АЛГОРИТМ ФОРМИРОВАНИЯ МНОГОКРИТЕРИАЛЬНОЙ СТРАТИФИКАЦИИ

М.А. ОРЛОВ
аспирант кафедры анализа данных и искусственного интеллекта, факультет компьютерных наук, Национальный исследовательский университет «Высшая школа экономики»

Адрес: 101000, г. Москва, ул. Мясницкая, д. 20
E-mail: ormian@mail.ru

Данная работа развивает подход к проблеме многокритериального ранжирования, называемый нами многокритериальной стратификацией. При таком подходе представляет интерес не столько полное упорядочивание вариантов, сколько разбиение множества вариантов на заданное число классов, упорядоченных по предпочтению. Ранжирование получается путем линейной свертки критериев с весами. При этом веса определяются, исходя из предположения о наличии некоторой структуры в данных, такой что можно выделить «параллельные» слои – страты.

В работе [6] авторы сформулировали задачу формирования оптимальной стратификации, как задачу оптимизации некоторой целевой функции относительно весов критериев, однако, предложенный в этой работе алгоритм решения получаемой задачи, использующий случайный поиск, оказался неэффективным по сравнению с другими методами стратификации.

В данной работе предлагается новый алгоритм оптимизации целевой функции многокритериальной стратификации на основе квадратичного программирования. Для всестороннего экспериментального исследования качества работы алгоритма предлагается усовершенствованная модель генерации искусственных стратифицированных данных. Новая модель генерации страт имеет больше параметров для настройки и позволяет гибко задавать геометрию страт: ориентацию, толщину, размах и интенсивность, что лучше учитывает структуру реальных данных. Предлагаемый алгоритм экспериментально сравнивается с существующими методами стратификации на искусственных данных, и показывается его преимущество в большинстве рассмотренных случаев. Рассматриваются два примера реальных данных – библиометрические показатели 118 научных журналов и характеристики публикационной активности 102 стран. На этих данных новый алгоритм приводит к хорошо интерпретируемым и адекватным результатам. Также оказалось, что на этих данных построенное алгоритмом многокритериальное разбиение наиболее согласовано с разбиениями, построенными по отдельно взятым критериям.

Ключевые слова: стратификация, многокритериальное ранжирование, взвешенная сумма, квадратичное программирование, оптимизация, библиометрия.

Введение

В работе [6] рассматривается проблема стратификации как поиск весов критериев таких, что элементы страт образуют компактные группы на оси обобщенного критерия. Для решения возникающей оптимизационной задачи определения оптимальных весовых коэффициентов критериев в этой работе использовался метод случайного поиска, имитирующий эволюцию. Предложенный метод стратификации экспериментально сравнивался с рядом других алгоритмов и, в общем и целом, уступил некоторым из них. В данной работе предлагается использовать более точный метод квадратичного программирования вместо случайного поиска. В такой модификации

метод стратификации оказывается вполне конкурентоспособным в вычислительных экспериментах. Применение метода к реальным данным приводит к хорошо интерпретируемым весовым коэффициентам критериев, а получаемые страты оказываются достаточно адекватны. Для проведения контролируемого вычислительного эксперимента на искусственных данных в настоящей работе предлагается усовершенствованная параметрическая модель генерации линейных страт. Этот механизм генерации позволяет в зависимости от параметров гибко задавать структуру страт, и тем самым моделировать многие ситуации, возникающие в задачах принятия решений.

Дальнейшее содержание включает четыре раздела. В первом разделе приводится краткий обзор существующих методов стратификации. Во втором разделе формулируется проблема многокритериальной стратификации. В третьем разделе приводится оптимизационная задача формирования многокритериальной стратификации и новый алгоритм её решения. Четвертый раздел содержит описание реальных и сгенерированных данных, а также результаты экспериментов.

Работа велась при частичной финансовой поддержке международной Лаборатории анализа и выбора решений НИУ ВШЭ. Автор выражает благодарность д.т.н. Б.Г. Миркину за поставленную задачу.

1. Постановка проблемы стратификации

Множество из N объектов, оцененных по M критериям, необходимо разделить на K непересекающихся групп, упорядоченных между собой таким образом, чтобы объекты из одной группы были бы как можно более схожи по предпочтению с объектами из той же группы и в основном лучше, чем объекты из групп с большим номером. Такое упорядоченное разбиение будем называть стратификацией, а полученные группы объектов стратами. Оценки объектов по критериям будем записывать в виде критериальной матрицы $X = \| x_{ij} \|$, где $i = 1, ..., N$ – объекты или варианты, $j = 1, ..., M$ – критерии, а x_{ij} – значение j – го критерия для i-го объекта. Искомое множество страт обозначим $S = \{S_1, ..., S_K\}$, где S_k – множество объектов, принадлежащих k-й страте ($k = 1, ..., K$). Объект из страты с номером k имеет более высокий ранг или предпочтительней чем объект из страты l, если $k < l$.

2. Критерий многомерной стратификации

В работе [6] предложена целевая функция для формирования оптимальной многокритериальной стратификации. Идеально, страты образуют параллельные гиперплоскости. Значения критериев i-го объекта удовлетворяют уравнению:

$$x_{i1}w_1 + x_{i2}w_2 + x_{iM}w_M = c_k + e_i,$$

если он принадлежит страте с номером k, где $c_k \in \{c_1, c_2, ..., c_K\}$ – искомые центры или «уровни» страт, искомые w_j – веса критериев, e_i – минимизируемая ошибка. Для получения стратификации, веса w, уровни страт с и разбиение S выбираются путем решения оптимизационной задачи (1):

$$\begin{cases} \sum_{k=1}^{K}\sum_{i \in S_k}^{N}\left(\sum_{j=1}^{M} x_{ij}w_j - c_k\right)^2 \underset{w,c,S}{\rightarrow} min \\ \sum_{j=1}^{M} w_j = 1,\ w_j \geq 0 \end{cases} \quad (1)$$

Смысл задачи: найти такую комбинацию весовых коэффициентов w, и такие уровни c_k агрегированного критерия $\sum_j w_j x_j$, чтобы значения агрегированного критерия на всех объектах k-й страты были как можно ближе к c_k [6].

