International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 06 | June 2021 www.irjet.net p-ISSN: 2395-0072

DATA CLASSIFICATION METHOD USING GENETIC ALGORITHM & K-MEANS ALGORITHM

Satendra Kumar[1], Prof. (Dr.) Arif Hakeem[2]

*[1](P.G Scholar) CSE Department, **SSSUTMS** Bhopal-Indore Road, Sehore (M.P), Pin - 466001*
*[2](Professor) CSE Departments, **SSSUTMS** Bhopal-Indore Road, Sehore (M.P), Pin – 466001*

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 data set. The large amount of data which needs certain powerful data analysis tools are 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.

In this work, 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.

Key Words: K- Means clustering algorithm, MATLAB.

1.1 INTRODUCTION

The information gathered from different fields is enormous. Proper storage and manipulation of this data is necessary for taking the decisions of future. A number of databases have been created for this purpose. These databases are used for the proper handling of information. Data mining is a procedure which is used to extract the important data and patterns from huge amount of accumulated information. There are other names for this process as well, such as knowledge discovery process, knowledge mining from data, knowledge extraction or data /pattern analysis.

There are various interesting patterns through which the huge data can be stored in an efficient manner within the databases, data warehouses and other repositories. The famous techniques are known as the knowledge discovery in databases (KDD). The integration of techniques is done from different aspects such as statistics, database knowledge, machine learning, neural networks, good performance calculation, pattern matching, and data extraction etc.

1.2 Knowledge Discovery from Data (KDD) Process

In KDD process, the execution of some steps is extremely imperative. The procedure which is used for the extraction of knowledge from the existing data is known as KDD. This approach basically highlights high-level applications in order to achieve specific data mining techniques. The data mining approach is extremely necessary for several regions. The fundamental role related to the presentation of regularly utilized object suites is performed by determining the correlations among the different kinds of domains existing within the data base. In order to identify frequently used object sets in KDD procedure, association rule is an additional significant aspect.

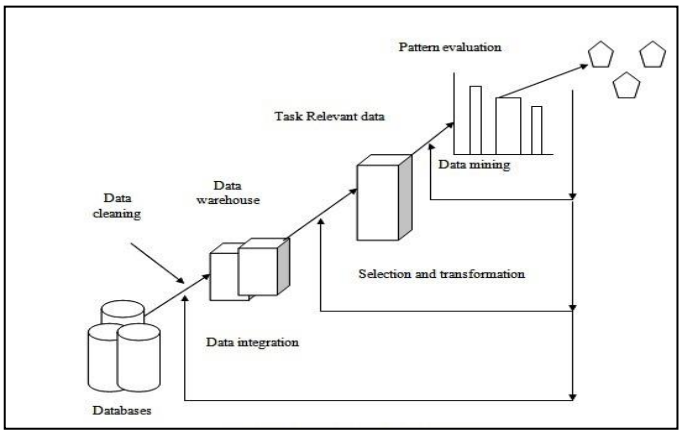

Fig 1.2 Data Mining in an Organization

The major aim of the KDD procedure is the extraction of knowledge from the information present in the enormous databases. The various stages involved in this process are [6].

1. **Data Cleaning:** The unwanted noise and contradictory information is eliminated here in the first stage.
2. **Data Integration:** The information gathered from numerous information sources is then merged.
3. **Data Selection:** Further, the information related to the testing process is extracted from the database.
4. **Data Transformation:** For transforming and consolidating information into suitable formats of data mining, data aggregation or summary operations are executed.

5. **Data Mining:** In this stage, smart techniques are applied for the extraction of data patterns. Therefore, this stage is extremely crucial.

6. **Pattern Evaluation:** The truthfully appealing patterns are recognized to represent knowledge based on appealer's ways.

7. **Knowledge Presentation:** The visualization and knowledge demonstration methods are applied in the final stage. These methods provide mined knowledge to customers.

Components involved in KDD process:

1. External and Internal Resources: First of all raw data is collected from number of resources either internal resources or external resources.

2. Extraction and Integration: In this step, first it extracts data from different resources and converts it into original format. Data cleaning is a process in which missing values are filled, it smooth noisy data and remove inconsistencies. In data integration, integration of multiple database, data cubes and files takes place.

3. Database: After that data is stored in database and data mart.

4. OLAP Application: OLAP can be used as for discovery in data mining for previously discerned relationship between data items.

