K-Medoid Algorithm and Application

Yinze Li

Internet of Things Engineering, Beijing University of Posts and Telecommunications
Beijing, China
2016213609@bupt.edu.cn

Abstract— **K-medoid algorithm is a clustering algorithm which upgrades from K-mean algorithm. This mechanism enables people to cluster objects into a number of groups without pre-label. The procedures of this algorithm are not sophisticated but it can cluster objects in many circumstances. The objects in one group always have some common features. If the algorithm is used for clustering people, we can prospect some traits of one people by analysing other people of the same group.**

Keywords— **K-medoid, clustering, data mining**

I. INTRODUCTION

As the development of the Internet and computing, the amount of data generated by users are increasing exponentially. However, behind the data, we can use some algorithms to find routines or patterns, which is called data mining. There are some phases for data mining, such as frequent pattern, classification and clustering. This paper introduces one classical methods of clustering, K-medoid.

Compared to K-mean algorithm, K-medoid solve the problem where the outline of a cluster is concave.[1] Instead of taking the mean value of the object in a cluster as a reference point, medoids can be used, which is the most centrally located object in a cluster.

By analyzing the object, we can give some recommend base on the common feature. The experiment of this paper analyzes an exam grade of a class. Then use the clustered data, we can give the students some special recommend teaching methods in order to promote their skills.

II. K-MEDOID ALGORITHM

First, there are two basic aspects, supervised learning and unsupervised learning. In supervised learning, training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations. And new data is classified based on the training set. Supervised learning is also called classification.[1]

However, in unsupervised learning, the class labels of training data are unknown. We only give a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data, which is also called clustering. In another words, we don't know the pattern of a set of objects, but the program can observe the attributes of objects and cluster them.[2]

K-medoid is a classical algorithm of partitioning methods of clustering. And there are four basic steps.

1) *Select initial medoids:* Assume that every object has two attributes. And then choose k of them as medoids, which is shown in figure 1, and purple circle objects are medoids.

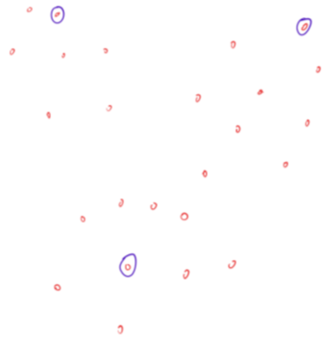

Fig. 1 Select initial medoids (k = 3)

2) *Assign objects:* Calculate the distance between every object and selected medoids. Then assign object itself to nearest medoids, showed on figure 2.

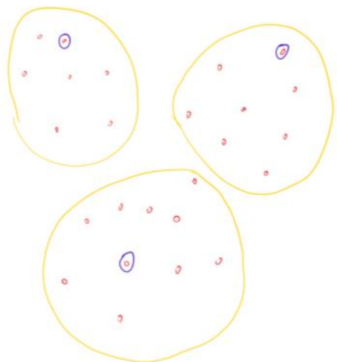

Fig. 2 Assign to medoids

3) *Select a new object:* Randomly select a new non-medoid object, showed in figure 3.

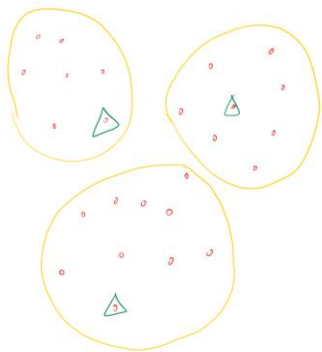

Fig 3. Randomly select a non-medoid object

4) *Calculate and swap:* calculate the total cost (the sum of the distance between the cluster head and other objects) of new selected object, if the cost is smaller than the old cluster medoid, then swap the cluster medoid, otherwise do nothing.

When step 4 finished, do step 2, step 3, step 4 again until the cluster medoid is not change.

