Volume 3, Issue 10, October 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Performance Evaluation of K-Means Algorithm and Enhanced Mid-point based K-Means Algorithm on Mining Frequent Patterns

Rachita Sony Krotha,
Department of IT,
GMRIT, Rajam, AP, India.

Satish Muppidi,
Department of IT,
GMRIT, Rajam, AP, India.

Abstract: ***Pattern and classification of stock data is very important for business development in decision making. Timely prediction of latest upcoming trends is also required in business. Clustering is used to generate groups of related patterns, while association provides a way to get generalized rules of dependent variables. Due to increase in the size and complexity of the data, it is impractical to manually analyze, explore, and understand the data. As a result, useful information is often overlooked. Data mining techniques are best suited for analysis of different types of classification, useful patterns extractions and predictions. The of aim this paper is to evaluate the performance of K-Means and proposed enhanced method of K-Means algorithm with improved initial centers using mid-point method for clustering and apply it on Most Frequent Pattern Mining Algorithm to generate frequent patterns***

Keywords: Clustering, K-Means, midpoint based K-Means, MFP, Association Analysis.

I. Introduction

Data mining is a process that uses a variety of data analysis techniques and tools to discover hidden relationships and patterns. The basic approach in Data mining is to summarize the data and to extract reasonable and previously unknown useful information. Data mining techniques like clustering and associations can be used to find meaningful patterns for future predictions. Clustering is used to generate groups of related patterns, while association provides a way to get generalized rules of dependent variables. Cluster analysis has wide applications including market/customer segmentation, pattern recognition, biological studies, and Web document classification. Association rules assist in marketing, targeted advertising, floor planning, inventory control, Banking services, Retail and Supermarkets etc. Nowadays due to the increase in the size and complexity of the data, it is impractical to manually analyze, explore, and understand the data. As a result, useful information is often overlooked. So extracting valid, previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions plays a major role. Today, many algorithms are proposed for clustering and association analysis such as partitional , hierarchical, density based and model based clustering algorithms and for mining frequent patterns, apriori, FP-growth, etc. But, sale data classification has different market trends. Some of the products will be fast selling items, some will be slow selling items and some will be very rare selling items. So, segment-by-segment forecasting needed to produce better results. Data mining techniques are best suited for analysis of different types of classification, useful patterns extractions and predictions. The information and knowledge gained can be used for decision making applications ranging from market analysis, fraud detection and customer retention, production control and science exploration.

II. Related work

Mining of association rules is a field of data mining that has received a lot of attention in recent years. Data mining researchers often try to find most feasible and efficient methods for extraction of useful patterns from stock data. In the paper titled frequent patterns mining of stock data using hybrid clustering association algorithm proposed a methodology where stock data is divided into clusters using K-means algorithm and then to generate the frequent patterns based on Most Frequent pattern Mining algorithm[1]. But K-means algorithm has some limitations the major being the cluster results heavily depend on the selection of initial centroids which caused to converge at local optimum. In another paper titled a mid-point based k-means clustering algorithm for data mining a new enhanced method for K-means algorithm is proposed where a systematic method to determine the initial centroid is explained[2]. In this research we applied this enhanced method for clustering instead of the proposed K-means algorithm in the earlier paper. After clustering is done using the enhanced method then it is applied on for mining patterns of huge stock data to show factors affecting the sale of products to find the frequencies of property values of the corresponding items.

III. Methodology

Proposed Architecture

The proposed architecture[1] is a two phase model. First clusters are generated using K-Means algorithm and then MFP is designed for counting frequencies of items under specified attributes.

