Proportional data clustering using K-means algorithm: A comparison of different distances

Conference Paper · March 2017

DOI: 10.1109/ICIT.2017.7915506

2 authors:

Jai Puneet Singh
Concordia University Montreal
15 PUBLICATIONS **29** CITATIONS

SEE PROFILE

Nizar Bouguila
Concordia University Montreal
321 PUBLICATIONS **3,346** CITATIONS

SEE PROFILE

Proportional Data Clustering using K-Means Algorithm: A comparison of different distances

Jai Puneet Singh *, Nizar Bouguila
CIISE, Concordia University, Montréal, Québec, Canada
jaipuneet.singh@mail.concordia.ca,nizar.bouguila@concordia.ca

***Abstract*—In this paper, we discuss proportional data clustering. It emerges in many applications such as document clustering and image classification using bag of visual words approach. When deploying mixture models for clustering, there is always a problem of initialization, and it is common to initialize using K-means algorithm. In proposed work, we present K-means clustering approach using different distance metrics. In particular, we propose the consideration of the Aitchison distance. Experimental results are presented using silhouette plots for showing divergence from the center, and confusion matrix is used to validate our clustering of synthetic and real data sets of images and texts. The algorithm with Aitchison distance metric results into lower error rates.**

Keywords: K-means, unsupervised learning, Aitchison distance, clustering.

I. Introduction

With the advancement of technology, huge amount of data is generated every day. Proportional data in particular are naturally generated by different fields such as geology, Bioinformatics, computer vision, etc. Proportional data can be defined as any vector $x = (x_1, x_2, ..., x_D)$ subject to unit sum constraint where $(x_1 + x_2 ... + x_D) = 1$ and $x_d \geq 0$, $d = 1, 2, ..., D$ [2] [3]. The clustering of this type of data requires efficient algorithms [4]. Cluster analysis is prevalent in any discipline that involves analysis of multivariate data [5]. The problem of unsupervised clustering using mixture models requires good initialization which is generally done with the help of K-means algorithm [6] [7] [8]. However, K-means uses Euclidean distance which finds shortest distance between two samples. Unfortunately, the Euclidean distance is not appropriate for proportional data [9]. Despite the fact that Aitchison and other distance metrics are appropriate for proportional data, they have not received much attention compared to Euclidean distance. However, some related works do exist. For instance, Kashima et al proposed a L1 distance based K-means algorithm to address the problem of proportional vector clustering [10]. Hijazi et al. used Dirichlet regression models for modeling compositional data and came with the conclusion that Dirichlet regression model is just an alternative to log-ratio analysis which can be done by Aitchison Log Ratio Analysis [11]. The goal of this work is to compare different distance metrics when clustering proportional data with K-means algorithm.

Aitchison introduced the log ratio analysis to model compositional data. In this work, we propose to use Aitchison distance metric for clustering of proportional data [9] [12]. Moreover, we compare it with several other distances in several applications. The rest of the paper is organized as follows: In section II, the proposed method is explained in details, and various distance metrics functions are introduced. In section III, outlier detection method with the proposed algorithm is shown. Section IV gives different experimental results on many synthetic and real data sets. Finally, in section V concluding remarks are drawn.

II. The Proposed Method

K-means clustering uses distance metrics to find nearest neighbors and mostly Euclidean distance has been used. The objective function of K-Means can be represented as follows:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left|\left| x_i^j - c_j \right|\right|^2 \quad (1)$$

In this equation x_i represents the data point and c_j represents the cluster center. This type of distance generates spherical shaped type of clusters as a result. Kashima et al. [13] proposed L_1 distance which is also known as Manhattan distances for optimization of K-means clustering on proportional data and successful improvement was seen. There are many different distances and divergence metrics. However in our paper, we concentrated on Euclidean log transformed data, Aitchison's and Kullback-Leibler divergence on proportional data. These types of distances have been known for a long time but they have not been explored for proportional data. The only drawback of these types of distances is that they don't accept 0 values. To normalize these type of data-sets Martin et al. [14] proposed to deal with zeros in compositional data which we have applied an approach before clustering. In this paper, we are proposing K-means clustering using Aitchison distance. As per our knowledge, this distance has not been applied on K-means in the past.

