ISSN: 2277-9655
Impact Factor: 3.785

IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

CLUSTER BASED BOOSTING TECHNIQUES FOR ACCURACY PREDICTION

L.Hemalatha*, Dr.I.Elizabeth Shanthi
*Research Scholar, Department of Computer Science, Avinashilingam University for Women Coimbatore, Tamilnadu -641043, India
Associate Professor, Department of Computer Science Avinashilingam University for Women Coimbatore, Tamilnadu -641043, India

DOI: 10.5281/zenodo.51531

ABSTRACT
In recent years encroachment of social media in human life is inexplicable. The greatest and most significant issue in latest technology is the fast retrieval of information from large databases. To overcome this, different techniques have been developed. Improving the retrieval process and its accuracy is a vital factor in the predictive analyses of Machine learning. Boosting is one of the techniques used for enhancing the accuracy of prediction in supervised learning. In this process, lot of noise data may arise which adds another problem in handling training data set. Extracting accurate data from the training dataset is an essential factor in applying clustering techniques. This paper surveys the various boosting/clustering techniques and compares them with a motive to suggest a better environment which is free from noise/errors so as to have maximum accuracy on the output data.

KEYWORDS: Data Mining, Clustering, Predictive, Boosting and accuracy.

INTRODUCTION
Boosting is the process used in machine learning algorithms which is used to enhance the improved accuracy. It works with many functions consecutively focusing on occurrences which are incorrect. In supervised learning system, boosting process converts weak learners to strong one. But when difficult functions over fit the data, boosting still has difficulties on certain data sets which contain label noise and troublesome areas. In reality, over fitting generally occurs when a model is excessively complex, such as having too many parameters relative to the number of observations. Boosting is an iterative process of machine learning ensemble Meta algorithm which is used to achieve improved predictive accuracy for functions that learns multiple functions using training data in the same supervised learning system. It then predicts the label for new data instances using a weighted vote over all the functions. By combining multiple functions together, boosting achieves more refined decision boundary on the training data than using a single function [1]. Boosting is provided to find a highly accurate prediction rule which is a difficult task. As boosting is an iterative process and each time when it calls the base learning algorithm, it generates a new weak prediction rule. After several iteration, the boosting algorithm will combine these weak rules into a strong prediction rule. This rule will be much more accurate than any one of the weak rules generated. In spite of its success, boosting such as adaboost has various problems. They are [2],

- Training data contains label noise where the labels of the instances provided are actually wrong.
- Training data with troublesome areas where the relevant features of the instances are different from the rest of the training data and
- Filtering out the data in subsequent functions when the training data contains troublesome areas and/or label noise.

In boosting, every learner learns the complete training data in which the learners are dependent on one another and are biased towards the data that are incorrectly predicted in the previous iteration. The training data contains label noise and so the boosting learning function also learns incorrectly. Even though the initial function is correct, it does not realize that the labels were incorrect. Hence boosting focuses subsequent functions on learning how to "correctly"

predict these instances assuming that the wrong labels provided are correct. In boosting, over fitting occurs due to overlapping regions. The samples that are located in the overlapping region are more likely to be classified unevenly. Adaboost works by increasing the weight of samples that are misclassified in the previous iterations. Due to the way it learns the subsequent function, Boosting works by filtering out some correctly classified instances and withheld the incorrect instances in the subsequent iterations. Boosting is an efficient method which converts any weak learning algorithm into strong learner in order to improve the accuracy. The limitations in boosting is to over fit on the training data and to filter out the correct data in the subsequent function as it concentrates on regions not predicted well by other learners [3].

