An Analysis of Decision Tree based Two-Step Clustering using Data Mining Technique

Ravinder Kaur[1] Anshu Sharma[2] Sarabjit Kaur[3]
[1]Student [2,3]Assistant Professor
[1,2,3]Department of Computer Science & Engineering
[1,2,3]CTITR, Jalandhar, India

Abstract— The classification of data patterns and distinguishing them into predefined set of classes is known as pattern recognition. A no of various methods for recognizing patterns studied under this paper. In this paper a new technique is proposed by using back propagation neural networks. This technique is enhancement of decision tree based two-step clustering to in order to achieve more accuracy and less elapsed time. Both the techniques such as existing two-step clustering and enhanced two step clustering are implemented on several data sets and results are compared with each other in terms of accuracy and elapsed time.

Key words: Data Mining, Tuberculosis, Decision Tree, Two-Step Clustering, Back Propagation Neural Networks

I. Introduction

Data Mining is known as the process of analyzing data to extract interesting patterns and knowledge. Data mining is used for analysis purpose to analyze different type of data by using available data mining tools [1]. This information is currently used for wide range of applications like customer retention, education system, production control, healthcare, market basket analysis, manufacturing engineering, scientific discovery and decision making etc. [2, 3].

This paper explores few data mining techniques in order to identify one that offers give best performance in application of classification of clusters in TB patients' data sets. A new technique is proposed that outperforms the results of two step clustering in term of accuracy and elapsed time. The proposed technique is enhancement of two-step clustering by using back propagation algorithm.

There are many preprocessing techniques for data mining but they have few limitations. Cluster analysis and standard statistical algorithm is used to arrange their score according to the level of their performance. Therefore K-mean clustering is used to analyze the performance of the decision support system [5, 6]. In this algorithm a set of beginning centroids are chosen from different parts of the test dataset and then optimal locations for the centroids are found by thoroughly exploring around the initial centroids [8].

A. Clustering

Data clustering is an unverified classification method and its objective is create groups of objects, or clusters, in a manner that place the objects in the identical cluster are very similar and place the objects in different clusters are quite distinct. Cluster analysis is commonly used in many applications such as market research, data analysis, and image processing and pattern recognition [9, 10]. In business, it is the type of data mining which can help marketers to attract their customers by improving obtaining patterns and distinguish groups of the customers [11].

1) Two Step Clustering Algorithm

Two step cluster analysis is technique of the arithmetical software set SPSS used for huge data bases, since ordered and *k* -means clustering do not scale scalable when *n* is very large. Two-step clustering is used to cluster data into different clusters and allocate classes based on variables. The SPSS Two-Step cluster technique is considered as a scalable cluster analysis algorithm that is designed to handle very huge data sets. It is capable to handle both regular and categorical variables and attributes [12]. It needs only one data pass. It is performed in two steps 1) pre-cluster the cases or records into several small sub-clusters 2) assemble the sub-clusters that are the output of pre-cluster step into the preferred number of clusters. It can also spontaneously select the number of groups. This clustering technique is very effective in classification of huge data sets and it has the ability to create clusters by using categorical and continuous variables and it is provided with spontaneous selection of number of clusters.

B. Classification

When we need to assign data items in a collection to target categories or classes, we require classification. The main objective of the classification is to a exactly forecast the object class for each case on the data set. We can take the examples such as we need a classification model that could be used for classify loan applicants as low, medium and high credit risks [14]. To create a momentous progress in the process of classification, many researchers have implemented various approaches and embrace various learning techniques which were much better than the traditional approaches. But very few researchers have used the process that could follow by clustering approach. Classification technique uses various mathematical techniques such as decision trees, statistics, neural networks and linear programming.

1) Decision Tree

Decision tree is a managed type of learning algorithm which has a pre-defined target variable and this algorithm is mostly used in classification problems. It can work for both regular and categorical output and input variables. According to this algorithm the data sample is isolated into two or more than two homogeneous groups based on most significant differentiator in variables of input data set.

2) Back Propagation Neural Network

The back propagation neural network is an ordinary way of teaching artificial neural networks used in combination with an optimization technique such as gradient descent [18]. This method computes the gradient of a loss function with according to all the masses computed in the network. After that the computed gradient is served to the optimization method which uses it to modernize the weights, in order to minimize the loss function. This algorithm is one of the most popular NN algorithms that is called back propagation algorithm [19]. This back propagation algorithm can be break down to four main stages. After selecting the masses of the network, the back propagation algorithm is used to calculate the necessary alterations.

