A Survey on Asymmetric spectral clustering

Shilpa Damor[1], Prof. Naimisha Trivedi[2]
[1, 2] Department of Information Technology
[1, 2] L.D.Engineering College,Ahmedabad, Gujarat, India

Abstract- *In this survey paper, There is large amount of data is present in the world. This data is coming from various sources like companies, organizations, social networking sites, image processing, world wide web, scientific and medical data etc. Peoples do not have time to look all this data. They attended towards the precious and interested information. Data mining is technique which is used to extract meaning full information from huge databases. Extracted information is visualized in the form of statics, graphs, and tables and vides etc. There are number of data mining techniques and asymmetric clustering is one of them. Asymmetric technique is type of unsupervised learning. In this, data sets which have similarity are placed in one cluster and others are in other clusters. From, number of years various asymmetric clustering technique are introduced which work well with datasets. These techniques do not work well with the complex and strongly coupled data sets. To reduce processing time and improve accuracy neural networks are combined with asymmetric clustering algorithms.*

Keywords- Data mining, asymmetric clustering, Two step clustering

I. INTRODUCTION

The sheer amount of data is stored in world today called big data. In 2001, it is assumed that about 8, 50,000 petabytes [1] of data is stored in the world and it is expected that it will be about 35 zettabyte in 2022[1]. Mostly, data is generated by the social websites, market analysis medical field, web mining and image processing etc. This data is stored in large databases in the forms of tables, images and videos etc. called data warehouses. The process of extracting useful patterns or knowledge from data base is called data mining. The extracted information is visualized in the form of charts, graph and tables etc. Data mining is also known by another name called KDD (knowledge discovery from the database). In data mining, frequent item set is used to find relations between numerous numbers of fields in data mining. Association rules are used to discover the frequent data item sets. The concept of association rules is used in various fields like retail stores, market strategy and stock market etc.

Classification of Data Mining System: Data mining system is classified according to following categories:

1. According to Data source to be mined: Data mine system can be classified according to kinds of mined techniques used like spatial data, multimedia data etc
2. According to Data models: Data mine systems may use many models like relational model, object oriented model and transactional models.
3. According to kind of Knowledge mined: Data mine system can be classified according to the type of knowledge is used like classification, prediction, cluster analysis and outlier analysis.
4. According to utilized Mining technique: Data mine system can be classified according to techniques used for data mining techniques like decision tree, neural network etc.
5. According to adapted applications: Data mine systems can be classified according to applications adapted like in finance, data mining system related to finance is used.

II. CLUSTERING IN DATA MINING

Clustering means putting objects having similar properties into one group and objects having dissimilar properties into another. For example, object having values above threshold values can be placed in one cluster and values below into another cluster. Clustering divide the large data set into groups or clusters according to similarity in properties.

Clustering is an unsupervised learning technique as there are no classifiers and their labels .It is form of learning by observation. Cluster analysis can be used in the areas such as image processing, analysis of data, market research (buying patterns) etc. Using clustering we can do outlier detection where outliers are values lying outside the cluster.

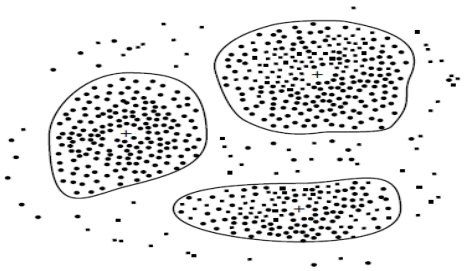

Figure 1. Clusters and Outliers

In above figure the dots which are outside the clusters represent outliers and there are clusters of object with similar properties.

