A SURVEY OF CLUSTERING ALGORITHMS IN ASSOCIATION RULES MINING

Wael Ahmad AlZoubi

Applied Science Department, Ajloun University College, Balqa Applied University.

ABSTRACT

The main goal of cluster analysis is to classify elements into groupsbased on their similarity. Clustering has many applications such as astronomy, bioinformatics, bibliography, and pattern recognition. In this paper, a survey of clustering methods and techniques and identification of advantages and disadvantages of these methods are presented to give a solid background to choose the best method to extract strong association rules.

1. INTRODUCTION

Clustering may be defined as the division of data into groups of similar objects. To get simplification in data representation as clusters, a lot of details when will be ignored. Clustering may be considered as a data modeling technique that gives brief summaries of the data. Ans so, clustering has direct relations with many topics and many applications depend on clustering. The applications of clustering usually deal with large datasets and data with many attributes. This survey concentrates on clustering algorithms from a data mining viewpoint. [2].

There are several cluster-based algorithms for mining association rules from transactional data as Cluster based rule mining (CBAR) algorithm [16], cluster decomposition rule mining (CDAR) algorithm [17], and Partition Algorithm for Mining Frequent Itemsets (PAFI) [18]. Although these algorithms have great influence on the process of mining association rules but they will not be studied in this paper, the concentration will be on clustering techniques rather than clustering algorithms.

This paper is organized as following. Section 2 talks briefly about the process on mining association rules from transaction dataset, section 3 discusses the methods to improve the process of frequent itemsets generation. In section 4 the requirements of clustering in mining of association rules are briefly explained. The unsupervised linear clustering algorithms are discussed, and the advantages and disadvantages of these algorithms have been summarized in section 5. Section 6 explains unsupervised nonlinear clustering algorithms as in section 5, and finally, section 7 concludes this paper.

2. ASSOCIATION RULES MINING

Efficiency in the process of association rules generation mostly depends on the number of database scans required to find out the frequent itemsets with respect to time, that is, the least time-consuming method is the best. Association rules have an important effect in the present market data that particularly requires extraction of the maximal frequent itemsets in an effective manner.

DOI: 10.5121/ijcsit.2019.11202

The two main steps of the process of mining association rules from large market basket databases are: the generation of frequent itemsets and the extraction of strong or confident association rules from the generated set of frequent itemsets.Frequent itemsets are those that have support greater than or equal the user defined minimum support threshold, which is the most time-consuming step and this operation is by far the most expensive phase of the mining process.While the second step is less time-consumingcomparing with the earlier step because each rule is a binary partitioning of a frequent itemset.Confident rules are the association rules that have confidence not less than the user defined confidence threshold [5].

3. Improvements Of Frequent Itemset Generation

The generation of frequent itemsets can be improved through one or more of the following actions: (i) reducing the number of itemsets by using some data pruning techniques, (ii) reducing the number of transactions in the database, or (iii) reducing the number of comparisons by using an efficient data structure to store the candidates or transactions[6].The third option will be selected in this paper to increase the performance and efficiency of frequent itemset generation, i.e. those that have support not less than a predefined support threshold.

Theprocess of association rules extraction from a dataset of transactions faces many challenges, and so it is very important to find an efficient technique to do so, which will be the clustering, i.e. grouping similar transactions or records together according to some criteria.

4. Requirements Of Clustering In Mining Association Rules

There are eight different requirements for efficient clustering process in association rule mining (ARM). They are: (1) Scalability: Data should be scalable, if not incorrect results may occur, (2) Clustering algorithm should be able to deal with various kinds of attributes, (3) Clustering algorithm should be able to find clustered data with the random form, (4) Clustering algorithm should be not sensitive to unprocessed data and outliers, (5) Clustering algorithm should be not sensitive to the organization of input records, (6) Clustering algorithm should be capable to process datasets of high dimensionality, (7) Integration of user-defined constraints, and (8) Interpretability and usability, i.e. the obtained results from clustering should be understandable and functional so that maximum knowledge about the input parameters can be obtained.

Clustering algorithms can be generally classified into two classes: (1) Unsupervised linear clustering algorithms and (2) Unsupervised non-linear clustering algorithms. These are the topics of the following sections.