3. Алгоритм решения оптимизационной задачи многокритериальной линейной стратификации

Для формирования стратификации предлагаемый алгоритм решает задачу (1) используя подход чередующейся минимизации, который заключается в поиске оптимального решения по одной переменной при фиксированных других. На таком принципе основан, например, один из вариантов алгоритма k-средних [11]. Решение задачи (1) ищется чередованием трех последовательных шагов:

1) При фиксированных весах w_j и центрах c_k найти оптимальное разбиение;

2) При фиксированных весах w_j и разбиении S найти оптимальные центры;

3) При фиксированных центрах c_k и разбиении S найти оптимальные веса w_j.

На первых двух шагах решение получается следующим образом: оптимальное разбиение находится присвоением каждому объекту номера страты с ближайшим центром, а центры вычисляются как средние значения взвешенного критерия для объектов внутри страт. На третьем шаге оптимальные

веса критериев получаются из решения задачи квадратичного программирования с ограничениями. Поскольку на данном этапе разбиение фиксировано, каждому объекту приписывается соответствующий ему центр страты $c_k(i) \in \{c_1, c_2, ..., c_K\}$, $i = 1 \dots N$.

Для удобства перепишем (1) в более компактном виде. Воспользуемся соотношением $c_k(i) = \sum_{j=1}^{M} c_k(i) w_j$, поскольку веса в сумме равны единице. Затем перепишем выражение внутри скобок (1) в следующем виде:

$$\sum_{k=1}^{K}\sum_{i \in S_k}^{N}\left(\sum_{j=1}^{M} x_{ij} w_j - c_k(i)\right)^2 = \sum_{k=1}^{K}\sum_{i \in S_k}^{N}\left(\sum_{j=1}^{M} x_{ij} w_j - c_k(i) w_j\right)^2 =$$

$$= \sum_{k=1}^{K}\sum_{i \in S_k}^{N}\left(\sum_{j=1}^{M} (x_{ij} - c_k(i)) w_j\right)^2 = \sum_{k=1}^{K}\sum_{i \in S_k}^{N}\left(\sum_{j=1}^{M} w_j \bar{x}_{ij}\right)^2$$

Сделав замену переменной $\bar{x}_{ij} = (x_{ij} - c_k(i))$ и введя обозначение $\bar{X} = \|\bar{x}_{ij}\|$ перепишем (1) в матричном виде:

$$\begin{cases} w^T \bar{X}^T \bar{X} w \underset{w}{\rightarrow} min \\ \sum_{j=1}^{M} w_j = 1 \\ w_j \geq 0 \end{cases} \qquad (2)$$

Для решения (2) можно воспользоваться одним из известных алгоритмов квадратичного программирования, например, методом активного множества (active-set algorithm) [7], имплементированным в пакете Matlab 7. Предлагаемый алгоритм стратификации на основе квадратичного програмирования Linstrat-Q приведен ниже.

Алгоритм Linstrat-Q

На входе:

- Объекты x_i, $i = 1 \dots N$;
- Число страт K;
- Число итераций T.

На выходе:

- Веса w;
- Центры страт c;
- Разбиение S.

1. Инициализировать веса w и центры страт с. Сгенерировать веса w случайно такие, что $w_j \geq 0$, $j = 1 \dots M$ и $\sum_{j=1}^{M} w_j = 1$. Вычислить свертку критериев с весами. Центры страт сгенерировать случайно равномерно из диапазона от минимального значения свертки критериев до максимального.

2. По заданным весам и центрам найти оптимальное разбиение:

$$x_i \in S_k \text{, где } k = argmin_k \left(\sum_{j=1}^{M} x_{ij} w_j - c_k\right)^2,$$

$$k = 1 \dots K,\ i = 1 \dots N.$$

3. По заданным весам и разбиению на страты найти оптимальные центры:

$$c_k = \frac{1}{|S_k|} \sum_{x_i \in S_k}^{N} \sum_{j=1}^{M} x_{ij} w_j .$$

4. По заданным центрам и разбиению найти оптимальные веса, решив задачу квадратичного программирования (2).

5. Завершить, если сделано T итераций, иначе перейти к п. 2.

Вопрос об инициализации алгоритмов, строящих разбиения, является одним из ключевых в кластерном анализе. В зависимости от начальных значений весов w и центров с могут получаться различные разбиения. Данная работа следует традиционному эвристическому подходу – задать начальные параметры случайно, найти решение для каждой такой инициализации, а затем выбрать решение, для которого значение целевой функции будет минимально. Другой так же важный вопрос – об определении оптимального числа страт – в данной работе не рассматривается

Продемонстрируем работу алгоритма по шагам на простом примере. Рассмотрим шесть объектов оцененных по двум критериям $x_1 = (0,1)$, $x_2 = (1,0)$, $x_3 = (1,2)$, $x_4 = (2,1)$, $x_5 = (2,3)$, $x_6 = (3,2)$.

Шаг 1. Произведем инициализацию весов и центров. Допустим, сгенерированные значения равны $w = (0.2, 0.8)$, $c_3 = 0.3$, $c_2 = 1.9$, $c_1 = 2.4$.

