ВЫСШАЯ ШКОЛА ЭКОНОМИКИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Б.Г. Миркин, М.А. Орлов

МЕТОДЫ МНОГОКРИТЕРИАЛЬНОЙ СТРАТИФИКАЦИИ И ИХ ЭКСПЕРИМЕНТАЛЬНОЕ СРАВНЕНИЕ

Препринт WP7/2013/06
Серия WP7
Математические методы
анализа решений в экономике,
бизнесе и политике

Москва
2013

УДК 316.342.5
ББК 60.5
М63

Редакторы серии WP7
«Математические методы анализа решений в экономике,
бизнесе и политике»
Ф.Т. Алескеров, В.В. Подиновский, Б.Г. Миркин

Миркин, Б. Г., Орлов, М. А. Методы многокритериальной стратификации и их
М63 экспериментальное сравнение [Текст] : препринт WP7/2013/06 / Б. Г. Миркин, М. А. Орлов ; Нац. исслед. ун-т «Высшая школа экономики». – М. : Изд. дом Высшей школы экономики, 2013. – 32 с. – (Серия WP7 «Математические методы анализа решений в экономике, бизнесе и политике»). – 50 экз.

При многомерном ранжировании интерес может представлять не столько сама упорядоченность рассматриваемых объектов, сколько ее разбиение на группы более или менее эквивалентных объектов – следуя социологии и минералогии, такие группы можно называть стратами. Такова, например, известная «одномерная» задача об ABC-разбиении. В данной работе предлагаются два метода автоматической стратификации (при заданном числе страт). Один подбирает веса критериев таким образом, чтобы объекты каждой отдельной страты были как можно ближе друг к другу на оси агрегированного критерия. Другой, наоборот, исходит из принципа несравнимости критериев и формирует страты из последовательных границ Парето. Эти методы экспериментально сравниваются с известными методами многомерного ранжирования, в связи с чем в работе рассматриваются некоторые механизмы генерации страт.

УДК 316.342.5
ББК 60.5

Миркин Б.Г. – Лаборатория анализа и выбора решений; Отделение прикладной математики и информатики НИУ ВШЭ, Москва, Россия; Department of Computer Science, Birkbeck University of London, UK.

Орлов М.А. – Отделение прикладной математики и информатики НИУ ВШЭ, Москва, Россия.

**Препринты Национального исследовательского университета
«Высшая школа экономики» размещаются по адресу: http://www.hse.ru/org/hse/wp**

Содержание

Введение....4

1. Проблема....8

2. Обзор методов многокритериального ранжирования....8
2.1. Ранжирование по влиянию....9
2.2. Разбиение по Парето....10
2.3. Правило Борда....10
2.4. Оптимизация линейных весов....10
2.5. Модифицированный метод *к*-средних....11

3. Предлагаемые методы стратификации....12
3.1. Метод линстрат....13
3.2. Метод паретострат....15

4. Экспериментальный анализ алгоритмов....16
4.1. Модели генерации данных....16
4.2. Множество сравниваемых методов....18
4.3. Критерии оценивания результатов....19

5. Некоторые результаты экспериментов....20
5.1. Верификация метода линстрат....20
5.2. Выбор функции расстояния между классами для метода паретострат....21
5.3. Сравнение результатов работы методов стратификации на синтетических данных....22
5.4. Сравнение результатов работы методов стратификации на реальных данных: оценка эффективности работы отделений банка....25

Заключение....28

Литература....29

Введение

При принятии решений и выборе одной из нескольких альтернатив часто приходится упорядочивать варианты по многим критериям [Алескеров, Хабина, Шварц 2006; Миркин 1974]. Несложно представить ситуации, в которых приходится осуществлять одновременно как разбиение, так и ранжирование, т.е. выделять упорядоченные однородные группы вариантов. Это соответствует выявлению в данных некоторой «вертикальной» иерархической структуры – ранжирования, и «горизонтальной» – слоев более или менее одинаковых объектов. Примером может служить разбиение фирм на классы по уровню риска банкротства или стран по уровню кредитного риска [De Smet, Montano, Guzman 2004]. Вертикальная составляющая представляет собой упорядочение групп: с высоким, средним и низким уровнем риска банкротства. А горизонтальная составляющая – это фирмы, входящие в одну и ту же группу. С одной стороны, подобное представление позволяет более компактно представить данные, с другой – может служить источником информации для выбора вариантов.

Мы будем называть проблему упорядоченного разбиения многокритериальных данных многокритериальной стратификацией. Слово «стратификация» происходит от латинского *stratum* – слой и греческого *facio* – делаю. Это понятие широко употребляется, например, в социологии. Социальная стратификация – это деление общества на специальные слои (страты) путем объединения различных социальных позиций с примерно одинаковым социальным статусом, отражающим сложившееся в нем представление о социальном неравенстве [Wikipedia]. В данной работе стратификация понимается в более широком смысле как упорядоченное разбиение множества объектов таким образом, чтобы упорядочение классов соответствовало рассматриваемым критериям.

Рассмотрим пример. В табл. 1 приведены цены на жилье и питание в десяти городах для иностранца [Burgess et al. 2006], эти же данные представлены графически на рис. 1 (сверху). Цены нормированы на максимальные значения по столбцу. Разделение на группы может быть неоднозначным; можно разделить варианты по стоимости жилья или по расходам на питание. Проблема в том, чтобы учитывать оба критерия сразу.

Таблица 1. Цены (в нормированных единицах) на жилье и питание в десяти различных городах

Город	Проживание	Питание
Москва (Moscow)	0,9749	0,7440
Лондон (London)	0,9479	0,7812
Токио (Tokyo)	1,0000	0,6764
Копенгаген (Copenhagen)	0,5602	1,0000
Нью-Йорк (New-York)	0,9749	0,6446
Пекин (Peking)	0,6924	0,4881
Сидней (Sydney)	0,4967	0,5318
Ванкувер (Vancouver)	0,3318	0,4775
Йоханнесбург (Johannesburg)	0,2322	0,4483
Буэнос-Айрес (Buenos-Aires)	0,3412	0,4178

Одним из классических алгоритмов кластеризации является *к*-средних [MacQueen 1967]. Этот метод хорошо изучен и имеет широкое применение на практике в задачах анализа данных и принятия решений. На рис. 1 (слева внизу) представлен результат разделения данных о городах на три кластера. Мы видим, что метод *к*-средних объединяет близкие точки в классы. Класс, помеченный «*», составляют {Joh, Van, Buen, Syd, Pek}, имеющие низкие цены на питание и жилье. Класс «+» составляет {Cop}, обладающий самыми высокими ценами на питание и умеренными на жилье, в класс «о» входят {Lon, Msk, Tok, NY} – это города с высокими ценами на жилье и умеренными ценами на питание. Такое разбиение не дает информации, какая из групп лучше, а какая хуже, и не позволяет упорядочить классы.

На рис. 1 (справа внизу) изображено разбиение методом стратификации «линстрат», предлагаемым в настоящей работе. Этот метод формирует агрегированный критерий как взвешенную сумму исходных критериев таким образом, чтобы веса индивидуальных критериев подбирались автоматически без какого-либо «учителя». Интуитивно суть метода состоит в следующем. Предположим, что может быть найден общий критерий, учитывающий цены на питание и жилье с некоторыми весовыми коэффициентами. Этот критерий можно интерпретировать как расходы на проживание. Веса подбираются из условия, чтобы варианты из одной страты имели как можно более близкие значения по общему критерию. На рис. 1 (справа внизу) видно, что разбиение на страты отлично от разбиения по методу *к*-средних. Первая группа «+» {Joh, Van, Buen} – горо-

да с самыми низкими ценами, вторая группа «*» включает в себя {Syd, Pek}, города с умеренными ценами, а третья «o» – города, в которых цены высокие либо на жилье, либо на питание {Cope, Lon, Msk, NY, Tok}. Теперь можно упорядочить классы с точки зрения общих расходов на проживание и питание.

