UDC 004

METHODS AND MEASURES SIMILARITIES ALGORITHM FOR FINDING A COMPLEX ARRAY

ALEXANDR ZALESKI, VALERY CHERTKOV
Polotsk State University, Belarus

This work describes the basic methods and algorithms for determining the measure of similarity. We identified two promising methods for determining the similarity measure of polynomials, namely the function of the correlation coefficient and functions based on the Hausdorff metric. The analysis of the main measure distances (metrics) is represented.

.

Cluster analysis is the common name of the computational procedures set used in the creation of the classification. As a result of the procedures the group which is very "similar" objects or "clusters" has been formed. More specifically, the cluster method is a multivariate statistical procedure, which collects data containing information on a sample of objects, and then arranges the objects in a relatively homogeneous group [1].

Examples of the cluster analysis use of are: informatics - simplifying information, data visualization, image segmentation, intelligent search. Economy is the analysis of markets and financial flows, elimination of laws on the stock exchanges. Marketing is market segmentation, analysis of consumer behavior, positioning products. Astronomy is the selection of stars and galaxies groups, the automatic processing of satellite images. Security systems are the recognition of biometric data. Applied Image Processing System is the machine vision system for technological processes, systems for diagnostics and condition monitoring objects, search and recognition of objects for processing medical images.

For the data cluster analysis similarity measure is used. There are four types of the subsections similarity: the correlation coefficients; distance measures; associative factors and probability coefficients of similarity. Although all similarity types of were used at one time, only the correlation coefficients and the distances are widely used [2].

1. The correlation coefficient is a measure of the mutual influence nature that indicates changes between two random variables. Defined by the expression (1):

$$r_{jk} = \frac{\sum_{i=1}^{n}\left(x_{ij}-\overline{x_i}\right)\left(x_{ik}-\overline{x_k}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_{ij}-\overline{x_i}\right)^2\sum_{i=1}^{n}\left(x_{ik}-\overline{x_k}\right)^2}}, \quad (1)$$

where x_{ij} – the value of i for the j variable object; x_j – the average of all values of the variables *j* object; *n* – number of variables.

Correlation coefficient ranges from –1 to +1, where zero indicates that there is no communication between objects.

The main drawback of the correlation coefficient as the similarity measure is that it is sensitive to the form by reducing the sensitivity to the magnitude of the differences between the variables. Furthermore, the correlation calculated in this manner has no statistical sense [3].

Despite these shortcomings, the coefficient is widely used in the cluster analysis applications. Hammer and Cunningham have shown that with proper use of cluster correlation coefficient method is superior to other similarity factors as to reduce the number of incorrect classifications.

2. Mery distance (metric) is widely popular. Two objects are identical when describing their variables take the same value. In this case, the distance between them is zero. Distance measures dependent on the choice of scale measurements and usually not bounded above [4]. One of the most famous distance is the Euclidean distance, defined as (2):

$$d_{ij} = \sqrt{\sum_{k=1}^{p}\left(x_{ik}-x_{jk}\right)^2} \quad (2)$$

where d_{ij} – the distance between objects iand j; x_{ik} – the value of the variable kfor the jobject,

$$d_{ij} = \sum\left|\left(x_{ik}-x_{jk}\right)\right| \ .$$

To give higher weights more distant from one another objects using squared Euclidean distance.

A well-known measure is also Manhattan distance or "taxicab geometry" (city-block) (3).

$$d_{ij} = \left(\sum_{k=1}^{p} \left|\left(x_{ik} - x_{jk}\right)\right|^{\gamma}\right)^{t/\gamma} \tag{3}$$

You can define other metrics, but most of them are private forms except from a special class of metric distance functions, known as Minkowski metric.

There distances are not the Minkowski metric, and the most important of them - Mahalanobis distance D^2. The expression of the metric (4):

$$d_{ij} = \left(x_i - x_j\right)^T S^{-1}\left(X_i - X_j\right) \tag{4}$$

where S – totalintra - variance - covariance matrix, x_i and x_j – vectors variables for objects I and *j*.