Шаг 2. Вычисляем свертку критериев с весами. Получаем значения для каждого объекта соответственно 0.8, 0.2, 1.8, 1.2, 2.8, 2.2. Далее для первого объекта вычисляем квадрат разности значения свертки и значения центра каждой страты $(c_k - x_1 w^T)^2$, $k = 1, 2, 3$. Минимальным является квадрат разности для третьей страты $(c_k - x_1 w^T)^2 = (0.3 - 0.8)^2 = 0.25$, следовательно, назначаем первый объект на страту с номером 3. Аналогично производим назначения для остальных объектов. Получаем, что объекты принадлежат соответственно стратам с номерами 3, 3, 2, 2, 1, 1.

Шаг 3. Теперь для каждой из страт вычисляем центры как средние значения взвешенных критериев объектов, принадлежащих заданной страте. Таким образом,

$$c_3 = \frac{1}{2}\cdot(0.8+0.2) = 0.5,\ c_2 = \frac{1}{2}\cdot(1.8+1.2) = 1.5,$$
$$c_1 = \frac{1}{2}\cdot(2.8+2.2) = 2.5\ .$$

Шаг 4. Вычисляем матрицу $\overline{X}$ для построения задачи квадратичного программирования (2):

$$\overline{X} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 2 \\ 2 & 1 \\ 2 & 3 \\ 3 & 2 \end{pmatrix} - \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \\ 1.5 & 1.5 \\ 1.5 & 1.5 \\ 2.5 & 2.5 \\ 2.5 & 2.5 \end{pmatrix} = \begin{pmatrix} -0.5 & 0.5 \\ 0.5 & -0.5 \\ -0.5 & 0.5 \\ 0.5 & -0.5 \\ -0.5 & 0.5 \\ 0.5 & -0.5 \end{pmatrix}$$

Матрица для задачи квадратичного программирования (2) будет иметь вид:

$$\overline{X}^T\,\overline{X} = \begin{pmatrix} 1.5 & -1.5 \\ -1.5 & 1.5 \end{pmatrix}$$

Решение задачи квадратичного программирования дает веса $w = (0.5, 0.5)$. Повторяем три вышеописанных шага заданное число итераций.

На шаге 2 для каждого из N объектов производится свертка с весами по M критериям и вычисляется наименьшая из K квадратов разностей, значит оценка сложности этого шага в худшем случае $O(NMK)$. На шаге 3 вычисляются средние значения взвешенных критериев внутри страт. Сложность этого шага $O(N)$. Далее, на шаге 4 требуется $O(NM)$ операций для вычисления матрицы $\overline{X}$. Произведение $\overline{X}^T\,\overline{X}$ требует $O(NM^2)$ операции. Обозначим $O(f(M))$ оценку сложности решения задачи квадратичного программирования (2), где f зависит от выбранного алгоритма оптимизации. При заданном T числе итераций оценка сложности алгоритма будет $O((NM(M+K)+f(M))T)$.

4. Экспериментальное сравнение алгоритмов стратификации

Для апробации рассматриваемых в работе алгоритмов и их экспериментального сравнения были взяты искусственные и реальные данные. Предложен усовершенствованный алгоритм генерации стратифицированных данных, позволяющий всесторонне контролировать конфигурацию страт. В качестве реальных данных были взяты библиометрические показатели журналов и стран.

Ниже, *в табл. 1* перечислены методы стратификации, используемые в работе для экспериментального сравнения. Более подробный обзор рассматриваемых методов приведен в работе [6].

Таблица 1.

Список методов стратификации и их аббревиатур, используемых в данной работе

Метод стратификации	Аббревиатура	Источник
Метод линейной стратификации Linstrat-Q с использованием квадратичного программирования	LSQ	Эта работа
Метод линейной стратификации Linstrat на основе эволюционной минимизации	LS	[6]
Стратификация с помощью правила Борда (Borda count)	BC	[2]
Метод ABC- классификации на основе линейной оптимизации весов (Linear weights optimization)	LWO	[13]
Ранжирование по влиянию (Authority ranking)	AR	[19]
Стратификация объединением границ Парето (Paretostrat)	PS	[6]

Все методы были имплементированы в среде Matlab 7. Для решения задачи квадратичного программирования (2) использовалась функция quadprog из библиотеки Matlab Optimization Toolbox. Для решения задачи линейного программирования для метода LWO использовалась функция linprog из вышеупомянутой библиотеки. Для BC, LWO, AR после получения одномерного ранжирования стратификация производилась по численным значениям агрегированного критерия. Для этого использовался алгоритм k-средних. Случайная инициализация для LS, LSQ, а также k-средних для BC, LWO, AR выполнялась 100 раз, и записывался результат, дающий наименьшее значение целевой функции. Число итераций T для алгоритма LS и LSQ в экспериментах было установлено равным 100.

Для оценки качества стратификации при контролируемом экспериментальном исследовании на искусственных данных использовался показатель точности стратификации, т. е. отношение объектов с правильно определенным номером страты к общему числу объектов:

$$accuracy = \frac{N_{correct}}{N} \qquad (3)$$

На реальных данных такая мера качества не применима, поскольку заведомо не известно к какой страте относится объект. Мы полагаем, что хорошее многокритериальное разбиение должно быть согласованно с разбиением по каждому отдельно взятому критерию. Поэтому для оценки качества работы алгоритма на реальных данных мы ис-

пользовали среднее расстояние от многокритериальной стратификации до 3 однокритериальных стратификаций по каждому отдельно взятому критерию. В качестве расстояния между стратификациями S и R использовалось нормированное расстояние Кемени-Снелла [4, 10]. Для стратификации S, заданной на объектах x_1, ..., x_N строилась матрица:

$$s_{ij} = \begin{cases} 1, S(x_i) > S(x_j) \\ 0, S(x_i) = S(x_j), \\ -1, S(x_i) < S(x_j) \end{cases} \quad (4)$$

где $S(x)$ обозначает номер страты, которой принадлежит объект x. Аналогично формируется матрица стратификации R. Расстояние вычисляется по формуле:

$$d_{RS} = \frac{1}{2N(N-1)} \sum_{i,j=1}^{N} | R_{ij} - S_{ij} |. \quad (5)$$

4.1 Модель генерации стратифицированных данных

Для того чтобы обеспечить возможность проведения контролируемых экспериментов по сравнению алгоритмов, в работе [6] был предложен случайный механизм генерации стратифицированных данных.

