ISSN:0975-9646

Centroid Based Clustering Algorithms- A Clarion Study

Santosh Kumar Uppada

PYDHA College of Engineering, JNTU-Kakinada
Visakhapatnam, India

Abstract— **The main motto of data mining techniques is to generate user-centric reports basing on the business requirements. The advent increase in thirst for knowledge discovery has increased the need of robust algorithms in the process of knowledge discovery. Mining in general is termed as intrinsic methodology of discovering interesting, formerly unknown data patterns. Clustering can be termed as a set-grouping task where similar objects are being grouped together. Clustering, a primitive anthropological method is the vital method in exploratory data mining for statistical data analysis, machine learning, and image analysis and in many other predominant branches of supervised and unsupervised learning. Cluster algorithms can be categorized based on the cluster models available depending on type of data we try to analyse. This paper focuses on different centroid based algorithms on par numerical and categorical data.**

Keywords— **Categorical data, k-means, k-medoids, CLARA, CLARANS, Euclidian, Manhattan, Minkowski, fuzzy, mushroom data, harmonic mean.**

I. INTRODUCTION

Clustering has a very prominent role in the process of report generation [1]. It is treated as a vital methodology in discovery of data distribution and underlying patterns. Underlying aspect of any clustering algorithm is to determine both dense and sparse regions of data regions. The main emphasis is on the type of data taken and the efficiency at which we can cluster algorithms with high accuracy and low I/O costs which generally hinges on the similarity or distance metrics. Clustering algorithms in general is a blended of basic hierarchical and partitioning based cluster formations [3]. In general cluster algorithms diversify from each other on par of abilities in handling different types of attributes, numerical and categorical data, and accuracy percentage and in handling of disk-non-migratory data [4].

Clustering algorithms in addition to aspect of handling numerical data, are forced to use combination of text and numerical data termed as categorical data [2]. Example of such need is the "MUSHROOM" data of popular UCI machine-learning repositories. For such data, it is very hectic to determine order or to quantify the dissimilarity factor as data here each tuple is a set of grilled mushrooms which is maintained using many categorical attributes.

Classification which is termed as a sophisticated mechanism of generalizing known structure to apply to upcoming data, can be treated as backbone for the kick-off of clustering. The basic difference between pure classification and clustering is that the classifications is a supervised learning process while the former is an unsupervised method of learning process.

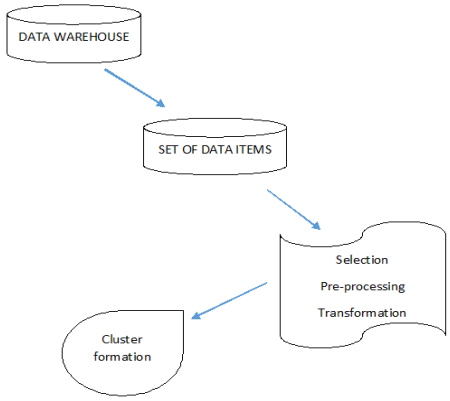

Figure 1. Cluster formation mechanism

Centroid based algorithm represents all of its objects on par of central vectors which need not be a part of the dataset taken. In any of the centroid based algorithms, main underlying theme is the aspect of calculating the distance measure [6] between the objects of the data set considered. The basic aspect of distance measure in general is derived using one of Euclidian, Minkowski or Manhattan distance measuring mechanism [5].

In general, mean is used in Euclidian distance measure, median in Manhattan and steepest descend method for calculating the distance measures [7].

