International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064

Clustering Algorithm Based on Efficient Centroids for Schema

Beninda Churchil[1], Naganathan E. R[2], John Aravindhar .D[3]

[1]M.Tech (CSE), Hindustan University, Chennai, India

[2]M.E, (Ph.D.), Associate Professor, Hindustan University, Chennai, India

[3]Ph.D, Professor, Hindustan University, Chennai, India

Abstract: ***Cluster analysis is an important process of grouping similar and dissimilar items. Clustering can be done efficiently by the help of finding centroid. Centroids are playing major role in efficient clustering. Clustering can be done without any outliers by grouping them efficiently based on the centroids. Clustering will be efficient for all kind of inputs. This is the main aim of our paper. We give the input in the form of two tables. We can check how efficient is the clustering if the input is in the form of tables. We use a algorithm named as Lloyd`s algorithm for making the clustering efficient. We need to find the average mean distance which should be minimum in all the cases. For this purpose we use a distance metric named as Earth mover distance metric which finds the mean value for all the entries in the table we give. Finally our clustering is analyzed with the performance check in the terms of precision and recall.***

Keywords: clustering, Lloyd`s algorithm, Earthmover distance, centroids.

1. Introduction

Data mining makes use of ideas, tools, and methods from other areas, especially computational area such as database technology and machine learning. It is not much concerned with all areas in which statisticians are interested. Mining essentially assumes that the data have already been collected, and is concerned with how to discover its secrets. It is not a one short activity, but rather an iterative and interactive process. There are clear overlaps between Statistics and Data mining. Data mining should be the nontrivial process of identifying valid, novel, potentially useful, and ultimately comprehensible knowledge from databases such knowledge should be useful in making crucial decisions.

They are more generally significant issues of existing and future KDD. They hide the shift from data mining to knowledge discovery, in particular, blocking the shift from hidden pattern mining to actionable knowledge discovery. The wide acceptance and deployment of data mining in solving complex enterprise applications is thus further restrained. Moreover, they are closely related and to some extent create a cause-effect relation, which is the involvement of domain intelligence contributing to actionable knowledge delivery. This paper explores the challenges and issues from the following aspects:

- Organizational and social factors surrounding data mining applications;
- Human involvement and preferences in the data mining process
- Actionable knowledge discovery supporting decision-making actions;
- Decision-support knowledge delivery facilitating corresponding decision-making,
- Consolidation of the relevant aspects for decision-support.

Clustering will be efficient for all kind of inputs. Tables can also be efficiently clustered with the centroid based clustering algorithms. Average mean distance is also found for clustering efficiently.

2. Related Works

2.1 Domain Driven Data Mining

The basic idea of domain driven data mining (DDDM) is as follows. On top of the data-centered framework, it aims to develop proper methodologies and techniques for integrating domain knowledge, human role and interaction, organizational and social factors, as well as capabilities and deliverables toward delivering actionable knowledge and supporting business decision-making action-taking in the KDD process. DDDM targets the discovery of actionable knowledge in the real business environment. Such research and development is very important for developing the next generation data mining methodologies and infrastructures. Most importantly, DDDM highlights the crucial roles of ubiquitous intelligence, including in-depth data intelligence, domain intelligence and human intelligence, and their consolidation, by working together to tell hidden stories in businesses, exposing actionable knowledge to satisfy real user needs and business operation decision making. End users hold the right to say “good” or “bad” to the mined results

2.2 Ubiquitous Intelligence

This section have stated the importance of involving and consolidating relevant ubiquitous intelligence surrounding data mining applications for actionable knowledge discovery and delivery. Ubiquitous intelligence surrounds a real-world data mining problem. *DDDM* identifies and categories ubiquitous intelligence into the following types. Data intelligence reveals interesting stories and/or indicators hidden in data about a business problem. The intelligence of data emerges in the form of interesting patterns and actionable knowledge. There are two levels of data intelligence:

2.3 General level of data intelligence

It refers to the patterns identified from explicit data, presenting general knowledge about a business problem, and

