UDC 336.774.3

EVALUATION OF INDICATORS CLUSTERING METHODS IN INTERRELATED COUNTERPARTIES FINANCIAL ASSESSMENT

D.M. Rahel
PhD in Economics, Associate Professor of the Department of Economics of the Belarusian State University of Informatics and Radioelectronics
ragel@bsuir.by

D.M. Rahel
In 2000 he graduated from the Faculty of Economics of the Belarusian State University of Informatics and Radioelectronics with a degree in Economic Informatics. In 2016 he graduated from the postgraduate course of the Academy of Management under the President of the Republic of Belarus. In 2018 he defended his PhD thesis. Research interests: data mining in marketing, process modeling, data analysis, statistical forecasting, macroeconomics.

Annotation. The article discusses approaches to clustering and classification of financial data that can be used in scoring indicators of various counterparties. The author considers two types of clustering of financial data using the Euclidean distance and the Mahalanobis distance and draws conclusions about their advantages and disadvantages.

Keywords: data analysis, financial data, big data, data array, scoring, analytics, data distribution, financial characteristics, financial analysis, clustering, Euclidean distance, Mahalanobis distance.

When evaluating counterparties, there is often a need to classify and search for a dependent in the data sets they provide. There are quite a large number of methods aimed at classifying and distributing large volumes of data. In this case, we will compare the reliability of the results that data clustering provides based on calculating the Euclidean distance between the data in the set under consideration and the results that can be obtained using the agglomerative clustering algorithm, which is calculated using the Mahalanobis distance formula.

Initial data set. The data set used contained data on 50 counterparties, which we need to classify and divide into groups depending on the values of the indicators. Thus, we have a table with 50 records, each of which contains the parameter X1 – the amount of revenue from sales and X2 – the amount of net profit. Based on these indicators, we need to classify the legal entities in question, which belong to the same market, and their business activities have the same scale according to preliminary estimates.

Segmentation of indicators based on Euclidean distance. The peculiarity of the algorithm is the search for the minimum distance between objects and on this basis the formation of similar groups of objects. The following characteristics were used to configure the algorithm: 2 clusters (based on empirical assessment), iterations are stopped at the moment of obtaining equivalent values of distances between groups.

Listing 1. Code for calculating Euclidean distance using nearest neighbor method

```
# Calculating Euclidean distance
matrix = pdist(X, metric='euclidean')
# Nearest Neighbor Method
link_matrix = sch.linkage(matrix, method='single')
# Number of clusters, in this case, 2 clusters
```

```
c = fcluster(link_matrix, t=2, criterion='maxclust')
```

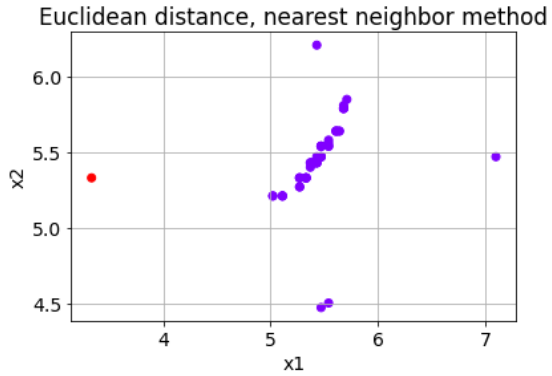

Figure 1. Results of data clustering based on Euclidean distance using the nearest neighbor method

Figure 1 shows the clusters formed based on the algorithm under consideration; as a result of the implementation, no clear cluster groups were formed and it is impossible to draw conclusions about the essential features and interrelationships of the counterparties considered in the analysis.

When implementing the algorithm for calculating the Euclidean distance based on the search for a distant neighbor, we see a similar picture of the final clustering of counterparties (Figure 2).

