UNSUPERVISED ITERATIVE CLUSTERING FOR HIGH DIMENSIONAL DATA FOR HIGH PREDICTION ACCURACY

A. Jenneth[1], Dr. G. Rasitha Banu[2], M. Thillainayaki [3], U. Prathibha[4]

ABSTRACT

The classification of the Imbalance data is a major research challenge in machine learning using parameter selection methods and data classification techniques. The challenges encountered in data learning and class imbalance learning are jointly dealt with the data streams that comprise distributions of very skewed class. The proposed idea in order to classify the majority class containing data points is to make a learning model which is predicted with wrong label by employing the existing data classification technique. Despite a considerable improvement in classifying the overfitting data in the cluster, it is still difficult to classify the high dimensional data. Hence, we propose a new technique termed as Unsupervised Iterative Clustering (UIC) to address the difficulties in handling high dimensional data. The method iterates selectively on new data points in the data streams to establish the cluster on each formed cluster using k Means Clustering Algorithm. We provide the experimental results on patient PIMA dataset. The outcomes reveal enhancement in the efficiency of the proposed method over the popular k means clustering algorithm.

Index terms - High Dimensional Data mining, K Means Clustering, Class imbalance problem, Iterative Clustering

I. INTRODUCTION

Clustering refers to a gathering of similar data points. Iterative clustering pertains to the process for improving the precision of prediction of the class labels with or without training data. More specifically the iterative clustering learns the multiple features in the data points and predicts the class labels for unique features or unique group of features [1]. The motivation of Cluster analysis on the data points to predict the over fitting of the data in the group. Mostly misclassification occurs due to an empty space phenomena as data tends to be sparse in a large stream of data. Hence, class representation is complex in large dimensional data as it grows exponentially. In order to resolve this issue, data classification mechanism has to be enhanced in managing the learning model through iterative mechanism. Even large density of data which fails in the distance based clustering [2] and density based clustering will be easily handled by the unsupervised iterative clustering mechanism. The learning model based on the iterative mechanism can generate the class with high accuracy. Unsupervised Iterative clustering is sculpted as iterative procedure to results of k means clustering methods. Firstly, the data are fragmented into clusters with varying k, where each set of clusters reduces the objective function to a minimum level. K means clustering creates the clusters based on the

[1,3,4]Asst. Professor, Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore - 641 021.

[2]Asst. Professor, Department of Health Informatics, Faculty of Public Health and Tropical Medicine, Jazan University, Kingdom of Saudi Arabia.

instance similarity of data points. Centroid and medoids in K-means iterations tend to converge to locations close to high centroid less data points which implies that using of clustering leads to data convergence with high data similarity in the cluster. The primary motivation of the work is to estimate the decision boundaries that maximize the clustering accuracy. The class is defined with the dynamic boundary extracted from the feature of the large data streams [3]. The cluster is iterated with a set of clusters from a small set of cluster which further increases as long based on information streaming and sparsity. The feature extracted from the large data stream can be considered as training data to the class labels for clustering. These data points will lead to an average increase in intracluster distance [4], [5].

The rest of work is organized as follows: Section 2 deals with related work. Section 3 describes the proposed system. Section 4 discusses the experimental result. Finally section 5 concludes the work.

2. Related Works

2.1. Subspace clustering algorithm

C.E. Muller [et al.] have proposed to detect the cluster and compared it with three different prototype models such as distance based, density based and clustering oriented approach [6]. The cell-based approach searches for cliques of stable or changing grid cells. It employs several approaches which are based on a distance approximation of the data space. First, archetypal model in this methodology for clustering was offered by CLIQUE. *** The CLIQUE algorithm is briefed in three steps given below: (1) Identify even units and locate subspaces comprising clusters. (2) find out clusters in the designated subspace and (3) generate minimal description for the clusters. Proficient subspace clustering is primarily processed on the basis of monotonicity property in trimming subspace. This pruning procedure can be applied in CLIQUE algorithm in order to eliminate noise and enhance the efficiency of the cluster excellence. Moreover, the resultant cluster is highly dependent on its cell properties but the result of the computation is much more efficient. Normally these processes do not get affected by the number of data objects, yet vary on grid size.

2. Hubness based clustering algorithm

N. Tomasev (et al.) have proposed Hubness Information k- Nearest Neighbor (HIKNN) for managing high dimensional data [7]. HIKNN algorithm is compared with other previous hubness based algorithm. Hub, is a data point that frequently occurs in k-nearest neighbor list, and rarely occurring points or outliers are called as anti-hubs. The search for the nearest neighbor is a very critical aspect in clustering algorithm. The k-nearest neighbor algorithm is the basic method to find the nearest neighbor. It is largely used as a classification method, very up-front and direct. The episode of Hubness is generally related to concentration of distances. Hubness aware approaches have three algorithms such as hw-kNN, h-FNN and NHBNN. Hubs can be categorized into two types namely good hubs and bad hubs.