В данной работе предлагается усовершенствованная версия алгоритма генерации, позволяющая более разносторонне контролировать геометрическую конфигурацию страт. Алгоритм генерирует страты в виде параллельных гиперплоскостей. Более точно, геометрия страт задается различными параметрами, к которым относятся: весовые коэффициенты w, уровни страт c, интенсивности страт θ, размах страт φ и толщина страт σ. Смысл этих величин раскрывается в нижеследующем алгоритме генерации стратифицированных данных. Список значений, задаваемых по умолчанию параметров генерации, приведен *в табл. 1 Приложения 1.* На *рис. 1* наглядно продемонстрировано то, как выглядят страты в зависимости от параметров генерации. Алгоритм генерации описан ниже.

Алгоритм генерации стратифицированных данных

На входе:

- Число объектов N, размерность M и число страт K;
- Уровни страт c;
- Весовые коэффициенты (ориентация страт) w;
- Толщина страт σ;
- Интенсивности страт θ;
- Размах страт φ.

На выходе:

- Значения критериев для каждого объекта;
- Индекс страты для каждого объекта.

1. Выбрать номер страты из мультиномиального распределения
$k \sim M(\theta_1, \theta_2, ..., \theta_K)$
2. Выбрать величину, моделирующую значение агрегированного критерия (значения свертки критериев с весами) на объекте, из нормального распределения $r \sim N(c_k, \sigma)$
3. Сгенерировать M-1 координат из равномерного распределения:
$x_j \sim U(c_k(1-\varphi), c_k(1+\varphi)/w_j), j = 1 \ldots M\text{-}1.$
4. Вычислить последнюю M-ю координату из уравнения гиперплоскости:
$x_M = (r - w_1x_1 + w_2x_2 + ... + w_{M-1}x_{M-1})/w_M.$
5. Завершить, если сгенерировано N объектов, иначе перейти к п. 1.

4.2. Реальные данные. Библиометрические показатели журналов и стран

Реальные данные были взяты с портала scimagojr.com, разработанного исследовательской группой SCImago (http://www.scimagolab.com/). Этот портал содержит в открытом доступе библиометрические показатели журналов и стран, полученные на основе данных из базы Scopus (http://www.elsevier.com/online-tools/scopus). Данные из этого источника широко используются в научных разработках (см., например, [1, 17, 18]).

4.2.1 Данные о библиометрических показателях журналов

Рассмотрим библиометрические показатели научных журналов из раздела Artificial Intelligence за 2012 год. Всего в этом разделе 118 журналов. В качестве критериев оценки престижа журнала рассмотрим три наиболее популярных библиометрических показателя:

1) Индекс SJR (Scientific Journal Ranking) [8]. Его значение отражает среднее число посещений журнала неким условным читателем, со-

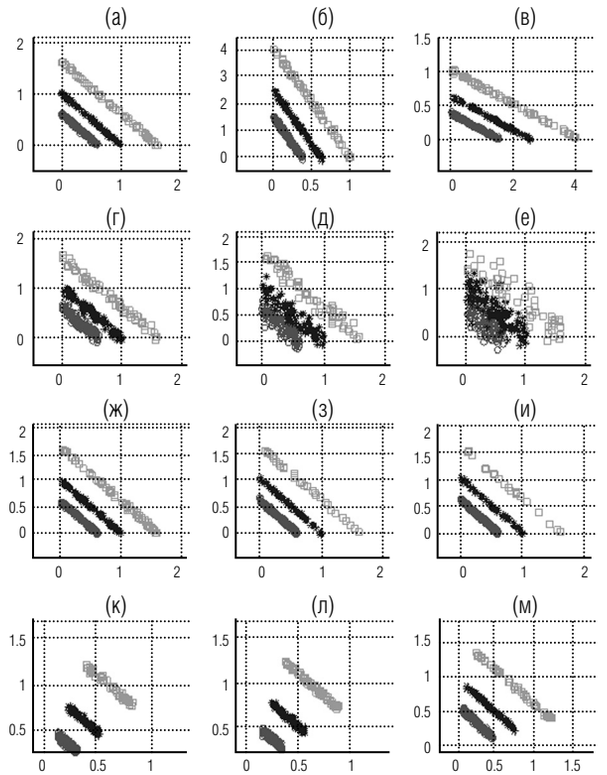

Рис. 1. Линейные страты для различных комбинаций параметров генерации в двумерном случае, $M = 2$. Приведены следующие варианты:

(1) страты в зависимости от ориентации w
(а) $w = (0.5, 0.5)$, (б) $w = (0.8, 0.2)$, (в) $w = (0.2, 0.8)$;
(2) страты в зависимости от толщины
(г) $\sigma = 0.05$,(д), $\sigma = 0.1$ (е), $\sigma = 0.2$;
(3) страты в зависимости от интенсивности
(ж) $\theta = (0.5, 0.3, 0.2)$, (з) $\theta = (0.7, 0.2, 0.1)$, (и) $\theta = (0.8, 0.15, 0.05)$;
(4) страты в зависимости от размаха
(к) $\varphi = 0.05$, (л) $\varphi = 0.1$, (м) $\varphi = 0.5$.

вершающим случайные переходы по ссылкам на документы журналов. Данный индекс развивает идею, лежащую в основе известного алгоритма пейджранк [12].

2) Индекс Хирша (H) [9]. Этот индекс характеризует количество статей h, опубликованных журналом, таких, что число их цитирований выше значения *h*.