Distance measure methodologies:

a. *Euclidian distance measure*:

Euclidian distance becomes a metric space, being processed using Pythagorean formula. The position in a Euclidian n-space is termed as Euclidian vector. The Euclidian measure is generally given by

$$d(p,q) = d(q,p) = \sqrt{(q_i - p_i)^2}$$

In general Euclidian distance is being used in nonlinear dimensionality measure. In general correlation analysis is been calculated using this metric. The main drawback of this measure is that it is sensitive to high noise and sensitive in determining correlation between similar trends [8].

b. Manhattan distance measure

Manhattan distance measure is a typical method that could be adapted even if a grid-like path is being traced in

the data sets [9]. It is typically the distance measure between corresponding correlated objects. The measure in general is given by

$$d = \sum_{i=1}^{n} |x_i - y_i|$$

In general Manhattan distance is a heuristic based algorithm which could determine distance metric for unevenly distributed objects. The major degrading aspect is that it is still sensitive in measuring correlation dissimilarity between similar trends.

c. Minkowski distance measure

In general Minkowski measure on Euclidian space is regarded as a generalization of both Euclidian and Manhattan distance measures. It can be treated as a power mean multiple of distance between objects. The Minkowski measure between two objects with order p is given by

$$d = \left(\sum_{i=1}^{n} | x_i - y_i|^p\right)^{1/p}$$

In typical cases, Minkowski distance measure suffers with limiting factor value being reached to infinity [10]. In such cases a new measure termed as Chebyshev measure is being calculated.

Here the following section examines some of the most predominant methods of centroid based algorithms. Section II deals about the primitive k-means algorithm, III defines the aspects of k-medoids, IV deals with the need of using CLARA algorithm and V defines a newer approach of CLARA termed using randomized search techniques termed as CLARANS, VI provides the aspects of dealing with K-Harmonic means, VII regarding the aspect of using fuzzy c-means algorithms. VIII summarizes all the clustering algorithms we have taken with tabulation of different aspects that are to be considered.

II. K-MEANS

The k-means algorithm deals with the process of defining clusters on par of centre of gravity of the cluster. It is a primitive algorithm for vector quantization originated from signal processing aspects. K-means algorithm is very widely used in pattern recognition, classifications of documents and image processing etc.

This approach starts with assigning some objects as pivot objects. The value of k which determines the number of clusters that one wish to have is given before-hand. Now the centroid of each such points are been calculated. Each data points are then assigned to a cluster whose centroid is the nearest possible. Arithmetic mean is being calculated separately on par of its dimensionality [11].

K-means algorithm can be treated as two-phase approach where in the first phase; k centroids are being identified depending on the k value that has been chosen in general the distance measure is being calculated using Euclidian distance metric. The second phase involves in determining the new centroids for which the dissimilarity measures are very less. The process loops in finding new centroids until the property of convergence is met [12].

Algorithm:

Input: $X = \{x_1, x2 \ldots x_n\}$, where n being a set of objects

K; Number of desired clusters

Output: A set of *k* clusters.

Steps:

1. Randomly choose k objects from X as primary centroids.
2. Repeat

2.1 Assign each data item *di* to the cluster which has the closest centroid;

2.2 Calculate the new mean of each cluster;

Until convergence criterion is met.

Advantages:

1. It is termed as the speediest centroid based algorithm.
2. It is very lucid and can sustain for large amount of data sets.
3. It reduces intra-cluster variance measure.

Disadvantages:

1. It suffers when there is more noise in the data.
2. Outliers can never be studied
3. Even though it reduces intra-cluster variance, it could not deal with global minimum variance of measure.
4. Very sensitive at clustering data sets of non-convex shapes.

III. K-MEDOIDS

In the method of clustering using k-medoid, each cluster is represented by nearest object towards centre. K-medoids chooses data points as centres which are also termed as exemplars. Here the formed clusters are termed as priori. Silhouette method is generally used in order to determine the value of k.

Process deals with applying the improved combination of k-medoids and 'Partitioning around Medoids' (PAM) algorithm on the data retrieved.

If $A_j$is a non-selected object; Ai being a medoid, then it is obvious that we take Aj as element belonging to the cluster A_i only if d (Aj, Ai) = Min A_ed (A_j, A_e) by aiming at considering minimum of the all medoids derived [12].

