УДК 519.254

КЛАСТЕРНЫЙ АНАЛИЗ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ ДАННЫХ

А.А. Рубчинский
Национальный исследовательский университет Высшая школа экономики
Россия, 101000, Москва, Мясницкая ул., 20
Университет «Дубна»,
Россия, 141980, Дубна, Университетская ул, 19
E-mail: arubchinsky@yahoo.com

Ключевые слова: кластерный анализ, автоматическая классификация, декомпозиция графов, волатильность, фондовый рынок, Государственная Дума

Аннотация. Целью данной работы является разработка трехуровневой схемы автоматической классификации, основанной на новом понятии – ***волатильности***, как отдельных кластеров, так и классификации в целом. Волатильность представляет собой точно определяемую и легко вычисляемую величину, которая определяет стабильность, точность, надежность некоторых подмножеств исходного множества вариантов – короче говоря, целесообразность их выбора в качестве кластеров. Предложенный алгоритм находит кластеры с уровнем волатильности, не превосходящим заданный. Рассмотрены как реальные примеры (фондовый рынок, выборные органы), так и модельные (двумерные множества точек). В обоих случаях приведены содержательные выводы, не получаемые на тех же данных другими методами.

1. Введение

При исследовании систем, в которых человеческий фактор является определяющим, традиционные или (как их теперь чаще называют) жесткие математические модели оказываются мало адекватными. Поэтому кластерный анализ, т.е. выбор нескольких групп похожих в определенном смысле объектов из рассматриваемого множества, является одним из наиболее подходящих инструментов исследования в подобных слабо структуризованных ситуациях. В задачах классификации часто требуется также, чтобы выделенные кластеры образовывали разбиение исходного множества, однако отказ от этого требования представляется более реалистичным. В предлагаемом подходе может быть получено как полное разбиение, так и отдельные кластеры; это не задается заранее, а определяется в процессе реализации алгоритма в зависимости от самих данных. В качестве одного из ответов возможно указание на полное отсутствие кластеров.

Неформальный характер кластерного анализа, его различные модификации, утверждения и приложения, многочисленные подходы и методы решения подробно описаны во многих публикациях [1-7]

Основное внимание в данной работе уделено формализации, точном определении и алгоритмах подсчета важного свойства подмножеств заданного множества объектов, которое описывает их стабильность, точность, значимость – по сути дела, целесообразность их выбора в качестве кластеров. Это свойство названо ***волатильностью***. Волатильность определена как для отдельных кандидатов в кластеры, так и для рассматриваемой задачи кластерного анализа в целом.

Обычно понятия волатильности, устойчивости и т.д. связаны с изменением рассматриваемых систем во времени. Однако в предложенном подходе это не так. Волатильность определена для фиксированной задачи классификации. Суть дела в том, что разработанный алгоритм выделения кластеров, как и некоторые другие алгоритмы, состоит из повторяющихся рандомизированных шагов. На каждом шаге строится семейство подмножеств (кандидатов в кластеры). Четкие кластеры с нулевой волатильностью, выделенные на каждом шаге, полностью совпадают друг с другом. Менее четкие кластеры на разных шагах могут слегка отличаться друг от друга и/или вообще отсутствовать (на некоторых шагах). Эти соображения позволяют сформулировать простой формальный алгоритм вычисления волатильности каждого кластера. Волатильность всей классификации определяется как взвешенная сумма волатильностей найденных кластеров. Алгоритм подробно описан в разделе 5

Представляется, что высокий уровень волатильности соответствует трудности задач классификации, и осознание этой зависимости и приводит к новому алгоритму. Алгоритм находит кластеры с произвольным уровнем волатильности (включая обычный случай нулевой волатильности), что позволяет ему справляться как с «трудными», так и с «легкими» (в указанном смысле) задачами. Более того, допустимый уровень волатильности является одним из очень немногих параметров алгоритма. Он же является единственным, который определяется человеческим решением, что является само по себе неожиданным в столь плохо структуризованных ситуациях.

Цель исследования состоит в разработке нового алгоритма кластерного анализа, удовлетворяющего следующим требованиям:

- результаты кластеризации не противоречат интуитивно ясным ответам в разнообразных двумерных случаях;
- не делается никаких предположений относительно исходного множества объектов геометрического, вероятностного и любого другого характера;
- число кластеров определяется только в процессе выполнения алгоритма;
- в алгоритм входит очень мало параметров с ясным объяснением;
- человеческое решение (если оно необходимо) делается только на заключительной стадии алгоритма.

Рассматриваемое формальное представление задач кластеризации и структура предложенного алгоритма описаны в разделе 2. Предложенный алгоритм дихотомии описан в разделе 3. Построение упомянутого семейства множеств – кандидатов в кластеры, повторяющееся на каждой внешней итерации алгоритма, описано в разделе 4. Окончательное выделение кластеров, сопровождаемое вычислением их волатильности, описано в разделе 5. Результаты применение предложенного алгоритма кластеризации и сравнение с другими методами даны в разделе 6.

2. Структура алгоритма

В предложенном подходе исходные данные о близости вариантов представлены в хорошо известном виде графа соседства (см., например, Luxburg, 2007). Суть дела в том, что для каждого объекта *a* несколько (4-5) ближайших объектов считаются близкими к *a*. Соответствующие вершины соединены в графе соседства ребром. Близость объектов определяется по исходному описанию: данной матрице схожести / несхожести, матрице «объект-свойства» и другими способами.

После построения графа соседства в алгоритме используется только информация, заключенная в этом графе. Никакие другие сведения (в частности, о точных расстояниях между исходными объектами) далее не используются. Конечно, обоснованием тако-

го подхода к представлению данных для кластерного анализа (как и к самому алгоритму) могут служить только экспериментальные результаты. Тем не менее, можно сказать, что модель графа соседства является одной из самых «мягких» моделей, и ей можно пользоваться, когда для каждого объекта каким-то образом (например, экспертным) указано несколько наиболее похожих на него объектов без каких-либо точных значений расстояния.

