Min-Max K-mean Clustering in Data Mining

TARIF, Nitin Sharma
Institute of Engineering & Technology, Alwar (Rajasthan)

Abstract - The data mining is the technique which is applied to extract useful data from the large amount of data. The prediction analysis is the technique of data mining which used to predict function from current data. This research work , is base on the prediction analysis. In this work, technique of SVM is applied for the prediction analysis. To increase the accuracy, the SVM is replaced with KNN classifier. The performance of proposed modal is tested in MATLAB and simulation results shows improvement in accuracy, execution time.

I. INTRODUCTION

Data mining (now and again called data or knowledge discovery) is the way toward analyzing data from alternate points of view and summarizing it into helpful information - information that can be utilized to build revenue, cuts costs, or both. Data mining software is one of various analytical tools for analyzing data. It allows clients to investigate data from a wide range of dimensions or angles, categorize it, and outline the relationships identified. Technically, data mining is the way toward discovering correlations or patterns among dozens of fields in extensive relational databases [1]. In spite of the fact that data mining is a moderately new term, the technology is definitely not. Companies have utilized powerful computers to sift through volumes of supermarket scanner data and investigate market research reports for quite a long time. In any case, continuous innovations in computer handling power, disk storage, and statistical software are dramatically increasing the accuracy of investigation while driving down the cost. Data mining is primarily utilized today by companies with a solid customer center - retail, financial, communication, and marketing associations. Clustering can in this manner be formulated as a multi-objective enhancement issue. The appropriate clustering algorithm and parameter settings (counting qualities, for example, the distance function to utilize, a density threshold or the quantity of expected clusters) rely on upon the individual data set and intended utilization of the results. Clustering plans to discover helpful groups of objects (clusters), where handiness is defined by the goals of the data examination [2]. Not surprisingly, there are a few unique notions of a cluster that demonstrate valuable by and by. Keeping in mind the end goal to visually illustrate the differences among these types of clusters, we utilize two-dimensional points, as our data objects. Classification comprises of predicting a specific result based on a given input. Keeping in mind the end goal to predict the result, the algorithm forms a preparation set containing a set of attributes and the respective result, ordinarily called goal or prediction attribute. The algorithm tries to find relationships between the attributes that would make it conceivable to predict the result. Next the algorithm is given a data set not seen some time recently, called prediction set, which contains the same set of attributes, except for the prediction attribute – not yet known [3]. The algorithm investigations the input and produces a prediction. The prediction accuracy defines how "good" the algorithm is. Classification process is utilized for the process of prediction on the basis of the given output for certain outcomes. Due to presence of different set attributes in the training set, outcome is predicted by the processed algorithm. The relationship between the attributes is discovered by the algorithm that will be helpful in the prediction of the outcome. A data set is provided by the algorithm known as the prediction set in which the same set of attributes are present but the prediction attribute is absent that is not well known. The input is analyzed by the algorithm that is helpful in the prediction process. The accuracy of the algorithm is defined by the prediction accuracy. SCRUM is a management methodology, created with a specific end goal to improve and maintain an existing system or a production model [4]. This methodology assumes the existence of a project and some source code sequences, which quite often exist in the object-oriented software development because of class libraries. SCRUM is not addressing to development efforts for totally new or legacy systems. In the SCRUM terminology, a SPRINT is an arrangement of development exercises which are embraced during a pre-decided timeframe, more often than not from one to four weeks. The interval depends on the complexity of the product, on the risk assessments and on the required degree of skills and expertise. The speed and the intensity of a SPRINT are dictated by its agreed duration [5]. The risk is assessed persistently and permanently, and adequate measures are gone for each risk event. The classification is the process of building a model of classes from a set of records that contain class labels. In order to find how attributes vector behaves has been done using decision tree algorithm for different instances. The classes of newly generated instances have also being found on the basis of training instances and target variable prediction rules has also been generated using it [6]. The critical distribution of data can be easily understand with the help of tree classification algorithm and the extension of ID3 algorithm comes in existence and name it J48. There are number of additional features such as accounting for missing values, decision trees

pruning, continuous attribute value ranges, derivation of rules, etc that has been added in J48. The C4.5 algorithm open source implementation has been done using J48 that is the simple C4.5 decision tree used for classification.

