Метод формирования многокритериальной стратификации и его использование в проблеме оценки научного вклада

Михаил Орлов
научный руководитель д. т. н. Миркин Б. Г.

НИУ-ВШЭ

ormian@mail.ru

18 декабря 2014 г.

Содержание

Понятие стратификации. В социологии.

Социальная стратификация — это деление общества на слои (страты) путём объединения различных социальных позиций с примерно одинаковым социальным статусом, выстроенное по горизонтали (социальная иерархия), вдоль своей оси по одному или нескольким стратификационным критериям (показателям социального статуса).

Рис.: Стратификация в социологии

Пример. Стратификация городов.

Таблица: Цены на жилье и еду в десяти городах мира для туриста. Значения цен приведены к диапазону от 0 до 100.

City	Housing	Food
Moscow	97	56
London	93	62
Tokyo	100	44
Copenhagen	43	100
New-York	97	39
Peking	60	12
Sydney	34	20
Vancouver	13	10
Johannesburg	0	5
Buenos-Aires	14	0

Пример. Стратификация городов.

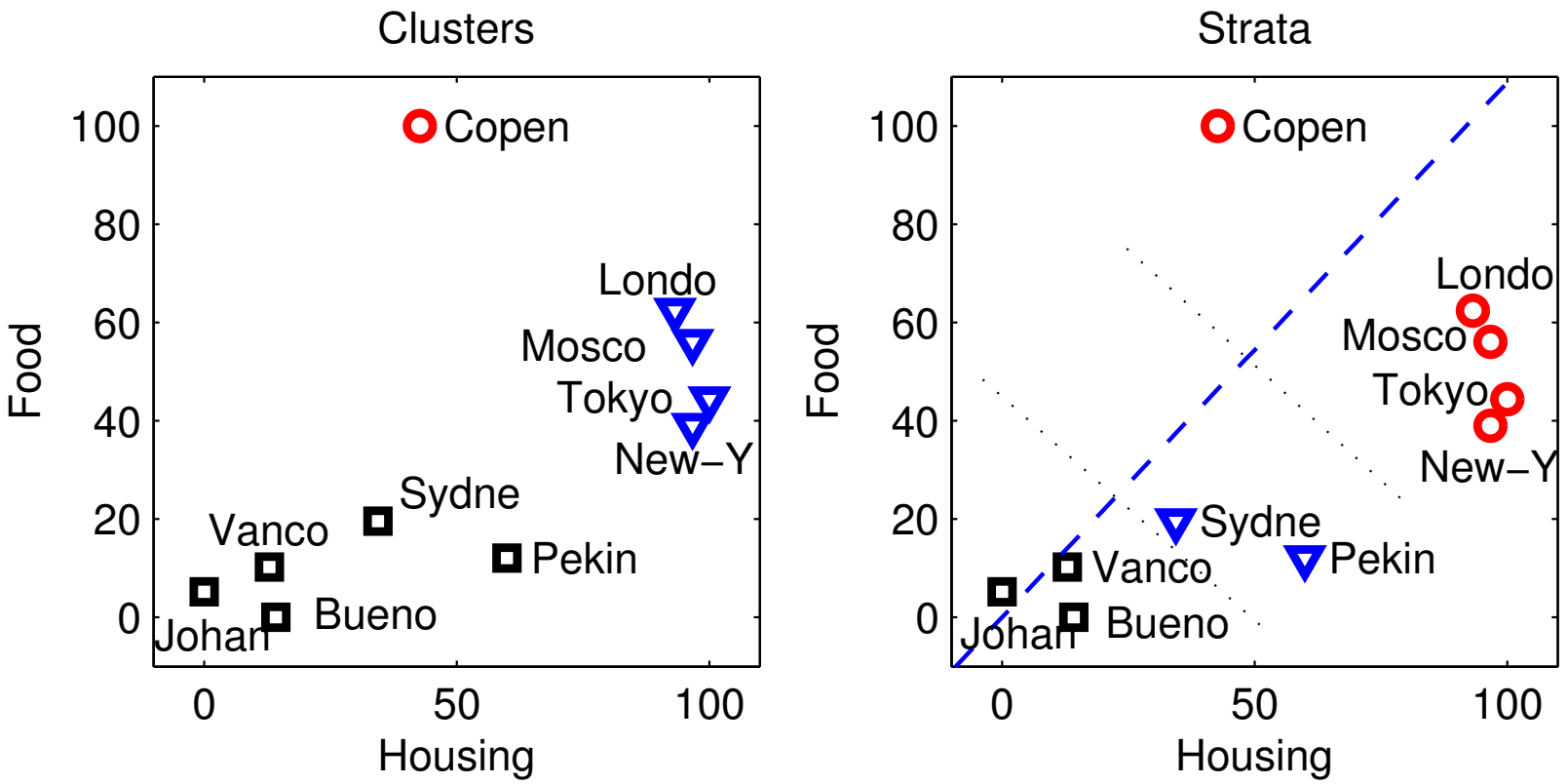

Рис.: Десять городов оцененных по двум критериям: цены на проживание и питание. Разбиение на три кластера (слева) разбиение на три страты (справа).

