Volume 3, Issue 11, November 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Survey on Overview of Clustering Techniques for Data Mining

Ms. K. R. Khandait, Ms. A. Bhattacharya

Computer Science and Engineering

GHRIETW, Nagpur, India

Abstract-Clustering is the unconfirmed categorization of patterns into groups .This groups are known as clusters. Group consists of objects that are very similar among themselves whereas less similar compared to objects of the other cluster. The present article provides a Review of the existing literature on clustering techniques for data mining. Typically, data mining is the process of nontrivial extraction of previously indefinite, potentially valuable and implicit information from data. Data clustering is nothing but the one of the technique for collecting similar data into groups. So, in this paper, four popular researched data clustering techniques are discussed- Hierarchical, Partitioning, Expectation-maximization and soft-computing clustering techniques. Again we have explained some essential clustering algorithm's applications.

Index Terms- Clustering, partitioning, data mining, hierarchical clustering, knowledge discovery.

I. INTRODUCTION

The data mining is the process of extracting or "mining" data from huge data sources. Mining finds a tiny set of valuable nuggets from a great deal of raw material. So working with both "data" and "mining" became a popular option. Expressions having slightly different meaning or nearly similar to data mining are knowledge extraction, knowledge mining from data, KDD process and pattern analysis. And thus for finding similarities in data and gathering similar data into groups, data clustering is very convenient technique . Clustering partitions a data set into several groups such that the similarity within a group is larger than that among groups [6].

The idea which is straightforward in nature and is very close to the normal human being's thinking is data grouping, or clustering ; whenever we are offered with a large amount of data, we usually like to summarize this huge data into a small groups or categories, to perform the further analysis of that data. Most of the time, data collected , appear to have some inherent properties that group themselves to natural groupings, in many problems. But, to find these groups or try to classify the data is not a simple job for humans till the data having low dimensionality (two or three dimensions at maximum.) Because of this only, soft computing methods have been proposed to overcome such problem. Such methods that are used to overcome the drawback of the low dimensionality problem are called as the "Data Clustering Methods".

The fixed procedure that we should follow for mining the data from huge databases is data collection first, after that data cleaning is required and once the data is cleaned after that only we should interpret the result. Another reason for clustering is to discover relevance knowledge in data. [7] implemented a Case Based Reasoning (CBR) system based on a Growing Cell Structure (GCS) model. Knowledge base can store data which is indexed or categorized by cases, is known as the Case Base. Cases in each group is assigned to a some group. Using a Growing Cell Structure (GCS) data can be added or removed based on the plan which are used. When a query is offered to the model, the system will get the most relevant cases from the case base based on how queries and cases are close to each other.

II. RELATED WORKS

As mentioned earlier, data clustering is apprehensive with the partitioning of a data set into several groups such that objects within the cluster are more similar where as objects of different cluster are less similar. This implies that the data set to be partitioned into natural grouping to some extent or when the data is homogeneously dispersed process of finding clusters of data will fail or will form the artificially introduced partitions. Problem that can arise is overlapping of the clusters, which can decrease the efficiency of the clustering technique and this decrease in efficiency results in clusters overlaps. Today, it can be found that many research work have done on clustering techniques in data mining domain. [1]-[8] offered a board view for the different clustering techniques and comparison between them considering the advantages and disadvantages of the same.

Different soft computing methodologies like Fuzzy sets are usually appropriate to handle the issues related to understandability of patterns, mixed media information, noisy data can provide approximate solutions faster[3]. Information overload problem can be solve by biomedical text retrieval ,which will help the researchers, by finding the extrapolative relationships among different pieces of extracted data, and after that we can apply the data mining algorithm to increase the efficiency of the information retrieval technique or the process[4]. The various algorithms are compared on the basis of factors like as size of datasets, member of cluster, type of dataset and type of software used. Performance, quality and accuracy of the algorithms are discussed with respect to all these four factors [5]. Reference [1]

has given a general schematic of the architecture of participation anticipating system in presidential election by using KNN, Classification Tree and Naïve Bayes and tools orange based on crisp and by anticipating the political behaviour of people in elections can determine the future prospect of each country domestic and foreign policies and characterize domestic and international relationships. Total four concepts are studied and given in detail and that are knowledge resource; knowledge types and/or knowledge datasets; data mining tasks; and data mining techniques and applications used in knowledge management and along with this terms data mining functionalities are surveyed [2]. Graham Cormode has given the key concepts in data mining along with that he has discussed the things like as clustering problems and some of the other clustering techniques.