II. LITERATURE REVIEW

Cheng-fa tsai, (2016) proposed a TSS-DBSCAN method is based on the new density-based clustering scheme. In order to reduce the expansion in the clustering as it is increasing frequently, they utilized the DBSCAN and a two-phase screening method in this paper. For the various applications, this method has been utilized for the improvement of the data clustering. On the basis of the obtained results, it is concluded that the proposed method has better performance as compared to other techniques in term of high noise filtering rate and clustering accuracy [7]. This method consumes less time and cost and considered as the best method in the world for the density-based clustering. The better efficiency has been showed by the proposed method TSS-DBSCAN and also provides the optimal data clustering technique.

Zakaria Gheid, (2016) proposed a novel method privacy-preserving k-means algorithm based on the multiparty additive scheme that is effective and cryptography-free [8]. It becomes the major privacy issue as running k-means has been utilized against distributed big data. In order to overcome this major issue, various techniques have been proposed so far. Cryptographic protocols were utilized but this method also contains many drawbacks such as degradation in the performance and not fulfils the requirement of the big data. Hence, for the given method horizontally partitioned data has been utilized. As per experiments, they concluded that proposed method provide better results against passive model.

Haohang Li, (2016) presented to examine the separation between bunches, two unique strategies has been utilized that is used to join the numeric and ostensible variables. For the numeric and ostensible variables, a separate measure has been utilized, and all the information is used to join the general separation measure [9]. In the second method, ostensible variables are converted into numeric variables and all the variables are utilized by the separation measure. In this paper author proposed the method that has been used for the Prolonged Supervised Clustering and for the classification of the algorithm. In this paper author also examined the comprehensive nature in the computational, the adaptability. On the basis of various aspects it measures the execution on various information sets. Author concluded that the proposed model not provide the optimal results as it hinder the exactness and dependability of the algorithm and hampering the involved information in the databases.

Kuan-Teng Liao, (2016) proposed a method that is based on the mechanism of the centroid boundary. In the process of clustering, the contribution of the involved error is used to position the centroid of the cluster in the average range. Therefore, the effectiveness of the clustering is degraded due to the present large average range. In order to decrease the range, author proposed the square root boundary mechanism [10]. This mechanism has been utilized to increase the effectiveness of clustering, as all the possible positions of centroids in the upper bound are limited by the mechanism. As per performed experiments, it is concluded that the proposed method provides the better performance in terms of reducing time and cost. This method also utilized for the uncertain data clustering as these two methods provide the UKmeans approaches.

Bogdan Neagu, (2016) used the 60 rural substations in order to test the capability of the proposed algorithm. As per performed examination, obtained results provide the efficiency of the proposed method using characteristics present in the load curves for the distribution of operators present in the pattern discovery [11]. On the basis o the obtained results, the capability of the proposed algorithm provide the effective and efficient results for the estimation of the losses energy by the distribution operators. It is concluded that the estimation of the power/energy losses in transformers can be obtained from the distribution substations and having knowledge of few indicators of the load curves, provide the simplification in distribution operators.

Ahmad M. Bakr, et.al, (2015) proposed improvement in the incremental DBSCAN algorithm for the building and updating of shaped clusters in huge datasets id. With the help of this proposed algorithm there is enhancement in the incremental clustering process. When this method is compared with the other incremental clustering algorithm an enhancement can be seen [12]. As per obtained experimental results, it is found that there is an increase in runtime speed of the incremental clustering process by factor up to 3.2 when the proposed method is compared with the existing methods. It is done when it is utilized on various size and dimensions of datasets. When the proposed algorithm is implemented on larger datasets, it shows better results and accuracy with higher dimensions. There are other enhancements also which are to be proposed in the future work. They also proposed methods to make an algorithm efficient enough to be working in a parallel manner. Within each partition in parallel manner, they applied the incremental DBSCAN algorithm.