5. Unsupervised Linear Clustering Algorithms

There are five main unsupervised linear clustering algorithms: (1) K-means clustering, (2) Fuzzy c-means clustering, (3) Hierarchical clustering algorithm (4) Gaussian (EM) clustering algorithm and (5) Quality threshold clustering algorithm. The following subsections explain briefly these algorithms.

5.1. K-Means Clustering Algorithm

K-means is one of the easiest unsupervised mining algorithms that explain the familiar clustering problem. K-means starts by dividing a given dataset into a certain number of clusters (k clusters), where k is a positive integer known previously. Other k integers are to be defined – different from the previous known integers – one for each cluster. These centers should be placed in a clever way since the difference in the location of centers will lead to unusual results. Therefore, these centers should be placed as much as possible far away from each other. Then each data point is taken and associated with the nearest center. K-means algorithm hasfour steps:

1. Associating each point with its nearest center.
2. Re-calculation of k new centers of the clusters generated from the previous step.
3. Associating all the original data points with the nearest new centers.
4. Repeating the second and third steps until the centers take their final locations and no extra changes are required.

The main goal of k-means algorithm is minimizing the error computed by formula 1, which is sometimes known as mean squared error (MSE) function:

$$MSE(X) = \sum_{i=1}^{c}\sum_{j=1}^{c_i}(\| y_i - x_j \|)^{2} \quad (1)$$

Where,

||y_i - x_j|| is the Euclidean distance between y_i and $x_{j.}$
c_i is the number of data points in i^{th} cluster.
c is the number of cluster centers.

MSE must be positive and close to zero to give the best quality of an estimator. [9]

5.2 Fuzzy *C*-Means Clustering Algorithm

In this algorithm, each data point is associated with a cluster center known previously depending on the distance between the data point and the cluster center, such that each data point is associated with the nearest cluster center. Membership and cluster centers are updated according to formula 2 and 3 given below:

$$\mu_{ij} = 1/\sum_{k=1}^{c}(d_{ij}/d_{ik})^{(2/m-1)} \quad (2)$$

$$v_j = \left(\sum_{i=1}^{n}(\mu_{ij})^m x_i\right)/\left(\sum_{i=1}^{n}(\mu_{ij})^m\right), \forall j = 1,2,...,c \quad (3)$$

Where *n* is the number of data points, $v_j$is the j^{th} cluster center, *m* is the fuzziness index such that $m \in [1,\infty)$, *c* is the number of cluster centers, μ_{ij} represents the membership ofi^{th} data to j^{th} cluster center, and d_{ij} represents the Euclidean distance between i^{th} data and j^{th} cluster center.

The central goal of fuzzy *c*-means algorithm is to minimize the Euclidean distance between i^{th} data and j^{th} cluster center[7].

$$J(U,V) = \sum_{i=1}^{n}\sum_{j=1}^{c}(\mu_{ij})^m \| x_i - v_j \|^2 \quad (4)$$

Where ||x_i– v_j|| is the Euclidean distance between i^{th} data and j^{th} cluster center.

Figure 1 presents a sample result of fuzzy *c*-means clustering.

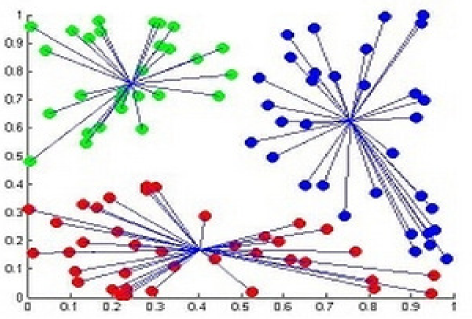

Figure 1. Result of Fuzzy c-Means Clustering [7]

5.3 HIERARCHICALCLUSTERING ALGORITHM

The main two kinds of hierarchical clustering algorithms are:

(i) Agglomerative Hierarchical clustering algorithm.
(ii) Divisive Hierarchical clustering algorithm.

Those two kinds of algorithms are reverse of each other. So, explaining one of them is enough to understand the other one, in the following subsection the agglomerative hierarchical clustering method has been discussed in some detail.

