

K-means with Three different Distance Metrics

Archana Singh
Amity University Uttar Pradesh
AIIT, I1 Block, 3rd Floor
Sector-125, NOIDA(UP)India

Avantika Yadav
Amity University Uttar Pradesh
ASET, E2 Block
Sector-125, NOIDA(UP)India

Ajay Rana
Amity University Uttar Pradesh
ASET, E2 Block, Ground Floor
Sector-125, NOIDA(UP)India

ABSTRACT

The power of k-means algorithm is due to its computational efficiency and the nature of ease at which it can be used. Distance metrics are used to find similar data objects that lead to develop robust algorithms for the data mining functionalities such as classification and clustering. In this paper, the results obtained by implementing the k-means algorithm using three different metrics Euclidean, Manhattan and Minkowski distance metrics along with the comparative study of results of basic k-means algorithm which is implemented through Euclidian distance metric for two-dimensional data, are discussed. Results are displayed with the help of histograms.

General Terms

Algorithms, Measurement, Performance.

Keywords

Centroids, clustering, distortion, metrics, similarity matrix.

1. INTRODUCTION

CLUSTERING is a technique to categorize the data into groups. Distance metrics plays a very important role in the clustering process. The more the similarity among the data in clusters, more the chances of particular data-items to belong to particular group. There are number of algorithms which are available for clustering. In general, K-means is a heuristic algorithm that partitions a data set into K clusters by minimizing the sum of squared distance in each cluster. The algorithm consists of three main steps: a) initialization by setting center points (or initial centroids) with a given K, b) Dividing all data points into K clusters based on K current centroids, and c) updating K centroids based on newly formed clusters. It is clear that the algorithm always converges after several iterations of repeating steps b) and c). In this paper, the simulation of basic k-means algorithm is done, which is implemented using Euclidian distance metric.

In the proposed paper, the k-means algorithm using Manhattan distance metrics and Minkowski distance metric is implemented and also the results obtained through both the methods with the basic k-mean's result are compared.

2. DISTANCE METRICS OVERVIEW

In order to measure the similarity or regularity among the data-items, distance metrics plays a very important role. It is necessary to identify, in what manner the data are inter-related, how various data dissimilar or similar with each other and what measures are considered for their comparison. The main purpose of metric calculation in specific problem is to obtain an appropriate distance /similarity function. Metric learning has emerged as a popular issue in many learning tasks and also it can be applied in a wide variety of settings,

since many learning problems involve a definite notion of distance or similarity [1,4]. A metric function or distance function is a function which defines a distance between elements/objects of a set [4,5]. A set with a metric is known as metric space. This distance metric plays a very important role in clustering techniques. The numerous methods are available for clustering. In the current paper, the solution of k-means clustering algorithm using Manhattan distance metric is proposed. Normally, the task is to define a function Similarity(X,Y), where X and Y are two objects or sets of a certain class, and the value of the function represents the degree of "similarity" between the two. Formally, a distance function is a function Dist with positive real values, defined on the Cartesian product $X \times X$ of a set X. It is called a metric of X if for each $x, y, z \in X$:

- $\text{Dist}(x,y)=0$ if $x=y$ (the identity axiom);
- $\text{Dist}(x,y) + \text{Dist}(y,z) \geq \text{Dist}(x,z)$ (the triangle axiom);
- $\text{Dist}(x,y)=\text{Dist}(y,x)$ (the symmetry axiom).

Metric space metric provides a set X.

2.1 Euclidean Distance

Euclidean distance computes the root of square difference between co-ordinates of pair of objects.

$$\text{Dist}_{XY} = \sqrt{\sum_{k=1}^m (X_{ik} - X_{jk})^2}$$

2.2 Manhattan Distance

Manhattan distance computes the absolute differences between coordinates of pair of objects

$$\text{Dist}_{XY} = |X_{ik} - X_{jk}|$$

2.3 Chebychev Distance

Chebychev Distance is also known as maximum value distance and is computed as the absolute magnitude of the differences between coordinate of a pair of objects.

$$\text{Dist}_{XY} = \max_k |X_{ik} - X_{jk}|$$

2.4 Minkowski Distance

Minkowski Distance is the generalized metric distance.

$$Dist_{XY} = \left(\sum_{k=1}^d |X_{ik} - X_{jk}|^{\frac{1}{p}} \right)^p$$

Note that when $p=2$, the distance becomes the Euclidean distance. When $p=1$ it becomes city block distance. Chebyshev distance is a variant of Minkowski distance where $p=\infty$ (taking a limit). This distance can be used for both ordinal and quantitative variables.

