ISSN:0975-9646

A Comprehensive Overview on Data Mining: Approaches and Applications

Deepti Mishra
Department of CSE
Noida International University , India

Dr. DevpriyaSoni
Department of CSE and IT
JIIT, Noida, India

Abstract - **Data Mining represents the process of decocting a previously unknown information from data. Advancement has lead to a sudden upsurge in large number of algorithms that effectively tackle the regular and computational task of data mining. Data mining can also be considered as retrieving knowledge from large amount of data. Data mining offers great promise in helping organizations to uncover patterns hidden in their data that can be used to analyse the relationship and behaviour of customers and products with future trends.The key idea of this article is to provide an overview of data mining algorithm.The developed software will provide useful and profitable results. The software helps to discover concealed facts and interesting knowledge, to provide help in decision making process.**

Keywords- **Data mining, clustering, classification, K-means, Decision Tree.**

I INTRODUCTION

Data mining can be performed using several models and technologies among which decision trees, neural networks, rule induction, nearest neighbour method and genetic algorithms belong.

Data mining in database systems refers to automatically extracting the predictive information that is not apparently visible. Now a days there is substantial increase in the amount of information or data. This accumulation of data is taking place at an intensive rate. It has been estimated that the amount of information in the world doubles every 20 months and the size and number of databases are increasing even faster. The increase in the use of electronic data gathering devices such as point of scale or remote sensing devices has contributed to this explosion of available data.

Rapid advances in both data storage and data capture technologies have resulted in a marked increase in the amount of data being stored in both the business and scientific sectors. Because of these advances millions of records being generated, each of them containing tens of hundreds of fields. Many of these databases are expanding on daily basis. Traditional analyses of these types of data sets, involving human experts who manually analyse the data, are clearly no longer adequate [1].

The field of data mining has developed in response to the need for machine-oriented, automated methods for analysing large data sets.

The methods and applications of data mining are still in their infancy. The next decade will see a revolution in the use of archived, simulation and near real time data to guide future decisions and research directions.

In this project, we have explored the different methods of data mining. The first and the simplest analytical step in data mining is to describe the data – summarize its statistical attributes (such as means and standard deviations)

This will remove noise and inconsistent data and data will be expressed in well understandable format.

II LITERATURE REVIEW

Data mining involves amassing information from data stored in a database. Identifying irrelevant data from databases is asignificant task.

Data mining is able to handle different kinds of data and has ability to mine data from different sources [1].

A new algorithm which is called as filtering algorithm and based on k-means algorithm has been presented. It builds kd-tree for data points [3].

Data mining contains many techniques, some are based on the concept of supervised learning i.e. classification and some on the concept of unsupervised learning i.e. clustering [4] [6].

Classification and clustering both are similar as they divide the data into groups. But somehow different as in classification all approaches performing it assume some knowledge of data.In clustering there are no class labels. K-means, CURE, Chameleon are some algorithms used to cluster data. These algorithm are based on some or other type of statistical methods [10][11][12]. Decision Tree, Bayesian classification are based on probability. Probability is defined in statistics [11][12].

Another technique of data mining is Association rule mining which discovers interesting association or correlation relationships among large set.

There are different types of algorithms for association rule mining and based on the values of support and confidence [5]. There are many applications of association rule such as catalog design, stock marketing, consumer and product relationship [7]. Apriori algorithm is the part of association rule which is based on candidate generation [8]. An algorithm has been also discussed for large data sets [9].

There are abundant applications of data mining such as fraud detection, anomaly detection, medical diagnosis, image and pattern recognition, detecting faults in industry application and classifying financial market trends, finding consumer and product relationship, outlier analysis, biometric analysis, weather forecasting, trend analysis of customers and travellers, working and analysis on big and high dimensional data, analysis of technical trends in scintific data.

III METHODOLOGY

The basic idea is to provide an outline of data mining algorithm. We bring an outline on different general data mining algorithms that we have implemented that is the clustering and classification algorithm.

The software developed will provide meaningful results. The software carry out the algorithms to learnconcealedsignificant facts and thought-provoking knowledge which will provide help in decision making process.It will detect anomaly present in the data and will provide valid data. The first and simplest analytical step is to describe the data and summarize its statistical attributes (such as means and standard deviations).