III. DISTANCE PROBLEM

About the distance of the objects, there are some calculation methods.[3]

A. Euclidean distance

Euclidean distance can be simply described as the geometric distance between the points in the multi-dimensional space. It should be noted that the Euclidean distance usually uses the original data, not the planned data, such as an attribute at 1-100. The value inside can be used directly, and it is not necessary to normalize it to the [0,1] interval. In this case, the original meaning of the Euclidean distance is eliminated. Because of this, the advantage is that the new object does not affect the distance between any two objects. However, if the metrics of the object attributes are different, such as taking the tense and the percentage system when measuring the score, the result is more significant.

$$\mathrm{Distance}(O_i,\ O_j)=\sqrt{\sum_{k=1}^{n}(O_{ik}-O_{jk})^2}$$

B. Manhattan distance

If the Euclidean distance is regarded as the linear distance of the multi-dimensional space object point, then the Manhattan distance is the distance of the polyline that passes from one object to another, and can sometimes be further described as the object in each dimension in the multi-dimensional space. The average difference, after the average difference is calculated, it should be noted that the Manhattan distance cancels the square of the Euclidean distance, thus weakening the influence of the outliers.

$$\mathrm{Distance}(P_i,\ P_j)=\frac{1}{n}\sum_{k=1}^{n}|P_{ik}-P_{jk}|$$

C. Chebyshev distance

Chebyshev distance is mainly expressed in the multi-dimensional space, the minimum distance consumed by an object to move from one location to another (this distance more vividly reflects the editing distance mentioned in the first section) Concept), so it can be simply described as using one-dimensional attributes to determine which cluster an object belongs to. This is like we are going to identify a rare phenomenon. If two objects have this rare phenomenon, then these two objects Should belong to the same cluster.

$$\mathrm{Distance}(Q_i,\ Q_j)=\mathop{Max}_{k=1}^{n}(Q_{ik}-Q_{jk})$$

D. Power distance

Power distance can be simply described as giving different weight values for different attributes, determining which belongs to that cluster, r, p are custom parameters, according to the actual situation, where p is used to control the progressive of each dimension Weight, r controls the progressive weight of the larger difference between objects. When p=r=1, it is the Manhattan distance. When p=r=2, it is the Euclidean distance. When p=r and tends to infinity, it is the Chebyshev distance. (It can be proved by the limit theory). Therefore, these distances are collectively referred to as the distance, and the shortcoming of the distance is: from the horizontal (dimensional), it treats different components differently, and this defect is cut.

$$\mathrm{Distance}(R_i,\ R_j)=\sqrt[r]{\sum_{k=1}^{n}(|R_{ik}-R_{jk}|)^p}$$

IV. EXPERIMENT

In the experiment, I write java code and get an exam grade table. Then I use my java program to analyse the students' grades of the class. There are six grades for one students, Chinese, mathematics, English, physics, chemistry and biology. For here, I use Euclidean distance to qualify the distance between two objects. And in order to work more efficiently, every time I choose the objects with the smallest cost rather than randomly.

Table 1 show the origin data. Table 2 shows the data which is divided into 4 clusters. Table 3 show the data which is divided into 6 clusters.

Table 1. Origin data (not complete)

ID	ch	math	en	phy	che	bio
1377160214	70	144	132	88	79	82
1377160219	73	130	127.5	79	77	77
1377160228	76	135	128.5	68	73	79
1377160302	75	120	130.5	73	68	82
1377160303	78	124	121	80	76	78
1377160308	98.5	88	127.5	81	63	67
1377160309	103.5	86	126.5	71	66	61
1377160321	71	132	126.5	84	80	70
1377160323	77	133	123.5	82	66	71
1377160324	74	118	134.5	77	81	67
1377160332	110	80	129.5	74	57	87
1377160402	111.5	108	97	60	64	63
1377160416	107.5	84	135	64	77	65
1377160431	107.5	91	128	64	60	67
1377160432	72	132	133	74	68	65
1377160502	109	100	101	50	55	68
1377160503	99.5	110	127	48	55	64
1377160506	95	90	124	63	69	65
1377160517	102	130	91	56	60	68
1377160526	105.5	120	92	66	57	69
1377160530	94.5	111	95	53	73	73
1377160531	100	116	98	55	72	66
1377160533	94.5	82	135	76	74	67
1377160610	92.5	120	122.5	47	67	47
1377160611	109	85	125	76	64	69
1377160615	96.5	101	116.5	44	53	46
1377160625	103.5	108	100	70	63	67
1377160629	105.5	94	112	45	51	67