As tremendous amount of work is being done on the clustering techniques in data mining domain , in which the fuzzy c-means is nothing but the popular technique of clustering and the reference[12] suggests that the the fuzzy c-means (FCM) algorithm is a useful tool for clustering real dimensional data, yet we cannot apply it to the incomplete data. In this clustering techniques, soft computing technique plays a very important role and it is because due to the fact that soft computing techniques allows the pattern to belong to more than 1 cluster. Reference [13] put forward the significance of soft computing including fuzzy logic (FL), artificial neural networks (ANNs), genetic algorithms (GAs), and (RS), rough sets highlighted. A study of the prose on "soft web mining", which is existing is provided with the commercially available systems. The term data clustering is indirectly refers to the text clustering and the reference [14] tells that the conceptual text clustering extends to web documents, Containing various markup language formats associated with the documents. Based on the markup languages like presentations, procedural and descriptive markup, the web document's text clustering is done efficiently using the concept-based mining model.

III. DATA MINING

Prediction and description are the two key objectives of any data mining. To complete these objectives, analysts usually use six data mining techniques: classification, clustering, regression, dependency, summarization, modelling, and deviation detection [8]. These are discussed next. Classification involves the mapping of data values into classes using a resulting function. Class representation of past trends enabling predictions based on categorization dependencies are provided by this. Regression is similar to classification in that it uses a function to provide grounded predictions. while, regression enables measures of correlation between two variables. This can be converted to the business background through uses such as predicting future sales. Clustering is a expressive tool used to collect common data objects. The concept has common applications in the marketing discipline, where group of customers with common attributes are formed [9]. Summarization uses an array of techniques to explain data. Simple techniques include tabulation of aggregates, whereas advanced techniques includes functional relationships, summary rules derived from a data source . Dependency (between variables): It offers stronger grounds for predicting uncertain outcomes such as human behaviour. Deviation detection examines the comparative dynamics of patterns in data, usually over a time series. This task delivers both a predictive and descriptive benefit. It highlights the seasonality or periodic market characteristics. Then we can make changes to that[15].

KDD PROCESS

Most of the time data mining term is treated as synonym for Knowledge Discovery from Data, or KDD. KDD subject is evolved, and continuously evolving, from the connection of research fields like databases, machine learning, statistics, artificial intelligence, pattern recognition, reasoning with doubts, data visualization, machine discovery. Theories, algorithms, and methods are included in KDD process from all these fields. It has applications in the area banking, marketing, telecommunications, finance and manufacturing.

IV. OVERVIEW OF CLUSTERING TECHNIQUES

A. *Hierarchical clustering algorithms*

The two-dimensional data set illustrated in Figure 1 for hierarchical clustering. shows seven patterns A, B, C, D, E, F, and G in three clusters are shown in the figure. A hierarchical algorithm form a dendrogram representing the nested collection of patterns and similarity levels at which groupings change. The respective dendrogram corresponding to the seven points are shown in Figure 2. This dendrogram is formed by using single-link algorithm. The dendrogram can break at various levels to form clustering of the data. Most hierarchical clustering algorithms are variants of the minimum-variance, complete-link and single-link algorithms. From these algorithms, the single-link and complete-link algorithms are most popular. These algorithms are differing in the way they characterize the similarity between a pair of clusters.

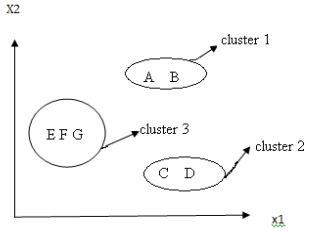

Figure 1. Three clusters with 7 patterns

In the single-link method, the distance between two clusters is the minimum of the distances between all pairs of patterns drawn from the two clusters, in single-link method. i.e., one pattern from the first cluster, the other from the second. And in the complete-link algorithm, the distance will be maximum between two clusters of all pair wise distances between patterns in the two clusters. In either cases, based on minimum distance criteria, both of the two clusters will be merged to form a larger cluster. The complete-link algorithm forms tightly bound or compact clusters. After comparison, the single-link algorithm suffers from a chaining effect. It has a tendency to produce clusters that are elongated [10].