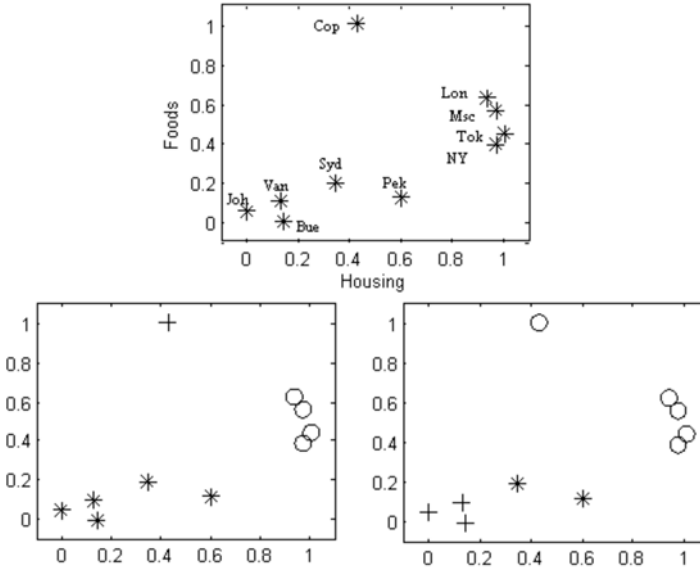

Рис. 1. Цены на жилье и питание в десяти городах (вверху). Разбиение на три группы методом *к*-средних (слева внизу) и методом linstrat (справа внизу)

Методы многокритериальной стратификации можно условно разделить с точки зрения того, каким образом в данных выделяется вертикальная иерархическая структура, т.е. упорядочивание:

1. По одному агрегированному критерию, являющемуся линейной сверткой критериев. В основе метода лежит постулат о том, что уменьшение одного критерия может быть «погашено» увеличением другого критерия («замещение»). При этом коэффициенты замещения критериев являются постоянными (не зависят от объектов).

2. Через индивидуальные критерии путем использования соответствующего отношения многомерного упорядочения. При этом критерии объявляются несравнимыми, так что нельзя заместить один другим.

3. В эту группу мы включаем любые другие методы, где критерии могут замещаться друг другом, но коэффициенты зависят от сравниваемых объектов, т.е. общий критерий является нелинейной комбинацией частных критериев.

Первый класс методов составляют всевозможные процедуры назначения весовых коэффициентов как с участием экспертов, так и автоматически. Например, веса критериев могут быть получены путем вычисления ранга объектов. Применительно к ранжированию конференций и авторов публикаций этот подход был разработан в статье [Sun et al. 2009].

Ко второму типу можно отнести ряд методов теории коллективного выбора [Алескеров, Хабина, Шварц 2006], которые позволяют упорядочивать варианты, используя ранжирования по отдельным критериям.

К третьему типу относятся, например, методы, разработанные в статьях [Ng 2007; Ramanathan 2006] применительно к задаче разделения ресурсов фирмы на группы по степени важности (ABC-анализ), весовые коэффициенты здесь находятся из решения задачи линейной оптимизации. Хотя ранг объекта в данном случае вычисляется путем линейной свертки с весами, но для каждого объекта набор весов индивидуален, т.е. зависит от местоположения сравниваемых вариантов. В работе [Белов, Коричнева 2012] для решения задачи ABC-анализа были предложены методы разбиения на основе критерия сходства результирующей многомерной классификации с результатами классификаций по каждому из критериев. Также к третьему типу принадлежат методы ранжирования с использованием информации о предпочтениях лица, принимающего решения (ЛПР). В работе [De Smet, Montano, Guzman 2004] предложен расширенный вариант алгоритма κ-средних, который использует метрику, учитывающую предпочтения ЛПР. В [Nemery, De Smet 2005] ранжирование вариантов и их объединение в кластеры производятся на основе матрицы парных сравнений вариантов.

В данной работе предлагаются новые методы многокритериальной стратификации как первого, так и второго типа. Метод первого типа использует линейную свертку весов. Веса находятся таким оброзом, чтобы страты наилучшим образом аппроксимировали данные по взвешенному критерию. Страты моделируются параллельными гиперплоскостями, ортогональными вектору весовых коэффициентов. Метод второго типа включает два этапа. Первый этап – это формирование разбиения Парето, при котором выделяется упорядоченное множество слоев объектов, не-

доминируемых по отношению Парето. На втором этапе полученные слои объединяются в нужное число страт путем склеивания близлежащих слоев на манер агломеративной кластеризации.

Для того чтобы организовать осмысленное сравнение предложенных методов с существующими методами, мы рассматриваем несколько вариантов механизма генерации стратифицированных данных, в некотором смысле моделирующих природу многокритериальных задач. Проводятся численные эксперименты по верификации предложенных методов и их сравнению с существующими методами. Основной вывод – несомненного чемпиона среди рассмотренных методов нет. Победитель в значительной степени определяется геометрической структурой многомерного отношения «больше» на заданном множестве объектов. Это предопределяет необходимость дальнейшей модификации предлагаемых алгоритмов.

1. Проблема

Проблема многокритериальной стратификации ставится следующим образом: по критериальной матрице $X = ||x_{ij}||$, где $i = 1,\ldots, n$ – варианты или объекты, $j = 1,\ldots, m$ – критерии, а x_{ij} – значение j-го критерия на i-м объекте, необходимо построить разбиение вариантов на заданное число K непересекающихся классов – страт, так, чтобы варианты, входящие в одну и ту же страту, были более или менее одинаковы, и в то же время в основном хуже, чем варианты из предыдущей страты. Иными словами, надо построить разбиение $S = \{S_1, ..., S_K\}$ множества объектов на K классов S_k ($k = 1,\ldots, K$), соответствующих стратам, и вектор рангов $c=(c_1,.., c_K)$ так, чтобы $c_k > c_j$, как только $k < j$.

2. Обзор методов многокритериального ранжирования

В принципе, результаты любого метода многокритериального ранжирования, позволяющего находить численное значение рангов объектов, т.е. переходить от многих критериев к одному интегральному, можно использовать для стратификации. При заданном интегральном критерии

разбиение на K страт легко найти применением одномерной процедуры κ-средних к данному критерию. Это позволяет включить в наш анализ и методы упорядоченного ранжирования, и упорядоченного разбиения (стратификации).

Сначала мы рассматриваем метод первого типа, использующий линейную свертку весов – ранжирование по влиянию (authority ranking), затем – два метода второго типа, т.е. ранжирование по несравниваемым критериям: разбиение Парето (partition via maximal element) и правило Борда (Borda count). Далее описываются методы линейной оптимизации весов (linear weight optimization) и модифицированный метод κ-средних (extended k-means), которые можно отнести к третьему типу в рассматриваемой классификации методов.

2.1. Ранжирование по влиянию

Этот метод был предложен в [Sun et al. 2009] для построения ранжирования авторов научных публикаций, участвующих в различных конференциях, как дальнейшее развитие методов, основанных на вычислении собственного вектора, соответствующего максимальному собственному числу определенной матрицы (см., например, [Берж 1962, Миркин 1974, Saaty 1980, Page and Brin 1999]). Правило, по которому строится ранг, опирается на два положения:

1) объекты, имеющие высокий ранг, имеют высокую оценку по критериям с большими весами;

2) вес критерия тем выше, чем больше его значения для объектов с высокими рангами.