Предыдущая работа

- Методы многокритериального ранжирования: Borda count, Authority ranking[Sun 2009].
- ABC классификация [Ng 2007, Ramanathan 2006]

Цель работы

Цель

- Предложить метод, позволяющий автоматически строить не столько ранжирование, сколько упорядоченное разбиение объектов на заданное число классов.
- Метод должен вписываться в популярные схемы ранжирования, т.е. использовать взвешенную сумму критериев (линейность), но веса подбирать автоматически, как бы моделируя структуру множеств Парето.
- Применить метод в проблеме оценки научного вклада ученого

Линейная стратификаци. Обозначения

- Объекты $X = (x_{ij})$, где $i \in 1, ..., N$, $j \in 1, ..., M$, x_{ij} значение j–го критерия для i–го объекта.
- Веса критериев $w = (w_1, w_2, ..., w_M)$, где $w_j \geq 0$ и $\sum_j w_j = 1$.
- Взвешенный критерий $f_i = \sum_{j=1}^{M} w_j * x_{ij}$.
- $S = \{S_1, ..S_k, ..., S_K\}, k = 1, ..., K$ непересекающиеся различные подмножества S -страты
- Центры страт c_k, общие значения взвешенного критерия, такие что $c_k > c_l$, когда $k < l$.

Линейная стратификация. Наглядное представление.

