Банк России

НЕКОТОРЫЕ ВОПРОСЫ КЛАССИФИКАЦИИ ОБЪЕКТОВ

Информационно-аналитический материал

Г. Гамбаров

Москва
2023

ОГЛАВЛЕНИЕ

Аннотация 1
1. Количественная оценка степени однородности совокупности 1
2. Подходы к классификации объектов в зависимости от вида данных 2
3. Использование контекстных мер близости в классификации на графах 4
Список литературы 7

Материал подготовлен Департаментом статистики.
107016, Москва, ул. Неглинная, 12, к. В
Официальный сайт Банка России: www.cbr.ru

Георгий Гамбаров
Банк России, Департамент статистики
д. э. н., доцент
gambarovgm@cbr.ru

Аннотация

В статье рассматриваются три задачи классификации финансовых объектов. Предложено определение однородности статистической совокупности, удобное для практического использования. Установлено соответствие между некоторыми видами исходных данных и методами классификации, сформулированы новые постановки задач классификации при изучении взаимосвязи показателей с использованием микроданных. Предложен метод классификации на графах.

Ключевые слова: классификация, однородность совокупности, микроданные, метод к-средних, контексты, теория графов, классификация на графах.

Key words: classification, group gomogeneity, microdata, k-mean method, context, graph theory, graph classification.

1. Количественная оценка степени однородности совокупности

Потребность оценки степени однородности статистической совокупности возникает во многих статистических исследованиях, в том числе при расчете процентных индикаторов. Более того, однородность совокупности является необходимым условием корректного применения практически всех статистических методов. При этом признание совокупности однородной или неоднородной осуществляется не на основе тех или иных критериев, а на основании субъективного анализа характеристик совокупности. Это ставит результат всего статистического исследования в зависимость от субъективных факторов, что в конечном итоге делает сложносопоставимыми результаты различных исследований.

Безусловно, определение однородности совокупности зависит от цели исследования: она может быть однородной для одной цели и неоднородной для другой. В то же время объективная количественная оценка степени однородности совокупности, не зависящей от субъективных факторов, позволяет сравнивать условия проведения различных исследований и подтверждать объективность их результатов.

Подробный обзор проблемы оценки степени однородности приведен в [4]. Для практических целей анализа финансовых процессов предлагается следующее определение: **однородной является совокупность, которую нельзя разбить на две статистически различимые совокупности**.

Из этого определения следует, что если совокупность можно разбить на две части, различие которых статистически значимо на заданном уровне значимости, то такая совокупность является неоднородной с соответствующей вероятностью.

При сравнении двух частей совокупности проверяется нулевая гипотеза о том, что обе совокупности взяты из одной генеральной совокупности, то есть являются реализацией одной случайной величины. Для проверки нулевой гипотезы рассчитывается определенная статистика, и если ее значение больше определенного порогового значения, то нулевая гипотеза отвергается. Это означает, что при заданном уровне значимости сравниваемые совокупности не принадлежат одной генеральной совокупности и различие совокупностей (средних величин) статистически значимо.

В случае когда элементы совокупности характеризуются одним показателем, используется t-критерий Стьюдента. Для его корректного применения требуется, чтобы сравниваемые

совокупности имели нормальное распределение. Однако с ростом числа наблюдений данное требование ослабевает, вследствие чего на практике t-критерий используется не только для нормальных распределений.

При сравнении двух совокупностей, элементы которых описываются несколькими переменными (показателями), вместо t-статистики используется T^2 Хотеллинга [5], равный:

$$T^2 = \frac{n_1 * n_2}{n_1 + n_2} * (\overline{X_1} - \overline{X_2})' * S * (\overline{X_1} - \overline{X_2}),$$

где n_1, n_2 – численность первой и второй совокупности соответственно;
$\overline{X_1}, \overline{X_2}$ – средние первой и второй совокупности соответственно;
S – общая ковариационная матрица, вычисляемая по формуле:

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

где S_1, S_2 – оценки ковариационных матриц первой и второй совокупности соответственно.

На основании T^2 рассчитывается величина F-статистики:

$$F = \frac{n_1 + n_2 - m - 1}{(n_1 + n_2 - 2) * m} * T^2,$$

где m – число показателей, характеризующих элемент совокупности.