3. CLUSTER ANALYSIS TECHNIQUE

As a data mining function, clustering can be used for distribution of data, to observe the characteristics of each cluster, and to focus on a particular set of clusters for further analysis. Clustering is one of the most fundamental issues in data recognition. It plays a very important role in searching for structures in data. It may serve as a pre-processing step for other algorithms, which will operate on the identified clusters.

In general, clustering algorithms are used to group some given objects defined by a set of numerical properties in such a way that the objects within a group are more similar than the objects in different groups. Therefore, a specific clustering algorithm needs to be provided with, a criterion to measure the similarity of objects, how to cluster the objects or points into clusters. The k-means clustering algorithm uses the Euclidean distance [1,4] to measure the similarities between objects. Both iterative algorithm and adaptive algorithm exist for the standard k-means clustering. K-means clustering algorithms need to assume that the number of groups (clusters) is known a priori.

An important step in clustering is to select a distance metric, which will determine how the Similarity of two elements is calculated.

3.1 Algorithm K-means : Basic Euclidean distance metric

Let $X = \{x_1, x_2, x_3, \dots, x_n\}$ be the set of data points and $V = \{v_1, v_2, \dots, v_c\}$ be the set of centers.

1. Select 'c' cluster centers randomly.
2. Calculate the distance between each data point and cluster centers using the Euclidean distance metric as follows

$$Dist_{XY} = \sqrt{\sum_{k=1}^m (X_{ik} - X_{jk})^2}$$

3. Data point is assigned to the cluster center whose distance from the cluster center is minimum of all the cluster centers.

4. New cluster center is calculated using:

$$V_i = \left(\frac{1}{c_i} \right) \sum_1^{c_i} x_i$$

where, 'ci' denotes the number of data points in ith cluster.

5. The distance between each data point and new obtained cluster centers is recalculated.
6. If no data point was reassigned then stop, otherwise repeat steps from 3 to 5.

3.2 Algorithm K-means: Manhattan distance metric

Let $X = \{x_1, x_2, x_3, \dots, x_n\}$ be the set of data points and $V = \{v_1, v_2, \dots, v_c\}$ be the set of centers.

1. Select 'c' cluster centers randomly.
2. Calculate the distance between each data point and cluster centers using the Manhattan distance metric as follows

$$Dist_{XY} = |X_{ik} - X_{jk}|$$

3. Data point is assigned to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
4. New cluster center is calculated using:

$$V_i = \left(\frac{1}{c_i} \right) \sum_1^{c_i} x_i$$

where, 'ci' denotes the number of data points in ith cluster.

5. The distance between each data point and new obtained cluster centers is recalculated.
6. If no data point was reassigned then stop, otherwise repeat steps from 3 to 5.

3.3 Algorithm K-means: Minkowski distance metric

Let $X = \{x_1, x_2, x_3, \dots, x_n\}$ be the set of data points and $V = \{v_1, v_2, \dots, v_c\}$ be the set of centers.

1. Select 'c' cluster centers randomly.
2. Calculate the distance between each data point and cluster centers using the Minkowski distance metric as follows

$$Dist_{XY} = \left(\sum_{k=1}^d |X_{ik} - X_{jk}|^{\frac{1}{p}} \right)^p$$

3. Data point is assigned to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
4. New cluster center is calculated using:

$$V_i = \left(\frac{1}{c_i} \right) \sum_1^{c_i} x_i$$

where, 'ci' denotes the number of data points in ith cluster.

5. The distance between each data point and new obtained cluster centers is recalculated.
6. If no data point was reassigned then stop, otherwise repeat steps from 3 to 5.

Advantages

1. Fast, robust and easier to understand.
2. Relatively efficient: $O(tknd)$, where n is number of objects, k is number of clusters, d is number of dimension of each object, and t is number of iterations. Normally, $k, t, d < n$.
3. Gives best result when data set are distinct or well separated from each other.

Disadvantages

1. The learning algorithm requires a priori specification of the number of cluster centers.

2. The use of Exclusive Assignment - If there are two highly overlapping data then k-means will not be able to resolve that there are two clusters.
3. The learning algorithm is not invariant to non-linear transformations i.e. with another representation of data different results are obtained (data represented in form of Cartesian co-ordinates and polar co-ordinates will give different results).
4. Euclidean distance measures can unequally weight underlying factors.
5. The learning algorithm provides the local optima of the squared error function.
6. Randomly choosing of the cluster center cannot lead us to the fruitful result.
7. This algorithm does not work well for categorical data i.e. it is applicable only when mean is defined.
8. Unable to handle noisy data and outliers.
9. Algorithm fails for non-linear data set.