2.4 In-depth level of data intelligence

It refers to the patterns identified in more complex data, using more advanced techniques, disclosing much deeper information and knowledge about a problem. Taking association rule mining as an example, a general level of data intelligences frequent patterns identified in basket transactions, while *associative classifiers* reflect deeper levels of data intelligence. Human intelligence refers to explicit or direct involvement of human knowledge or a human as a problem-solving constituent, etc., and implicitor indirect involvement of human knowledge or a human as a system component. Domain intelligence refers to the intelligence that emerges from the involvement of domain factors and resources in pattern mining, which wrap not only a problem but its target data and environment. The intelligence of domain is embodied through the involvement into KDD process, modeling and systems.

- Network and web intelligence refers to the intelligence that emerges from both web and broad-based network information, facilities, services and processing surrounding a data mining problem and system.
- Organizational Intelligence refers to the intelligence that emerges from involving organization-oriented factors and resources into pattern mining. The organizational intelligence is embodied through its involvement in the KDD process, modeling and systems.

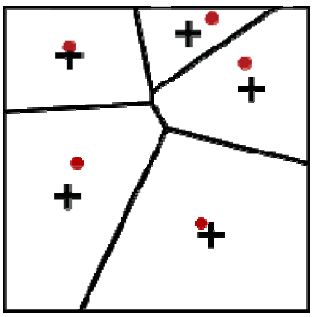
.

3. Search Related Terminologies

This section introduces general search-related terminology that reappears throughout this system. In our paper clustering reappears and we chose specific clustering algorithm for clustering that's based on centroids. In the chosen algorithm the centroids will be changing often. Centroids can be chosen based on the closest mean of the entries of the table. The closest mean will be the minimum one. This will be found with the help of Earth mover distance formula. The main Concept introduces a technique of clustering efficiently based on centroids for inputs in the kind of tables.

3.1 Lloyd`s Algorithm

Lloyd's algorithm is known for finding evenly-spaced sets of points in subsets of Euclidean spaces and partitions of these subsets into well-shaped and uniformly sized convex cells. Like the closely related *k*-means clustering algorithm, it repeatedly finds the centroid of each set in the partition, and then re-partitions the input according to which of these centroids is closest. However, Lloyd's algorithm differs from *k*-means clustering in that its input is a continuous geometric region rather than a discrete set of points. Thus, when re-partitioning the input, Lloyd's algorithm uses Voronoi diagrams rather simply determining the nearest center to each of a finite set of points as the *k*-means algorithm does. Although the algorithm may be applied most directly to the Euclidean plane similar algorithms may also be applied to higher-dimensional spaces or to spaces with other non-Euclidean metrics. Lloyd's algorithm can be used to construct close approximations to centroidal Voronoi tessellations of the input, which can be used for quantization, dithering and stippling. Other applications of Lloyd's algorithm include smoothing of triangle meshes in the finite element method. This Lloyd`s algorithm is used efficiently for the process of forming clusters. The below figure explains how the clustering process takes place along with finding the centroid.



3.1.1 Finding the centroid

This section indicates the next step of the process. For efficient clustering using Lloyd`s algorithm, there is a need to find the average mean distance among all the inputs. There are 3 steps in finding the centroid.

1. To find the average mean distance, first the entries in the table were divided into groups.
2. Then the average mean distance has been found for both the groups they divided as.
3. The Earth mover distance formula has been used to find the average mean. Then the clustering process is done.

3.1.2 Earth mover distance formula

Earth mover's distance (EMD) is a measure of the distance between two probability distributions over a region *D*. In mathematics, this is known as the Wasserstein metric. Informally, if the distributions are interpreted as two different ways of piling up a certain amount of dirt over the region *D*, the EMD is the minimum cost of turning one pile into the other; where the cost is assumed to be amount of dirt moved times the distance by which it is moved.

The above definition is valid only if the two distributions have the same integral (informally, if the two piles have the same amount of dirt), as in normalized histogram or probability density functions. In that case, the EMD is equivalent to the 1st Mallows distance or 1st Wasserstein distance between the two distributions.

