УДК 330

А.А. Макарова*

ИСПОЛЬЗОВАНИЕ СРЕДСТВ МНОГОМЕРНОЙ КЛАССИФИКАЦИИ В БАНКОВСКОЙ СФЕРЕ

В статье описаны средства многомерной классификации, в частности метод кластерного и дискриминантного анализа, а также их применение относительно банковского сектора. Выделены основные методы классификации и описано применение математического инструментария при анализе банковских данных в условиях современного совершенствования банковской системы.

Ключевые слова: кластерный анализ, дискриминантный анализ, дискриминантные переменные, классификация с обучением, банковская сфера.

Сегодня в банковской системе широко используются различные многомерные методы. Важными среди них являются методы классификации, поскольку в работе банка требуется постоянный анализ и сравнение различных характеристик, а также объектов или клиентов. Ключевыми для классификации являются методы кластерного и дискриминантного анализа.

Кластерный анализ позволяет структурировать данные группы (кластера): разделение клиентов, депозитов, кредитов и других объектов в группы.

В последнее время статистический метод моделирования кластерного анализа все чаще применяется к финансовой отчетности, что позволяет решать две основные проблемы – обработка отсутствующих данных и поиск однородных групп данных. Этот подход является достаточно гибким и позволяет обрабатывать большие и сложные структуры данных.

Многие задачи в банковском секторе можно решить с использованием кластеризации, например:

1) необходимо найти лучшие из филиалов крупной банковской сети, которая простирается по всей стране – для сравнения необходимо разделить филиалы на группы;

2) для филиала банка необходимо определить оптимальную конфигурацию фронт-линии – для этого нужно выделить группы схожих дней и найти наилучшую фронт-линию для каждого из них;

3) в отделении банка требуется понять, какую роль выполняют сотрудники – здесь актуален вопрос определения групп работников, выполняющих подобные операции.

Методы кластерного анализа можно разделить на несколько групп: агломеративные (объединяющие) методы – последовательно объединяющие отдельные объекты в кластеры и дивизимные – разделяющие группу на отдельные объекты.

Итеративные методы – методы, в которых кластеры формируются на основе задаваемых условий разбиения (или параметров). Пользователь в процессе работы алгоритма имеет возможность изменять условия для получения желаемого разбиения. Итеративные алгоритмы могут привести к образованию пересекающихся кластеров.

Макарова Анастасия Александровна (zvezdanastya2@mail.ru), кафедра математики и бизнес-информатики, Самарский государственный университет, 443011, Российская Федерация, г. Самара, ул. Акад. Павлова, 1.

На рис. 1 представлена дендрограмма – график, отражающий последовательное объединение двух кластеров в один и расстояния между ними.

На дендрограмме видно, что на первом шаге в один кластер объединяются n2 и n3, расстояние между которыми 0,15. Далее на втором шаге к объектам n2 и n3 присоединяется n1. Расстояние от первого объекта до кластера, содержащего n2 и n3, – 0,3 и т. д.

Существует три агломеративных метода. В методе «ближнего соседа» при объединении в кластер двух объектов выбирается минимальное расстояние из двух. В методе «дальнего соседа» – максимальное. А в методе «средней связи» расстояние рассчитывается как среднее арифметическое из двух расстояний.

Кроме агломеративных существуют иерархические дивизимные методы. Их основная предпосылка заключается в том, что изначально все объекты относятся к одному кластеру, а уже в процессе классификации отделяются от кластера группы схожих объектов. Таким образом, на каждом шаге число кластеров увеличивается, а мера расстояния уменьшается.

На рис. 2 представлена дендрограмма для дивизимных иерархических методов.

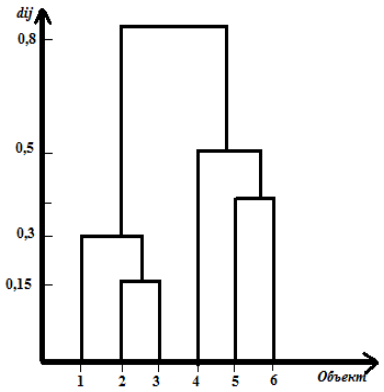

Рис. 1. Пример дендрограммы иерархического агломеративного кластерного анализа

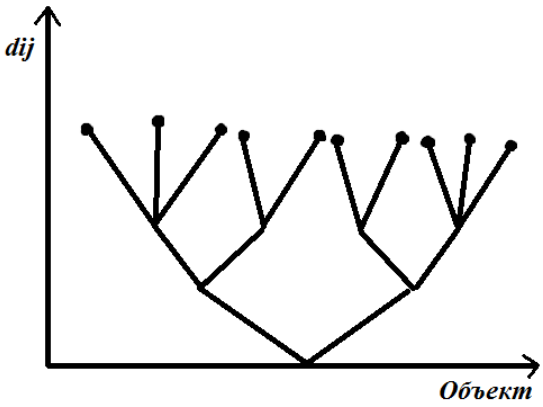

Рис. 2. Дендрограмма иерархического дивизимного алгоритма

При оценке полученных результатов применяются критерии качества кластеризации, которые представляются тремя функционалами.

