Метод формирования многокритериальной стратификации и его верификация

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

НИУ-ВШЭ

ormian@mail.ru

9 октября 2016 г.

Содержание

Актуальность работы

Методы стратификации могут служить полезным инструментом для анализа и принятия решений, когда необходимо распределить элементы по группам в соответствии со степенью успешности достижения нескольких, не обязательно однонаправленных, критериев. Примеры применения:

- разбиение фирм по уровню риска банкротства или стран по уровню кредитного риска [De Smet, Montano, Guzman 2004]
- многокритериальная ABC классификация для управления материально техническими запасами [Ng 2007; Ramanathan 2006], [Белов, Коричнева 2012]
- ранжирование и информационный поиск [Skoutas et. al. 2010]

Степень разработанности темы

- Методы многокритериального ранжирования. Обзор в [Алескеров, Хабина, Шварц 2006]
- Подходы на основе свертки с весами: многомерная ABC-классификация [Ng 2007, Ramanathan 2006], authority ranking[Sun 2009], метод главных компонент
- Подходы использующие предпочтения ЛПР: orderd k-means [De Smet, Montano, Guzman 2004], [Nemery, De Smet 2005]

Вывод

Отсутствует такая формулировка проблемы многокритериальной стратификации, которая позволяла бы автоматически получать одновременно ранжирование и разбиение исходя из "геометрической" структуры данных

Объект и предмет исследования

Объект исследования

Многокритериальное ранжирование

Предмет исследования

Разработка модели, методов и программного обеспечения для решения задачи формирования линейной стратификации

Цель и задачи исследования. Методы исследования

Цель исследования

Разработка модели, методов и программных средств для агрегирования заданных критериев в выпуклую комбинацию, а объектов – в заданное число страт таким образом, чтобы объекты каждой страты лежали в одной гиперплоскости

Задачи исследования

- Разработать «геометрический» критерий стратификации и провести анализ его сходства и различия с другими критериями агрегирования данных
- Разработать эффективные методы оптимизации сформулированного критерия
- Разработать комплекс программ для решения задачи стратификации и генерации синтетиических данных для проведения вычислительных экспериментов
- Провести апробацию разработанных методов на реальных данных

Основные результаты исследования

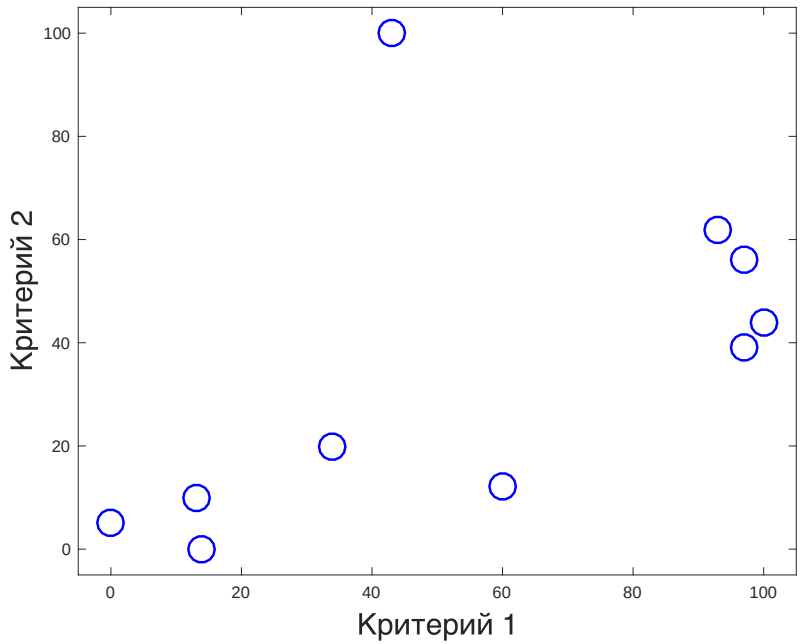
- Предложена оптимизационная модель многокритериальной линейной стратификации
- Разработан алгоритм решения оптимизационной задачи линейной стратификации
- Разработан параметрический алгоритм генерации линейных стратифицированных данных
- Разработан комплекс программ для численного решения задачи стратификации, генерации синтетиических данных и проведения вычислительных экспериментов
- Алгоритм и программное обеспечение верифицированы на синтетических и реальных данных
- Принято участие в разработке методики оценки уровня научных результатов с использованием таксономии предметной области; проведены экспериментальные расчеты
- Алгоритм и программное обеспечение, примененные для оценки разных аспектов научного вклада на выборке ведущих специалистов в области машинного обучения и анализа данных, привели к согласованным результатам.

