

ПРАКТИЧЕСКИЙ АНАЛИЗ ПРОБЛЕМЫ КЛАСТЕРИЗАЦИИ БАНКОВ САНКТ-ПЕТЕРБУРГА

Г.С. Осипов, д-р техн. наук, доцент
Н.С. Вашикидзе, старший преподаватель
Г.В. Филиппова, старший преподаватель
Н.Л. Рауш, старший преподаватель
Сахалинский государственный университет
(Россия, г. Южно-Сахалинск)

DOI:10.24412/2500-1000-2026-6-2-98-106

Аннотация. приведена постановка задачи кластеризации сложных многомерных объектов, рассматриваются основные показатели деятельности банков Санкт-Петербурга, проведен разведывательный анализ данных и снижение размерности задачи методами главных компонент и IsoMap, проведено аналитическое исследование данных о работе банков с помощью самоорганизующейся карты Кохонена. Осуществлена кластеризация банков методом k -means, найдено оптимальное число кластеров по критерию максимизации индекса силуэтов.

Ключевые слова: снижение размерности данных; визуализация; кластеризация банков.

Текущее положение исследований и практических реализаций систем искусственного интеллекта в области машинного обучения «без учителя» направлено на решение проблемы разбиения множества сложных многомерных объектов на группы схожих по характеристикам, чтобы перейти от узконаправленного воздействия на конкретные объекты к управлению группой подобных объектов. Это позволяет оптимизировать процессы принятия решений и управления множеством структурированных объектов.

Объектом исследования являются банки Санкт-Петербурга, а предметом – нахождение оптимального числа кластеров, на которые целесообразно разбить объекты с целью оптимизации управляющих воздействий на целые группы из них.

1. Постановка задачи кластеризации

Дано множество объектов $X = \{x_1, x_2, \dots, x_n\}$, где каждый объект описывается набором признаков. Требуется разбить множество на k кластеров $C_1, C_2, \dots, C_k$, удовлетворяющих условиям:

$$C_i \cap C_j = \emptyset \quad (i \neq j);$$

$$\bigcup_{i=1}^k C_i = X.$$

Кластеризация относится к классу задач машинного обучения без учителя.

2. Исходные данные

Основные показатели банков [1]:

1. Активы нетто.
2. Чистая прибыль.

3. Капитал (по форме 123).

4. Кредитный портфель.

5. Вклады физических лиц.

6. Вложения в ценные бумаги.

На рисунке 1 представлен набор данных для кластеризации банков.

Банк	Активы нетто	Чистая прибыль	Капитал (по форме 123)	Кредитный портфель	Вклады физических лиц	Вложения в ценные бумаги
SEB	15038899	210155	14022867	0	195	0
Александровский	16289972	-27526	2449708	5950965	7698160	6618561
Балтинвестбанк	15433330	-375209	-25920461	895128	298394	2471662
Банк «Викинг»	1460316	8489	815482	493593	20467	0
Банк «Объединенный капитал»	52537889	120600	7234017	1225248	23303713	13395492
Банк Оранжевый	553057	-26354	421391	0	83445	308334
Банк «ПСКБ»	77640674	185077	5061682	8711485	20749140	6140926
Банк «РОССИЯ»	1394975492	1618194	139950029	859594918	224726591	106647078
БСПБ	1342782400	6068559	207884661	829756695	332386651	164056891
ВТБ	35218820422	81087940	2229073599	20428409087	8587907652	7100795819
ГОРБАНК	5971751	71872	2988557	493787	2540865	1809125
Мобильная карта	70183054	5465371	17177369	0	0	0
МТС Деньги (ЭКСИ-Банк)	5648810	-65729	691619	128260	6072	0
Петербургский Расчетный Центр	773347	14250	698255	0	0	208
Платежный Конструктор	11936900	1302985	2437708	106598	0	198529
Санкт-Петербургский Банк Инвестиций	2365688	26553	1680124	7783	0	226
Сити Инвест Банк	6145859	606329	2165091	1628579	0	1834710
ФИНСТАР БАНК	9775149	69968	1495071	3642560	3078181	3513666

Рис. 1. Таблица исходных данных

3. Разведывательный анализ

3.1. Снижение размерности

Предварительный (разведывательный) анализ полного набора исходных данных (из

рисунка 1) основан на применении метода главных компонент. На рисунке 2 представлена проекция шестимерных данных на первые две главные компоненты (1ГК и 2ГК).

