IJournals: International Journal of Social Relevance & Concern
ISSN-2347-9698
Volume 4 Issue 10 October 2016

Hubness Based Clustering for Improved Accuracy Prediction in Medical Data Set

L.Hemalatha[1]; Dr.I.Elizabeth Shanthi[2]

Research Scholar, Department of Computer Science, Avinashilingam University, Coimbatore[1]

Associate Professor, Department of Computer Science, Avinashilingam University, Coimbatore[2]

Email ID: hemamphil2015@gmail.com[1] ; shanthianto@gmail.com[2]

ABSTRACT

The huge amount of data springs up naturally in various domains, which confronts a great challenge for the data mining techniques in terms of efficiency and effectiveness. To achieve accurate information from the collected data various techniques are evolved. The analysis of cluster or clustering is a process of collecting objects in a manner that objects in the same gathering are similar than to those in different gatherings. Boosting is an iterative process which aims to improve the predictive accuracy of the learning algorithms. Boosting engages by learning in various functions concentrating on incorrect instances where the previous functions predicted the wrong label. The clustering becomes difficult when the sparsity of the data is increased and a difficulty also arises in grouping the data points. So the system is enhanced by implementing the hubness which is the tendency of high-dimensional data to contain points (hubs) that frequently occur in k nearest-neighbor lists of other points that can be successfully exploited in clustering. When clustering is performed on basis of hubness then the problem of high dimensionality is diminished. The aim of this paper is to form cluster without label noise and to improve the predictive accuracy by applying hubness phenomena.

Key Words: Data Mining, Clustering, Boosting Accuracy Prediction.

1. INTRODUCTION

Data mining is the process of finding insightful, intriguing enlightening, justifiable knowledge and extracting various models from substantial large scale of database information. The tools of data mining predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The necessity of data mining arises from the natural evolution of information and data base technology [1]. It is not an easy thing to identify the information from the massive databases. Mostly many tools and techniques are used for large set of numerical and categorical value of data. The inability of human is to understand and to decide if it is a complex data. Boosting is an iterative process which aims to improve the predictive accuracy of the learning algorithms. It helps the user to predict the accurate label for the classes. Medical field is one where prediction saves lives. Diabetes is a common sickness where many people fall prey off. Due to the wrong communication by the past diabetic patients, most of the people come to a quick and wrong conclusion that leads other side effects. Hence there is better way to overcome this problem that is to get a thorough knowledge about the disease so as to predict it at the initial stage and prevent it before.

Boosting has problems in the way it learns subsequent functions. These functions are trained focusing on all the incorrect instances in the training data where the initial function did not predict the

correct label. These additional functions subsequently accommodate highly dissimilar training data. This can result in subsequent functions with an increased complexity and likelihood of over fitting. The data point separation and grouping of data in accurate manner is difficult with the high dimensional data which is omnipresent and abundant. However, not all phenomena that arise are necessarily detrimental to clustering techniques.

In the proposed work the hubness phenomena is applied on clustering process. The hubness is the tendency of some data points in high-dimensional data sets to occur much more frequently in k-nearest neighbor lists of other points than the rest of the points from the set, can in fact be used for clustering. Hubness is a good measure of point centrality within a high-dimensional data cluster and that major hubs can be used effectively as cluster prototypes. In addition, new clustering algorithms are proposed and the clustering task is evaluated [2]. This algorithm frequently offers improvements in cluster quality and homogeneity. 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 [3].

2. RELATED WORK

The literature survey provides a look about the different research works.

2.1 A clustering method based on boosting

D. Frossyniotis, A. Likas, and A. Stafylopatis (2004) describes that boosting enhances the classification process and ready to provide better results. As described in [4], using supporting terms for better bunching quality there were augmentation for acquiring change in results. In order to upgrade the way of parceling this paper proposed a solid multi clustering arrangement which relies on upon general guidelines of boosting by boosting a fundamental gathering calculation. This various bunching approach performs emphasis of the technique of preparing cases which gives distinctive grouping Fundamental apportioning was refined by using cycles of vital bunching calculation and accumulation of various grouping results. This allotment total is acquired by using weighted voting. Each segment has a weight which shows its quality measure. Test result exhibited that the strategy is promising and give strength and improved execution.