4. RESULTS AND EXPERIMENTS

Results that are obtained after the implementation of K-means using 3 various distance metrics are shown using histograms. All the experiments are performed on dummy data. The results obtained by using Euclidean distance metric i.e. basic k-means are shown in fig 4.1.

The comparative graph of distortion in both techniques is displayed in fig 4.9. Through graph it is clear that the k-means using Euclidian distance metric gives better results as compared to Manhattan k-means[4,7].

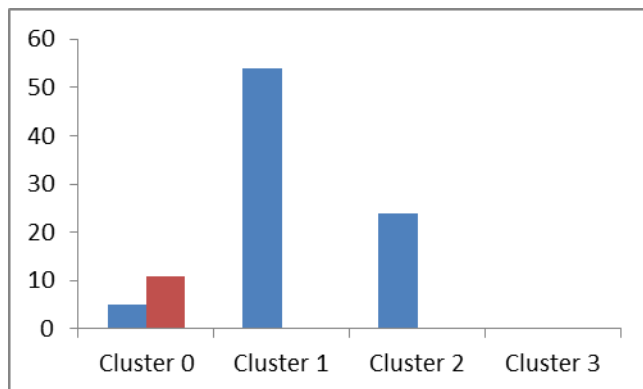


Fig: 4.1 Output Basic K-means

The results obtained by using Manhattan distance metrics are shown in figure 4.2. In both the methods the clusters have almost same values but the distortion in Manhattan k-means is more as compared to basic k-means. And it is known that minimum distortion is considered best for good clustering of data-items.

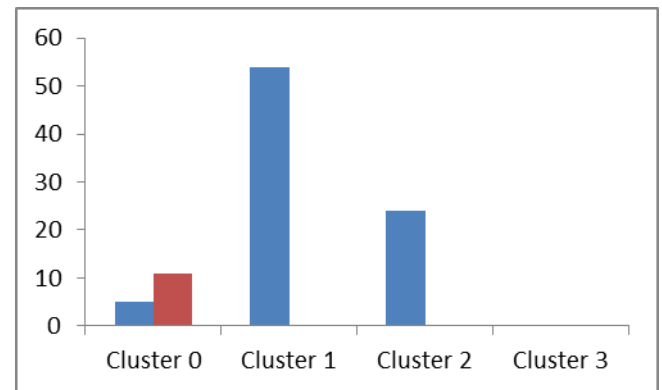


Fig: 4.2 Output Manhattan k-means

The results obtained by K-means based on Minkowski distance metrics for different values of P are displayed in figures from fig 4.3 to 4.8. And the results at P=1 are same as results using Manhattan distance metric because formula for Manhattan distance metric is derived by taking P=1. Similarly, the results at P=2 are same as results using Euclidian distance metric because formula for Euclidian distance metric is derived by taking P=2 in Minkowski distance metric formula.