It should be noted that $A_e$and d (A_a, A_b) determines the distance or dissimilarity between the objects taken into account. The average of the dissimilarity that has been derived generally improves the quality of the clustering.

Let us assume A_1, A2…$A_k$are the k-medoids that has been selected at any taken stage. Let C_1, C2…$C_k$are their respective clusters.

For a non-selected object A_j, where j≠1, 2…k;

If A_j€ A_m then Min $_{(1\leq i\leq k)}$ d (A_j, A_i) = d (A_j, A_m)

If an unselected object A_h becomes a medoid replacing eAisting medoid A_i. The new set ofmedoid then will be

$K_{med'}$= {A_1, A_2...A_{i-1}, A_h, A_{i+1}…A_k}

Note that here A_i is been replaced by the new medoid formed. For the same measure to occur, let us take $C_h$to be a new cluster with A_h as its representative.

There generally raises three different cases here, each of which can defined as

For a non-selected object,

1. A_j€ C_i before swapping and A_j€ $C_h$when,

Min d (A_j, A_e) = d (A_j, A_i), minimum taken for all $A_e$in K_{med}Min$_{e\neq i}$d (A_j, A_e) = d (A_j, A_h), minimum taken for all A_e in $K_{med'}$

Here the cost for such change is given by

$$C_{jih} = d(A_j, A_h) - d(A_j, A_i)$$

2. A_j€ C_i before swapping and A_j€ C_g and g ≠ hwhen,

Min d (A_j, A_e) = d (A_j, A_i), minimum taken for all A_e inK_{med}Min$_{e\neq i}$d (A_j, A_e) = d (A_j, A_g); g ≠ h, minimum taken for all A_e in K_{med},Here the cost for such change is given by C_{jih} = d (A_j, A_g) – d (A_j, A_i)

3. A_j€ C_g before swapping and A_j€ $C_h$when,

Min d (A_j, A_e) = d (A_j, A_i), minimum taken for all $A_e$in K_{med}Min$_{e\neq i}$d (A_j, A_e) = d (A_j, A_h), minimum taken for all A_e in K_{med},

Here the cost for such change is given by

$$C_{jih} = d(A_j, A_h) - d(A_j, A_g)$$

PAM algorithm:

Input: database of objects S.

Choose k representative objects at random and term the as K_{med}

Mark selected objects for future reference

For all selected objects A_i

For all non-selected objects A_h

Compute C_{ih}

Now, select minimums of I and h to derive Min$_{i,h}C_{ih}$

If the minimum measures is less than zero,

Swap A_i and A_h to be non-selected and selected

Find clusters further up to the value of k taken before-hand.

The total cost of swapping two medoids is the sum of costs for all non-selected objects, which are being calculated as per the above cases [13].

Advantages:

1. K-medoids algorithm is very robust to noisy data.
2. Outliers can be studied i.e. it is not sensitive to outliers.
3. Pairwise dissimilarity measure comes into picture in case of squared Euclidian distance measures.

Disadvantages:

1. Different initial set of medoids effect the shape and effectiveness of the final cluster.
2. Clustering depends on the units of measurement, difference in nature of objects differ the efficiency.
3. Sensitive at clustering non-convex shaped clusters.

IV. CLARA

CLARA stands for clustering large applications and is been given by Kauffman and Rousseau in 1990. CLARA in generally used in reducing the computational efforts that one come across using k-medoid algorithm [14]. In contrast to the basic methodologies of finding the object representatives for the entire dataset, CLARA attempts in measuring the same on sample data [17]. Once the process of identifying the key objects is done, PAM algorithm is applied on the sample on par of deriving optimal set of medoids. In case the samples are drawn at random of best level, the medoids thus formed would represent for the whole data set [16].

The cost of dissimilarity is being given by

$$Cost(M, D) = \frac{\sum_{i=1}^{n} dissimilarity(O_i, rep(M, O_i))}{n}$$

Algorithm:

Input: Database with O objects

Repeat for t times

Draw sample S, being a subset of O at random

Derive PAM(S, k) where k is the predefined medoid number

Classify the entire dataset O into k segments

Calculate the average dissimilarity rate to choose the best medoid at each clustering step.