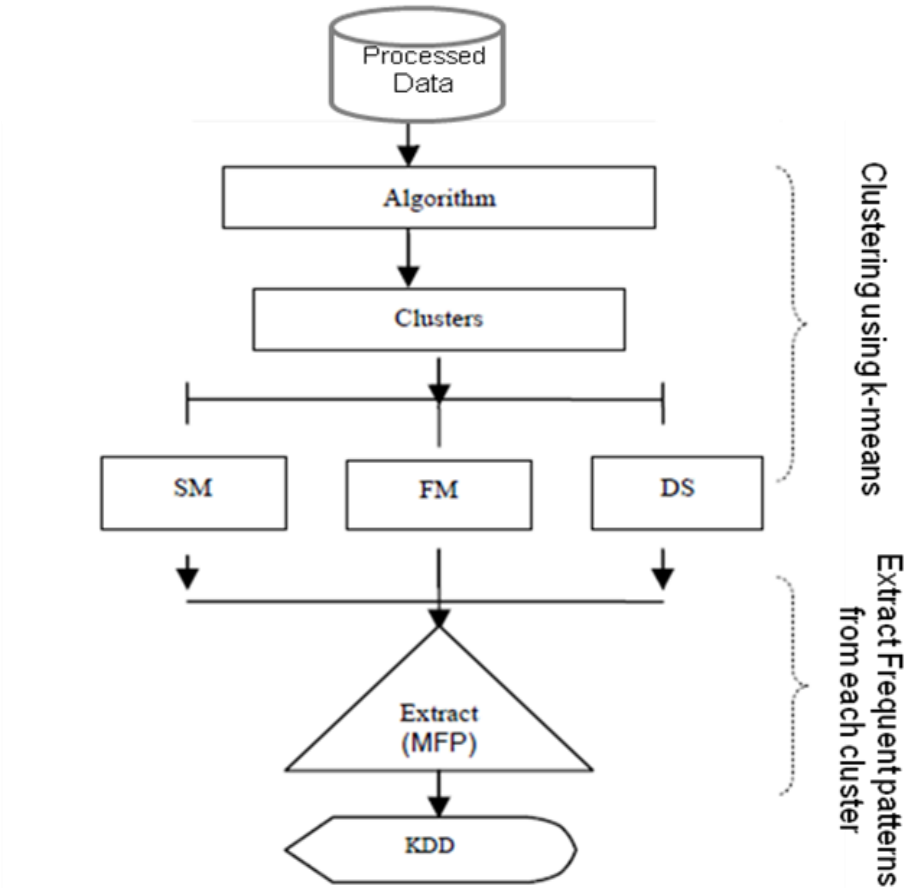

Fig 1: Proposed Architecture

The New proposed Architecture:

Here in this work we followed the same proposed model but instead of the K-Means algorithm used for clustering we have used the enhanced K-Means with improved initial center using Mid-point Method.

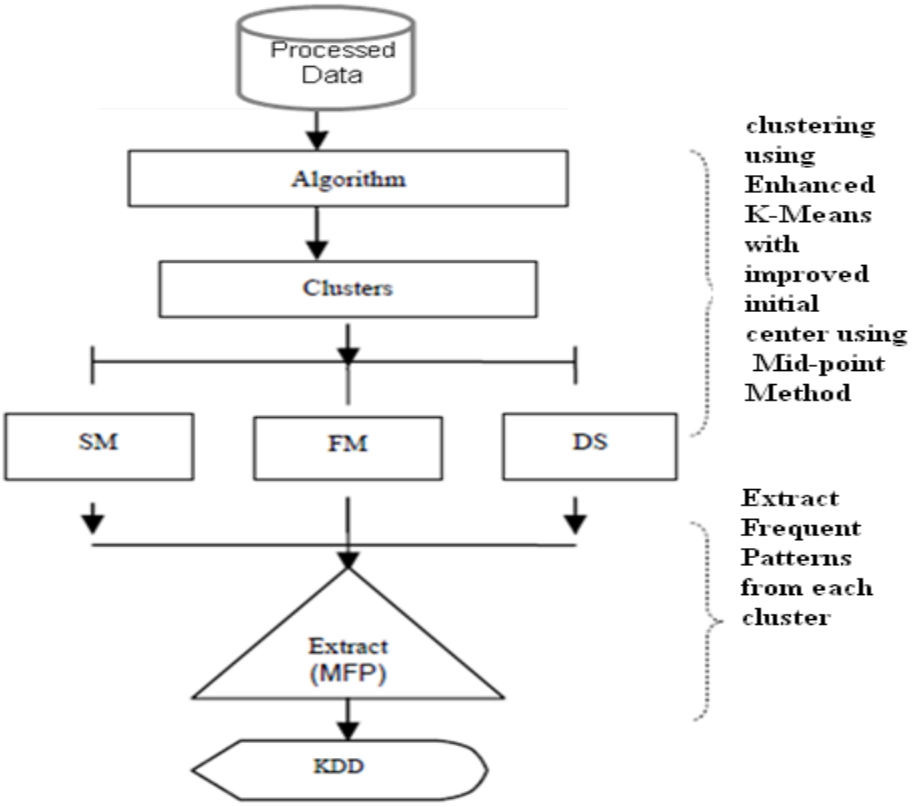

Fig 2:The New Proposed Architecture

K-Means clustering:

The inputs of this algorithm are the number of clusters to be formed i.e., k and the data to be clustered. The algorithm starts with an initial set of cluster centers, chosen at random or according to some heuristic procedure. In each iteration, each instance is assigned to its nearest cluster center according to the Euclidean distance between the two. Then the cluster centers are re-calculated. The center of each cluster is calculated as the mean of all the instances belonging to that cluster. This algorithm aims at minimizing an objective function, in this case a squared error function. The objective function is:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

Where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the *n* data points from their respective cluster centers.