3) Импакт-фактор журнала (I) [20]. Этот индекс рассчитывается отнесением числа цитирований статей, опубликованных в журнале в таком-то году, за последующие два года, к числу статей, опубликованных в журнале в том самом году.

Более подробное обсуждение вышеперечисленных и других наукометрических показателей, их достоинств и недостатков может быть найдено в ряде работ, например, в [1, 3, 5].

4.2.2 Данные о библиометрических показателях публикационной активности стран

Из упомянутого выше источника были взяты данные о публикационной активности 102 стран за 2012 год, в предметной области – искусственный интеллект (Artificial intelligence). Критерии, используемые для оценки стран, следующие**:**

1) Общее число документов опубликованных за 2012 (D);

2) Число цитируемых документов, опубликованных в 2012 году (CD);

3) Общее количество цитирований в 2012 году, полученных документами, опубликованными в этом же году (C).

4) Самоцитирование документов в 2012 году (country self-citations) (SC);

5) Среднее число цитирований в 2012 году документов, опубликованных в этом году (CPD);

6) H-индекс на уровне страны (H).

4.2.3 Нормировки данных

Для реальных данных мы использовали два способа нормировки: стандартная нормировка (6) (приведение значений критериев к диапазону [0,1]), и статистическая нормировка (7) (приведение значений к нулевому среднему и единичному стандартному отклонению):

$$z_{ij} = \frac{x_{ij} - min_j(x_{\cdot j})}{max_j\left(x_{\cdot j}\right) - min_j(x_{\cdot j})} \qquad (6)$$

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \qquad (7)$$

В формуле (6) $min_j(x_{\cdot j})$ и $min_j(x_{\cdot j})$ максимальное и минимальное значение по *j*-му критерию. В формуле (7) μ_j и σ_j среднее значение и стандартное отклонение значений *j*-го критерия.

4.3 Результаты экспериментов на искусственных данных

В этом разделе приведены результаты экспериментов по стратификации, в зависимости от параметров генерации страт: размерности, общего количества сгенерированных объектов, интенсивности, размаха и толщины страт. Критерий оценки качества работы алгоритма – точность стратификации (3). Результаты экспериментов приведены в *табл. 2-5 Приложения 1.*

Предлагаемый алгоритм LSQ оказался достаточно устойчивым к увеличению размерности (*табл. 2, Приложение 1*). В тоже время точность стратификации методом LS значительно снизилась уже при размерности больших пяти. Точность стратификации заметно снизилась с ростом размерности также для методов LWO и PS. Алгоритм AR показал обратную тенденцию: точность стратификации с увеличением размерности возросла.

Снижение точности стратификации для LSQ при увеличении размерности может быть компенсировано увеличением объема выборки (*табл. 3, Приложение 1*). Такой же эффект «компенсации» наблюдается и для LWO. Методы AR и BC являются относительно нечувствительными к объему данных. А для LS и PS снижение точности стратификации при увеличении размерности не удается компенсировать увеличением объема данных.

Алгоритмы линейной стратификации LSQ и LS показали высокую устойчивость к изменению интенсивностей страт θ по сравнению с другими алгоритмами (*табл. 4, Приложение 1*).

При малых значениях размаха φ все алгоритмы показали достаточно высокие результаты (*табл. 5, Приложение 1*). Особенно сильно этот параметр влияет на работу алгоритма PS. Не так заметно, но, тем не менее, снижается и точность для AR и BC. В меньшей степени размах влияет на LWO и совсем не влияет на качество работы LS и LSQ.

Утолщение страт влияет на качество работы всех алгоритмов (*табл. 6, Приложение 1*). Алгоритмы LSQ, LS, показывающие отличную точность при небольших значениях σ, с увеличением σ начинают уступать алгоритму LWO, который оказался наиболее устойчивым к увеличению толщины страт.

Таким образом, практически во всех экспериментах LSQ входит в группу лидеров. Это выделяет его из всех других алгоритмов, которые могут оказаться чувствительными к тому или иному параметру.

4.4. Результаты экспериментов на реальных данных

В этом разделе приведены результаты экспериментов по стратификации на реальных данных, описанных в п. 4.2. Рассмотрена стратификация 118 научных журналов и 102 стран по библиометрическим показателям. Обсуждается согласованность разбиений. Оценкой степени согласованности служит среднее расстояние Кемени-Снелла (5) между многокритериальной стратификацией и стратификациями по каждому отдельному критерию.

4.4.1 Результаты стратификации академических журналов по библиометрическим показателям

Как видно из *табл. 2, Приложения 2*, наиболее согласованное разбиение для обоих типов нормировки получилось методами линейной стратификации LS и LSQ. Метод AR нашел согласованное разбиение только при стандартной нормировке.

Наш алгоритм LSQ сформировал одинаковые стратификации для обеих нормировок. При этом наибольшие веса получили индекс SJR и импакт-фактор (*см. табл. 1, Приложения 2*). То есть по этим двум критериям можно получить хорошо стратифицированное множество журналов. В то же время индекс Хирша получил небольшой вес – таким образом, алгоритм как бы присоединяется к критике этого индекса (см., например, сборник «Игра в цифирь» 2011). В первую страту вошли 6 журналов, во вторую 42 и в третью 70. В первой страте журналы, высоко ценимые в сообществе исследователей:

1. IEEE Transactions on Pattern Analysis and Machine Intelligence;
2. International Journal of Computer Vision;
3. Foundations and Trends in Machine Learning;
4. ACM Transactions on Intelligent Systems and Technology;
5. IEEE Transactions on Evolutionary Computation;
6. IEEE Transactions on Fuzzy Systems.

Несколько неожиданным является попадание в эту группу относительно нового журнала Foundations and Trends in Machine Learning. Этот журнал публикует высококачественные монографические обзоры по актуальным проблемам.