II. Tuberculosis

Tuberculosis is a very common disease which is caused by mycobacterium and established as severe disease with really serious effects. This disease classically distresses the lungs, but it also can distress any other organ of the body. This disease is typically cured with a schedule of drugs taken for six months to two years depending on the stage of disease [20]. TB is spread by means of air from the infected person to the normal one. The TB bacteria are spread into the air when an infected person with TB disease of the lungs, speaks, sings, or sneezes. The normal people nearby can breathe in these bacteria and infected air and become infected**.** TB is not spread by sharing food, drink and sharing toothbrushes and shaking someone's hand [22].

III. Related work

According to Bellaachia A. et al. presented an analysis of the prediction of survivability rate of breast cancer patients using data mining techniques. SEER public datasets has been used in this project. This preprocessed dataset consists of 151, 886 records which have 16 fields from the SEER. At the end existing techniques have been compared with the achieved prediction performance [3]. Oyelade et al. described the ability of the student performance of high learning. This paper analyzed student result based on cluster analysis and use standard statistical algorithm to arrange their score according to the level of their performance. In this paper K-mean clustering is implemented to analyze student result. The model was combined with deterministic model to analyze student’s performance of the system [4]. Ranjini K. et al. identified the clustering and their classification. The partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often proximity according to some defined distance measure. This paper explains the implementation of agglomerative and divisive clustering algorithms applied on various types of data [7]. Hatamlou A. introduced a novel binary search algorithm for data clustering that not only finds high quality clusters but also converges to the same solution in different runs. In this algorithm a set of initial centroids are chosen from different parts of the test dataset and then optimal locations for the centroids are found by thoroughly exploring around of the initial centroids [8].

Khanna S. et al. presented classification on diabetes dataset taken from SGPGI, Lucknow. The classifier is further trained on the basis of weights assigned to different attributes which are generated by means of expert guidelines. The accuracy of classifier is verified by kappa statistics and accuracy, evolution criteria for classifiers [15]. Yadav A. et al. described that huge data is available in medical field to extract information from large data sets using analytic tool. In this paper a real data set has been taken from SGPGI. The main focus of this paper is to develop a novel technique based upon foggy k-mean clustering. The result of the experiment depicts that foggy k-means clustering algorithm has excellent result on datasets which are real as compared to simple k-means clustering algorithm and provides an enhanced result to the real world problem [16]. Chakrabotry S. et al. proposed generic methodology of incremental K-mean clustering is proposed for weather forecasting. This research has been done on air pollution of west Bengal dataset. This paper generally uses typical K-means clustering on the main air pollution database and a list of weather category will be developed based on the peak mean values of the clusters. Thus it is able to predict weather information of future.

IV. Proposed Methodology

The proposed technique that is followed by two-step clustering algorithm and back propagation algorithm is performed by using decision tree classifier. The clustering is performed the various ways and there are so many functions available for measuring the quality of clusters.

$$F(O, C) = \sum_{i=0}^{k} \sum_{oi\in ci} d(O_{i\,,}\ C_i)^2 \qquad (1)$$

These function called fitness function. A widely-known function which is mostly used is total mean-square quantization error function which is given below:

$$d(O_i, O_j) = \sqrt{\sum_{p=1}^{d}(O_i^p - O_j^p)^2} \qquad (2)$$

After that using the Euclidean distance formulas to computes the root of square difference between co-ordinates of pair of objects.

$$Dist_{xy} = \sqrt{\sum_{k=1}^{m}\left(X_{ik} - X_{jk}\right)^2} \qquad (3)$$

Two step clustering algorithm has three phases are as following.

A. *Phase1: Data Preprocessing*

- Read the data set s, remove redundancy from the data set.

  a = xlsread (‘data set name’);
  [row, column] = Size(0)
- Find the number of members and attributes of the dataset.

  r = random (1, Size (data input) *R + mu (i, 1);
  Where i =number of rows and r is redundancy free data.
- Apply the normalization to calculate similarity b/w the attributes.

  function clus_normalization
  data.x= (data.x- repmat (min (data.x)))

B. *Phase 2: First Level Clustering*

- Select the centroid point randomly from the data set.

  index = randparam (data)
- Calculate similarity in the matrix by calculating distance to the points.
- Show output of first level clustering.
  Plot Result.

