ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 8, August 2016

Accurate Prediction of Labels by Boosting the Cluster

[1]Dr.P.Kalyani [2]N.Meenatchi
[1]Associate professor [2]M.Phil. Research Scholar,
[1]Dept of Information Technology, [2] Department of Computer Science,
[1][2]S.N.R Sons College, Coimbatore
[1]Kal.y@rediffmail.com[2]meenuraj02@gmail.com

***Abstract*:--The huge amount of data springs up naturally in various domains, which confronts a great challenge for the tralatitious data mining techniques in terms of efficiency and effectiveness. In order to achieve accurate information from the collected data various techniques gets evolved. 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. It is widely recognized that the boosting methodology provides superior results for classification problems. Boosting process possesses some limitations. Different approaches introduced to overcome the problems in boosting such as over fitting and troublesome area problem to improve performance and quality of the result. Cluster based boosting address limitations in boosting for supervised learning systems. In this paper, we propose the boost-clustering algorithm which constitutes a novel clustering methodology that exploits the general principles of boosting in order to provide a consistent partitioning of a dataset. The methodology is implemented in dot net and the experimental results show that the proposed methodology supports data in various environments even in presence of noise. The good performance in clustering the data gets obtained from large data set effectively.**

***Key words:* -- boosting, clustering, effective label, high dimensional data.**

I. INTRODUCTION

Data clustering, also called unsupervised learning, is one of the key techniques in data mining that is used to understand and mine the structure of unlabeled data. The idea of improving clustering by side information, sometimes called semi-supervised clustering or constrained data clustering, has received significant amount of attention in recent studies on data clustering. Often, the side information is presented in the form of pairwise constraints: the must-link pairs where data points should belong to the same cluster and the cannot-link pairs where data points should belong to different clusters [1]. There are two major approaches to semi-supervised clustering: the constraint-based approach and the approach based on distance metric learning. The first approach employs the side information to restrict the solution space, and only finds the solution that is consistent with the pairwise constraints. The second approach first learns a distance metric from the given pairwise constraints, and computes the pairwise similarity using the learned distance metric. The computed similarity matrix is then used for data clustering.

Classifiers in the data mining can be divided by their learning process or representation of extracted knowledge. Support vector machine (SVM), decision trees like ID3, C4.5, k-nearest neighbor classifiers, and Probability based classifiers like Naive Bayes. Boosting means, once learning process is completed and classifier is learned, boosting generates subsequent classifiers by learning incorrect predicted examples by previous classifier. All generated classifiers then used for classification of the test data. Feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in learning.Data Dimensionality is crucial for learning and prediction systems. Term Curse of High Dimensionality means when data becomes more dimensional, complexity in learning increases. Boosting may face an issue Accuracy degradation due to irrelevant features. This can be overcome by implementing in the cluster process to predict the accurate label for the data [2].

Another problem with boosting is 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. This can

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 8, August 2016

result in complexity and higher probability of overfitting because some correctly predicted instances may be similar to the incorrectly classified instances in the heterogeneous region. To address the limitations of boosting, we propose a novel approach called cluster based boosting (CBB). CBB uses clusters to partition data and then the boosting is applied on the clusters containing highly similar data which helps to reduce the complexity and to mitigate the over fitting.

Based on these facts, we present in this paper a boosting based clustering algorithm which builds forward stage-wise additive models for data partitioning and overcomes previously explained problems in a theoretical framework. Needless to say, data clustering plays an important and essential role in many computer vision applications. For example in unsupervised or weakly-supervised object recognition problems, the visual words (or parts) are usually constructed by clustering a set of descriptor responses to some selected image regions (usually extracted by interest point/region detectors), for example refer to [3] and references therein.

In Section 3 we will explain the works related to our proposed idea. In section 4 the methodology of proposed work gets discussed and the experimental results are shown in section 5.

II. RELATED WORKS

L. Reyzin and R. Schapire clarified that boosting method gets prepared on mistaken ordered occurrences for consequent capacity learning [4]. It is acknowledged in this paper the boosting does not for the most part over fit the preparation data even with the classifiers with broad size. It is elucidated using edges the classifier finishes on preparing case by Schapire et al. Edge contains to the sureness of the consistency of the aggregated classifier. This paper analyzed Breiman's bend gv computation for expanding edges moreover it clears up why boosting is impenetrable to over fitting and how it refines as far as possible for careful expectations.

