Fuzzy based Probability Factor Calculation for Number of Cluster Estimation to K-Mean by using Apriori

Pratishtha Singh Baghel
CSE Department BUIT,
Barkatullah University, Bhopal

Divakar Singh, Ph.D
CSE Department BUIT,
BARKATULLAH UNIVERSITY, Bhopal

ABSTRACT
Data mining is a powerful and a new field having various techniques. It converts the raw data into useful information in various research fields. Clustering is used to collect similar data in a group. It is a process of putting similar data into groups. A popular technique for clustering is K-means in which data are partitioned into K clusters. In this method, the number of clusters is pre defined and the technique is highly dependent on the initial identification of elements that represent the clusters well. But we cannot changethe number of cluster at mid of execution of algorithm. But in k-mean, important factor is that how many clusters we should take, it may be less and it may be more. There is not any mechanism to estimate the number of clusters in k-mean. It totally depends upon user, how many he takes. But for large amount of data user can't decide how much data have similar. For example, if maximum data have common similarities, so why we take more cluster. For this it may be minimum number of s for better evaluation and better performance. similarly if we have a amount of dissimilar data so we should take more cluster in k-mean. For this we are using a priori to generate association rules and with the help of association rule we put the values in my proposed equation and calculate the probability factor to give us the estimated number of cluster sfor k-mean.

Keyword:
Data mining, clustering, a priori, k-means, association rules, probability factor.

1. INTRODUCTION
Data mining also known as Knowledge Discovery in Database (KDD). It is a powerful technology with great potential l for organizing the most important information in data warehouses [7][8][10]. Data mining tools predict proactive knowledge-driven decisions [11] [9]. Data mining majorly uses the concept of association rules. For implementing association rule, Apriori [12] is a classical algorithm used here.

Clustering is a technique which partition data elements such that elements have similar property assigned to the same cluster while elements with other properties are assigned to other clusters. Clustering performs efficient search in a data set.

2. K-MEANS CLUSTERING ALGORITHM
This part describes the novel k-means clustering algorithm. The scheme is to classify a given set of data into k number of clusters, in which the value of k is fixed in advance. The algorithm consists of two different phases: the first phase is to describe k centroids, one for each cluster. The next stage is to take each point belonging to the given data set and associate it to the nearest centroid. Euclidean distance is normally considered to decide the distance between data points and the centroids. When all the points are integrated in some clusters, the first step is done and an early grouping is finished. At this point we need to recalculate the new centroids, as the enclosure of new points may guide to a change in the cluster centroids. Once we find knew centroids, a new binding is to be created between the same data points and the nearest newly centroid, create a loop. As an outcome of this loop, the k centroids may modify their location in a step by step method. Finally, a condition will be reached where the centroids do not shift to any further extent. This signifies the convergence principle for clustering. Pseudocode for the k-means clustering algorithm is listed as Algorithm 1 [6].

Algorithm 1: The k-means clustering algorithm

Input: D = {d1, d2,.....,den} //set of n data items.

K// Number of desired clusters

Output: A set of k clusters.

Steps:

1. Arbitrarily choose k data-items from D as initial centroids;

2. Repeat

Assign each item di to the clusters, which has the closest centroid;

Calculatethe new mean for each cluster;

Until convergence criteria is met.

3. LITERATURE SURVEY

3.1 Hybrid Clustering, Association Mining Approach
For prediction of factors affecting the sale of products, Hybrid clustering ,association mining approach[1] uses mining patterns of huge stock data. It classifies stock data and finds a compact form of associated patterns of sale. The objectives of this research is to get better decision making for improving quality, sale and services as to identify the reasons of dead stock, fast-moving, and slow-moving products which is useful mechanism for business, investment, support and surveillance. In the first phase, on the basis of product categories and sold quantities, it divides the stock data in three different clusters i.e. Slow-Moving (SM), Dead-Stock (DS) and using K-means algorithm. In the second phase Most Frequent Pattern (MFP) algorithm has been proposed to find the frequencies of property values of the corresponding items. This technique is simple by using a matrix and counting of attribute values. Some of the limitation ofthe study includes: it requires proper data with specific attributes.

