ISSN 1691-5402
ISBN 978-9984-44-071-2
Environment. Technology. Resources
Proceedings of the 8th International Scientific and Practical Conference. Volume I1

THE CHOICE OF METRICS FOR CLUSTERING ALGORITHMS

Peter Grabusts
Rezekne Higher Educational Institution
Atbrivoshanas al. 90, Rezekne LV 4601, Latvia
Ph.: +(371)64623742, e-mail: peter@ru.lv

Abstract. *Methods of data analysis and automatic processing are treated as knowledge discovery. In many cases it is necessary to classify data in some way or find regularities in the data. That is why the notion of similarity is becoming more and more important in the context of intelligent data processing systems. It is frequently required to ascertain how the data are interrelated, how various data differ or agree with each other, and what the measure of their comparison is. An important part in detection of similarity in clustering algorithms play the accuracy in the choice of metrics and the correctness of the clustering algorithms operation.*

Keywords: *metric, clustering algorithms.*

Introduction

Nowadays the concept of *regularity or similarity* is acquiring more and more attention in the representation of intelligent data processing system operation. In many cases it is necessary to ascertain in what manner the data are interrelated, how various data differ or agree with each other, and what the measure of their comparison is. In various dictionaries the term *regularity* or similarity is interpreted as similarity, conformity with a law or conclusion by analogy. Regularity can be considered determined correctly if it explains the results of all experiments that relate to the given area of operation.

The main purpose of metric learning in a specific problem is to learn an appropriate distance/similarity function. Metric learning has become a popular issue in many learning tasks and can be applied in a wide variety of settings, since many learning problems involve a definite notion of distance or similarity [1]. A metric or distance function is a function which defines a distance between elements of a set [2, 3]. A set with a metric is called a metric space. In many data retrieval and data mining applications, such as clustering, measuring similarity between objects has become an important part. Normally, the task is to define a function Sim(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 is a function D with nonnegative real values, defined on the Cartesian product X x X of a set X. It is called a metric on X if for every x,y,z ε X:

- D(x,y)=0 if x=y (the identity axiom);
- D(x,y) + D(y,z) ≥ D(x,z) (the triangle inequality);
- D(x,y)=D(y,x) (the symmetry axiom).

A set X provided with a metric is called a metric space.

Distance metrics overview

Euclidean distance is the most common use of distance – it computes the root of square differences between coordinates of a pair of objects:

$$D_{XY} = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2} \quad (1)$$

Manhattan distance or city block distance represents distance between points in a city road grid. It computes the absolute differences between coordinates of a pair of objects:

$$D_{XY} = \sum_{k=1}^{m} |x_{ik} - x_{jk}| \quad (2)$$

Chebyshev distance is also called Maximum value distance. It computes the absolute magnitude of the differences between coordinates of a pair of objects:

$$D_{XY} = \max_k |x_{ik} - x_{jk}| \quad (3)$$

Minkowski distance is the generalized metric distance:

$$D_{XY} = \left(\sum_{k=1}^{d} |x_{ik} - x_{jk}|^{1/p}\right)^p \quad (4)$$

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

The distance measure can also be derived from the correlation coefficient, such as the Pearson correlation coefficient. Correlation coefficient is standardized angular separation by centering the coordinates to its mean value. It measures similarity rather than distance or dissimilarity:

$$r_{ij} = \frac{\sum_{k=1}^{d}(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j)}{\sqrt{\sum_{k=1}^{d}(x_{ik}-\bar{x}_i)^2 \sum_{k=1}^{d}(x_{jk}-\bar{x}_j)^2}}, \quad (5)$$

where $\bar{x}_i = \frac{1}{d}\sum_{k=1}^{d} x_{ik}$. Noticing that the correlation coefficient is in the range of [-1, 1], with 1 and -1 indicating the strongest positive and negative correlation respectively, we can define the distance measure as

$$D_{XY} = (1 - r_{ij})/2 \quad (6)$$

When using correlation coefficients for distance measures, it should be taken into consideration that they tend to detect the difference in shapes rather than to determine the magnitude of differences between two objects.

