Sanjeeva Rao Sanku* et al.
(IJITR) INTERNATIONAL JOURNAL OF INNOVATIVE TECHNOLOGY AND RESEARCH
Volume No.4, Issue No.6, October – November 2016, 4486-4489.

A Theoretical Analysis On Clustering Concepts In Data Mining

Mr. SANJEEVA RAO SANKU
Asst.Professor, Dept.of CSE
TeegalaKrishnaReddy Engineering College
Hyderabad.

Mr. LALBAHADUR KETHAVATH
Asst.Professor, Dept.of CSE
TeegalaKrishnaReddy Engineering College
Hyderabad.

Abstract: **Clustering mechanism is the unsupervised classification of patterns observations data items or feature vectors into different clusters. This type of clustering problem has been addressed in many contexts and by researchers in different domains, this makes us to understand its broad appeal and usefulness as one of the steps analyzing the whole data. As we all know that there will be huge assumptions in solving the clustering problems which makes it very complex and the clustering process became very slow. Here in this paper we are concentrating on overview of pattern clustering methods from a statistical pattern recognition perspective with a goal of providing useful advice and references to fundamental concepts accessible to the broad community of clustering practitioners. We also presents the study of different clustering algorithms as well as the current development in these mechanisms.**

Keywords— **Data mining; cluster; classification;**

I. INTRODUCTION

Analyzing the data requires many computing operations to be performed. These operations may be performed in designing them or dynamically they can be implemented whenever it requires. These analysis procedures are explanatory or confirmatory. In these two mechanism the main approach is to classify or to group the data. This was done based on the good-nes-soft to a postulated model or natural groupings which is known as clustering. Clustering is a kind of concept which deals with grouping of similar or related data together to form a group or cluster. The following figure 1.a explains the input pattern which is given and figure 1.b explains the formation of the clusters.

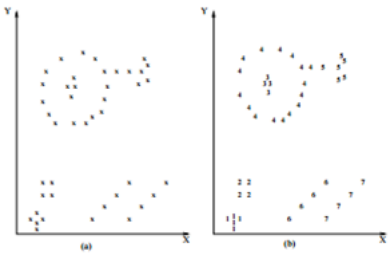

Fig. 1. Data clustering.

In the above figure, the points which belongs to the same cluster given as the same name. There are different techniques are existed to implement the clustering process. These techniques implements proximity in between the data elements and making them as clusters. The main uses of clustering can be had in the concepts like exploratory pattern analysis grouping decision making and machine learning situations including data mining document retrieval image segmentation and pattern classification.

Custering Tasks:

The following are the different activities which are involved in forming the given patterns into clusters. They are

- Pattern representation.
- Defining the pattern similarity measure appropriate to the data domain.
- Clustering or grouping.
- Data abstraction if needed.
- Checking the output.

All these steps are represented in the following figure 2.

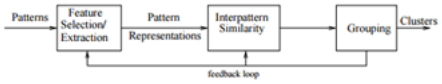

Figure 2: Steps involved in creation of clusters.

The main step in the above steps to form the clusters are pattern proximity(similarity). Pattern proximity is usually measured by a distance function defined on pairs of patterns. There exists different mechanism in implementing the distance measure in different areas. Apart from all of them, widely used Euclidean distance is implemented to show the differences between the two patterns. Here similarity measure is used to make the given patterns into clusters.

II. CLUSTERING TECHNIQUES:

There exists different clustering mechanisms which will be implemented in forming the clusters. The following figure 3 gives the information about the different clustering algorithms.

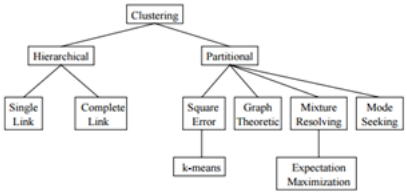

Figure 3: Clustering Taxonomy.

As per the above figure 3, all the clustering algorithms will fall into two categories like hierarchical and partitional. Hierarchical algorithms are single link and complete link algorithms. Where as the partitional algorithms are K-means, Graph theoretic, mode seeking and expectation maximization. All these algorithms are categorized into two approaches, they are Agglomerative Vs Devisive and Monothetic Vs Polythetic.

Agglomerative vs divisive: This approach takes the input as a pattern and considers as singleton cluster and continues merging the items into the cluster until it meets the stopping criteria. A divisive method begins with all patterns in a single cluster and performs splitting until a stopping criterion is met.