Table 2. Four clusters

ID	ch	math	en	phy	che	bio	
Cluster 0	medoid 1377160402						
1377160402	111.5	108	97	60	64	63	503.5
1377160502	109	100	101	50	55	68	483
1377160517	102	130	91	56	60	68	507
1377160526	105.5	120	92	66	57	69	509.5
1377160530	94.5	111	95	53	73	73	499.5
1377160531	100	116	98	55	72	66	507
1377160625	103.5	108	100	70	63	67	511.5
1377160630	106	108	94	65	70	63	506
1377160701	105	122	90	67	60	60	504
1377160724	106	115	99	53	67	54	494
1377160804	103.5	96	96	61	72	73	501.5
1377160805	102	113	93	71	52	77	508
Cluster 1	medoid 1377161028						
1377160308	98.5	88	127.5	81	63	67	525
1377160309	103.5	86	126.5	71	66	61	514
1377160332	110	80	129.5	74	57	87	537.5
1377160416	107.5	84	135	64	77	65	532.5
1377160431	107.5	91	128	64	60	67	517.5
1377160506	95	90	124	63	69	65	506
1377160533	94.5	82	135	76	74	67	528.5
1377160611	109	85	125	76	64	69	528
1377160806	107.5	89	123	54	71	72	516.5
1377160812	104.5	83	122.5	77	61	74	522
1377160827	102.5	81	129	79	63	74	528.5
1377161028	99	87	129	70	63	73	521
Cluster 2	medoid 1377160735						
1377160503	99.5	110	127	48	55	64	503.5
1377160610	92.5	120	122.5	47	67	47	496
1377160615	96.5	101	116.5	44	53	46	457
1377160629	105.5	94	112	45	51	67	474.5
1377160631	99	95	130.5	40	40	54	458.5
1377160632	98.5	75	112	44	49	50	428.5
1377160633	89	117	121.5	40	47	42	456.5
1377160722	95.5	115	108.5	42	48	38	447
1377160730	91.5	82	122.5	45	28	53	422
1377160735	96	105	124.5	41	53	54	473.5
1377160817	104	91	119	42	37	76	469
1377160819	103	100	114.5	39	46	68	470.5
1377160820	96	117	120.5	46	47	63	489.5
1377160821	99.5	120	121	49	67	64	520.5
1377160822	96.5	99	124	43	64	57	483.5
1377160931	66	109	122.5	43	40	42	422.5
1377160932	102.5	101	105.5	39	50	60	458
1377161003	86.5	80	111.5	46	49	52	425
1377161004	94.5	109	116.5	41	37	50	448
1377161007	90	89	90.5	47	28	65	409.5
1377161017	94	93	119.5	38	63	69	476.5
Cluster 3	medoid 1377160219						
1377160214	70	144	132	88	79	82	595
1377160219	73	130	127.5	79	77	77	563.5
1377160228	76	135	128.5	68	73	79	559.5
1377160302	75	120	130.5	73	68	82	548.5
1377160303	78	124	121	80	76	78	557
1377160321	71	132	126.5	84	80	70	563.5
1377160323	77	133	123.5	82	66	71	552.5
1377160324	74	118	134.5	77	81	67	551.5
1377160432	72	132	133	74	68	65	544

From table 2, we can conclude that in cluster 0, students get not bad score in total, but should enhance their English skills, so English teacher can take more care of them. In cluster 1, students should enhance their mathematics. In cluster 2, student's physics skill is really bad. And in cluster 3, students' Chinese is considerably bad.