Другим преимуществом такого представления является расширение потенциальной области применения. В рамках рассматриваемого подхода оказывается и задача о декомпозиции произвольных неориентированных графов, которая сама по себе является важной в теоретическом и практическом отношении.

Алгоритм представляет собой трехуровневую процедуру. На ***внешнем*** уровне по семействам множеств-кандидатов в кластеры, найденных при каждом из r прогонов процедуры дивизимно-агломеративного кластерного алгоритма (далее ДАК-алгоритма) строятся финальные кластеры. Алгоритм построения финальных кластеров описан в разделе 5.

Построение семейства кандидатов в кластеры ДАК-алгоритмом представляет собой ***промежуточный*** уровень предложенной трехуровневой процедуры. Краткое описание ДАК-алгоритма дается в разделе 4.

Сам ДАК-алгоритм основан на алгоритме дихотомии графов предложенном в [8] (см. раздел 3). Алгоритм дихотомии многократно применяется при реализации ДАК-алгоритма; он представляет собой ***внутренний*** уровень общей трехуровневой процедуры. Этот алгоритм является новой модификацией частотного подхода к кластерному анализу, предложенному в статье [9]. Как и некоторые другие алгоритмы, включая спектральные [4], он также дает некоторую аппроксимацию решения известной задачи о разрезе с лучшим делением (Ratio-cut problem), упоминаемой в разделе 3. В отличие от упомянутых в [4] подходов, предложенный алгоритм позволяет выявить недостатки дихотомий, полученных не только приближенным, но и точным решением данной *NP*-полной задачи. Однако в рамках ДАК-алгоритма на промежуточном уровне даже очевидно «неправильные» дихотомии приводят к правильным классификациям, а их построение не требует точного решения указанной *NP*-полной задачи. Эти два уровня (внутренний и промежуточный) были описаны в препринте [8]. Использование внешнего уровня и понятия волатильности оказалось связанным с реальными задачами классификации социально-экономических данных, которые не всегда укладываются в описанную в цитированной работе двухуровневую схему.

3. Внутренний этап процедуры

В статье [9] был предложен принципиально новый подход к декомпозиции графов и тем самым – к задаче автоматической классификации. Эта идея была подробно обсуждена в работе [8], поэтому здесь мы сразу переходим к предложенному варианту частотного алгоритма. Суть дела в том, что в ранее предлагавшихся частотных алгоритмах пути, соединяющие очередную пару вершин, проводятся независимо от предыдущих путей. Однако учет всех ранее проведенных путей позволяет получать разрезы между двумя группами вершин, в которых все ребра имеют одинаковую максимальную частоту. Тогда одновременное и однократное удаление всех ребер с максимальной частотой определяет искомую дихотомию графа.

Минимаксный алгоритм подробно описан в препринте [8], поэтому здесь он не излагается. Важно, что, в отличие от ранее известных вариантов частотных алгоритмов, он реализует приближенное решение некоторой оптимизационной задачи на графе, вы-

ражающей разумное (хотя, как и в других случаях, неполное) представление о правильности классификации (подробнее см. в комментариях в конце этого раздела). Укажем также, что параметрами алгоритма являются два целых числа:

- максимальное начальное значение частоты f;
- число повторений для набора статистики T.

В работе [8] установлена связь между предложенным алгоритмом и известными оптимизационными постановками поиска сбалансированного разреза в графе. Приведем основные выводы.

a) Разрез, найденный минимаксным алгоритмом, приближенно максимизирует введенную декомпозиционную функцию графа

$$D(A,B) = \frac{|A|\times|B|}{d(A,B)}, \tag{1}$$

где $d(A,B)$ – число ребер в разрезе, отделяющем A от B.

b) Введенная в уже цитированном обзоре [4] задача минимизации

$$R(A,B) = d(A,B)\times\left(\frac{1}{|A|}+\frac{1}{|B|}\right)\to \min \tag{2}$$

на множестве всех разрезов исходного графа (называемая RatioCut Problem), эквивалентна максимизации функции $D(A,B)$ на том же множестве. Это сразу следует из сравнения формул (1) и (2).

Тем самым предложенный алгоритм является весьма эффективным (в экспериментальном плане) приближенным методом решения той же RatioCut Problem, для которой используются спектральные и ядерные (kernel) методы. Однако основной вопрос, связанный с данной *NP*-полной оптимизационной задачей, состоит не в поиске ее приближенных решений, а в том, насколько функции D и R адекватны задачам декомпозиции графа соседства, или, точнее, насколько ее приближенная максимизация предложенным минимаксным алгоритмом позволяет находить интуитивно правильные дихотомии. Естественно, этот вопрос является содержательным и ответ на него может быть дан только примерами.