Предположим, что объекты имеют ранг r_i, а критерии имеют веса w_i. Сформулированные положения могут быть представлены в виде системы линейных уравнений:

$$\begin{cases} r_i = x_{i1}w_1 + x_{i2}w_2 + ... + x_{im}w_m, i = 1...n; \\ w_j = x_{1j}r_1 + x_{2j}r_2 + ... + x_{nj}r_n, j = 1...m. \end{cases} \quad (1)$$

В первых n уравнениях системы (1) исходные переменные заменим нормированными $a_{ij} = \dfrac{x_{ij}}{\sum_{j=1}^{m} x_{ij}}$. Полученную из данных коэффициентов матрицу обозначим A, а в оставшихся m уравнениях произведем замену

$b_{ji} = \dfrac{x_{ji}}{\sum_{i=1}^{n} x_{ji}}$, которые будут составлять матрицу B. В матричном виде уравнения записываются как

$$\begin{cases} r = Aw; \\ w = Br. \end{cases} \quad (2)$$

Подставляя w из второго уравнения в первое, получаем, что искомый ранг r является собственным вектором матрицы AB, соответствующим ее максимальному собственному значению, равному единице [Sun et al. 2009].

2.2 Разбиение по Парето

Рассмотрим векторное отношение R «больше» такое, что для двух вариантов $x, y \in X$ имеет место xRy если и только если $x_j \geq y_j$ для всех j, причем хотя бы одно неравенство строгое. На каждом шаге находим множество недоминируемых объектов $x = \{b \in X \mid \neg\exists y \in X : yRb\}$, которые объединяются в класс C_l. Полученный класс исключается из рассмотрения, и процедура повторяется для оставшихся объектов $X \backslash C_l$, пока множество X не пусто [Aleskerov, Ersel, Yolalan 2004].

2.3. Правило Борда

Для каждого объекта x_i вычисляются оценки по всем критериям $r_i(x_j) = \left|\{x_p \in X : x_{ij} > x_{pj}, p \neq i\}\right|$. Затем итоговый ранг объекта x_i подсчитывается как сумма оценок по всем критериям $r(x_i) = \sum_{j=1}^{M} r_j(x_i)$.

2.4. Оптимизация линейных весов

Подход на основе линейной оптимизации весов был предложен в работе [Ramanathan 2006] применительно к задаче многокритериального ABC-анализа. Для того чтобы вычислить ранги, поочередно решается задача линейного программирования (3) относительно весов w_{ij} для каждого объекта x_i:

$$\begin{cases} \sum_{j=1}^{M} x_{ij} w_{ij} \to \max; \\ \sum_{j=1}^{M} x_{pj} w_{ij} \le 1,\ p = 1 \ldots n; \\ w_{ij} \ge 0. \end{cases} \qquad (3)$$

Решая оптимизационную задачу (3), получаем весовые коэффициенты w_{ij} для каждого критерия, по которым вычисляются искомые ранги $r_i = \sum_{j=1}^{M} x_{ij} w_{ij}$.

2.5. Модифицированный метод к-средних

Метод, разработанный в статье [De Smet, Montano, Guzman 2004] является адаптацией известного алгоритма кластеризации *к*-средних для стратификации. Модифицированный метод *к*-средних использует функцию расстояния между объектами, основанную на структуре предпочтений: $< P, I, J >$, где отношения P – строгого предпочтения (асимметричное), I – безразличия (рефлексивное и симметричное) и J – несравнимости (иррефлексивное и симметричное). Структура предпочтений считается заданной на множестве X, если для любых двух элементов x_i, x_j из X имеет место только одно из отношений: $x_i P x_j$, $x_j P x_i$, $x_i I x_j$, $x_i J x_j$. Чтобы определить структуру предпочтений $< P, I, J >$, часто используют информацию от ЛПР о сравнении локальных предпочтений, включая методы AHP [Saaty 1980], см. также [Подиновская, Подиновский 2011], и Electre [DeSmet, Gilbart 2001]. Однако можно использовать и векторное отношение «больше».

После того, как предпочтения заданы, для каждого объекта x_i строится профиль $P(x_i) =< P_1(x_i),\ P_2(x_i),\ P_3(x_i),\ P_4(x_i) >$,

где: $P_1(x_i) = \{x_j \in X \mid x_i J x_j\}$ – множество вариантов, несравнимых с x_i;

$P_2(x_i) = \{x_j \in X \mid x_i P x_j\}$ – множество вариантов, строго доминируемых x_i;

$P_3(x_i) = \{x_j \in X \mid x_i I x_j\}$ – множество вариантов, безразличных x_i;

$P_4(x_i) = \{x_j \in X \mid x_j P x_i\}$ – множество вариантов, строго доминирующих x_i.

Далее, используя профиль каждого объекта, определяем расстояние между x_i и x_j:

$$d(x_i, x_j) = 1 - \frac{\sum_{k=1}^{4} \left| P_k(x_i) \cap P_k(x_j) \right|}{n}. \quad (4)$$

Если задано множество объектов $C = \left\{ x_{i_1}, x_{i_2}, ..., x_{i_p} \right\}$, то центром для C называется

$$c = \arg\min\left(\sum_{j=1}^{p} d(x_{ij}, p)\right). \quad (5)$$

То есть центр – это тот элемент, до которого суммарное расстояние от всех других объектов в классе наименьшее. Для нахождения кластеров используется известный алгоритм *к*-средних, с той особенностью, что в качестве расстояния между объектами используется (4), а в качестве центра кластера – объект, удовлетворяющий (5).

3. Предлагаемые методы стратификации

В настоящей работе предлагается два метода многокритериальной стратификации: линстрат и паретострат, которые в какой-то мере аппроксимируют слои несравнимых по Парето объектов. В методе паретострат это достигается «склеиванием» близлежащих «тонких» слоев отношения Парето. В методе линстрат страты ищутся в форме линейных образований, определяемых автоматически подбираемыми коэффициентами. Параметры страт определяются из решения оптимизационной задачи. Критерием оптимизации является «толщина» получаемого образования (чем меньше, тем лучше). В рамках предлагаемой классификации методов ранжирования и стратификации линстрат относится к первому типу, поскольку использует линейную свертку, а паретострат – ко второму, так как ранжирование производится на основе несравнимых независимо упорядоченных критериев.

Для простоты изложения ограничимся случаем трех страт, хотя обобщение методов на большее число страт не представляет сложности. Выбор данного числа страт объясняется тем, что на практике часто удобно разделять объекты на три группы: «плохие», «средние» и «хорошие» (АВС-классификация).

3.1. Метод линстрат

Линстрат позволяет находить страты в виде слоев, расположенных на гиперплоскостях, задаваемых вектором весов *w,* таким что $\sum_{j=1}^{M} w_j = 1$ и $w_j \geq 0$. В идеальном случае проекции точек одного класса на направление *w* должны совпадать. Значения проекций – это ранги объектов и, таким образом, объекты из одного класса разделяют общее значение ранга, который назовем уровнем страты. Будем искать такое направление *w* и разбиение на классы, чтобы ранги объектов группировались как можно ближе друг к другу внутри классов, как показано на рис. 2.

