AN INSIGHT OVERVIEW OF ISSUES AND CHALLENGES ASSOCIATED WITH CLUSTERING ALGORITHMS

Richa Loohach*

Dr. Kanwal Garg**

ABSTRACT

Data mining is the process of taking out of concealed prognostic information from a huge amount of databases. It is an influential technology which helps companies to focus on important information in their data warehouses. There are different steps in data mining process like Anomaly detection, Association rule learning, Clustering, Classification, Regression, Summarization. This paper is mainly concerned about clustering which is the procedure of organising the objects in groups whose members contains some kind of similarity. In the present review work, the author will make an attempt for identifying the major issues and challenges associated with different clustering algorithms and to select optimal clustering algorithm for the prediction of future.

Keywords: *Clustering Algorithms, Data Mining*

*Research Scholar, Department of Computer Science Applications, Kurukshetra University, Kurukshetra

**Assistant Professor, Department of Computer Science Applications, Kurukshetra University, Kurukshetra

1. DATA MINING

Data mining is the process to extract the hidden predictive information from large amount of databases which help companies to focus on important information in their data warehouse. Data mining tools predicts future trends and behaviours which allow business to make practical, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were time consuming to resolve. Data mining contributes by searching in databases to evaluate hidden patterns, predictive information which experts may miss because this information may lies outside their expectations [9]. Data mining algorithms represent techniques that have been implemented as established, consistent, understandable tools that consistently outperform older statistical methods. Before Data mining process occur we apply pre-processing of data in which we first select the data, after this pre-processing of data is done in which we assemble large amount of target data set. Pre-processing is necessary to analyze the multivariate data set. Subsequent to this data cleaning is done in which we remove noise and missing data from target dataset. Data mining may be applied as per the steps given below:

1. Anomaly detection: This is the identification of the unusual records or data errors.

2. Association rule learning: It Searches the relationships between variables. This is sometimes referred to as market basket analysis.

3.Clustering: This is the process of finding groups and structures in the data that are in some way or another "similar", without using known structures in the data.

4. Classification: This is the task of generalizing known structure to apply to new data.

5. Regression: this process is used to search a function which modals the data with the least error.

6. Summarization: It provides a more compact representation of the data set, including visualization and report generation [10].

As in the above discussed six steps of data mining clustering is most important so author discuss it in detailed manner in following section of clustering.

2. CLUSTERING

Clustering is the procedure of organising the objects in groups whose members contains some kind of similarity. So a cluster is the collection of objects which are alike and are different from the objects that belongs to other clusters. Core objective of clustering is to find out the inherent grouping in a set of unlabeled data[2]. There is no standard to find the best clustering algorithm which is independent of the dataset. It depends on user who must supply the criterion in such a way that the result of clustering will suits their needs. Clustering algorithms can be applied in many fields like in marketing to find groups of customers with similar behaviours and their buying habits, in biology for classification of plants and animals, or in library for ordering books etc. The major requirements for a clustering algorithm are: it should be scalable, it can deal with different types of attributes, it can discover clusters with arbitrary shape, there should be minimal requirements for domain knowledge to determine input parameters, it should have ability to deal with noise and outliers; it should be insensitive to order of input records etc. There are some problems with clustering techniques like these do not address all the requirements effectively (and simultaneously); there is time complexity problem with large number of dimensions and large number of data items, the effectiveness of the clustering method depends on the distance function used (for distance-based clustering); defining a new distance function if required is not always easy especially in multi-dimensional spaces, the result of the clustering algorithm can be interpreted in different ways[11].

3. REVIEW OF LITERATURE

Research in the various techniques in clustering is started in early 1990s. Now a days we have lots of clustering algorithms which are useful in different areas .we have different kind of clustering algorithms from which we can select the best suited algorithm according to our requirement . H.G Wilson et.al [2] compares hierarchical and partitional clustering techniques for multispectral image classification. Of all clustering procedures, the hierarchical nearest neighbour linkage had the lowest classification accuracy, whereas the combinatorial K-means partitional procedure produced the best classification result. Tung-Shou Chen et.al [8] introduces hierarchical K-means regulating divisive or agglomerative approach. The result indicates divisive hierarchical K-means is superior to hierarchical clustering on cluster quality and is superior to K-means clustering on computational speed. Peng Jin et.al [5] proposed clustering algorithm for data mining based on swarm intelligence called Ant-Cluster. The results illuminate that Ant-Cluster has better performance than k-means algorithm. Ren Jingbiao and Yin Shaohong[6] proposed an improved K-Harmonic means clustering algorithm. Through the regulation of distance metric parameters can achieve better clustering effects than the traditional K-harmonic means, and has an advantage both in run time and number of iterations. Srinivas and C. Krishna Mohan[4] propose an efficient hybrid clustering algorithm by combining the features of leader's method which is an incremental clustering method and complete linkage algorithm which is a Bo Ji and Yangdong Ye [1] give an improved sIB algorithm (CW-sIB) for high dimension document clustering using combination weighting. The experiments have shown that the proposed CW-sIB algorithm is superior to the sIB algorithm. Hemanta Kumar Kalita et.al [3] proposes a new, semiautomatic clustering algorithm called – Ordering of Points to Identify Clustering Structure Based on Perimeter of Triangle: OPTICS (BOPT) for finding out clusters by ordering a database of points using a unique method – triangulation of points. Shi Na et.al [7] proposes an improved k-means algorithm; the improved method avoids computing the distance of each data object to the cluster centres repeatedly, saving the running time.