Пример 1. Дихотомии, полученные предложенным минимаксным алгоритмом для шести различных двумерных множеств, представлены на рис. 1. На нем, как и на рисунках из [8], представляющих результаты классификации, оставлены только те ребра, которые соединяют вершины, попавшие в разные классы; условная линия, пересекающие эти ребра, разделяет найденные классы. Сам же алгоритм дихотомии относит каждую точку к одному из двух классов. Во всех шести случаях не только использовалась одна и та же программа, но и немногие изменяемые параметры были одними и теми же: $f = 10$, $T = 1000$. При других f и T результаты были теми же самыми. Они не зависят и от начального значения случайного датчика. Во всех случаях результаты не противоречат интуитивному представлению о правильности классификации, которое в данном случае не вызывает сомнений ■

Однако так бывает не всегда, что и инициировало разработку описываемого в следующем разделе ДАК-алгоритма, в котором предложенный метод дихотомии используется в качестве основного шага на дивизимном этапе (см. раздел 4). Для объяснения необходимости более глубокого анализа рассмотрим следующий

Пример 2. Рассматриваются два двумерных множества, показанные на рис. 2a и 2c. Результат дихотомии для множества рис. 2a показан на рис. 2b. Как и во всех 6-и случаях, показанных на рис. 2, он не зависит от инициализации случайного датчика. Найденный минимаксным алгоритмом разрез максимизирует декомпозиционную функцию (1) на множестве всех разрезов графа соседства и определяет интуитивно верную классификацию на два класса. Естественно, этот же разрез минимизирует функцию (2).

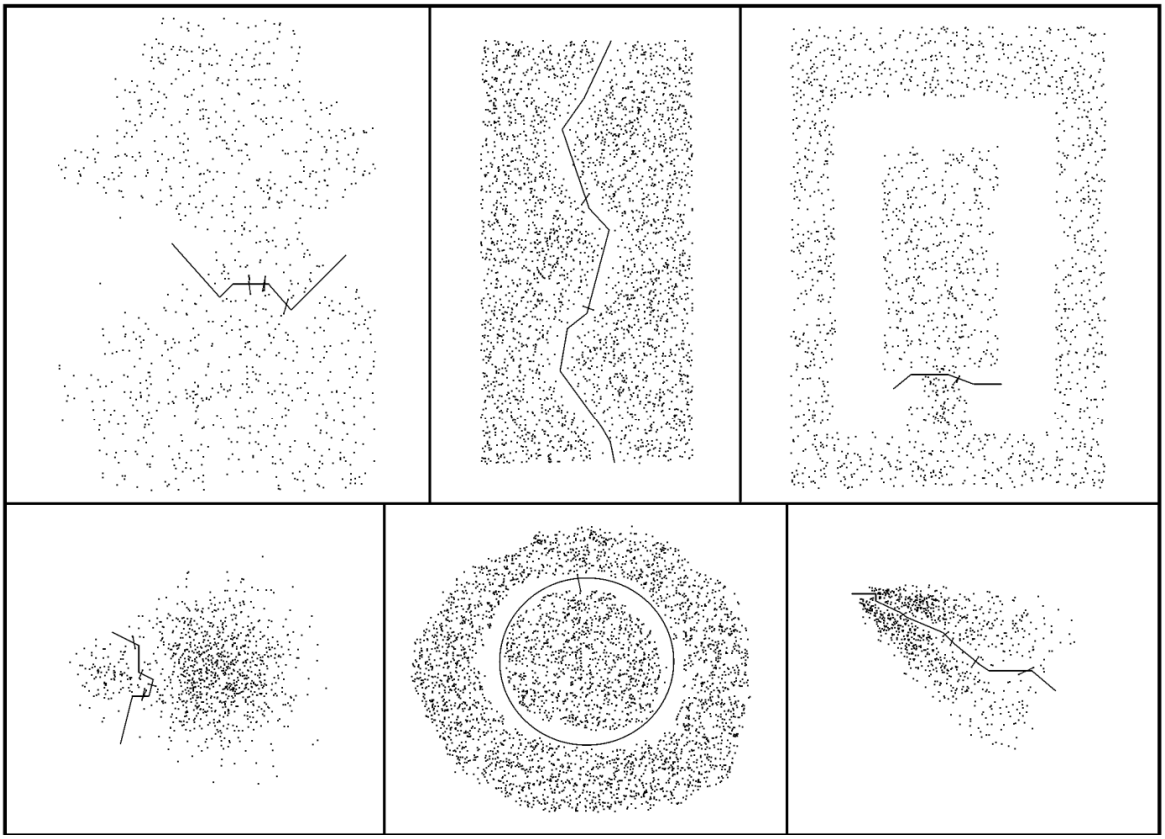

Рис. 1. Решение шести простых двумерных задач классификации.