4.4.2 Результаты стратификации стран по библиометрическим показателям

Так же, как и в предыдущем эксперименте, наиболее согласованные разбиения получились по методам LS и LSQ (*см. табл. 2, Приложения 3*). Веса критериев записаны в *табл. 1, Приложения 3*, из них ненулевые веса получили, главным образом, два критерия: самоцитирование и индекс Хирша. Остальные критерии на данной выборке ведут себя неустойчиво, и не позволяют разделить ее на

«параллельные» слои. Это, прежде всего, характеристики общего количества статей (критерии 1 и 2 из п. 4.2.2), и характеристики цитирования в том же году, также носящие случайный характер: действительно, для действенного учета новой публикации необходимо время на развитие и адаптацию ее идей, а также на публикацию полученных результатов. Расчеты для данных за 2009 год (при этом в расчет берется цитирование работ за более долгий период 2009 - 2012) приводят к аналогичным результатам, при этом среднее число цитирований получает несколько больший вес – 13% вместо 5%, а веса критериев 1 и 2 остаются нулевыми.

Первую страту составили две страны: Китай, США. Во второй страте оказались 17 стран: Испания, Англия, Франция, Тайвань, Япония, Индия, Германия, Канада, Италия, Южная Корея, Австралия, Гонконг, Голландия, Сингапур, Швейцария, Израиль. Остальные 83 страны сформировали третью страту.

Заключение

В работе рассмотрена проблема многокритериальной стратификации на основе автоматического определения весовых коэффициентов критериев таким образом, чтобы получаемые страты образовывали компактные множества на оси обобщенного критерия. Предложен алгоритм, использующий квадратичную оптимизацию. Этот алгоритм экспериментально сравнен с группой существующих подходов. Как оказалось, в большинстве случаев предложенный алгоритм приводит к наилучшим решениям. Применение алгоритма к реальным данным приводит к достаточно хорошо интерпретируемым результатам.

В дальнейшем мы планируем проверить работу алгоритма на различных реальных данных, особенно в применении к динамическим рядам многомерных критериев с целью дальнейшего изучения проблем интерпретации и устойчивости получаемых решений. Заслуживает внимания, в частности, проблема устойчивости алгоритма относительно изменения толщины страт.

Существенным недостатком подхода к стратификации на основе рассматриваемого в работе критерия линейной стратификации является то, что данный критерий не учитывает расстояния между стратами. То есть, из нескольких решений, позволяющих получить одинаково плотные группировки на оси взвешенного критерия, не будет сделан выбор в пользу того, которое обеспечивает, максимальный зазор между стратами. В настоящее время исследуется возможность сформулировать критерий стратификации таким образом, чтобы преодолеть вышеупомянутый недостаток. ■

Литература

1. Алескеров Ф.Т., Писляков В.В., Субочев А.Н. Построение рейтингов журналов по экономике с помощью методов теории коллективного выбора. WP7/2013/03. – М.: Изд. дом ВШЭ, 2013.
2. Алескеров Ф.Т., Хабина Э.Л., Шварц Д.А. Бинарные отношения, графы и коллективные решения. М.: ГУ-ВШЭ, 2006.
3. Игра в цыфирь, или как теперь оценивают труд ученого. М.: Московский Центр непрерывного математического образования, 2011.
4. Миркин Б.Г. Проблема группового выбора. М.: Наука, 1974.
5. Миркин Б.Г. О понятии научного вклада ученого и его измерителях // Управление большими системами: сборник трудов. 2013. С. 1-16.
6. Миркин Б.Г., Орлов М.А. Методы многокритериальной стратификации и их экспериментальное сравнение. WP7/2013/03. М.: Изд. дом ВШЭ, 2013.
7. Gill P.E., Murray W., Wright M.H. Numerical linear algebra and optimization. Vol. 1. Addison Wesley, 1991.
8. Gonzalez-Pereira B., Guerrero-Bote V., Moya-Anegon F. A new approach to the metric of journals scientific prestige: The SJR indicator // Journal of Informetrics. 2010. Vol. 4, issue 3. P. 379-391.
9. Hirsch J.E. An index to quantify an individual's scientific research output. 2005 (Retrieved from arXiv 05 February 2014).
10. Kemeny J., Snell L. Mathematical models in the social sciences. Boston: Ginn, 1962.
11. Mirkin B.G. Clustering: A data recovery approach. CRC Press, 2012.
12. Page L., Brin S., Motwani R., Winograd T. The PageRank citation ranking: Bringing order to the Web. Technical Report. Stanford InfoLab, 1999.

13. Ramanathan R. Inventory classification with multiple criteria using weighted linear optimization // Computers and Operations Research. 2006. No. 33. P. 695-700.
14. SCImago Journal & Country Ranking. (2007). SJR – SCImago Journal & Country Rank. http://www.scimagojr.com. (дата обращения 14.01.2014).
15. SCImago Lab. http://www.scimagolab.com/ (дата обращения 14.01.2014).
16. Scopus. http://www.elsevier.com/online-tools/scopus (дата обращения 22.01.2014).
17. Siebelt M., Siebelt T., Pilot P., Bloem R.M., Bhandari M., Poolman R.W. Citation analysis of orthopedic literature; 18 major orthopedic journals compared for Impact Factor and SCImago // BMC Musculoskeletal Disorders. 2010.
18. Spreckelsen C., Deserno T.M., Spitzer K. Visibility of medical informatics regarding bibliometric indices and databases // BMC Medical Informatics and Decision Making. 2011.
19. Sun Y., Han J., Zhao P., Yin Z., Cheng H., Wu T. RankClus: Integrating clustering with ranking for heterogeneous information network analysis // Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology (EDBT 2009). March 23-26 2009, Saint-Petersburg, Russia. P. 565-576.
20. Garfield E. The Thomson Reuters Impact Factor. Thomson Reuters, 1994.