3.2 Dynamic Model based Clustering
A new method of data mining is proposed [2] to determine the structure and number of clusters, and refining groups in multivariate different data set including groups, completely and partly overlapped group structures by using a dynamic model based clustering. This method works for high

dimensional data without data reduction, in which some of variables including completely overlapped situations. It is called dynamic model based clustering since the structure of model changes dynamically at every stage of refinement process. A classification tree is obtained by method at the end of this technique. The classification tree consists of groups of levels or stages. There is whole data group at level 0 or stage 0. There are homogeneous groups on the leaves of the classification tree. The depth of the classification tree is determined by the last stage of the data mining method, proposed for refining groups in multivariate data set using dynamic model based clustering. The depth of the classification tree of glass identification data set is obtained as 4.

3.3 Clustering Centroid Value For Spatio-Temporal Data Mining

The main work of this research [3] gives the knowledge about the moving region data centroid. Clustering centroid value for spatio-temporal data mining is the main aim of this work, Using advanced k-means, k-means algorithm and Avg Centroid (AC) clustering. In this paper the real time data of the hurricane Indian Ocean 2001 to 2010 maximum wind details are focused. The clustering is taking as, the first window form the basis of the pixel coordinate value of the screen by using the selection window method. Form basis of the selection window, the data mining retrieves clustering data. There are more steps in enhanced k-means algorithm, but result is accurate. Iteration also repeated very minimum times. The comparative study of the AC clustering values, k-means and enhanced k- means algorithms is presented in this paper.

3.4 Management of Data Replica of the Cluster System

The management of data replica focuses on the management of the replica of the database data. This research [4] introduces cloud storage technology available now firstly, and gives the related concepts of cloud storage cluster architecture in the second part, then explains the important effect of the management of data replica technology in cloud storage cluster system .In the third part, this research introduces a model of data replica adjusted to the need, and proves it to be effective to the system performance of the cluster by the way of experiments and analysis in the fourth part. still in the initial stage data replica present in the management of the non-database.

3.5 Ant Colony Optimization with Different Favor (ACODF) FOR Data Clustering

In this paper [5], they propose a novel algorithm known ant colony optimization with different favor (ACODF) for data clustering. Clustering algorithms generally make a distance metric based similarity measure in direct to partition the database such that data points in the similar partition are more comparable than points in different partitions. The presentation of this method is improved than the Fast SOM combines K-means approach (FSOM+K-means) and Genetic K-Means Algorithm (GKA).

The ACODF algorithm has the following three important popular strategies: (a) using ACO with different favour to resolve the clustering difficulty, (b) adopting simulated annealing concept for ants to decreasingly visit , the amount of cities to get local optimal solutions, (c) utilizing tournament selection strategy to decide a path. They evaluate his ACODF method with the FSOM+K-means approach and GKA. Throughout experiments, they demonstrate that ACODF efficiently finds correct clusters in large high dimensional data sets. In addition, in all the cases they calculated, his method produces much smaller errors than both the FSOM+K-means approach and GKA.

After reviewing all researchers work mentioned above, practical understanding on how to make clustering could be understand, Number of algorithm are used for more efficiently modifying or adding steps, but none of existing algorithm talks about how many cluster should be taken. This problem is being conceptualized in proposed research work. Because if less number of dissimilar element available then no need to take more clusters. And if more numbers of dissimilar elements, why should to go for less number of cluster.

4. PROPOSED WORK

The main aim of research work is to propose a novel method that is not proposed yet by any researcher. In this work a method is proposed by which identification of fuzzy based probability factor can get more simpler and accurate. If number of clusters are estimated according to data then it may be more beneficial for reducing complexity and unnecessary calculation. This factor is directly proportional to similarity of the element. If all elements have similar property, then no need of taking more clusters, because it may be more complex and may take more time. So before clustering a reference or an idea of number of clusters less or more is the heart of proposed work.

4.1 Algorithm

Step 1: Select any data set for which we have to decide number of cluster.

Step 2: Generate frequent item sets by using apriori.

Step 3: Find ‘Li’ that is ‘n’ in n-item set for i^{th} item set generated by apriori.

Step 4: Find ‘Mi’ that is frequency of i^{th} item set generated by apriori.

Step 5: Find ‘N’ that is total number of transaction in dataset.