Advantages:

1. CLARA can easily handle noisy data.
2. Deals with larger data sets than PAM algorithm.
3. Sampling helps in improving computational speed and efficiency.

Disadvantages:

1. Efficiency is effected by the sample size.
2. For certain data objects, biasing of samples causes degradation of cluster efficiency.

V. CLARANS

CLARANS stands for clustering large applications based on randomized search. CLARANS at times is termed as an enhanced version of primitive CLARA and thus termed as Randomized "CLARA". CLARANS is similar to PAM and CLARA with modification that a randomized iterative optimization is being applied [14]. CLARANS does not have restrictions on search terms as of CLARA for any subset of objects.

The basic CLARANS method starts with PAM and selects few pairs say (i, h) instead of working on the whole dataset [15]. Like PAM, it starts with randomly selected medoids and then checks for minimal dissimilarity measure i.e. it looks for medoid which is very nearer, termed as "Max-Neighbour" pair for swapping. If the cost is derived to be negative, it just updates the medoid set and process continues. The process comes to an end after reaching optimal medoid set termed as "Num-Local" is achieved.

Algorithm:

Input: Data set as O, K being value of clusters needed, M- max-neighbour value, L as num-local value.

Select k representatives at random.

Mark all selected and non-selected sets of objects and call it as current

Set x=1

Repeat while x is less than or equal to L

Set y=1

For all m ≤ M

Select a pair (i, j) such that O_j is termed as selected and O_h termed as non-selected, compute the cost as C_{ij}.

If the cost factor is negative, update current

Else increment m by 1

Now compute cost of clustering with min-cost

If current-cost < min-cost

Mark the new value as min-cost and mark current node as best-node

Increment value of x

Save and return the best-node.

Advantages:

1. Termed as best than many medoid-based algorithms like k-medoids, PAM and CLARA.
2. CLARANS is more flexible, efficient and scalable.
3. It well defines major aspects of outliers
4. Robust to noisy data.

Disadvantages:

1. CLARANS assumes every object of entire data set to fit into main memory and hence very sensitive to input order.
2. The trimming aspect of "Max-neighbour" driven searching, degrades the efficiency of finding a real local minimum value termed as "Loc-Min".
3. R^* trees are now being used to minimize the limitations of CLARANS.

VI. GENERALIZED K-HARMONIC MEANS

K-Harmonic means has a built-in dynamic weighting function which tries in boosting the data objects which are not nearer in the current iteration. In this approach, weighting function is automatically changed after each iteration. The K-Harmonic means is insensitive to the data set's initialization. First the initializations are made then the convergence factor is being calculated which enhances the efficiency [18]. Harmonic value will always tend to be minimum and thus behaves as min () function rather than as avg () function.

K-Harmonic means are generally been used in many real-world datasets. The dynamic weighting approach here makes the algorithm being build robust to inefficient initializations and noise in the data sets.

Input:

$X = \{x_1, x2 \ldots x_n\}$, where n being a set of objects

K; Number of desired clusters

Output:

A set of *k* clusters.

Steps:

1. Choose k objects from X as primary centroids at random.
2. Repeat

2.1 Assign each data item *di* to the cluster which has the closest centroid;

2.2 Calculate the new harmonic mean of each cluster;

Until convergence criterion is met.

The K-Harmonic mean is given by

$KHM(X, C) = \sum_{i=1}^{n} k/P$ where P is given by

$$P = \sum_{j=1}^{h} ||x_i - c_j||^{-p}$$

Advantages:

1. KHM allows us in deriving high quality clusters with low dimensionality.
2. Robust to noisy data.
3. Very scalable and efficient algorithm for trendy data sets.

Disadvantages:

1. Process involves computational complexity.
2. Sensitive to outliers.
3. Sensitive in clustering efficiently for non-polygon structure.

