A Survey on Enhancement in Asymmetric Clustering Algorithm

Shilpa H Damor[1], Prof. Naimisha Trivedi[2]
[1, 2] Department of Information technology
[1, 2] L. D. College of Engineering, Ahmedabad – 380015

Abstract- *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 in asymmetric clustering algorithms.*

Keywords- Data mining, clustering, asymmetric clustering, data sets.

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.

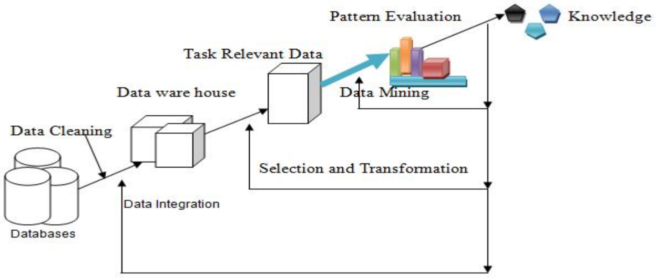


Figure 1.1: Data Mining Process Ease of Use

We know that these days Informational technology is mounting and databases created by organizations and companies like telecommunications, banking, marketing, transportation, manufacturing, and social networking sites etc. are becoming huge day by day. Knowledge discovery process is used to store this data in databases and efficiently access the interested or useful data from databases.

1.2 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

1.3 CLASSIFICATION OF DATA MINING SYSTEM:

Data mining system is classified according to following categories:

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

According to Data models: Data mine systems may use many models like relational model, object oriented model and transactional models**. 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.

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.

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.

1.4 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.

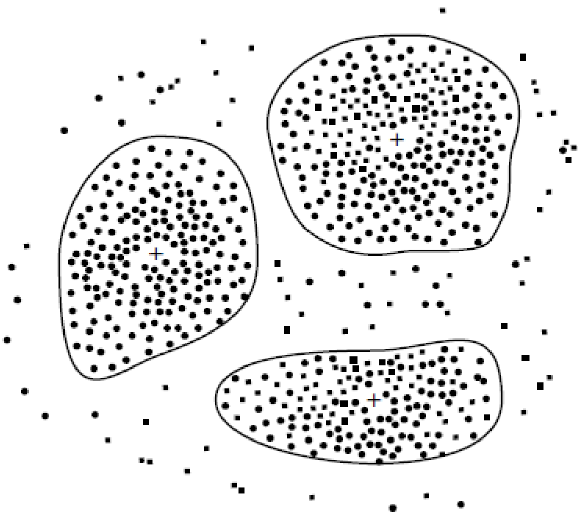
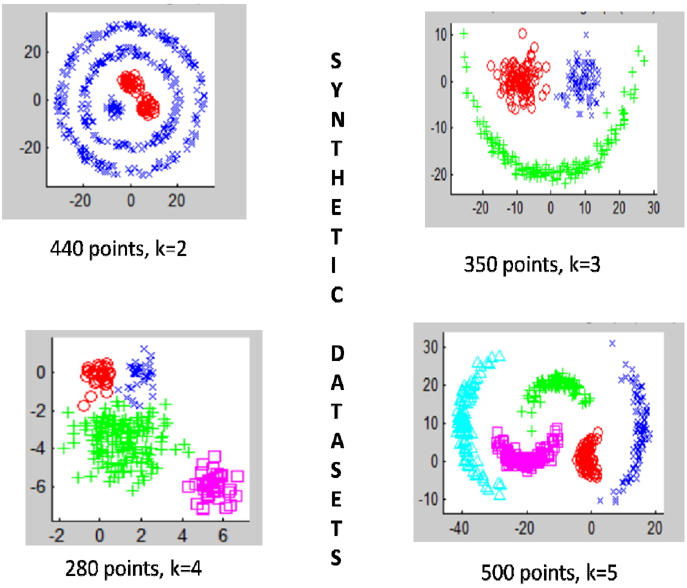


Figure 1.2 Clusters and Outliers

II. WHICH TYPES OF DATASETS USED?

We experimented with five different data sets containing points in two dimensions whose geometric shape are shown in below Figure. All data sets, DS1 to das 4, have different clusters that are of different size, shape, and density, and contains noise and also different regions of the clusters have different densities.

A particularly challenging feature of this data set is that clusters are very close to each other and they have different densities. The size of these data sets ranges from 280 to 500 points, and their exact size is indicated in Figure.