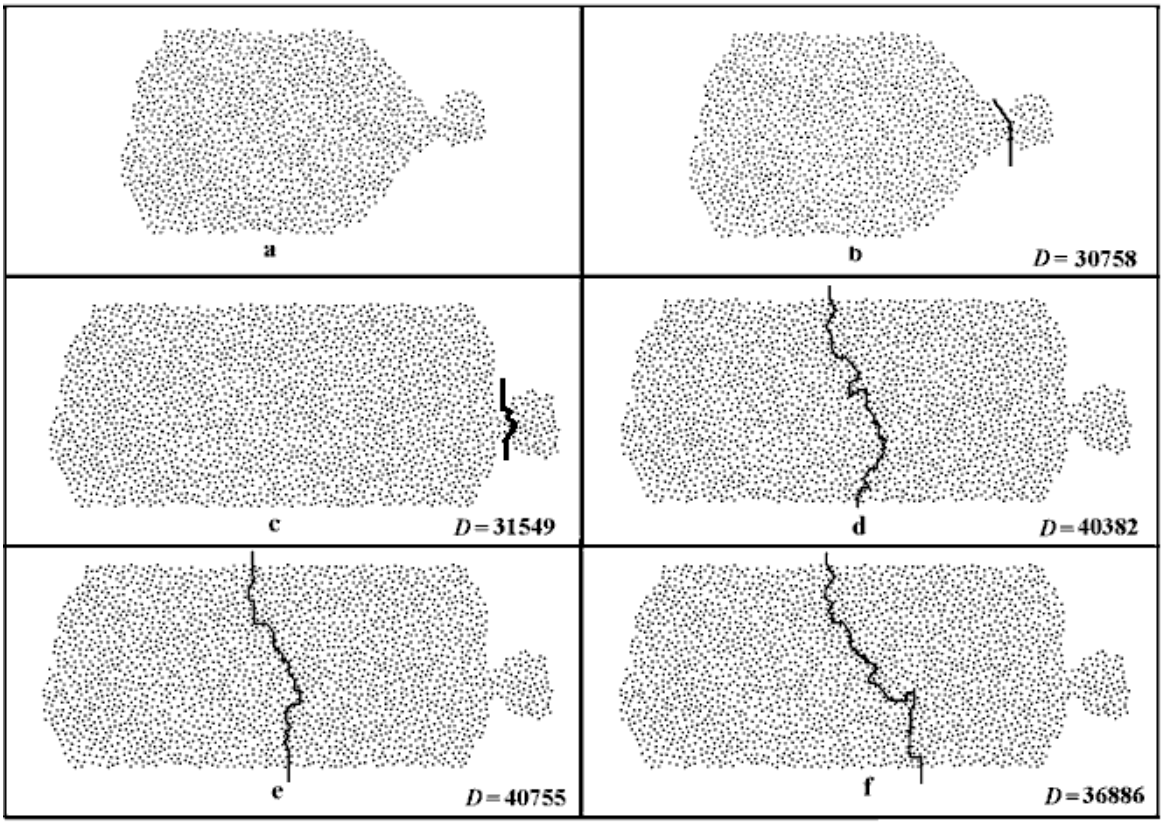

Рис. 2. Другие примеры дихотомии

В то же время при использовании того же самого алгоритма для похожего множества, показанного на рис. 2c, результат заметно зависит от инициализации случайного датчика, как это видно на рис. 2d, 2e и 2f. Во всех этих случаях найденное решение не совпадает с интуитивно очевидным. Наконец, значение декомпозиционной функции для «правильного» разреза равно 31549, а для неверного разреза, найденного минимаксным алгоритмом и показанного на рис. 2d, эта функция равна 40382. В других двух случаях эта функция также существенно больше, чем ее значение на правильном разрезе. Причем речь идет о точных, а не приближенных значениях декомпозиционной функции. Как уже говорилось, все то же самое относится и к минимизации критерия RatioCut. Это еще раз говорит о той осторожности, с которой надо относиться к достаточно популярным критериям сбалансированности (как и ко многим другим формальным моделям классификации) ■

Причина неудачи критерия (1) (и эквивалентного ему критерия (2)) в рассматриваемом случае вполне понятна. Отношение максимального и минимального числа точек, принадлежащих «правильным» классам, в примере рис. 2c существенно больше, чем в примере рис. 2a и во всех примерах рис. 1. Поэтому произведение числа точек $|A|\times|B|$, являющееся числителем в (1), настолько мало по отношению к произведению мощностей подмножеств, близких к половине исходного множества, что это уже не компенсируется знаменателем, равным сравнительно небольшому числу ребер в «правильном» разрезе.

Принимая во внимание результаты десятков вычислительных экспериментов с различными данными, приходим к следующим неформальным выводам.

1) Известный критерий RatioCut (и, следовательно, аппроксимирующие его спектральные и ядерные методы) могут давать неверные результаты во многих относительно простых случаях.
2) Все стохастически устойчивые дихотомии, найденные предложенным минимаксным алгоритмом, интуитивно правильны и максимизируют критерий (1).
3) Все стохастически неустойчивые дихотомии, найденные предложенным минимаксным алгоритмом, интуитивно неправильны, а числовые значения критерия (1) превосходят значения на «правильном» разрезе.

Однако само понятие устойчивости не является точно определенным. Между очевидно устойчивыми и очевидно неустойчивыми ситуациями есть «серая зона» слабой нестабильности. Подобно разнообразным ситуациям такого типа, встречающимся во многих разделах чистой и прикладной математики, такие промежуточные ситуации в некотором смысле неизбежны, а все самое интересное и важное происходит именно в таких промежуточных зонах. Реальные примеры в разделе 6 показывают, что такие явления действительно встречаются в кластерном анализе, что не только инициирует, но и в некотором смысле оправдывает разработку предложенного подхода к задачам классификации, не только объясняющего, но и использующего в алгоритмах нестабильность кластеров.

4. Промежуточный этап процедуры

В настоящем разделе описывается ДАК-алгоритм промежуточного этапа (см. раздел 2). Его блок-схема показана на рис. 3. Входом алгоритма является исходный неориентированный граф. Выход алгоритма будет определен далее. Единственным параметром ДАК-алгоритма является максимальное число K частей, на которые делится исходное множество на дивизимном этапе.

Рис. 3. Блок-схема ДАК-алгоритма.

Кратко рассмотрим этапы блок-схемы рис. 3 по отдельности. Подробно они описаны в [8].

1) ДИВИЗИМНЫЙ ЭТАП. Исходный граф последовательно делится на две подграфа минимаксным алгоритмом дихотомии, упомянутым в разделе 3. Для деления на каждом шаге выбирается тот из уже построенных графов, у которого число вершин максимально (а при равенстве – тот, который был построен раньше). Таким образом, на i-ом шаге получается классификация исходного множества на $i+1$ класс (i = 1, …, K–1). Выходом этапа является семейство вложенных классификаций $D = (D_2, D3, ..., D_K)$ на 2, 3, …, K классов.