If the domain *D* is discrete, the EMD can be computed by solving an instance transportation problem, which can be solved by the so-called Hungarian algorithm. In particular, if *D* is a one-dimensional array of "bins" the EMD can be efficiently computed by scanning the array and keeping track of how much dirt needs to be transported between consecutive bins. For example:

$$EMD_0 = 0$$
$$EMD_{i+1} = (Ai + EMD_i) - B_i$$
$$TotalDistance = \sum | EMD_i |$$

An early application of the EMD in computer science was to compare two grayscale images that may differ due to dithering, blurring, or local deformations. In this case, the region is the image's domain, and the total amount of light (or ink) is the "dirt" to be rearranged.

The EMD is widely used in content-based image retrieval to compute distances between the color histograms of two digital images. In this case, the region is the RGB color cube, and each image pixel is a parcel of "dirt". The same technique can be used for any other quantitative pixel attribute, such as luminance, gradient, apparent motion in a video frame, etc.

More generally, the EMD is used in pattern recognition to compare generic summaries or surrogates of data records called signatures. A typical signature consists of list of pairs $((x_1,m_1), \ldots (x_n,m_n))$, where each x_i is a certain "feature" (e.g., color in an image, letter in a text, etc.), and m_i is "mass" (how many times that feature occurs the record). Alternatively, x_i may be the centroid of a data cluster, and m_i the number of entities in that cluster. To compare two such signatures with the EMD, one must define a distance between features, which is interpreted as the cost of turning a unit mass of one feature into a unit mass of the other. The EMD between two signatures is then the minimum cost of turning one of them into the other.

In our paper clustering reappears and we chose specific clustering algorithm for clustering that's based on centroids. In the chosen algorithm the centroids will be changing often. Centroids can be chosen based on the closest mean of the entries of the table. The closest mean will be the minimum one. This will be found with the help of Earth mover distance formula.

The main Concept introduces a technique of clustering efficiently based on centroids for inputs in the kind of tables.

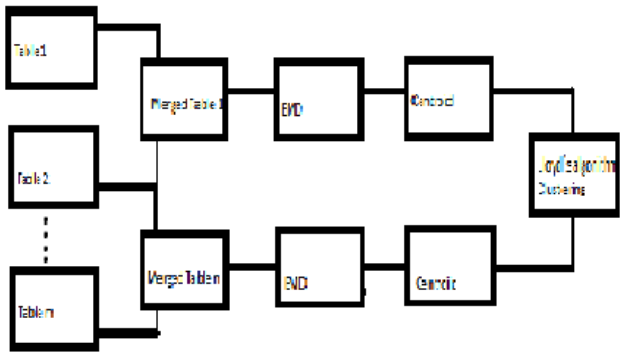


Figure 1: System architecture

The systems architect establishes the basic structure of the system, defining the essential core design features and elements that provide the framework. The system architecture provides the whole structure of the work starts with the input, which was n number of tables. The tables were merged at the first step. We will get n number of merged tables. Then from this merged tables we find the average mean distance using the formula Earth mover. An average mean has been found based in which the clustering has been done using the algorithm Lloyd`s. The clustering has been done efficiently based on the centroids. Some of the following points should be considered in order to efficient clustering.

1. The table should contain a limited number of entries.
2. The entries in the table must be either continuous or discrete.
3. Merge the tables in any one of the criteria which makes the process simple.
4. Use the EMD formula in order to get a closest mean.
5. The closest mean obtained from the EMD formula must be the minimum distance among all the entries in the table.

4. Framework Modules

The framework modules explain the modules that come under our proposed work. These modules explain from step by step how the cluster forms. The process of creating table, comparing and merging the tables, finding the average distance mean and forming the cluster are some of the steps include in our paper.

Merging of Tables: Here the tables were created and the entries are given. The entries that are given in the table must be either discrete or continuous. In case of numeric inputs, the input must be considered as continuous as well as discrete in means of ranges. In case of alphabetic inputs their type was considered as discrete. Then the entries are compared in any of the criteria and they are merged.