Наглядное представление

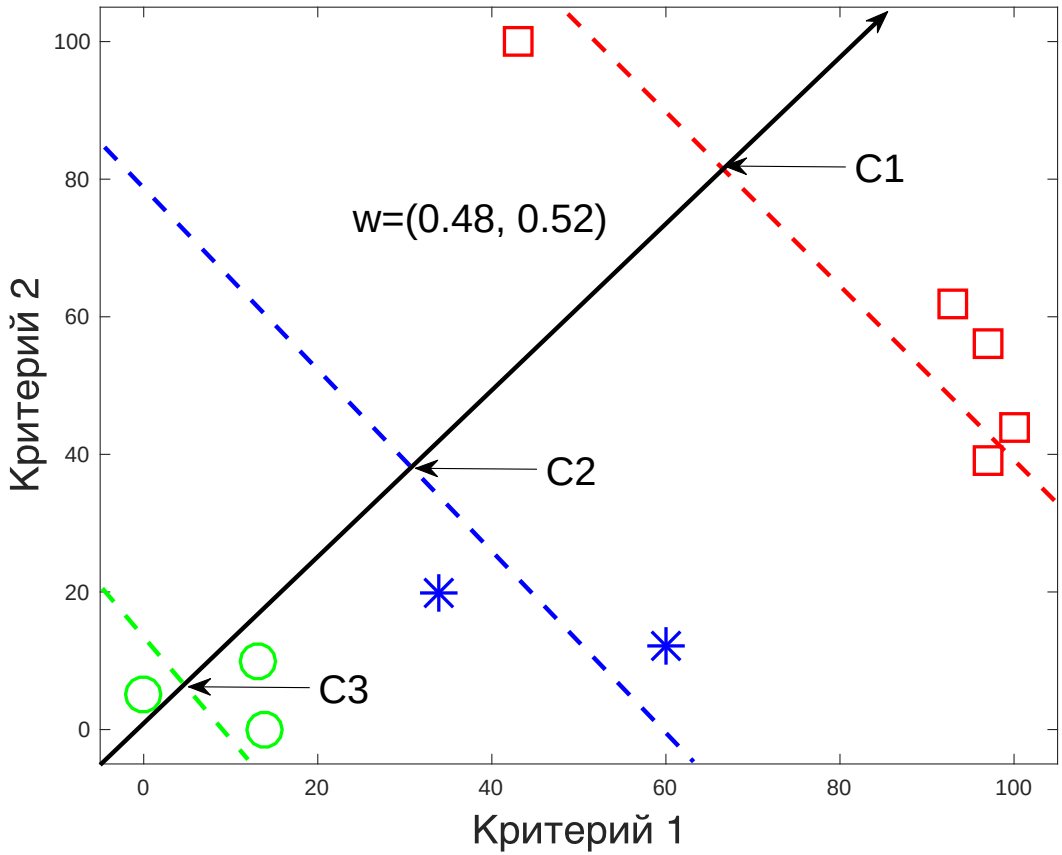
Задано желаемое число страт. Например 3.

Найти:

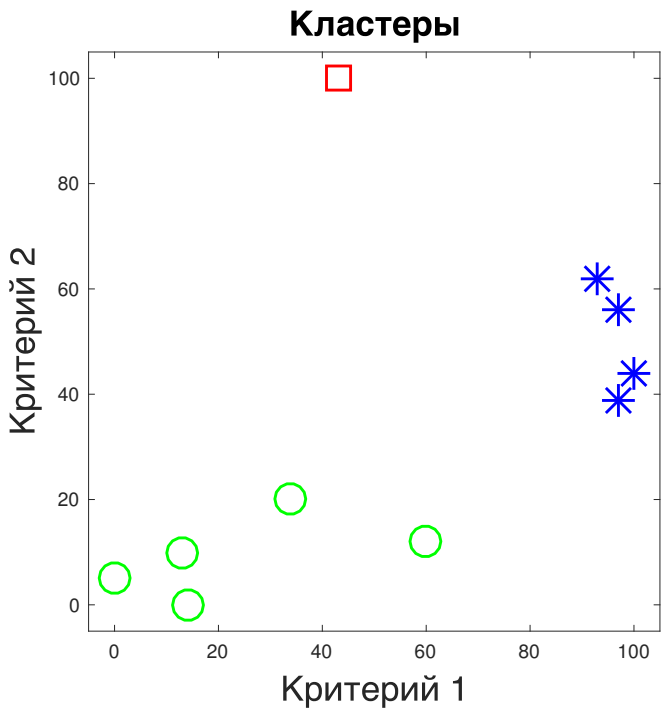
- Линейные страты: т.е. выпуклую комбинацию криериев
- Направление страт: по возможности, объекты лежат в гиперплоскости страт

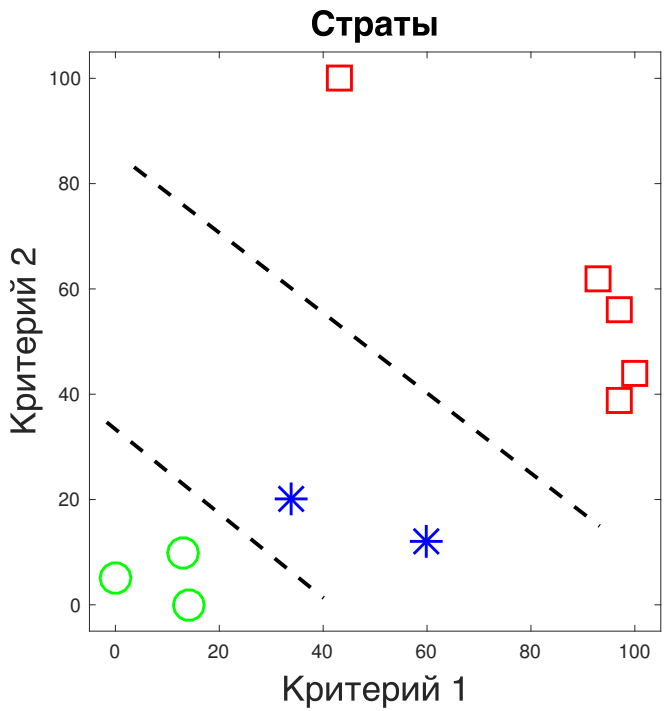


Наглядное представление



Различие страт и кластеров





Формулировка задачи многокритериальной стратификации

Формулировка задачи

Дано N объектов, оцененных по M критериям x_{ij}, $i = 1..N, j = 1..M$. Найти такие веса w_j, центры $c_k, k = 1..K$ и разбиение $S = \{S_1, S_2, ..., S_k\}$, чтобы значения взвешенного критерия $f_i = \sum_{j=1}^{M} x_{ij} w_j$ для объектов каждой страты ($i \in S_k$) были как можно ближе к значению центра страты c_k. То есть ошибка e_i для $f_i = c_k + e_i, i \in S_k$ была бы минимальна.

Оптимизационная задача

Оптимизационная задача линейной стратификации. ЛИНСТРАТ.

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

Алгоритм решения оптимизационной задачи

1. Инициализировать случайно w и c
2. При заданных w и c найти разбиение $x_i \in S_k$, $k = argmin_k(\sum_{j=1}^{M} x_{ij} w_j - c_k)^2$
3. При заданных w и S найти центры $c_k = \frac{1}{|S_k|} \sum_{x_i \in S_k}^{N} \sum_{j=1}^{M} x_{ij} w_j$
4. При заданных S и c найти w

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

5. Остановить итерации, когда разница значений целевой функции для последовательных итерации не станет меньше заданного значения

Пример. Увеличение значений признака может привести к уменьшению его веса.

