Аннотация. *В статье на примере 50 однотипных контрагентов рассматриваются основные подходы к кластеризации схожих данных, характеризующих хозяйственную деятельность этих субъектов. Сравнение проводится между базовыми алгоритмами кластеризации с использованием евклидова расстояния и типовым агломеративным алгоритмом на основе расстояния Махаланобиса, который учитывает ковариационные оценки элементов из приведенного набора. Исследуется последовательное применение разных типов алгоритмов в отношении компаний, а также изменение зависимости при применении различных аналитических процедур. По итогам анализа делаются выводы о преимуществах и недостатках методов, отмечается целесообразность последовательного применения алгоритмов различного типа для выявления скрытых причин, которые оказывают влияние на деятельность юридических лиц и не являются очевидными при проведении стандартного финансового анализа работы предприятия.*

Ключевые слова: *анализ данных, финансовые данные, большие данные, массив данных, прогнозирование, аналитика, распределение данных, финансовые характеристики, финансовый анализ, банковский скоринг, кредитная политика, управляющая компания, дочерняя компания, кластеризация, евклидово расстояние, расстояние Махалонобиса, связанные контрагенты, скоринг, скоринговые процедуры, анализ связанности.*

Для цитирования: *Рагель Д. Методы кластеризации показателей при финансовой оценке взаимосвязанных контрагентов // Наука и инновации. 2025. №10. С. 66–69. https://doi.org/10.29235/1818-9857-2025-10-66-69*

Дмитрий Рагель,
доцент кафедры экономики Белорусского государственного университета информатики и радиоэлектроники, кандидат экономических наук;
ragel@mail.ru

Методы кластеризации показателей при финансовой оценке взаимосвязанных контрагентов

УДК 336.774.3

При проведении анализа эффективности работы контрагентов зачастую возникает необходимость в классификации и поиске зависимых показателей в предоставленных наборах финансовых данных, которые могли бы указывать на связанность различных компаний, что является существенным фактором оценки их результативности. При этом на основании чисто финансовых инструментов его непросто обнаружить из-за высокой трудоемкости аналитических операций. Менее сложно и более эффективно применение процедур нефинансового характера в сочетании со стандартными инструментами анализа хозяйственной деятельности субъектов хозяйствования.

Существует достаточно большое количество методов, ориентированных на распределение массивов однотипной информации, которые указывают на имеющиеся зависимости и помогают оценить их характер. Однако такой тип методов не очень распространен, так как не представляет собой исключительно финансовый инструмент и по этой причине не дает однозначных выводов, характеризующих экономическую деятельность компаний. В то же время на основании общих классификационных подходов можно сделать выводы о взаимосвязи субъектов или разбить их на определенные категории для осуществления дальнейшего анализа. Для этого предусмотрено достаточно большое количество методик, но в данном случае сравним два наиболее распространенных типа и установим достоверность результатов с применением расчета евклидова расстояния между имеющимися данными с использованием алгоритма агломеративной кластеризации, рассчитываемого по формуле Махаланобиса. В результате сравнения необходимо понять и сравнить их полезность с точки зрения финансового анализа и результатов, которые можно получить и проинтерпретировать в рамках оценочных скоринговых процедур.

Формула евклидова расстояния позволяет найти численную характеристику, определяющую разницу между сравниваемыми векторами данных рассматриваемых контрагентов (1):

$$d(A,B) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}, \quad (1)$$

где x_1 и x_2, y_1 и y_2 – характеристики сравниваемых векторов данных.

Далее при помощи отбора на основании алгоритмов ближайшего либо дальнего соседа можно описать кластеры контрагентов, то есть выделить типы значений для того, чтобы сделать выводы и классифицировать их по ходу дальнейшего анализа. С помощью алгоритма ближайшего соседа осуществляется поиск прилегающих значений и с учетом этого производится поэтапный отбор минимальных расстояний и разбиение на кластеры (2):

$$d(X_1, X_2) = \min d(x, y), \quad (2)$$

где x, y – сравниваемые характеристики элементов массива с данными;

X_1, X_2 – сравниваемые в ходе анализа контрагенты.

При поиске дальнего соседа проводится поэтапный отбор максимальных расстояний между сопоставимыми значениями (3):

$$d(X_1, X_2) = \max d(x, y). \quad (3)$$

В случае расстояния Махаланобиса учитывается сходство рассматриваемых выборок, расчет производится с учетом корреляционных характеристик их значений (4). По итогам разбиения последних на классы можно говорить о некоторой зависимости самих анализируемых совокупностей значений:

$$d(A,B) = \sqrt{(x-y)^T S^{-1}(x-y)}, \quad (4)$$

где x, y – векторы данных,
S^{-1} – обратная матрица ковариации.

На основании этого можно проанализировать данные контрагентов с целью разделения их на классы для дальнейшей оценки взаимосвязанности. Для этого были отобраны сведения, характеризующие экономическую эффективность ряда субъектов хозяйствования, занимающихся сходной деятельностью на идентичных рынках, при этом перед началом исследования был осуществлен поиск связанности между некоторыми из рассматриваемых предприятий.