Monothetic vs polythetic: Another approach which relates to the sequential or simultaneous use of features in the clustering process. Many of the algorithms are polythetic because of giving the input as all features for computation of distances between patterns and decisions are based on those distances. Whereas, simple monothetic algorithm reported, considers features sequentially to divide the given collection of patterns.

III. CLUSTERING ALGORITHMS:

Hierarchical Algorithms: The most popular hierarchical algorithm are single link and complete link algorithms. The hierarchical algorithm will take the dataset from the figure 4 and performs the operations.

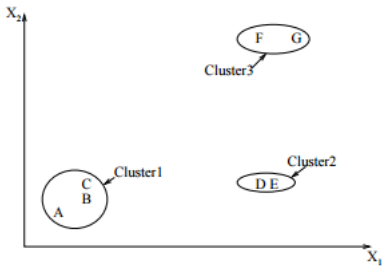

Figure 4: Example pattern to implement the Hierarchical clusters.

In the single link method the distance between two clusters is the minimum of the distances between all pairs of patterns drawn from the two clusters whereas the complete link algorithm is the distance between two clusters is the maximum of all pairwise distances between patterns in the two clusters. In either case two clusters are merged to form a larger cluster based on minimum distance criteria. Given figure 5 is a dendogram which represent the cluster formation by using the single link approach.

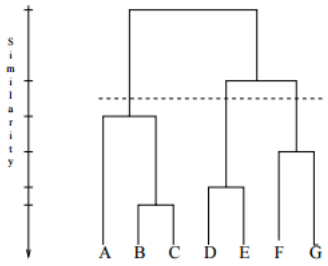

Figure 5: Dendogram using the single link cluster.

Single Link Algorithm:

Steps:

- Place each pattern in its own cluster Construct a list of inter pattern distances for all distinct unordered pairs of patterns and sort this list in ascending order.
- Step through the sorted list of distances forming for each distinct dissimilarity value d_k a graph on the patterns where pairs of patterns closer than d_k are connected by a graph edge. If all the patterns are members of a connected graph stop, Otherwise repeat this step.
- The output of the algorithm is a nested hierarchy of graphs which can be cut at a desired dissimilarity level forming a partition clustering identified by simply connected components in the corresponding graph.

Partitional Algorithms:

A partitional clustering algorithm obtains a single partition of the data instead of a clustering structure. Partitional methods have advantages in applications involving large data sets for which the construction of a dendrogram is computationally prohibitive. A problem accompanying the use of a partitional algorithm is the choice of the number of desired output clusters.

Squared Error Clustering Method:

- Select an initial partition of the patterns with a fixed number of clusters and cluster centers.
- Assign each pattern to its closest cluster center and compute the new cluster centers as the centroids of the clusters. Repeat this step until convergence is achieved ie until the cluster membership is stable.
- Merge and split clusters based on some heuristic information optionally repeating step.

kMeans Clustering Algorithm:

- Choose k cluster centers to coincide with k randomly chosen patterns or k randomly defined points inside the hyper volume containing the pattern set.
- Assign each pattern to the closest cluster center.
- Recompute the cluster centers using the current cluster membership.
- If a convergence criterion is not met go to step 2. Typical convergence criteria are no (or minimal) reassignment of patterns to new cluster centers or minimal decrease in squared error.

Algorithm:

Initialize $\mathbf{m}_i$, i = 1,…,k, for example, to k random $\mathbf{x}^t$

Repeat

For all $\mathbf{x}^t$ in X

$b_i^t \leftarrow 1$ if $\| \mathbf{x}^t - \mathbf{m}_i \| = \min_j \| \mathbf{x}^t - \mathbf{m}_j \|$

$b_i^t \leftarrow 0$ otherwise

For all $\mathbf{m}_i$, i = 1,…,k

$\mathbf{m}_i \leftarrow$ sum over t $(b_i^t \mathbf{x}^t)$ / sum over t (b_i^t)

Until $\mathbf{m}_i$ converge.

The vector **m** contains a reference to the sample mean of each cluster. **x** refers to each of our examples, and **b** contains our "estimated [class] labels"

Nearest Neighbor Clustering:

Since proximity plays a key role in our intuitive notion of a cluster nearest neighbor distances can serve as the basis of clustering procedures. An iterative procedure was proposed. it assigns each unlabeled pattern to the cluster of its nearest labeled neighbor pattern provided the distance to that labeled neighbor is below a threshold. The process continues until all patterns are labeled or no additional labeling occur. The mutual neighborhood value can also be used to grow clusters from near neighbors.