5.4 AGGLOMERATIVE HIERARCHICAL CLUSTERING

This method sometimes called bottom up clustering that starts by grouping similar data pointstogether. This type of clusteringbegins by considering each object as a cluster. Then, pairs of clusters are consecutivelycombined until all clusters have been combined into one big cluster covering all objects.Many techniques may be used to compute the distance between each pair of data points; some of these techniques are [11]:

(i) Single-nearest distance: the distance between two clusters is computed by a single element pair, specifically those two elements (one in each cluster) that are closest to each other.

(ii) Complete: The distance between two clusters is defined as the maximum value of all pairwise distances between the elements in some cluster and the elements in another cluster. It tends to produce more compressed clusters.

(iii) Average: average distance among data points in each cluster

(iv) Centroid distance: divide the average distance by the number of data points in the cluster.

(v) Ward's method: sum of squared Euclidean distance is minimized.

The results of hierarchical clustering are usually displayedas a dendrogram. Dendrogram is a Greek concept means drawing tree, it consists of two parts, the first part is Dendron which means tree and the second part is gramma which means drawing. This diagrammatic representation is frequently used in different contexts. The number of clusters is calculated exactly depending on the dendrogram graph, as in figure 2.

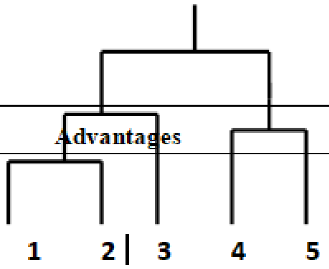

Figure 2 Dendrogram formed from the data set of size $N = 5$

5.5 Gaussian (Expectation Maximization Em) Clustering Algorithm

nGaussian points are supposed firstly, then the data points will be fit into the ***n*** Gaussians by expecting the classes of all data points and then making the best use of the maximum probability of Gaussian centers. The main advantage of this algorithm is that it gives convenient result for the real-world data set, while it suffers from high complexity. [10]

5.6 Quality Threshold (Qt) Clustering Algorithm

One of the requirements of the QT algorithm is an earlier identification of the threshold distance within the cluster and the minimum number of elements in each cluster. Each data point is used to find its candidates [14]. Candidate data points are those which are within the range of the threshold distance from the given data point. A cluster is formed from grouping data points with large number of candidates such that the candidate data points for each data point are found in the previous way. This process is repeated with the minimized set of data points – as the data points that belong to the formed cluster are deleted – until no more cluster can be formed satisfying the minimum size constraint.

5.7 Unsupervised Linear Clustering Algorithms Summary

Table 1 presents the unsupervised clustering algorithms discussed inprevious sections

Algorithm	Advantages	Disadvantages
K- means Algorithm	– Fast, robust and easier to understand. – Relatively efficient: O (tknd), where n is number of objects, k is number of clusters, d is the dimension of each object, and t is number of iterations. Normally, k, t, d << n. – Gives best result when data set are distinct or well separated from each other.	– The learning algorithm requires previous specification of the number of cluster centers. – The use of exclusive assignment - If there are two highly overlapping data then k-means will not be able to resolve that there are two clusters. – The learning algorithm is not invariant to non-linear transformations i.e. with different representation of data we get – Applicable only when mean is defined i.e. fails for categorical data. – Unable to handle noisy data and outliers. Algorithm fails for non-linear data set

Fuzzy c-means Algorithm	– Gives best result for overlapped data set and comparatively better then k-means algorithm. Data point is assigned membership to each cluster center because of which data point may belong to more than one cluster center.	– The number of clusters must be specified before beginning. – Better results have been gotten if lower value of β are used but at the expense of more number of iteration. Euclidean distance measures can unequally weight underlying factors. – Algorithm can never undo what was done previously. – Time complexity of at least O(n^2 logn) is required, where ***n*** is the number of data points. – Based on the type of distance matrix chosen for merging different algorithms can suffer with one or more of the following:
Agglomerative Hierarchical clustering	– No prior information about the number of clusters required. – Easy to implement and gives best result in some cases.	– Sensitivity to noise and outliers – Breaking large clusters – Difficulty handling different sized clusters and convex shapes – No objective function is directly minimized – Sometimes it is difficult to identify the correct number of clusters by the Dendogram
Gaussian (EM) clustering	Gives very useful result for the real-world data set.	Algorithm has high complexity
Quality Threshold (QT) clustering [12]	– Quality Guaranteed - Only clusters that pass a user-defined quality threshold will be returned. – Number of clusters is not specified previously. – All possible clusters are considered - Candidate cluster is generated with respect to every data point and tested in order of size against quality criteria	– Computationally Intensive and Time Consuming – Increasing the minimum cluster size or increasing the number of data points can greatly increase the computational time. – Threshold distance and minimum number of elements in the cluster must be defined firstly.

