ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

Performance Analysis of K-Mean Clustering on Normalized and Un-Normalized Information in Data Mining

Richa Rani[1], Manju Bala [2]

PG Student, Dept. of CSE, JCDM College of Engineering, Sirsa, India [1]

Assistant Professor, Dept. of C.S.E., JCDM College of Engineering, Sirsa, India [2]

ABSTRACT: Data mining is used for extract the useful information and Clustering is the concept used to groups which can be creating by identifying similar kind of data and this can done by identify one or more attributes or classes. There are different types of clustering techniques such as K-Means clustering etc. The analysis has been done using the K-Means Clustering technique and by normalizes the data using data mining normalization techniques. Furthermore, the research work is about the study of data such as Normalized data and Un-normalized Data and analyzes the Data using Clustering Algorithm such as K-Means Clustering algorithm. The data mining means extract the useful information from the large dataset and clusters the records. The need for data mining is that there have been too much data, too much technology but don't have useful information. Data clustering is a process of putting similar data into groups.

KEYWORDS: Data Mining.

I INTRODUCTION

Data mining means extracting of useful information from the large pool of dataset. Dataset is the collection of data and Data Mining is the process of analyzing data from different perspectives and summarizing it into useful information [1]. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. This process is known as KDD (Knowledge Discovery in Database) and clustering method is the example which is required for make data clusters [2].

Knowledge Discovery in Databases:

The main idea in KDD is to discover a high level knowledge (abstract knowledge) from lower levels of relatively raw data, or to discover a higher level of interpretation and abstraction than those previously known. Data mining is the part of knowledge discovery in database. It deals with the mining hierarchy in which it involve text mining as well as web mining. Thus, starting from the most general mining hierarchy, firstly knowledge discovery in database is placed then data mining is placed and then text mining is placed in the hierarchy. It is a process of extracting hidden information, useful knowledge or interesting relations from some data. Obviously, the nature of this data determines the hierarchy levels at which it mine. Hence data mining is the nontrivial extraction of implicit, previously unknown, and potentially useful information from a large amount of data.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

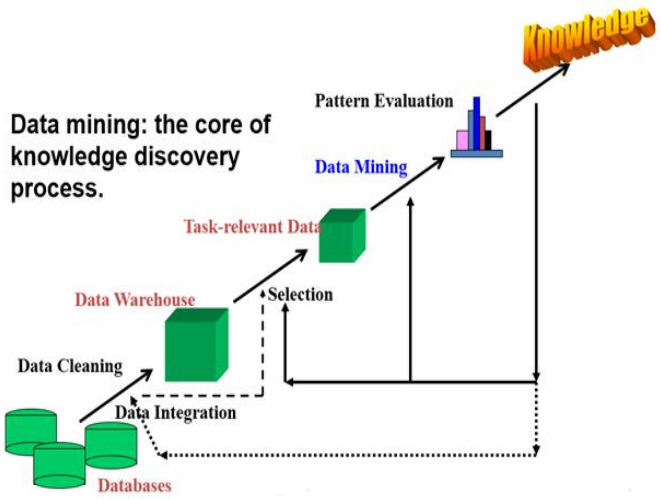

Fig. 1 Data Mining Process

II LITERATURE REVIEW

Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning,pattern recognition,image analysis, information retrieval, and bioinformatics. [1]
There are various method in clustering these are followed:

- PARTITIONING MATHOD
  - oK-mean method
  - oK-Medoids method

- HIERARCHICAL METHODS
  - o Agglomerative
  - o Divisive

- GRID BASED

- DENSITY BASED METHODS
  - oDBSCAN

.I use K-mean Clustering. K-mean is a widely used partitioned clustering method in the industries. The K-means algorithm is the most commonly used partitioned clustering algorithm because it can be easily implemented and is the most efficient one in terms of the execution time.[4]

It is a centroid based technique. This algorithm takes the input parameters k and partition a set of n objects into k clusters that the resulting intra-cluster similarity is high but the inter-cluster similarity is low. The method can be used by cluster to assign rank values to the cluster categorical data is statistical method. K mean is mainly based on the distance between the object and the cluster mean. Then it computes the new mean for each cluster.