In order to illustrate different metrics the following two points will be applied: point X has coordinate (1, 2, 3, 4) and point Y has coordinate (5, 6, 7, 8).

For example, the Euclidean distance between point X and Y is:

$$D_{XY} = \sqrt{(2-3)^2 + (3-5)^2 + (4-7)^2 + (5-9)^2} = 5{,}5$$

The Manhattan distance between point X and Y is:

$$D_{XY} = |2-3| + |3-5| + |4-7| + |5-9| = 10$$

The Chebyshev distance between point X and Y is:

$$D_{XY} = max\{|2-3|, |3-5|, |4-7|, |5-9|\} = max\{1,2,3,4\} = 4$$

The Minkowski distance of order 3 between point X and Y is:

$$D_{XY} = (|2-3|^3 + |3-5|^3 + |4-7|^3 + |5-9|^3)^{1/3} = \sqrt[3]{100} = 4{,}6$$

The correlation coefficient between point X and Y. The mean value of each object is:

$$\bar{x}_X = \frac{1}{4}(2+3+4+5) = 3{,}5$$

$$\bar{x}_Y = \frac{1}{4}(3+5+7+9) = 6$$

$$r_{XY} = \frac{(2-3{,}5)(3-6)+(3-3{,}5)(5-6)+(4-3{,}5)(7-6)(5-3{,}5)(9-6)}{([(2-3{,}5)^2+(3-3{,}5)^2+(4-3{,}5)^2+(5-3{,}5)^2][(3-6)^2+(5-6)^2+(7-6)^2+(9-6)^2])^{0{,}5}} = \frac{4{,}5+0{,}5+0{,}5+4{,}5}{\sqrt{5\times 20}} = 1$$

The summary of the metrics under consideration is shown in Table 1.

Table 1.

Proximity measures and their applications

Measure	Metric	Examples and applications
Euclidean distance	Yes	K-means with its variants
Manhattan distance	Yes	Fuzzy ART, clustering algorithms
Chebyshev distance	Yes	Fuzzy C-means clustering
Minkowski distance	Yes	Fuzzy C-means clustering
Pearson correlation	No	Widely used as the measure for microarray gene expression data analysis

Cluster analysis method

As a data mining function, clustering can be used as a standalone tool to gain insight into the distribution of data, to observe the characteristics of each cluster, and to focus on a particular set of clusters for further analysis [4, 5]. Clustering is one of the most fundamental issues in data recognition. It plays a significant role in searching for structures in data. It may serve as a pre-processing step for other algorithms, which will operate on the detected 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 particular clustering algorithm needs to be given a criterion to measure the similarity of objects, how to cluster the objects into groups. The k-means clustering algorithm uses the Euclidean distance to measure the similarities between objects [5]. 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.

The standard k-means clustering algorithm is a general clustering algorithm to cluster *N* objects into *m* groups of the given number *m*. The method minimizes the total squared Euclidean distance D of the form:

$$D = \sum_{i=1}^{N} \sum_{j=1}^{m} M_{ij} \left\| x_i - c_j \right\| \qquad (6)$$

where x_i ; i=1, 2, … *N* are the *N* objects and c_j ; j=1, 2, … *m* are the *m* centres. M_{ij} is the cluster membership function, which is defined by N×m matrix of 0's and 1's with exactly one 1 per row that identifies the groups to which a given object belongs. In this algorithm, the similarity of objects is defined by the Euclidean distance: the smaller distance between two objects is, the more similar they are. The algorithm works in this way. At the beginning, the *m* centres c_j are set to some initial data points. If the training data was not ordered in any way, the first *m* training data is usually chosen as the initial set of function centres. Otherwise, *m* data points would be selected randomly. At step 2, each of the training patterns is assigned to the closest centre. At step 3, the centres are adjusted by taking the arithmetic average in each cluster group. Steps 2 and 3 will be repeated until each training pattern stays in its group, i.e., no reassignment of any pattern to a different group than its previous group (see Table 2.).