Figure 2: Merging of Tables

Centroid computation by AMD: Here the centroid of the cluster is computed by finding the average mean distance. This average mean distance was calculated by means of a formula called Earth mover. This mean will be an intermediate number that will be easy for finding the cluster in the last stage.

$$EMD_0 = 0$$
$$EMD_{i+1} = (A_i + EMD_i) - B_i$$
$$TotalDistance = \sum | EMD_i$$

Clustering based on centroids: Here clustering is done based on centroids. The Cluster process has been upgraded with the technique called Lloyd`s algorithm. Here the centroid never remains in the same point. The centroid will move according to the closest mean found by the EMD.

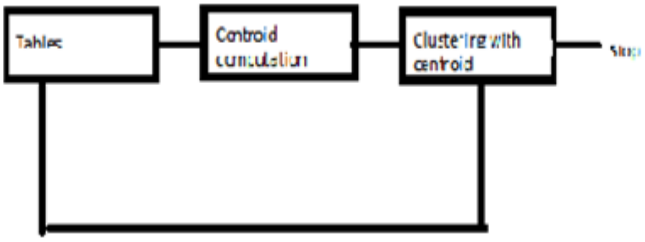


Figure 3: Clustering

Performance analysis: After the clustering has done, the performance is analyzed by comparing the existing system. Here we are using two basic measures namely precision and recall. These measures will be too less on giving continuous inputs.

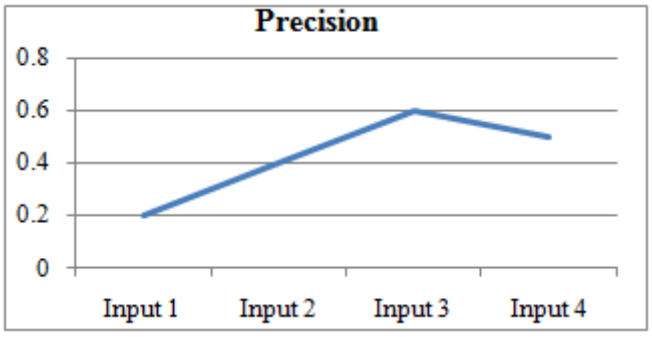


Figure 4: Performance using precision

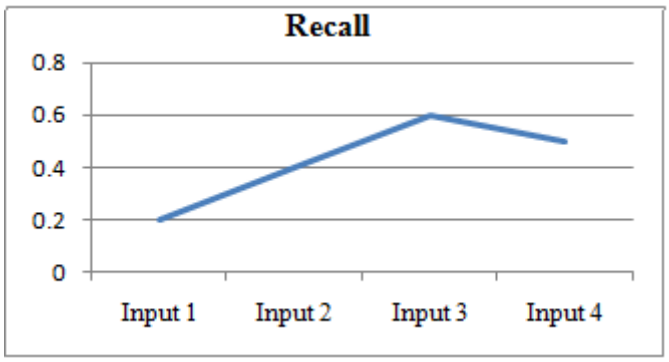
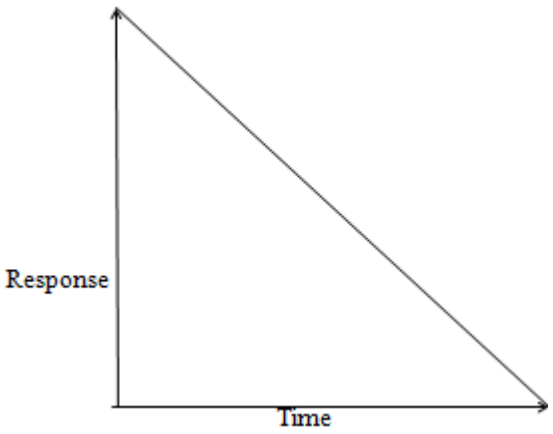


Figure 5: Performance using recall

After analyzing the performance using the measures precision and recall, the response time is measured. The response time will be very less indicating our clustering algorithm which is based on centroids is very efficient.