ПРИЛОЖЕНИЕ 1.
Результаты экспериментального исследования алгоритмов стратификации на искусственных данных

Таблица 1.
Значения параметров генерации страт по умолчанию

Параметр генерации страт	Значение по умолчанию
Число объектов, N	$N = 300$
Число критериев, M	$M = 5$
Число страт, K	$K = 3$
Центры страт, c	$c = (0.3, 0.5, 0.8)$
Интенсивности страт, θ	$\theta = (0.33, 0.33, 0.33)$
Веса критериев, w	$w = (0.2, 0.2, 0.2, 0.2, 0.2)$
Толщина страт, σ	$\sigma = 0.01$
Размах страт, φ	$\varphi = 1$

Таблица 2.
Точность стратификации (среднее значение и стандартное отклонение для 10 генераций) различными методами в зависимости размерности данных M

Метод	Размерность данных		
	M = 3	M = 5	M = 20
LSQ	**1.00±0.00**	**1.00±0.00**	0.84±0.02
LS	**1.00±0.00**	**1.00±0.00**	0.44±0.08
BC	0.90±0.02	0.73±0.03	**0.88±0.01**
LWO	**1.00±0.00**	0.98±0.01	0.45±0.04
AR	0.60±0.02	0.61±0.04	**0.88±0.03**
PS	**0.99±0.00**	0.36±0.02	0.32±0.04

Таблица 3.
Точность стратификации (среднее значение и стандартное отклонение для 10 генераций) в зависимости от количества сгенерированных объектов N при размерности $M = 20$

Метод	Число сгенерированных объектов		
	N = 200	N = 300	N = 500
LSQ	0.65±0.27	0.75±0.22	**1.00±0.00**
LS	0.49±0.11	0.43±0.11	0.44±0.10
BC	**0.87±0.03**	**0.88±0.02**	0.88±0.02
LWO	0.38±0.04	0.44±0.03	0.63±0.04
AR	**0.88±0.05**	**0.89±0.04**	0.88±0.02
PS	0.33±0.04	0.33±0.02	0.33±0.02

Таблица 4.
Точность стратификации (среднее значение и стандартное отклонение для 10 генераций) в зависимости от интенсивностей θ

Метод	Паттерн интенсивности страт		
	$\theta = (0.33, 0.33, 0.33)$	$\theta = (0.5, 0.3, 0.2)$	$\theta = (0.6, 0.4, 0.1)$
LSQ	**1.00±0.00**	**1.00±0.00**	**1.00±0.00**
LS	**1.00±0.00**	**1.00±0.00**	**1.00±0.00**
BC	0.75±0.03	0.77±0.03	0.56±0.11
LWO	0.99±0.01	0.95±0.03	0.88±0.05
AR	0.60±0.04	0.67±0.03	0.65±0.06
PS	0.36±0.02	0.20±0.03	0.09±0.02

Таблица 5.
Точность стратификации (среднее значение и стандартное отклонение для 10 генераций данных) в зависимости от размаха страт φ

Метод	Размах		
	$\varphi = 0.01$	$\varphi = 0.1$	$\varphi = 1$
LSQ	**1.00±0.00**	**1.00±0.00**	**1.00±0.00**
LS	**1.00±0.00**	**1.00±0.00**	**1.00±0.00**
BC	0.97±0.01	0.95±0.01	0.74±0.03
LWO	0.99±0.00	0.99±0.01	0.98±0.01
AR	0.67±0.04	0.65±0.05	0.59±0.02
PS	0.96±0.02	0.86±0.04	0.36±0.02

Таблица 6.
Точность стратификации (среднее значение и стандартное отклонение для 10 генераций данных) в зависимости от толщины σ

Метод	Толщина страт		
	$\sigma = 0.01$	$\sigma = 0.1$	$\sigma = 0.2$
LSQ	**1.00±0.00**	**0.99±0.01**	0.76±0.09
LS	**1.00±0.00**	**0.99±0.01**	0.77±0.09
BC	0.76±0.03	0.73±0.03	0.71±0.03
LWO	**0.98±0.01**	0.96±0.01	**0.90±0.02**
AR	0.59±0.03	0.58±0.04	0.59±0.03
PS	0.35±0.02	0.38±0.02	0.42±0.03

ПРИЛОЖЕНИЕ 2.
Анализ многокритериальной стратификации журналов по библиометрическим показателям

Таблица 1.
Веса критериев важности научных журналов, найденные методом линейной стратификации LSQ

Критерий	Веса критериев, найденные при стратификации	
	Стандартная нормировка (7)	Статистическая нормировка (8)
Индекс SJR	0.38	0.47
Индекс Хирша, H	0.14	0.14
Импакт фактор, I	0.47	0.39

Таблица 2.
Средние значения расстояний КемениСнелла между упорядоченными разбиениями журналов по отдельным критериям и разбиениями, найденными методами многокритериальной стратификации

Метод	LSQ	LS	BC	LWO	AR	PS
Среднее расстояние. (Стандартная нормировка)	**0.12**	**0.12**	0.17	0.13	**0.12**	0.23
Среднее расстояние. (Статистическая нормировка)	**0.12**	**0.12**	0.17	0.15	0.20	0.23

ПРИЛОЖЕНИЕ 3.
Анализ многокритериальной стратификации стран по публикационной активности

Таблица 1.
Веса критериев оценки публикационной активности стран, найденные методом линейной стратификации LSQ

Критерий	Веса критериев, найденные при стратификации	
	Стандартная нормировка (7)	Статистическая нормировка (8)
1	0	0
2	0	0
3	0	0
4	**0.63**	**0.52**
5	0.05	0.07
6	0.33	0.41

Таблица 2.
Средние значения расстояний Кемени-Снелла между упорядоченными разбиениями стран по отдельным критериям и разбиениями, найденными методами многокритериальной стратификации