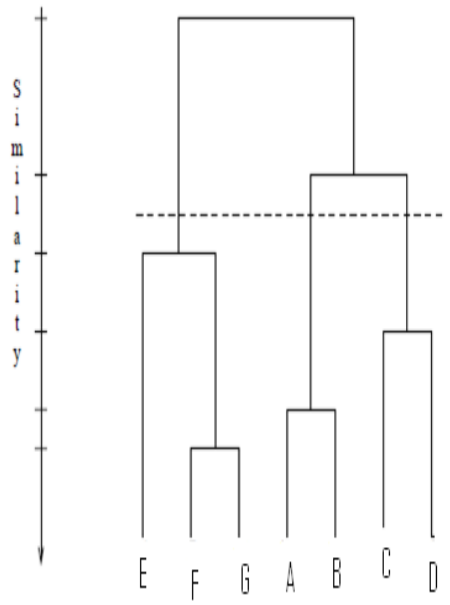

Figure 2. Dendrogram using single-link method

B. ***Partitioning Methods***

Partitioning methods transfer objects by moving them from one cluster to another, starting from an initial partitioning. In Such methods, pre-setting of the number of cluster is required by the user. An exhaustive enumeration process is required to achieve global optimality in partitioned-based clustering. Certain greedy heuristics are used in the form of iterative optimization. Points between the *k* clusters iteratively relocates by relocation method. Some of the popular types of Partitioning methods are:

1) Graph-Theoretic Clustering and

2) Error Minimization Algorithms.

Clusters (using graphs) are formed by the graph theoretic methods. The instances represented as nodes are connected by the edges of the graph. A well-known graph-theoretic algorithm is based on the Minimal Spanning Tree (MST). The edges whose weight is significantly larger than the average of nearby edge lengths, are called as inconsistent edges. There is some connection between graph theoretic clustering and hierarchical methods :

a) The sub-graphs of the MST, of Single-link clusters of the data instances. Every sub-graph is a connected component, namely a set of instances in which each object is connected to at least one other member of the set, so that the set is maximal with respect to this property. According to some similarity threshold, These sub-graphs are formed.

b) A maximal complete sub-graph is a sub-graph such that each node is connected to every other node in the sub-graph and the set is maximal with respect to this property. Complete-link clusters are maximal complete sub-graphs, formed using a similarity threshold.

Algorithms, which tend to work well with remote and dense clusters, are the most sensitive and normally used methods. In this first thing is to find a clustering structure that minimizes a certain error criterion measuring, "distance" of each instance to its representative value. The K-means and K-medoid are the two algorithm which comes under this category: K-means algorithm takes the input parameter, k, and partition the set of n objects into k cluster so that resulting intra-cluster similarity is high but similarity is less in intra-cluster. Similarity among cluster is measured with respect to the mean value of the objects in cluster, which can be viewed as a cluster's centroid or center of gravity. K-means algorithm is sensitive to outliers because an object with extremely large value may distort the dividation of the data.[11]

Compared to K-means, In K-medoid, instead of taking mean value of the objects in cluster as reference-point, we can pick actual object to represent clusters, using one representative object per cluster. Remaining clustering is with representative object to which it is mainly similar.

C. ***Expectation-Maximization clustering algorithm***

EM is well-established clustering algorithm in the statistics community. EM is the distance-based algorithm that assumes the dataset can be modeled as a linear combination of multivariate normal distributions and the algorithm finds the distribution parameters that maximize the model quality measure called log likelihood.

EM is useful to cluster data compared to others because it is linear when we consider database size, it can accept the most wanted number of cluster, it converges fast and gives the better initialization, again it has strong statistical basis, it is robust to noisy data and finally it also can handle the high dimensionality [5].

D. Soft-computing Methods

Fuzzy clustering and Simulated Annealing for Clustering are reviewed in this paper. As soft-computing technique is very important technique because it allows pattern to belong to more than one cluster whereas the hard – computing technique is limited to the pattern belonging to the single cluster only.