440 points, k=2

350 points, k=3

280 points, k=4

500 points, k=5

Figure 2. Synthetic datasets from datasets 1 to datasets

III. RESEARCH METHODOLOGY

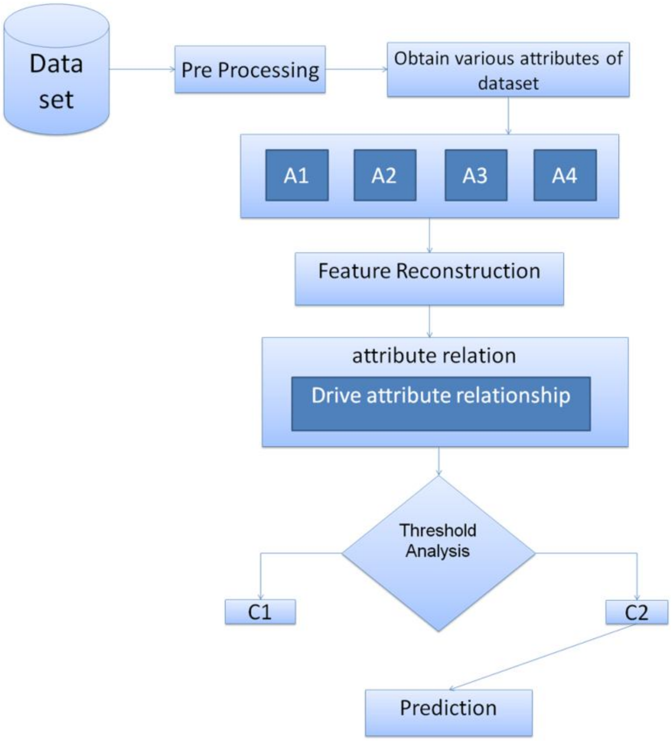


Figure 3.1.methodology

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

2. 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.

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

4. 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

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

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

3.1 PROPOSED ALGORITHM

Here is the explanation about algorithm and also explain about how to k-mean, normalization ,affinity and mean shift steps are works.

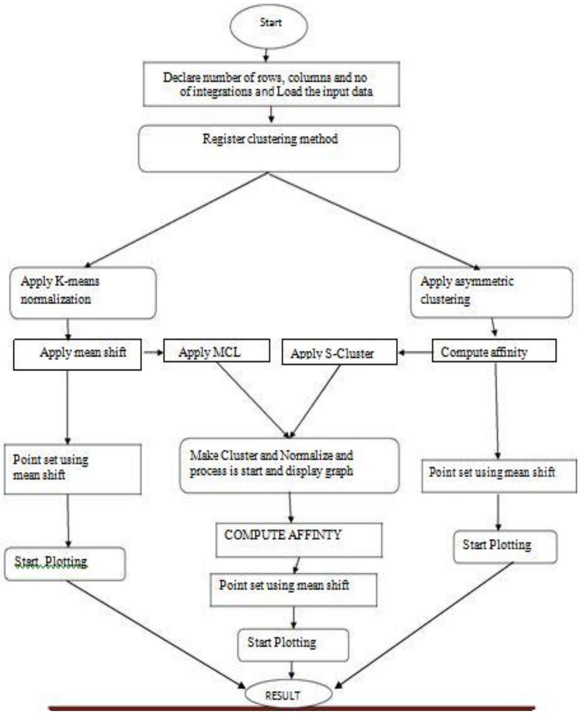

Start
Declare number of rows, columns and no of integrations and Load the input data
Register clustering method
Apply K-means normalization
Apply asymmetric clustering
Apply mean shift
Apply MCL
Apply S-Cluster
Compute affinity
Point set using mean shift
Make Cluster and Normalize and process is start and display graph
Point set using mean shift
COMPUTE AFFINTY
Start Plotting
Point set using mean shift
Start Plotting
Start Plotting
RESULT

Figure 3.2 flowchart of proposed algorithm

Working of K-mean and Normalization Step

Input: Data set $P = \{p1, p2, . . . , pn\}$, cluster number k.

Step 1 Compute the distance matrix W, construct similarity matrix S according to W,
where $W(i, j\ )$ is the distance between pi and pj , $i = 1, 2, . . . , n$;

Step 2 Calculate the Laplacian matrix, $L = D - S$;

Step 3 Compute the first k eigen vectors $\{v1, . . . , vk\}$ of the generalized eigen problem $Lv = \lambda Dv$;

Step 4 Let $V \in Rn\times k$ be a matrix composed of the vectors $\{v1, . . . , vk\}$ as columns;

Step 5 For $i = 1, . . . , n$, let $yi \in R1\times k$ be the vector corresponding to the ith row of V;

Step 6 Cluster the points $\{yi \in R1\times k \mid i = 1, 2, . . . , n\}$ with the k-means algorithm into clusters $C1, . . . ,Ck$, if $yi \in Cj$ then $pi \in Pj$, $1 \le i \le n$, $1 \le j \le k$.

Output: k clusters $P1, \dots, Pk$.

