IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-6, June 2018 E-ISSN: 2347-2693

Prediction Analysis Technique based on Clustering and Classification

Bhupendra Kumar Jain[1*], Manish Tiwari[2]

[1,2] Geetanjali Institute of Technical studies, Udaipur, India

Available online at: www.ijcseonline.org

Accepted: 10/Jun/2018, Published: 30/Jun/2018

***Abstract*-** The data mining is the technique to analyze the complex data. The prediction analysis is the technique which is applied to predict the data according to the input dataset. In the recent times, various techniques have been applied for the prediction analysis. In this paper, k-mean and SVM classifier based prediction analysis technique is improved to increase accuracy and execution time. In the prediction analysis based technique, k-mean clustering algorithm is used to categorize the data and SVM classifier is applied to classify the data. The back propagation algorithm has been applied with the k-mean clustering algorithm to increase accuracy of prediction analysis. The proposed algorithm is implemented in MATLAB and it is been tested that accuracy of clustering is increased, execution times is reduced for prediction analysis

***Keywords*-** K-mean, SVM, Prediction, categorization, Classification

1. INTRODUCTION

The large amount of data which needs certain powerful data analysis tools are thus put for the here which is also known as the data rich but information poor condition. There is an increase in the growth of data, its gathering as well as storing it in huge databases. It is no more in the hands of humans to do it easily or without the help of analysis tools [1]. There are certain data archives created here which can be visited when the data is required. The insightful, interesting and novel patterns of data are discovered from large-scale data sets using the data mining. The knowledge discovery in databases process is a very important step in data mining. The data mining and KDD are often termed as synonyms. There are databases, data warehouses, internet, information repositories involved within the data sources. The two high-level primary goals of data mining in practice have a tendency to be prediction and description [2]. As expressed before, prediction involves utilizing a few variables or fields as a part of the database to predict unknown or future values of different variables of interest, and description focuses on discovering human-interpretable patterns describing the data. In spite of the fact that the boundaries amongst prediction and description are not sharp (a portion of the predictive models can be descriptive, to the degree that they are understandable, and the other way around), the distinction is valuable for understanding the overall discovery goal. The relative importance of prediction and description for particular data-mining applications can differ considerably. The goals of prediction and description can be accomplished utilizing a variety of particular data-mining methods [3]. The data clustering is an unsupervised classification method. Its main objective is to create group of objects or clusters in such a manner that the objects which have similar properties can be grouped together. Here the distinct objects are thus present in different groups as per their properties. Within the data mining research area, cluster analysis a very old and efficient area for study. For the purpose of knowledge discovery, this step is the starting point in this direction. The data objects are grouped within a set of disjoint classes called clusters using the clustering method. There is a higher resemblance of objects which are present within a same class as compared to the two objects which belong to separate classes [4].

I.1. K-mean Clustering

The K-Means calculation utilizes a recursive system. Along these lines, its functionality it is known like k-means calculation; it is characterized from the core calculation as Lloyd's calculation, especially in the Data mining community. K-means clustering is a strategy for vector quantization, initially from flag processing, that is popular for cluster investigation in data mining. K-means clustering aims to partition n observations into k clusters in which every observation has a place with the cluster with the nearest mean, serving as a prototype of the cluster [5]. This results in a partitioning of the data space into Voronoi cells. The calculation has a loose relationship to the k-nearest neighbor classifier, a popular machine learning method for arrangement that is frequently confused with k-means in light of the k in the name. One can apply the 1-nearest neighbor classifier on the cluster centers acquired by k-means to classify new data into the existing clusters [6].

I.2. SVM Classifier

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. The steps are explained below [7]:

Set up the training data: The training data of this exercise is formed by a set of labeled 2D-points that belong to one of

two different classes; one of the classes consists of one point and the other of three points [8].

Set up SVM's parameters: In this tutorial we have introduced the theory of SVMs in the simplest case, when the training examples are spread into two classes that are linearly separable. However, SVMs can be used in a wide variety of problems (e.g. problems with non-linearly separable data, a SVM using a kernel function to raise the dimensionality of the examples, etc). As a consequence of this, we have to define some parameters before training the SVM.