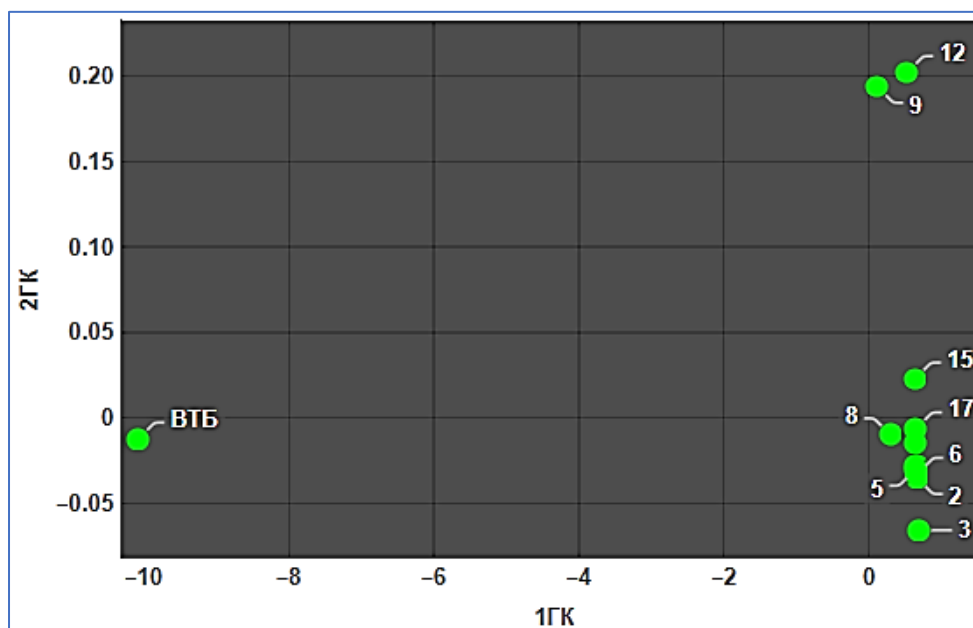


Рис. 2. Проекция исходных данных на первые две главные компоненты

Очевидно, ВТБ отстоит от основной группы остальных банков (которые имеют произвольную нумерацию) на существенное расстояние.

Метод главных компонент является линейным методом и ищет линейные структуры в данных, а нелинейный метод IsoMap ориентирован на поиск многообразий, что может повысить качество визуализации за счет

использования не евклидова, а геодезического расстояния между объектами.

Результат снижения размерности задачи методом IsoMap представлен на рисунке 2. Оба метода показывают схожие результаты, свидетельствующие о том, что ВТБ необходимо определить в уникальный кластер, и продолжить анализ данных для остальных банков.