A) Fuzzy Clustering: conventional clustering approaches generate partitions in which each instance belongs to one and only one cluster. So the cluster in hard clustering are incoherent. Fuzzy clustering extends this idea. In this case, each prototype is linked with every cluster using some sort of membership function, namely, each cluster is a fuzzy group of all the patterns. Bigger membership values indicate higher assurance in the assignment of the pattern to the cluster. Fuzzy partition by using a threshold of the membership value is formed due to hard-clustering . The most popular fuzzy clustering algorithm is the fuzzy *c*-means (FCM) algorithm. Because it is better one than hard *K*-means algorithm at avoiding (local minima), FCM algorithm can still join to local minima of the squared error criterion.

B) Simulated annealing (SA): In this method, sequential stochastic search technique designed to avoid local optima[11]. This is accomplished by accepting with some probability a new solution for the next iteration of lower quality (as measured by the criterion function). The probability of acceptance is governed by a critical parameter called the temperature (by analogy with annealing in metals), which is typically specified in terms of a starting (first iteration) and final temperature value.

V. CONCLUSION

In this paper, four popular data clustering algorithms in data mining are reviewed namely hierarchical clustering, partitioning clustering, expectation maximization clustering and the soft computing methods. Soft computing methodologies, containing genetic algorithms, simulated annealing, fuzzy clustering, neural networks are recently useful to solve the problems of data mining from large databases.. They attempt to provide us the low cost solutions, and hence the process will speed -up. Data mining is the very good area for studying the clustering algorithms.

REFERENCES

[1] Amin Sangar, Seyyed Khaze and Laya Ebrahimi," *participation anticipating in elections Using data mining methods*", International Journal on Cybernetics & Informatics Vol.2, No.2, 2013

[2] Tipawan Silwattananusarn and Dr. Kulthida Tuamsuk, "*Data Mining and Its Applications for Knowledge Management* : A Literature Review from 2007 to 2012", International Journal of Data Mining & Knowledge Management Process, Vol.2, No.5, 2012.

[3] Sushmita Mitra, Sankar Pal, and Pabitra Mitra,"*Data Mining in Soft Computing Framework :A Survey*", IEEE transactions on neural networks, vol. 13, no. 1, 2002.

[4] Sumit Vashishta, Dr. Yogendra Kumar Jain," *Efficient Retrieval of Text for Biomedical Domain using Data Mining Algorithm*", International Journal of Advanced Computer Science and Applications, Vol. 2, No. 4, 2011.

[5] Osama abu abbas,"*Comparision between data clustering algorithms*", International arab journal of information technology, vol. 5,No. 3,july 2008.

[6] U. Fayyad, R. Uthurusamy, "*Data mining and knowledge discovery in databases*," Commun. ACM, vol. 39, pp. 24–27, 1996.

[7] J. Huband, J. Bezdek, R. Hathaway, " *Visual Assessment of Cluster Tendency for Large Data Sets*", pattern recognition, vol. 38, no.11, pp. 1875 -1886, 2005.

[8] Hemlata Sahu, Shalini Shrma, Seema Gondhalakar," *A Brief Overview on Data Mining Survey*", international Journal of Computer Technology and Electronics Engineering Volume 1, Issue 3.

[9] Madhuri A. Tayal. ,M.M.Raghuwanshi," *review on various clustering methods for the image data* ", Journal of Emerging Trends in Computing and Information Sciences, vol 2, 2010 .

[10] A.K. Jain, M.N. Murty, and P.J. Flynn, "*Data Clustering: A Review*", ACM Computing Surveys, Vol. 31, No. 3, September 1999.

[11] Lior Rokach, Oded Maimon, *Clustering Methods*, data mining and knowledge discovery handbook, pp 321-551, 2010.

[12] Richard J. Hathaway, and James C. Bezdek, "*Fuzzy c-Means Clustering of Incomplete Data*", IEEE transactions on systems, man, and cybernetics—part b: cybernetics, vol. 31, no. 5, 2001.

[13] Sankar K. Pal, Varun Talwar, and Pabitra Mitra, "*Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions*", IEEE transactions on neural networks, vol. 13, no. 5, 2002.

[14] V.M.Navaneethakumar, Dr.C.Chandrasekar, "*A Consistent Web Documents Based Text Clustering Using Concept Based Mining Model*", International Journal of Computer Science Issues, Vol. 9, Issue 4, No 1, July 2012.

[15] B.S.Vamsi Krishna, P.Satheesh, Suneel Kumar R," *Comparative Study of K-means and Bisecting k-means Techniques in Wordnet Based Document Clustering*", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-1, Issue-6, 2012.