Clustering is one of the techniques which are used in several exploratory pattern analysis, grouping, decision making and machine learning situations including data mining, document retrieval, image segmentation and pattern classification. However in many such problems, there is little prior information which is available in the data and the decision maker must make a few possible assumptions about the data. There are some restrictions in clustering methodology that is particularly appropriate for the exploration of interrelationships among the data points to make an assessment perhaps preliminary of their structure. Many research communities make use the term clustering to describe the required methods for grouping of unlabeled data. These communities have different types of terminologies and assumptions for the components of the clustering process and the contexts in which clustering are used. Thus we face a dilemma regarding the scope of this survey [4]. The production of a truly comprehensive survey would be a huge task given the sheer mass of literature in this area. The approach of the survey might also be questionable given the need to reconcile very different vocabularies and assumptions regarding clustering in the various communities. With the help of the cluster object in the dataset the predictive analyses method can be over taken for analyzing the cluster data with the given attributes and to find out the accurate result in the dataset by the help of the boosting technique [5]. This paper provides the cluster boosting object with various analyses of technique that which technique shows the accurate result.

LITERATURE REVIEW

P. Velvizhy, S. Abayambigai, and A. Kannan [6] proposed a general method for boosting X-means algorithm which learns any weak learning algorithm and convert them into strong learner to improve the accuracy. To address the limitations of boosting such as over fitting on the training data and filtering the correct data, cluster based boosting (CBB) approach is used. Initially X-means algorithm is used to cluster the data and the clusters are selectively boosted based on the structure information additionally provided by clusters and previous function accuracy on the member data. Cluster Based Boosting can be applied to the high dimensional data, with the help of dimensionality reduction technique. Global Redundancy Minimization frame work was applied which considers the redundancy of the feature with all other features. The features are selected to contribute more mutual information for prediction. This frame work can be performed with feature selection techniques, variable selection techniques and attribute selection techniques. The experimental results are tested on various dataset. These results compare the efficiency of Global redundancy framework and Cluster Based Boosting with Global redundancy minimization framework than classifier with global redundancy minimization framework.

L. Dee Miller et al [7] discussed the problem of boosting technique that boosting cannot handle the data which contains noise and troublesome area. Noisy data is the training data that is incorrectly labeled and troublesome areas are areas of instances whose relevant features are different from the training data. Author of the paper observed that boosting the incorrectly predicted data for subsequent function learning creates some problem and to overcome this boosting method is used for predicted data by the initial function for subsequent function learning. This paper discusses about the partition of the training data into clusters and then integrates them directly in to the boosting process. Cluster based boosting applied selective boosting strategy that includes high learning rate, low learning rate and not boosting strategy on each cluster according to the previous function accuracy on member data. When data has noisy label and troublesome area, the problem of over fitting in subsequent functions, filtering subsequent functions are addressed.

A. Ganatra and Y. Kosta [8] studied the ensembles of classifiers are obtained by generating and combining base classifiers, which are constructed using suitable machine learning methods. The target of these ensembles is to increase the predictive accuracy with relevant to the base classifiers. One of the most popular methods for creating ensembles is boosting and adaboost is the most prominent member. Boosting is a general approach for improving classifier

ISSN: 2277-9655
Impact Factor: 3.785

performances. Boosting is an eminent method in the machine learning community for improving the performance of any learning algorithm. It refers to the problem of producing a very accurate prediction rule by combining rough and moderately inaccurate rules-of-thumb. It resembles Bagging and other committee based methods. Many weak classifiers are combined to produce strong classifiers. Sequentially apply weak classifiers to modified versions of data. Predictions of these classifiers are combined to produce a powerful classifier.

Reshma Y et al [9] analyze the Boosting is the repetitive process to perk up the accuracy in functions for prediction that supervised learning system learn using training data. In this prediction process, boosting considers multiple functions rather that considering only single function from the same supervised learning system. Using a weighted vote over all the functions, boosting process predicts the label for new data instances. By considering and merging two or more functions together , boosting prepare to get fine grained decision boundary on training data than using single function. Boosting for supervised learning having certain limitations like e.g. because of problematic data difficulty arises in analyzing the data , over-fitting of training data , wrong label prediction by initial function etc. Previous works reveal that boosting is resistant to over fitting problem. Also in case of wrong label prediction from function, boosting improves higher accuracy when multiple functions are used to decide the labels for clusters. Hence there must be proposed system that works on these problems. Also clustering can be achieved on problematic dataset and cluster based boosting (CBB) approach should be adopted to achieve this. Along with CBB, the outlier detection shall be achieved so that data will be easy to analyze and smart clusters can be formed.