N. Tomasev and D. Mladenic has been projected that Hubness information k- Nearest Neighbor (HIKNN) for overseeing high dimensional information. HIKNN rule was compared with alternative previous hubness based algorithm. Hubs could be a data point that often occurred in k-nearest neighbor list and barely occurring points or might outliers referred as anti-hubs. The search for nearest neighbor is a very vital aspect in clustering algorithm. The k-nearest neighbor algorithm is the essential strategy for easy to discover the closest neighbor. It is comprehensively utilized as a characterization technique and exceptionally clears. The phenomenon of hubness is ordinarily connected with grouping of separations. Hubness aware methodologies have three algorithms, for example, hw-kNN, h-FNN, NHBNN. Hubs can be classified into two sorts. Initial one is good hubs and another is bad hubs. This classification can be founded on the quantity of label matches and mismatches in the k-events [5]. To start with methodology is hw-kNN. This technique diminishes the effect of bad hubs and it is extremely easy to implement. Bad hubness can be distinguished by its weight. If a point shows a bad hubness, give its vote as lesser weight. Second approach is h-FNN. This algorithm consolidates weight with fuzzy votes. It utilizes a threshold parameter. The anti-hubs are getting decided by utilizing the threshold parameter. One noteworthy drawback in this algorithm as it has not clarified a reasonable method for managing with anti-hubs. Third approach is NHBNN. This algorithm utilizes the Naïve Bayes standard to take into consideration further advancement. It also has not given a detailed description of managing with anti-hubs. Both h-Fnn and NHBNN does not handle with anti-hubs. In High dimensional information, the greater part of the focuses may have a place with either hubs or to anti-hubs yet few may neither has a place with hubs nor to anti-hubs. These focuses have not taken consideration into the past algorithm. The accompanying data based voting methodology has taken in to consideration. HIKNN handles anti-hubs through data based structure. The overall occurrence of informativeness is taken into consideration. It had well generalized and may be over fitting on the dataset. It was parameter free. It has enhanced the general order precision.

E-shopping is the major looming trends among people. They wish to share their experience in the form of rating and reviews in public network. Even though the Recommendation System gives the best and good results it suffer from classification and over-fitting problem. The personalization can't be predicted by social resemblance alone, it also in need of personal characteristics. To overcome the problems in RS, the recommended model called iterative recommended system, which integrates user's profile, interpersonal, intrapersonal curiosity and interpersonal impact [8]. The system makes use of traditional boosting approach and the proposed iterative commend System to restore correctness and robustness. AdaP-Boost algorithm selects model from the dataset and integrate predictions for each user. The AdaP-Boost uses much iteration and certainly adopts guessing of products for recommendations based on other guessing to make it constant with each other.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 8, August 2016

A. Vezhnevets and O. Barinova recommended that for the most part boosting faces over-fitting issue on some dataset and capacities outstandingly on some another datasets [6]. Creators of this paper found that this issue happens due to nearness of covering classes. To beat this issue boosting, 'confounding specimens' are find using Bayesian classifier and expelled amid boosting stage. Creators experimented with proposed approach; they evacuated confounding cases and did the investigation on the aftereffects of AdaBoost without befuddling delineations. To recognize the confounding occurrences creators' used flawless Bayesian classifier, cases which are misclassified this classifier are considered as befuddling examples. For boosting reason AdaBoost calculation is used, confounding examples are expelled from boosting process. An aftereffect of the trials exhibited perception about covering classes was right.

A. Ganatra and Y. Kosta portrays that groups of classifiers are won by delivering and consolidating base classifiers, created using other machine learning strategies [7]. The goal of these troupes is to expand the prescient exactness concerning the base classifiers. A champion amongst the most standard systems for making outfits is boosting, a gathering of procedures, of which AdaBoost is the most unmistakable part. Boosting is a general approach for upgrading classifier exhibitions. Boosting is an entrenched procedure in the machine learning group for upgrading the execution of any learning calculation. It is a strategy to consolidate feeble classifiers delivered by a powerless learner to a solid classifier. Boosting worries to the general issue of conveying a to a great degree exact conjecture guideline by joining unpleasant and tolerably erroneous dependable guidelines. Boosting Methods join various feeble classifiers to convey an advisory group. It looks like Bagging and other advisory group based techniques. Various feeble classifiers are joined to make a successful capable board of trustees. Successively apply frail classifiers to changed renditions of information. Expectations of these classifiers are joined to deliver a capable classifier i.e. to enhance the prescient exactness concerning base classifiers, outfit classifiers are utilized. Paper [11] portrayed the advancement of the boosting and assessment of boosting calculations with various parameters. Tests demonstrated that boosting has unrivaled expectation abilities than sacking as groups the specimens all the more accurately.