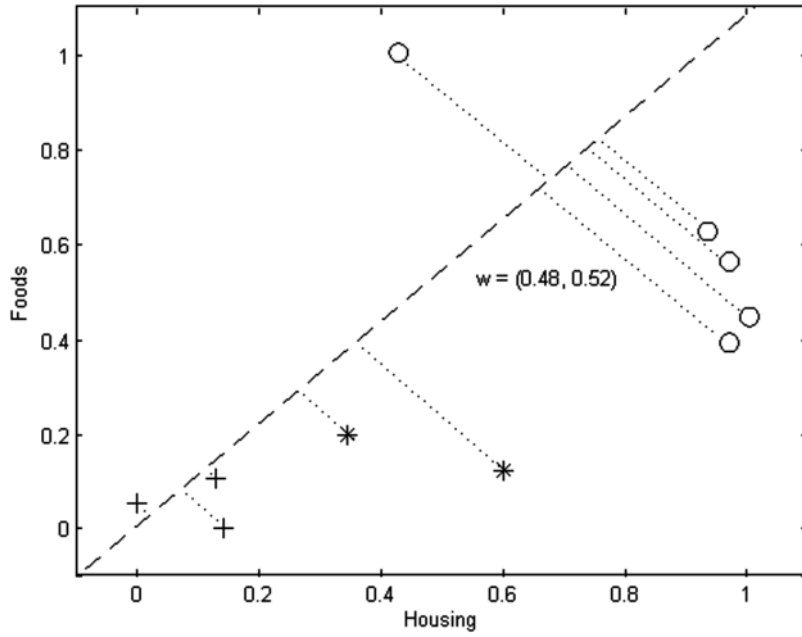

Рис. 2. Проекция объектов из табл. 1 на направление, заданное вектором весовых коэффициентов *w* = (0,48, 0,52)

Это приводит к следующей модели. Даны *N* объектов, оцененных по *M* критериям, результаты записаны в виде критериальной матрицы $X=||x_{ij}||$ – оценка *i*-го объекта по *j*-му критерию. Для того чтобы ранжи-

ровать варианты, используем взвешенный критерий. Считаем, что каждый объект получен из одного из K классов, для которых характерно некоторое общее значение ранга (уровня страты) $c_k \in \{c_1, c_2, ..., c_K\}$, причем $c_1 > c_2 > ... > c_K$. Модель страты может быть записана в виде $w_1 x_{i1} + w_1 x_{i2} + ... + w_M x_{iM} = c_k + e_i$ – минимизируемая ошибка. Чтобы найти наилучшее разделение на классы, необходимо подобрать значения w, c и S, для которых сумма квадратов невязок будет минимальна:

$$\begin{cases} \sum_{k=1}^{K} \sum_{i \in S_k} (x_i w - c_k)^2 \xrightarrow[w,c,S]{} \min; \\ \sum_{j=1}^{M} w_j = 1, w_j \geq 0. \end{cases} \quad (6)$$

По сути, наш алгоритм решения задачи (6), называемый далее линстрат, представляет собой случайный поиск оптимального решения по весам w (например, можно использовать инспирированные природой методы случайного поиска на основе эволюции популяции решений [Fogel 1995] или оптимизации роя частиц [Kennedy, Eberhart 2001]). На каждой итерации поиска для фиксированных значений w применяется одномерная процедура k-средних для нахождения оптимального разбиения и уровней страт. Ниже алгоритм линстрат описан более подробно.

Алгоритм линстрат:

Вход: объекты X = (x_1, x_2, ..., x_N).

Выход: страты S, уровни страт c и веса w.

1. Сгенерировать популяцию векторов весовых коэффциентов $w^{(l)}$, такую, что веса удовлетворяют ограничениям $\sum_{j=1}^{M} w_j^{(l)} = 1, w_j^{(l)} \geq 0$, l=1...L.

2. Для каждого члена популяции найти оптимальное разбиение S и уровни c применением одномерной процедуры k-средних к рангам объектов $r_i^{(l)} = \sum_{j=1}^{M} x_{ij} w_j^{(l)}$, l=1...L.

3. Найти наилучший элемент популяции, проверить, лучше ли он, чем рекорд, запомненный с предыдущих итераций; если лучше, то заменить им рекорд, после чего заменить наихудший элемент популяции на рекорд.

4. Видоизменить популяцию весов:

$w^{(l)} = w^{(l)} + \varepsilon^{(l)}$, где l=1...L, $\varepsilon^{(l)} \sim N(0, \alpha)$, α – параметр алгоритма.

5. Привести веса в соответствие ограничениям неотрицательности и единичной суммы следующим образом. Веса, значения которых получи-

лись отрицательными, заменить нулевыми и нормировать каждый полученный набор из популяции весов на его сумму.

6. Перейти к пункту 1 и повторить заданное число итераций.

Описанный алгоритм линстрат требует задания трех параметров: численность популяции L, число итераций и параметр α. В наших экспериментах мы установили значения парметров 300, 100 и 0,01 соответственно. Напомним, что число страт везде полагается равным трем.

3.2. Метод паретострат

Чтобы получить страты из слоев несравнимых по Парето объектов (полученных как описано в п. 2.2), можно воспользоваться идеей агломеративной кластеризации. Будем объединять группы объектов, находящиеся близко друг к другу с точки зрения некоторой метрики $d(C_i, C_j)$, задающей расстояние между группами точек C_i и C_j. Чтобы расстояние между стратами было наибольшим, а расстояние внутри страт наименьшим, необходимо найти две пары соседних классов, имеющих максимальное расстояние $d(C_i, C_{i+1})$ и $d(C_j, C_{j+1})$ и , затем построить страты $S_1 = \{C_1, C_2, ..., C_i\}$, $S_2 = \{C_{i+1}, C_{i+2}, ..., C_j$, $S_3 = \{C_{j+1}, C_{j+2}, ..., C_S\}$.

Алгоритм паретострат:

Вход: объекты $X = \{x_1, x_2, ..., x_N\}$.

Выход: страты $S = \{S_1, S_2, S_3\}$.

1. Найти разбиение Парето $C_1, C_2, ..., C_S$

2. Найти расстояния между соседними классами

$R_k = d(C_k, C_{k+1}), p = 1 \ldots s-1$

3. Вычислить индексы $i = \arg\max(R_k), j = \arg\max(R_p),\ p > i$

4. Построить страты $S_1 = \{C_1, C_2, ..., C_i\}$, $S_2 = \{C_{i+1}, C_{i+2}, ..., C_j\}$,

$S_3 = \{C_{j+1}, C_{j+2}, ..., C_S\}$.

В качестве функции расстояния между C_i и C_j можно взять одну из известных метрик:

– single-link $d(C_i, C_j) = \min(d(x,y))$, где $x \in C_i$, $y \in C_i$

– complete-link $d(C_i, C_j) = \max(d(x,y))$, где $x \in C_i$, $y \in C_i$

– average-link $d(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_x \sum_y d(x,y)$, где $x \in C_i$, $y \in C_i$

Какое из вышеуказанных расстояний лучше взять – открытый вопрос. В разделе 5.3 описаны результаты эксперимента по сравнению указанных метрик и выбору наилучшей с точки зрения стратификации.

4. Экспериментальный анализ алгоритмов

Чтобы проверить корректность работы предложенных методов стратификации и сравнить их с существующими методами, были проведены эксперименты на синтетических и реальных данных.

Первый эксперимент был нацелен на проверку корректности линстрат. Нас интересовала точность предсказания параметров модели, по которым были сгенерированы данные, а именно проверялась точность восстановления значений уровней страт $c = (c_1, c_2, c_3)$ и весов w.

Второй эксперимент по верификации методов проводился для выбора наилучшей функции расстояния между группами объектов в методе паретострат.

Далее все описанные в работе методы сравнивались с точки зрения правильной стратификации объектов для различных моделей генерации стратифицированных данных в зависимости от добавленного шума.

В заключительном эксперименте все рассматриваемые в работе методы сравниваются на реальных данных.

4.1. Модели генерации данных

В работе рассмотриваются три различных модели генерации синтетических данных, которые различаются формой страт.

4.1.1. Линейные страты

Первый тип страт полностью соответствует модели линейных страт. Точки принадлежат одной из трех гиперплоскостей, заданных вектором нормали w, которые описываются уравнениями $w_1 x_{i1} + w_1 x_{i2} + ... + w_M x_{iM} = c_k + e_i$, где $e_i = N(0,\sigma)$. На рис. 3 показано, как выглядят сгенерированные данные в двумерном случае при различных значениях добавленного шума.