Regions classified by the SVM: The method is used to classify an input sample using a trained SVM. In this example we have used this method in order to color the space depending on the prediction done by the SVM. In other words, an image is traversed interpreting its pixels as points of the Cartesian plane.Each of the points is colored depending on the class predicted by the SVM;in green if it is the class with label 1 & in blue if it is the class with label-1 [9].

The rest of the paper is organized as follow : in the section 1 the introduction is given related to data mining, classification and clustering techniques which are used in data Mining. In the section 2, the literature review is given related to the techniques which are proposed by the authors previously. In the section 3 the proposed methodology is described in the last section results of the proposed technique is highlighted.

II. LITERATURE REVIEW

Doreswamy, et.al, Medical informatics primarily deals with finding solutions for the issues identified with the diagnosis and prognosis of different deadly diseases utilizing machine learning and data mining approaches. One such disease is breast cancer, killing millions of people, particularly women. In this paper we propose a bio inspired model called BATELM which is a mix of Bat algorithm (BAT) and Extreme Learning Machines (ELM) which is first of its kind in the study of non image breast cancer data analysis [10]. Here we make utilization of BAT to optimize the parameters of ELM so that the prediction task is completed efficiently. The fundamental aim of ELM is to predict the data with least error

R. Karakis, et.al, Axillary Lymph Node (ALN) status is an extremely important factor to survey metastatic breast cancer. Surgical operations which might be vital and cause some adverse effects are performed in determination ALN status [11]. The motivation behind this study is to predict ALN status by methods for selecting breast cancer patient's essential clinical and histological feature(s) that can be obtained in every healing center. 270 breast cancer patients' data are collected from Ankara Numune Educational and Research Hospital and Ankara Oncology Educational and Research Hospital. It is concluded from LR and GA based MLP, that menopause status and lymphatic invasion are the most significant features for determining ALN status.

Marjia Sultana, et.al, Heart disease is considered as one of the major reasons for death all through the world. It can't be effectively predicted by the medical specialists as it is a troublesome task which demands expertise and higher knowledge for prediction. The heart disease turns into a plague all through the world. It can't be effortlessly predicted as it is a troublesome task that demands expertise and higher knowledge for prediction. Data mining extracts hidden information that assumes an important role in settling on choice [12]. This paper addresses the issue of prediction of heart disease as per input attributes on the premise of data mining strategies.

Kamaljit Kaur et.al, the new system, called the Credit Based Continuous Evaluation and Grading System (CBCEGS), assesses a student on the premise of her persistent evaluation during the semester, joined with her performance at last semester examination. This multistage examination design gives a chance to students to improve their performance. In the event that a student can't perform well in tests during the semester, she can improve her performance at last semester test. In any case, it doesn't appear to be so natural [13]. In specific courses, because of their difficulty level, for example, mathematics, a student will most likely be unable to improve her knowledge at last despite hard work.

III. PROPOSED WORK

The prediction analysis is the technique to predict the situations according to the input dataset. The prediction analysis required two phases, in the first phase the k-mean clustering is applied which will cluster the similar and dissimilar type of data. The SVM classifier is applied which will classify the data. The k-mean clustering consists of three steps, in the first step the arithmetic mean of the whole dataset is calculated which will be the central point. In the second step, Euclidian distance is calculated from the central point. In the last step, the data will be clustered according to their similarity. The clustered data will be given as input the SVM classifier for the classification. In this work, the k-mean clustering algorithm will be improved to increase cluster quality which increase classification quality. The back propagation algorithm is applied with the k-mean clustering algorithm which increase cluster quality. The back propagation algorithm will calculated the Euclidian distance in the dynamic manner and Euclidian distance at which maximum accuracy is achieved is the final distance for the data clustering.