According to the graph we obtain for the response time, we will get large response in less time. As time increases the response get lower. As the input was given in the form of tables, the clustering process takes place efficiently and we find clustering works efficiently which is based in clustering.

5. Future Work

Accordingly, we have concluded that when input is given in the form of tables clustering process works efficiently. We derived a clustering algorithm that is based on centroids which is working efficiently. In future, we can try by giving another input.

6. Conclusion

This paper presented a novel algorithm for extracting actionable clusters using Lloyd`s clustering algorithms. The clusters are of different size and different density. The proposed algorithm used one additional centroid, the distance measurement depends on the density of data objects from all clusters mean. Also this algorithm uses another measure to find the average mean distance using EMD. These experimental results demonstrated that our scheme could do better than the traditional K-means algorithm. While our proposed algorithm solve the problems when clusters are of differing Sizes and Densities, the traditional K-means failed.

References

[1] Macqueen J. B., “Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability”. Berkeley, University of California Press, 1:281-297, 1967.

[2] Johnson S. C., “Hierarchical Clustering Schemes”. Psychometrika, 2:241-254, 1967.

[3] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. 1996. “From data mining to knowledge discovery: an overview.” in U. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthursamy, eds, ‘Advances in Knowledge Discovery and Data Mining’, AAAI-Press, pp.1–34.

[4] Longbing Cao, 2012, Domain-Driven Data Mining: Challenges and Prospects, IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No. 6,pp.755-769.

[5] S. M. Savaresi and D. Boley. 2004, ‘A comparative analysis on the bisecting K-means and the PDDP clustering algorithms’. Intelligent Data Analysis , 8(4), pp.345–362.

[6] G. Hamerly and C. Elkan. Alternatives to the k-means algorithm that find better clusterings. In Proc. of the 11th Intl. Conf. on Information and Knowledge Management, pp:600–607, McLean, Virginia, 2002. ACM Press

[7] Geng, L., Hamilton .H .J. 2006. “*Interestingness Measures for Data Mining: A Survey*”, ACM Computing Surveys, Vol. 38, No. 3, Article.9, pp.1-32.

[8] Tapas Kanungo,2002 “An Efficient k-means clustering algorithm analysis and implementation”.

[9] P.S. Bradley and U. Fayyad, ªRefining Initial Points for K-means Clustering,º Proc. 15th Int'l Conf. Machine Learning, pp. 91-99, 1998.

[10] P.S. Bradley, U. Fayyad, and C. Reina, ªScaling Clustering Algorithms to Large Databases,º Proc. Fourth Int'l Conf. Knowledge Discovery and Data Mining, pp. 9-15, 1998.
[11] V. Capoyleas, G. Rote, and G. Woeginger, ªGeometric Clusterings,º J. Algorithms, vol. 12, pp. 341-356, 1991.
[12] J.M. Coggins and A.K. Jain, ªA Spatial Filtering Approach to Texture Analysis,º Pattern Recognition Letters, vol. 3, pp. 195-203,1985.
[13] S. Dasgupta, ªLearning Mixtures of Gaussians, ºProc. 40th IEEE Symp. Foundations of Computer Science, pp. 634-644, Oct. 1999.
[14] S. Dasgupta and L.J. Shulman, ªA Two-Round Variant of EM for Gaussian Mixtures,º Proc. 16th Conf. Uncertainty in Artificial Intelligence (UAI-2000), pp. 152-159, June 2000.
[15] Q. Du, V. Faber, and M. Gunzburger, ªCentroidal Voronoi Tesselations: Applications and Algorithms,º SIAM Rev., vol. 41,pp. 637-676, 1999.
[16] R.O. Duda and P.E. Hart, Pattern Classification and Scene Analysis. New York: John Wiley & Sons, 1973.
[17] M. Ester, H. Kriegel, and X. Xu, ªA Database Interface for Clustering in Large Spatial Databases,º Proc. First Int'l Conf. Knowledge Discovery and Data Mining (KDD-95), pp. 94-99, 1995.