In the K-means algorithm, we change distances which are used in the steps 2 and 4 of Algorithm 1. The distance metrics that we have explored are given in Table 1.

III. Outlier Detection

Outlier detection is a deeply researched topic in both communities of statistics and data mining [15]. Outlier detection methods are categorized as external and internal methods [16]. In our case, we used internal outlier detection techniques where after K-means clustering with various distance metrics, distance is compared with the other data in same group with centroids of particular group as shown in Algorithm 2.

Algorithm 1 K-Means Algorithm

```
Set the Initial number of centroids randomly or sequentially
Calculate the distance between each data point and cluster centers
repeat:
    Assign the minimum distance data points to cluster center whose distance is minimum to that point.
    Recalculate the cluster center using:
$c_i = \frac{1}{m_i}\sum_{j=1}^{m_i} x\left(i\right)$; $m_i$ represents total number of data points in $\left(i\right)$ cluster
    Re-calculate the distance between each data point and newly obtained cluster center
until : No data point is reassigned.
```

TABLE I: Different Distance Metrics used in K-Means

S.No.	Distance Name	Distance Metrics
1	Euclidean Distance	$d_E^2\left(x, y\right) = \sum_i \left(x_i - y_i\right)^2$
2	Euclidean Distance log transformed Data	$d_{EL}^2\left(x, y\right) = \sum_i \left(\log x_i - \log y_i\right)^2$
3	J-divergence	$d_{jd}^2\left(x, y\right) = \sum_i \left(\log x_i - \log y_i\right)\left(x_i - y_i\right)$
4	Jeffery's-Matusita Distance	$d_m^2\left(x, y\right) = \sum_i \left(\sqrt{x_i} - \sqrt{y_i}\right)^2$
5	Manhattan Distance (L1 Distance)	$d_{L1}^2\left(x, y\right) = \sum_i \left\|x_i - y_i\right\|$
6	Kullback-Leibler Divergence	$d_{KL}\left(x, y\right) = \sum_i \left(x_i \log \frac{x_i}{y_i} + y_i \log \frac{y_i}{x_i}\right)$
7	Aitchison's Distance	$d_{AD}\left(x, y\right) = \frac{1}{D}\sum_{i<j}\left(\log \frac{x_i}{x_j} - \log \frac{y_i}{y_j}\right)$ $d_{AD}^2\left(x, y\right) = \sum_{k=1}^{D}\left(\log \frac{x_i}{g\left(x_j\right)} - \log \frac{y_i}{g\left(y_j\right)}\right)$
8	Cosine Distance	$d_C\left(x, y\right) = 1 - \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}}$
9	Mahalonbis Distance	$d_m^2\left(x, y\right) = \left(x - y\right)^T S^{-1}\left(x - y\right)$

IV. Experiments with Synthetic and Real Data

A. Synthetic Data sets:

We have taken synthetic data set to validate our proposed method. We have generated samples of data using 2 different mixtures of Dirichlet distribution [17] [18] [1] by using different α parameters. This synthetic data set consist of 400 vectors with a dimension of 100. The Dirichlet distribution [19] can be expressed as:

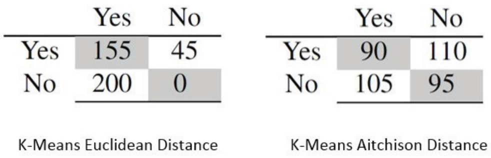
$$p\left(X, \theta\right) = \frac{1}{\beta\left(\alpha\right)}\prod_{i=1}^{K} x_i^{\alpha_i - 1} \tag{2}$$

$$\beta\left(\alpha\right) = \frac{\prod_{i=1}^{k}\Gamma\left(\alpha_i\right)}{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)} \tag{3}$$

Over here in above equation $\alpha = \left(\alpha_1, ..., \alpha_K\right)$. When we did K-means clustering with Euclidean and Aitchison distance metrics the results obtained are shown by confusion matrix in Fig 1. which shows that Aitchison distance is much better for proportional data clustering.