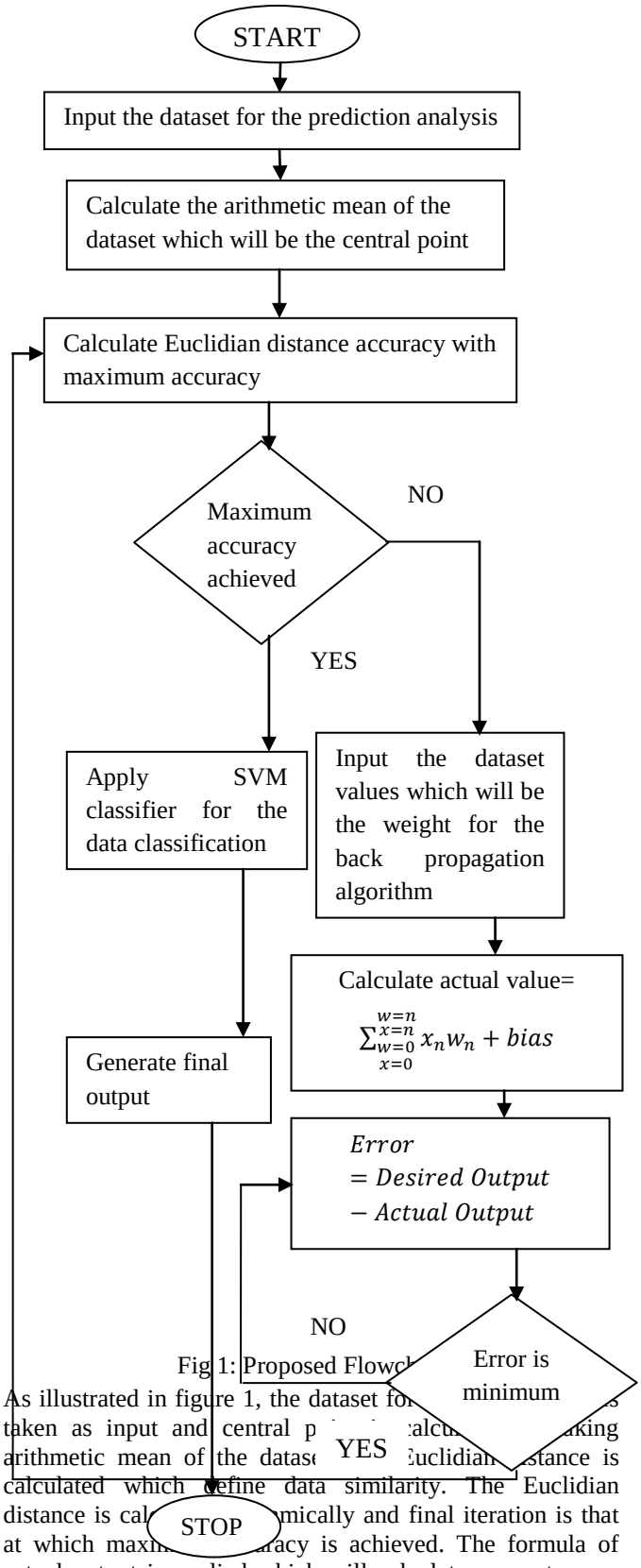

Fig 1: Proposed Flowch

As illustrated in figure 1, the dataset fo taken as input and central p calcu king arithmetic mean of the datas Euclidian stance is calculated which define data similarity. The Euclidian distance is cal mically and final iteration is that at which maxi racy is achieved. The formula of actual output is applied which will calculate error at every iteration and when the error is reduced to minimum the maximum accuracy is achieved. When the maximum accuracy is achieved the SVM classifier is been applied to classify the input data.