Fuzzy Clustering:

Algorithm:

Select an initial fuzzy partition of the N objects into K clusters by selecting the NxK membership matrix U. An element u_{ij} of this matrix represents the grade of membership of object x_i in cluster c_i. Typically u_{ij}€[0.1].

- Using U find the value of a fuzzy criterion function eg a weighted squared error criterion function associated with the corresponding partition. One possible fuzzy criterion function is

$$E^2(\mathcal{X}, \mathbf{U}) = \sum_{i=1}^{N} \sum_{k=1}^{K} u_{ij} \|\mathbf{x}_i - \mathbf{c}_k\|^2,$$

where $\mathbf{c}_k = \sum_{i=1}^{N} u_{ik}\mathbf{x}_i$ is the k^{th} fuzzy cluster center.

- Repeat step2 until entries in U do not change significantly.

An Evolutionary Algorithm for Clustering:

1. Choose a random population of solutions. Each solution here corresponds to a valid k-partition of the data. Associate a finitness value with each solution. Typically finiteness is inversely proportional to the squared error value. A solution with a small squared error will have a larger finitness value.
2. Use the evolutionary operators selection recombination and mutation to generate the next population of solutions. Evaluate the finiteness values of these solutions.
3. Repeat step 2 until some termination condition is satisfied.

IV. CONCLUSION

In this paper, we have presented the concept of clustering and different mechanisms of the clustering. Apart from this we have even analyzed different clustering algorithms and presented the concept of different algorithms.

V. REFERENCES

[1]. E. Aarts and J. Korst Simulated Annealing and Boltzmann Machine A Stochastic Approach to Combinatorial Optimization and Neural Computing. John Wiley Sons. Computing Reviews.

[2]. ACM CR Classications. ACM Computing Reviews.

[3]. K. S. Al Sultan, A Tabu Search Approach to Clustering Problem, Pattern Recognition.

[4]. Taghizadeh Yazdi, Mohammad Reza. *Journal of Organizational Change Management* , 2015, Vol. 28 Issue 3, p469-485, 17p. Publisher: Emerald Group Publishing Limited.

[5]. PANDIAN, P. SENTHIL; SRINIVASAN, S. *Journal of Multiple-Valued Logic & Soft Computing.* 2016, Vol. 26 Issue 3-5, p205-220. 16p. 1 Chart..

[6]. Pan Su; Changjing Shang; Qiang Shen. *Journal of Intelligent & Fuzzy Systems.* 2015, Vol. 28 Issue 6, p2409-2421. 13p. DOI: 10.3233/IFS-141518.

[7]. Aparna, K.; Nair, Mydhili K. *International Journal of Technology*. 2016, Vol. 7 Issue 4, p691-700. 10p. DOI: 10.14716/ijtech.v7i4.1579.

[8]. Nanda, Satyasai Jagannath; Panda, Ganapati. In Swarm and Evolutionary Computation. June 2014 16:1-18 Language: English.

DOI: 10.1016/j.swevo.2013.11.003, Database: ScienceDirect

[9]. MENÉNDEZ, HÉCTOR D.; BARRERO, DAVID F.; CAMACHO, DAVID.*International Journal of Neural Systems*. May2014, Vol. 24 Issue 3, p1-19. 19p. DOI: 10.1142/S0129065714300083.

[10]. Marques, Ana Rita; Neto, Jose Soares Ferreira; Ferreira, Fernando. InAquaculture. 1 October 2016 463:106-112 Language: English. DOI: 10.1016/j.aquaculture. 2016.05.012, Database: ScienceDirect.

AUTHOR's PROFILE

SANJEEVA RAO SANKU is working as Asst. Professor in COMPUTER SCIENCE AND ENGINEERING department in TEEGALA KRISHNA REDDY ENGINEERING COLLEGE, Meerpet, Hyderabad. He has 10+ years of teaching experience. His interesting areas are Data Mining, Computer Networks, Image Processing. He has attended number of workshops in various engineering colleges and universities. He has conducted number of workshops in various colleges.

LALBAHADUR KETHAVATH is working as Asst. Professor in COMPUTER SCIENCE AND ENGINEERING department in TEEGALA KRISHNA REDDY ENGINEERING COLLEGE, Meerpet, Hyderabad. He has 10+ years of teaching experience. His interesting areas are Data Mining, Data bases, Computer Networks. He has attended number of workshops in various engineering colleges and universities. He has conducted number of workshops in various colleges.

