k-medianoids Clustering Algorithm

James Cha[a], **Teryn Cha**[b], **Sung-Hyuk Cha**[c]
[a] Nothern Valley Regional High School at Demarest, NJ, USA
[b] Computer Science, Essex County College, Newark, NJ, USA
[c] Computer Science Department, Pace University, New York, NY, USA

Abstract

One of the simplest and popular clustering method is the simple k-means clustering algorithm. One of the drawbacks of the method is its sensitivity to outliers. To overcome this problem, the k-medians clustering algorithm is used. Another limitation of the simple k-means clustering algorithm is the Euclidean space assumption. The k-medoids has been used to overcome this assumption. Here a combined method called the k-medianoids clustering algorithm is proposed. A *medianoid* is a kind of median that does not require the Euclidean space assumption and is formally defined. The proposed method is demonstrated using nucleotide sequences.

Introduction

One of the simplest and popular clustering method is the simple k-means clustering algorithm. It has been widely used in various applications, including image processing, pattern recognition, and data mining. It has also been used in various fields, such as marketing, where it has been used to segment customers into different groups based on their demographic and purchasing behavior.

This simple method had been proposed by multiple researchers independently such as in (Forgy 1965; Friedman and Rubin 1967; Macqueen 1967). The earliest one goes as far back as 1956 (Steinhaus 1957). It is also known as Lloyd's algorithm or Voronoi iteration (Lloyd 1982) and the most popular and earliest one.

One of the major drawbacks of the k-means algorithm is that k-means algorithm is too sensitive to outliers. A variation to the k-means clustering method can be made to mitigate the outlier problem. One of them is k-medians clustering (see (Jain and Dubes 1988) for details). The membership of each instance to clusters is reliant to the median statistical parameter instead of the mean. It is closely related to the general but hard p-median facility location problem (Hakimi 1964).

Another problem of the simple k-means clustering algorithm is the Euclidean space assumption. If the data space is non-Euclidean such as binary or data is represented as a graph or string, mean is often meaningless or cannot be computed. The k-medoids has been used to overcome this assumption. The medoid was utilized instead of mean in k-means method to cluster and it is called partitioning around medoids, PAM in short (Kaufman and Rousseeuw 1987) (see (Kaufman and Rousseeuw 1990, Ch. 2) for further description).

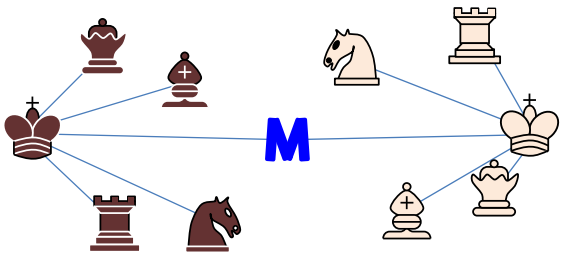
The k-medoids clustering algorithm still suffers from the outlier problem and the k-median clustering may not be useful if the data space is non-Euclidean. In order to incorporate the advantages of both k-medians and k-medoids to overcome some limitations of k-means algorithm, the k-medianoids clustering algorithm is proposed here.

Medianoids

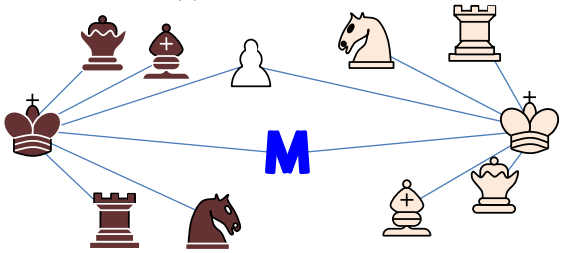
In order to compute the medianoid, two extreme points that make a diameter of the cluster are first identified as given in eqn (1).

$$\{k_1, k_2\} = \underset{x,y \in \mathcal{X}}{\operatorname{argmax}}(d(x, y)) \tag{1}$$

Then the medianoid is defined recursively as follows.