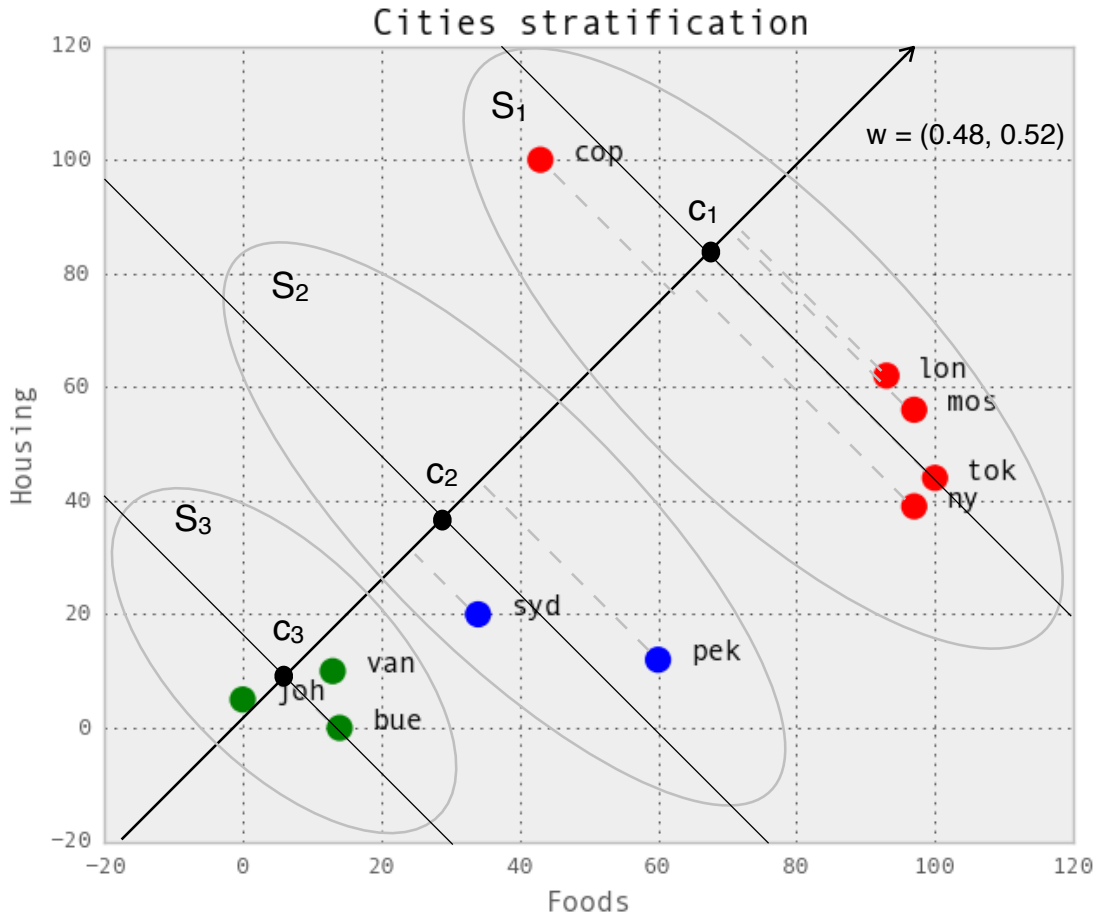

Рис.: Наглядное представление страт на примере городов

Линейная стратификация. Формулировка проблемы.

Формулировка проблемы

Найти такие веса w, центры c и разбиение, чтобы значения проекций объектов на ось взвешенного критерия были как можно ближе к соответствующим центрам. То есть невязка e_i для $f_i = c_{k(i)} + e_i$ была бы минимальна.

Оптимизационная задача

Оптимизационная задача линейной стратификации

$$\begin{aligned} \min_{w,c,S} \quad & \sum_{k=1}^{K} \sum_{i \in S_k} \left(\sum_{j=1}^{M} x_{ij} w_j - c_k \right)^2 \\ \text{such that} \quad & \sum_{j=1}^{M} w_j = 1 \\ & w_j \geq 0, j \in 1...M. \end{aligned} \tag{1}$$

Схема алгоритма решения оптимизационной задачи линейной стратификации

1. Инициализировать веса и центры
2. При заданных весах и центрах найти разбиение. Назначить объект на страту с ближайшим по оси взвешенного критерия центром
3. При заданных весах и разбиении найти центры как средние значения взвешенного критерия внутри страт
4. При заданном разбиении и центрах найти веса решив задачу квадратичного программирования относительно весов
5. Остановить итерации, когда разница значений целевой функции для последовательных итерации не станет меньше заданного значения

Верификация алгоритма на синтетических данных

Методы для сравнения:

- Linstrat QP
- Linstrat Evolutionary
- Borda count
- Linear Weights Optimization
- Authority ranking

Параметры генерации синтетических данных:

- (а,б,в) ориентация страт
- (г,д,е) толщина страт
- (ж,з,и) интенсивность страт
- (к,л,м) размах страт
- размерность данных
- размер выборки