VII. FUZZY C-MEANS

In these types of clustering algorithms, allocation of data points to clusters is "Fuzzy" rather than being hard. Therefore the fuzzy clustering is also termed as "Soft clustering". Fuzzy c-means (FCM) is a typical clustering algorithm that allows certain data points to reside in one or more clusters. This clustering algorithm is being effectively used in pattern recognition [21]. The clustering technique relies on minimization of objective function.

Any point *x* has a set of coefficients giving the degree of being in the *k*th cluster $w_k(x)$. With fuzzy *c*-means, the centroid of a cluster is the mean of all points, weighted by their degree of belonging to the cluster:

The clustering algorithm in case of fuzzy c- means has its centroid being the mean of all objects weighted by the degree of belongingness to a specific cluster [19].

For any point x belonging to cluster k, w_k (x) is the degree of closeness which is inverse to distance from cluster's centre. The factor m which gives the closest centre's weight also have a significant mark. Fuzzy C-means algorithm is identical to K-means algorithm at remarkable level.

$$c_k = \frac{\sum_x w_k(x)^m x}{\sum_x w_k(x)^m}.$$

Algorithm:

Choose cluster count value.

Assign coefficients with values for which reside in the clusters.

Repeat until convergence level is achieved

Compute centroid of each cluster

Come coefficients of belonging to cluster.

Advantages:

1. Minimizes intra-cluster variance.
2. Data points are flexible to reside in more than one cluster.
3. Efficient algorithm in avoiding hard coding of data point values.
4. Very effective in pattern recognition.
5. Objective function minimization is achieved.

Disadvantages:

1. The local minimum is directly taken as minimum directly which is problematic for huge data objects.
2. Results depends on the initial choice of weights.
3. Sensitive to noisy data.
4. Very sensitive at clustering data sets of non-convex shapes.

Table 1. Comparative aspects of major centroid based algorithms

S.No	Sensitive to noise	Outlier	Structure-centric	Minimize Intra-cluster variance	complexity
k-means	Very high	Very Sensitive	Yes	No	$O(n^{dk+1} \log n)$
k-medoids	Optimum	sensitive	Yes	No	$O(k(n-k)^2)$
CLARA	Optimum	Kick-off to study	No	Yes	$O(ks^2 + k(n-k))$
CLARANS	Very low	Deals with outliers	No	Yes	$O(n^2)$
k-Harmonic means	High	sensitive	Yes	No	$O(n^k \log n)$
Fuzzy c-means	Optimum	Kick-off to study	Yes	No	$O(n^{(k-c)} \log n)$

VIII. CONCLUSION

This paper has a predominant study of basic clustering algorithms, basic types of data that are very much needed, and process of generating reports basing on the user requirements. As discussed it is very much needed to classify the data first which comes under supervised learning technique. Process is then extended in segregating data objects basing on its properties. This novel method of unsupervised learning is also been mentioned.

Clustering as discussed is done on the basis of similarity and dissimilarity measures. In general centroid based algorithms are selected to improve the quality and efficiency of the clustering. In any of the centroid based algorithms, distance measure takes a vital role and for that in addition to the pre-existing metrics like Euclidian, Manhattan and Minkowski, we have a measure termed as Chebyshev distance measure in certain limiting cases.

Algorithms are being discussed on par of primitive numerical data to categorical data and new methods of clustering which allows data points to reside in more than one cluster.

The table 1 here deals with the aspects of comparative study of different algorithms basing on aspects like sensitivity to outliers, noise in data, intra-cluster variance measures, fuzzing coding of data objects and computational complexity [22].