Метод	LSQ	LS	BC	LWO	AR	PS
Среднее расстояние. (Стандартная нормировка)	**0.10**	**0.10**	0.26	0.21	0.16	0.18
Среднее расстояние. (Статистическая нормировка)	**0.10**	**0.10**	0.26	0.17	0.12	0.18

AN ALGORITHM FOR MULTICRITERIA STRATIFICATION

Mikhail ORLOV
Post-graduate Student, Department of Data Analysis and Artificial Intelligence,
Faculty of Computer Science,
National Research University Higher School of Economics
Address: 20, Myasnitskaya street, Moscow, 101000, Russian Federation
E-mail: ormian@mail.ru

This paper elaborates an approach to the problem of multicriteria ranking referred to as multicriteria stratification. The target of stratification is an ordered partition with predefined number of classes – strata rather than a complete ranking of the set of objects. Ranking is computed by means of linear convolution of criteria with some weights. These weights are based on assumption that data can fit some linear structure so that «parallel» layers can be identified - strata.

In the paper [6] the authors formulated the problem of multicriteria stratification as a task of minimization of a cost function depending on criteria weights; however the algorithm proposed in that paper to address the emerging task based on random searching has demonstrated low performance in comparison to some other stratification approaches.

In this paper a new algorithm based on quadratic programming is proposed to optimize the multicriteria stratification target function. A more sophisticated synthetic data generator for a comparative study of the stratification algorithm has been developed. The new data generator has more parameters to tune and allows more flexible control of geometry of synthetic strata: orientation, thickness, spread and intensity of layers that enables to pay due regard to real data structure.

The novel algorithm has been compared experimentally with existing stratification approaches by involving synthetic data, and its competitiveness has been shown in the majority of case studies. Two real-world datasets have been processed – bibliometrical indicators of 118 scientific journals and parameters of publication activities of 102 countries. The new algorithm applied to handle these data has produced sensible and well interpretable outputs. Furthermore, on these data the proposed algorithm found the most coherent multicriteria stratification to those computed by each single criterion.

Key words: stratification, rank aggregation, multicriteria, weighted sum, quadratic programming, optimization, bibliometrics.

References

1. Aleskerov F.T., Pisljakov V.V., Subochev A.N. (2013) *Postroenie rejtingov zhurnalov po jekonomike s pomoshh'ju metodov teorii kollektivnogo vybora* [Rankings of economic journals constructed by the Social Choice Theory methods]. WP7/2013/03. – Moscow: HSE. (in Russian)
2. Aleskerov F.T., Habina E.L., Shvarc D.A. (2006) *Binarnye otnoshenija, grafy ikollektivnye reshenija* [Binary relations, graphs and group choice]. Moscov: HSE. (in Russian)
3. MCNMO (2011) *Igra v cyfir', ili kak teper' ocenivajut trud uchjonogo* [The game of numbers, or how scientific research is evaluated today] // Moscow: MCNMO. (in Russian)
4. Mirkin B.G. (1974) *Problema gruppovogo vybora* [Group choice problem]. Moscow: Nauka. (in Russian)
5. Mirkin B.G. (2013) O ponjatii nauchnogo vklada i ego izmeriteljah [On the concept of scientific contribution and its metrics] // *Upravlenie bol'shimi sistemami: sbornik trudov* [Big systems management: Proceedings], pp. 1-16. (in Russian)
6. Mirkin B.G., Orlov M.A. (2013) *Metody mnogokriterial'noj stratifikacii i ih jeksperimental'noe sravnenie* [Methods for multicriteria stratification and experimental comparisons]. WP7/2013/03. – Moscow: HSE. (in Russian)
7. Gill P.E., Murray W., Wright M.H. (1991) *Numerical linear algebra and optimization*, vol. 1. Addison Wesley.
8. Gonzalez-Pereira B., Guerrero-Bote V., Moya-Anegon F. (2010) A new approach to the metric of journals scientific prestige: The SJR indicator // *Journal of Informetrics*, vol. 4, no. 3, pp. 379-391.
9. Hirsch J.E. (2005) *An index to quantify an individual's scientific research output* (Retrieved from arXiv 05 February 2014).
10. Kemeny J., Snell L. (1962) *Mathematical models in the social sciences.* Boston: Ginn.

11. Mirkin B.G. (2012) *Clustering: A data recovery approach*. CRC Press.
12. Page L., Brin S., Motwani R., Winograd T. (1999) *The PageRank citation ranking: Bringing order to the Web*. Technical Report. Stanford InfoLab.
13. Ramanathan R. (2006) Inventory classification with multiple criteria using weighted linear optimization // *Computers and Operations Research*, no. 33, pp. 695-700.
14. SCImago Journal & Country Ranking. (2007). SJR — SCImago Journal & Country Rank. http://www.scimagojr.com. (Accessed 14 January 2014)
15. SCImago Lab. http://www.scimagolab.com/ (Accessed 14 January 2014).
16. Scopus. http://www.elsevier.com/online-tools/scopus (Accesed 22 January 2014).
17. Siebelt M., Siebelt T., Pilot P., Bloem R.M., Bhandari M., Poolman R.W. (2010) *Citation analysis of orthopedic literature*; 18 major orthopedic journals compared for Impact Factor and SCImago // BMC Musculoskeletal Disorders.
18. Spreckelsen C., Deserno T.M., Spitzer K. (2011) *Visibility of medical informatics regarding bibliometric indices and databases* // BMC Medical Informatics and Decision Making.
19. Sun Y., Han J., Zhao P., Yin Z., Cheng H., Wu T. (2009) RankClus: integrating clustering with ranking for heterogeneous information network analysis // Proceedings of the *12th International Conference on Extending Database Technology: Advances in Database Technology (EDBT 2009)*. March 23-26 2009, Saint-Petersburg, Russia, pp. 565-576.
20. Garfield E. (1994) *The Thomson Reuters Impact Factor*. Thomson Reuters.