Table 3. Six clusters

ID	ch	math	en	phy	che	bio	
Cluster 0	medoid 1377161028						
1377160308	98.5	88	127.5	81	63	67	525
1377160309	103.5	86	126.5	71	66	61	514
1377160332	110	80	129.5	74	57	87	537.5
1377160416	107.5	84	135	64	77	65	532.5
1377160431	107.5	91	128	64	60	67	517.5
1377160506	95	90	124	63	69	65	506
1377160533	94.5	82	135	76	74	67	528.5
1377160611	109	85	125	76	64	69	528
1377160806	107.5	89	123	54	71	72	516.5
1377160812	104.5	83	122.5	77	61	74	522
1377160827	102.5	81	129	79	63	74	528.5
1377161028	99	87	129	70	63	73	521
Cluster 1	medoid 1377160735						
1377160503	99.5	110	127	48	55	64	503.5
1377160615	96.5	101	116.5	44	53	46	457
1377160629	105.5	94	112	45	51	67	474.5
1377160631	99	95	130.5	40	40	54	458.5
1377160735	96	105	124.5	41	53	54	473.5
1377160817	104	91	119	42	37	76	469
1377160819	103	100	114.5	39	46	68	470.5
1377160820	96	117	120.5	46	47	63	489.5
1377160821	99.5	120	121	49	67	64	520.5
1377160822	96.5	99	124	43	64	57	483.5
1377160932	102.5	101	105.5	39	50	60	458
1377161017	94	93	119.5	38	63	69	476.5
Cluster 2	medoid 1377160219						
1377160214	70	144	132	88	79	82	595
1377160219	73	130	127.5	79	77	77	563.5
1377160228	76	135	128.5	68	73	79	559.5
1377160302	75	120	130.5	73	68	82	548.5
1377160303	78	124	121	80	76	78	557
1377160321	71	132	126.5	84	80	70	563.5
1377160323	77	133	123.5	82	66	71	552.5
1377160324	74	118	134.5	77	81	67	551.5
1377160432	72	132	133	74	68	65	544
Cluster 3	medoid 1377160633						
1377160610	92.5	120	122.5	47	67	47	496
1377160633	89	117	121.5	40	47	42	456.5
1377160722	95.5	115	108.5	42	48	38	447
1377160931	66	109	122.5	43	40	42	422.5
1377161004	94.5	109	116.5	41	37	50	448
Cluster 4	medoid 1377160402						
1377160402	111.5	108	97	60	64	63	503.5
1377160502	109	100	101	50	55	68	483
1377160517	102	130	91	56	60	68	507
1377160526	105.5	120	92	66	57	69	509.5
1377160530	94.5	111	95	53	73	73	499.5
1377160531	100	116	98	55	72	66	507
1377160625	103.5	108	100	70	63	67	511.5
1377160630	106	108	94	65	70	63	506
1377160701	105	122	90	67	60	60	504
1377160724	106	115	99	53	67	54	494
1377160804	103.5	96	96	61	72	73	501.5
1377160805	102	113	93	71	52	77	508
Cluster 5	medoid 1377161003						
1377160632	98.5	75	112	44	49	50	428.5
1377160730	91.5	82	122.5	45	28	53	422
1377161003	86.5	80	111.5	46	49	52	425
1377161007	90	89	90.5	47	28	65	409.5

From table 3, the whole data is divided into 6 clusters. And we can see that, in cluster 0, students really did bad at mathematics. In cluster 1, students' physics are very poor. In cluster 2, Chinese is significantly bad while maths is much better. In cluster 3, students did bad at physics, chemistry and biology. In cluster 4, students did almost average in every subject. In cluster 5, students' almost all courses are not very good, while some of them are good at English.

V. Conclusions

The K-medoid algorithm can help people classify data, analyze data and see the routine and pattern behind data. The clustered data can help teacher know the students' study status of a class, and teach them in a specific way. Furthermore, K-medoid can be used in other situations to create more value.

References

[1] Jiawei Han, Micheline Kamber and Jian Pei, *Data Mining: Concepts and Techniques*, 3rd ed., Morgan Kaufmann, 2011.

[2] WikiPedia. (2019) K-medoid. [Online]. Available: https://en.wikipedia.org/wiki/K-medoids

[3] Data Novia. (2018) Clustering Distance Measures. [Online]. Available: https://www.datanovia.com/en/lessons/clustering-distance-measures/