C. *Phase 3: Second Level Clustering*

- Apply normalization to calculate best distance.
  function_kmean (idx, ctrs, sumd)
- Show output of clustering accuracy to best distance.

We propose a decision tree system in a medical context. This system aims to assist physicians in their practice to provide treatment for patients with tuberculosis. In the medical field, doctors are sometimes faced with issues critics during their practice that requires decision making about the disease diagnosis or a drug to prescribe. However, the knowledge of medical experts is not only based on rules, but also on a mixture of knowledge and experiences.

To develop this decision support system, we have proposed a new technique that is enhanced two-step clustering by using back propagation neural networks. This proposed technique is tested on given datasets and it is analyzed that the results are showing superiority over the existing two step clustering. In this we can say our proposed technique is better than the existing technique.

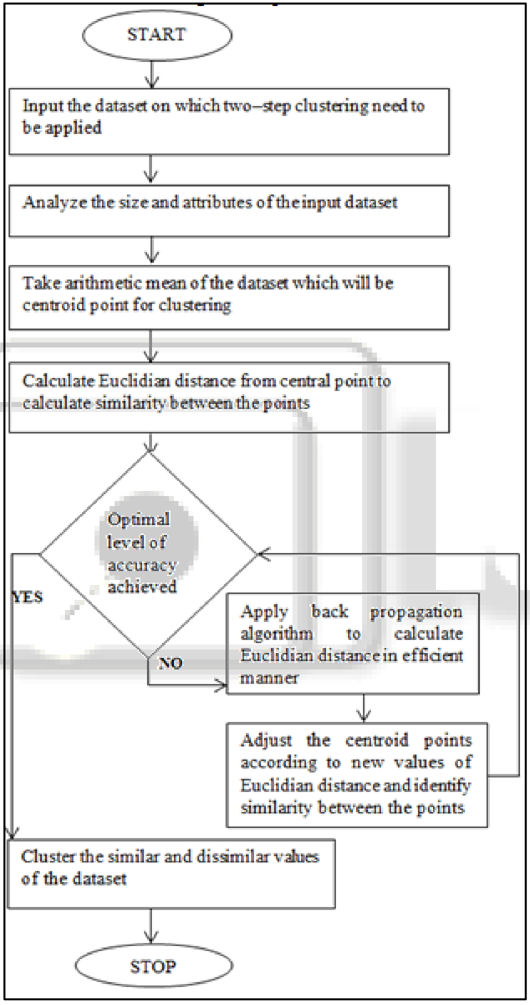
START
Input the dataset on which two–step clustering need to be applied
Analyze the size and attributes of the input dataset
Take arithmetic mean of the dataset which will be centroid point for clustering
Calculate Euclidian distance from central point to calculate similarity between the points
Optimal level of accuracy achieved
YES
NO
Apply back propagation algorithm to calculate Euclidian distance in efficient manner
Adjust the centroid points according to new values of Euclidian distance and identify similarity between the points
Cluster the similar and dissimilar values of the dataset
STOP

Fig. 1: Proposed Methodology

In this methodology, back propagation algorithm is also included which helps to find the Euclidian distance in an efficient manner. When the optimal level accuracy is not achieved by the only Euclidian distance method then the back propagation technique is introduced. Then after the centroid point is adjusted in accordance to the new distance values and similarity between the points is identified which helps to find the optimal level accuracy.

V. Results and Discussion

In this study, first of all the tuberculosis data has been taken and existing technique that is decision tree based two-step clustering is employed on given data sets and results are calculated. After that enhanced technique that two-step clustering based on back propagation neural networks is employed. Now let’s discuss the visual and quantitative results of the employed techniques.

Firstly the graph of given tuberculosis data set is plotted as in Fig 2.where x-axis represent the data. Once data set is plotted, the clustering algorithm is applied on it which turns the data sets into three clusters as shown in Fig 3.