Working of affinity and mean shift step

Input: Data set $P = \{p1, p2, \dots, pn\}$, > 0, $\delta > 0$, user-specified upper threshold
Cmax ≥ 2 for cluster number to be testified, user-specified maximum number of neighbors Kmax ≥ 2.

Step 1 Calculate the distance matrix W;

Step 2 For $i = 1, 2, \dots, n$, sort the ith row of W, then calculate $pi\ K$, which is the Kth neighbor of pi , $K = 2, \dots, K$max;

Step 3 For $K = 2, \dots, K$max run step 4~5;

Step 4 Calculate the similarity matrix S, where $S(i, j) = \exp( )$;

Step 5 For every $k = 2, \dots, C$max, make use of the Meilˇ*a*–Shi spectral clustering algorithm to cluster the data set P into k clusters and calculate the value of index Ratio(k) for obtained clusters;

Step 6 To determine whether the candidate cluster number $2 \leq k \leq C$max is an -reasonable and δ-stable cluster number according to the results of step 4 and step 5;

Output: The set of reasonable and δ-stable cluster numbers

3.2 EXPLANATION OF PROPOSED FLOWCHART

1. Declare rows, columns, integration and load dataset: **-** This is the first set of algorithm in which the number of rows and columns are defined for the dataset. The second condition is defined to define number of iteration to define cluster quality. In the step of the flowchart the dataset will be loaded to perform clustering operation

2. Register Clustering Method: - To register clustering method is the second step of the flowchart in which we defined the two clustering method. The first method is K-mean clustering and second method is asymmetric clustering. According to the selected method the operation of clustering will be performed on the dataset.

3. Apply K-mean Normalization and asymmetric Clustering: - When the clustering method is registered, it may be K-mean normalization method which is selected for clustering with the normalization equation. The normalization equation when implemented with k-mean the cluster quality can be improved. The second method is of asymmetric clustering which is implemented to cluster the asymmetric data from the loaded dataset.

4. **Apply mean shift and affinity metrics : -** In this step, two operations are performed. In the first step mean shift algorithm is applied on the loaded dataset. In the mean shift algorithm, the mean value is calculated on the dataset and left shift operation is performed to simplify the operation of clustering. The second method is of affinity metrics, it is equation which is applied to find relationship between various elements of the dataset.

5. Apply MCL and S-clustering : - The MCL is the markov clustering algorithm, which is the unsupervised clustering graph based algorithm. This algorithm is fast and reliable and has good cluster quality. The main concept behind this algorithm is mathematical theory behind it, its position in cluster analysis and graph clustering, issues concerning scalability, implementation, and benchmarking, and performance criteria for graph clustering in general. The second method is S-clustering which is applied to cluster the data on the basis of graph methods

6. Plot and make clustering and normalize :- In the previous step, two methods are applied which are MCL and S-cluster, to cluster the data. In this method clustered data will be plotted. When the data is plotted, the method of normalization will be applied on the plotted data to improve the cluster quality.

7. Start of iteration, mean shift insertion and affinity insertion: **-**In these steps of flowchart, the iterations which are defined in start of flowchart. The process of mean shift and affinity metrics is calculated and which are inserted on every iteration and with each iteration cluster quality had been improved.

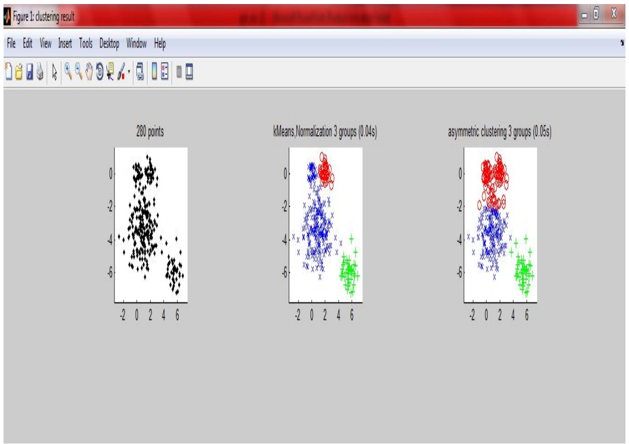
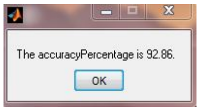


Figure 1: clustering result
File Edit View Insert Tools Desktop Window Help
280 points
kMeans,Normalization 3 groups (0.04s)
asymmetric clustering 3 groups (0.05s)
0
-2
-4
-6
-2 0 2 4 6



The accuracyPercentage is 92.86.

Figure 3.3 output of proposed algorithm

IV. CONCLUSION

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.spectral clustering to make this kind of clustering more attractive.

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

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

[8] 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

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

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

[11] 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