(a) odd number case



(b) even number case

Figure 1: Medianoids

Table 1: Pairwise Distance Matrix for SSU rDNA sequences retrieved from (Sokolova and Hawke 2016)

	PN	PM	P6	MM	AM	NC	PP	GA	HA	TC	SL	AP	AR	LO
PN	0	0.001	0.02	0.016	0.351	0.314	0.227	0.228	0.216	0.265	0.226	0.458	0.401	0.38
PM	0.001	0	0.022	0.017	0.353	0.316	0.225	0.23	0.218	0.267	0.228	0.455	0.403	0.383
P6	0.02	0.022	0	0.031	0.346	0.31	0.221	0.226	0.214	0.259	0.225	0.467	0.393	0.386
MM	0.016	0.017	0.031	0	0.352	0.32	0.233	0.234	0.219	0.269	0.238	0.479	0.409	0.4
AM	0.351	0.353	0.346	0.352	0	0.083	0.397	0.37	0.372	0.341	0.375	0.522	0.476	0.521
NC	0.314	0.316	0.31	0.32	0.083	0	0.365	0.342	0.343	0.33	0.347	0.487	0.441	0.468
PP	0.227	0.225	0.221	0.233	0.397	0.365	0	0.069	0.066	0.205	0.156	0.446	0.405	0.335
GA	0.228	0.23	0.226	0.234	0.37	0.342	0.069	0	0.055	0.199	0.157	0.437	0.402	0.342
HA	0.216	0.218	0.214	0.219	0.372	0.343	0.066	0.055	0	0.181	0.139	0.434	0.381	0.332
TC	0.265	0.267	0.259	0.269	0.341	0.33	0.205	0.199	0.181	0	0.126	0.452	0.362	0.357
SL	0.226	0.228	0.225	0.238	0.375	0.347	0.156	0.157	0.139	0.126	0	0.455	0.363	0.326
AP	0.458	0.455	0.467	0.479	0.522	0.487	0.446	0.437	0.434	0.452	0.455	0	0.319	0.412
AR	0.401	0.403	0.393	0.409	0.476	0.441	0.405	0.402	0.381	0.362	0.363	0.319	0	0.423
LO	0.38	0.383	0.386	0.4	0.521	0.468	0.335	0.342	0.332	0.357	0.326	0.412	0.423	0

$$\text{medianoid}(\mathcal{X}) = \begin{cases} x \in \mathcal{X} & \text{if } |\mathcal{X}| = 1 \\ \underset{x \in \mathcal{X}}{\operatorname{argmin}}(d(k_1, x) + d(k_2, x)) & \text{if } |\mathcal{X}| = 2 \\ \text{medianoid}(\mathcal{X} - \{x_1, x_2\}) & \text{if } |\mathcal{X}| > 2 \end{cases} \quad (2)$$

where $x_1 = \underset{x \in \mathcal{X}}{\operatorname{argmin}}(d(k_1, x))$ and $x_2 = \underset{x \in \mathcal{X}}{\operatorname{argmin}}(d(k_2, x))$. The medianoid in odd and even number samples are depicted in Figure 1 (a) and (b) respectively.

Two kings pick the closest element one by one, recursively. When only one left, it is the medianoid. When only two remain, pick one whose sum of distances to kings is smaller is the medianoid.

Experiments on Microsporidia SSU rDNA

While k-means algorithm is guaranteed to converge in Euclidean space, k-medianoids clustering algorithm may not converge. To disprove the convergency of the k-medianoids clustering algorithm, SSU rDNA sequences of fourteen species of microsporidia from crustaceans and fish, retrieved from (Sokolova and Hawke 2016), are considered.

Table 1 shows the pairwise distance matrix among 14 species where Kimura-2 distance (Kimura 1980) is used. Kimura 2-parameter (K2P) is a model for comparing the evolutionary distances between sequences.