IV. RESULTS AND DISCUSSION

The proposed algorithm is been implemented in MATLAB by considering the dataset which is described in the table 1. The Breast cancer dataset is taken for the classification from the UCI repository

Parameter	Values
No of instances	569
Attributes	32
Missing values	Yes
Area	Life
Association rules	Classification

Table 1: Dataset Parameters

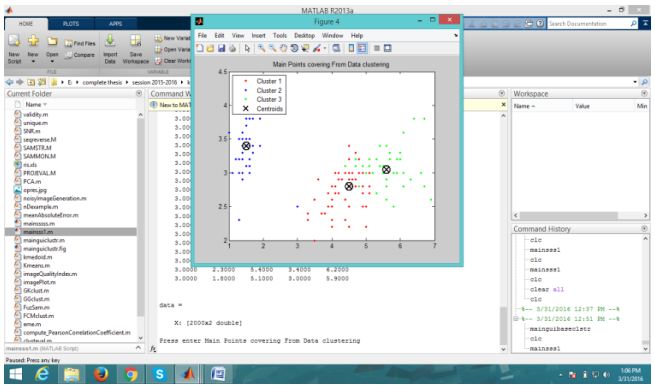

Fig 2: Data Clustering

As shown in figure 2, the k-mean algorithm is applied with the back propagation for the data clustering

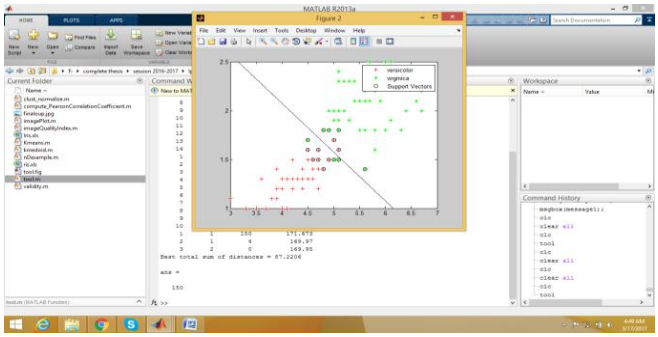

Fig 3: Data Classification

As shown in figure 3, the SVM classifier is been applied which will classify the data which is output of data clustering

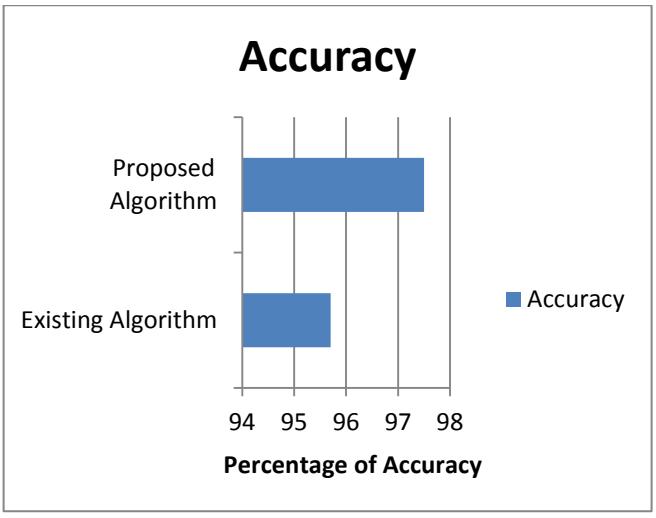

Fig 4: Accuracy Comparison

As shown in figure 4, 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.

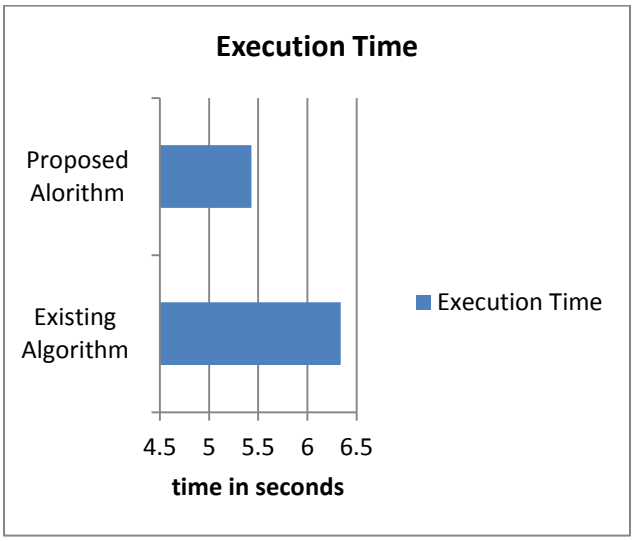

Fig 3: Execution time

As illustrated in figure 3, 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