III. RESEARCH METHODOLOGY

In this work, data regarding placement of the students based on some criteria on which student's selection is done is being collected. That involves data based on the selection of

students without giving any inputs for placements i.e. raw data of final year students who are eligible for placements based on certain criteria. Sectioning of students based on this is done by using k-mean clustering for generating clusters of similar and dissimilar type of data. Neural network classifier is used that consists of units arranged in layers, which convert an input vector into some output. In this each unit takes an input, applies a function to it and then passes the output to the next layer. Neural network is used as it is relevant for applying scrum practices i.e. the result one iteration at the end of sprint is used as input in the next iteration in next sprint for generation of final outcome at the end of sprint. The method is proposed without affecting the current work and comparing the result with and without using scrum practices.

In this work, the existing scrum analysis system will be improved using the KNN classifier. In the existing system the following steps are followed:-

1. The dataset will be taken as input for the classification

2. In the second step, the technique of k-mean clustering is applied which will cluster the similar and dissimilar type of data

3. In the final step, the technique of SVM classifier will be applied which will classify the similar and dissimilar data into two classed

In the proposed system the improvement in the existing system will be proposed in which following steps are applied

1. The dataset will be taken as input for the classification

2. In the first step, the technique of k-mean clustering will be applied which will cluster the similar and dissimilar type of data

3. In the last step, the technique of KNN classifier will be applied which will classify the similar and dissimilar type of data

Pseudo code of proposed technique

Step 1: Input the dataset for the classification

Step 2: Classify (X,y,x)

Here X is the training set, y is the trained set and x is the number of samples

Step 3: for i=1: size of dataset

Calculate Euclidian distance (X(i) X(i+1)

End of for loop

Step 4: Check the number of points which belongs to which class, the maximum points belongs to class will be the final class for the classification

IV. EXPERIMENTAL RESULTS

The proposed approach is implemented in MATLAB and the results are analyzed by comparing with existing approach in terms of accuracy and execution time.

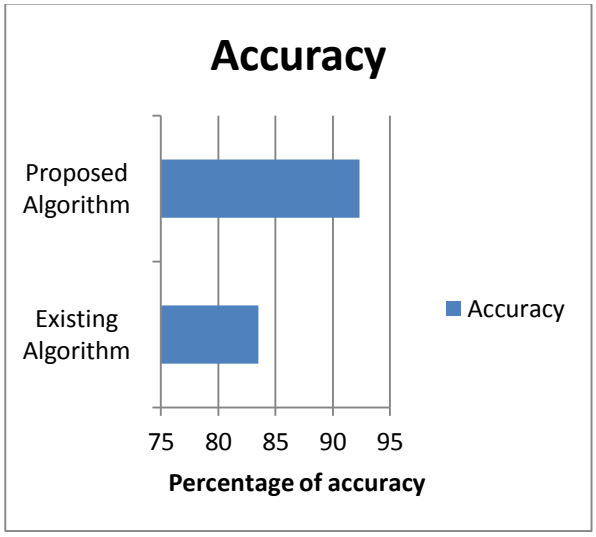

Fig 1: Accuracy Comparison

As shown in figure 1, the accuracy of proposed and existing algorithm is been compared and it is been analyzed that proposed algorithm has high accuracy due to clustering of uncluttered points from the dataset