In contrast to the Minkowski metric and Euclidean, this metric is related to the correlations of the variables with the help of the covariance matrix of dispersions.

Function based on the Hausdorff metric.(5)

$$R^H = 1 - \frac{1}{l}\max_{ij}\left|o_{ij} - b_{ij}^*\right| \tag{5}$$

where $i \in 0...N-1, j \in 0...N-1$

Lack of distance measures is that the similarity score is strongly dependent on differences in the data shifts. Moreover, the metric distances vary under the influence of variables measuring scale transformations.

3. The coefficients of associativity apply when it is necessary to establish the similarity between objects described by binary variables, where 1 indicates the presence of a variable, and 0 – in its absence. There are three measures that are commonly used: simple coefficient, Jaccard coefficient and the Gower coefficient.

4. Probabilistic similarity coefficients. The great difference of this type described above lies in the fact that, the similarity between two objects is not calculated. The formation of clusters is computed information gain from the merger of two objects, and those associations that give a minimum prize shall be treated as a single object. Probability measures are only suitable for binary data and are attached directly to the source data prior to processing. What are some things discover each other similarity or difference, is a very important moment for the classification process.

Similarities

The problem is not a simple recognition of similar or dissimilar things, and in what place they occupy in the concept of scientific research. Today, cluster analysis is one of the most efficient large amounts of data processing tools and is used all over the place, which applies computer technology [5].

Algorithms squared error

The task of clustering can be regarded as the construction of the optimal partition of objects into groups. This optimality can be defined as a requirement to minimize the mean square error of the partition (6):

$$e^2(X, L) = \sum_{j=1}^{K}\sum_{i=1}^{n_j}\left\|x_i^{(j)} - c_j\right\|^2 \tag{6}$$

where c_j-"center of mass" of the cluster j (point with the average values for the characteristics of the cluster).

Algorithms square error related to the type of flat algorithms. The most common search algorithm is k-means. This algorithm builds a predetermined number of clusters located as far as possible from each other. The algorithm is divided into several stages:

1. Randomly select k points, which are the initial "centers of mass of" clusters.
2. Classify each object to the cluster with the closest "center of mass".
3. Calculate the "centers of mass" of clusters according to their current structure.
4. If the stopping criterion of the algorithm is not satisfied, return to the n. 2.

As a stopping criterion of the algorithm is typically selected minimum change in the mean square error. It is also possible to stop the algorithm if step 2 was not the objects move from cluster to cluster. The disadvantages of this algorithm include the need to specify the number of clusters to split. [3]

The most promising method for finding similarity measures polynomials are a function of the correlation coefficient and the function based on the Hausdorff metric.

1. Finding a similarity measure by the correlation coefficient. Formula features presented in the article number 1.

Its popularity is method to two factors: the correlation coefficients are relatively easy to count, their use requires no special mathematical training. In combination with the ease of interpretation, the ease of use factor has led to its widespread use in the analysis of statistical data.

Coefficient is shown in figure 1.