2.2 Nearest neighbor voting in high dimensional data learning from past occurrences

N. Tomasev and D. Mladenic (2013) 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 antihubs. 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 [2]. 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 antihubs 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 antihubs. 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 antihubs. Both h-Fnn and NHBNN does not handle with antihubs. In High dimensional information, the greater part of the focuses may have a place with either hubs or to antihubs yet few may neither has a place with hubs

nor to antihubs. These focuses have not taken consideration into the past algorithm. The accompanying data based voting methodology has taken in to consideration. HIKNN handles antihubs through data based structure.

2.3 Survey on Hubness - Based Clustering Algorithms

Nikita Dhamal, Antara Bhatttacharya (2012) proposed that the research on subspace clustering calculations which give all the more grouping proficiency. The advantage of the proposed calculations is their bunch proficiency on high dimensional information. The K-closest neighbor calculations depend on ways which don't utilize parameterized groups of the likelihood utilized for order or relapse [5]. They don't make presumption of circulation of components. The fluffy based methodology is fundamentally utilized when there is irregularity as a part of area or when the information is incompletely uncovered. The dense closest neighbor calculation is a novel request autonomous calculation for finding a preparation set steady subset for the NN guideline, called FCNN principle which is a half breed approach. Half and half techniques scan for a little subset of the preparation set that, in the meantime it increases both uproarious and repetitive occasion's annihilation. Fitness improvement technique and protection strategies are joined to pick up the same point of mixture strategies.

2.4 Comprehensive evolution and evaluation of boosting

A. Ganatra and Y. Kosta (2010) have portrayed that groups of classifiers are won by conveying and combining base classifiers, made utilizing other machine learning techniques [6]. The objective of these troupes is to grow the judicious precision concerning the base classifiers. A champion amongst the most standard frameworks for making outfits is boosting, a social event of methods, of which AdaBoost is the most unmistakable part. Boosting is a dug in system in the machine learning bunch for redesigning the execution of any learning computation. Boosting stresses to the general issue of passing on a to an extraordinary degree accurate guess rule by joining obnoxious and fairly incorrect tried and true rules. Progressively apply delicate classifiers to changed interpretations of data. Desires of these classifiers are joined to convey a fit classifier i.e. to improve the insightful precision concerning base classifiers, outfit classifiers are used. This paper depicted the progression of the boosting and appraisal of boosting estimations with different parameters.

3. RESEARCH METHODOLOGY

The major steps involved in the proposed research design are shown in Figure 3.1

3.1 Formation of Cluster using K-Means clustering

Boosting of cluster includes the formation of cluster using K-Means clustering and establishment of cluster based boosting mechanism. The attribute relation file format is obtained from the CSV file format of the input data. The change in the format helps in processing the dataset in the WEKA tool. Then the data can be used for the purpose of clustering using K-means. Clustering is the process of partitioning a group of data points into a small number of clusters. Here k, the number of clusters is formed by grouping the collected data based on the attributes or features by using k- means algorithm. This forms k centers, each for one cluster. The better decision is to place them however much as could reasonably be expected far from each other. The following step is to take every point having a place with given information set and partner it to the closest center. When there is no data pending, the initial step is completed. Now we have to re-figure k new centroids as barycenter of the cluster coming about because of the past stride. After we have these k new centroids, another coupling must be done between the same information that set focuses and the closest new focus and this forms a looping process. As a result of this loop we may see that the k focuses change their area orderly until no more changes are possible. This calculation helps in minimizing the target capacity which is known as squared mistake capacity. This process is performed using WEKA tool.