В результате кластеризации формируются обучающие выборки, с помощью которых осуществляется классификация с обучением или дискриминантный анализ, который позволяет в дальнейшем проводить сортировку по известным кластерам.

Дискриминантными переменными являются признаки, применяемые для отличия одного класса от другого. Таким образом, общий вид функции $f(x)$ представляет собой линейную комбинацию

$$f(x) = a_1 x_1 + a_2 x_2 + \cdots + a_k x_k,$$

где x_1, x_2 … x_n – дискриминантные переменные.

Необходимо найти значения коэффициентов дискриминантной функции a_j. При этом следует помнить, что внутригрупповая вариация для рассматриваемых объектов должна быть минимальной, а межгрупповая – максимальной. В таком случае будет достигнуто наилучшее разбиение двух групп. Другими словами, величина F должна быть максимальной.

Прежде чем приступить к процедуре классификации, необходимо определить границу, которая разбивает рассматриваемую группу на две. Данной величиной может быть значение, которое равноудалено от $\bar{f}_1$ и $\bar{f}_2$:

$$C = \frac{1}{2}(f_1 + f_2).$$

Величина C является константой дискриминации. Если граница между рассматриваемыми группами выбрана правильно, то суммарная вероятность ошибочной классификации будет минимальна.

Дискриминантный анализ, а именно его метод, впервые был применен в области банковской деятельности в рамках кредитного анализа. Именно в кредитном анализе виден подход метода, подразумевающий применение прошлого опыта, когда необходимо определить различия между заемщиками, которые вернули кредит в срок, и теми, кто этого не смог сделать. Данная информация используется в решении о кредитоспособности новых потенциальных заемщиков.

Иными словами, цель применения метода заключается в построении модели, которая позволяет предсказать, к какой из групп относятся данные потребители, исходя из набора прогнозирующих переменных, измеренных в интервальной шкале.

Примером применения дискриминантного анализа в банковской сфере являются выдачи кредитов. В оценке финансового состояния клиентов при выдаче им кредита банк разделяет их на надежных и ненадежных по нескольким основаниям: достижение совершеннолетнего возраста, средний душевой доход, величина прожиточного минимума, уровень безработицы, наличие обеспечения, страхования и многое другое.

Зачастую перед банком стоит сложный выбор при выдаче кредитных денежных средств клиентам, организациям или же в масштабе районов городского округа. В рассмотренной задаче имелось 12 районов городского округа Самара, которые необходимо было распределить по двум выборкам (кластерам) с целью дальнейшей дискриминации.

Также имелись данные о среднедушевом денежном доходе, средней заработной плате работников предприятий и организаций, величине прожиточного минимума и уровне безработицы для каждого района соответственно.

На основании исходных данных по каждому признаку рассчитаны среднее значение $\bar{x}_j$ и среднее квадратическое отклонениеσ по формуле

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - x)^2}.$$

Оценка сходства между объектами зависит от абсолютного значения признака и от степени его вариации в совокупности. Для устранения такого влияния к таблице исходных данных было применено нормирование:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}.$$

Для дальнейшего применения алгоритмов кластерного анализа необходимо было рассчитать симметричную матрицу расстояний D по формуле евклидового расстояния:

$$d_{ij}(X_i, X_j) = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2}.$$

Полученная матрица расстояний позволила использовать агломеративные методы кластерного анализа. В результате были получены дендрограммы для каждого метода, что позволило визуализировать разделение исходных данных на два кластера.

Для оценки полученных результатов использовались функционалы, или критерии качества разбиения (табл. 1).

Так, в методе «ближнего соседа» сумма функционалов равна 1816543,26, в методе «дальнего соседа» – 2110617,51, а в методе «средней связи» –2475497,47. Минимальная сумма функционалов получена в методе «ближнего соседа», или методе одиночной связи. Следовательно, для дальнейшей дискриминации необходимо использовать разбиение на кластеры данным способом.

Используем следующий алгоритм дискриминантного анализа.

1. Вычислить средние значения признаков для каждого множества (обучающей выборки), записать векторы средних значений $\bar{X}_1$ и $\bar{X}_2$. Вычислить вектор разности $(\bar{X}_1 - \bar{X}_2)$.
2. Вычислить матрицы ковариаций для каждой выборки S_1 и S_2.
3. Вычислить несмещенную оценку обобщенной матрицы ковариаций используя формулу

$$\hat{S} = \frac{1}{(n_1 + n_2 - 2)(n_1 S_1 + n_2 S_2)}.$$

4. Вычислить $\hat{S}^{-1}$.
5. Вычислить вектор коэффициентов дискриминантной функции А.
6. Вычислить константу дискриминации С.
7. Сравнить значение дискриминантной функции тестируемых объектов с величиной С.