5. UNSUPERVISED NON-LINEAR CLUSTERING ALGORITHMS

The main unsupervised non-linear clustering algorithms are: (1) MST based clustering algorithm, (2) Kernel k-means clustering algorithm, and (3) Density based clustering algorithm.

6.1 MINIMUM SPANNING TREE (MST) BASED CLUSTERING ALGORITHM

MST based clustering algorithm [15] starts by constructing MST using Kruskal's algorithm, which is a greedy algorithm in graph theory introduced by Joseph Kruskal in 1956 that aims to find the cheapest link available between two points, and then set a threshold value and step size. After that the edges whose lengths are greater than the threshold values are removed from the MST. Then the percentage between the intra-cluster distance, i.e. the distance between clusters, and inter-cluster distance, i.e. distance between data points within a cluster, is calculated and the ratio is recorded in addition to the threshold.

The threshold value is changed by adding the step size, and this process is repeated every time the threshold value is changed until the threshold value takes the maximum value and no edges can be deleted. In such case, all the data points belong to a single cluster. Finally, we obtain the minimum value of the recorded ratio and form the clusters corresponding to the stored threshold value.

The MST based clustering algorithm has two exceptional cases: (1) when the threshold value equals zero, this means that each point will be in a single cluster, and (2) when the threshold value takes a maximized value, this means that all the points remain within a single cluster.
So, the MST based clustering algorithm looks for that best value of the threshold for which the Intra-Inter distance ratio is minimized. The initial threshold value should not be equal zero to decrease the number of iterations.

6.1 KERNEL *K*-MEANS CLUSTERING ALGORITHM

Kernel *k*-Means Clustering algorithm differs from the *k*-means in that a kernel method is used rather than the Euclidean distance to calculate the distance between clusters and within a cluster.

6.2 DENSITY BASED CLUSTERING

One of the most well-known algorithms that represent density-based clustering is DBSCAN algorithm. DBSCAN (**D**ensity-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise) [13] has played a very important task in finding nonlinear shapes structure based on the density. It uses the concept of density reachability, i.e. a point "*p*" is said to be density reachable from a point "*q*" if point "*p*" is within ε distance from point "*q*" and "*q*" has sufficient number of points in its neighbors which are within distance ε, and density connectivity, i.e.a point "*p*" and "*q*" are said to be density connected if there exist a point "*r*" which has sufficient number of points in its neighbors and both the points "*p*" and "*q*" are within the ε distance. This process takes the shape of series, such that, if "*q*" is neighbor of "*r*", "*r*" is neighbor of "*s*", "*s*" is neighbor of "*t*" which in turn is neighbor of "*p*" implies that "*q*" is neighbor of "*p*".

6.3 UNSUPERVISED LINEAR CLUSTERING ALGORITHMS SUMMARY

The advantages and disadvantages of the previous unsupervised clustering algorithms are displayed in table 2.

Table 2 Advantages and disadvantages of Unsupervised Non-Linear Clustering Algorithms

Algorithm	Advantaged	Disadvantages
MST based clustering algorithm	– Comparatively better performance then k-means algorithm.	– Threshold value and step size needs to be defined firstly.
DBSCAN	– Does not require a-priori specification of number of clusters. – Able to identify noise data while clustering. – DBSCAN algorithm can find arbitrarily size and arbitrarily shaped clusters.	– DBSCAN algorithm fails in case of varying density clusters. – Fails in case of neck type of dataset. – Does not work well in case of high dimensional data.
Kernel *k*-means	– Kernel k-means can identify the non-linear structures. – Kernel k-means is best suited for real life datasets.	– Number of cluster centers need to be predefined. – It is complex in nature and time complexity is large.