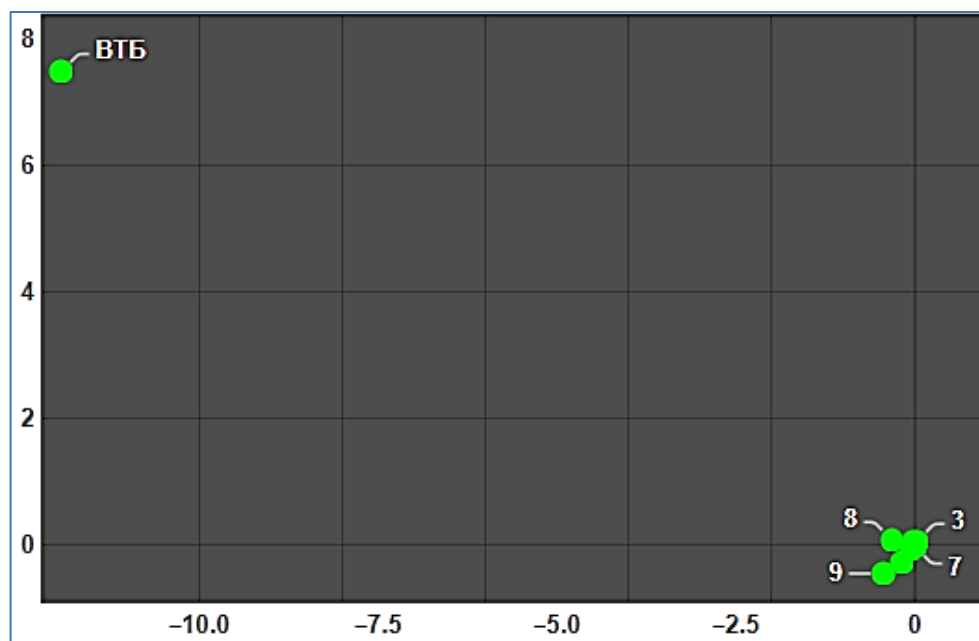


Рис. 3. 2D проекция данных, полученная методом IsoMap

В дальнейшем этот метод будет основным для задачи снижения размерности и визуализации.

Результаты второго этапа снижения размерности группы банков без ВТБ представлен на рисунке 4.

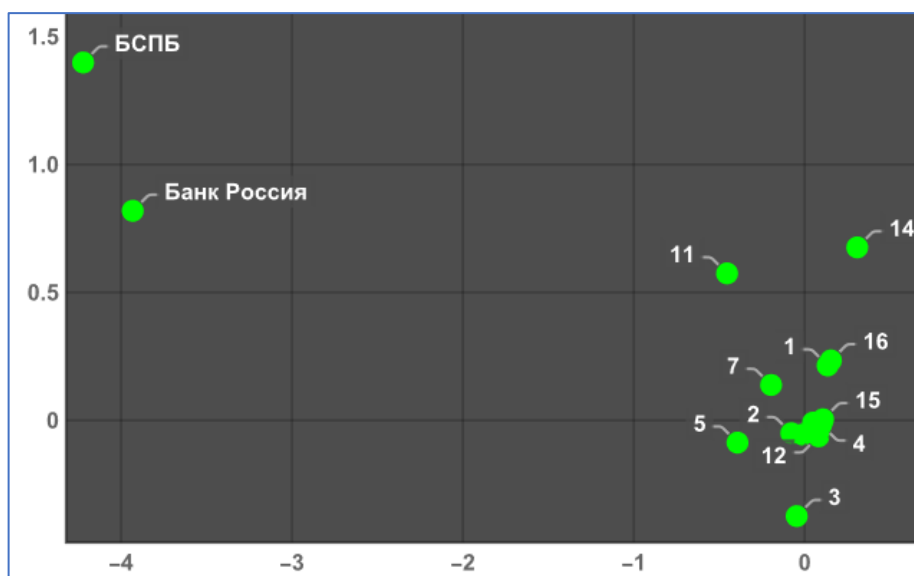


Рис. 4. 2D проекция показателей группы банков (без ВТБ)

Рассуждая аналогично, исключим из кластеризации еще два банка – БСПБ и Банк Россия. Результат приведены на рисунках 5 и 6.

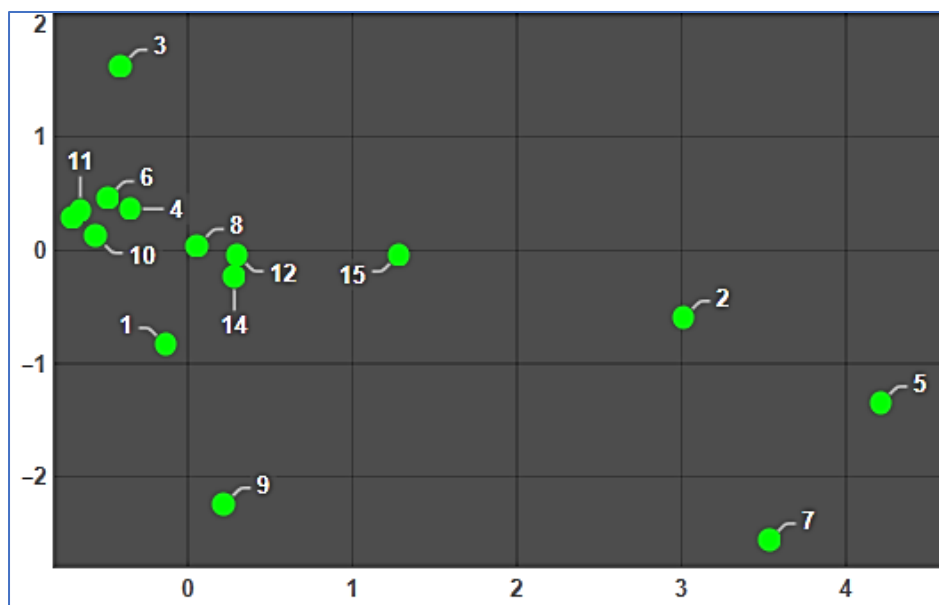


Рис. 5. Расположение на плоскости показателей основной группы банков

Рисунок 6 представляет трехмерную проекцию.