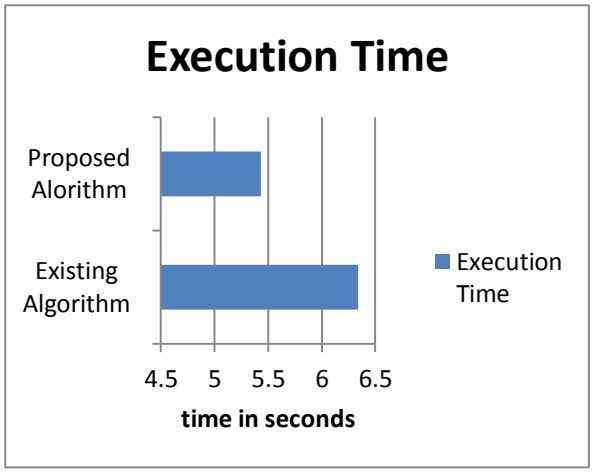

Fig 2: Execution time

As illustrated in figure 2, the execution time of proposed and existing algorithm is been compared and due to used of back propagation algorithm execution time is due in the proposed work.

V. CONCLUSION

In this work,it is concluded that classification techniques are applied for the prediction analysis. In the prediction analysis future possibilities are predicted from the current data. In this research work, the SVM classifier is replaced with KNN classifier for the prediction analysis. The performance of proposed modal is tested and simulation results shows that accuracy is increased with reduction in execution time

VI. REFERENCES

[1]. Sven Overhage, Sebastian Schlauderer, "Investigating the Long-Term Acceptance of Agile Methodologies: An Empirical Study of Developer Perceptions in Scrum Projects", 2012 45th Hawaii International Conference on System Sciences

[2]. Damian A. Tamburri, Ivan S. Razo-Zapata, Hector Fernandez, Cedric Tedeschi, "Simulating Awareness in Global Software Engineering: A Comparative Analysis of Scrum and Agile Service Networks", 2012, IEEE

[3]. Merem Elallaoui, Khalid Nafil, Raja Touahni, "Automatic generation of UML sequence diagrams from user stories in Scrum process", 2015, IEEE

[4]. Jeff Sutherland, Anton Viktorov, Jack Blount, Nikolai Puntikov, "Distributed Scrum: Agile Project Management with Outsourced Development Teams", 2007, 40th Hawaii International Conference on System Sciences

[5]. M. Mahalakshmi, Dr.M.Sundararajan, "Tracking the Student's Performance in Web-Based Education Using Scrum Methodology",

[6]. Geir K. Hanssen, Børge Haugset, Tor Stålhane, Thor Myklebust, Ingar Kulbrandstad, "Quality Assurance in Scrum Applied to Safety Critical Software", 2016, LNBIP 251, pp. 92–103

[7]. CHENG-FA TSAI, YAO CHIANG, "ENHANCEMENT OF DATA CLUSTERING USING TSS-DBSCAN APPROACH FOR DATA MINING", Proceedings of the International Conference on Machine Learning and Cybernetics, vol. 4, pp. 4, 2016.

[8]. Zakaria Gheid, Yacine Challal, "Efficient and Privacy-Preserving k-means clustering For Big Data Mining", IEEE TrustCom-BigDataSE, vol. 4, pp. 6, 2016.

[9]. Haohang Li, Shen Wang, Rui Tang, "Research on the high robustness data classification and the mining algorithm based on hierarchical clustering and KNN", IEEE, vol. 4, pp. 5, 2016.

[10]. Kuan-Teng Liao, Chuan-Ming Liu," An Effective Clustering Mechanism for Uncertain Data Mining Using Centroid Boundary in UKmeans", 2016, IEEE

[11]. Bogdan Neagu, Gheorghe Grigoraș, Florina Scarlatache, Cristina Schreiner, Romeo Ciobanu, "Patterns Discovery of Load Curves Characteristics Using Clustering Based Data Mining", IEEE, vol.4, pp.5, 2016.

[12]. Ahmad M. Bakr, Nagia M. Ghanem, Mohamed A. Ismail," Efficient incremental density-based algorithm for clustering large datasets", 2015, Elsevier B.V.