In the previous times, various clustering algorithms had been developed to cluster data when diversion is seen in the given datasets. Now days, data in the world is increasing day by day like in social networking sites, market analysis, medical field, image processing and world wide web etc. volume of data is continuously increasing. To, store and efficiently access this data we need to make cluster of similar and dissimilar. There are many clustering techniques are used to perform various type of operation on data in databases. The clustering algorithms which are recommended can give good performance like provide good efficiency on the numeric and pure categorical type of data. This proposed algorithm perform good operation on simple and statistical database and but will not perform desired operation on data which are complex and of mixed category like plant dataset which we consider for this work. In previous year an efficient clustering algorithm is proposed which works in two steps to find clusters for complex type of data. The two step algorithm works as every dataset had some of the attributes and then to cluster data the relationship between the attributes are maintained and similarity between attributes are derived, on the basis of similarity derived, the data will be clustered. This algorithm will also be applied on hierarchical and partitioning methods. This method shows relationships between the items or objects and tries to improve the weakness of using single clustering algorithms. In this research enhancement in asymmetric clustering is done on the basis on two- step algorithm. Clusters of complex data is made the conclusion taken from the output and enhancement will be main agenda of this work.

III. PROBLEM FORMULATION

Cluster analysis is being broadly used in several applications like basket analysis, e-commerce, image processing, scientific and medical field, data analysis, and word wide web etc. Today in business, stock market clustering can support marketers to determine interest's vendors and customers based on their record of purchasing patterns and distinguish groups of their customers who are interested in goods. In medical science, cluster analysis can be used to derive new plant like testing new hybrid species or estimating the conditions in which they grow well and observing soil and water quality. Animal taxonomies, classify their genetic factors with similar functionality. In geology, expert can use clustering technique to recognise areas of similar interests, lands, similar, houses and infrastructure in a city or in country etc. Data clustering technique is also useful in organising data on the World Wide Web for interested knowledge or data. Clustering is an unsupervised classification technique that aims at generating collections of items, or clusters in that way that object with similar properties are grouped together in same cluster and objects with different cluster are quite distant. Mining arbitrary shaped clusters in large data sets is an open challenge in data mining. The number of solutions of these problems has been proposed with high time complexity. Computational cost can be saved by using some algorithms by shrinking a data set size to a smaller amount data examples and user defined threshold ratios can affect the clustering performances. The CLASP(clustering algorithm for arbitrary shaped clusters) algorithm is an effective and efficient algorithm for mining arbitrary shaped clusters which automatically shrinks the size of a data set while effectively preserving the shape information of clusters in the data set with representative data examples. After this it changes the locations of these data examples to improve their intrinsic relationship and make the cluster structures more clear and distinct for clustering. At last, it does agglomerative clustering to find the cluster structures with the help of pk metric called mutual k-nearest neighbour-based similarity metric. In this work, the enhancement of the asymmetric clustering algorithms to increase the quality of cluster and improve the efficiency of algorithms.

IV. TWO STEP CLUSTERING ALGORITHM

The proposed two step asymmetric algorithms works very efficiently with simple databases. The accuracy of asymmetric clustering will be reduced, when the data will be highly coupled together. enhancement neural networks are used along with two step asymmetric algorithms techniques to improve the performance and accuracy of cluster. The proposed algorithm along with existing algorithms is implemented to results in terms of accuracy and quality. Main work is on that to enhancement in asymmetric clustering to remove noise from the datasets to reduce escape time.

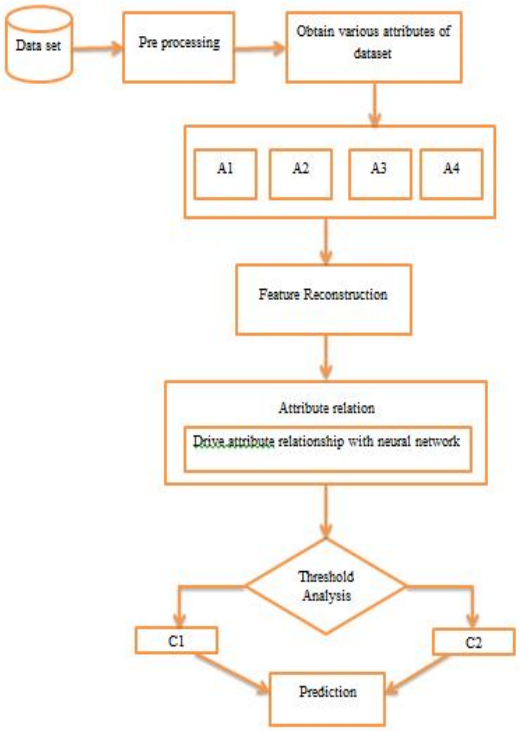

Figure.2 System's flow diagram

DATASET and Pre-processing: In the first step of flowchart, the dataset is extracted and then pre-processed to perform clustering on dataset.

