ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

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

Richa Rani[1], Mrs. Manju Bala[2]

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

Asst Professor, CSE, JCDM College of Engineering, Sirsa, India[2]

Abstract: The research includes the concept of data mining, clustering and clustering techniques. 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, 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 basic principles of data mining is to analyze the data from different angle, categorize it and finally to summarize it. 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, K-Means clustering algorithm, Normalized data, Un-normalized Data.

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]. The information which has been analyzed can be used for increase revenue of Organization and industry level. 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]. This Refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process.

Knowledge Discovery in Databases

Knowledge discovery in databases and data mining are sometimes confused with each other. 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 not evident where to put web mining. Obviously web data is more than textual data.

It is used to check whether it is a sub-division of data mining or a sub-division of knowledge discovery. 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.

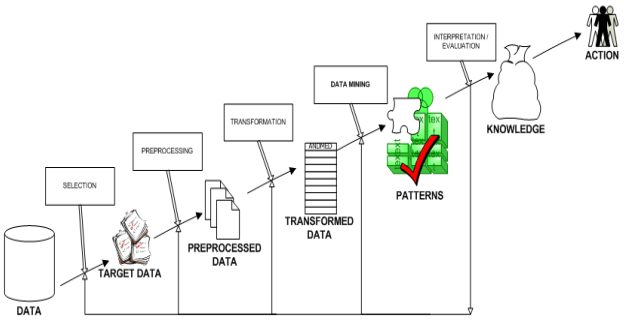

Fig. 1 Data Mining Process

Association

Association (or relation) technique is for make a simple correlation between two or more items, frequently of the same type to identify patterns. For example, when tracking people's buying habits, it can identify that a customer always buys cream when they buy strawberries, and therefore suggest that the next time that they buy strawberries they might also want to buy cream. Building association or relation-based data mining tools can be achieved simply with different tools.

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

Database

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

C_1

Itemset	Support
{1}	2
{2}	3
{3}	3
{5}	3

C_2

Itemset	Support
{1 3}*	2
{1 4}	1
{3 4}	1
{2 3}*	2
{2 5}*	3
{3 5}*	2
{1 2}	1
{1 5}	1

C_3

Itemset	Support
{1 3 4}	1
{2 3 5}*	2
{1 3 5}	1

Fig. 2 Multicolumn Association

Classification

Classification technique for classify the objects in predefined classes. This can be used for build up an idea of the type of customer, item, or object by describing multiple attributes to identify a particular class. For example, it can easily classify cars into different types (sedan, 4x4, convertible) by identifying different attributes (number of seats, car shape, driven wheels). Given a new car, it can apply it into a particular class by comparing the attributes with known definition. This principle can be applicable to customers for classifying them by age and social group.

Clustering

In clustering, the groups can be creating by identifying similar kind of data and this can done by identify one or more attributes or classes. Clustering is using one or more attributes as basis for identifying a cluster of correlating results [1]. Clustering is useful to identify different information because it correlates with other examples so it can see where the similarities and ranges agree. Clustering can work both ways. It can be assume that there is a cluster at a certain point. For example, a sample of sales data compares the age of the customer to the size of the sale.

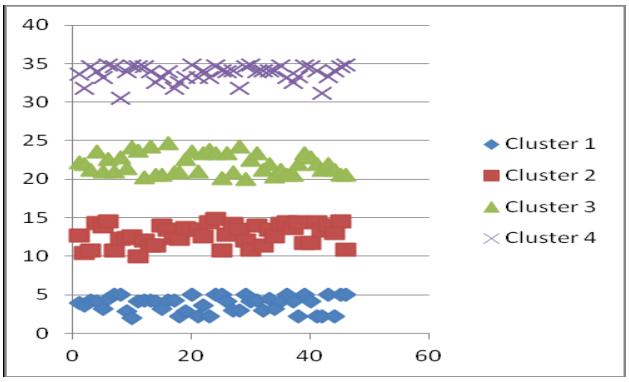

Fig 3 Example of Clustering

It is not unreasonable to expect that people in their twenties before marriage and kids, fifties, and sixties when the children have left home, have more disposable income. In the example, it can be identified two clusters, one around the US$2,000/20-30 age group, and another at the US$7,000-8,000/50-65 age group. In this, it can be hypothesized and proved hypothesis with a simple graph that can be create using any suitable graphing software for a quick manual view. More complex determinations require a full analytical package and especially for automatically decisions on nearest neighbor information. Plotting clustering in this way is a simplified example and called nearest neighbor identity. It can be identify individual customers by their literal proximity to each other on the graph. It's highly likely that customers in the same cluster also share other attributes and you can use that expectation to help drive, classify, and otherwise analyze other people from your data set.

Process of Data Mining

To extract the information, the first step is to process the data and for this, there is need to follow the process of Data mining. Data mining is for processing data and identifying patterns and trends in that information for decision making. Big data caused an explosion in the use of more extensive data mining techniques because the size of the information is much larger and because the information tends to be more varied and extensive in its very nature and content. With large data sets, it is no longer enough to get relatively simple and straightforward statistics out of the system. With 30 or 40 million records of detailed customer information, knowing that two million of them live in one location is not enough.