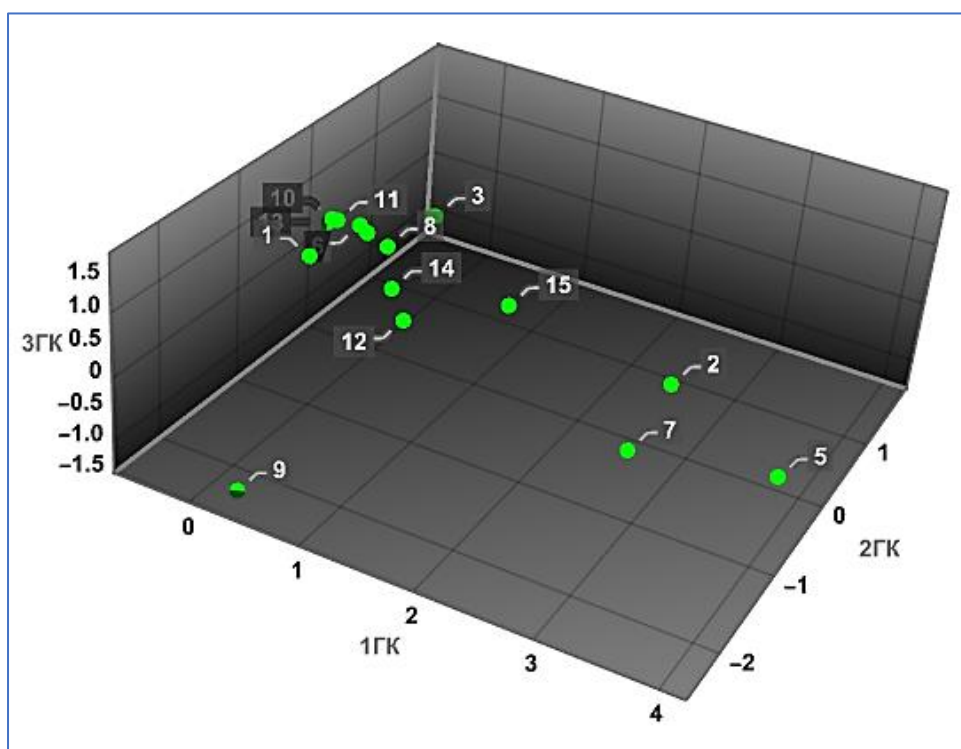


Рис. 6. 3D проекция показателей банков

Кластеризацию проведем для оставшиеся 15 банков.

3.2. Анализ с помощью Карты Кохонена

Карта Кохонена позволяет многомерные данные представить в удобном для восприятия виде, показывая, какие группы (кластеры) существуют в данных и как они связаны друг с другом за счет того, что близкие типы объектов расположены на соседних нейронах.

В качестве примера на рисунке 7 приведена самоорганизующаяся карта с полным списком банков (с ВТБ). Банк представлен в правом нижнем углу (голубые и синие клетки). Все остальные банки сосредоточены левее и выше. ВТБ и остальные банки разделены яркой границей – коричневые, красные и желтые ячейки, что свидетельствует о хорошей разделимости этих групп.