Obtain various attributes: After the pre-processing phase, the dataset contains various attributes, in this phase relationship between various attributes are established.A1, A2, A3…. is the number of different clusters of particular attributes.

Feature Reconstruction: in this step, again pre-processing technique is performed on the clusters to remove noise from the attributes in the clusters

Drive relationship and training dataset:- to drive relationship between various attributes of the dataset, technique of neural network will be applied. To apply neural network we need an trained dataset. The trained dataset will establish relationship between various attributes

Threshold Analysis: in this step, threshold analysis done and values above threshold valuesare stored in one cluster and values below threshold values stored in another cluster called C1 andC2

Prediction: this step predicts the increased efficiency of the work in research.

V. SUMMARY AND CONCLUSIONS

To extract useful or interested information from large set of databases data mining techniques are used. KDD (knowledge discovery from databases) is data mining method to extract information from data warehouses. Association rule is method to place the frequent item sets together to ado analysis like in basket analysis, retail stores and stock market etc. Asymmetric clustering is unsupervised technique of data mining. Clustering is technique in which large datasets are divide in to small datasets in this way that objects and items with having similar properties into one group and objects having dissimilar properties into another.

There are number of algorithms that work well with simple datasets in the term of accuracy and performance but, when these algorithms has to work with mixed and tightly coupled different data sets their performance in the term of accuracy is decreased. Neural networks can be combined with these existing asymmetric algorithms to improve and accuracy and reduce escape time.

REFERENCES

[1] Cloud Computing Principles and Paradigms, Edited by, Rajkumar Buyya, James Broberg, Andrzej Goscinski, by John Wiley & Sons, Inc publications 2011.

[2] Hao Huang† Yunjun Gao‡,_ Kevin Chiew§ Lei Chen# Qinming He " Towards Effective and Efficient Mining of Arbitrary Shaped Clusters" Department of Computer Science and Engineering, Hong Kong University of Science and Technology, China, ICDE Conference 2014

[3] Gunnar Carlsson et.al , "Hierarchical Quasi-Clustering Methods for Asymmetric Networks", Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR:W&CP volume 32, 2014

[4] R.Jensi and Dr.G.Wiselin Jiji, "A Survey On Optimization Approaches To Text Document Clustering", International Journal on Computational Sciences & Applications (IJCSA) Vol.3, No.6, December 2013

[5] Mahendra Pratap Yadav, Mhd Feeroz and Vinod Kumar Yadav (2012) "Mining the customer behavior using web usage mining In e- commerce" Coimbatore, India. IEEE-201S0

[6] Neelamadhab Padhy , Dr. Pragnyaban Mishra and and Rasmita Panigrahi "The Survey of Data Mining Applications And Feature Scope" International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012

[7] Satoshi Takumi and Sadaaki Miyamoto," Top-down vs Bottom-up methods of Linkage for Asymmetric Agglomerative Hierarchical Clustering", 2012 International Conference on granular Computing

[8] S.R.Pande, Ms..S.S.Sambare, V.M.Thakre, "Data Clustering Using Data Mining Techniqes", IJARCCE Vol. 1, issue 8, October 2012

[9] Ming-Yi Shih, Jar-Wen Jheng and Lien-Fu Lai, "A Two-Step Method for Clustering Mixed Categroical and Numeric Data", Tamkang Journal of Science and Engineering, Vol. 13, No. 1, pp. 11-19 ,2010

[10] Wilhelmiina Hamalainen, Matti Nykanen (2008) "Efficient discovery of statistically significant association rules", Eighth IEEE International Conference on Data Mining.

[11] Jiawei Han J and Kamber M, Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann, San Francisco, CA, 2012.