Parameter	Existing Technique	Proposed Technique
Accuracy	95.7 percent	97.8 percent
Execution Time	6.4 second	5.4 second

Table 1: Comparison

As shown in table 1, the proposed and existing techniques are compared in terms of accuracy and execution time. It is analyzed that accuracy of proposed technique is high and execution time is less as compared to existing technique.

V. CONCLUSION

In this paper, it is been concluded that prediction analysis the efficient technique for the complex data analysis. The back propagation algorithm is applied with the k-mean clustering algorithm to increase accuracy of data clustering. The SVM classifier is applied which will classify clustered output. It is been analyzed that proposed algorithm is testing in MATLAB and it is analyzed that accuracy is increased upto 20 percent and execution time is reduced upto 10 percent

REFERENCES

[1] Rupali, R.Patil, "Heart disease prediction system using Naive Bayes and Jelinek-mercer smothing," 2014, International Journal of Advanced Research in Computer and Communication Engineering, vol. 3, no. 5

[2] Shamsher Bahadur Patel, Pramod Kumar Yadav and Dr. D. P. Shukla, "Predict the diagnosis of heart disease patients using classification mining Techniques," 2013, IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS), vol. 4, no. 2, pp. 61-64

[3] Jyoti Soni, Ujma Ansari and Dipesh Sharma, "Prediction data mining for medical diagnosis: An overview of heart disease prediction," 2011, International Journal of Computer Applications (0975-8887), vol. 17

[4] John G. Cleary and Leonard E. Trigg," K: An Instance-based learner using an entropic distance measure," 1995, Proc. 12th International Conference on Machine Learning, pp. 108-114

[5] S. Vijayarani and M. Muthulkshmi, "Comparative analysis of Bayes and Lazy classification algorithms," 2013, International Journal of Advanced Research in Computer and Communication Engineering, vol. 2, no. 8

[6] R. Vijaya Kumar Reddy, K. Prudvi Raju, M. Jogendra Kumar, CH. Sujatha, P. Ravi Prakash, "Prediction ofheart disease using decision tree approach," 2016, International Journal of Advanced Research in Computer Science and Engineering, vol. 6, no. 3

[7] Promad Kumar Yadav, K. L. Jaiswal, Shamsher Bahadur Patel, D. P. Shukla, "Intelligent heart disease prediction model using classification algorithms," 2013, UCSMC, vol. 3, no. 08, pp. 102-107

[8] Gaurav Taneja and Ashwini Sethi, "Comparison of classifiers in data mining," 2014, International Journal of Computer Science and Mobile Computing, vol. 3, pp. 102-115

[9] Sheweta Kharya, "Using data mining techniques for diagnosis of cancer disease," 2012, UCSEIT, vol. 2, no. 2

[10] Doreswamy, Umme Salma M," BAT-ELM: A Bio Inspired Model for Prediction of Breast Cancer Data", 2015, IEEE

[11] R. Karakis, M. Tez, Y. Kilic, Y. Kuru, and I. Guler," A genetic algorithm model based on artificial neural network for prediction of the axillary lymph node status in breast cancer," 2013, Engineering Applications of Artificial Intelligence, vol. 26, no. 3, pp. 945–950

[12] Marjia Sultana, Afrin Haider and Mohammad Shorif Uddin," Analysis of Data Mining Techniques for Heart Disease Prediction", 2016, IEEE

[13] Kamaljit Kaur and Kuljit Kaur," Analyzing the Effect of Difficulty Level of a Course on Students Performance Prediction using Data Mining", 2015 1st International Conference on Next Generation Computing Technologies (NGCT)