Algorithm 2 Outlier Detection Algorithm

```
To perform K-means Algorithm (Algo. 1 )
INPUT: D-Dimensional Data $X_n, n = 1, ..., N$, No of Clusters and choose distance metrics (Aitchison Distance) for K-means.
Find distance between obtained centers and points using distance metrics.
Sort in descending order the distance obtained.
$d_{max} = max_i\left\{\left\|x_i - c_i\right\|\right\}, i = 1...N$
Highest distance between center and points is an outlier.
```



	Yes	No
Yes	155	45
No	200	0

K-Means Euclidean Distance

	Yes	No
Yes	90	110
No	105	95

K-Means Aitchison Distance

Fig. 1: Confusion matrices when clustering synthetic data set using Euclidean and Aitchison distances

B. Real Data sets:

In our experiments we have taken three real data sets for clustering. The data sets are:

- Data set 1 (Text Documents) [20]
  Instances: 3430
  Vocabulary words: 6906
  number of words in collection: 467714.
- Data set 2 (Spambase) [21]
  Instances: 4601
  Attributes: 57
  Labeled into into 2 groups (Spam and Non Spam Email).
- Data set 3 : Human Face Identification [22]
  Number of Facial expression: 400 images
  Dictionary size: 1000
  classification is 40.

The first experiment is on document clustering using bag of words approach. In the bag of words approach we have taken the KOS blog entries [20] as our data set which contains 3430 documents and number of words in the vocabulary is 6906. The problem arises with value of zeros in the datasets. So, before performing clustering we remove 0 values by using exponentially small value around $2^{(-52)}$. Processing the zeros is an important task as calculation involves log. After processing, we normalize data using following equation:

$$x_i = \frac{x_i}{x_1 + x_2 .... + x_D} \tag{4}$$

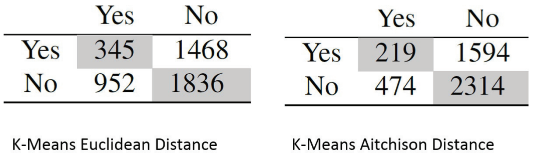
The second experiment is on visual objects clustering using bag of visual words approach. The real image data set contains facial expression in 400 images of 256 pixels each. In order to generate bag of visual words, SIFT descriptor [23] has been used as a feature extractor for each image. After this process, each image is represented as a proportional vector.

Fig. 2: Data set 3 contains 400 facial images of 40 individuals, each showing 3 different but similar facial ex- pressions [22]

High dimensional clusters are very difficult to visualize. The best method to visualize high dimensional clustering is with the help of Silhouette. Peter J.Rousseew [24] explains how the cluster analysis can be done with the help of silhouette. Silhouette is a method which is used to find the consistency of clustered data. The values obtained for each vector can be combined and statistical values can be obtained to validate the clustering operation. The statistical values ranges between -1 to 1, where it shows that how an object is well matched to its own cluster. Higher silhouette value means the clustering is appropriate. If the negative value is obtained it means more clustering can be done and the points does not lie within the same clusters. After finding the silhouette value it is necessary to find the group wise summary statistics which gives us the clear view of the clustering. The statistical value obtained helps us to determine the concentration and pertinent of a cluster. Table 2. shows the statistical values for clustering using different distance metrics of KOS blog entries. Our results for text clustering shows that the Aitchison distance metrics is most appropriate among all, as the higher silhouette value the more appropriate is the clustering. It is followed by Kullback distance metrics and then Euclidean log distance metrics. The Euclidean, Matisuita and the cosine show the poorest results for proportional data clustering, while, Cosine distance metrics gives good results when data set is in form of frequency. For Spambase [21],we have obtained good results as shown in Fig. 6.