The medoid of the fourteen microsporidia species is PN since the sum of distances to the rest of species is minimum. To compute the medianoid, two extreme points need to be identified first. They are AM and AP since the distance between them is the maximum by eqn (1). By eliminating the closest one recursively, two species remain and they are PM and PP. The medianoid is PM by eqn (2);
d(PM, AM) + d(PM, AP) $<$ d(PP, AM) + d(PP, AP).

Suppose a user wishes to find two clusters, i.e., $k = 2$. Figure 2 (a) shows the steps of the k-medoids clustering algorithm where initial cluster medoids are M = {P6, PP}. The algorithm terminates after two steps and finds two clusters with their respective medoids: M = {PN, HA}.

Figure 2 (b) shows the steps to find two clusters by the k-medianoids clustering algorithm. Suppose that initial cluster medoids are M = {P6, PP}. Then every odd number step will have the same clusters with identical medianoids, M = {MM, PP}. Every even number step will have the same clusters with identical medianoids, M = {P6, GA}. It will not terminate and fall into an infinite loop.

	step	Cluster 1	Cluster 2
0	initial	P6	PP
1	Expectation	{PN, PM, P6, MM, AM, NC, AR}	{PP, GA, HA, TC, SL, AP, LO}
	Maximi.	PN	HA
2	Expectation	{PN, PM, P6, MM, AM, NC}	{PP, GA, HA, TC, SL, AP, AR, LO}
	Maximi.	PN	HA

(a) k-medoids

	step	Cluster 1	Cluster 2
0	initial	P6	PP
1	Expectation	{PN, PM, P6, MM, AM, NC, AR}	{PP, GA, HA, TC, SL, AP, LO}
	Maximi.	MM	PP
2	Expectation	{PN, PM, P6, MM, AM, NC}	{PP, GA, HA, TC, SL, AP, AR, LO}
	Maximi.	P6	GA
3	Expectation	{PN, PM, P6, MM, AM, NC, AR}	{PP, GA, HA, TC, SL, AP, LO}
	Maximi.	MM	PP
∞		⋮	⋮

(b) k-medianoids

Figure 2: Clustering illustration for SSU rDNA

Conclusion

The k-medianoids clustering algorithm was proposed and tested on the DNA gene sequence clustering. The major contribution is that the concept of the medianoid was defined recursively. Further theoretical analysis remains to be future work.

References

Forgy, E. W. 1965. Cluster analysis of multivariate data : efficiency versus interpretability of classifications. *Biometrics* 21:768–769.

Friedman, H. P., and Rubin, J. 1967. On some invariant criteria for grouping data. *Journal of the American Statistical Association* 62(320):1159–1178.

Hakimi, S. L. 1964. Optimum locations of switching centers and the absolute centers and medians of a graph. *Operations Research* 12(3):450–459.

Jain, A. K., and Dubes, R. C. 1988. *Algorithms for Clustering Data*. Prentice-Hall, Inc.

Kaufman, L., and Rousseeuw, P. J. 1987. Clustering by means of medoids. Faculty of mathematics and informatics, Delft University of Technology.

Kaufman, L., and Rousseeuw, P. J. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution* 16:111–120.

Lloyd, S. P. 1982. Least squares quantization in pcm. *IEEE Transactions on Information Theory* 28(2):129–137.

Macqueen, J. 1967. Some methods for classification and analysis of multivariate observations. In *In 5th Berkeley Symposium on Mathematical Statistics and Probability*, 281–297.

Sokolova, Y., and Hawke, J. P. 2016. Perezia nelsoni (microsporidia) in agmasoma penaeiinfected atlantic white shrimp litopenaeus setiferus (paenaidae, decapoda) and phylogenetic analysis of perezia spp. complex. *Protistology* 10(3):67–78.

Steinhaus, H. 1957. Sur la division des corps matériels en parties. *Bulletin de l'Académie Polonaise des Sciences, Classe 3* 4:801–804.