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

Drawbacks of k-means clustering

The term «cluster analysis» actually comprises a set of different classification algorithms. A common question frequently asked by researchers is: how to organize the data observed into clear structures? A viewpoint exists that unlike many other statistical procedures, methods of cluster analysis are commonly used when the researcher has not got any prior hypotheses

regarding classes but is still at the descriptive stage of investigation. It should be noted that cluster analysis determines the most possible meaningful decision [6].

Table 2.

K-means clustering procedure

Step 1. Initialise the function centres Set the initial function centres to the first m training data or to the m randomly chosen training data. Step 2. Group all patterns with the closet function centre For each pattern x_i, assign x_i to group j*, where $\left\| x_i - c_{j^*} \right\| = \min_j \left\| x_i - c_j \right\|$ Step 3. Compute the sample mean for the function centre For each group c_j, $c_j = \frac{1}{m_j} \sum_{x_i \in group\ j} x_i$ where m_j is the number of patterns in group j. Step 4. Repeat by going to step 2, until no change in cluster assignments

Cluster analysis is used to automatically generate a list of patterns by a training set. All the objects of this sample are presented to the system without the indication to which pattern they belong. The cluster analysis is based on the hypothesis of compactness. It means that methods of cluster analysis enable one to divide the objects under investigation into groups of similar objects frequently called clusters or classes. Given a finite set of data *X*, the problem of clustering in *X* is to find several cluster centres that can properly characterize relevant classes of *X*. In classic cluster analysis, these classes are required to form a partition of *X* such that the degree of association is strong for data within blocks of the partition and weak for data in different blocks.

Similar to other clustering algorithms, k-means clustering has many drawbacks:

- Cluster number, k, must be determined beforehand.
- It is difficult to determine the contribution each attribute makes to the grouping process, since it is assumed that each attribute has the same weight.
- By using the same data, we may never know the real cluster. If the number of data is a few, by inputting data in a different order, a result may be a different cluster.
- In case there are not many numbers of data, the cluster will be significantly determined by the initial grouping.
- Weakness of arithmetic mean is not robust to outliers. As a result, the centroid may be pulled away from the real data by very far data.
- It is sensitive to initial condition, since different initial condition may lead to different result of cluster. The algorithm may be trapped in the local optimum.
- As a result one gets a circular cluster shape which is based on distance.

Materials and methods

The purpose of the experimental part was to test the operation of the k-means algorithm by applying different metrics. Three different metrics have been chosen: Euclidean distance, Manhattan distance and Pearson correlation. In the course of the experiments in order to determine cluster centres in k-means clustering algorithm sequentially all three metrics have been used. The results obtained have been analyzed and the clustering correctness has been tested.

During the experiment the well-known Fisher's IRIS data set was employed [7], containing three species classes of 50 elements each: setosa, versicolor and virginica. Each species has

four attributes: SL - sepal length, SW - sepal width, PL - petal length, PW - petal width. However, it is uncommon to use this data set in cluster analysis, since the data set contains only two clusters with rather obvious separation. One of the clusters contains the Iris setosa species, while the other cluster contains both Iris virginica and Iris versicolor and is not separable without the species information Fisher used.
The experimental part has been carried out in Matlab environment [8].

Results and discussion

The results of the experiments are shown in Table 3.

Table 3.