1.3 Data Mining Methods

The two high-level primary goals of data mining in practice have a tendency to be prediction and description. As expressed before, prediction involves utilizing a few variables or fields. These variables are used as a section of database for predicting unidentified or prospect values of several variables of interest. The description is mainly focused on the discovery of human-interpretable patterns which describe information.

The objectives of forecasting and depiction can be performed by employing an array of specific data-mining techniques.

a. Classification: Classification is identified as a learning function. This approach is used for the mapping of (classification) data object into one or more already defined classes.

b. Regression: Regression is identified as learning a function. This approach maps a data item to an authentic-valued forecasting variable. Regression applications are of several types such as forecasting the determination of biomass existing in a backwoods gives distantly sensed microwave dimensions, the probability

c. Clustering: Clustering is a universal explanatory function. In this approach, a fixed suite of classes or clusters are identified for portraying the information. The classes can be jointly restricted and meticulous or may include a more affluent demonstration such as hierarchical or overlap classes.

d. Summarization: Summarization includes several methods to find a compressed depiction for a subset of information. An easy illustration would categorize the mean and standard deviations for all domains.

e. Change and deviation detection: This approach gives attention to the discovery of most important transforms in the data from earlier measurement.

1.4 Clustering in Data Mining

The clustering can be employed by the specialists for identifying the lands which have similarities, similar houses within a city and other properties which come used the geology filed. For the purpose of information discovery, the data clustering can provide help in documents classification on the Internet.

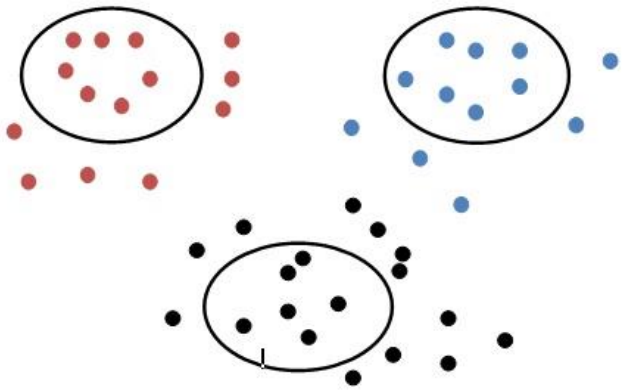

Fig.1.4 Clustering

1.5 K-Mean Clustering Algorithm

The K-Means calculation utilizes a recursive system. The functioning of this algorithm is recognized as k-means calculation beside these lines. This algorithm is typified from the interior calculation similar to the Lloyd's calculation, particularly in the Data mining area. K-means clustering is an approach which is used for vector quantization, primarily from flag processing. Flag processing is commonly used for cluster analysis in data mining. The main aim of this algorithm is the portioning of n observations into k number of clusters. In these clusters, each observation comprises a place with the adjacent mean and serves as an archetype of the cluster. This phenomenon partitions data space into Voronoi cells. The computation has a wobbly association to the k-nearest neighbour classifier. This is a well-liked machine learning technique which is used for agreement. This technique is normally puzzled with k-means in radiance

of the k in the name. The 1-nearest neighbour classifier can be applied on the cluster centre. K-means acquire these centres for classifying novel information into the accessible clusters. This is identified as adjacent centroid classifier calculation.

1.6 Classifiers

a. Principal component analysis (PCA): This is an arithmetical process that utilizes an orthogonal alteration for converting a set of observations from probably connected variables into a suite of linearly unconnected variables identified as principal components. The principal components are less or equal to the amount of authentic variables.

Algorithm:

- Mean centre the data (optional)
- Compute the covariance matrix of the dimensions
- Find eigenvectors of covariance matrix
- Sort eigenvectors in decreasing order of Eigen values
- Project onto eigenvectors in order
- Assume data matrix B is of size $m \times n$
- For each dimension, compute mean μ_i
- Mean centre B by subtracting μ_i from each column i to get A
- Compute covariance matrix C of size $n \times n$
  - If mean centred, $C = A^T A$
- Find eigenvectors and corresponding Eigen values (V,E) of C
- Sort Eigen values such that $e_1 \geq e_2 \geq \cdots \geq e_n$
- Project step-by-step onto the principal components $\overrightarrow{v_1}$, $\overrightarrow{v_2}$, … etc.

b. SVM: Support Vector Machine is considered a discriminative classifier properly described via a separating hyper plane. Following are the steps used here:

1. **Set up the training data:** A set of labelled 2D-points are used for the formation of a training data of this exercise. This training data belongs to one out of two different classes. One class comprises one point while other class comprises three points.
2. **Set up SVM's parameters:** In this study, the hypothesis of support vector machine is commenced in the simplest case. In this case, the training patterns are separated into two linearly separable classes.
3. **Train the SVM:** A method called CvSVM: train is used for the training of SVM model.
4. **Support vectors:** This approach uses a couple of technique for getting knowledge regarding support vectors.