Величина F подчиняется распределению Фишера с m и $(n_1 + n_2 - m - 1)$ степенями свободы. Если величина F больше соответствующего порогового значения, то нулевая гипотеза о том, что обе совокупности взяты из одной генеральной совокупности, отклоняется.

Корректное использование T^2 Хотеллинга предполагает нормальное распределение обеих совокупностей и равенство ковариационных матриц, однако на практике используется и для совокупностей, не имеющих нормального распределения аналогично тому, как это делается при использовании t-критерия Стьюдента.

Описанный подход наиболее эффективен в иерархических алгоритмах классификации. Степень однородности сформированных групп на каждом этапе объединения или разделения элементов совокупности позволяет определить оптимальное число классов.

2. Подходы к классификации объектов в зависимости от вида данных

Поведение сложных объектов определяется структурой связи *существенных (внутренних)* показателей их функционирования. Изменения условий функционирования объектов со сходной структурой связи существенных показателей должны быть похожи. Изменения условий функционирования объектов с различной структурой связи существенных показателей, скорее всего, будут различаться.

Существенные показатели не всегда известны, а известные показатели не всегда поддаются измерению. Поэтому в статистических исследованиях связи показателей и при построении их моделей рассматриваются не только существенные показатели.

Статистические методы выявления связи показателей и построение их моделей основаны на обобщении согласованности изменений рассматриваемых показателей у элементов совокупности – объектов. Связь показателей тем сильнее, чем более похожи объекты, точнее говоря, чем больше похожи структуры связи их существенных показателей. Этим объясняется требование однородности совокупности при построении моделей связи показателей.

Если исходная совокупность оказалась неоднородной, то для получения корректных выводов ее следует разбить на однородные группы. Однако близость объектов по показателям (даже по существенным показателям) не гарантирует сходства поведения объектов: близкими

значениями показателей могут обладать объекты с совершенно различной структурой связи существенных показателей. Объекты с близкими значениями в периоде, на котором формировалась группа, могут иметь заметно различающиеся показатели в следующем периоде. Только у объектов со сходной структурой связи существенных показателей изменения внешних условий приведут к похожим изменениям показателей.

Вероятность формирования групп со сходной структурой существенных показателей тем выше, чем больше число существенных показателей рассматривается в исследовании. Важные условия нахождения однородных групп – наличие достаточного числа объектов, временных периодов и выбор релевантного метода классификации.

Развитие технических возможностей сбора и обработки статистических данных повлияло на отношение к их возможностям в вопросах классификации. Так, если раньше считалось, что формальные методы классификации бесполезны при проведении типологических группировок, то после появления и использования микроданных актуален вопрос: «В какой степени методы классификации могут быть полезны для выявления типологических группировок?».

Выбор метода классификации, способного объединить объекты с близкими существенными показателями, зависит от количества периодов наблюдения за объектами: чем их больше, тем выше вероятность получения желаемого результата.

Рассмотрим некоторые подходы к классификации объектов в наиболее типичных ситуациях.

Классификация на основе данных за один период времени

Для решения задачи построения пространственной модели связи показателей используется информация о N объектах, описываемых m показателями (переменными), значения которых известны на один период.

Наличие лишь одного периода классификации объектов по близости значений показателей приводит к высокой вероятности ошибки первого рода: объекты с идентичной внутренней структурой связей будут признаны неоднородными. Особенно сильно это проявляется при использовании стандартных процедур (метод к-средних [2] и другие).

К несколько лучшим результатам приводит использование алгоритма с последовательным перемещением объекта из класса (группы) в класс и фиксацией разбиения с наибольшим значением специального функционала качества – алгоритм «Перемещение» [2]. К приемлемым результатам приводит использование в качестве функционала средневзвешенного значения качества групповых моделей связи (например, регрессии), в котором весом выступает численность класса. Правда, в этом случае вероятна ошибка второго рода: в одну группу могут попасть неоднородные объекты. Тем не менее для некоторых целей, не предполагающих предсказания неизвестных значений показателей, пространственные модели могут оказаться полезными.