Clustering results by applying different metrics

Distance	Euclidean	Manhattan	Correlation
Cluster centres	50.06 34.28 14.62 2.46 68.50 30.74 57.42 20.71 59.02 27.48 43.94 14.34	50 34 15 2 57 27 42 13 65 30 54 19	0.68 0.24 -0.29 -0.63 0.62 -0.35 0.34 -0.61 0.69 -0.23 0.20 -0.66
Cluster1 contains:	Records from cluster 1 – **50** Records from cluster 2 – 0 Records from cluster 3 – 0	Records from cluster 1 – **50** Records from cluster 2 – 0 Records from cluster 3 – 0	Records from cluster 1 – **50** Records from cluster 2 – 0 Records from cluster 3 – 0
Cluster2 contains:	Records from cluster 1 – 0 Records from cluster 2 – **48** Records from cluster 3 – **2**	Records from cluster 1 – 0 Records from cluster 2 – **39** Records from cluster 3 – **11**	Records from cluster 1 – 0 Records from cluster 2 – **47** Records from cluster 3 – **3**
Cluster3 contains:	Records from cluster 1 – 0 Records from cluster 2 – **14** Records from cluster 3 – **36**	Records from cluster 1 – 0 Records from cluster 2 – **4** Records from cluster 3 – **46**	Records from cluster 1 – 0 Records from cluster 2 – **3** Records from cluster 3 – **47**
Correct-ness:	For cluster 1 – **100 %** For cluster 2 - **96 %** For cluster 3 - **72 %**	For cluster 1 – **100 %** For cluster 2 - **78 %** For cluster 3 - **92 %**	For cluster 1 - **100 %** For cluster 2 - **94 %** For cluster 3 - **94 %**

The above table shows that all metrics correctly recognize cluster 1 records. Cluster 2 records are best recognized by Euclidean distance, whereas cluster 3 records – by correlation. The following figure in the form of a chart shows potentialities of different metrics in clustering (see Fig. 1.).

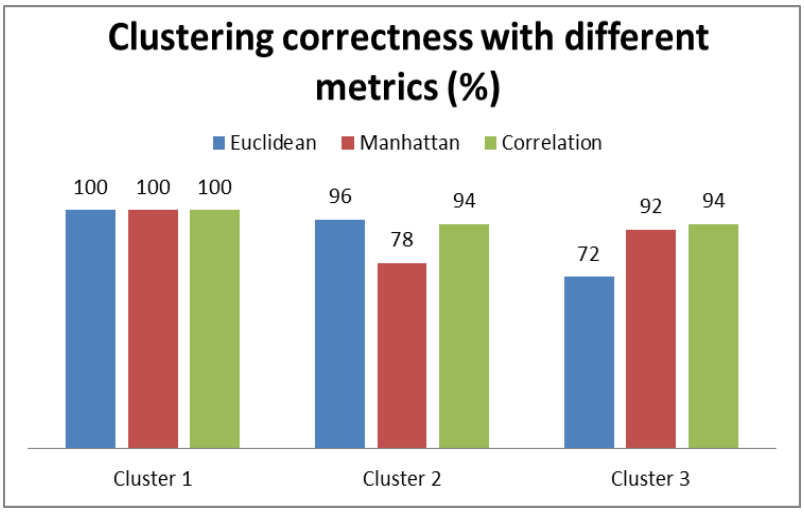

Fig.1. **Clustering correctness**

The visualization of clustering may be useful when analyzing results. For data visualization purposes, 2D projections can be used showing the distribution of particular parameters with respect to each other, while dendrogram graphs are normally used for visualization of the formation of clusters (see Fig.2., Fig.3. and Fig.4.).