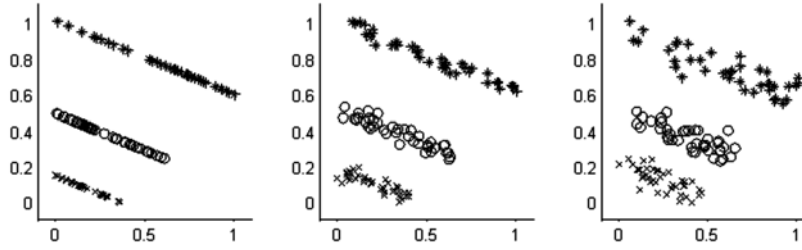

Рис. 3. Линейные страты при трех различных значениях дисперсии шума $\sigma = \{0{,}00,\ 0{,}02,\ 0{,}05\}$

4.1.2. Сферические страты

Вторая модель формирования страт генерирует точки, лежащие на поверхности сфер: $w_1 x_{i1}^2 + w_1 x_{i2}^2 + ... + w_M x_{iM}^2 = c_k^2 + e_i$, где $e_i = N(0,\sigma)$ $c = (c_1, c_2, c_3)$, $x_i \in [0,1]$. Рисунок 4 наглядно демонстрирует, как выглядят сферические страты в двумерном случае.

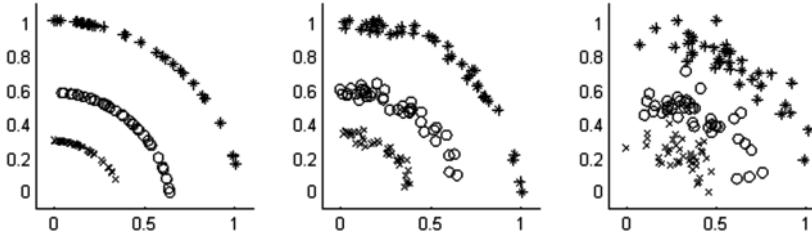

Рис. 4. Сферические страты при трех различных значениях дисперсии шума $\sigma = \{0{,}00,\ 0{,}02,\ 0{,}05\}$

4.1.3. Точечные страты

Третий тип данных – это страты, образующие облака точек вокруг центров с координатами $\{wc_1, wc_2, wc_3\}$, лежащих на одной прямой, которая проходит через 0 (рис. 5).

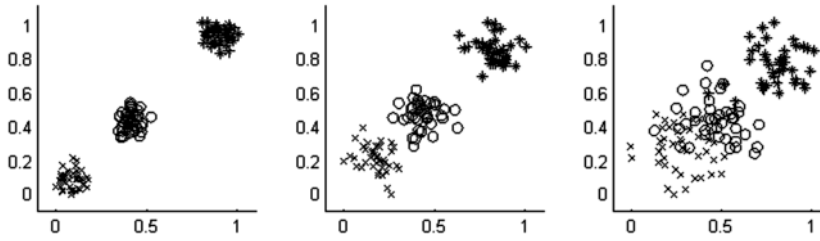

Рис. 5. Точечные страты при трех различных значениях дисперсии шума $\sigma = \{0{,}03,\ 0{,}06,\ 0{,}12\}$

4.1.4. Реальные данные. Оценка эффективности работы отделений банка

Этот набор данных взят из работы [Aleskerov, Ersel, Yolalan 2004]. Он содержит информацию о деятельности 23 отделений банка, оцененной по четырем критериям (табл. 13):

– сумма вкладов (Total deposits of the bank, D);

– сумма кредитов (Total credits of the bank, C);

– объем операций с ценными бумагами (Total securities trading of the bank, S);

– Объем торговых операций с валютой (Total FX trading volume of the bank, FX).

4.2. Множество сравниваемых методов

В табл. 2 представлен список методов многокритериальной стратификации и ранжирования, на которых проводились экспериментальные исследования. Если тот или иной метод использовался для ранжирования, то в обозначениях к аббревиатуре добавляется «+*R*». Например, *LS*+*R* означает, что метод стратификации линстрат используется для получения рангов объектов, с помощью найденных весовых коэффициентов. Если с помощью метода было получено разбиение на страты, то к аббревиатуре добавляется «+*S*». Например, *BC*+*S* означает, что страты были получены из рангов, найденных по правилу Борда.

Таблица 2. Методы многокритериальной стратификации и ранжирования, используемые в данной работе

Название метода	Аббревиатура	Описан в пункте
Линстрат	LS	3.1
Ранжирование по влиянию	AR	2.1
Оптимизация линейных весов	LW	2.4
Разбиение Парето (паретострат)	PS	2.2, 3.2
Правило Борда	BC	2.3
Расширенный *к*-средних	EK	2.5

4.3. Критерии оценивания результатов

Для сравнения точности стратификации использовался критерий, показывающий долю правильно классифицированных объектов (т.е. с правильно назначенным номером страты) $N_{correct}$ среди всей выборки N:

$$accuracy = \frac{N_{correct}}{N}. \quad (7)$$

При измерении различий между двумя строгими или нестрогими ранжированиями объектов – R, S использовалось нормированное расстояние Хэмминга. Оно вычисляется по формуле (8):

$$d_H = \frac{1}{N(N-1)} \sum_{i,j=1}^{N} |r_{ij} - s_{ij}|. \quad (8)$$

где N – число элементов, r_{ij} и s_{ij} – элементы 1/0 матриц бинарных отношений, соответствующих ранжированиям R и S. Обычно для измерения расстояния между стратификациями используется нормированное расстояние между упорядоченными разбиениями Кемени – Снелла [Kemeny, Snell, 1962]. Для стратификации S, заданной на объектах $x_1,...,x_N$ строится матрица:

$$\overline{s}_{ij} = \begin{cases} 1, S(x_i) < S(x_j); \\ 0, S(x_i) = S(x_j); \\ -1, S(x_i) > S(x_j). \end{cases} \quad (9)$$

Здесь $S(x)$ обозначает номер страты, которой принадлежит объект x. Расстояние между стратификациями вычисляется по формуле

$$d_{KS} = \frac{1}{N(N-1)} \sum_{i,j=1}^{N} | \overline{r}_{ij} - \overline{s}_{ij} |. \qquad (10)$$

Однако меры (8) и (10) пропорциональны друг другу, $d_H = d_{KS}/2$ [Миркин, 1974].

5. Некоторые результаты экспериментов

В этой части приведены некоторые результаты экспериментального исследования методов стратификации. На синтетических данных верифицируется правильность работы метода линстрат. Рассматривается вопрос выбора наилучшей функции расстояния между слоями несравнимых объектов для метода паретострат. Проводится сравнение качества работы предлагаемых методов с известными методами, как на искусственных, так и на реальных данных.

5.1. Верификация метода линстрат

При формулировке метода линстрат предполагается, что стратифицированные данные соответствуют модели линейных страт (4.1.1). Чтобы убедиться в том, что метод линстрат правильно восстанавливает параметры модели линейных страт, мы случайным образом сгенерировали пять векторов весовых коэффициентов и, соответственно, пять значений уровней страт.

Таблица 4. Верификация правильности работы алгоритма линстрат. Значения уровней страт $c = (c_1, c_2, c_3)$, заданных при генерации данных и значения, полученные методом линстрат

Истинные значения уровней	**Полученные уровни страт**
0,7623 0,4109 0,0724	0,7634 0,4105 0,0723
0,7565 0,5872 0,4177	0,7648 0,5938 0,4220
0,9204 0,8533 0,5308	0,9221 0,8551 0,5311
0,5210 0,4296 0,3269	0,5165 0,4264 0,3242
0,5366 0,1470 0,1280	0,5393 0,1486 0,1289

Таблица 5. Верификация работы алгоритма линстрат. Значения весовых коэффициентов критериев, заданных при генерации данных, и значения, полученные методом линстрат

Истинные веса	**Полученные веса**
0,23 0,04 0,03 0,23 0,48	0,23 0,04 0,03 0,23 0,47
0,11 0,05 0,38 0,32 0,13	0,11 0,06 0,38 0,33 0,13
0,02 0,09 0,21 0,37 0,30	0,02 0,10 0,21 0,37 0,30
0,01 0,46 0,32 0,06 0,14	0,02 0,47 0,30 0,07 0,14
0,28 0,43 0,24 0,01 0,04	0,29 0,42 0,23 0,02 0,04

Из полученных результатов видно, что метод достаточно хорошо восстанавливает значения весов критериев и уровней страт.