This classification can be established based on the quantity of label matches and mismatches identified in the k-occurrences.

3. Proposed System

3.1. Class imbalance problem[8]

We consider the Class imbalance problem, in which the

primary class of attention is rare. The data set distribution represents substantial characteristics of the negative class and a marginal positive class. In fraud detection applications, the 'fraud' is identified as the class of attention (positive class), which happens much less often than the "nonfraudulant" class which is 'negative'. Class imbalance problem may lead to some sort of misclassification.

Training the data to make down a classifier and evaluating the correctness of the ensuing learned model can produce indefinite overoptimistic estimations that are carried out by the over-specialization of the learning process to the data. But, the better option is to assess the classifiers correctness on a test set involving class-labeled records that were not applied to train the model. The building blocks applied in computing several estimation measures are the four terms discussed below:

"True positive (TP): denotes the positive records that are appropriately categorized by the classifier.

"True negatives (TN): denotes the negative records that are appropriately categorized by the classifier.

"False positives (FP): denotes the negative records that are inappropriately branded as positive.

"False negatives (FN): demotes the positive records that were inappropriately branded as negative.

The confusion matrix is a beneficial instrument for investigating the ability of the classifier to distinguish between records of different classes. When TP or TN is obtained it is indicated that the classifiers gets things correctly while if it is FP or FN, it is indicated that the classifier gets things wrong.

		Predicted class		
		Yes	No	Total
Actual class	Yes	TP	FN	P
	No	FP	TN	N
	Total	P'	N'	P+N

Figure : Confusion matrix, shown with totals for positive and negative tuples

accuracy = TP + TN / P + N

precision = TP / TP + FP

recall = TP / P

3.2 Formation of Cluster Using K-Means Clustering

The data collected are categorized into a number of clusters based on the attributes /features, using k-means algorithm in which the clustering is done around corresponding cluster centroid. Then the interspace between each entity and the centroids is computed and the value is based on the selection of a particular attribute. As we are not certain about the position of the centroid, we require to fine-tune the position of the centroid, on the basis of the recent updated data. Finally, the algorithm finds the closest centroid. i.e. the minimum distance is calculated. The main benefit of k-Means clustering is that the clusters are generated on the basis of instance resemblance, not using the instance labels.

3.3. Unsupervised Iterative Clustering

The clusters formed in the previous module are given as input to this iterative mechanism, and this process uses selective iteration to enhance prognostic precision on challenging training data and predict the correct label. This structure uses cluster types such as HES

Heterogeneous Struggling, and HOS Homogeneous Struggling, which help to diminish the filtering problem in the consequent functions. The cluster type is calculated by the confined estimate metric from the minority label. First, the training data are divided into sets of clusters with varying k. In this case firstly, each set of clusters diminishes the objective function; secondly, unsupervised iterative learning picks the set of clusters that has bottommost BIC (Bayesian information criterion), and thirdly, unsupervised iterative learning acquires the initial function by all the training data. After selective enhancement, the set of functions dispenses the weighted vote using minority label estimate (MLE) and is applied to envisage the labels for a new instance. The degree of learning is used to regulate the update of the weights for the instances that are not appropriate.

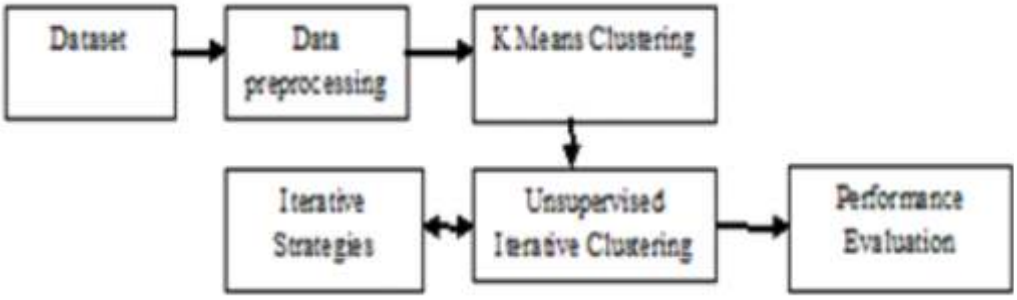

Figure 3.1. Architecture of Unsupervised Iterative Clustering

There exist two kinds of ways to use the subsequent functions: 1) restricted and 2) unrestricted.

Pseudo code:

Begin

Initialize center for random data points

Form cluster ()

Repeat

Until no reputation of data points

Count the probability of positive and negative labels in the iterated conditions

If (no. of positive data points > negative data points)

Class label = positive

Else

Class label = negative

End if

End

Restricted way of usage counts only the number of subsequent functions learned on the cluster to which the new instance would be allotted and neglects labels from other clusters. Unrestricted way of usage counts the labels from subsequent functions learned from all the clusters.