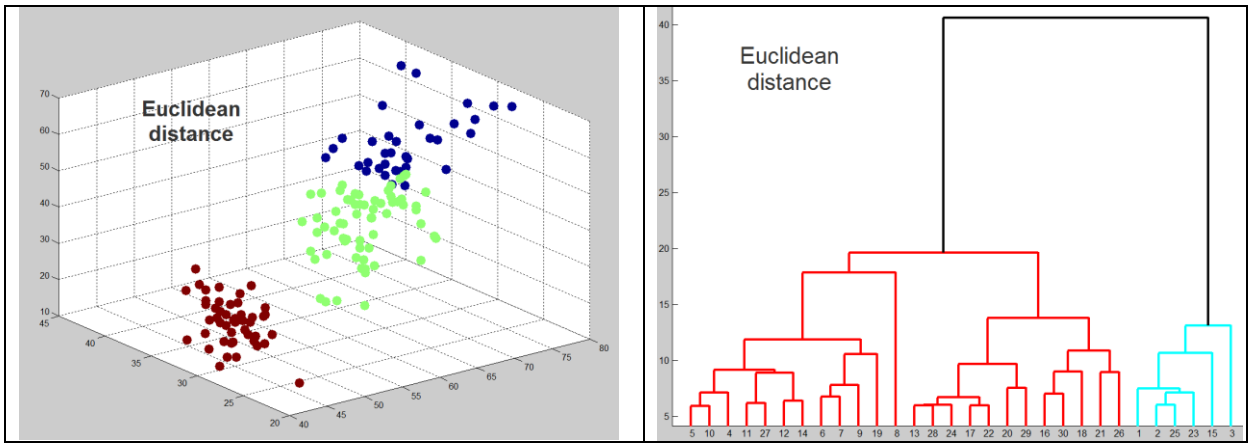

Fig.2. **Clustering results for Euclidean distance**

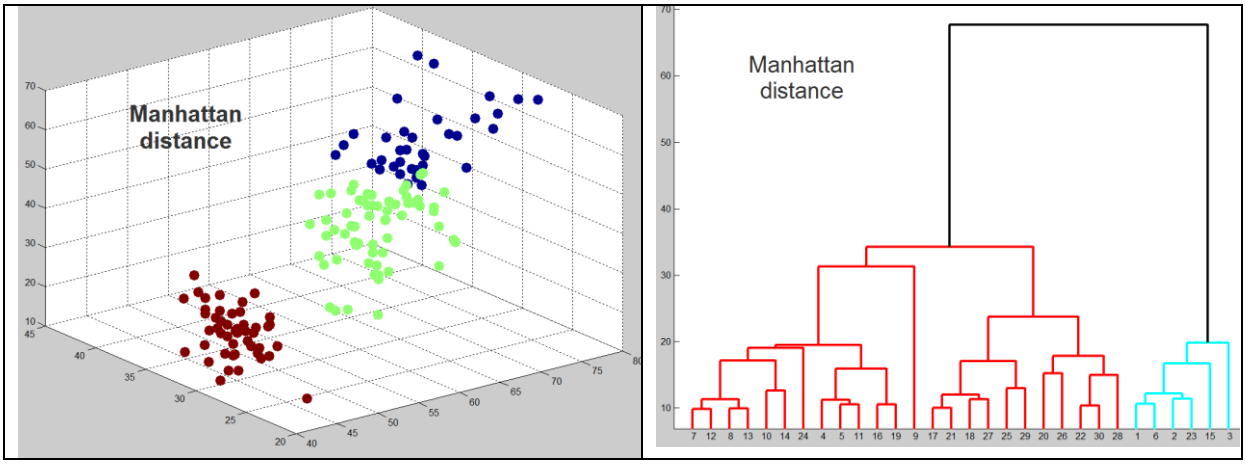

Fig.3. **Clustering results for Manhattan distance**

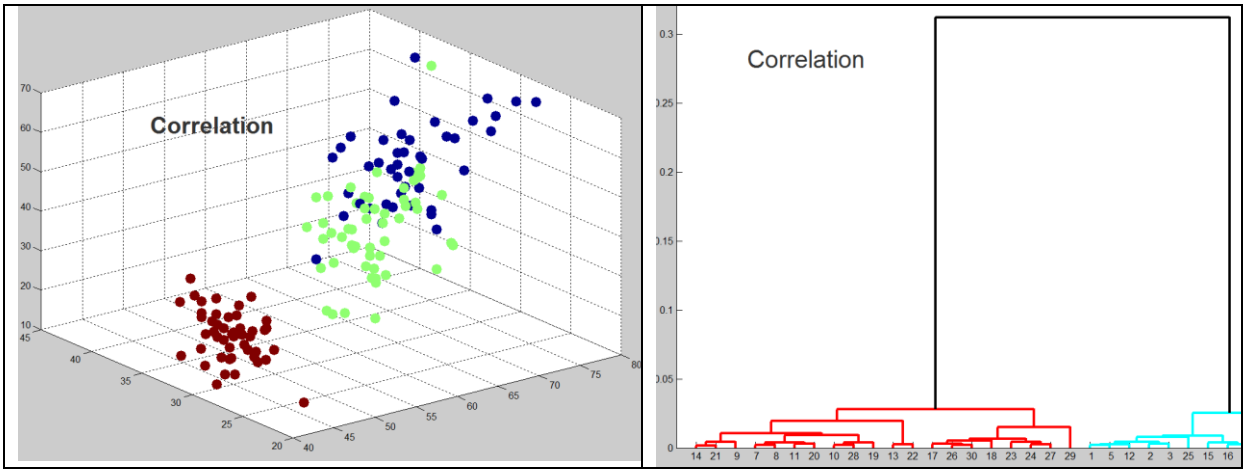

Fig.4. **Clustering results for Correlation**

Conclusions

Based on the tables and figures it can be concluded that the results obtained by applying all three metrics are very similar. None of the metrics shows dominance that could allow to consider it as the best metric. Traditionally Euclidean distance is used in clustering algorithms, however, the choice of other metric in definite cases may be disputable. It depends on the task, the amount of data and on the complexity of the task.

Further research will be dedicated to the analysis of the evaluation of clustering results' correctness.

References

1. Agrawal R., Faloutsos C., Swami A. Efficient similarity search in sequence databases. Proc. 4th Int. Conf. On Foundations of Data Organizations and Algorithms, 1993. – Chicago. pp. 69-84.
2. Li M., Chen X., Ma B., Vitanyi P. The similarity metric. IEEE Transactions on Information Theory, 2004, vol.50, No. 12, pp.3250-3264.
3. Vitanyi P. Universal similarity, ITW2005, Rotorua, New Zealand, 2005.
4. Kaufman L., Rousseeuw P.J. Finding groups in data. An introduction to cluster analysis. – John Wiley & Sons, 2005.
5. Xu R., Wunch D.C. Clustering. – John Wiley & Sons, 2009, 358 p.
6. Everitt B.S. (1993). Cluster analysis. Edward Arnold, London, 170 p.
7. Fisher R.A. The use of multiple measurements in taxonomic problems. Ann. Eugenics, 1936,7(2), p.179-188.
8. http://www.mathworks.com/

Anotācija