2) АГЛОМЕРАТИВНЫЙ ЭТАП. Каждая из классификаций D_j на j классов определяет подсемейство классификаций: на j классов (сама D_j), на (j–1) классов (полученная из D_j путем объединения двух подграфов, связанных максимальным числом ребер); и т.д., в соответствии с обычной агломеративной схемой (последовательно соединяя подграфы, связанные максимальным числом ребер), вплоть до классификации на 2 класса. Обозначим полученные классификации через $C^j_j, C^j_{j-1}, ..., C^j_2$ и представим их все следующим образом:

$$C_2^2, C_2^3, ..., C_2^{k-1}, C_2^k$$
$$C_3^3, C_3^4, ..., C_3^k$$
$$\cdots\cdots\cdots\cdots$$
$$C_3^3, C_3^4$$
$$C_k^k$$

Все классификации, расположенные в 1-ой строке, являются классификациями на 2 класса, во 2-ой строке – на 3 класса, и т.д., вплоть до последней строки, содержащей единственную классификацию C^k_k на K классов. Все указанные в приведенном списке классификации образуют выход агломеративного этапа.

3) ЭТАП ВЫДЕЛЕНИЯ МНОЖЕСТВ. Все различные множества, вошедшие хотя бы в одну из найденных выше классификаций, образуют выход ДАК-алгоритма, т.е. всего промежуточного уровня трехуровневой схемы. Заметим, что этих множеств будет не так уж много: по построению, все они либо совпадают с классами единственной классификации C^k_k на K классов, либо являются объединениями некоторых из этих классов. Нетрудно понять, что число всех таких множеств не может быть очень велико:

самая грубая оценка дает $K^3/4$, но эксперименты показывают, что их обычно гораздо меньше.

5. Внешний этап процедуры

Обозначим через *r* число прогонов ДАК-алгоритма. Поскольку на каждом прогоне используются случайные числа (например, для выбора последовательных пар вершин в минимаксном алгоритме построения равномерного разреза), на каждом прогоне выходом ДАК-алгоритма будут, вообще говоря, различные множества, хотя многие из них в достаточно простых случаях будут совпадать. Степень совпадения множеств (далее точно определяемая) и послужит основой для алгоритма построения финальных кластеров.

Введем необходимые понятия и обозначения. Обозначим через U_i множество всех потенциальных кластеров, найденных при *i*-ом прогоне ДАК-алгоритма. Для краткости будем называть просто кластерами.

Пусть F – произвольное семейство кластеров, принадлежащих различным множествам U_i. Представим F в следующем виде:

(3) $\mathrm{F} = \langle F_{i1}, ..., F_{id} \rangle$, где $F_{ik} \in U_{ik} (k = 1, ..., d)$, и $s < t$ влечет $i_s < i_t$.

Обозначим

(4) $A(\mathrm{F}) = \cap F_j, B(\mathrm{F}) = \cup F_j, \alpha(\mathrm{F}) = |A(\mathrm{F})| / |B(\mathrm{F})|,$

где пересечение и объединение берется по всем множествам *F* из семейства F. Ясно, что $\alpha(\mathrm{F})$ не может превзойти 1. Семейства F, такие, что $\alpha(\mathrm{F}) > 0.5$, называются **α-стабильными**. Близость $\alpha(\mathrm{F})$ к 1 означает стабильность некоторого кластера при всех рассматриваемых прогонах с номерами $i_1, ..., i_d$. Это понятие иллюстрируется на рис. 4.

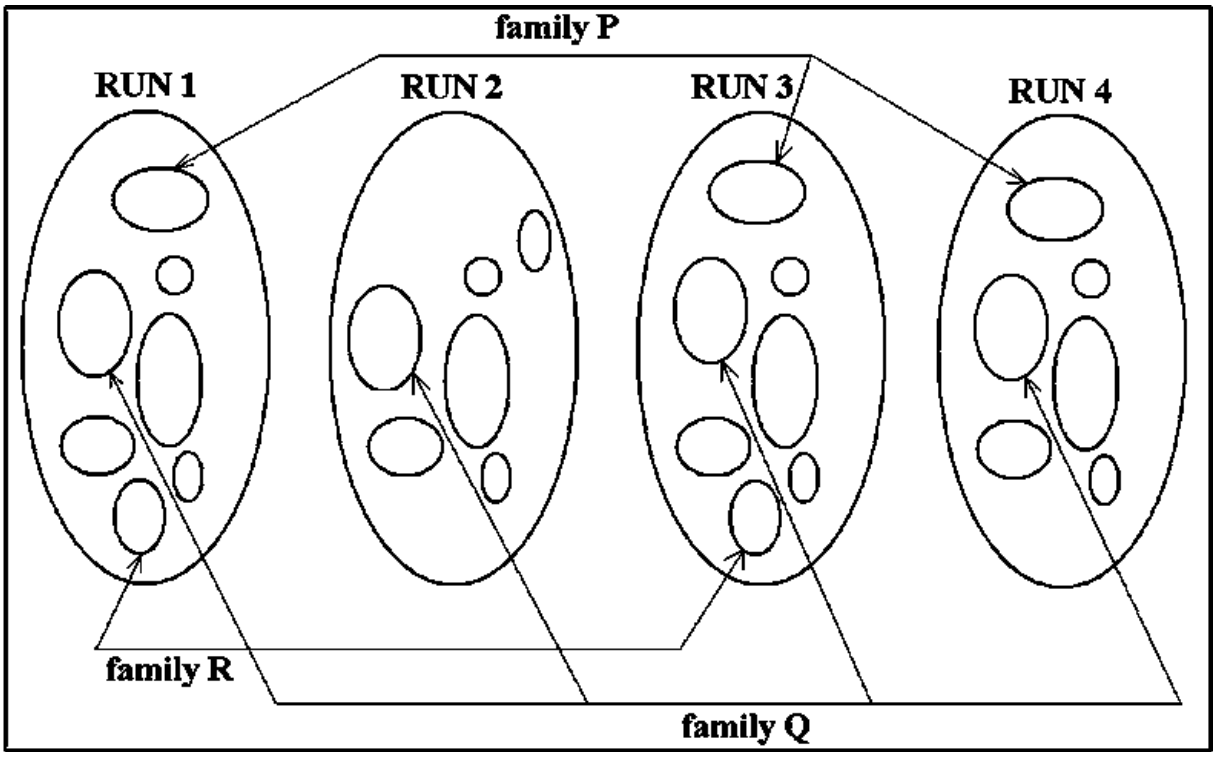

Рис. 4. Гипотетические результаты 4-ех прогонов.