1.7 PROBLEM FORMULATION

Following are the various research gaps of existing work which is fulfilled in this research

1. To study and analyze various predictions based technique for Data mining.

2. To propose improvement in K-mean and SVM based prediction techniques for data classification.

3. The proposed improvement will be based on back propagation algorithm to increase accuracy of classification.

4. To implement proposed technique and compare with existing in terms of accuracy, execution time.

1.8 HYBRID ALGORITHM

INPUT: Dataset

OUTPUT: Clustered Data

Start ( )

1. Read dataset and dataset has number of rows "r" and number of columns "m"
2. For (i=0 ;i=r; i++) /// selection of medoid point
    1. For (j=0; j=m; j++)
    2. Select k=data (i, j);
    End
3. Calculation of Euclidian distance()
    1. For (i=0; i=r; i++)
    2. For (j=9; j=m; j++)
    3. A(i)=data(i);
    4. B(i)=data(j);
    5. Distance =sqrt[(A(i+1)-A(i)^2) –(B(j+1)-B(j)^2);
    End
4. Normalization ()
    1. For (k=0; k=data; k++)
    2. Swap k(i+1) and k(i);
    end
5. Repeat step 3 to 4 until all points get clustered.

1.9 Tool Description

The MATLAB is the tool which is used to perform mathematical complex computations. In this MATLAB simplified C is used as the programming language. The MATLAB has various inbuilt toolboxes and these toolboxes are mathematical toolbox, drag and drop based GUI, Image processing, neural networks etc. The MATLAB is generally used to implement algorithms, plotting graphs and design user interfaces. The MATLAB has high graphics due to which it is used to simulate networks.

The MATLAB has three Commands which are used frequently and these commands are:

1. CLC= The 'clc' stands for clear command window
2. Clear all:- The 'clear all' command is used to de-allocate the variable from the workspace
3. Close all:- The close all is the command which is used to close all the interfaces and return you to default MATLAB interface.