Metrikas izvēle klasterizācijas algoritmiem

Intelektuālās analīzes datortehnoloģijas patlaban piedzīvo uzplaukuma periodu. Tas galvenokārt saistīts ar jaunu ideju realizēšanos vairāku zinātņu nozaru saskarsmes punktos tādos kā mākslīgais intelekts, statistika, datu bāzu metodes u.c. Daudzos gadījumos ir nepieciešams kaut kādā veidā klasificēt datus vai atrast likumsakarības tajos, tāpēc jēdziens „likumsakarība" iegūst arvien lielāku nozīmi intelektuālās datu apstrādes sistēmu kontekstā. Bieži ir nepieciešams noskaidrot – kādā veidā dati ir saistīti savā starpā, kāda ir dažādu datu līdzība vai atšķirība, kāds ir šo datu salīdzināšanas mērs. Tādiem nolūkiem var izmantot dažādus klasterizācijas algoritmus, kas datus sadala grupās pēc noteiktiem kritērijiem – metrikas. Ar metriku šajā kontekstā tiek saprasta distance (attālums) starp klasterā ietilpstošajiem punktiem.

Darbā tika pārbaudīta klasiskā klasterizācijas algoritma k-means darbība ar dažādām metrikām: Eiklīda attālumu, Manhetenas attālumu un Pīrsona korelācijas koeficientu. Eksperimentu gaitā k-means klasterizācijas algoritmā klasteru centru noteikšanai secīgi tika izmantotas minētās trīs metrikas. Iegūtie rezultāti tika analizēti un tika pārbaudīts klasterizācijas korektums. Tradicionāli klasterizācijas algoritmos izmanto Eiklīda attālumu, taču citas metrikas izvēle atsevišķos gadījumos var būt diskutējama. Tas atkarīgs no risināmā uzdevuma, datu apjoma un sarežģītības. Tika konstatēts, ka klasterizācijas rezultāti visu triju metriku izmantošanā ir ļoti līdzīgi. Nevienai no izvēlētajām metrikām nebija izšķirīga pārsvara, kas varētu garantēti pasludināt to par labāko.