Step 6: Add=0;

```
For(i = 1 ; i < N; i++)
{
    Prod=Li*Mi ;
    Add=Add+Prod ;
}
```

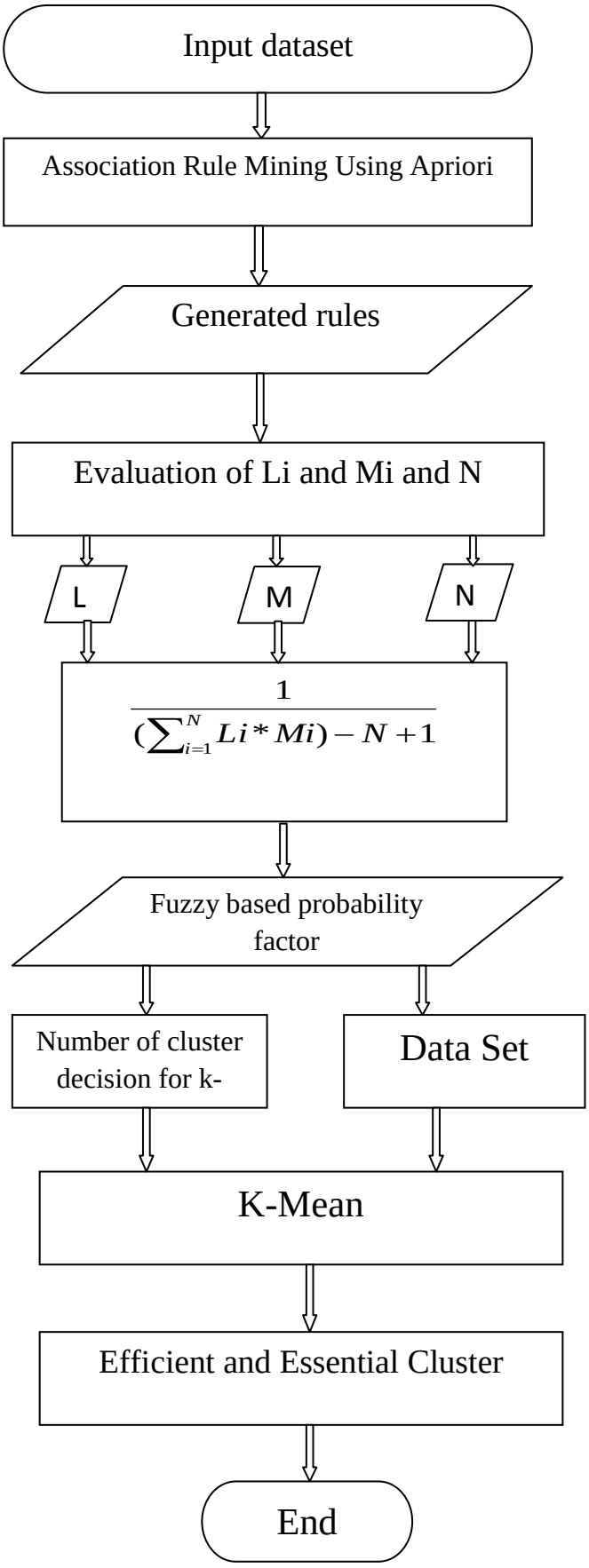
Step 7: Add=Add-N+1;

Step 8: Prob_fact=1/Add;

Step 9: Select number of cluster for that data set for which we have calculated Probability factor, by using reference of Prob_fact;

Step 10: Exit.

4.2 Flow Chart



5. RESULT AND ANALYSIS

In fig 2 seems that probability factor for three dataset with different support count values.

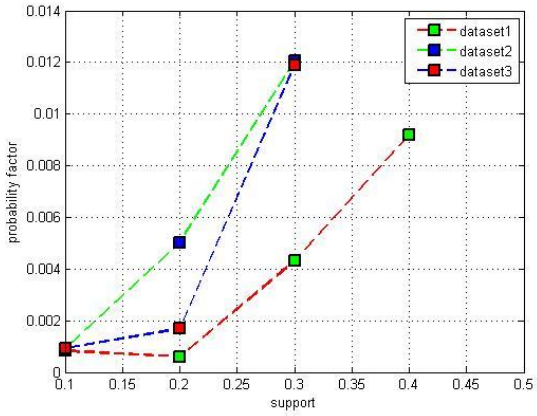


Fig 2: probability factor with different support count.