4. ISSUE AND CHALLENGES

Since there are various clustering algorithms available to automate or semi-automate the clustering procedure it is very difficult to choose the suitable algorithm for a particular dataset. As different algorithms applied on a dataset may produce different kind of results with different clusters. Each algorithm has its own run time, complexity, error frequency, resources used etc to complete the procedure of clustering. Another issue may be that the outcome of a clustering algorithm mainly depends on the type of dataset used. As the size and dimensions of dataset increases day by day this makes it difficult to handle for a particular clustering algorithm. Also the complexity of data set increases, which include data like audios, videos, pictures and other multimedia data which form very heavy database, this in turn create the time complexity of a clustering algorithm. Furthermore clustering algorithms do not concentrate on all of the requirements simultaneously and effectively which makes the result uncertain. Most of the clustering algorithms depends on the distance function used in the algorithm and if the given distance function do not perform efficiently then a new distance function may required which is difficult to formulate especially for multi-dimensional data this increases the tediousness of work. Also the output of a clustering algorithm can be interpreted in different ways which may create confusion for understanding the result by users. So we need an immense concern to choose a clustering algorithm for the dataset. The selection of a clustering algorithm may based on the type of dataset, time requirement, efficiency needed, accuracy required, error tolerance etc. so the main challenge is to choose the correct type of clustering algorithm for the data set which are based on user requirements among many known clustering algorithms so that user can get the desired results which helps in further research for data mining process.

5. CONCLUSION

This paper deals with study of different kind of clustering algorithms. It first defines the data mining process which is the method of finding predictive information from a huge amount of databases. Then it defines the clustering process which is the procedure of assemblage of the objects in groups whose members contain some kind of resemblance. After that a detailed study of clustering algorithms and their comparison in different perceptions are examined. This paper highlights the concerned issues and challenges which may be helpful for the upcoming researchers to carry on their work.

REFERENCES

[1] Bo Ji and Yangdong Ye "An improved sIB algorithm for document clustering
Using combination weighting measures" 978-1-4244-8728-8/11/$26.00 ©2011 IEEE

[2] H. G. Wilson, B. Boots, and A. A. Millward "A Comparison of Hierarchical and partitional Clustering Techniques for Multispectral Image Classification", 0-7803-7536-X (C) 2002 IEEE

[3] Hemanta Kumar, Kalita Dhruba Kumar and Bhattacharyya Avijit Kar "A New Algorithm for Ordering of Points to Identify Clustering Structure Based On Perimeter of Triangle: OPTICS (BOPT)", 15th International Conference on Advanced Computing and Communications

[4] M. Srinivas and C. Krishna Mohan, "Efficient Clustering Approach using Incremental and Hierarchical Clustering Methods", 978-1-4244-8126-2/10/$26.00 ©2010 IEEE

[5] Peng Jin, Yun-long Zhu, Kun-yuan Hu , "A Clustering Algorithm for Data mining based on Swarm Intelligence", Proceedings of Sixth International Conference on Machine Learning Cybernetics, Hong Kong, 19-22 August 2007

[6] Ren Jingbiao and Yin Shaohong "Research and Improvement of Clustering Algorithm in Data Mining", 2010 2nd International Conference on Signal Processing Systems (ICSPS)

[7] Shi Na, Liu Xumin and Guan yong "Research on k-means Clustering Algorithm An Improved k-means Clustering Algorithm", Third International Symposium on Intelligent Information Technology and Security Informatics 978-0-7695-4020-7/10 $26.00 © 2010 IEEE

[8] Tung-Shou Chen, Tzu-Hsin Tsai, Yi-Tzu Chen, Chin-Chiang Lin, Rong-Chang Chen, Shuan-Yow, Li and Hsin-Yi Chen " A combined K-Means and Hierarchical clustering method for improving the clustering efficiency of microarray" ,Proceedings of 2005

International Symposium on Intelligent Signal Processing and Communication Systems December 13-16, 2005 Hong Kong

[9]http://www.thearling.com/text/dmwhite/dmwhite.htm (28/11/2011)

[10]http://en.wikipedia.org/wiki/Data_mining retrieved on 28/11/2011

[11]http://home.dei.polimi.it/matteucc/Clustering/tutorial_html retrieved on 28/11/2011