Limitations of K-Means:

Though being a popular algorithm it has some disadvantages being:

1. Final cluster results depend upon the selection of initial centroids
2. Computationally expensive
3. Sensitive to outliers

Enhanced K-Means with improved initial center using Mid-point Method:

This algorithm overcomes the first limitation. Here a systematic method to determine the initial centroids is explained [2]. This method is quite efficient to produce clusters using k-mean method, as compared to taking the initial centroids randomly.

The enhanced method:

Input: D= Set of n data points.
K=desired number of clusters
Output: k number of initial centroids
Steps:

1. In the given data set D, if the data points contain both the positive and negative attribute values then goto step 2, else goto step 4.
2. Find the minimum attribute value in the given dataset D.
3. For each data point attribute, subtract with the minimum attribute value.
4. For each data point calculate the distance from origin.
5. Sort the distances obtained in step 4. Sort the data points in accordance with the distances.
6. Partition the sorted data points into k equal sets.
7. In each set, take the middle point as the initial centroid.

Most frequent patterns method:

Let we have set X of N items in a Dataset having set Y of attributes. This algorithm counts maximum of each attribute values for each item in the dataset. First, it reads all datasets in a cluster. For each item in the cluster, find number of occurrences of each attribute for that item. Now, Find attribute name of the item having maximum count. That means this attribute name is the most frequent attribute value for that item. In the same way, we have to find maximum count for all attributes. All these attribute names of maximum counts of an item gives a most frequent pattern. In this way, we can find most frequent patterns for all items[1].

This algorithm can be explained briefly as below.

Input: *Datasets (DS)*
Output*: Matrix Most Frequent Pattern (MFP):*

```
MFP (DS)
Begin
    For each item Xi in DS
        a. for each attribute
            i. count occurrences for Xi
                C=Count (Xi)
            ii. Find attribute name of C having maximum count
                Mi=Attribute (Ci)
        Next [End of inner loop]
        b. Find Most Frequent Pattern
            i. MFP=Combine (Mi)
    Next [End of outer loop]
```


IV. Experimentation and Results

We developed three programs for this work. The first one is to preprocess the text documents. The second one is to calculate the mid-point method and the third one is for MFP. All of our programs are implemented using java on Intel P III processor with 512MB RAM. We have conducted experiments on sample data to compare the efficiency and accuracy of the proposed mid-point based K-means clustering method over the original K-Means clustering method.