5.2. Выбор функции расстояния между классами для метода паретострат

Целью данного эксперимента является определение наилучшей функции расстояния между слоями несравнимых по Парето элементов. Сравниваются три функции расстояния между слоями недоминируемых по Парето объектов: single link, complete link и average link. Качество разделения на страты оценивается с точки зрения правильной классификации (7). В каждом эксперименте генерируются линейные страты в соответствии с моделью, описанной в пункте 4.1.1, со случайными весами и фиксированными уровнями страт c = (0,3, 0,5, 0,8). Общее количество сгенерированых объектов в каждом эксперименте равно 90, по 30 на каждую страту. Эксперимент повторялся 50 раз. В табл. 6 приведены значения среднего и дисперсии точности стратификации, в зависимости от интенсивности добавленного шума σ.

Из табл. 6 видно, что расстояния complete link и average link дают практически одинаковые результаты, лучшие по сравнению с результатом single link. Все дальнейшие результаты стратификации методом паретострат, приведенные в работе, получены с использованием метрики average link.

Таблица 6. Значения среднего и дисперсии точности стратификации методом паретострат для трех различных функций расстояния между слоями несравнимых по Парето вариантов

Метрика	**σ = 0,05**		**σ = 0,1**		**σ = 0,2**		**σ = 0,3**	
	mean	**std**	**mean**	**std**	**mean**	**std**	**mean**	**std**
Single link	0,86	0,12	0,74	0,13	0,60	0,08	0,51	0,09
Complete link	0,93	0,04	0,82	0,07	0,70	0,06	0,62	0,06
Average link	0,93	0,05	0,82	0,07	0,71	0,06	0,61	0,07

5.3. Сравнение результатов работы методов стратификации на синтетических данных

Для всех типов данных генерировалась выборка из 90 объектов (по 30 объектов на каждую страту). Значения уровней страт во всех случаях фиксированы и равны $c = (0{,}8, 0{,}5, 0{,}3)$. Точность стратификации вычислялась в зависимости от весовых коэффициентов (ориентации страт) и уровня добавленного шума (толщины страт). Каждый эксперимент повторялся 50 раз.

В табл. 7, 8 и 9 приведены значения среднего и дисперсии точности стратификации для модели линейных страт (п. 4.1.1), в зависимости от уровня добавленного шума, для трех различных наборов весовых коэффициентов.

Таблица 7. Среднее значение и дисперсия точности стратификации (50 повторений эксперимента), в случае линейных страт $w = (0{,}2, 0{,}2, 0{,}2, 0{,}2, 0{,}2)$ для четырех значений уровня добавленного шума σ

Метод стратификации	**σ = 0,05**		**σ = 0,1**		**σ = 0,2**		**σ = 0,3**	
	mean	**std**	**mean**	**std**	**mean**	**std**	**mean**	**std**
LS+S	0,99	0,05	0,93	0,09	0,69	0,10	0,62	0,08
PS+S	0,97	0,03	0,91	0,04	0,75	0,06	0,64	0,05
BC+S	0,83	0,04	0,80	0,04	0,71	0,05	0,64	0,05
AR+S	1,00	0,00	0,97	0,03	0,75	0,06	0,63	0,05
LW+S	0,99	0,01	0,94	0,03	0,79	0,06	0,69	0,05
EK+S	0,64	0,09	0,61	0,09	0,52	0,12	0,45	0,09

Таблица 8. Среднее значение и дисперсия точности стратификации (50 повторений эксперимента) в случае линейных страт $w = (0{,}5, 0{,}2, 0{,}1, 0{,}1, 0{,}1)$ для четырех значений уровня добавленного шума σ

Метод стратификации	**σ = 0,05**		**σ = 0,1**		**σ = 0,2**		**σ = 0,3**	
	mean	**std**	**mean**	**std**	**mean**	**std**	**mean**	**Std**
LS+S	0,98	0,07	0,86	0,14	0,65	0,10	0,55	0,08
PS+S	0,95	0,45	0,85	0,06	0,73	0,06	0,63	0,06
BC+S	0,83	0,04	0,79	0,04	0,69	0,04	0,63	0,05
AR+S	0,77	0,04	0,76	0,05	0,70	0,04	0,65	0,05
LW+S	0,98	0,01	0,92	0,03	0,76	0,04	0,66	0,06
EK+S	0,63	0,11	0,60	0,10	0,52	0,07	0,46	0,09

Таблица 9. Среднее значение и дисперсия точности стратификации (50 повторений эксперимента), в случае линейных страт $w = (0{,}8, 0{,}05, 0{,}05, 0{,}05, 0{,}05)$ для четырех значений уровня добавленного шума σ

Метод стратификации	**σ = 0,05**		**σ = 0,1**		**σ = 0,2**		**σ = 0,3**	
	mean	**std**	**mean**	**std**	**mean**	**std**	**mean**	**std**
LS+S	0,77	0,15	0,72	0,11	0,58	0,07	0,52	0,07
PS+S	0,90	0,06	0,79	0,06	0,68	0,06	0,59	0,05
BC+S	0,77	0,04	0,75	0,06	0,66	0,05	0,59	0,05
AR+S	0,61	0,04	0,59	0,05	0,59	0,05	0,58	0,04
LW+S	0,98	0,02	0,80	0,03	0,71	0,05	0,64	0,06
EK+S	0,57	0,13	0,56	0,10	0,48	0,09	0,42	0,07

Из табл. 7–9 можно сделать заключение, что в большинстве случаев лучшие результаты дает метод линейной оптимизации весов LW+S. Метод линстрат LS+S дает приемлемые результаты только при небольших значениях шума и при равномерно распределенных между критериями весах.

В табл. 10, 11 приведены значения среднего и дисперсии точности стратификации для сферических и точечных страт, модели генерации которых описаны в п. 4.1.2 и 4.1.3.

Таблица 10. Среднее значение и дисперсия точности стратификации (50 повторений эксперимента) для сферических страт с параметрами w = (0,2, 0,2, 0,2, 0,2, 0,2), для четырех уровней добавленного шума σ

Метод стратификации	σ = 0,001		σ = 0,01		σ = 0,05		σ = 0,1	
	mean	std	mean	std	mean	std	mean	std
LS+S	0,63	0,05	0,62	0,05	0,62	0,05	0,60	0,06
PS+S	0,65	0,07	0,68	0,06	0,65	0,06	0,63	0,06
BC+S	0,73	0,05	0,72	0,05	0,70	0,05	0,66	0,04
AR+S	0,55	0,05	0,55	0,05	0,55	0,04	0,54	0,05
LW+S	0,34	0,02	0,35	0,04	0,34	0,02	0,35	0,06
EK+S	0,46	0,08	0,45	0,08	0,45	0,07	0,42	0,08

Таблица 11. Среднее значение и дисперсия точности стратификации (для 50 повторений эксперимента) для точечных страт с параметрами w = (0,2, 0,2, 0,2, 0,2, 0,2) для четырех уровней добавленного шума σ

Метод стратификации	σ = 0,05		σ = 0,1		σ = 0,2		σ = 0,3	
	mean	std	mean	std	mean	std	mean	std
LS+S	1,00	0,00	0,98	0,01	0,85	0,04	0,70	0,04
PS+S	1,00	0,00	0,86	0,12	0,64	0,07	0,63	0,05
BC+S	1,00	0,00	0,99	0,01	0,87	0,03	0,74	0,04
AR+S	1,00	0,00	0,97	0,03	0,76	0,05	0,62	0,06
LW+S	1,00	0,00	0,96	0,02	0,79	0,04	0,68	0,05
EK+S	1,00	0,00	0,98	0,01	0,75	0,12	0,58	0,15

Из табл. 10–11 видно, что для сферических страт все методы работают хуже, чем для страт в форме облаков точек. В этих двух случаях лучше всех работает метод на основе правила Борда BC+S.