[1] Narendra Sharma, Aman Bajpai 2012,"Comparison the various clustering algorithms of weka tools", IJETAE.
Author has explained that how the data classified using different cluster techniques. They analyzed the data from different perspectives and summarizing it into useful information. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it,and summarize the relationships identified. Weka is a data mining tools. It is contain the many machine leaning algorithms. It is provide the facility to classify our data through various algorithms. They have been used the various

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

clustering algorithms. Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar to each other than to those in other clusters. Author's main aim is to analyze data by different- different clustering algorithms of weka and find out which algorithm will be most suitable for the users.

[2] Manish Verma, Mauly Srivastava,Nidhi Gupta(2012),"A Comparative Study of Various Clustering Algorithms in Data Mining",(IJERA).

Author has been assimilated the knowledge about clustering of data in data warehouse. Author explained clustering that the Data clustering is a process of putting similar data into groups. A clustering algorithm partitions a data set into several groups such that the similarity within a group is larger than among groups. This paper reviews six types of clustering techniques- k-Means Clustering, Hierarchical Clustering, DBScan clustering, Density Based Clustering, Optics, EM Algorithm. These clustering techniques are implemented and analyzed using a clustering tool WEKA. Performance of the 6 techniques are presented and compared. Author analyzed the results after testing the algorithms and running them under different factors and situations. Author concluded that the Performance of K-Means algorithm increases as the RMSE (root mean square error) decreases and the RMSE decreases as the number of cluster increases and performance of K-Means algorithm is better than Hierarchical Clustering algorithm but all the algorithms have some ambiguity in some noisy data when clustered.

[3] Dogdas, T.,Akyokus, S.(2013), "Document clustering using GIS visualizing and EM clustering method",IEEE.

The author has been explained the concept of EM Clustering algorithm and considered the multidimensional projection method. Author uses expectation-maximization clustering algorithm and a simple multidimensional projection method for visualization and data reduction. The multidimensional data is projected into a 2D Cartesian coordinate system. They run EM and K-Means algorithms on the transformed data. The system uses Microsoft Spatial Data Base Engine as a GIS tool for visualization and also used Expectation-Maximization (EM) and K-Means clustering algorithms of the Microsoft Analysis Services. The simple multidimensional projection method used in this paper tries to preserve the similarity relationships in original datasets.

[4] Swasti Singhal, Monika Jena(2013),A Study on WEKA Tool for Data Preprocessing, Classification and Clustering, IJITEE.

Author has explained the basic principles of data mining. Data mining concept is used to analyze the data from different angle, categorize it and finally to summarize it. They explained the data mining has been increasingly become very interesting and popular in terms of all application. The need for data mining is that there is too much data, too much technology but difficult to analyze the information. Data mining software allows user to analyze data. They introduce the key principle of data pre-processing, classification, clustering and introduction of WEKA tool. Weka is a data mining tool. They has been described the steps of how to use WEKA tool for these technologies. It provides the facility to classify the data through various algorithms.

[5] Vats, P. Hmritm, Mandot, M.,(2014), "A Comparative Analysis of Various Cluster Detection Techniques for Data Mining", ICESC(IEEE).

The author has been assimilated the knowledge about knowledge discovery in data mining. Data mining is a knowledge discovery technique which is used for exploring the new facts and relationships among data and enables a user to uncover hidden information among available datasets. Cluster detection is one of the major techniques, which is used for data mining. In the Cluster detection techniques, User performs mining of data by searching for cluster of elements that are similar to each other. Each implementation of the cluster detection techniques adopts a method of comparing the value of individual datasets with those in their centroids. In this paper, Author has enlisted a few of them. Based on certain parameters, they have carried out a comprehensive analysis of various clustering techniques.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

III PROBLEM FORMULATION

The clustering implement on given set of records e.g. objects, observations and organize them into clusters (groups, classes). The quality of a clustering result also depends on both the similarity measure used by the method and its implementation. The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns.
The problem of clustering if the information is irrelevant or noisy, unreliable, then knowledge discovery during training is more difficult. The irrelevant information means that the computation is now proportional to the problem size instead of the problem dimension. However, the actual cluster centers are not necessarily located at one of the data points, but in most cases it is a good approximation, especially with the reduced computation this approach introduces.

1. Irrelevant information in dataset.
2. Cost to group the information.
3. Computation Time to generate the Clusters.
4. Difficult to analyze the noisy Data.

