Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 2, No. 1, February : 2024

A NOVEL K-MEDIAN CLUSTER ANALYSIS IN DATA MINING

Dr.R.Suresh Kumar, Assistant Professor, Computer Science Department, Sree Saraswathi Thyagaraja College, Pollachi. sureshdhanya2002@yahoo.com

Abstract

There are standard methods in literature in cluster analysis of data mining of quantitative (numeric) and categorical (qualitative nominal, qualitative ordinal and binary) data using any of the distance functions: Euclidian, Manhattan, Chebycheve (for quantitative data) and categorical □2 and modular (for categorical data) like K-mean, K-median, Minimum variance, agglomerative, divisive hierarchical and modular algorithms. K-median cluster analysis is just an algorithm using the Manhattan distance function and in K-mean algorithm Euclidian distance is used but no median and thus the terminology is misleading. K-Phototype cluster analysis is introduced for mixed data by introducing a distance function Sn + □Sc where Sn is the Euclidian distance, Sc is distance in the categorical sense for quantitative and categorical attributes respectively and □ is a weight fixed from nature of the attributes with the same algorithm more or less. K-modes algorithm is introduced by Huange considering only categorical attributes. Modes is taken to be the minimum distance in categorical or □2. K-means cluster analysis is organized converting all the attributes of categorical nature into binary through which the attributes are multiplied like anything and thus the algorithm is complicated.

$$D(X, Y) = \sum_{i=1}^{n} d(X, Y) \quad \text{where } d(X, Y) \text{ is}$$

categorical distance or □2 distance. Minimum of D(X, Y) is mode of the cluster. Just like K-means algorithm K-mode algorithm is formulated. Categorical data only is considered in k-mean algorithm and K-mode algorithm.

Now, in this note, medianal approach is introduced for either quantitative, categorical or mixed data (k-median algorithm). This terminology may be replaced by K-median if the latter is terminologised properly in later course of time. Mathematically if the components(attributes) are sequencable the resultant data (true data) is squencable (example: countability of Nr, r Є N). Thus any data having many attributes is sequencable and thus identification of the median of the data is possible like the classical problem of countability in mathematics. This sequencability also may be operated, in many versions accordingly suitable to the investigation under consideration. One such algorithm is investigated here. Let us take any two particular data xi and xj at random or otherwise, and find the distance of all the others from them. The whole data is bifurcated into three namely in between xi and xj and other sides of xi and xj according as d(xi , x) □ d(xi , xj), d(xi , x) □ d(xi , xj) and d(xj , x) □ d(xi , xj) respectively. The process is repeated for these three sub data and again continued until the whole data is sequenced fully. Just like median, quartile, descile, percentile, … k-ile is defined for the above sequenced data (trile, quartile, pentile, …). Thus the data is clustered into k-clusters. Two criterions are involved here: distance function and technique of sequencing by proper choice of them, better results may be attained. Another innovation is on the sick data in the bulk. This may be properly analyzed and removed and then cluster algorithm is used.

Keywords - cluster, medoids, centriod, similariy, sequence, minimal, k-ile.

1. INTRODUCTION

There are standard methods in literature in cluster analysis of data mining of quantitative (numeric) and categorical (qualitative nominal, qualitative ordinal and binary) data using any of the distance functions. The standard distance functions are

1. Euclidean D(x,y) = (∑(xi - yi)2) ½. This is the distance function mostly used in mathematics as well as applications.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 2, No. 1, February : 2024

2. Manhattan $D(x,y) = \sum |x_i - y_i|$ this is more prominent theoretically.
3. Chebychev $D(x,y) = \text{Max} |x_i - y_i|$. This is usable in applications.

The above three distance functions are used for quantitative data.

4. Categorical

$D(x,y) = (\text{number of } x_i \text{ and } y_i \text{ are different } -y_i) / N$

5. χ^2 $d(x,y) = \sum_{j=1}^{m} \delta(x_j, y_j)$

where $\delta(x_j, y_j) = \begin{cases} 0 & \text{if } (x_j = y_j) \\ 1 & \text{if } (x_j \neq y_j) \end{cases}$

This distance is called dissimilarity distance. The distance χ^2 is defined by the formula

$$d_{\chi^2}(X, Y) = \sum_{j=1}^{m} \frac{(n_{x_j} + n_{y_j})}{n_{x_j} \cdot n_{y_j}} \delta(x_j, y_j)$$

6. Modular A mode of the data is a vector $Q = \{ q_1, q_2, \ldots, q_m \}$ that minimizes

$D(Q, X) = \sum_{i=1}^{n} d(X_i, Q)$ where d is defined by either

dissimilarity or χ^2.