Таким образом, среди сравниваемых методов абсолютного лидера, с точки зрения точности стратификации, нет. Тот или иной метод дает лучшие результаты в зависимости от типа данных или значения добавленного шума. В случае линейных страт в большинстве случаев лучшие результаты показал алгоритм на основе линейной оптимизации LW+S. На стратах, имеющих сферическую форму, все методы работают несколько хуже, чем на линейных стратах. Здесь наибольшую эффективность показал метод BC+S, в то время как линейные методы и EK+S оказались малоэффективными. Случай, когда страты имеют форму облаков точек,

оказался благоприятным для всех методов. В последнем случае наиболее устойчивым к шуму оказался BC+S.

Из данных, представленных в табл. 7, 8 и 9, видно, что не все из рассматриваемых методов являются устойчивыми к изменению ориентации страт. Точность стратификации алгоритмов LS+S и AR+S с изменением весовых коэффициентов ухудшилась, при этом LW+S показал стабильно высокие результаты. Методы на основе процедур многокритериального ранжирования PS+S и BC+S являются достаточно устойчивыми к изменению ориентации линейных страт (т.е. весовых коэффициентов), а метод стратификации EK+S в линейном случае оказался наименее эффективным.

5.4. Сравнение результатов работы методов стратификации на реальных данных: оценка эффективности работы отделений банка

В табл. 13 показаны результаты стратификации отделений банка по эффективности работы различными методами. Чтобы оценить, насколько похожие результаты дают те или иные методы, использовалось расстояние Хэмминга (8), оно же – половина расстояния Кемени – Снелла (9).

Таблица 13. Значения критериев оценки работы отделений банков и номеров страт, полученных различными методами стратификации

	Критерии				Методы стратификации					
Branch	**D**	**C**	**S**	**FX**	**LS+S**	**PS+S**	**BC+S**	**AR+S**	**LW+S**	**EK+S**
1	3,42	1,26	0,81	0,51	3	3	2	3	2	2
2	1,71	0,68	0,82	0,51	3	3	3	3	3	3
3	3,72	0,64	3,38	1,22	2	1	1	2	1	2
4	2,08	1,85	0,68	1,30	2	3	2	3	3	3
5	3,63	1,63	1,68	0,98	2	3	1	2	2	2
6	5,83	2,71	3,34	2,99	1	1	1	1	1	1
7	1,97	5,78	4,06	1,55	1	1	1	1	1	1
8	3,12	1,12	1,91	0,22	3	3	2	2	2	2
9	4,02	2,15	2,74	2,98	1	2	1	1	1	1
10	1,62	1,50	1,09	1,21	2	3	2	3	3	3

Branch	Критерии				Методы стратификации					
	D	C	S	FX	LS+S	PS+S	BC+S	AR+S	LW+S	EK+S
11	1,93	1,00	0,19	1,61	2	3	3	3	3	2
12	2,72	0,66	2,87	0,88	2	2	2	2	2	2
13	2,78	0,81	1,65	0,71	3	3	2	3	2	2
14	2,36	2,20	1,23	1,68	2	2	1	2	2	1
15	2,13	3,93	3,07	1,47	1	1	1	2	1	1
16	1,07	1,06	1,58	0,57	3	3	3	3	3	2
17	1,31	1,16	0,26	0,21	3	3	3	3	3	3
18	2,24	0,95	1,21	1,08	2	3	2	3	3	2
19	2,66	5,49	0,92	2,64	1	1	1	1	1	1
20	2,93	1,08	2,60	0,67	2	3	2	2	2	2
21	1,66	1,73	0,62	0,24	3	3	3	3	3	3
22	4,06	0,71	0,81	0,70	3	2	2	3	2	2
23	3,11	1,54	0,98	3,85	1	1	1	2	1	1

Таблица 14. Значения парных расстояний Кемени – Снелла для разбиений на страты различными методами

	LS+S	PS+S	BC+S	AR+S	LW+S	EK+S
LS+S	–	0,38	0,36	0,34	0,44	0,38
PS+S	–	–	0,34	0,36	0,26	0,34
BC+S	–	–	–	0,32	0,24	0,32
AR+S	–	–	–	–	0,26	0,34
LW+S	–	–	–	–	–	0,26
EK+S	–	–	–	–	–	–

Значения расстояний Кемени – Снелла между стратификациями в табл. 14 показывают, что наиболее схожими оказались LW+S и BC+S.

Поскольку предложенные методы LS и PS, помимо стратификации, позволяют получить ранжирование вариантов, было интересно сравнить их с методами многокритериального ранжирования, а также с ранжированиями, полученными по значениям каждого из критериев оценки отделений банка. В табл. 15 указаны все ранжирования отделений банка, а в табл. 16 результаты работы методов сравниваются между собой с помощью расстояния Хэмминга.

Таблица 15. Ранжирование отделений банков по отдельным критериям и методами многокритериального ранжирования и стратификации

Branch	D	C	S	FX	BC+R	PM+R	LS+R	AR+R	LW+R
1	6	12	18	19	8	3	19	12	14
2	19	21	17	20	23	3	22	23	22
3	4	23	2	10	3	1	8	6	6
4	16	7	20	9	16	3	9	16	16
5	5	9	9	13	5	2	10	11	8
6	1	4	3	2	1	1	1	1	1
7	17	1	1	7	11	1	2	2	2
8	7	14	8	22	9	2	18	14	13
9	3	6	6	3	2	1	5	5	3
10	21	11	14	11	18	3	11	18	18
11	18	17	23	6	20	3	13	17	19
12	11	22	5	14	14	2	12	8	12
13	10	19	10	15	15	2	16	15	15
14	13	5	12	5	13	2	7	13	10
15	15	3	4	8	10	1	6	7	5
16	23	16	11	18	21	3	17	20	21
17	22	13	22	23	22	4	23	22	23
18	14	18	13	12	17	2	15	21	17
19	12	2	16	4	6	1	3	3	4
20	9	15	7	17	12	2	14	9	9
21	20	8	21	21	19	4	21	19	20
22	2	20	19	16	4	2	20	10	11
23	8	10	15	1	7	1	4	4	7

Таблица 16. Значения парных расстояний Хэмминга между упорядочениями объектов методами многокритериального ранжирования и стратификации, а также по отдельным критериям. Здесь $\sum$ – суммарное расстояние (сумма значения по столбцу выше строки $\sum$) между результатами многокритериального ранжирования и ранжирования по отдельному критерию

Branch	**D**	**C**	**S**	**FX**	**BC+R**	**PS+R**	**LS+R**	**AR+R**	**LW+R**
D	–	0,49	0,36	0,40	0,13	0,29	0,40	0,26	0,24
C	–	–	0,45	0,34	0,40	0,41	0,28	0,35	0,33
S	–	–	–	0,39	0,32	0,24	0,30	0,26	0,24
FX	–	–	–	–	0,33	0,25	0,11	0,26	0,25
$\sum$	–	–	–	–	1,18	1,19	1,09	1,13	1,06
BC+R	–	–	–	–	–	0,22	0,28	0,15	0,13
PS+R	–	–	–	–	–	–	0,22	0,18	0,15
LS+R	–	–	–	–	–	–	–	0,18	0,17
AR+R	–	–	–	–	–	–	–	–	0,08
LW+R	–	–	–	–	–	–	–	–	–

Суммарное расстояние от упорядочивания по каждому из критериев до отдельного многокритериального ранжирования оказалось наименьшим для метода LW+R. То есть этот метод дает наиболее согласованное с каждым отдельным критерием ранжирование. Самые близкие между собой ранжирования дали методы LW+R и AR+R.