Рис. 4 показывает, что кластеры из семейства P появляются 3 раза из 4-ех, кластеры из семейства Q – 4 раза из 4-ех, и кластеры из семейства R – появляются 2 раза из 4-ех. Этот пример приводит к другому понятию стабильности. Обозначим через $c(\mathrm{F}) = |\mathrm{F}| = d$ (см. (7)). Положим $\beta(\mathrm{F}) = c(\mathrm{F}) / r$. Это число указывает на долю прогонов, в которых рассматривается некоторое семейство кластеров F, по отношению ко всем *r* прогонам. Наконец, положим

(5) $$\gamma(\mathrm{F}) = \alpha(\mathrm{F}) \times \beta(\mathrm{F}).$$

Число $V(\mathrm{F}) = 1 - \gamma(\mathrm{F})$ называется ***волатильностью семейства*** F. Предположим, что все семейства F (см. (3)) упорядочены по возрастанию волатильности: $\mathrm{F}_1, \mathrm{F}_2, ...,$ так что $s < t$ влечет $\mathrm{F}_s < \mathrm{F}_t$. Положим $C_i = A(\mathrm{F}_i),\ i = 1, 2, ...,$ (см (4)). Предположим также, что задано число V^* – максимально допустимый уровень волатильности.

Следующие шаги алгоритма построения кластеров определяют предложенное решение задачи кластеризации.

Алгоритм построения кластеров

1) Найти все α-стабильные семейства $\mathrm{F}_1, \mathrm{F}_2, ..., \mathrm{F}_m$ (см. (4)).

2) Выделить из них все семейства F, такие что $V(\mathrm{F}) \le V^*$ (они названы допустимыми).

3) Упорядочить допустимые семейства F_i в порядке возрастания $V(\mathrm{F}_i)$.

4) Положить $C_i = A(\mathrm{F}_i),\ i = 1, 2, ..., k$ (здесь *k* – число допустимых семейств).

5) Положить $D_1 = C_1$ и текущее значение $i_c = 1$.

6) Если множества $D_1, ..., D_t$ уже найдены, рассмотреть последовательно все $i > i_c$ до тех пор, пока не будет выполнено хотя бы одно из следующих двух условий:

- C_i содержит D_j ($j \in 1, \ldots, t$) или D_j содержит C_i ($j \in 1, \ldots, t$) или C_i не пересекается с D_j ($j = 1, \ldots, t$);
- $i = k+1$.

В 1-ом случае положить $D_{t+1} = C_i$, $i_c = i$, $t = t+1$ и вернуться к шагу 6. Во 2-ом случае перейти к следующему шагу 7.

7) Рассмотреть все кластеры $D_1, ..., D_t$ и удалить из них кластеры, содержащие другие кластеры из этого же списка.

8) Остановка алгоритма ■

Построенные множества $D_1, ..., D_S$ образуют выход внешнего уровня трехуровневого алгоритма кластеризации. Другими словами, они являются найденными кластерами. Эти кластеры могут образовывать полное разбиение исходного множества объектов, или могут образовывать его неполное разбиение. Волатильность $V(D)$ кластера *D* определяется как волатильность семейства F такого, что $D = A(\mathrm{F})$. Волатильность всей задачи кластеризации определяется как взвешенная сумма волатильности всех найденных кластеров:

$$V = \sum_{i=1}^{S} V(D_i) |D_i| \Big/ \sum_{i=1}^{S} |D_i| \ .$$

5.1. Параметры трехуровневой процедуры. Имеется очень немного, особенно для столь универсальной схемы, параметров. Два параметра *f* и *T* предложенного частотного алгоритма на внутреннем уровне не оказывают заметного влияния на результаты. Во всяком случае, результаты при любых *f* в пределах 5-25 и любых *T* в пределах 500 – 3000 практически совпадают. Единственный параметром ДАК-алгоритма (про-

межуточный уровень) является число частей K. Этот параметр является существенным. Грубо говоря, при слишком малом K правильные классификации могут не быть найденными, а при слишком большом K могут появиться «лишние» устойчивые классификации. Однако при всех промежуточных K находятся одни и те же правильные классификации. В общем случае K можно выбирать «с запасом», заметно бóльшим, чем предполагаемое число классов. Наконец, на внешнем уровне при построении финальных кластеров задается два параметра – число прогонов ДАК-алгоритма r и допустимый уровень волатильности V^*. Во всех экспериментах произвольный выбор r в пределах 5 – 10 не оказывал влияния на результаты. Однако допустимый уровень волатильности V^* в наиболее сложных случаях действительно может оказаться существенным: он влияет на финальное число кластеров. Более того, подсчитанный реальный уровень волатильности классификации (а не задаваемый заранее допустимый уровень V^*) является в некоторых случаях важной характеристикой рассматриваемой ситуации (см. ниже **Кластеры в депутатском корпусе**).

6. Примеры

Модельные примеры, относящиеся в основном к двумерным множествам точек, были подробно рассмотрены в препринте [9] и здесь не приводятся (за исключением рис. 1). Там же приведены результаты сравнений с основными известными методами кластерного анализа, которые оказались неработоспособными в достаточно простых ситуациях (показанных на рис. 1 и некоторых других). В этом разделе рассмотрены реальные данные, относящиеся к фондовым рынкам и голосованиям в выборных органах.