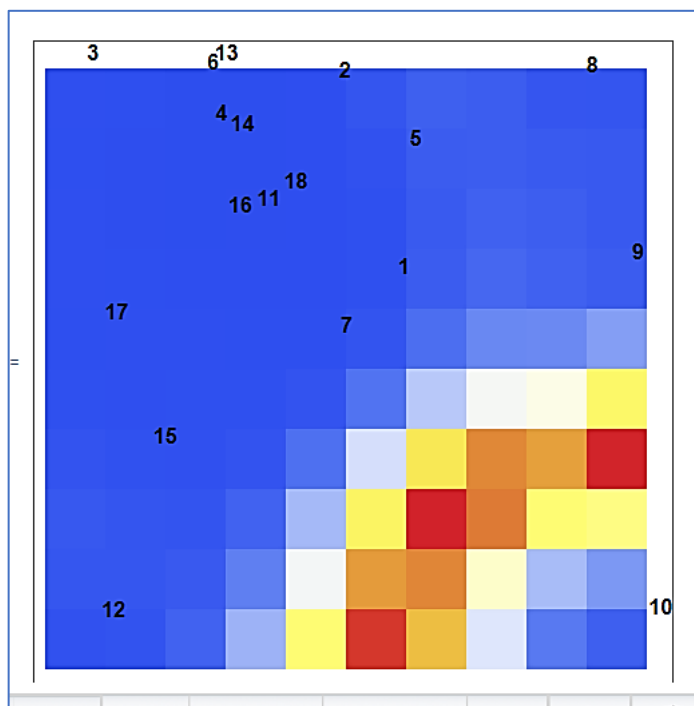


Рис. 7. Карта Кохонена 10×10

На рисунке 8 приведена карта после исключения из рассмотрения ВТБ. В этом варианте хорошо «видны» два банка – БСПБ и Банк

Россия в нижнем правом углу. Основная группа отделена от этих банков клетками «теплого цвета».

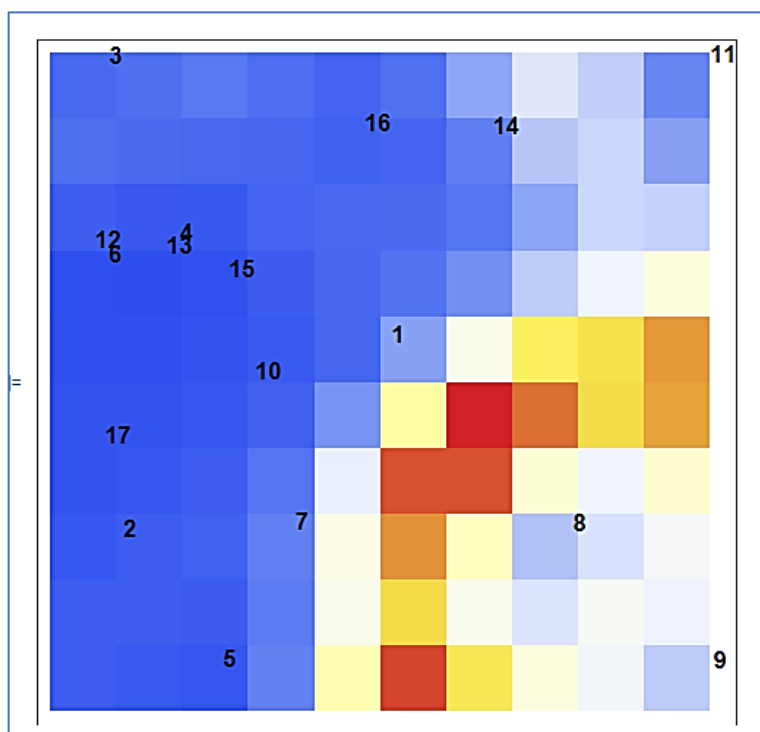


Рис. 8. Карта Кохонена для банков без ВТБ

Окончательно карта после дополнительного исключения и Банка Россия и БСПБ приведена на рисунке 9.