C. Outlier detection results:

We have used Haberman's survival dataset for finding outliers. The data set consists of 306 instances and 3 attributes. Through the results obtained by k-means with Aitchison distance was able to determine 5 outliers correctly whereas k-means with Euclidean distance could determine 3 outliers correctly. Figure 6 shows the confusion matrix after performing the experiment and figure 7 shows graphical output.



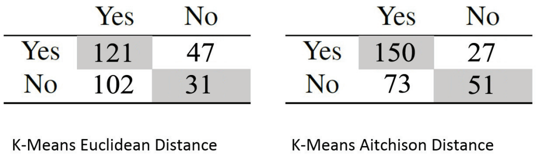
	Yes	No
Yes	345	1468
No	952	1836

K-Means Euclidean Distance

	Yes	No
Yes	219	1594
No	474	2314

K-Means Aitchison Distance

Fig. 3: The spambase data set K-means clustering with Euclidean and Aitchison distance representation by confusion matrix



	Yes	No
Yes	121	47
No	102	31

K-Means Euclidean Distance

	Yes	No
Yes	150	27
No	73	51

K-Means Aitchison Distance

Fig. 4: The confusion matrices when clustering after outlier detection

D. Error percentage:

To find error percentage Silhouette method is used as given by equation 5. In this equation, $a(i)$ is the average dissimilarity within the same cluster, $b(i)$ is lowest average dissimilarity of i to any other cluster where as i is a datum. It is used to find total sum of the silhouette values of each cluster on a data set. The error percentage is used to the dissimilarity of the clustered data as given by equation 6. It represents group wise statistics where mean of each cluster silhouette value is calculated and $\bar{x}$ gives the sum of the mean of silhouette value of each cluster. In equation 7, we find error percentage or dissimilarity percentage of K-means clustering based on distance metrics.

$$s\left(i,j\right)=\begin{cases}1-a\left(i\right) & :a\left(i\right)<b\left(i\right)\\ 0 & :a\left(i\right)=b\left(i\right)\\ b\left(i\right)/a\left(i\right) & :a\left(i\right)>b\left(i\right)\end{cases} \tag{5}$$

$$\bar{x}=\frac{1}{N}\sum_{j=1}^{N}s\left(i,j\right) \tag{6}$$

$$e=\frac{N-\bar{x}}{N}\times 100 \tag{7}$$

The above process was performed on facial image data set [22] of 400 images which has 40 different classes of 10 images each. The error percentage or dissimilarity measure obtained by clustering into 40 different clusters was 24.75 % by k-means using Aitchison distance and 32.66 % by k-means using euclidean distance. Table 2 provides with the statistical values which are mean of each silhouette value that clearly determines Aitchison distance is much better distance metric when compared to other distance metrics which is followed by Euclidean log in which log transformed data has been considered while processing K-Means clustering.

S.No.	Cluster	Euclidean Dist.	Aitchison Dist.	Kullback Div.	Cosine	Euclidean Log Dist	Matusita Dist.
1.	Cluster 1	-0.0795	1.000	0.2720	0.0047	0.4205	-0.1169
2.	Cluster 2	0.0477	-0.2244	-0.0805	-0.0994	0.2389	0.1666
3.	Cluster 3	0.1952	0	0.1569	0.0869	-0.2761	-0.0789
	Sum	0.1750	0.7756	0.3484	-0.0078	0.3832	-0.0293

TABLE II: The Statistical Value of different clusters by distance metrics for K Blog Entries Bag of words