The last three distances are for categorical data.

K-mean algorithm is the very standard method to cluster a data. This is a simple but effective clustering technique. The data is partitioned into K-disjoint clusters. The algorithm may be explained by steps in the following way.

1. Choosing a value for K, the total number of clusters desired to be determined, choose K data points with in the data set. They are called initial cluster centres or cluster seeds.
2. Using Euclidian distance, the distance between every seed and all data points are worked out. Then the distance with the minimum is clustered around that seed. So clusters are some way fixed.
3. Now the mean of each cluster may be found out. They may be used as the cluster center are seed. Operation is repeated. We get clusters in a new form. We have improved the clusters one step ahead.
4. The process is repeated until the process stagnates ie we get the same elements in and after operation. This is the most commonly used algorithm due to Lloyd (1982). K-median, minimum variance, agglomerative, divisive hierarchical and modular algorithms are recent techniques but less practiced. K-median cluster analysis is just an algorithm using the Manhattan distance but no median is used is the speciality. Thus the terminology is misleading. K-Phototype cluster analysis is introduced for mixed data by introducing a distance function Sn + □Sc where Sn is the Euclidian distance and Sc is the distance in the categorical sense for quantitative and categorical attributes respectively and □ is a weight. □ is fixed from nature of the attributes. The K-Phototype algorithm is more or less the same as K-means algorithm. Recently K-modes algorithm is introduced by Huange considering only categorical attributes. Mode is used instead of mean but defined differently. Mode is taken to be the minimum of distances in categorical or □2 senses. The K-means algorithm is improved to categorical data also. K-means cluster analysis is organized converting all the attributes of categorical nature into binary. But due to this attributes are multiplied like anything, thus the algorithm becomes complicated.

$D(X, Y) = \sum_{i=1}^{n} d(X, Y)$ where $d(X, Y)$ is categorical

distance or □2 – distance. Minimum of d(X, Y) is the mode of the cluster. Just like K-means algorithm K-mode algorithm is formulated. Usually categorical data only is considered in K-mode algorithm. We understand now K-mean algorithm is the sum and substance or essence of cluster analysis techniques. We note that medianal approach is not at all used so far in cluster analysis technique.

Now, in this note, medianal approach is introduced for either quantitative, categorical or mixed data (k-median algorithm). This terminology is replaced by K-median if the latter is terminologized properly in latter course of time. We found a misleading use of terminology of median earlier.

Mathematically if components (attributes) are sequencable, the resultant data (the true data) is sequencable (example : countability Nr , r Є N) we know that cross product of countable set is countable as N2 is countable. Thus any data having many attributes is sequencable. Thus identification of median of data is possible, like the classical problem of countability in mathematics the sequencability of the large data may be explained thus. The sequencability may be operated in many versions accordingly suitable to the investigation under consideration: one such algorithm is investigated here.

Let us take any two particular xi and xj at random or otherwise and find the distance of all the other data from them. The distance bifurcate the data into three as

1. d(xi , x) □ d(xi , xj) and d(xj , x) □ d(xi , xj)
2. d(xi , x) □ d(xi , xj)
3. d(xj , x) □ d(xi , xj) respectively,

but not satisfy in two conditions simultaneously. If cases 2 and 3 satisfy together, take the minimum and allow x to include in that minimum class. For example, if d(xi , xj) =10, d(xi , x)=13 and d(xj , x)=19, x Є {x'/d(xi, x') > d(xi, xj)}. This is a partition of the data. This is unambiguous. This is due to triangular in equality among axioms of distance.

The process is repeated for these three sub data and again continued until the whole data is sequenced fully in the sense that the bifurcated classes are singleton.

Just like median, quartile,decile, percentile, …, k-ile is defined for the above sequenced data (median, trile, quartile, pentile …). This k-ile is the boundary of clusters in the sequenced data. Thus the data is clustered into k-clusters. Two criterions are involved here: distance function and technique of sequencing. By proper choice of them, better results may be attained. The same algorithm is used for mixed data. Identifying the distance function as the sum of distances in quantitative attributes and categorical attributes, a weightage may be given somewhere. Usually clusters are defined to be centered at centroids. In our terminology we use medoid. Medoid is identified by finding 2k-ile here, the odd-iles are medoids while even-iles are the boundary of the clusters in the sequenced data. Thus the clusters and medoids (centroids) are known.