Используя средние значения $\bar{U}_1$ и $\bar{U}_2$, вычисляем константу дискриминации С. Для этого складываем средние значения, делим пополам и получаем С = 0,00449. Данная величина представляет собой границу, которая равноудалена от центров двух множеств (табл. 2).

Таблица 1

Функционалы разбиения

Методы		«ближнего соседа»	«дальнего соседа»	«средней связи»
Функционалы	F_1	908159,33	1055175,10	1955086,34
	F_2	224,60	267,32	219,19
	F_3	908159,33	1055175,10	520191,94

Таблица 2

Вектор дискриминации

Районы городского округа Самара	U
Алексеевский	-0,033
Сызранский	0,029
Исаклинский	-0,028
Пестравский	-0,025
Нефтегорский	-0,043
Кинельский	-0,019
Безенчукский	-0,023
Борский	-0,021
Сергиевский	0,021

Процедура дискриминантного анализа закончена. Сравнивая значения константы дискриминации С и вектора дискриминации F, делаем вывод о том, что у Сызранского и Сергиевского районов более высокие финансовые показатели. Республика Тыва, Читинская область, республика Саха, Чукотский автономный округ, Камчатская, Магаданская и Сахалинская области отнесутся же ко второму кластеру с менее высокими финансовыми показателями.

В результате дискриминации банковская организация при решении вопроса о выдаче кредитных денежных средств районам городского округа Самара отдаст предпочтение Красноярскому краю и Калининградской области.

Дискриминантный анализ, а также кластерный анализ относятся к методам многомерной классификации, но основываются на определенных предпосылках.

Основное различие заключается в том, что в ходе дискриминантного анализа не образуются новые кластеры, а формируется правило, в котором новые элемен-

ты присоединяются к одному из существующих множеств (кластеров). Основой для отнесения каждой единицы совокупности к определенному множеству является величина дискриминантной функции, которая рассчитывается на соответствующем значении дискриминатной функции.

Существуют две основные проблемы дискриминантного анализа – определение набора дискриминантных переменных и выбор формы дискриминантной функции. Также существуют различные критерии для последовательного отбора переменных, позволяющие получать наилучшие разбиения множеств.

Современное совершенствование банковской системы предполагает использование математического инструментария при анализе банковских данных. Многомерные математические методы находят широкое применение при решении различных задач анализа банковских данных.

Методы многомерной кластеризации и дискриминации способствуют улучшению качества работы банков, так как в их основе лежит структурирование объектов различной природы, а также их дискриминация.

Библиографический список

1. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Основы моделирования и первичной обработки данных. М.: Финансы и статистика, 2006. 487 с.

2. Айвазян С.А., Мхитарян А.С. Прикладная статистика. Основы эконометрики: учебник для вузов: в 2 т. Т. 1: Теория вероятностей и прикладная статистика. 2-е изд., испр. М.: ЮНИТА-ДАНА, 2001. 656 с.

3. Буреева Н.Н. Многомерный статистический анализ с использованием ППП «STATISTICA». Нижний Новгород, 2007. 112 с.

4. Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы: учебник. М.: Финансы и статистика, 2000. 352 с.

5. Трусова А.Ю. Многомерные статистические методы: учеб. пос. для студентов факультета экономики и управления: в 2 ч. Самара: Изд-во «Самарский университет», 2008. Ч. 1. 67 с.

References

1. Ayvazyan S.A., Enyukov I.S., Meshalkin L.D. Applied statistics. Basics of modeling and preprocessing of data. M., Finansy i statistika, 2006, 487 p. [in Russian].

2. Ayvazyan S.A., Mkhitaryan A.S. Applied statistics. Basics of econometrics: Textbook for Institutes of Higher Education: in 2 Vol. Vol.1: Probability theory and applied statistics. 2nd edition, revised. M., IuNITI-DANA, 2001, 656 p. [in Russian].

3. Bureeva N.N. Multivariate statistical analysis with the use of application software package «STATISTICA». Nizhny Novgorod, 2007, 112 p. [in Russian].

4. Dubrov A.M., Mkhitaryan V.S., Troshin L.I. Multidimensional statistical methods: textbook. M., Finansy i statistika, 2000, 352 p. [in Russian].

5. Trusova A.Yu. Multidimensional statistical methods: training manual for the students of the Faculty of Economics and Management: in 2 parts. Samara, Izd-vo «Samarskii universitet», 2008, part 1, 67 p. [in Russian].

A.A. Makarova*

USE OF MEANS OF MULTIDIMENSIONAL CLASSIFICATION IN THE BANKING SECTOR

The article describes means of multidimensional classification, namely the method of cluster and discriminant analysis, also their use in respect of the banking sector. The key methods of classification are identified and the use of mathematical tools in the analysis of bank data in the conditions of modern improvement of the banking system is described.

Key words: cluster analysis, discriminant analysis, discriminant variables, classification learning, banking.

* *Makarova Anastasia Alexandrovna* (zvezdanastya2@mail.ru), Department of Mathematics and Business Informatics, Samara State University, Samara, 443011, Russian Federation.