Развитие пакетов статистических методов и появление возможности обработки микроданных позволили уточнить выбор наиболее адекватного способа классификации при исследовании взаимосвязи финансовых показателей, а также формулировать задачи, решение которых ранее было невозможно. Каждый объект в этом случае представляет собой совокупность элементов (микрообъектов), описываемых своими переменными. Это позволяет для каждого объекта строить пространственную модель связи, после чего сходство объектов можно определять по сходству параметров этих моделей.

Классификация на основе данных за несколько периодов времени

В этом случае рассматривается задача построения модели связи показателей на основе данных о значениях m показателей совокупности N объектов, измеренных за несколько (T) периодов времени. Подобную постановку имеет, например, задача построения панельной регрессии.

При нескольких показателях значительно лучше результаты, когда для каждого объекта строится модель динамики его показателей, например VAR-модели, после чего проводится классификация объектов, описываемых параметрами полученных моделей. Для классификации можно использовать стандартные методы, в том числе метод к-средних. Следует отметить, что данный подход эффективен в той степени, в которой характеризующие объекты показатели являются существенными в описанном выше смысле. Тем не менее данный подход заметно эффективнее предыдущих.

К наилучшим результатам приводит наличие микроданных. В этом случае для каждого элемента каждого объекта можно построить динамическую модель связи переменных, описывающих элементы объекта. Например, в задаче классификации регионов (объектов) при наличии соответствующих данных можно строить модели отдельных организаций (элементов). После чего по параметрам полученных моделей проводить классификацию всех элементов (организаций) всех объектов. Далее для каждого объекта (региона) определяется доля его элементов в каждом из полученных классов. В результате каждый объект характеризуется вектором, длина которого равна числу полученных классов, а значения компонент вектора равны доле его элементов в соответствующем классе. Классификация объектов осуществляется по компонентам данных векторов.

Задача оценки неизвестных значений показателя

Потребность в решении подобной задачи возникает в том числе при заполнении пропусков в данных. Обычно для ее решения проводят разбиение исходной совокупности на однородные группы. В группе, в которую попал оцениваемый объект, строят модель зависимости показателя с неизвестным значением от остальных показателей. Модель строится по объектам с известными значениями. Далее в полученную модель подставляют известные значения оцениваемого объекта и вычисляют его неизвестное значение.

Заслуживает внимания несколько иной способ решения данной задачи. Вначале по показателям, значения которых известны у оцениваемого объекта, находят его ближайших соседей – объекты с минимальным расстоянием до оцениваемого объекта. По отобранным таким образом объектам строится модель связи показателя с неизвестным значением с остальными показателями. Полученную модель можно использовать для оценки неизвестного значения.

Преимущество подобного подхода связано с тем, что оцениваемый объект находится в центре группы, по которой строится модель связи показателей. В то время как в традиционном подходе он может оказаться на периферии группы, что увеличит погрешность оценки его значения. Однако при использовании предлагаемого подхода следует тщательно проверять однородность полученной таким образом группы. Для этой цели может быть использовано описанное в первой части определение однородности.

3. Использование контекстных мер близости в классификации на графах

В последнее время появились работы с описанием использования методов теории графов для анализа рынков [10, 11]. Построенные сети достаточно подробно описывают операции участников торгов. В то же время при большом количестве участников торгов анализ и интерпретация их операций сильно затруднены. В этой связи перспективно агрегирование участников, осуществляющих сходные операции, в группы с последующим анализом взаимодействия полученных групп. Однако подобная классификация в этом случае осложнена нестандартным представлением информации об объектах (участниках) [7].

В связи с этим для обобщения и интерпретации результатов, полученных с помощью методов теории графов, предлагается метод классификации, использующий контекстные меры близости. В контекстных мерах близости описание единиц совокупности осуществляется не значениями показателей (переменных), а контекстами – например, наборами номеров объектов, упорядоченных по некоторому свойству. Одним из таких свойств является близость объектов, определяемая традиционными методами.

Виды контекстных мер близости

Контекстные меры близости отражают степень совпадения формализованного описания объектов. Их вычисление включает два этапа: построение контекстов и анализ их сходства. Простейшим контекстом объекта является упорядоченный набор его ближайших соседей.

Анализ сходства контекстов можно проводить методами многокритериального оценивания. Однако значительно проще сходство контекстов описывать с помощью линейной параметрической формулы. Варьирование параметров этой линейной формы позволяет достигать различных целей исследования. Выделим два критерия сходства контекстов пары объектов (i-го и j-го):

1) s_1 – число совпадений в контекстах, то есть число элементов, находящихся в контекстах и i-го, и j-го объектов;

2) s_2 – порядок расположения совпадающих элементов в контекстах.