Listing 2. Program code for calculating the Euclidean distance using the distant neighbor method

```
# Calculating Euclidean distance
matrix = pdist(X, metric='euclidean')
# Distant Neighbor Method
link_matrix = sch.linkage(matrix, method='complete')
# Number of clusters, in this case, 2 clusters
c = fcluster(link_matrix, t=2, criterion='maxclust')
```

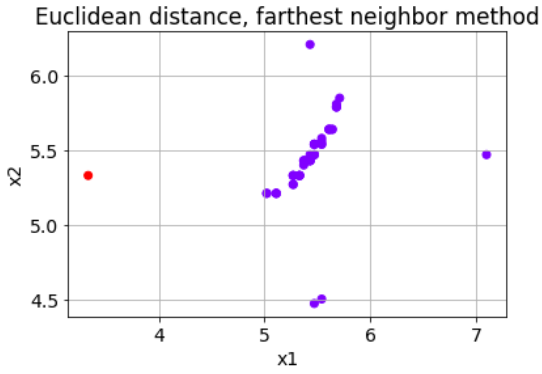

Figure 2. Results of data clustering based on Euclidean distance using the farthest neighbor method

Data segmentation based on Mahalanobis distance. When using the agglomerative clustering algorithm, which assumes the correlation between the data under consideration and, due to this, is invariant to the scale of the data volume under consideration, we obtained a picture that is more definite (Figure 3).

Listing 3. Program code for calculating the Mahalanobis distance

```
# Calculate the covariance matrix
cov_matrix = np.cov(X.T)
# Calculate the inverse covariance matrix
icov_matrix = np.linalg.inv(cov_matrix)
# Function to calculate Mahalanobis distance
```

```
def mah_dist(x, y, icov_matrix):
    d = x - y
    return np.sqrt(np.dot(np.dot(d, icov_matrix), diff.T))
# Calculate the Mahalanobis distance matrix between all pairs of points
dist_matrix = cdist(X, X, metric=lambda u, v: mah_dist(u, v, icov_matrix))
# Hierarchical Agglomerative Clustering Using Nearest Neighbor Method
c = linkage(dist_matrix, method='single')
```

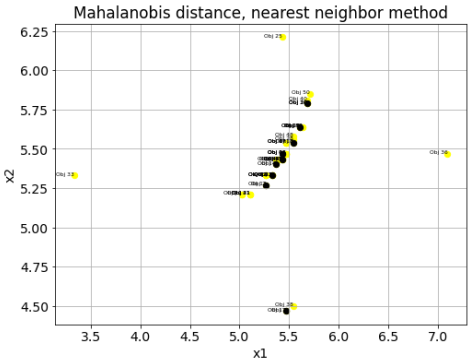

Figure 3. Results of data clustering based on Mahalanobis distance using nearest neighbor method

In this case, we see a clearer division of groups, similar companies were in the same cluster, which allows us to make initial conclusions about their connectivity and specify further analytical procedures. From this point of view, this type of algorithm is more effective in implementing scoring procedures.

Conclusion. During the study of the features of assessing the connectivity of counterparties, we considered two types of algorithms - based on the calculation of the Euclidean distance and agglomerative algorithms with the calculation of the Mahalanobis distance. The second type of algorithms, which takes into account the correlation estimates of indicators, showed its higher efficiency in implementing scoring procedures for assessing counterparties.

References

[1] Cainiao Smart Logistics [Электронный ресурс]. Режим доступа: https://www.cainiao.com/en/index.html. – Дата доступа: 12.01.2025

[2] Data Science and Big Data Analytics. A Step by Step Guide to learn Data Science from Scratch with Python Machine Learning and Big Data / Andrew Park. – Published by Andrew Park, 2021. – 124 p.

[3] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. Applied Logistic Regression // John Wiley & Sons. - 2013. - Vol. 398, No. 56. – P. 38–60.

[4] Nicholson W.L. Exploring Data Analysis. – Nobel Press, 2012. – 421 p.

ОЦЕНКА МЕТОДОВ КЛАСТЕРИЗАЦИИ ПОКАЗАТЕЛЕЙ ДЛЯ ФИНАНСОВОЙ ОЦЕНКИ ВЗАИМОСВЯЗАННЫХ КОНТРАГЕНТОВ

Д.М. Рагель
к.э.н., доцент кафедры экономики
Белорусского государственного
университета информатики и
радиоэлектроники

Аннотация. В статье изложен подход к анализу характера распределения различных типов финансовых данных. Предлагается подход, при котором можно проанализировать совокупность однотипных данных на основании построения ряда распределения и анализа значений в различных интервалах ряда. Характер распределения значений позволит сделать выводы о нормальности деятельности изучаемых контрагентов и, с учетом этого, провести дальнейший анализ данных.

Ключевые слова: анализ данных, финансовые данные, большие данные, массив данных, прогнозирование, аналитика, распределение данных, финансовые характеристики, финансовый анализ, кластеризация, Евклидово расстояние, расстояние Махаланобиса.