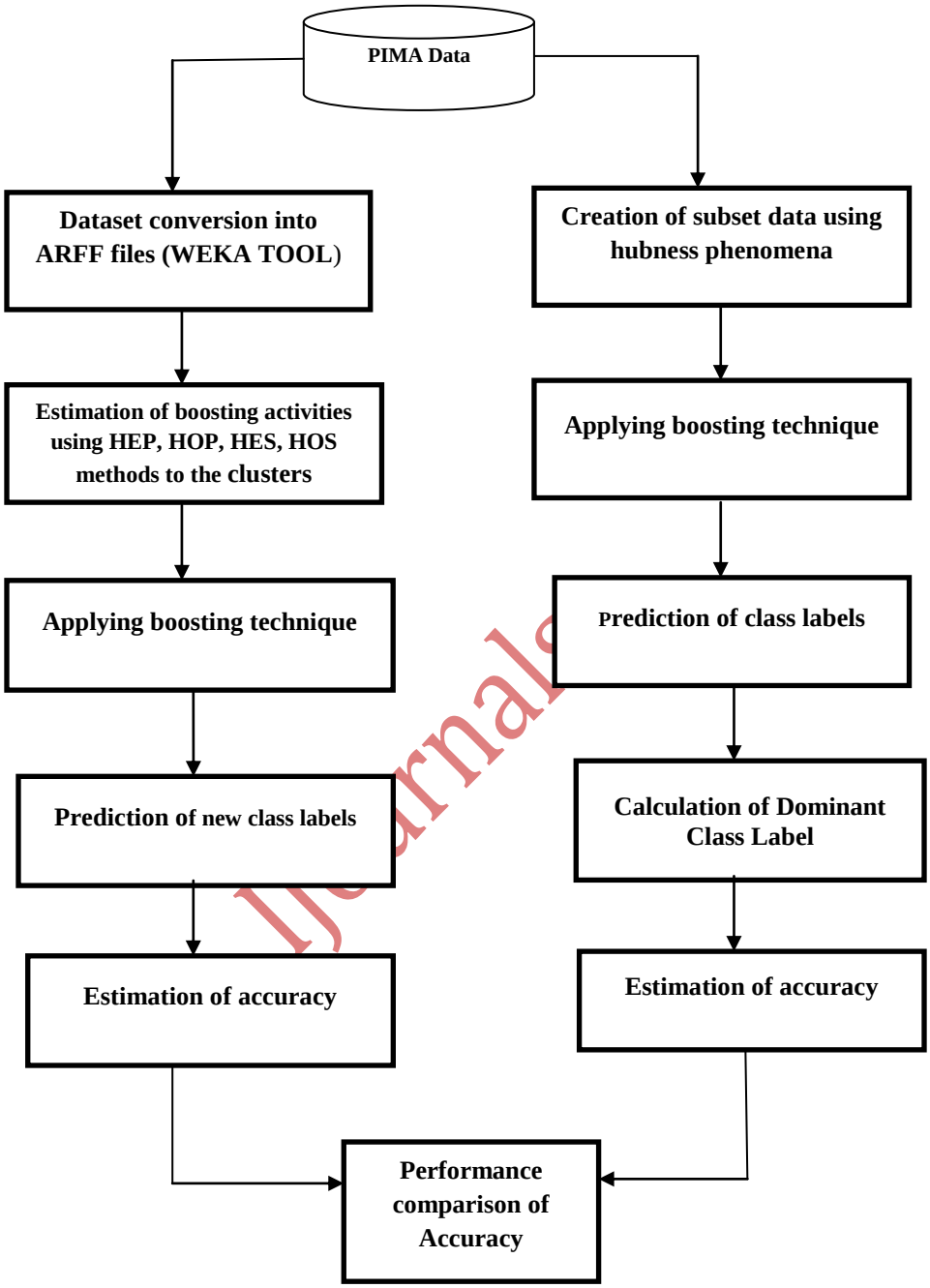

Fig 3.1 Overview of Research Design

3.2 Establishment of Cluster Based boosting mechanism

The clusters formed in the previous module are given as input to this boosting mechanism, this uses selective boosting to improve predictive accuracy on problematic training data and predict the correct label. The structure of cluster types is HES Heterogeneous Struggling, HEP Heterogeneous Prospering, HOS Homogeneous Struggling, and 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. Initially the training data is broken into set of clusters with varying k where each set of clusters minimizes the objective function and then boosting the cluster chooses the set of clusters with the lowest BIC (Bayesian information criterion). Finally, boosting the cluster learns the initial function using all the training data. After particular boosting, the arrangement of functions is attributed out the weighted vote MLE (Maximum Likelihood Estimate) and used to anticipate the marks for another instance. The learning rate used to control the redesign of the weights for the incorrect occurrences. There are two diverse ways that these resulting capacities can be utilized as restricted and unrestricted. Restricted just numbers the consequent capacities learned on the cluster to which the new example would be allocated and slights votes from different groups. Unlimited numbers the votes from ensuing capacities gained from every one of the groups. Next these groups are intended to break the preparation information into various regions since every cluster exemplifies just the name occasions with a high level of closeness. We utilize altogether different strategies for learning and deliver capacities with changing many-sided quality permitting us to survey and investigate our methodologies all the more thoroughly.

3.3 Prediction of new disease class labels

This module is used to predict the new class instances for the class label. Each member in a cluster is predicted whether it belongs to old label or new label. The cluster is classified by using decision tree concept which depends on the attributes. This approach builds the tree from the top down, with no backtracking. The classification made by the decision tree includes three different nodes. The root node is the one where there is no incoming node; they will have only the outgoing nodes. The node is known as internal node when they have one incoming node and two outgoing nodes. The node present in the terminal possesses a single incoming node. Each leaf node present in the decision tree has a class label. The different attributes present in the nodes help to classify the decision. The choice tree prompting calculation must give a strategy to indicating the test condition for various trait sorts and a target measure for assessing the integrity of every test condition. The decision tree process includes the input data, feature set for classification and output feature as target. The tree is begun to construct with a root node where the growth of tree is processed until the end condition is satisfied. The attribute value plays an important role in classifying the metrics and prediction of class label value predicts whether they are affected with diabetics or not.