III. PROPOSED WORK

The boosting process of the cluster approach includes the partitions of training data into clusters that contain highly similar member data to break up and localize the problematic training data. The boosting of cluster then uses these clusters integrated into boosting to improve the subsequent functions as opposed to previous work that has used clusters only for preprocessing. First, the boosting of clusters evaluates each cluster separately to identify whether the problematic training data should be used to learn subsequent functions. This allows for more selective boosting to accommodate different types of problematic training data. Next, boosting the cluster learns subsequent functions separately on each cluster using only the member data in that cluster. This allows for less complex subsequent functions and helps to mitigate over fitting from being propagated into boosting. Last, boosting learns subsequent functions starting with all the cluster members—not just those deemed incorrect by the initial function. This allows for more inclusive boosting that can accommodate problematic training data deemed correct.

In this module we collect the data from the UCI bench mark resource in which we extract a particular dataset named, PIMA Indians diabetes (http://archive.ics.uci.edu/ml/machine-learning-databases/pima-indians-diabetes/pima-indians-diabetes.names) to predict the diabetics by considering several attributes like Plasma glucose, Diastolic blood pressure, Age etc., the attribute called Class variable is used to predict the value of a particular person whether they have diabetes if the class value 1 is interpreted as positive for diabetes and the value 0 is interpreted as negative for diabetes. And finally this module loads the data attributes into the database.

Clustering is the process of partitioning a group of data points into a small number of clusters. Here k number of clusters gets formed by grouping the collected data based on the attributes or features by using k- means algorithm. The idea is define k centers, each for one cluster. The better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early grouping is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing an objective function knows as squared error function.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 8, August 2016

Algorithmic steps for k-means clustering

Let $X = \{x_1,x_2,x_3,\ldots\ldots\ldots,x_n\}$ be the set of data points and $V = \{v_1,v_2,\ldots\ldots\ldots,v_c\}$ be the set of centers.
1) Randomly select *'c'* cluster centers.
2) Calculate the distance between each data point and cluster centers.
3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

Where, *'c_i'* represents the number of data points in i^{th} cluster.
5) Recalculate the distance between each data point and new obtained cluster centers.
6) If no data point was reassigned then stop, otherwise repeat from step 3).

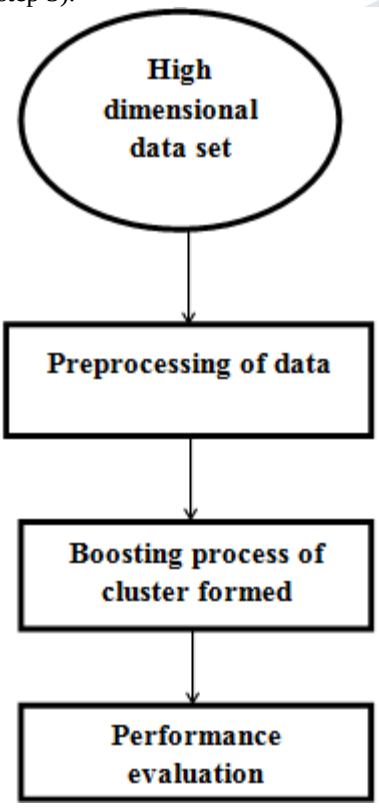

Fig 1: Process of proposed methodology

The clusters formed in the previous module is give as input to this boosting mechanism, this mechanism uses selective boosting to improve predictive accuracy on problematic training data and to predict the correct label, this structure uses cluster types such as HES Heterogeneous Struggling, HEP Heterogeneous Prospering, HOS Homogeneous Struggling, HOP Homogeneous Prospering which helps to mitigate the filtering problem in subsequent functions. The cluster type is computed using the localized estimate metric from the minority label. First, the training data is broken into sets of clusters with varying k where each set of clusters minimizes the objective function; Second, CBB chooses the set of clusters with the lowest BIC (*Bayesian information criterion)*, Third, CBB learns the initial function using all the training data. After selective boosting, the set of functions is assigned the weighted vote (MLE) and used to predict the labels for a new instance. The learning rate used to control the update of the weights for the incorrect instances. There are two different ways that these subsequent functions can be used: restricted and unrestricted. Restricted only counts the subsequent functions learned on the cluster to which the new instance would be assigned and disregards votes from other clusters. Unrestricted counts the votes from subsequent functions learned from all the clusters. Next; these clusters are designed to break the training data into different areas since each cluster encapsulates only the label instances with a high degree of similarity.We use very different methods for learning and produce functions with varying complexity allowing us to assess and analyze our approaches more comprehensively.

This module is used to predict the new class instances for the class label. Each member in a cluster get predicted whether it belongs to old label or new label. The cluster gets classified by using decision tree concept which depends on the attributes. This approach builds the tree from the top down, with no backtracking. Information Gain is used to select the most useful attribute for classification. Entropy gets calculated to find the homogeneity of sample. A completely homogeneous sample has entropy of 0. An equally divided sample has entropy of 1.The information gain is based on the decrease in entropy after a dataset is split on an attribute.