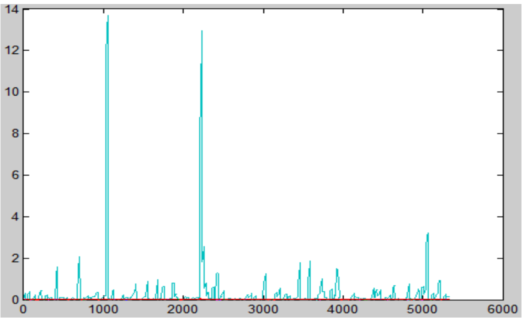

Fig. 2: Ploting the data in graph

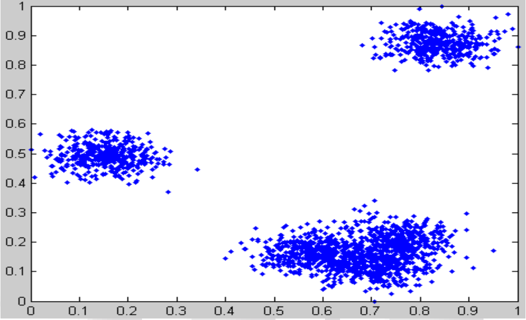

Fig. 3: Cluster the data set

This cluster algorithm is scalable upto any number of cluster sets. This is called first level clustering. Now, second level clustering or normalization is performed on clusters defined by the first level clustering which futher decompose it into the clusters as shown in the Fig 4 and in Fig 5. Decision tree classification is plotted. There is a drawback of this technique that is the intersection of the clusters occuring here.

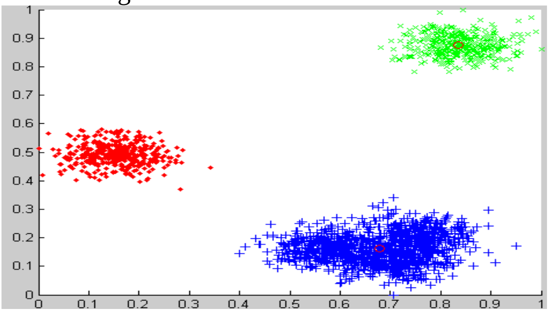

Fig. 4: Again cluster the dataset and show the centroid

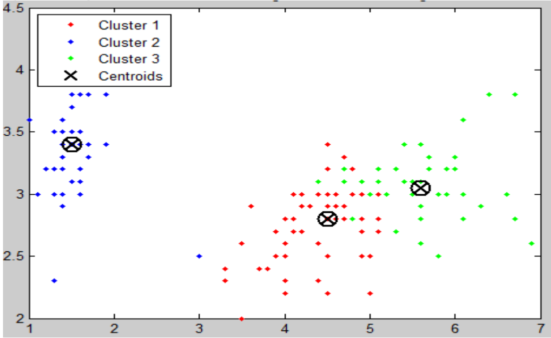

Fig. 5: Decision tree classification is plotted

The decision tree based two-step clustering is enhanced to decision tree based two-step clutering by using back propagation neural networks. It removes the drawback of existing technique and overcomes the problem of intersecting clusters along with better accuracy and elapsed time. In Fig.6 it is clearly seen that there is no intersection of clusters.

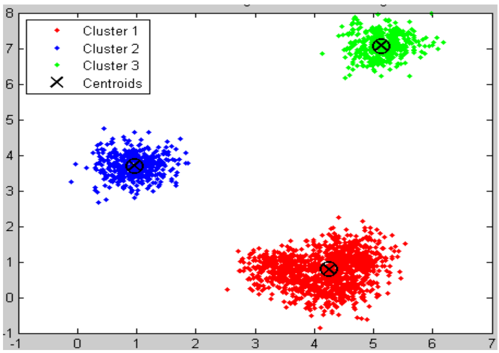

Fig. 6: Output of Decision tree based two-step clustering approach using back propagation neural networks.

VI. CONCLUSION AND FUTURE SCOPE

In this paper a new technique is proposed that is enhancement of two-step clustering by using back propagation neural networks and then the results are compared with existing two step clustering technique. After comparing the results of both the techniques it is observed that the results of proposed technique are much better and showing superiority over the existing technique. In future the proposed methodology can be used in various real life applications of data mining, web security, medical diagnosis etc. This methodology works as a sequential learning machine taking the input patterns sequentially for recognition but as a future prospect, work should be carried to generate high level networks for recognizing concurrent patterns.

REFERENCES

[1] Ray S., and Turi R.H., “Determination of Number of Clusters in K-Means Clustering and Application in Colour Image Segmentation”, School of Computer Science and Software Engineering Monash University, Wellington Road, Clayton, Victoria, 3168, Australia, 1999.

[2] Nazeer A.K.A., and Sebastian M.P., “Improving the Accuracy and Efficiency of the k- means Clustering Algorithm,” Proceedings of the World Congress on Engineering,Vol IWCE, pp.1 - 3, 2009.