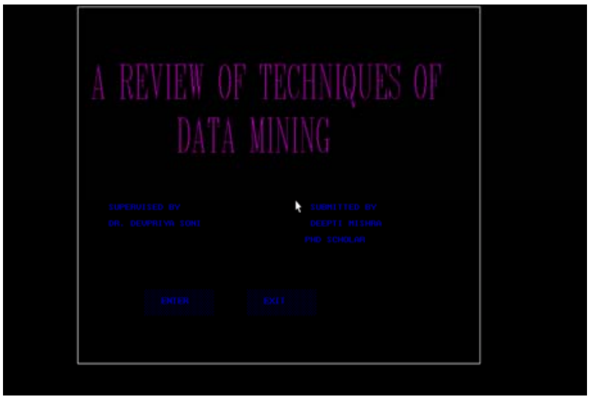
But data description alone cannot provide an action plan. So there is a predictive model which is based on resulted patterns determined by the results, and then test that model based on results.

The latter step is to verify the model.

This work describes different approaches in context by pointing out common aspects and differences, after that with thoroughly investigate there strength and weaknesses.

IV EXPERIMENTAL RESULTS

We implemented the algorithms in C. We conducted all experiments on windows processor intel p3. The software is user friendly and robust. We evaluated methods on artificial data sets.



A REVIEW OF TECHNIQUES OF
DATA MINING

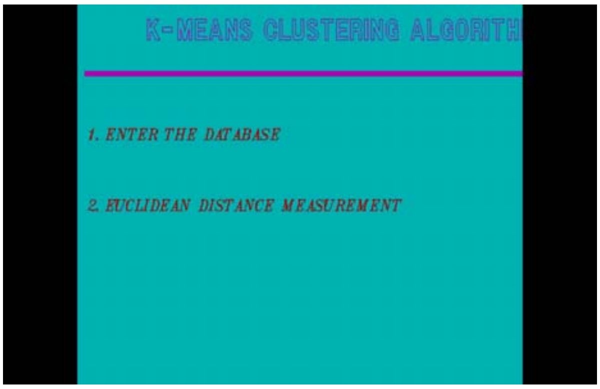
SUPERVISED BY
DR. DEVPRIYA SONI

SUBMITTED BY
DEEPTI MISHRA
PHD SCHOLAR

Fig1 Experimental results 1

REVIEW OF TECHNIQUES of
DATA MINING
CLUSTERING , CLASSIFICATION
AND
ASSOCIATION RULES TECHNIQUES

Fig 2 Experimental results 2

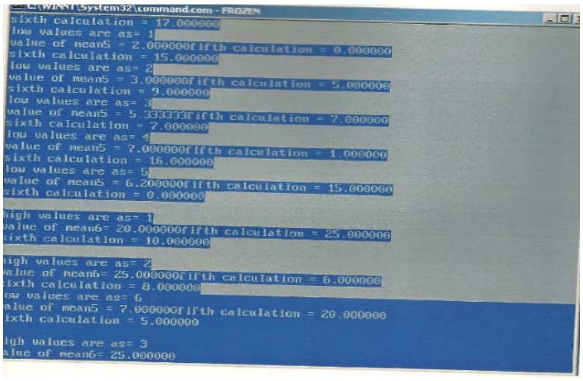


K-MEANS CLUSTERING ALGORITH

1. ENTER THE DATABASE

2. EUCLIDEAN DISTANCE MEASUREMENT

Fig 3 Experimental results 3



C:\WINNT\System32\command.com - FROZEN

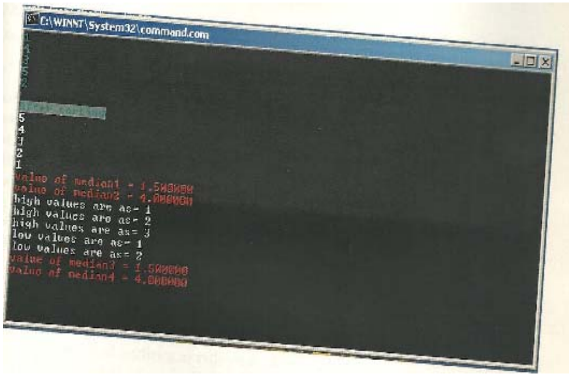
```
sixth calculation = 17.000000
low values are as= 1
value of mean5 = 2.000000fifth calculation = 0.000000
sixth calculation = 15.000000
low values are as= 2
value of mean5 = 3.000000fifth calculation = 5.000000
sixth calculation = 9.000000
low values are as= 3
value of mean5 = 5.333333fifth calculation = 7.000000
sixth calculation = 7.000000
low values are as= 4
value of mean5 = 7.000000fifth calculation = 1.000000
sixth calculation = 16.000000
low values are as= 5
value of mean5 = 6.200000fifth calculation = 15.000000
sixth calculation = 0.000000

high values are as= 1
value of mean6= 20.000000fifth calculation = 25.000000
sixth calculation = 10.000000

high values are as= 2
value of mean6= 25.000000fifth calculation = 6.000000
sixth calculation = 8.000000
low values are as= 6
value of mean5 = 7.000000fifth calculation = 20.000000
sixth calculation = 5.000000

high values are as= 3
value of mean6= 25.000000
```