Another innovation is on the sick data in the bulk. On probability grounds, some entries of the data may be thrown out from the data. Such entries are called sick data. Here also, in mining, this concept may be introduced and investigated. It is desirable to exclude the sick data from the clusters. Class, group and cluster are used synonymsly in data mining.

2. PRE-REQUISITES

In Cluster analysis classification of data characteristically, clustering reveals pattern, trend or characteristics of data. It analyses similarity and dissimilarity among entries. Similar entries according to assigned characteristic are in the same cluster and dissimilar entries differently. By this version, the data is partitioned into classes. The classes are always disjoint. Thus mathematically, there is a hidden equivalence relation in this clustering. In this sense, the clustering reveals pattern. But the cluster expert need not know the pattern or relation behind the clustering. But he could manage clustering.

Median is known from the formula

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 2, No. 1, February : 2024

$M = l + \frac{N/2 - m}{f}$ .c with usual terminology for a classified data, likewise quartile is $Qr = l + \frac{rN/4 - m}{f}$ .c

But for raw data that entry which comes in the middle when data is arranged in ascending or descending order if the median and likewise for quartiles and so on. This is for a frequency distribution or raw data. While attributes are many or in bulk, this is not possible immediately, but by the sequencing technique this may be done. A data with n-attributes may be expressed as n-tupled set.

3. MEDIAN FOR A DATA WITH MANY ATTRIBUTES

We consider a bulk data for cluster analysis. Though many devices are evolved so far particularly using the K-mean algorithm but the median is not considered. This is due to the difficulty in identifying the median for a data with many attributes. We have over come the difficulty here. Let the data fully categorical. Now we define a distance function for these data.

D(X, Y) = sum of the dissimilarity between X and Y.

$$\text{Thus } d(X, Y) = \sum_{i=1}^{n} \delta(x_i, y_i) \quad \text{where}$$

$$\delta(x_i, y_i) = \begin{cases} 0 & \text{if } (x_i = y_i) \\ 1 & \text{if } (x_i \neq y_i) \end{cases}$$

There are n attributes. So any observation may be expressed as n-tuples. Thus the distance between any two observations is noted.

Now let us take at random or otherwise two observations xi, xj from the data and calculate the distances of every observation from both of them. These come in three categories:

1. d(x, xi) □ d(xi , xj) and d(x,xj) □ d(xi , xj)
2. d(x, xi) □ d(xi , xj)
3. d(x, xj) □ d(xi , xj) respectively. But not satisfying two conditions simultaneously. If they satisfy two conditions together, the minimality condition is used to find the clusters as explained in introduction.

	Sex Male/ Female	**Marital Status Married/ Unmarried**	**Education Educated/ Uneducated**	**Economic Status Rich / Poor**	**Employment Employed/ Unemployed**
1	M	M	L	R	E
2	F	M	L	P	E
3	M	U	L	P	U
4	M	U	I	P	U
5	F	U	L	R	E
6	M	U	I	P	E

7	F	M	L	P	U
8	F	M	I	P	E
9	F	U	I	R	E
10	M	M	L	R	E
11	F	U	L	P	E
12	M	U	L	R	U
13	F	U	L	P	E
14	F	M	I	P	U
15	F	U	L	P	U

Considering these three compartments of data as the whole data the process may be repeated. We get 9 minor compartments in the next processing the compartments are 27 in number and so on. Finally, we get each compartment contain atmost one. Thus data is sequenced.

Example : Consider the data in following table. The following facts may be noted

1. D(1, 10) = 0 + 0 + 0 + 0 + 0 = 0
2. D(2, 8) = 0 + 0 + 1 + 0 + 0 =1
3. D(2, 5) = 0 + 1 + 0 +1 + 0 = 2
4. D(2, 6) = 1 + 1 + 1 + 0 + 0 = 3
5. D(1, 4) = 0 + 1 + 1 + 1 + 1 = 4
6. D(8, 12) = 1 + 1 + 1 + 1 + 1 = 5

The distances between all the points are noted in the TABLE 1.