Анализ фондовых рынков. Использование кластерного анализа для исследования фондовых рынков достаточно хорошо известно (см, например, статью [10]). В цитированных там публикациях фондовому рынку сопоставляется рыночный граф $G\theta$, определяемый следующим образом. Вершины графы соответствуют различным акциям; ребро соединяет две вершины тогда и только тогда, когда коэффициент корреляции между курсами акций, соответствующими этим двум вершинам, не меньше θ. Предполагается, что клики в рыночном графе (при различных значениях θ) соответствуют акциям с «близким поведением». Конструкция рыночного графа совпадает с хорошо известной конструкцией порогового графа в автоматической классификации, в которой роль матрицы несхожести играет корреляционная матрица.

В настоящей работе корреляционной матрице сопоставляется другой хорошо известный в автоматической классификации граф – упоминаемый в начале раздела 2 граф соседства. Эта конструкция представляется более гибкой. Во-первых, она позволяет учесть различный масштаб в различных частях исходного множества объектов; во вторых, в ней не требуется выделение в качестве компонент графа его клик, что представляется слишком жестким требованием; в третьих, не требуется определять «правильный» уровень θ или рассматривать несколько таких уровней.

Результаты применения разработанного подхода представлены в виде групп акций на фондовом рынке США и России. Исходными матрицами несхожести служили корреляционные матрицы с октября 2008 по октябрь 2010 года: для 500 случайно выбранных (из более чем 5000) компаний США, 151 российских компаний и 266 шведских компаний. На шведском фондовом рынке кластеры не были обнаружены, что является достаточно важным содержательным фактом. Наиболее интересными выводами из полученных результатов являются следующие.

1) При определении кластеров область деятельности никак не учитывалась – использовались только попарные коэффициенты корреляции. Однако оказалось, что все найденные кластеры состоят в основном из акций компаний, занятых в одной и той же или достаточно близких видах деятельности. При этом существует достаточно много фирм, чьи акций имеют близкие коэффициенты корреляции, занятых различными видами деятельности. Таким образом, предложенный метод кластерного анализа учитывает не только близость внутри одной группы объектов, но и их расстояние относительно других групп, что и позволяет получить достаточно разумные кластеры.

2) Среди найденных кластеров оказался лишь один с 0-ой волатильностью; волатильность остальных меняется от 0,125 и достигает значительного уровня 0,583.

3) Из 500 компаний в кластеры вошло 84. Остальные компании не образуют кластеров, т.е. среди них нет ни одного α-стабильного семейства.

4) Отличие российского фондового рынка от рынка США состоит в том, что среднее значение коэффициентов корреляций двух найденных там кластеров (оба состоят из акций компаний, работающих в электроэнергетике) оказалось низким (около 0,2). Тщательная формальная проверка подтвердила эти данные. Кластер 2 с теми же самыми вершинами появляется 4 раза из 5 прогонов алгоритма. Расстояние его вершин от других ближайших к ним превосходит расстояние между вершинами из кластера. Однако такое значительное отличие от кластеров на рынке США нуждается в содержательном объяснении, которого пока еще нет.

Приведем найденные на рынке США кластеры в виде списков соответствующих видов деятельности, с указанием волатильности V.

Кластер 1. $V = 0$; {Gold mining, Gold mining, Gold mining, Gold mining, Gold mining, Color Metallurgy, Color Metallurgy}.

Кластер 2. $V = 0{,}125$; {Hotels, Building, Investment, Investment, Finances, Investment, Investment, Real Estate, Real Estate, Real Estate, Finances, Real Estate, Real Estate}.

Кластер 3. $V = 0.167$; {Insurance, Insurance, Insurance, Insurance, Insurance, Insurance}.

Кластер 4. $V = 0{,}167$; {Index, Index, Index, Investment, Investment, Investment, Investment, Investment, Investment}.

Кластер 5. $V = 0{,}286$; {Oil and Gas, Coal Mining, Oil and Gas, Wholesale Trading, Oil and Gas}.

Кластер 6. $V = 0{,}356$; {Bank, Finances, Medical Equipment, Bank, Insurance, Semiconductors, Bank, Investment, Services, Investment, Finances, Bank, Finances, Finances, Bank, Bank, Bank, Information Technologies, Pharmaceutics, Finances, Bank, Bank}.

Кластер 7. $V = 0{,}559$; {Index, Aluminum, Pharmaceutics, Index, Fertilizations, Index, Telecommunications, Semiconductors, Insurance. Information Technologies, Investment, Investment, Retail Trading, Food and Beverage Sale, Transport, Oil and Gas}.

Кластер 8. $V = 0{,}583$; {Coal Mining, Oil and Gas, Oil and Gas, Oil and Gas, Coal Mining}.

Кластеры на фондовом рынке при тех же самых исходных данных по США и России строились в работах [11, 12]. В этих работах использовались другие подходы и были выделены другие кластеры. Однако очевидный кластер из фондового рынка США, содержащий акции 7-и компаний, занятых добычей и производством золота, выделен не был. Как на российском рынке, так и на рынке США выделенные кластеры состояли из акций компаний, действующих в различных сферах. Типичным примером является кластер, состоящий из акций следующих компаний: БанкВТБ, НорильскийНикель, Газпром, Лукойл, Роснефть, Сбербанк. Кластеры, в которых большинство акций относится к близким видам деятельности, практически не было выделено. Наконец, различные пороги θ определяют различные семейства кластеров.