CLUSTERING IN BOOSTING

Boosting is a method to improve the accuracy of any given machine learning algorithm. Boosting predicts the label for new data instances using a weighted vote over all the functions [7]. By combining multiple functions together, boosting achieves a more refined decision boundary on the training data than using a single function. Clustering is a method to group the data in a field. It helps to render the data in the similar group, which is used for grouping the similar objects [10, 11]. With the help of the cluster object in the dataset the predictive analyses method can be over taken for analyzing the cluster data with the given attributes and to find out the accurate result in the dataset by the help of the boosting technique. This paper surveys the various clustering method for boosting the object with various analysis for showing the accurate result.

A. Hierarchical Based Boosting

Hierarchical clustering is a method of cluster analysis which constructs a hierarchy of clusters. It is the connectivity based clustering algorithms. The hierarchical algorithms construct clusters gradually by step by step. In hierarchical clustering, the data are not partitioned into a particular cluster with in a single step. It takes a series of partitions, in which a single cluster containing all objects to n clusters each has a single object. With the help of this cluster it sub divided into many groups, which value is belongs to which group. This is based on the tree structure the data are stored or grouped in the manner of hierarchy order [12]. This method helps to boost the data from the given dataset and verify the accurate data. Here the drawback is it contains the duplicate data in another hierarchical tree.

B. Spectral Clustering Algorithm

The spectral clustering algorithm is an algorithm for grouping N data points in an I-dimensional space into several clusters. Each cluster is parameterized by its similarity, which means that the points in the same group are same and points in different groups are different to each other. Spectral clustering is appealingly simple: Given some data, you build an affinity (or kernel) matrix, analyze its spectrum, and frequently get a perfect clustering from the free dominant eigen vectors [13]. This simple algorithm or its slightly multiple variants which yield so good results are widely appreciated for applications. Spectral clustering is a popular method to perform data clustering as one of the most basic tasks of machine learning. These methods possess some important advantages, such as the ability to cluster non-vector data, and often yield superior empirical performance. In spectral clustering algorithm with computational complexity linear in the number of data points that is directly applicable to large-scale datasets [14]. In this algorithm it boost the data in parameterized by its similarity and group it's by using the objects of the parameter which is given by the initial level. Using this method it can track the non-vector data.

C. Density-Based Clustering

In density-based clustering, clusters are defined as areas of higher density than the remaining data set. The sparse areas objects - that are required to separate clusters - are usually considered to be more noise and border points. There are two major approaches for density-based methods. The first approach pins density to a training data point and is reviewed in the sub-section Density-Based Connectivity. In this clustering technique density and connectivity are measured in terms of local distribution of nearest neighbors. So defined density-connectivity is a uniform relation and all the points reachable from core objects can be factorized into maximal connected components serving as clusters [15, 16]. It boosts the data based on the training tuple.

D. Grid Based Algorithms

In Grid-based clustering the data space is quantized into finite number of cells which form the grid structure and perform clustering on the grids. Grid based clustering arranges the infinite number of data records in data streams to finite numbers of grids. This clustering has the fastest processing time that typically depends on the size of the grid instead of the data. This methods use the single uniform grid mesh to partition the entire problem domain into cells and the data objects located within a cell are represented by the cell using a set of statistical attributes from the objects. As these algorithms have a fast processing time, they go through the data set once to compute the statistical values for the grids and the performance of clustering depends only on the size of the grids which is usually much less than the data objects. All these methods use a uniform grid mesh to cover the whole problem to boost the data [17, 18]. Based on the problems with highly irregular data distributions, the resolution of the grid mesh must be too fine to obtain a good clustering quality. It significantly increases the computation load for clustering ad it provide an accurate result when compared to the above mentioned.