TABLE 1

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	10	11	12	13	14	15
1	0	2	3	4	2	3	3	3	3	0	3	2	3	4	4
2		0	3	4	2	3	1	1	3	2	1	4	1	2	2
3			0	1	2	2	2	4	4	3	2	1	2	3	1
4				0	4	1	3	3	3	4	3	2	3	2	2
5					0	3	3	3	1	2	1	2	1	4	2
6						0	4	2	2	3	2	3	2	2	3
7							0	2	4	3	2	3	2	1	1
8								0	2	3	2	5	2	1	3
9									0	3	2	3	2	3	3
10										0	3	2	3	4	4

11										0	3	0	3	1
12											0	3	4	2
13												0	3	1
14													0	2
15														0

There are only 6 possibilities. Now the data is equipped with a distance function but the data is not sequenced. But using distance function we can sequence the data.

Let us consider the following data

TABLE II

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	0	2	3	4	2	3	3	3	3	0	3	2
2		0	3	4	2	3	1	1	3	2	1	4
3			0	1	2	2	2	4	4	3	2	1
4				0	4	1	3	3	3	4	3	2
5					0	3	3	3	1	2	1	2
6						0	4	2	2	3	2	3
7							0	2	4	3	2	3
8								0	2	3	2	5
9									0	3	2	3
10										0	3	2
11											0	3
12												0

The one which comes in the middle is the median. But here we get two middle ones. In this case, the median is not uniquely fixed.

4. k-MEDIAN ALGORITHM

Like median, quartile, decile and percentile, k-ile may defined.

$$k_i = l + \frac{i.\,N/K - m}{f} .c$$

for a classified data with usual terminology. We have sequenced out data. Thus the data is raw.
k1 = first k-ile is the N/Kth entry in the sequenced data.

k2 = second k-ile is the 2N/Kth entry in the sequenced data.
k3 = third k-ile is the 3N/Kth entry in the sequenced data. And so on
kr = rth k-ile is the rN/Kth entry in the sequenced data.

Thus k-iles are found out already. The k-iles are the boundaries of the clusters. There are k-clusters. In the above example the clusters and their formulation are given from the tables. In the first example,
S1= {14,15,4} S2={3,7,11,13,6,5,2,1,10,12} S3={8,9}

S12={14,15,4} S21={6,3,13,5,12} S22={11,2} S23={7,1,10} S32={8,9}
The sequenced data is <14,15,4,6,3,12,5,11,13,2,7,1,10,8,9>
Identifying the median in certain cases is difficult. Fortunately in the above example N=15, so 8th entry is median in the sequence, 5th entry and 10th entry are boundaries of trile. There is no difficult. This median is 11 and boundaries of trile are 3 and 2. So clusters are {14,15,4,6,3}, {12,5,11,13,2} , {7,1,10,8,9}
In the second example,
S1= <4> S2=<3,7,11,6,5,2,1,10,12> S3=<8,9>
S21= <6,3,5,12> S22=<11,2> S23=<7,1,10> S32=<8,9>
The sequenced data is <4,6,3,12,5,11,2,7,1,10,8,9>

	Sex Male/ Female	**Martial Status Married/ Unmarried**	**Education Educated/ Uneducated**	**Economic Status Rich/ Poor**	**Employment Employed/ Unemployed**
1	M	M	L	R	E
2	F	M	L	P	E
3	M	U	L	P	U
4	M	U	I	P	U
5	F	U	L	R	E
6	M	U	I	P	E
7	F	M	L	P	U
8	F	M	I	P	E
9	F	U	I	R	E
10	M	M	L	R	E
11	F	U	L	P	E
12	M	U	L	R	U

5. MEDOID

In clustering theories, the data is partitioned into k-clusters. Which are disjoint among themselves. The most popular method is K-means algorithm. All the other algorithms are some way indebted to K-means algorithm. Here we have introduced k-median algorithm instead to K-means algorithm. Here we have introduced k-median algorithm instead. This method is fully independent of earlier attempts. While k-median algorithm is utilised to cluster a data, and the centre of the cluster (centroid) is identified. Here also, the centre is essential, but it is called medoid of the cluster. In section IV, we found k-iles to cluster the data into k-clusters. Now we find 2k-iles then we get 2k-iles. The even numbered ones are boundaries of the cluster and the odd numbers ones are medoids of the respective clusters. In the first problem medoids are 4,11 and 10 respectively. In the second problem, median as well as medoids are not fixed uniquely. Some more investigation and ‘thicker and sharper’ definitions are desirable here.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 2, No. 1, February : 2024

6. EXAMPLE

In data mining what we receive is bulk data. At this stage, investigate such a data in full strength with heavy length and breadth is not advisable. But now, we discuss a data which is slightly bigger in size compared the problem discussed earlier. Here we use data of 30 farmers with five attributes they are farmers' status poor or rich, they crops cholam or paddy or wheat or ragi , after sales profit or loss, their lands small or big , using natural water or irrigated, and their market distance is near to their place or too long. For this example also we have to clusters these 30 farmers' data using our above new approach explained in section III and IV. After that, finally we find the medoids and clusters for the example given below as given in the section V. The example is given in the following table.

