data is first and foremost before running and analysis.

- Removal of noise and outliers - Will improve the performance of mining
- Sampling is employed for data selection - Processing entire Data might be expensive
- Dealing with High-dimensional data
  - Curse of dimensionality
  - Data Normalization
  - Different features have different range values - e.g. human age, height, weight.
- Feature Selection - Remove unnecessary features – redundant or irrelevant

III. CLUSTERS IN DATA MINING

First, there has been considerable previous work on using boosting to improve clustering. An Agglomerative hierarchical method places each sample in its own cluster and gradually merges these clusters into larger clusters until all samples are ultimately in a single cluster. First, the clusters created provide additional structure for the subsequent functions since these clusters include both correct and incorrect instances from previous functions. A cluster is a subset of objects which are "similar". A subset of objects such that the distance between any two objects in the cluster is less than the distance between any object in the cluster and any object not located inside it. Cluster works can be defined as

- No class Labels – so, no prediction
- Groupings in the data (descriptive)
- Can be used to summarize the data
- Can help in removing outliers and noise
- Image segmentation, document clustering, gene expression data etc..

A connected region of a multidimensional space containing a relatively high density of objects. A good clustering method will produce high quality clusters in which:
• The intra-class that is, intra-cluster likeness is high.
• The inter-class similarity is low.

Here are many different method and algorithms for clustering in data mining:
• For numeric and/or symbolic data
• Exclusive vs. overlapping
• Crisp vs. soft computing paradigms
• Hierarchical Vs. flat (non-hierarchical)
• Access to all data or incremental learning

• Semi-supervised mode
• Algorithms also vary by:
• Measures of similarity
• Linkage methods
• Computational efficiency

IV. BOOSTING IN DATA MINING

Boosting generates successive classifiers by learning wrong 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. Cluster based boosting address limitations in boosting for supervised learning systems. Boosting is a machine realizing en same for mostly reducing bias, and also difference in supervised learning, and a family of machine learning algorithms which disciple weak learners to strong ones. The final clustering solution is produced by compilation the obtained partitions using weighted voting, where the weight of each partition is a measure of its quality. Boosting is based on the question posed by Kern sand V alien. Can a set of weak learners create a single strong student? A weak learner is clear to be a classifier which is only slightly correlated with the true classification it can label examples better than random guessing. Further extensions to margin theory have examined how the margin distribution (including margin average and variance) is connected to the predictive accuracy. In contrast, a strong learner is a classifier that is arbitrarily well-correlated with the truefrom:editor@ijesc.org classification. Boosting can be categories in two methods. They are Binary categories and Multi-Class categories.

Boosting for binary categorization
We use Ada Boost for face detection as an example of Binary categories. The two categories are faces versus background.

- ✓ Form a large set of simple features
- ✓ Initialize weights for training images

Boosting for multi-class categorization: Compared with binary categorization, Multi-class categorization looks for common features that can be shared across the categories at the same time. They turn to be more generic edge like structure. During learning, the detectors for each category can be trained jointly. Compared with training discretely, it generalizes better, needs less training data, and requires less number of features to achieve same performance.

V. 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. Literature theoretically proves that boosting is over-fitting resistant. 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. Evaluation on the different real datasets proves that boosting yields higher predictive accuracy than using single classifier. The specialized inputs, synergies, and increased access to information/ public goods that accompany cluster manufacturing all boost productivity of plants.

VI. CONCLUSION

In this paper, we discussed various boosting problem and proposed solutions and also described some clustering techniques In order to performance enhancement in our work, we integrate the boosting methodology with fuzzy c means (FCM), Expectation Minimization and Hierarchical algorithm which are different available clustering techniques and analyzed it the outputted results. Our CBB approach partitions the training data into clusters containing highly similar member data and integrates these clusters directly into the boosting process. Use of boosting is advantageous for more accurate results in machine learning. Now that we have established the feasibility and effectiveness of CBB, we intend to continue our investigation down several avenues. Other interesting future research issues concern the specification of alternative ways for evaluating how well a data point has been clustered, as well as the experimentation with other types of basic clustering algorithms.

VII. REFERENCES

[1]. Amir Saffari and Horst Bischof, "Clustering in a Boosting Framework," Institute for Computer Graphics and Vision, Graz University of Technology in 2005

[2]. L. Dee Miller and Leen-Kiat Soh, "Cluster-Based Boosting", in June 2015.

[3]. Yogesh D. Ghait, "Efficient Clustering for Cluster based Boosting", in 2016.

[4]. D. Frossyniotis, A. Likas b, A. Stafylopatis, "A clustering method based on boosting" in 2004.

[5]. Rutuja Shirbhate and Dr. S. D. Baber, "Cluster based boosting for high dimensional data", in July 2016.