2.0 Result Analysis

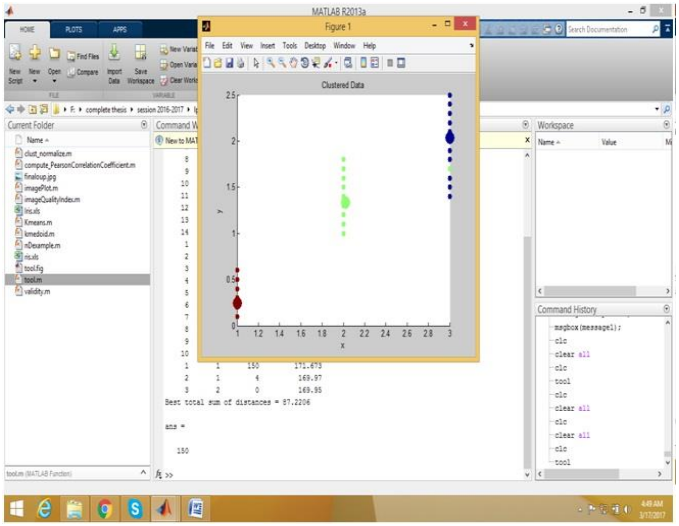

Figure 2.0: Clustered output

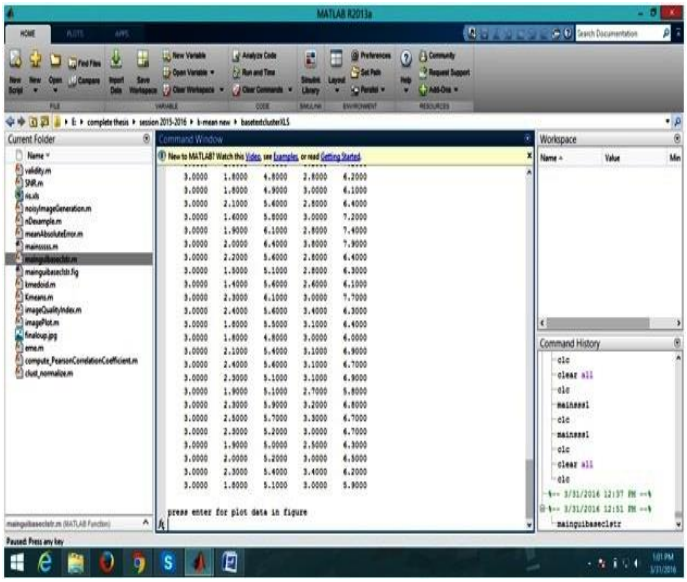

Fig : DATASET Clustere

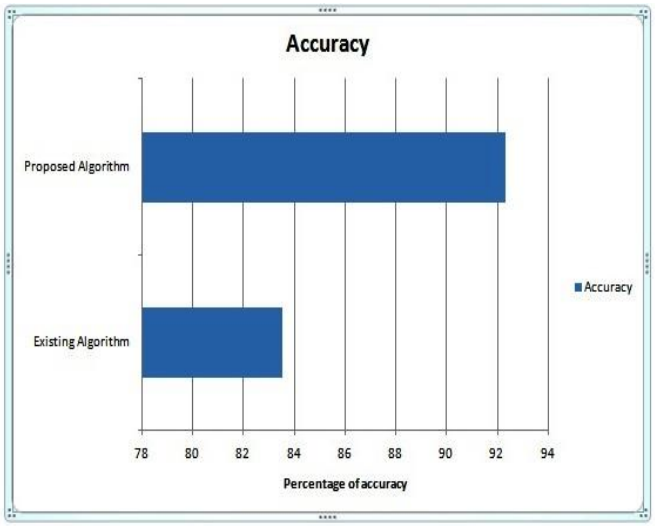

Fig.: Shows Accuracy Comparison

2.1 Future Scope

Following are the various futures prospective of this research.

1. The proposed technique can be applied on the other datasets to test the performance of the improved algorithm.

2. The proposed prediction analysis technique can be compared with the other prediction techniques.

2.2 Conclusion

Data mining is a procedure which is used to extract the important data and patterns from huge amount of accumulated information. There are other names for this process as well, such as knowledge discovery process, knowledge mining from data, knowledge extraction or data /pattern analysis. A number of data mining technologies are used to analyze different types of information. Data mining approach is used in various regions for example user preservation, education structure, manufacture management, medical management, mechanized engineering, decision making and a lot more.

References

[1] Micheline Kamber and Jian Pei Jiawei Han," Data Mining Concepts and Techniques", 2012, 3rd ed.

[2] Mohammed Abdul Khaled, Sateesh Kumar Pradhan and G.N. Dash," A survey of data mining techniques on medical data for finding locally frequent diseases," 2013, International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, pp. 149-153

[3] Abdur Razzak, "A questionnaire survey on infectious disease among hospital patients in Kushtia and Jhenaidah,

Bangladesh," 2011, International Journal of Genetics and Molecular Biology, vol. 3, no. 9, pp. 120-xxx

[4] D. P. Shukla, Shamsher Bahadur Patel and Ashish Kumar Sen, "A literature review in health informatics using data mining techniques," 2014, International Journal of Software & Hardware Research in Engineering, vol. 2, no. 2

[5] V. Gayathri, M.Chanda Mona and S.Banu Chitra, "A survey of data mining techniques on medical diagnosis and research," 2014, International Journal of Data Engineering (OOE) Singapore Journal of Scientific Research (SJSR), vol. 6, pp. 301-310

[6] M.Akhil Jabbar, Priti Chandra and B.L Deekshatulu, "Heart disease prediction system using associative classification and genetic algorithm," 2012, International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies-ICECIT

[7] R. Chitra and V.Seenivasagam, "Review of heart disease prediction system using data mining and hybrid intelligent," 2013, ICTACT Journal on Soft Computing, vol. 03, no. 04

[8] Abhishek Taneja, "Heart disease prediction system using data mining techniques," 2013, Oriental Journal of Computer Science & Technology, vol. 6, pp. 457-466

[9] Hlaudi Daniel Masethe and Mosima Anna Masethe, "Prediction of heart disease using classification algorithms," 2014, Proceeding of the World Congress on Engineering and Computer Science, vol. II, San Francisco, USA

[10] 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

[11] 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

[12] 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

[13] 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

[14] 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

[15] 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

[16] 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

[17] 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

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

[19] M. Hall, E. Frank, G.Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The weka data mining software: An update," 2009, SIGKDD explorations, vol. 11

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

[21] R. Karakis, M. Tez, Y. Kilic, Y. Kuru, and I. Guler," A genetic algorithm model based on artificial neural network.