II. LITERATURE REVIEW

Evolution Strategy (ES) implementing variable length genomes is developed to address the problem of dynamic partitional clustering given by author Lee and Anton son (2000) [1]. As opposed to static, dynamic partitional clustering does not require the prior specification of the number of clusters. The proposed ES implements variable length genomes that allow the algorithm to effectively search for both optimal clusters center position and cluster number. Cluster number is optimized during runtime, such clustering re referred to as dynamic. The proposed ES is developed as a general framework for dynamic partitioning; it would be interesting to observe its performance using different fitness criteria. Furthermore, Agglomerative clustering of a search engine query log given by author Beefer man et al. (2000) [2]. It introduces a technique for mining a collection of user transactions with an Internet search engine to discover clusters of similar queries and similar URLs. The information that the user exploit is click-through data each record consists of a user's query to a search engine along with the URL which the user selected from among the candidates offered by the search engine. By viewing this dataset as a bipartite graph, with the vertices on one side corresponding to queries and on the other side to URLs, one can apply an agglomerative clustering algorithm to the graph's vertices to identify related queries and URLs and the author stated a comparison of document clustering techniques given by author Steinbach and Karypis (2000) [3].It presents the results of an experimental study of some common document clustering techniques: agglomerative

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

hierarchical clustering and K-means. The results indicate that the bisecting K-means technique is better than the standard K-means approach and as good as or better than the hierarchical approaches that the user tested. [1]

Author explained similarity measures in web-page clustering. Clustering of web documents enables automated, categorization, and facilitates certain types of search assimilated by the author Strehl and Ghosh (2000) [4]. Any clustering method has to embed the documents in a suitable similarity space. Here the comparative study of the impact of similarity metrics on cluster quality, possibly because the popular cost criteria do not readily translate across qualitatively distance metrics. Then it compares four popular similarity measures in conjunction with several clustering techniques on high dimensional sparse data representing web documents. Performance is measured against a human-imposed classification into news categories and industry categories. Here a number of experiments are conducted that use t-tests to assure statistical significance of results. Furthermore, author proposed a novel clustering algorithm called Maximum Variance Clustering Algorithm(MVC) for partition based clustering that minimizes the within cluster scatter with a constraint on the cluster variance and minimizes the sum-of-squared-error criterion while imposing a hard constraint on the cluster variance given by Veenman(2002) [5].Accordingly, in contrast to many other cluster algorithms, this method finds the number of clusters automatically and uses a maximum variance constraint instead of the number of clusters as parameter and they have been classify the documents when there are a small number of features available in the document given by the author Zhang and Yang's(2003) [6]. The algorithm first extracts and ranks salient phrases as candidate cluster names, learned from the training data. It then assigns documents to relevant salient phrases to form clusters. Lastly, the candidate clusters are merged to generate the final clusters. Clustering has been effective for displaying web search results and should be just as effective in database query results.[2]

Author describes clustering approach for students to help in the evaluation of learning process given by the author Agathe(2004) [7]. Here it shows how clustering techniques can be applied to student answers generated from a web-based tutoring tool. In particular it is interested in extracting clusters of students based on the mistakes they made using the tool, with the aim of obtaining pedagogically relevant information and providing this feedback to the teacher. Furthermore, a novice approach of clustering model, called a general c-means algorithm (GCM), by extending the definition of mean from a statistical point of view given by author Jian Yu (2005) [8] . In order to study GCM, author make two assumptions about the partitional clustering algorithms. The first one is that each subset is often expected to have a different prototype than others when the data set is clustered into c (c>1) subsets in general cases. According to this assumption, it defines the undesirable solution of GCM (USGCM). And, the second assumption is USGCM should not be stable to fixed point of GCM. Author explained clustering by compression given by Cilibrasi& Vitanyi (2005) [9]. The method does not use subject-specific features or background knowledge, and works as follows: Firstly, it determines a parameter-free, universal, similarity distance, the normalized compression distance (NCD), computed from the lengths of compressed data files. Secondly, it applies a hierarchical clustering method. To extract a hierarchy of clusters from the distance matrix, it determines a dendrogram by a new quartet method and a fast heuristic to implement it. [3]