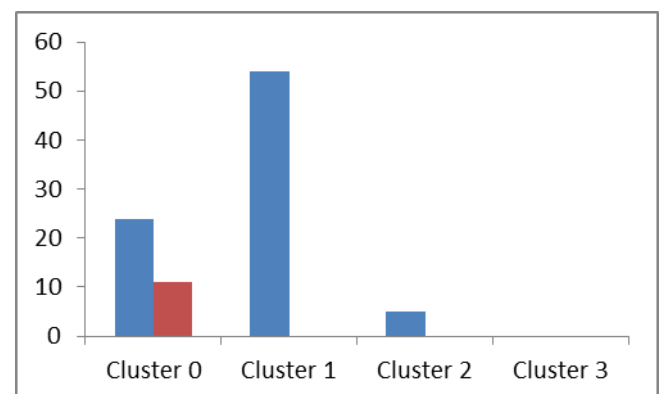


Fig: 4.3 Output Minkowski Distance at P=4



Fig: 4.4 Output Minkowski Distance at P=6

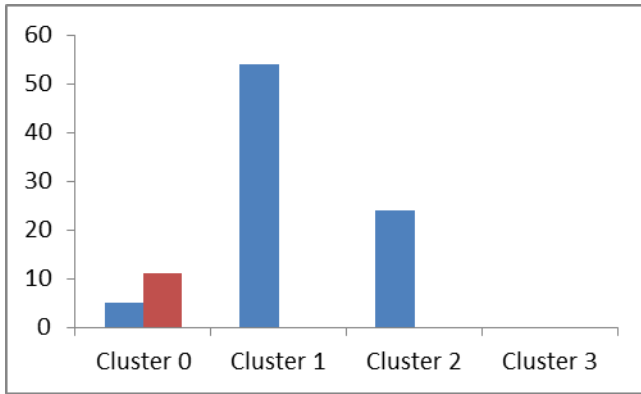


Fig: 4.5 Output Minkowski Distance at P=8

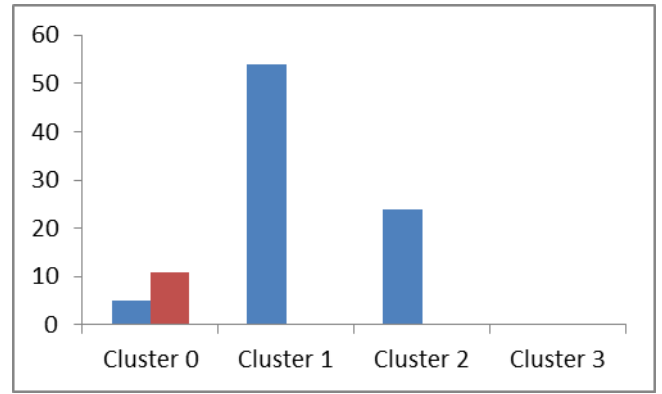


Fig: 4.8 Output Minkowski Distance at P=14



Fig: 4.6 Output Minkowski Distance at P=10

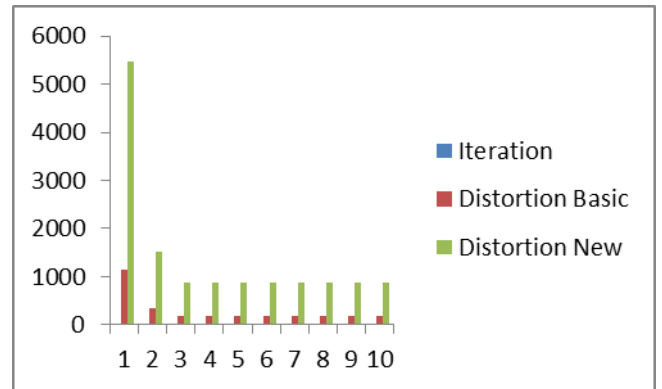


Fig: 4.9 Comparative graph of distortion in basic k-means and Manhattan K-means

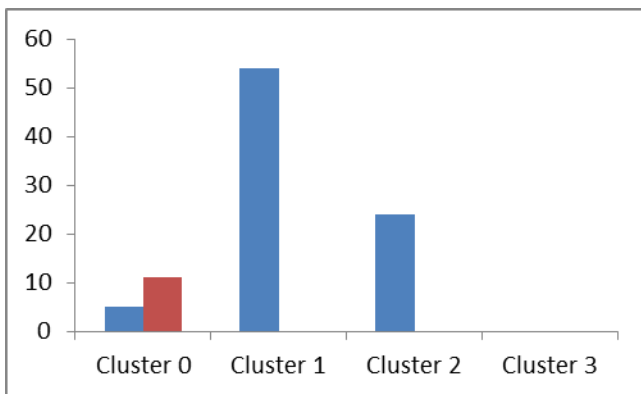


Fig: 4.7 Output Minkowski Distance at P=12

The comparative graph of distortion in K-means algorithm, using Minkowski distance metric is displayed in fig: 4.10. As the value of P increases, value of the distortion decreases and the results starts converging at P=10 and greater.

Minkowski distance metric gives similar results for larger values of P but it converges slowly as compared to basic k-means algorithm. So, in terms of performance basic k-means gives best results.