CONCLUSION

For large data sets, analyzing and acquiring the needed information is a crucial factor. Irrelevant duplication of data occurs in many level; those data are the noise data, which is unwanted one. Boosting is one of the techniques to find out the accurate data from the large dataset. For finding out the data clustering of data mining technique is used. With the combination of the clustering and boosting we can get accurate results. Many clustering approaches are applied for the boosting the predict data with the help of hierarchy, spectral, Density and Grid methods. From analysis done Grid based clustering approach provides a better result for boosting the accurate predicted data. The future work may be extended by comparing different types of clustering methods.

REFERENCES

[1]. R. Schapire and Y. Freund, Boosting: Foundations and Algorithms. Cambridge, MA, USA: MIT Press, 2012.
[2]. D. Mease and A. Wyner, "Evidence contrary to the statistical view of boosting," Journal of Machine Learning Research, vol. 9, pp. 131–156, Feb. 2008.
[3]. A.Vezhnevets and O. Barinova, "Avoiding boosting overfitting by removing confusing samples," in Proc. Eur. Conf. Mach. Learn., 2007, pp. 430–441.
[4]. G. P. Babu, M. N. Murty, "Clustering with Evolution Strategies", *Pattern Recognition*, vol. 27, pp. 321-329, 1994.
[5]. D. Frossyniotis, A. Likas, and A. Stafylopatis, "A clustering method based on boosting," Pattern Recog. Lett., vol. 25, pp. 641–654, 2004
[6]. P. Velvizhy, S. Abayambigai, and A. Kannan, "Enhanced Cluster Based Boosting in High Dimensional Data".
[7]. L. Dee Miller and Leen- Kiat Soh, Member, IEEE "Cluster-Based Boosting".
[8]. A.Ganatra and Y. Kosta, "Comprehensive evolution and evaluation of boosting," Int. J. Comput. Theory Eng., vol. 2, pp. 931–936, 2010.
[9]. Ms. Reshma Y. Nagpure, Prof. P. P. Rokade, "Refined Clustering technique based on boosting and outlier detection".
[10]. B. Wu and R. Nevatia, "Cluster boosted tree classifier for multiview, multipurpose object detection," in Proc. IEEE Int. Conf. Comput.Vis., 2007, pp. 1–8.
[11]. Pavel Berkhin, "A Survey of Clustering Data Mining Techniques", pp.25-71, 2002.

[12]. Cheng-Ru Lin, Chen, Ming-Syan Syan , "Combining Partitional and Hierarchical Algorithms for Robust and Efficient Data Clustering with Cohesion Self-Merging" IEEE Transactions On Knowledge And Data Engineering, Vol. 17, No. 2,pp.145-159, 2005.
[13]. Auffarth B. (2007) Spectral Graph Clustering. Retrieved February12, 2013
[14]. Tung F., Wong A. & Clausi D. (2010) Enabling Scalable Spectral Clustering for Image Segmentation.
[15]. Zheng Hua, Wang Zhenxing, Zhang Liancheng, Wang Qian, "Clustering Algorithm Based on Characteristics of Density Distribution" Advanced Computer Control (ICACC), 2010 2nd International Conference on National Digital Switching System Engineering & Technological R&D Center, vol2", pp.431-435, 2010.
[16]. Pragati Shrivastava, Hitesh Gupta. "A Review of Density-Based clustering in Spatial Data", International Journal of Advanced Computer Research (ISSN (print), pp.2249-7277, September-2012.
[17]. MR ILANGO, Dr V MOHAN, "A Survey of Grid Based Clustering Algorithms", International Journal of Engineering Science and Technology, pp.3441-3446, 2010.
[18]. Gholamreza Esfandani, Mohsen Sayyadi, Amin Namadchian, "GDCLU: a new Grid-Density based CLUstring algorithm", IEEE 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, pp.102-107, 2012.