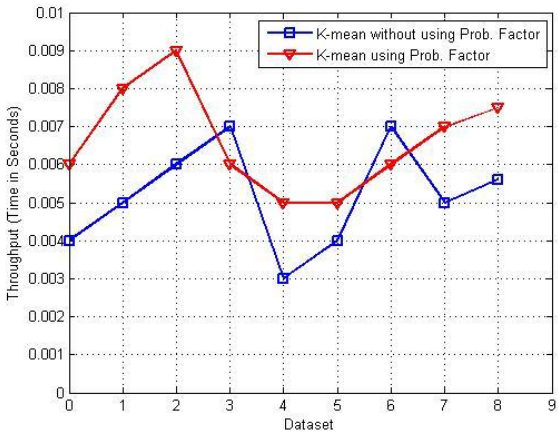


Fig 3: Throughput comparison of k-means with probability factor and without probability factor.

Fig 3 shows a comparison of throughput of k-mean with probability factor and without using probability factor.

Here when using the probability factor before clustering in k-means, then it gives better throughput as comparable to k-means without using probability factor.

6. CONCLUSIONS

This work proposes a method by which calculating fuzzy based probability factor. This factor is directly proportional to similarity of the element. Because many researchers have worked to improve k-means for a better

Result, but the main facot is 'number of cluster' that have to select, because the number of cluster can't change in mid of the process. So number of cluster selection is an important factor for k-means. And apriori plays an important role in estimating the cluster for k-mean.

7. FUTURE WORK

In future it will more efficient and accurate and with the help of this approach exact number of clusters will be decided. Because currently that is estimating the probability to take number of clusters.

8. REFERENCES

[1] Aurangzeb Khan, Khairullah khan, Baharum B. Baharudin, "Frequent Patterns Minning Of Stock Data Using Hybrid Clustering Association Algorithm", 2009 International Conference on Information Management and Engineering.

[2] [2]. Tayfun Servi, Hamza Erol, "A Data Mining Method For Refining Groups In Data Using Dynamic Model Based Clustering", 978-1-4799-0661-1 / 13 / 2013 IEEE.

[3] Dr.S.Santhosh Baboo, K.Tajudin, "Clustering Centroid Finding Algorithm (CCFA) using Spatial Temporal Data Mining Concept", "Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME) February 21-22".

[4] Guangbin Bao,Chaojia Yu, Hong Zhao ,Hong Zhao, "The Model of Data Replica Adjust to the Need Based on HDFS Cluster", 2012 Fifth International Conference on Business Intelligence and Financial Engineering.

[5] Cheng-Fa Tsai, Han-Chang Wu, and Chun-Wei Tsai, "A New Data Clustering Approach for Data Mining in Large Databases", Proceedings of the International Symposium on Parallel Architectures, Algorithms and Networks (ISPAN02).

[6] Margaret H. Dunham, Data Mining- Introductory and Advanced Concepts, Pearson Education, 2006.

[7] Berry, M. J. A. and Linoff, G. Data mining techniques for marketing, sales and customer support, USA: John Wiley and Sons,1997

[8] Fayyad, U. M; Piatetsky-Shapiro, G.; Smyth, P.; and Uthurusamy, R.. Advances in Knowledge Discovery and Data Mining. Menlo Park, Calif.: AAAI Press 1996.

[9] Dr. Gary Parker, vol 7, Data Mining: Modules in emerging fields, CD-ROM, 2004.

[10] Jiawei Han and Micheline Kamber , Data Mining Concepts and Techniques, published by Morgan Kauffman, 2nd ed 2006.

[11] Literature Review: Data mining, http://nccur.lib.nccu.edu.twlbitstream/ 140.1 I 9/3523 I/S/35603 I OS.pdf, retrieved on June 2012.

[12] Divakar Singh, A. Shrivastava, algorithm for frequent item set based on Apriori: SFIT, "3rd International conference on Electronics Computer Technology (ICECT)", 8-10 April 2011.