Another problem is that Structure of database; Real life Data may not always contain clearly identifiable clusters.Also the order in which the tuples are arranged may affect the results when an algorithm is executed if the distance measure used is not perfect.With a structureless data(for eg. Having lots of missing values), even identification of appropriate number of clusters will not yield good results.

IV OBJECTIVES

The objective of cluster analysis is to assign observations to groups (clusters) so that observations within each group are similar to one another with respect to variables or attributes of interest and the groups themselves stand apart from one another. This provides measures and criteria that are used for determining whether two objects are similar or dissimilar.
The K-Means Clustering algorithm will be implemented with dataset normalization concept and improved the clusters in less time. The result will be analyzed by WEKA tool with backend relational database and results will be generated.

V METHODOLOGY

1. Study of K-Means Clustering Algorithm.
2. Identify and Analyze benefits of Algorithm.
3. Study of Relational Database and Normalization Techniques.
4. Study of Normalization Advantages.
5. Design an efficient Dataset with Normalization Technique.
6. Use Dataset with K-Means Clustering Algorithm for enhanced performance.
7. WEKA and SQL Server Tool will be used for Implementation.
8. Analyze the Results.

VI FACILITIES REQUIRED FOR PROPOSED WORK

a. Windows Vista/7 OS

Windows 7 was primarily intended to be an incremental upgrade to the operating system, intending to address Windows Vista's critical reception (such as performance improvements), while maintaining hardware and software compatibility. Windows 7 was a major success for Microsoft. It support for systems using multiple heterogeneous graphics.
Windows Vista include an updated graphical user interface and visual style

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

b. WEKA Tool

WEKA Explorer is an application that provides the following functionality of Dataset Management, loading data, feeding them to classifiers, filters, storing the results of classification, apportioning data between training and testing subsets..

WEKA is a software tool that was developed at the University of Waikato in New Zealand and written on Java . WEKA is platform-independent, open source and user friendly with a graphical interface that allows for quick set up and operation, WEKA is a collection of machine learning algorithms for data mining tasks.

Fig. 2 Main Window of WEKA Tool

WEKA tool contains Attribute-relationship file format (.arff) and .csv file of the data set. Data set consists of attribute names, types, values and the data. In WEKA, the data objects are called as instances and features of data are considered as attributes.

c. SQL Server Express 2008

Microsoft SQL Server is a relational database management system developed by Microsoft. As a database server, it is a software product with the primary function of storing and retrieving data as requested by other software applications which may run either on the same computer or on another computer across a network (including the Internet).

SQL Server 2008 Express includes the Import and Export Wizard, making it much easier to transfer data into and out of SQL Server 2008 Express databases. which is a freely distributable database that shares the same.

VII CONCLUSION

The clustering involves partitioning a given dataset into some groups of data whose members are similar in some way. The usability of cluster analysis has been used widely in data recovery, text and web mining, pattern recognition, image segmentation and software reverse engineering. In this dissertation, K-Means clustering algorithmic discussed on normalized and un-normalized data. The Normalization techniques of data mining have been implemented such as Min-Max, Z-Score and Decimal Scaling Normalization. Time complexity can be analyzed by identify the timing of clusters calculation and particular normalization techniques can be used for further prediction and analysis in data mining.

In future work, we can work on the time complexity factor. Outlier detection and removal is another area where work can be done. There must be some method to detect the outliers and can be removed if desired. The research could be extended in this direction to revise the clustering algorithm, which can reduce the complexity of the proposed algorithm.

REFERNCES

[1]Alexander Strehl, Joydeep Ghosh et al.: Impact of Similarity Measures on Web-page Clustering , AAAI-2000: Workshop of Articial Intelligence for Web Search, July 2000.

[2]Lee Y. C., Antonsson K. E. : Dynamic Partitional Clustering Using Evolution Strategies, IEEE Computer, page 2716-2721, 2000.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

[3]Michael Steinbach, George Karypis et al.: A comparison of document clustering techniques, In KDD Workshop on Text Mining 2000.
[4]Doug Beeferman : Adam Berger "Agglomerative Clustering of a Search Engine Query Log, sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2000.
[5]VeenmanJ.Cor, Marcel J. T. et al. : A Maximum Variance Cluster Algorithm, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. ,September 2002
and Machine Intelligence, VOL. 27, NO. 8, August 2005.