Используемый набор данных содержал в себе информацию о 50 компаниях, которые необходимо разбить на группы в зависимости от значений показателей. Таким образом, имелась в наличии таблица с 50 записями, каждая из которых содержала параметр X_1 – объем выручки от реализации и X_2 – размер чистой прибыли. На их базе требовалось классифицировать юридические лица со сходными характеристиками по предварительным оценкам.

При выполнении кластеризации на основании евклидова расстояния применялся ряд допущений и предпосылок, связанных с особенностью алгоритма, заключающегося в поиске минимального расстояния между объектами и на основе этого формировании их сходных групп. Для настройки алгоритма использовались следующие ограничения: 2 кластера выбирались вследствие эмпирической оценки, остановка итераций происходила в момент получения равноценных значений расстояний между группами.

Листинг 1. Программный код для расчета евклидова расстояния с использованием метода ближайшего соседа

```
# Расчет Евклидова расстояния:
matrix = pdist(X, metric='euclidean').

# Метод ближайшего соседа:
link_matrix = sch.linkage(matrix, method='single').

# Количество кластеров, в данном случае 2:
c = fcluster(link_matrix, t=2, criterion='maxclust').
```

На *рис. 1* отображены сформированные кластеры, не создающие четких групп и не имеющие существенных особенностей и взаимосвязанности контрагентов. При реализации алгоритма расчета евклидова расстояния путем поиска дальнего соседа наблюдается сходная картина *(рис. 2)*. Такой результат закономерен, так как алгоритм поиска не меняется и в модель оценки не вводятся дополнительные факторы, которые могут каким-то образом скорректировать итоги реализуемых процедур. Это достаточно важное условие при анализе набора данных, имеющих определенные экономические и общественные характеристики.

Листинг 2. Программный код для расчета евклидова расстояния с использованием метода дальнего соседа

```
# Расчет Евклидова расстояния:
matrix = pdist(X, metric='euclidean').

# Метод дальнего соседа:
link_matrix = sch.linkage(matrix, method='complete').

# Количество кластеров, в данном случае 2:
c = fcluster(link_matrix, t=2, criterion='maxclust').
```

Сегментация данных на основании расстояния Махаланобиса

При использовании алгоритма агломеративной кластеризации, подразумевающего учет корреляции между данными и за счет этого инвариативного к масштабу их объема, мы получили картину, которая меняется из-за того, что при расчете расстояния учитывается ковариация между элементами, и именно такой тип оценки позволяет проводить дальнейший анализ на предмет их связанности *(рис. 3)*. Предоставляется возможность рассматривать контрагентов с точки зрения происходящих в их деятельности изменений и с учетом этого корректировать последующие аналитические процедуры.

Листинг 3. Программный код расчета расстояния Махаланобиса

```
# Расчет матрицы ковариации:
cov_matrix = np.cov(X.T).

# Расчет обратной ковариационной матрицы:
icov_matrix = np.linalg.inv(cov_matrix).

# Функция для расчета расстояния Махаланобиса:
def mah_dist(x, y, icov_matrix):
d = x – y
return np.sqrt(np.dot(np.dot(d, icov_matrix), diff.T)).

# Вычисление матрицы расстояний Махаланобиса
между всеми парами точек:
dist_matrix = cdist(X, X, metric=lambda u, v: mah_
dist(u, v, icov_matrix)).

# Иерархическая агломеративная кластеризация
с использованием метода ближайшего соседа:
c = linkage(dist_matrix, method='single').
```

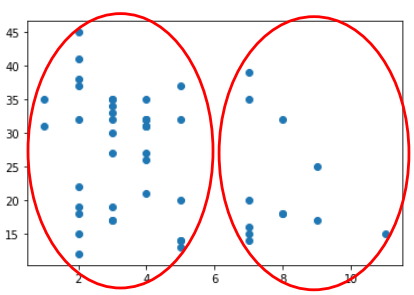

Рис. 1. Результаты кластеризации данных на основании расчета евклидова расстояния с использованием метода ближайшего соседа

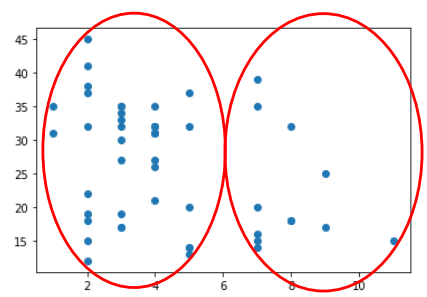

Рис. 2. Результаты кластеризации данных на основании расчета евклидова расстояния методом дальнего соседа