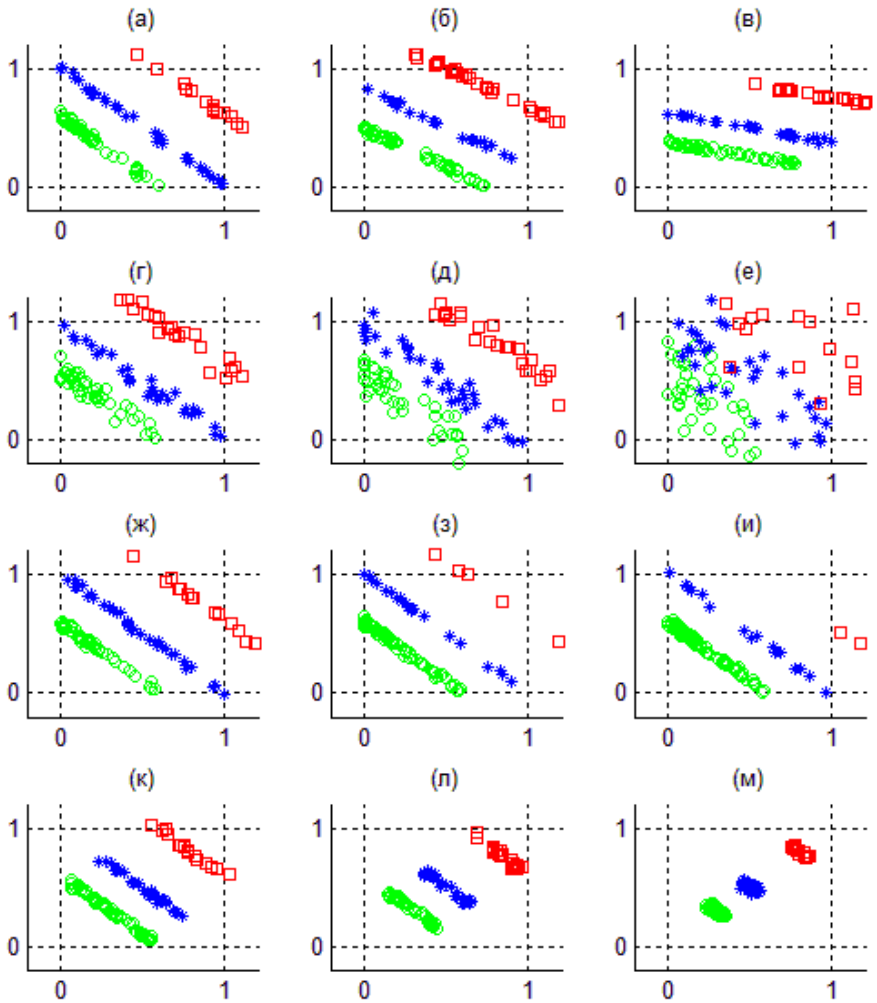

Рис.: Синтетические данные

Приложения метода линейной стратификации

Приложения

- Стратификация и ранжирование научных журналов по библиометрическим показателям
- Стратификация и ранжирование страт стран по различным показателям: экономическим, научным и т. д.
- Построение рейтингов университетов по многим критериям
- Оценка научного вклада

Три подхода к оценке научного вклада

Уровень цитирования

- Общее число цитирований
- Число работ получивших не менее 10 цитирований
- H-индекс. Число работ h получивших по меньшей мере h цитирований

Уровень достижений или заслуг

- Число организованныйх или соорганизованных конференций.
- Число подготовленных кандидатов наук (PhD)
- Участие в рецензировании журналов в роли главного редактора, зам. главного редктора (в любое время) или членство в редколлегии на момент 2013 года.

Таксономический ранг

- Таксономический ранг, на основе классификации предметной области [Миркин 2013]

Таксономический ранг. Пояснение.