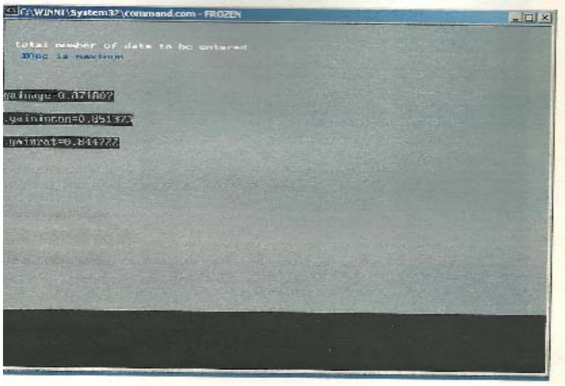
Fig 4 Experimental results of K-means algorithm



C:\WINNT\System32\command.com

```
5
4
3
2
1
value of median1 = 1.500000
value of median2 = 4.000000
high values are as= 1
high values are as= 2
high values are as= 3
low values are as= 1
low values are as= 2
value of median3 = 1.500000
value of median4 = 4.000000
```

Fig 5 Experimental results of K-median algorithm



C:\WINNT\System32\command.com - FROZEN

```
total number of data to be entered
 10pc is maximum

gainage=0.871807

.gaininmon=0.851373

.gainrat=0.844772
```

Fig 6 Experimental results of Decision tree algorithm

V CONCLUSION

Although the data mining has come a long way since the term was coined. Currently data mining is little more than a set of tools that can be used to uncover previously hidden information in database. No one data mining technique is always superior to others in terms of grouping data. However there are advantages and disadvantages to the use of each. Data mining is the field that will continue to attract the interest of researchers. Despite being extensively studied data mining can still be considered infancy as far as its utility applications a concered.

The paper provide a framework on classification and clustering algorithms. This paper emphasis on modication and development of exixting and new algorithms that can serve as a advanced/ sophisticaed tool in carving out the obscure knowledge and trends thereby easing the process of serendipitous new discoveries.

REFERENCES

[1] Chen M.S., Han J., Yu P.S., " Data mining an overview from a database perspective", IEEE Trans. Knowledge and Data Engineering, 8:866-883,1996.

[2] Fayyad U.M., Piatetsky G-Shapiro, Smyth P., Uthuruswamy R., " Advances in knowledge discovery and data mining", AAAI/MIT Press, 1996.

[3] Kanungo T., Mount D.M., Netanyahu N.S., Piatko C.D., Silverman R., Angela Y.Wu, "An efficient K-means clustering algorithm analysis and implementation", IEEE Trans Pattern Analysis and machine Intelligence, vol 24, no 7, jul 2002.

[4] Jain A.K., Murty M.N., Flynn P.J., " Data Clustering : A review", ACM computing survey, vol 31, no. 3, sep 1999.

[5] Hipp J., Guntzer U., Gholamreza, " Algorithm for association rule mining – A general survey and comparison", ACM SIGKDD, jul 2000.

[6] Han J., Kamber M., " Data Mining: Concept and Techniques".

[7] Srikant R., Aggarwal R., "Mining Generalized Association Rules", VLDB, 1995.

[8] Han J., Pei J., Yin Y., " Mining frequent patterns without candidate generation", ACM SIGMOID, 2000.

[9] Aggarwal R., Imielinski T., Swami A., " Mining Association Rules between sets of items in large datasets", ACM SIGMOD, 1993.

[10] Dr.D.C.Sancheti and V.K.Kapoor, "Statistics

[11] David J.Hand, " Statistics and data mining: Interesting Disciplines", ACM SIGKDD, vol 1, jun 1999.

[12] Negi B.S., " Statistics".

[13] Hawkins D.M., "Identification of outlier", Chapman and Hall, Reading, London.

[14] Barnett, V., Lewis T., " Outliers in statistical data ", JohnWiley, 1994.