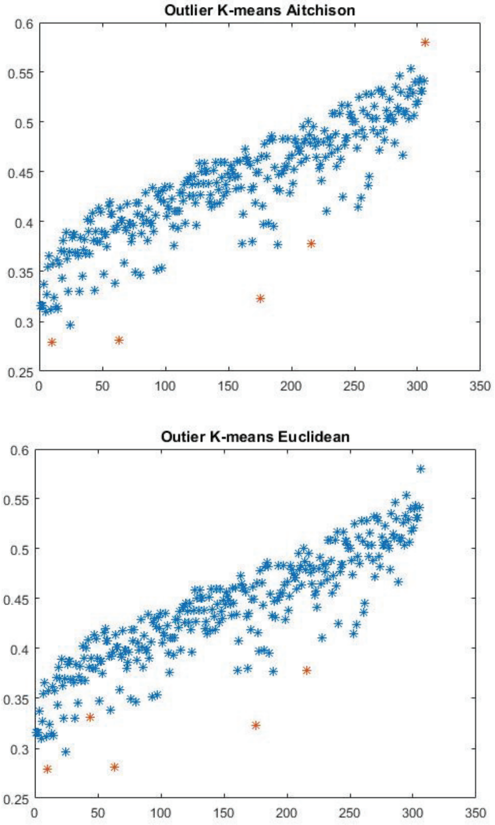


Fig. 5: Haberman's survival Data set results of an outlier [25]

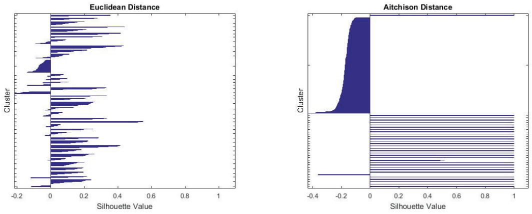


Fig. 6: Silhouette value for each points with Euclidean and Aitchison distance of image clustering of 400 Image Dataset

V. Conclusion

Different distance metrics have been investigated for the K-means clustering of proportional data. Aitchison distance has been shown performs its best with respect to clustering. This distance can be used with different types of clustering methodologies. The proper initialization for mixture models is a crucial step in unsupervised learning [26]. Using, Aitchison distance as initialization step can give better mixture results applicable to proportional data. It has also been observed that Aitchison distance performs better for sparse data sets and for high dimension when compared with Euclidean distances for this particular type of data. By finding the K-means score it has been seen that Aitchison's distance is more viable solution as a distance metric for doing K-means clustering for proportional data.

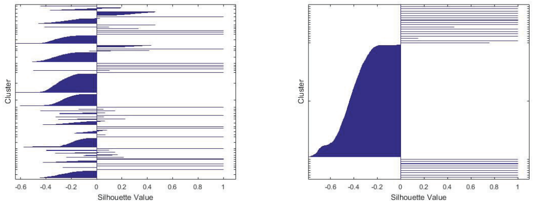


Fig. 7: Silhouette value for each points with Euclidean and Aitchison distance of text clustering of K blog Entries when cluster (N) is 50

VI. Acknowledgment

The authors would like to thank Mitacs Inc. for providing the research support and Concordia University for providing the facilities to conduct a research.

References

[1] T. Bdiri, N. Bouguila, and D. Ziou, "Object clustering and recognition using multi-finite mixtures for semantic classes and hierarchy modeling," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1218–1235, 2014.

[2] N. Bouguila and D. Ziou, "A countably infinite mixture model for clustering and feature selection," *Knowledge and information systems*, vol. 33, no. 2, pp. 351–370, 2012.

[3] W. Fan, N. Bouguila, and D. Ziou, "Unsupervised hybrid feature extraction selection for high-dimensional non-gaussian data clustering with variational inference," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 7, pp. 1670–1685, 2013.

[4] J. Martín-Fernández, C. Barceló-Vidal, V. Pawlowsky-Glahn, A. Buccianti, G. Nardi, and R. Potenza, "Measures of difference for compositional data and hierarchical clustering methods," in *Proceedings of IAMG*, vol. 98, 1998, pp. 526–531.

[5] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[6] W. Fan, N. Bouguila, and D. Ziou, "Unsupervised anomaly intrusion detection via localized bayesian feature selection," in *Data Mining (ICDM), 2011 IEEE 11th International Conference on*. IEEE, 2011, pp. 1032–1037.

[7] N. Bouguila and D. Ziou, "Mml-based approach for finite dirichlet mixture estimation and selection," in *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Springer, 2005, pp. 42–51.