3.4 Hubness in Clustering

The high dimensional data is difficult to form clusters with accurate predictions. This problem is overcome by using the hubness phenomenon. Hubness, which is the tendency of some data points in high-dimensional datasets to occur much more frequently in k-nearest neighbor lists of other points than the rest of the points from the set, can in fact be used for clustering [3]. Hubness is a demonstration of high dimensional information to contain centre points that focuses every now and again happen in k-closest neighbor arrangements of different focuses. The hubness phenomenon supports in the high dimensional data as the points which occurs frequently in the neighboring list grouped as clusters. The dataset to form clusters are picked in a random manner where it contains maximum positive points and with minimum negative points. The iterative process is carried out in choosing the random points from the dataset. The class label is predicted in each iteration and the new class label is chosen for the particular data from the major result obtained from the prediction. The prototype of the cluster is supported in prediction of patient's disease.

3.5 Performance Evaluation

The parameters used to evaluate the performance of proposed method are Precision, Recall and Accuracy. The system is evaluated against the confusion matrix properties such as true positive, true negative, false positive and false negative which is compared using actual values and predicted values. The overall performances of the existing and proposed methods are calculated using the same measures.

4. RESULTS AND DISCUSSIONS

The first step is to upload the dataset which is needed to be preceded into cluster formation. The chosen dataset cannot be processed in their original form in WEKA tool. It should be converted into ARFF file format to be loaded into WEKA tool. The data file which is uploaded should be in CSV format which is converted into ARFF file format.

The file to be processed is chosen from the option open file. The open file option helps the user to upload the required ARFF file. The PIMA dataset file is uploaded. In the preprocessing stage, the attributes present in the dataset are separated and displayed automatically on the screen. The minimum and maximum value for the selected attribute is displayed. The mean and the standard deviation for the selected attribute are displayed on the screen.

The cluster option is chosen from the WEKA tool and the trainer set is chosen from the cluster mode. This is chosen by the user to decide the type of cluster formation that is required for the further process. The simple K-means cluster process is chosen for this work to support the cluster formation. The cluster formed is chosen and uploaded for the processing. When the file is uploaded the user selects the start button and the process of cluster formation starts. The number of clusters is formed as specified by the user. The result of K-means clustering is processed to the next level of cluster file uploading process.