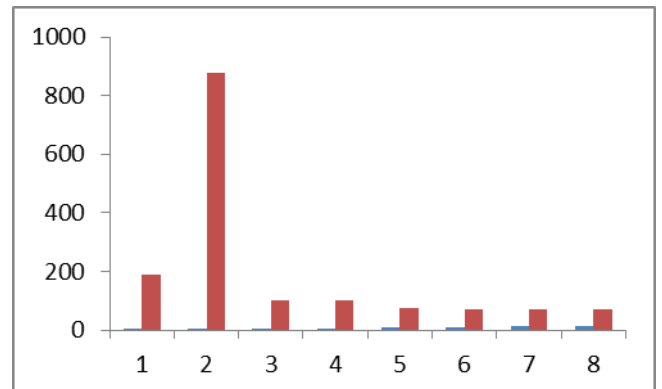


Fig: 4.10 Comparative graph of distortion in Minkowski K-means for different values of P

5. CONCLUSION

K means is a heuristic algorithm that partitions a data set into K clusters by minimizing the sum of squared distance in each cluster. During the implementation of k-means with three different distance metrics, it is observed that selection of distance metric plays a very important role in clustering. So, the selection of distance metric should be made carefully. The distortion in k-means using Manhattan distance metric is less than that of k-means using Euclidean distance metric.

As a conclusion, the K-means, which is implemented using Euclidean distance metric gives best result and K-means based on Manhattan distance metric's performance, is worst.

6. REFERENCES

- [1] Agrawal R., Faloutsos C., Swami A. *Efficient similarity search in sequence databases*. Proc. 4 Th Int. Conf. On Foundations of Data Organizations and Algorithms, 1993. – Chicago. pp. 69-84.
- [2] Archana Singh, Jyoti Agarwal, Ajay Rana January 2013. Performance Measure of Similis and FPGrowth Algorithm. *International Journal of Computer Applications (0975 – 8887) Volume 62– No.6*.
- [3] Archana Singh, Megha Chaudhary, Dr (Prof.) Ajay Rana Gaurav Dubey 2011 .Online Mining of data to Generate Association Rule Mining in Large Databases. *IEEE-International Conference on Recent Trends in Information Systems*.
- [4] *Fast Distance Metric Based Data Mining Techniques Using P-trees: k-Nearest-Neighbor classification and k-Clustering* : A Thesis Submitted to the Graduate Faculty Of the North Dakota State University.
- [5] Joaquin Perez Ortega, Ma. Del Rocio Boone Rojas and Maria J. Somodevilla Garcia. *Research issues on K-means Algorithm: An Experimental Trial Using Matlab*.
- [6] Jiawei Han and Micheline Kamber, “*Data Mining: Concepts and Techniques*,” Morgan Kaufmann Publishers, August 2000. ISBN 1-55860-489-8
- [7] Kaufman L., Rousseeuw P.J. *Finding groups in data. An introduction to cluster analysis*. – John Wiley & Sons, 2005.
- [8] Li M., Chen X., Ma B., Vitanyi P. *The similarity metric*. *IEEE Transactions on Information Theory*, 2004, vol.50, No. 12, pp.3250-3264.
- [9] Mike Estlick, Mirian Leeser, James Theiler and John Szymanski, “*Algorithmic Transformations in the Implementation of K-means Clustering on Reconfigurable Hardware*,” FPGA 2001, February 11-13, 2001, Monterey, CA, USA.
- [10] S. J. WAN, S. K. M. WONG, and P. PRUSINKIEWICZ, “*An Algorithm for Multidimensional Data Clustering*.”
- [11] S. Z. Selim and M. A. Ismail. “k-Means-Type Algorithms: A Generalized convergence theorem and characterization of the local optimality.” *IEEE transaction, Pattern Analysis Machine Intelligence*.” PAMI-6, 1, 1994, 81-87.
- [12] T.V.Vijay Kumar, Archana Singh, Gaurav Dubey 2012. Mining Queries for Constructing Materialized Views in a Data Warehouse. *Advances in Intelligent Systems and Computing* Volume 167, 2012, pp 149-159 Springer.
- [13] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman and Angela Y. Wu. *An Efficient k-means Clustering Algorithm: Analysis and Implementation*.
- [14] Tan, Steinbach, Kumar Ghosh. *The k-means algorithm - Notes*.
- [15] Vitanyi P. Universal similarity, ITW2005, Rotorua, New Zealand, 2005.
- [16] Xu R., Wunch D.C. *Clustering*. – John Wiley & Sons, 2009, 358 p.
- [17] X. Wu and I. H. Witten, “*A Fast k-Means Type Clustering Algorithm*.” Dept of Computer Science, Univ. of Calgary, Canada, May 1985.
- [18] Yair Bartal, Moses Charikary, Danny Razz, “*Approximating min-sum k-Clustering in Metric Spaces*,” Symposium on Theory of computing, July 6-8, 2001, Hersonissos, Crete, Greece.