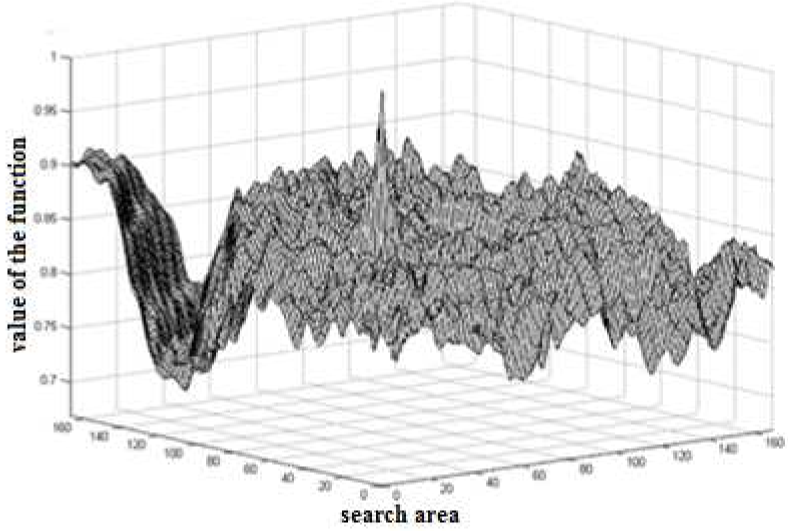

Fig. 1. The graph of the correlation function

2. Finding a similarity measure by a function based on the Hausdorff metric. The function presented in the article number 6.

Hausdorff metric used as a measure of similarity, are very sensitive to noise. Therefore, in itself does not guarantee symmetric rigor of the evaluation itself similarities, but at the same time measure satisfying combination of other assets, but not a metric, can in most cases give a better estimate of the similarities than the other measure, which is a metric [6].

Hausdorff metric is shown in Figure 2.

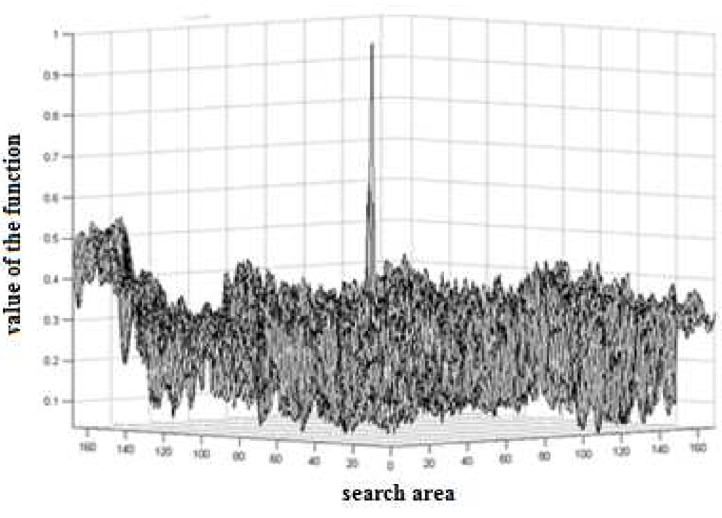

Fig. 2. Schedule Hausdorff metric function

Conclusion. In this scientific article we discussed the main methods for determining similarity measures, and identified two promising method for determining the similarity measure on the basis of two polynomials function of the correlation coefficient and the Hausdorff metric. The selection of these methods was due to the implementation of the criteria to measures computes the similarity: symmetric, be normalized values average rating, resistance to noise, the monotony, the speed calculation of similarity measures.

REFERENCES

1. Воронцов, К.В. Алгоритмы кластеризации и многомерногошкалирования. Курс лекций / К.В. Воронцов. – М. : МГУ, 2007. – 170 с.
2. Котов, А. Кластеризация данных / А. Котов, Н. Красильников. – 2006. – 213 с.
3. Чубукова, И.А. Курс лекций «DataMining» / И.А. Чубукова // Интернет-университет информационных технологий – Режим доступа: www.intuit.ru/department/database/datamining. – Дата доступа: 20.01.2017г.
4. Васильев, Н. Метрические пространства / Н. Васильев. – Квант, 1990. – 326 с.
5. Харин, Ю.С. Теория вероятностей, математическая и прикладная статистика: учебник / Ю.С. Харин, Н.М. Зуев, Е.Е. Жук. – Минск : БГУ, 2011. – 464 с.
6. Метрические пространства. – Режим доступа: http://dfgm.math.msu.su/files/ivanov-tuzhilin/2014-2015/METRGEOM2014-1.pdf. – Дата доступа: 20.01.2017.