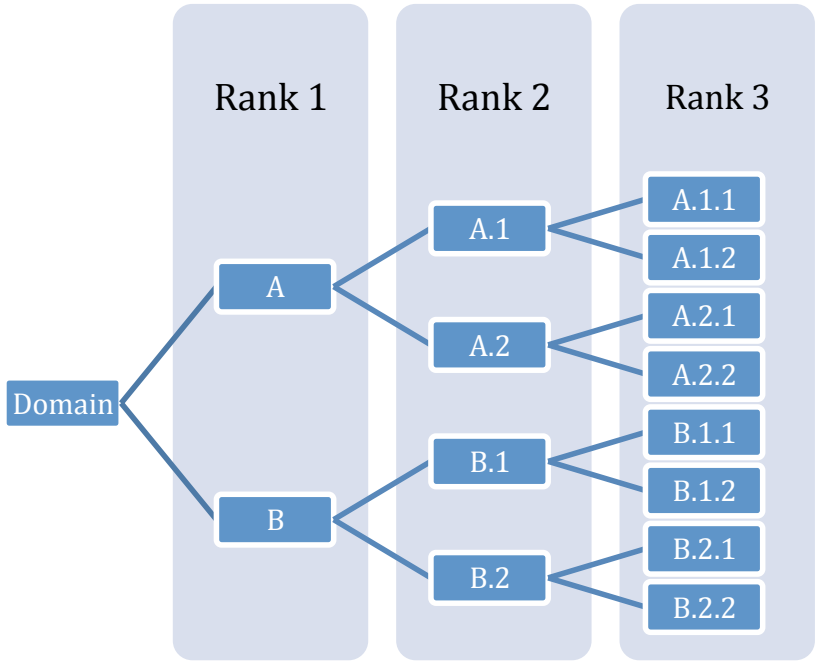

Рис.: Наглядный пример определения таксономического ранга

Таксономический ранг

Таксономический ранг - уровень результатов в иерархической структуре предметной области.

Цели и гипотезы эмпирического исследования.

Цели эмпирического исследования

- Проверить возможность оценки вклада ученого путем отображения на таксономию предметной области
- Соотнести таксномический ранг ученого и стратификации по численным критериям цитирования и заслуг

Гипотезы эмпирического исследования

- Ученые высокго ранга могут иметь высокие значения по одним показателям, компенсирующие низкие по другим для ученых одного ранга. Уровень вклада ученого может быть выражен интегральным значением путем свертки критериев с некоторыми весами
- Таксономический ранг ученого может быть определен через линейную стратификацию по числовым показтелям

Используемые данные. Характеристики выборки.

Множество ученых

30 высоко цитируемых ученых в области анализа данных и машинного обучения из Европы, Индии, Китая, России и США.
Критерии отбора:

- Наличие профиля на Google Scholar
- Наличие резюме в открытом доступе с информацией об организованных конференциях, подготовленных PhD студентах и участиях в редколлегиях журналов

Используемая таксономия

- Модифицированная классификация компьютерных наук Ассоциации вычислительных машин, версия 2012 г.

Результаты анализа данных и стратификации 1

Таблица: Веса отдельных критериев полученные при стратификаци, и веса полученные при стратификации по аггрегированным критериям.

Citation		Merit	
Citations	0.5	PhDs	0.22
I10	0.5	Conf	0.10
Hirsch	0.0	Editorial	0.69

Таблица: Таблицы сопряженности стратификаций по показателям заслуг и цитирования в сравнении с таксономическим рангом

		Citation			Merits		
	S#	S_1	S_2	S_3	S_1	S_2	S_3
Tax	S_1	0	5	6	1	1	9
	S_2	2	1	3	0	4	2
	S_3	2	5	6	1	2	10

Результаты анализа данных и стратификации 2

Таблица: Парные кореляции между значениями критериев и стратификациями

Pearson				Spearman			
Criterion	Tr	Cr	Mr	Stratification	Tr	Cs	Ms
Tr	-	-0.12	-0.04	Ts	-	-0.12	-0.02
Cr	-	-	0.31	Cs	-	-	0.25
Mr	-	-	-	Ms	-	-	-

Вывод

Все три способа ранжирования взаимно некореллированы. Это означает, что для всесторонней оценки вклада ученого необходимо учитывать по меньшей мере все три рассмотренных аспекта.

Полученные результаты

Полученные результаты

- Предложен алгоритм формирования линейной стратификации
- Алгоритм верифицирован на синтетических данных и показал свое преимущество в большинстве рассмотренных случаях
- Алгоритм апробован на реальных данных в задаче оценки вклада ученого

Дальнейшая работа.

Дальнейшая работа

- Экспериментальное изучение метода стратификации на реальных данных. Поиск приложений метода.
- Улучшение метода стратификации, включая проблему выбора числа страт, инициализации параметров страт, интрепретацию страт и весов
- Теоретическое обоснование корректности и сходимости алгоритма

Конец.