Заключение

В работе рассматривается задача многокритериальной стратификации, суть которой заключается в разделении объектов на упорядоченные по предпочтительности классы – страты, таким образом, чтобы варианты, входящие в одну и ту же страту, были более или менее одинаковы и в то же время в основном лучше, чем варианты из предыдущей страты.

Основными результатами работы можно считать следующие:

1. Предложены два новых метода стратификации, в некоторой мере аппроксимирующие слои несравнимых по Парето объектов. Метод линстрат ищет страты в виде линейных образований, при этом позволяет автоматически находить веса критериев и ранги объектов. Метод паретострат позволяет из разбиения Парето сформировать необходимое чис-

ло страт, что достигается «склеиванием» близлежащих «тонких» слоев отношения Парето.

2. Предложены три модели генерации стратифицированных данных, которые имитируют ситуацию, возникающую в многокритериальных задачах на реальных данных.

3. Для этих моделей генерации страт проведено экспериментальное сравнение методов стратификации. Оказалось, что единого победителя нет – метод-победитель зависит от формы генерируемых страт.

Преимуществом критерия линстрат является его ясность и прозрачность. Мы ожидали, что метод линстрат окажется наилучшим для случая линейных страт. Однако в некоторых экспериментах метод линстрат оказался не только не лучше, но даже и хуже, чем другие методы. По нашему мнению, это связано с недостатками используемого метода эволюционной оптимизации, слишком сильно опирающегося на вероятностные механизмы. Мы надеемся, что использование более регулярных подходов к оптимизации улучшит метод и поставит его на должную высоту. Уже сейчас видно, что надежда имеет основания – использование модификаций метода Ньютона действительно может резко повысить эффективность метода линстрат. Более адекватные методы будут представлены в наших следующих разработках.

Литература

Алескеров Ф.Т., Хабина Э.Л., Шварц Д.А. (2006) Бинарные отношения, графы и коллективные решения. М.: ГУ ВШЭ, 2006.

Белов В.В., Коричнева Ю.Л. (2012) Многомерная ABC-классификация. Критерии качества и канонические алгоритмы // Бизнес-информатика. 2012. № 1 (19). С. 9–16.

Берж К. Теория графов и ее применения. М., 1962.

Миркин Б.Г. (1974) Проблема группового выбора. М.: Наука, 1974.

Миркин Б.Г. (2011) Методы кластер-анализа для поддержки принятия решений: обзор: препринт WP7/2011/03. М.: Изд. дом Высшей школы экономики, 2011.

Подиновский В.В., Подиновская О.В. (2011) О некорректности метода анализа иерархий // Проблемы управления. 2011. № 1. С. 8–13.

Aleskerov F., Ersel H., Yolalan R. (2004) Multicriteria ranking approach for evaluating bank branch performance // International Journal of Information Technology & Decision Making. No. 3(2). P. 321–335.

Burgess A., Davies U., Doyle M. et al. (2006) The Economist's pocket world in figures // The Economist in association with Profile Books Ltd. L., 2007.

De Smet Y., Montano Guzman L. (2004) Towards multicriteria clustering: an extension of the k-means algorithm // European Journal of Operational Research. No. 158. P. 390–398.

De Smet Y., Gilbart F. (2001) A class definition method for country risk problems. Technical report IS-MG.

Fogel D.B. (1995) Evolutionary Computation. Toward a New Philosophy of Machine Intelligence // IEEE Press. Piscataway.

Kemeny J., Snell L. (1962) Mathematical Models in the Social Sciences. Ginn, Boston.

Kennedy J., Eberhart R.C. (2001) Swarm Intelligence. Morgan Kaufmann Publishers. San Francisco. Calif. USA.

MacQueen J. (1967) Some methods for classification and analysis of multivariate observations // Le Cam, J. Neyman (eds.) 5th Berkeley Symp. Math Statist. Prob. No. 1. P. 281–297.

Mirkin B. (2005) Clustering for Data Mining: A Data Recovery Approach. Boca Raton. Fl., Chapman and Hall/CRC. 2005.

Nemery Ph., De Smet Y. (2005) Multicriteria ordered clustering. Technical Report TR/SMG/2005-003. Universite Libre de Bruxelles, 2005.

Ng W.L. (2007) A simple classifier for multiple criteria ABC analysis // European Journal of Operational Research. No. 177. P. 344–353.

Page L., Brin S., Motwani R., Winograd T. (1999) The PageRank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford InfoLab.

Ramanathan R. (2006) Inventory classification with multiple criteria using weighted linear optimization // Computers and Operations Research. No. 33. P. 695–700.

Saaty T.L. (1980) The analytic hierarchy process. McGraw-Hill: N. Y., 1980.

Sun Y., Han J., Zhao P. et al. (2009) RankClus: integrating clustering with ranking for heterogeneous information network analysis // Proc. EDBT. 2009. P. 565–576.

Wikipedia: Социальная стратификация, http://ru.wikipedia.org/wiki/Социальная_ стратификация (дата обращения: 22.12. 2012).

Zopounidis C., Doumpos M. (2002) Multicriteria classification and sorting methods: A literature review // European Journal of Operational Research. No.138. P. 229–246.

Mirkin, B. G., Orlov, M. A. Methods for multicriteria stratification and experimental comparisons [Text] : Working paper WP7/2013/06 / B. G. Mirkin, M. A. Orlov ; National Research University "Higher School of Economics". – Moscow : Publishing House of the Higher School of Economics, 2013. – 32 p. – (Series WP7 "Mathematical methods for decision making in economics, business and politics"). – 50 copies.

At multivariate ranking, one may concentrate not on the problem of strict ordering of the objects but rather on tying them up in groups of more or less similar entities. Following Sociology and Mineralogy, such tied groups can be referred to as strata. A popular "univariate" ABC-classification problem, in fact, partitions the set in three strata. This work proposes two novel algorithms for automatic stratification with a prespecified number of strata. One of our algorithms chooses criteria weights in such a way that the objects within each single stratum are located on the axis of the aggregate criterion as close to each other as possible. The other, in contrast, takes all the criteria to be incomparable, so that the strata approximate Pareto boundaries of the vector preference relation. These methods are experimentally compared with a set of known ranking methods, for which we propose some strata generation mechanisms.

Mirkin B. G. – International Laboratory of Decision Choice and Analysis; Department of Applied Mathematics and Informatics NRU HSE, Moscow Russia; Department of Computer Science, Birkbeck University of London, UK.

Orlov M. A. – Department of Applied Mathematics and Informatics NRU HSE, Moscow, Russia.

Препринт WP7/2013/06
Серия WP7
Математические методы анализа решений
в экономике, бизнесе и политике

Б.Г. Миркин, М.А. Орлов

Методы многокритериальной стратификации и их экспериментальное сравнение

Зав. редакцией оперативного выпуска *А.В. Заиченко*
Технический редактор *Ю.Н. Петрина*

Отпечатано в типографии
Национального исследовательского университета
«Высшая школа экономики» с представленного оригинал-макета

Формат 60×84 $^{1}/_{16}$. Тираж 50 экз. Уч.-изд. л. 1,9
Усл. печ. л. 1,86. Заказ № . Изд. № 1560

Национальный исследовательский университет
«Высшая школа экономики»
125319, Москва, Кочновский проезд, 3
Типография Национального исследовательского университета
«Высшая школа экономики»