Кластеры в депутатском корпусе. Деятельность Государственной Думы рассмотрена за 5-месячный период с 01.09.2001 по 31.01.2002. Этот период представляется важным, поскольку в нем произошло важное политическое событие – создание партии «Единая Россия» 01.12.2001. Было построено пять классификаций по результатам голосований в течение каждого из 5 месяцев этого периода. В каждом месяце учитывались все (а не только политически значимые) голосования (от 200 до 400 в месяц). Каждому i-му депутату (i = 1, 2, …, 479) был сопоставлен вектор $v_i = (v_1^i, v_2^i, ..., v_n^i)$, где n – число голосований в данном месяце,

$$v_j^i = \begin{cases} 1, \text{если } i \text{ - ый депутат голосовал за } j \text{ - ое предложение;} \\ -1, \text{если } i \text{ - ый депутат голосовал против } j \text{ - го предложения;} \\ 0, \text{в противном случае (ввоздержася или не участвовал).} \end{cases}$$

Несходство d_{st} между s-ым и t-ым депутатами определяется как обычное евклидово расстояние между векторами v_S и v_t. Матрица несхожести $D = (d_{st})$ была исходной для определения депутатских кластеров и их волатильности предложенным трехуровневым методом, описанным выше. Волатильность кластеров описана (в зависимости от месяца) в следующей таблице:

Таблица 1. Волатильность депутатских кластеров

Сентябрь	0; 0; 0; 0; 0
Октябрь	0; 0; 0; 0; 0; 0; 0; 0
Ноябрь	0; 0; 0,022; 0,200; 0,260; 0,315
Декабрьr	0; 0; 0; 0,010; 0,012; 0,074; 0,125
Январь	0; 0; 0; 0,020; 0,035; 0,060; 0,144

Заметим, что число кластеров не является постоянным. Более того, они не всегда соответствуют депутатским фракциям. Однако наиболее важным фактом в полученных данных является резкий рост волатильности непосредственно перед возникновением «Единой России» и незначительное уменьшение волатильности в последующие два месяца. Достаточно подробно работа Государственной Думы 3-его созыва проанализирована в книге [13]. Однако рассмотренные там индексы и показатели не продемонстрировали всплеска, падения или других особых элементов поведения в окрестности политической «особой точки» 01.12.2001. Можно сказать, что примененный подход может оказаться полезным в ситуациях двухпартийных парламентов или абсолютного большинства одной из партий, когда обычные индексы влияния и согласованности мало информативны, несмотря на достаточно острую политическую борьбу.

7. Заключение

В большинстве случаев волатильность ассоциируется с хаосом и, как следствие, она имеет негативную коннотацию. Но в предложенном формальном использовании данного термина это не так. Можно предположить, что высокая волатильность – в данном случае отсутствие кластеров – фондового рынка указывает на его регулярный характер (как на шведском фондовом рынке). Наоборот, наличие четких кластеров с нулевой или низкой волатильностью указывает на наличие специальных видов деятельности (как добыча золота) или на некоторые специальные действия по согласованию между фирмами, образующими кластер (особенно при малом размере этого кластера). Можно предположить, что тщательное отслеживание фондового рынка и его отдельных кластеров с точки зрения

волатильности может помочь в предсказании так называемых «черных лебедей» на этих рынках. Но сейчас это не более чем предположение, для обоснования которого понадобится много аналитической и экспериментальной работы.

Ситуация с выборными органами является другой. Можно предположить, что как очень низкий, так и очень высокий уровень волатильности может быть нежелателен. И здесь требуется большая аналитическая и экспериментальная работа для выработки и проверки различных предположений.

Список литературы

1. Браверман Э.М., Мучник И.Б. Структурные методы обработки эмпирических данных. М.: Наука, 1983. 464 с.
2. Filippone M., Camastra F., Masulli F. Rovetta S. A Survey of Kernel and Spectral Methods for Clustering // Pattern Recognition, 2008. Vol. 41, No. 1. P. 176-190.
3. Gordon A.D. Classification. Chapman & Hall/CRC, 1999. 293 p.
4. Luxburg U. A Tutorial on Spectral Clustering // Statistics and Computing. 2007. Vol. 17, No. 4. P. 395-416.
5. Mirkin B. Mathematical Classification and Clustering. Kluwer Academic Publishers, 1996.
6. Mirkin B. Clustering for Data Mining: A Data Recovery Approach. Chapman & Hall/CRC, 2005.
7. Mirkin B. Core Concepts in Data Analysis: Summarization, Correlation, Visualization. Springer, 2010.
8. Rubchinsky A. Divisive-Agglomerative Classification Algorithm Based on the Minimax Modification of Frequency Approach: Preprint / NRU HSE / WP7/2010/07. M.: 2010. 48 p.
9. Girvan, M., Newman, M.E.J. Community structure in social and biological networks // Proc. Natl. Acad. Sci. USA, 2002. Vol. 99. P. 7821-7826.
10. Wei-Qiang Huang, Xin-Tian Zhuang, Shuang Yao. A network analysis of the Chinese stock market // Physica A. 2009. Vol. 388. P. 2956-2964.
11. Визгунов А.Н., Гольденгорин Б.И, Замараев В.А. и др. Применение рыночных графов к анализу фондового рынка России // Журнал Новой экономической ассоциации. 2013. Т.15, №3. С. 66-81
12. Кочетуров А.А., Бацын М.В., Пардалос П.М., Гольденгорин Б. И. Анализ финансовых рынков средствами модели о *p*-медианах. Доклад // Лаборатория алгоритмов и технологий анализа сетевых структур, НИУ ВШЭ, 2013 // http://fs.nashaucheba.ru/docs/46/index-6217357.html#689383.
13. Алескеров Ф.Т., Благовещенский Н.Ю. Сатаров Г.А. и др. Влияние и структурная устойчивость в Российском парламенте (1905-1917 и 1993-2005 гг.). М.: ФИЗМАТЛИТ, 2007. 312 с.