REFERENCES

1. A survey on different issues of different clustering algorithms by M. Vijaya Lakshmi, M. Renuka Devi, www.ijarcsse.com, volume 2, issue 3.
2. S. AnithaElavarasi and Dr. J. Akilandeswari and Dr. B. Sathiyabhama, January 2011, A Survey On Partition Clustering Algorithms
3. S. AnithaElavarasi and Dr. J. Akilandeswari (2011) A Survey On Partition Clustering Algorithms, International Journal of Enterprise Computing and Business Systems.
4. Pavel Berkhin, Survey of Clustering Data Mining Techniques, Accrue Software, Inc.
5. Practical approach towards data mining and its analysis, AnuradhaSrinivas (2013).
6. Bailey, Ken (1994). "Numerical Taxonomy and Cluster Analysis". *Typologies and Taxonomies*. p. 34.ISBN 9780803952591.
7. The choice of metrics for clustering algorithms, Peter Grabusts, Environment. Technology. Resources Proceedings of the 8th International Scientific and Practical Conference. Volume I1 © RēzeknesAugstskola, Rēzekne, RA Izdevniecība, 2011.
8. Aloise, D.; Deshpande, A.; Hansen, P.; Popat, P. (2009). "NP-hardness of Euclidean sum-of-squares clustering". *Machine Learning* **75**: 245–249. doi:10.1007/s10994-009-5103-0.
9. For most large underdetermined systems of linear equations the minimal $\ell 1$-norm solution is also the sparsest solution; See Donoho, David L, Communications on pure and applied mathematics, 59, 797 (2006) http://dx.doi.org/10.1002/cpa.2013.
10. AR Mohazab, SS Plotkin,"Minimal Folding Pathways for Coarse-Grained Biopolymer Fragments" Biophysical Journal, Volume 95, Issue 12, Pages 5496–5507.
11. Hartigan, J. A.; Wong, M. A. (1979). "Algorithm AS 136: A K-Means Clustering Algorithm". *Journal of the Royal Statistical Society, Series C* **28** (1): 100–108. JSTOR 2346830.
12. Kaufman, L. and Rousseeuw, P.J. (1987), Clustering by means of Medoids, in Statistical Data Analysis Based on the $L1$–Norm and Related Methods, edited by Y. Dodge, North-Holland, 405–416.
13. H.S. Park , C.H. Jun, A simple and fast algorithm for K-medoids clustering, Expert Systems with Applications, 36, (2) (2009), 3336–3341.
14. R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan,"Automatic Subspace Clustering of High Dimensional Data forData Mining Applications," Proc. 1998 ACM-SIGMOD, pp. 94–105,1998.
15. CLARANS: A Method for Clustering Objects for Spatial Data Mining, Raymond T. Ng and Jiawei Han, Member, IEEE Computer Society
16. Chih-Ping, Wei, Yen-Hsien, Lee, and Che-Ming, Hsu, Empirical Comparison of Fast Clustering Algorithms for Large Data Sets.
17. The R Development Core Team, R: A Language and Environment for Statistical Computing.
18. K-Harmonic Means-A Data Clustering Algorithm Bin Zhang, Meichun Hsu, UmeshwarDayal Hewlett-Packard Research Laboratory[2010].
19. K-Harmonic Means-A Data Clustering Algorithm Bin Zhang, Meichun Hsu, UmeshwarDayal Hewlett-Packard Research Laboratory[2010].
20. FCM: THE FUZZY c-MEANS CLUSTERING ALGORITHM JAMES C. BEZDEK Mathematics Department, Utah State University, Logan, UT 84322, U.S.A. ROBERT EHRLICH Geology Department, University of South Carolina, Columbia, SC 29208, U.S.A. WILLIAM FULL[2012].
21. J. C. Bezdek (1981): "Pattern Recognition with Fuzzy Objective Function Algoritms", Plenum Press, New York
22. Comparison of Cluster Algorithms for the Analysis of Text Data Using Kolmogorov Complexity byTinaGeweniger, Frank-Michael Schleif, Alexander Hasenfuss, Barbara Hammer, Thomas VillmannAdvances in Neuro-Information Processing, Lecture Notes in Computer Science Volume 5507, 2009, pp 61-69.