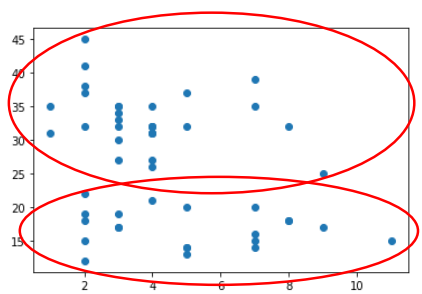

Рис. 3. Результаты кластеризации данных на основании расчета расстояния Махаланобиса с использованием метода ближайшего соседа

На *рис. 3* представлено разделение групп с учетом связанности их деятельности и согласованности в динамике представленных данных. Однако только на основании реализации этого алгоритма кластеризации нельзя утверждать о взаимозависимости рассматриваемых значений отдельных показателей либо общих результатов деятельности контрагентов, но при этом можно определить направление проведения дальнейшего анализа. Предположим, что на начальном этапе сходные по деятельности, либо объединенные по каким-либо не выявленным в данный момент признакам компании оказались в одном кластере. Это позволяет сделать первоначальные выводы об их связанности и конкретизировать дальнейшие аналитические процедуры. В таком случае агломеративные алгоритмы более эффективны, и их реализация более целесообразна в рамках скоринговых процедур, так как дает возможность обратить внимание на возможность существования связанности между оцениваемыми субъектами. Кроме того, по итогам реализации алгоритмов прослеживаются различия в разбиении на кластеры имеющихся данных, и с учетом этого можно сделать выводы о наличии дополнительных факторов, под воздействием которых картина распределения изучаемых элементов изменилась. Следовательно, скрытые причины, которые в данный момент не конкретизированы, оказывают существенное влияние на работу организаций и, базируясь на этом, следует принять решение о необходимости дальнейших аналитических процедур, а также о качестве инструментария для последующего анализа.

Надо заметить, что результаты последовательного использования алгоритмов кластеризации разного типа еще не говорят о наличии связанности, они просто обращают внимание на определенную синхронность динамики работы некоторых из рассматриваемых субъектов, а также на наличие факторов, которые являются существенными для проведения оценки хозяйственной деятельности, но не учтены на данном этапе аналитики. Ее применение на примере большого объема информации позволит сформировать оценочную шкалу, на основании которой можно будет охарактеризовать силу влияния скрытых факторов на характер работы рассматриваемых контрагентов.

В ходе исследования особенностей оценки их связанности рассмотрены два типа алгоритмов: на основании евклидова расстояния и с расчетом расстояния Махаланобиса. Второй тип, учитывающий корреляционные оценки показателей, продемонстрировал более высокую эффективность при реализации скоринговых процедур оценки компаний, так как по его результатам можно сделать выводы о наличии дополнительных факторов, оказывающих влияние на элементы рассматриваемого набора данных. Если мы ведем речь о субъектах хозяйствования, то использование агломеративных алгоритмов позволяет сделать начальные выводы об их связанности, что является существенным признаком при банковской оценке их деятельности. Помимо этого следует отметить, что проведение процедур кластеризации на основании евклидова расстояния и далее агломеративной процедуры, например, по расчету расстояния Махаланобиса, дает возможность судить о силе воздействия невыявленных факторов на экономическую эффективность рассматриваемой группы предприятий. И

Summary. The article examines the main approaches to clustering of similar data characterizing the economic activity of these entities using fifty similar counterparties as an example. A comparison is made between the basic clustering algorithms using the Euclidean distance and the typical agglomerative algorithm based on the Mahalanobis distance, which considers the covariance estimates of elements from the set under consideration. The article examines the consistent application of different types of algorithms to economic entities. In addition, it examines the change in dependence when applying various analytical procedures. Based on the analysis, conclusions are made about the advantages and disadvantages of the above algorithms, and it is noted that it is advisable to consistently apply algorithms of various types to identify hidden causes that affect the activities of legal entities and that are not obvious when conducting a standard financial analysis of the activities of business entities.

Keywords: data analysis, financial data, big data, data array, scoring, analytics, data distribution, financial characteristics, financial analysis, bank scoring, credit policy, management company, subsidiary, clustering, Euclidean distance, Mahalanobis distance, related counterparties, scoring, relatedness analysis.

https://doi.org/10.29235/1818-9857-2025-10-66-69

Статья поступила в редакцию
17.01.2025 г.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Мыльников Л.А. Статистические методы интеллектуального анализа данных.– СПб., 2021.
2. Управление банковскими рисками: учебник / Е.В. Бережная, С.В. Зенченко, М.В. Сероштан, О.В. Бережная.– 2-е изд.– М., 2022.
3. Управление кредитным риском в банке: подход внутренних рейтингов: практическое пособие для вузов / М.В. Помазанов; под научной редакцией Г.И. Пеникаса.– 2-е изд., перераб. и доп.– М., 2023.
4. Волков А.А. Управление рисками в коммерческом банке: практ. руководство / А.А. Волков.– 3-е изд., испр. и доп.– М., 2015
5. Nicholson W.L. Exploring Data Analysis.– Oakland, 2012.