	POOR /RICH	CROPS C/P/W/R	PROFIT/ LOSS	SMALLER/BIG FAR	NATURAL WATER/ IRRIGATED	MARKET NEAR / FAR
1	P	C	F	S	N	E
2	R	C	L	B	I	F
3	P	P	L	B	I	E
4	P	W	F	S	N	F
5	R	R	F	S	I	F
6	P	W	L	S	N	F
7	R	C	F	B	N	E
8	R	P	L	S	N	E
9	R	R	L	B	I	F
10	P	W	L	S	I	E
11	P	W	F	B	I	F
12	R	C	F	S	N	E
13	P	P	L	S	I	E
14	R	R	F	S	N	E
15	P	R	F	S	I	F
16	R	R	F	B	I	F
17	P	C	L	B	I	F
18	P	C	F	S	I	F
19	P	C	L	B	N	E
20	R	P	L	S	I	F
21	P	P	L	B	N	E
22	R	W	L	S	I	F
23	R	P	F	B	I	E
24	P	P	F	S	N	F
25	R	W	L	S	N	E
26	R	W	F	S	N	E
27	R	R	L	B	I	F
28	P	W	F	S	N	F
29	P	P	F	B	I	E
30	P	P	F	S	N	E

7. CONCLUSION

One method of cluster analysis is introduced and evolved here. Instead of mean median is introduced to cluster the data with many attributes. To find out median is difficult. We have introduced a method to evolve the cluster practice. Sequencing of many attributed data is introduced here. This makes the

data sequenced and thus raw. The median is evolved. Likewise deciles, quartitles and so on, k-ile is introduced now to cluster the data into compartments which are themselves boundaries of the cluster. Likewise the centroids, which are termed medoids are detected. Likewise, as any other method of clustering analysis here also the data is clustered. In any such partition there is one inherent similarity. Thus they are called equivalence classes. Here again in any problem the similarity may be inherent. Identifying the inherent similarity, similarity characteristics is again another task.

We have introduced one method to cluster a data using median. We expressed this notion is not unique. Median can be defined differently to cluster a data. One suggestion is forwarded now. Again we hint another approach. The median of data with many attributes X1, X2, … Xn may be defined to be M = (x1,x2,…xn) where xi is median of ith attribute. In this way k-iles and corresponding medoids may be defined and evolved. The suitable algorithm may be selected and used here. According to the purpose and characteristic of the data under investigation, it is envisaged that these approaches will suit some occasions of clustering analysis.

REFERENCES

[1] Gupta, G. K. (2014). Introduction to data mining with case studies. PHI Learning Pvt. Ltd..
[2] Huang, Z. (1997). A fast clustering algorithm to cluster very large categorical data sets in data mining. DMKD, 3(8), 34-39.
[3] Serban, G., & Moldovan, G. S. (2006). A comparison of clustering techniques in aspect mining. Informatica, 51(1).
[4] Rai, P., & Singh, S. (2010). A survey of clustering techniques. International Journal of Computer Applications, 7(12), 1-5.
[5] Roiger, R., & Geatz, M. Data mining: A tutorial-based primer. 2003. Boston MA: Addision Wesley.
[6] Rao, I. K. R. (2003). Data Mining and Clustering Techniques DRTC Workshop on Semantic Web.
[7] K.K.Velukutty. and R.Suresh Kumar., An Innovative Measure For Centrality, Proc.Nat.Modern Techniques And Applications in Mathematics, STC, pp.5-6. 2008.
[8] Kalra, M., Lal, N., & Qamar, S. (2018). K-mean clustering algorithm approach for data mining of heterogeneous data. In Information and Communication Technology for Sustainable Development (pp. 61-70). Springer, Singapore.
[9] Maione, C., Nelson, D. R., & Barbosa, R. M. (2019). Research on social data by means of cluster analysis. Applied Computing and Informatics, 15(2), 153-162.
[10] Faizan, M., Zuhairi, M. F., Ismail, S., & Sultan, S. (2020). Applications of Clustering Techniques in Data Mining: A Comparative Study. ALGORITHMS, 11(12).