Author suggested the idea of automatic database clustering given by author Sylvain et al. (2006) [10].This is an important area because with data proliferation, human attention has become a precious and expensive resource. Here, goal is to develop an automatic and dynamic database clustering technique that will dynamically re-cluster a database with little intervention of a database administrator (DBA) and maintain an acceptable query response time at all times. The framework for an automatic and dynamic mixed database clustering technique currently under development called Auto Clust also presented. Auto Clust mines closed item sets to create clusters of attributes and uses data mining clustering to perform record clustering within each attribute cluster. Auto Clust is triggered when a drop in the query response time is detected. Furthermore, author assimilated the knowledge about the different types of normalization given by author Al Shalabi, L. ; Shaaban, Z. 2006[11]. Each of which was tested against the ID3 methodology using the HSV data set. Number of leaf nodes, accuracy, and tree growing time are three factors that were taken into account. Comparisons between different learning methods were accomplished as they were applied to each normalization method. A simple matrix was designed to check for the best normalization method based on the factors and their priorities. Recommendations were concluded and author presented an approach of web data clustering using Fuzzy C-Mean (FCM) and proximity hints from text as well as hyperlink-structure assimilated by the Agarwal (2008) [12]. In this study, FCM clustering along with proximity hints (P-FCM) to the Web pages for clustering is used. It provides proximity hints using a new approach of combining textual information, hyperlink structure and co-citation relations into a single similarity metric. It provides the result of Web-basedexperiments to show the significance of proximity hints during P-FCM functioning. These observations suggest that with the combination of textual and hyperlink-structure information, the user can improve the clustering done by FCM. [4]

III.PROPOSED METHODOLOGY

Problem Statement

Data clustering is concerned with the partitioning of a data set into several groups such that the similarity within a group is larger than that among groups. This implies that the data set to be partitioned has to have an inherent grouping to some extent; otherwise if the data is uniformly

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

distributed, trying to find clusters of data will fail, or will lead to artificially introduced partitions. Another problem that may arise is the overlapping of data groups. Overlapping groupings sometimes reduce the efficiency of the clustering method, and this reduction is proportional to the amount of overlap between groupings.

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.

Objectives

Data mining may be viewed as the extraction of patterns and models from observed data or a method used for analytical process designed to explore data. There are many different methods, which may be used to predict the appropriate class for the objects. The majority of data mining techniques can deal with different data types. There are number of techniques and many variations of the methods, one of the techniques from the mentioned group is almost always used in real world deployments of data mining systems. The objective of clustering analysis is to find segments or clusters and to examine their attributes and values. Numerous algorithms are used for clustering. The clustering technique defines the classes and puts objects in each class accordingly.

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.

IV. RESULTS

This shows the clustering of data using Clustering algorithm. The different clusters have been generated on different data with different normalization techniques. The results before normalization are:

Attribute	Full Data (2515)	Cluster# 0 (1733)	1 (782)
SrNo	1258	1642.8569	405.1138
G1	276.2056	217.3434	406.6508
G2	279.0325	219.873	410.1367
G3	273.3176	214.5524	403.5479
G4	276.189	217.1784	406.9632

```
Time taken to build model (full training data) : 0.16 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      1733 ( 69%)
1       782 ( 31%)
```

Fig 4 Two Clusters before Normalization on Data

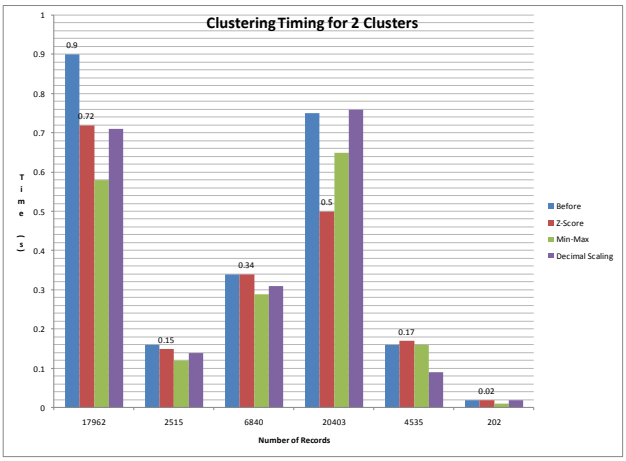

Fig 5 Comparison Clustering Timing Graphs for 2 Clusters

V. 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

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

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.

REFERENCES

[1] Alexander Strehl, Joy deep 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., Anton son K. E. : Dynamic Partitional Clustering Using Evolution Strategies, IEEE Computer, page 2716-2721, 2000.
[3] Michael Steinbach, George Karypis et al.: A comparison of document clustering techniques, In KDD Workshop on Text Mining 2000.
[4] Doug Beefer man : 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
[6] Zhang and Yang's: Classification of documents, 2003.
[7] Mercer Agathe: Introduced an idea for "Clustering Students to Help Evaluate Learning, 2004
[8] Yu Jian: General C-Means Clustering Model, IEEE Transactions on Pattern analysis and Machine Intelligence, VOL. 27, NO. 8, August 2005.
[9] Rudi Cilibrasi, Paul M.B. Vitanyi: Clustering by compression. IEEE Transactions on Information Theory, Vol. 51, No. 4, pp 1523-1545, 2005.
[10] Sylvain Guinepain: Automatic Database Clustering Using Data Mining, 23rd International Workshop on Database and Expert Systems Applications 2006, pp. 124-128, doi:10.1109/DEXA.2006.32