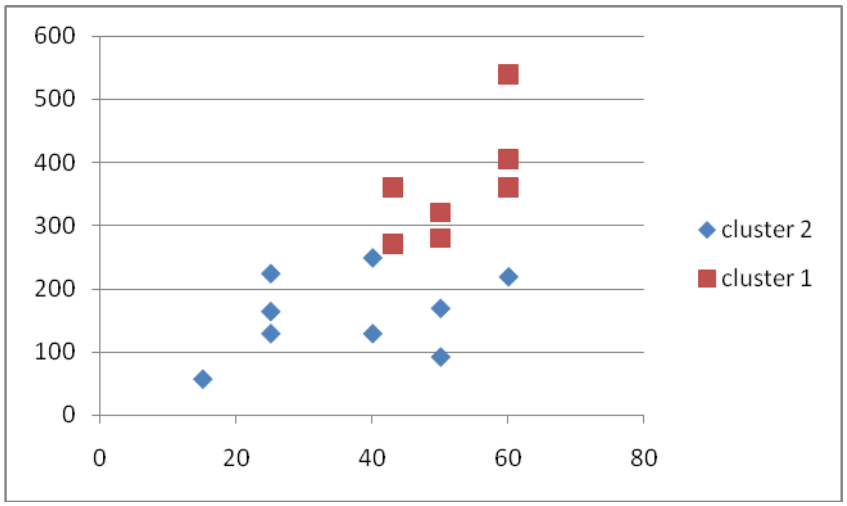

Fig 4: Clustering using K-Means

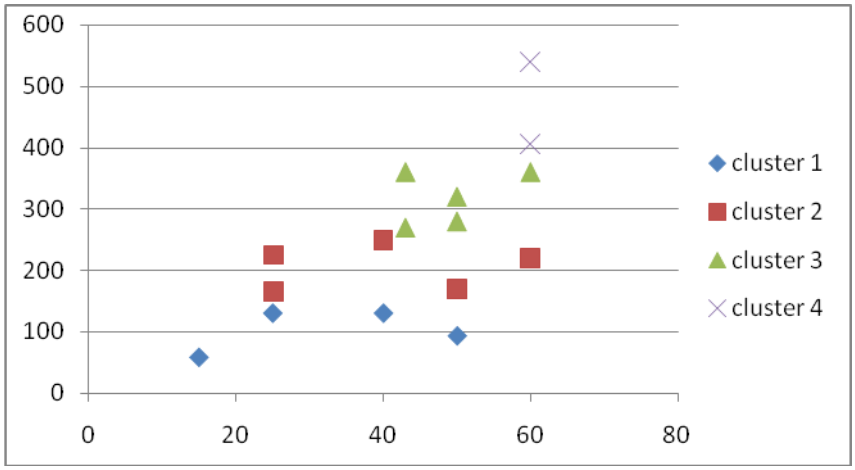

Fig 5: Clustering using enhanced K-Means mid-point method

V. Conclusion

If we observe the above figures it is clearly seen that in K-Means the outliers are not identified and it is restricted to data based on the selection of initial centroids and also it is computationally very expensive. Whereas the proposed midpoint method overcomes these limitations. When applied on Most Frequent Patterns algorithms it generated only most frequent patterns instead of mining all unnecessary frequent patterns. The problem of pattern discovery from stock data mining is addressed. It is clear that proposed approach is very efficient for mining patterns of huge stock data and predicting the factors affecting the sale of products. In future we try to make the process more flexible and efficient. We will also try to implement the same process in document classification and retrieval in unstructured data.

.

References

[1] Aurangazeb Khan, Khairulllah Khan,behram B.Baharuddin "Mining Frequent Patterns Minning Of Stock Data Using Hybrid Clustering Association Algorithm" International Conference 2009.
[2] Neha Aggarwal,Kirti Agarwal " A Mid-point based K-mean clustering algorithm for Data Mining", IJCSE 2012.
[3] Jiawan Han, Micheline Kamber "Data Mining Concepts and Techniques" 2nd edition 2004.
[4] Chaturvedi J. C. A, Green P, "K-modes clustering," J. Classification, (18):35–55, 2001.
[5] Fahim A.M, Salem A. M, Torkey A and Ramadan M. A, "An Efficient enhanced k-means clustering algorithm," Journal of Zhejiang University, 10(7):1626–1633, 2006.
[6] Huang Z, "Extensions to the k-means algorithm for clustering large data sets with categorical values," Data Mining and Knowledge Discovery, (2):283–304, 1998.
[7] Margaret H. Dunham, Data Mining- Introductory and Advanced Concepts,Pearson Education,2006.
[8] Pang-Ning Tan, Michael Steinback and Vipin Kumar, Introduction to Data Mining, Pearson Education, 2007.

[9] Yuan F, Meng Z. H, Zhang H. X and Dong C. R, "A New Algorithm to Get the Initial Centroids," Proc. of the 3rd International Conferenceon Machine Learning and Cybernetics, pages 26–29, August 2004.
[10] Herbert Schildt, "Java 2 The Complete Reference", Osborne.
[11] Gebouw D, B-3590 Diepenbeek, Belgium "Building an Association Rules Framework to Improve Product Assortment Decisions" 2004.
[12] Abubakar, Felix "Customer satisfaction with supermarket retail shopping" 2002.
[13] Brijs, Bart, Gilbert, Koen, Geert "A Data Mining Framework for Optimal Product Selection in Retail Supermarket Data: The Generalized PROFSET Model" 2000.
[14] R. C. Wong, A. W. Fu, K. Wang"Data Mining for Inventory Item Selection with Cross-Selling Considerations" 2005.
[15] Dr. Sankar Rajagopal, "Customer data clustering using data Mining technique" November 2011.
[16] L. Frans, Wei, Paul, "Towards an agent based framework for onlineafter sales services" 2006.
[17] S. Kotsiantis, Kanellopoulos "Association Rules Mining: A Recent Overview" GESTS International Transactions on Computer Science and Engineering, 2006,**32**(1), 71- 82.
[18] R. Agrawal, T. Imielinski and A. Swami, "Mining Association Rules between Sets of Items in Large Databases", Proceedings of ACM SIGMOD Conference, Washington DC, USA, May 1993.