Порядок расположения совпадающих элементов s_2 определяется с помощью числа инверсий – числа перестановок двух соседних элементов, в результате которых совпадающие элементы одного контекста располагаются на тех же местах, что и во втором контексте [4].

Нормированные критерии s_1, s_2, изменяющиеся от 0 до 1, вычисляются по формулам:

$$u_1 = \frac{s_1}{k},$$

$$u_2 = \frac{2 * s_2}{k * (k-1)},$$

где k – длина контекста (число элементов в контексте).

Контекстная мера связи i-го и j-го объекта имеет вид:

$$\mu_{ij} = \beta * u_1 + (1-\beta) * u_2, \qquad (1)$$

где β – свободный параметр.

Изменяя β в пределах от 0 до 1, а также варьируя длину контекста k, можно получать различные меры близости, ориентированные на получение классов с различными свойствами. Чем больше совпадений в контекстах, тем более схожи сами элементы, поэтому критерий s_1 ориентирован на выделение компактных кластеров. Критерий s_2 отвечает за получение групп цепочной формы. Снижая относительный вес критерия s_1, можно от выделения групп шарообразной формы перейти к формированию групп с цепочной формой. Оптимальная длина контекста выбирается для каждой конкретной совокупности на стадии исследования результатов классификации.

Классификация на графах

После построения графа для каждой вершины (объекта) формируется его контекст – упорядоченный набор его ближайших соседей. На первом месте контекста располагается номер его ближайшего соседа; на втором месте – номер следующей по близости вершины и так далее. Выбор длины контекста зависит от степени полноты графа, но даже при высокой степени длина контекста не должна быть большой. Например, при классификации участников рос-

сийского рынка межбанковских кредитов оптимальная длина контекста выбиралась равной четырем. Близость соседа определялась объемом предоставленного кредита.

Меры сходства рассчитываются по формуле (1), в которой величина β выбирается в зависимости от цели исследования: если требуются компактные классы, то величина β близка к 1; если предполагается построение моделей связи показателей, например регрессий, то величина β выбирается ближе к 0,6. На основании мер сходства могут быть вычислены расстояния между i-м и j-м объектами:

$$R_{ij} = \frac{1}{1 + \alpha * \mu_{ij}} \text{ или } R_{ij} = e^{-\alpha * \mu_{ij}}.$$

Далее для классификации объектов может использоваться стандартный метод классификации.

Список литературы

1. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Классификация и снижение размерности: Справочное издание / Под ред. С.А. Айвазяна. – М.: Финансы и статистика, 1989. – 607 с.
2. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. – М.: ЮНИТИ, 1998. – 1022 с.
3. Бураковская Ю.Б. Теория графов. Часть 1. – Томск.: Томский политехнический университет, 2008. – 201 с.
4. Гамбаров Г.М. Проблемы статистического анализа и оценки стоимости финансовых активов. – М.: МЭСИ, 2010. – 141 с.
5. Кендалл М., Стюарт А. Многомерный статистический анализ и временные ряды. М.: Наука, 1976. – 672 с.
6. Мандель И.Д. Кластерный анализ. – М.: Финансы и статистика, 1988. – 176 с.
7. Матула Д.В. Методы теории графов в алгоритмах кластер-анализа. – В кн.: Классификация и кластер / Под ред. Дж. Вэн Райзина. – М: Мир, 1980. – С. 83–111.
8. Мандель И.Д. Кластерный анализ. – М.: Финансы и статистика, 1988. – 176 с.
9. Общая теория статистики. Учебник для вузов / под редакцией И.И. Елисеевой. 5-е изд. перераб. и доп. – М.: Финансы и статистика, 2005. – 572 с.
10. Alves, J. Brinkhoff, S. Georgiev, J.-C. H´eam, I. Moldovan, et al. Network analysis of the eu insurance sector. Technical report, European Systemic Risk Board, 2015.
11. M. Boss, H. Elsinger, M. Summer, and S. Thurner. Network topology of the interbank market. Quantitative Finance, 4 (6):677–684, 2004.