The implementation starts with the process of dataset given as input to the hubness phenomenon, where the clusters are formed by the random data points where it contains maximum positive points and with minimum negative points that improve the centroid approach. The data points for analyzing are taken in a random manner and the labels are predicted. The positive predicted values and negative predicted values are counted and compared for the class variable 0 or 1. The percentage measures for cluster based boosting and hubness clustering are calculated using the metrics precision, recall and accuracy. A sample iteration result is shown in Table 4.1.

Table 4.1 Performance Evaluation of Cluster based boosting and Hubness based clustering

Metrics	CBB	HBC
Precision	0.7	0.87
Recall	0.68	0.72
Accuracy	0.69	0.79

The performance comparison chart is drawn between the performance of cluster based boosting and the hubness based clustering. The graphical implementation of the Table 4.1 is shown in Fig 4.1. The chart shows better accuracy prediction of 10% more (approximately) than the existing method.

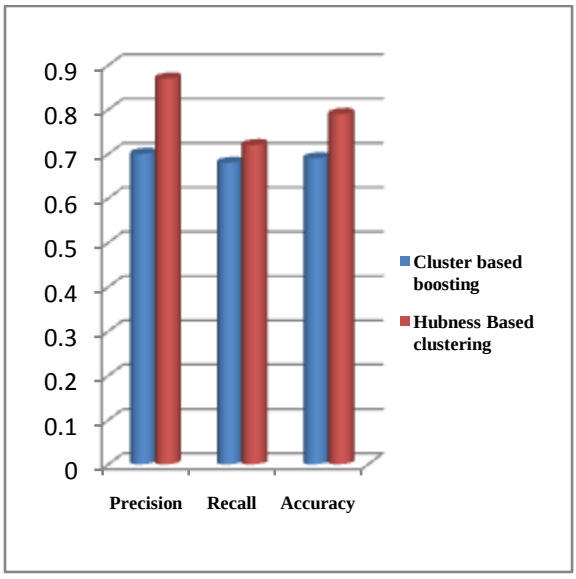

Fig 4.1 Performance Comparison

5. CONCLUSION AND FUTURE WORK

This work hubness based clustering is employed to improve prediction accuracy than the existing methodology that uses boosting process when dealing with the high dimensional noisy data. The high dimensional data in the real world data sets introduce noise where the phenomenon of hub performs better analysis of these data. The existing cluster algorithm that used boosting reduced the performance while dealing with the high dimensional data but the hubness phenomenon helps to provide better clusters. The separation of data as cluster can be obtained in an effective manner. The number of clusters while analyzing the high dimensional data varies automatically. The experiment is demonstrated using the PIMA dataset in order to analyze the label prediction in an accurate manner. The performance comparison shows that the accuracy of the hubness phenomenon in dealing the high dimensional data is improved 10% better than the existing method.

Future research can be moved towards the improvisation of hubness as directed in kernel mapping and the shared neighbor clustering approach. Further it can progress in the direction of handling arbitrarily shaped clusters.

REFERENCES

[1] Han, J., Kamber, Data Mining: Concepts and Techniques. Morgan Kauffman, San Francisco (2006)

[2]R.Shenbakapriya,M.Kalimuthu,P.Sengottuvelan, "Improving Clustering Performance on High Dimensional Data using Kernel Hubness", International Conference on Simulations in Computing Nexus, ICSCN-2014.

[3] Nenad Tomasev, Milo s Radovanovi_c, Dunja Mladeni_c, and Mirjana Ivanovi, "The Role of Hubness in Clustering High-Dimensional Data", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, NO. 3, March 2014.

[4] D. Frossyniotis, A. Likas, and A. Stafylopatis, "A clustering methodbased on boosting," Pattern Recog. Lett., 2004

[5] NikitaDhamal, AntaraBhatttacharya, "Survey on Hubness - Based Clustering Algorithms", International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 (2012).

[6] A. Ganatra and Y. Kosta, "Comprehensive evolution and evaluation of boosting," Int. J. Comput. Theory Eng.,2010.