Process:

- First the entropy of the total dataset is calculated.
- The dataset is then split on the different attributes.
- The entropy for each branch is calculated. Then it is added proportionally, to get total entropy for the split.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 8, August 2016

- The resulting entropy is subtracted from the entropy before the split.
- The result is the Information Gain, or decrease in entropy.
- The attribute that yields the largest IG is chosen for the decision node.

When a branch set with 0 then it's a leaf node. Otherwise they need to split depend on the attributes. The prediction rules get understandable from the training data set. The whole data set get analyzed and predict the label. The boosting approach will carry out this prediction of class label value whether they are affected with diabetics or not.

Performance Evaluation

The System is evaluated against the following properties:

- Precision is the probability that a (randomly selected) retrieved record is relevant to the search.
  Precision rate= true positive/true positive+false positive (1)
- Recall is the probability that a (randomly selected) relevant record is retrieved the query that are successfully retrieved.
  Recall rate= true positive/true positive+false negative
  (2)
- Accuracy is calculated using the formula,
  Accuracy rate = true positive+true negative/true positive+true negative+false positive +false negative.
  (3)

IV. EXPERIMENTAL RESULTS

The data set used is collected from the UCI bench mark resource in which extracted a particular dataset named, PIMA Indians diabetes to predict the diabetics by considering several attributes like Plasma glucose, Diastolic blood pressure, Age etc.The data set contains 8 attributes and the class value. The attributes are present in numeric value. Some attributes are age, diabetics pedigree function, body mass. The models which are proposed in this work get implemented by using the dot net language. The 2GB system is used to implement the experiment. The performance analysis is made for both the method and they get compared.

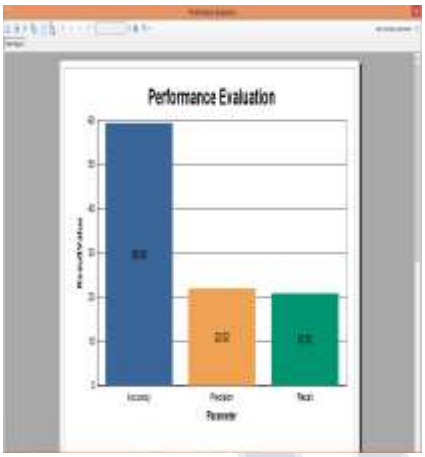

Fig 2: performance evaluation of process

V. CONCLUSION

Data mining possess high importance in dealing the high dimensional real time noisy data. The extraction of useful information from the large amount of data is a tedious task. The hubness phenomenon is implemented to improve the accuracy of the cluster formed.A general boosting framework has been proposed to improve the accuracy of any given clustering algorithm. Such performance improvement is achieved by iteratively finding new data representations that are consistent with both the clustering results from previous iterations. Empirical study shows that our proposed boosting framework is able to improve the clustering performance of several popular clustering algorithms.Boosting proved advantageous for more accurate results in machine learning. Cluster based boosting approach addresses limitations in boosting on supervised learning algorithms

REFERENCES

[1] Yi Liu, Rong Jin, and Anil K. Jain, "BoostCluster: Boosting Clustering by Pairwise Constraints", 2007.

[2] Rutuja Shirbhate, Dr. S. D. Babar, "Cluster based boosting for high dimensional data", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) 2016.

[3] R. Fergus. Visual Object Category Recognition. PhD thesis, Robotics Research Group, Department of Engineering Science, University of Oxford, 2005.

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

[5] N. Tomasev and D. Mladenic, "Nearest neighbor voting in high dimensional data: Learning from past occurrences," Computer Science and Information Systems, vol. 9, no. 2, pp 691–712, 2012.

[6] A. Vezhnevets and O. Barinova, "Avoiding boosting over fitting by removing confusing samples," in Proc. Eur. Conf. Mach. Learn.,2007, pp. 430–441.

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

[8] Mr.D.Ravi, R.Deepika, R.Jai Gayatiri, S.Jaya Surya, "Cluster based boosting algorithm for efficient recommender system", IJRTER-2016.

[9] Tran T.N., Wehrens R, and Buydens L.M.C, 'Knn Density Based Clustering for High Dimensional Multispectral Images,'Proc.Second GRSS/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas, 89(7), 147-151(2003).

[10] Arthur and Vassilvitskii, 'K-Means++: The Advantages of Careful Seeding,' Proc. 18th Ann. ACM-SIAM Symp. Discret Algorithms (SODA), 8(3), 1027-1035(2007).