[8] A. Sefidpour and N. Bouguila, "Spatial color image segmentation based on finite non-gaussian mixture models," *Expert Systems with Applications*, vol. 39, no. 10, pp. 8993–9001, 2012.

[9] J. Aitchison, "The statistical analysis of compositional data," 1986.

[10] S. Boutemedjet, D. Ziou, and N. Bouguila, "Model-based subspace clustering of non-gaussian data," *Neurocomputing*, vol. 73, no. 10, pp. 1730–1739, 2010.

[11] R. Hijazi, "An em-algorithm based method to deal with rounded zeros in compositional data under dirichlet models," in *Proceedings of the 4th International Workshop on Compositional Data Analysis*, 2011, pp. 1–5.

[12] N. Bouguila and D. Ziou, "A probabilistic approach for shadows modeling and detection," in *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, vol. 1. IEEE, 2005, pp. I–329.

[13] H. Kashima, J. Hu, B. Ray, and M. Singh, "K-means clustering of proportional data using l1 distance," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. IEEE, 2008, pp. 1–4.

[14] J. A. Martın-Fernandez, J. Palarea-Albaladejo, and R. A. Olea, "Dealing with zeros," *Compositional data analysis: Theory and applications*, pp. 43–58, 2011.

[15] S. Chawla and A. Gionis, "k-means-: A unified approach to clustering and outlier detection." in *SDM*. SIAM, 2013, pp. 189–197.

[16] K.-A. Yoon, O.-S. Kwon, and D.-H. Bae, "An approach to outlier detection of software measurement data using the k-means clustering method," in *First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007)*. IEEE, 2007, pp. 443–445.

[17] N. Bouguila and D. Ziou, "On fitting finite dirichlet mixture using ecm and mml," in *International Conference on Pattern Recognition and Image Analysis*. Springer, 2005, pp. 172–182.

[18] N. Bouguila, D. Ziou, and R. I. Hammoud, "On bayesian analysis of a finite generalized dirichlet mixture via a metropolis-within-gibbs sampling," *Pattern Analysis and Applications*, vol. 12, no. 2, pp. 151–166, 2009.

[19] N. Bouguila and W. ElGuebaly, "Discrete data clustering using finite mixture models," *Pattern Recognition*, vol. 42, no. 1, pp. 33–42, 2009.

[20] "Kos blog enteries," http://dailykos.com, accessed: 2016-08-10.

[21] "Landsat satellite data set," https://archive.ics.uci.edu/ml/datasets/ Spambase, accessed: 2016-09-01.

[22] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on*. IEEE, 1994, pp. 138–142.

[23] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th ACM international conference on Multimedia*. ACM, 2007, pp. 357–360.

[24] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[25] "Haberman's survival data set," https://archive.ics.uci.edu/ml/datasets/ Haberman's+Survival, accessed: 2016-09-01.

[26] N. Bouguila, "Bayesian hybrid generative discriminative learning based on finite liouville mixture models," *Pattern Recognition*, vol. 44, no. 6, pp. 1183–1200, 2011.

[27] "30 coast images dataset," http://groups.csail.mit.edu/vision/SUN/, accessed: 2016-08-10.

[28] D. Birant and A. Kut, "St-dbscan: An algorithm for clustering spatial–temporal data," *Data & Knowledge Engineering*, vol. 60, no. 1, pp. 208–221, 2007.

[29] P. Thomas and D. Lovell, "Compositional data analysis (coda) approaches to distance in information retrieval," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 2014, pp. 991–994.

[30] M. Senoussaoui, P. Kenny, P. Dumouchel, and T. Stafylakis, "Efficient iterative mean shift based cosine dissimilarity for multi-recording speaker clustering," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7712–7715.

[31] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 03 1951. [Online]. Available: http://dx.doi.org/10.1214/aoms/1177729694

[32] R. H. Hijazi and R. W. Jernigan, "Modelling compositional data using dirichlet regression models," *Journal of Applied Probability & Statistics*, vol. 4, no. 1, pp. 77–91, 2009.