Minority Label index =

{Homogenous if minority |heterogeneous

The parameter makes use of a range to hold data sets with changing label distributions. Data sets with a greater slanting in the direction of the majority label require a congruently lesser threshold. Also these clusters are intended to divide the training data into various areas as each cluster condenses only the label instances with a more grade of similarity. We apply

different approaches for learning and create functions with changing intricacy letting us to evaluate and investigate our methodologies more systematically.

4. Experimental Results

The Experimental analysis of the proposed Clustering model using unsupervised iterative clustering approach is evaluated with diabetes dataset collected from UCI repository. The dataset is initially preprocessed using k- means clustering to establish initial clusters for the attribute to predict diabetes by considering several attributes like Plasma glucose, Diastolic blood pressure, age etc..

The method iteratively updates the weights on data points of interrelated class labels till the learning process identifies the right labels for the training data.

We make use of a cross-validation technique to decrease variance in the predictive Precision. The objects in the data set are arbitrarily separated into 10 different folds of nearly equal size. Then, Iterative process is applied to compute the predictive accuracy. In the first iteration, the data points in the first fold are treated as the test data, while all other data points in the residual folds are considered as the training data. The process then executes each pair of algorithms in the investigation, on the training data and estimates the predictive precision on the test data.

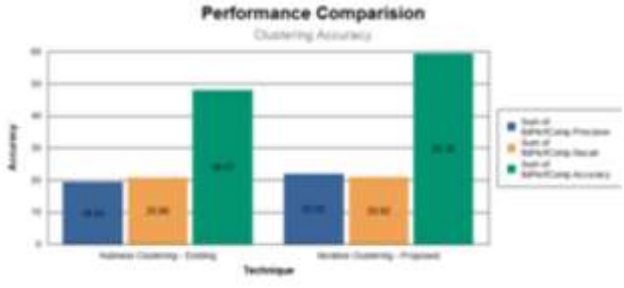

Figure 4.1: Performance Evaluation of the Unsupervised Iterative Clustering Algorithm

Technique	Precision	Recall	Accuracy
Iterative Clustering- Proposed	22.015	20.922	59.361
Hubness – Clustering Existing	19.433	20.799	48.073

Table 4.1: Performance evaluation of UIC algorithm

The performance value of the unsupervised iterative clustering method produces values described in performance tables 4.1.

Conclusion

We designed and implemented a novel Unsupervised Iterative Clustering (UIC) to address the difficulties in handling high dimensional data. The method iterates selectively on each formed cluster using k Means Clustering Algorithm. We provide the experimental results on patient PIMA dataset. The Proposed approach divides the training data into clusters containing high similarity measures between the data points. The Iterative process eliminates the over-fitting and class imbalance problem with high learning rate. The proposed learning outperforms other clustering systems.

The proposed idea is to make a learning model to classify the majority class containing data points, which is predicted a wrong label. Despite a considerable improvement in classifying the overfitting data in the cluster still it has difficulty in classifying the high dimensional data. Hence we propose a new technique termed as Unsupervised Iterative Clustering (UIC) to address the difficulties in

handling the high dimensional data. The method iterates selectively on each formed cluster using k Means Clustering Algorithm. We provide the experimental results on patient PIMA dataset. The outcomes reveal enhancement in the efficiency of the proposed method over the popular k means clustering algorithm.

The results show the superiority of the efficiency of the proposed method over the popular k-means clustering algorithm.

Reference

1. C.E. Muller, S. Gunnemann, I. Assent, and T. Seidl, "Evaluating clustering in subspace projections of high dimensional data, "Proceedings of the VLDB Endowment, vol. 2, pp. 1270-1281, 2009.

2. 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.

3. http://www.tutorialspoint.com/data_mining/data_mining_tutorial.pdf.

4. Yogesh D. Ghait, "Efficient Clustering for Cluster based Boosting", National Conference on Advancements in Computer & Information Technology (NCACIT-2016) 0975 - 8887.

5. L. Dee Miller and Leen-KiatSoh, "Cluster-Based Boosting", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 6, JUNE 2015.

6. E. Muller, S. Gunnemann, I. Assent, and T. Seidl, "Evaluating clustering in subspace projections of high dimensional data,"Proceedings of the VLDB Endowment, vol. 2, pp. 1270-1281, 2009

7. NenadTomasev, Milo s Radovanovi_c, DunjaMladeni_c, and MirjanaIvanovi, "The Role of Hubness in Clustering High-Dimensional Data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2014.

8. Jiawei Han, Micheline Kamber and Jian Pei, Data Mining Concepts and Techniques 3rd edition Han, et al. Morgan Kaufman Publishers