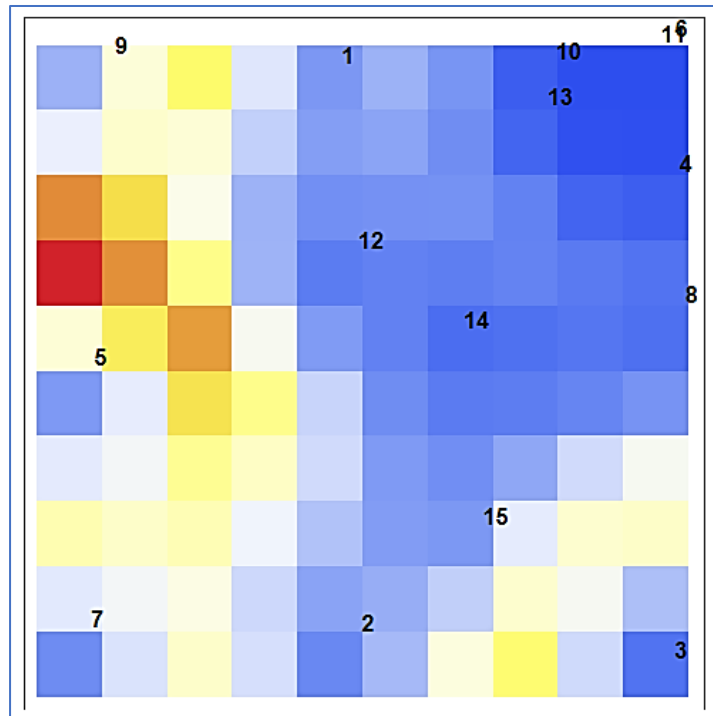


Рис. 9. Карта Кохонена для 15 банков

4. Кластерный анализ

Кластеризация осуществлялась методом *k-means* [2], а для оценки качества кластеризации применялся метод силуэтов [3], который позволяет получить полное графическое представление объектов кластеров в виде столбиковой диаграммы, высота которых указывает на

индекс силуэта, характеризующего компактность и отделимость кластеров.

4.1. Метод силуэтов

Компактность кластеров оценим как среднее расстояние от объекта $x_i \in c_k$ до других объектов из этого кластера c_k :

$$a(x_i, c_k) = \frac{1}{|c_k|} \sum_{x_j \in c_k} \|x_i - x_j\|$$

Отделимость кластеров может быть определена как среднее расстояние от $x_i \in c_k$ до объектов из другого (ближайшего) кластера c_l ($k \neq l$).

$$b(x_i, c_k) = \min_{c_l \in C \setminus c_k} \left\{ \frac{1}{|c_l|} \sum_{x_j \in c_l} \|x_i - x_j\| \right\}$$

Тогда, очевидно, совокупная оценка качества кластерной структуры найдется так:

$$Sil(C) = \frac{1}{N} \sum_{c_k \in C} \sum_{x_i \in c_k} \frac{b(x_i, c_k) - a(x_i, c_k)}{\max\{a(x_i, c_k), b(x_i, c_k)\}}$$

Таким образом индекс силуэта показывает, насколько хорошо объект вписывается в свой кластер и насколько плохо в другие.

4.2. Определение оптимального числа кластеров

Оптимальное число кластеров определяется максимальным значением совокупного индекса силуэтов для всей структуры кластеров.

На рисунке 10 представлены значения индексов силуэтов для 15 банков (нумерация банков здесь не представлена).

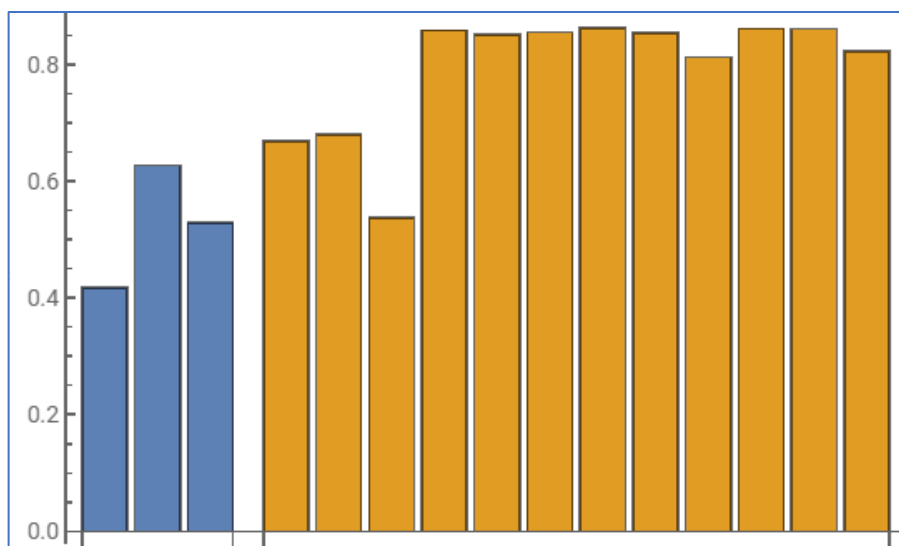


Рис. 10. Диаграмма индекса силуэтов для 15 банков

Максимальное значение индекс силуэтов достигает при числе кластеров равным 2. В данном случае он равен 0.74 (см. рисунок 11). В первом кластере 3 банка и 12 во втором.