[3] Bellaachia A., and Guven E., “Predicting Breast Cancer Survivability using Data Mining Techniques,” Washington DC 20052, pp.1-4, 2010.

[4] Oyelade O.J., and OladipupoO.O., and Obagbuwa I.C.,“Application of k-Means Clustering algorithm for Prediction of Students’ Academic Performance,” International Journal of Computer Science and Information Security, Vol.7, No.1, pp.292-295, 2010.

[5] Salvador S., and Chan P., “Determining the Number of Clusters / Segments in Hierarchical Clustering/ Segmentation Algorithms”, 2010.

[6] Yedla M., and Srinivasa T.M., “Enhancing K-means Clustering Algorithm with Improved Initial Center,”International Journal of Computer Science and Information Technologies, Vol.1 (2), pp.121-125, 2010.

[7] Ranjini K., and Rajalingam D., “Performance Analysis of Hierarchical Clustering Algorithm “, International Journal Advanced Networking and Applications, Vol.03, Issue.01, pp. 1006-1011, 2011.

[8] Hatamlou A., “In search of optimal centroids on data clustering using a binary search algorithm,” Pattern Recognition Letter, Vol.33, No.13, pp.1756-1760, 2012.

[9] Osamor V.C., Adebiyi E.F., and Oyelade J.O., and Doumbia S., “Reducing the Time Requirement of K-Means Algorithm,” PLoS ONE,Vol.7, Issue 12, pp.56-62, 2012.

[10] Rauf A., and Sheeba.,and Mahfooz S., and Khusro S., andJaved H., “Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity,” Middle-East Journal of Scientific Research, pp. 959-963, 2012.

[11] Sundar B., and Devi V.T., and Saravan N., “Development of a Data Clustering Algorithm for Predicting Heart,” International Journal of Computer Applications, Vol.48, No.7, pp.0975-888, 2012.

[12] Agrawal K.C., and Nagori M., “Clusters of Ayurvedic Medicines Using Improved K-means Algorithm,” International Conf. on Advances in Computer Science and Electronics Engineering, 2013.

[13] Hatamlou A., “Black hole: A new heuristic optimization approach for data clustering,” Information Science, Vol.222, pp.175-184, 2013.

[14] Kaur D., and JyotK.,“Enhancement in the Performance of K-means Algorithm”, International Journal of Computer Science and Communication Engineering, Vol.2, Issue.1, 2013.

[15] Khanna S., and Agarwal S., “An Integrated Approach towards the prediction of Likelihood of Diabetes,” Machine Intelligence Research and Advancement, pp.294-298, 2013.

[16] Yadav A.K., and Tomar D., and Agarwal S., “Clustering of Lung Cancer Data Using Foggy K-Means,” International Conference on Recent Trends in Information Technology (ICRTIT), pp.13-18, 2013.

[17] Chakrabotry S., and Nigwani N.K., and Dey L., “Weather Forecasting using Incremental K-means Clustering”, 2014.

[18] Sa L.C., and Ibrahim A., and Hossain E.D., and Hossin M.B., "Student Performance Analysis System (SPAS)," In Information and Communication Technology for The Muslim World (ICT4M), 2014 The 5th International Conference on, pp.1-6, 2014.

[19] Shukla M., and Agarwal S., “Hybrid approach for tuberculosis data classification using optimal centroid sselection based clustering,” Engineering and Science (SCES), pp.1-5, 2014.

[20] Lakshmanan S.B., and Srinivasan V., and Ponnuraja C., “Data Mining with Decision Tree to Evaluate the Pattern on Effectiveness of Treatment for Pulmonary Tuberculosis: A Clustering and Classification Techniques,” Scientific Rresearch Journal (SCIRJ), Vol.3, Issue.4, pp.2201-2796, 2015.

[21] Rajalakshmi K., and DhenakaranS.s., and RoobinN,. “Comparative Analysis of K-Means Algorithm in

Disease Prediction”, International Journal of Science, Engineering and Technology Research (IJSETR), Vol.4, Issue 7, 2015.

[22] Garg S., and Rupal N., “A Review on Tuberculosis Using Data Mining Approaches,” International Journal of Engineering Development and Research, Vol.3, Issue 3, pp.1-4, 2015.

[23] Dogra A. D., and Wala T., “A Review Paper on Data Mining Techniques and Algorithms,” International Journal of Advanced Research in Computer Engineering & Technology, Vol.4, Issue 5, pp.1976-1979, 2015.