7. SUMMARY

This paper explains the different cluster-based algorithms and techniques and compares between them to enable the researcher to select the best one suitable to his/her data, besides that this paper talks about the requirements of clustering for association rules extraction.

As mentioned in the tables in the previous section, every technique has its benefits and drawbacks, this may give flexibility to the data scientist to select the most convenient method according to the data available. clustering analysis can be used to increase some valued visions from the available data by putting the data points into the most appropriate cluster.The most well-known clustering technique is k-means, K-Means isfast, because it computes the distances between points and group centers; very few computations,ittherefore has a linear complexity.

The future work will study extensively the clustering algorithms in terms of their efficiency, usability and flexibility.

REFERENCES

1) Dhillon, I. S., Guan, Y. and Kulis, B. Kernel k-means: spectral clustering and normalized cuts. Proceeding of KDD '04 Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. Seattle, WA, USA — August 22 - 25, 2004.

2) Berkhin, P. A Survey of Clustering Data Mining Techniques. United States, North America: Springer, 2006. PP. 25 – 71.

3) AlZoubi, W. A. An Improved Clustered Based Technique for Frequent Items Generation from Transaction Datasets. CCIT 2018.

4) Moreira, A. Density-based clustering algorithms – DBSCAN and SNN. Version 1.0, 25.07.2005, University of Minho – Portugal.

5) Han, J., Cheng, H., Xin, D., & Yan, X. 2007. Frequent pattern mining: current status and future directions. Data Mining Knowledge Disc (2007), pp. 55–86.

6) Astashyn, A. 2004. Deterministic Data Reduction Methods for Transactional Datasets. Master Thesis. Polytechnic University. http://photon.poly.edu/~hbr/publi/alex_msthesis.pdf.

7) Pal N. R., Pal K., Keller J. M., and Bezdec J. C.2006. A possibilistic fuzzy c-means clustering algorithm. IEEE Transactions on Fuzzy Systems. Issue 4, Volume 13, August2005, pp. 517 – 530.

8) Alfred R. &Dimitar, K. 2007. A Clustering Approach to Generalized Pattern Identification Based on Multi-instanced Objects with DARA. In Local Proceedings of ADBIS. Varna. pp. 38 – 49.

9) Khan S. and Ahmad A. Cluster center initialization algorithm for K-means clustering. Pattern Recognition Letters. Volume 25, Issue 11, August 2004, pp. 1293 – 1302.

10) Fraley, C. Algorithms for Model-Based Gaussian Hierarchical Clustering. SIAM Journal on Scientific Computing, 1998, Vol. 20, No. 1. pp. 270-281.

11) Eyal Salman, H., Hammad, M., Seriai, A. and Al-Sbou, A. Semantic Clustering of Functional Requirements Using Agglomerative Hierarchical Clustering. Information 2018, 9, 222; doi:10.3390/info9090222. www.mdpi.com/journal/information.

12) Heyer L, Kruglyak S, Yooseph S (1999) Exploring expression data: identification and analysis of co-expressed genes. Genome Res 9:1106–1115.

13) Rodriguez A, Laio A. Clustering by fast search and find of density peaks. Science 27 Jun 2014: Vol. 344, Issue 6191, pp. 1492-1496 DOI: 10.1126/science.1242072.

14) Song M, Christian W. Günther, Wil M. P. van der Aalst. Trace Clustering in Process Mining. International conference on Business Process Management (BPM 2008): Business Process Management Workshops pp 109-120.

15) T. Asano, B. Bhattacharya, M. Keil, and F. Yao. Clustering algorithms based on minimum and maximum spanning trees. In Proceedings of the 4th Annual Symposium on Computational Geometry, pages 252-257, 1988.

16) Tsay, Y.-J. & Chiang, J.-Y. 2005. CBAR: an efficient method for mining association rules. Knowledge-Based Systems 18 (2005), pp. 99–105.

17) Tsay, Y.-J. &Chien.-C, Y.-W. 2004. An efficient cluster and decomposition algorithm for mining association rules. Information Sciences 160 (2004) 161–171.

18) Hanirex, K &Rangaswamy, D. 2011. Efficient algorithm for miningfrequent itemsets using clustering techniques.International Journal on Computer Science and Engineering (IJCSE), Vol. 3 No. 3 Mar 2011, pp. 1028 - 1032.