CalinskiHarabasz	51.0884
DaviesBouldin	0.416082
Dunn	1.01414
RSquared	0.549869
Silhouette	0.740435
StandardDeviation	13 376 528.

Рис. 11. Различные показатели качества кластеризации

При других значениях числа кластеров значение этого индекса меньше. На рисунке 12 приведен точный график зависимости индекса силуэтов от числа кластеров.

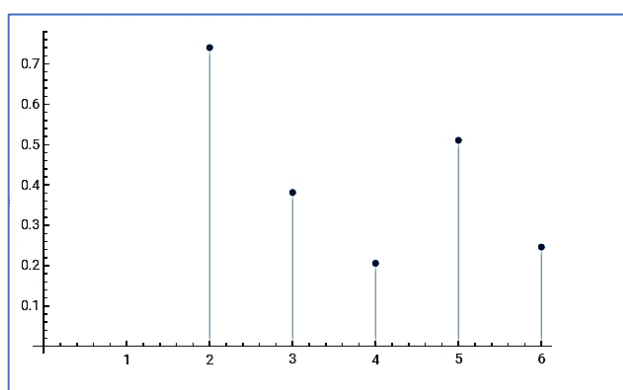


Рис. 12. График зависимости индекса силуэтов от числа кластеров

На рисунке 13 представлено окончательное разбиение 15 банков Санкт-Петербурга на 2 кластера (без обозначения наименования банков).



Рис. 13. Схематичное изображение двух кластеров на плоскости

Заключение

В статье проведен практический анализ проблемы кластеризации многомерных объектов. В качестве объектов выступают банки Санкт-Петербурга, работа которых оценивается по 6 цифровым показателям. Исследованы методы снижения размерности задачи, такие как метод главных компонент и IsoMap. Предложены рекомендации по априорной визуализации показателей банков в пространстве размерности 2 с помощью самоорганизующейся карты Кохонена, которая позволяет представить многомерные данные в удобном для

восприятия виде, не нарушая исходную топологию кластеров. Отработана методика определения оптимального числа кластеров путем сведения проблемы к задаче максимизации индекса силуэтов кластерной структуры.

Проведенное исследование практически реализовано в системе символьной математики и компьютерной алгебры Wolfram Mathematica. Результаты исследования направлены на разбиение объектов на группы с целью выработки рациональных управляющих воздействий на схожие по показателям банки.

Библиографический список

1. Рейтинги банков. – [Электронный ресурс]. – Режим доступа: https://www.banki.ru/banks/ratings/?PAGEN_1=7&after-login=esia.
2. Ломакин Н.И., Марамыгин М.С., Черная Е.Г., Кузьмина Т.И., Бестужева Л.И., Борискина Т.Б. Кластеризация в исследовании закономерностей развития и устойчивости российских коммерческих банков системами искусственного интеллекта // *Фундаментальные исследования*. – 2024. – № 2. – С. 17-25.
3. Гусев В.М. Сравнительный анализ методов оценки качества кластеризации на примере алгоритма k-means и набора данных Iris / В.М. Гусев // *Флагман науки*. – 2026. – № 4(39). – С. 384-395.

PRACTICAL ANALYSIS OF THE PROBLEM OF CLUSTERING OF ST. PETERSBURG BANKS

G.S. Osipov, *Doctor of Technical Sciences, Associate Professor*

N.S. Vashakidze, *Senior Lecturer*

G.V. Filippova, *Senior Lecturer*

N.L. Rausch, *Senior Lecturer*

Sakhalin State University

(Russia, Yuzhno-Sakhalinsk)

Abstract. *Setting out the task of clustering complex multidimensional objects, considering the main performance indicators of banks in St. Petersburg, conducting an intelligence analysis of data and reducing the dimension of the task using the methods of the main components and IsoMap, conducting an analytical study of data on the work of banks using the Kohonen self-organizing map. Banks were clustered using the k-means method. The optimal number of clusters was found according to the criterion of maximizing the silhouette index.*

Keywords: *data dimensionality reduction; visualization; bank clustering.*