object	x_1	x_2	$f_1 = 0.09x_1 + 0.91x_2$	stratum
1	0	1	0.91	3
2	0	2	1.82	2
3	0	3	2.73	1
4	10	0	0.90	3
5	20	0	1.82	2
6	30	0	2.73	1

object	x_1	x_2	$f_2 = 0.90x_1 + 0.10x_2$	stratum
1	0	80	7.79	3
2	0	90	8.77	3
3	0	100	9.74	3
4	10	0	9.03	3
5	20	0	18.05	2
6	30	0	27.08	1

Пример. Отличие ЛИНСТРАТ от метода главных компонент (PCA using SVD)

LS: $w = (0.33, 0.67)$
ошибка 0.0%

PCA: $w = (0.77, 0.23)$
ошибка 13.4%

Label	x	y	LS	PCA
C1	2	0	0.67	1.54
C2	0	1	0.67	0.23
B1	6	0	2.00	4.63
B2	5	0.5	2.00	3.97
B3	3	1.5	2.00	2.66
B4	1	2.5	2.00	1.34
A1	4	2	2.67	3.54
A2	2	3	2.67	2.23

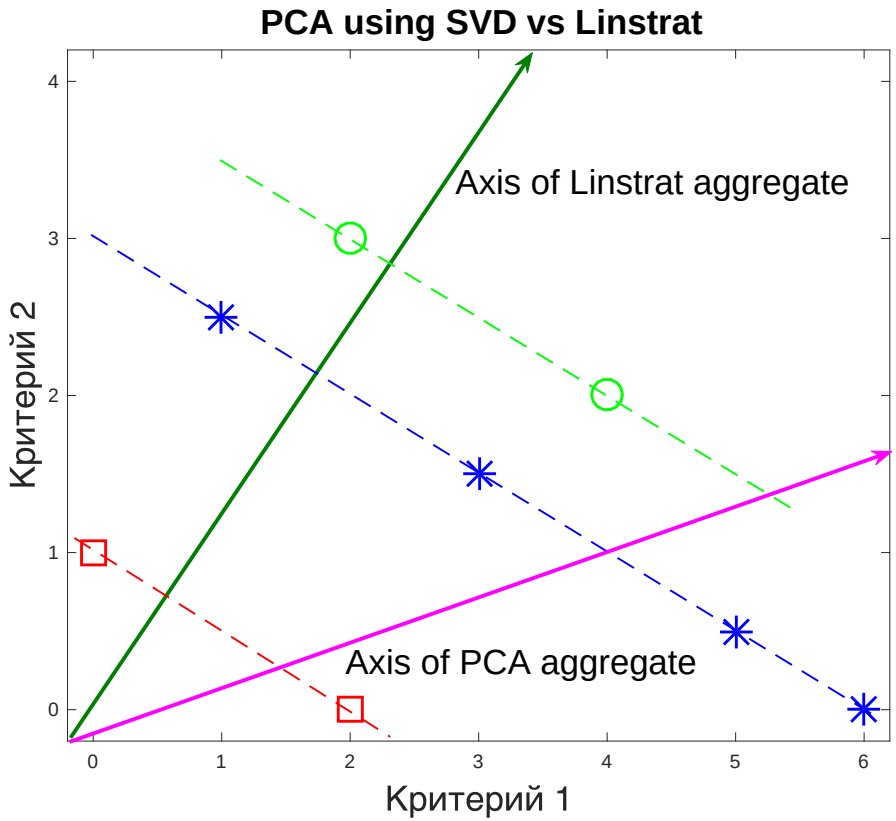


Рис.: Различие PCA и Linstrat

Верификация метода ЛИНСТРАТ

Данные:

- Библиометрические показатели журналов и стран
- Синтетические страты
- Числовые показатели оценки научного вклада ученого

Верификация метода ЛИНСТРАТ

Методы для сравнения:

- Linstrat QP [Орлов 2014]
- Linstrat Evolutionary [Миркин, Орлов 2013]
- Borda count [Алескеров, Хабина, Шварц 2006]
- Linear Weights Optimization [Ramanathan 2006]
- Authority ranking [Sun et. al 2009]
- Pareto bound stratification [Миркин, Орлов 2013]

Метрики для сравнения и нормировки

Метрики сравнения алгоритмов:

- Точность стратификации. Доля объектов с правильно определенной стратой
- Среднее расстояние Кемени-Снелла между многокритерильной стратификацией и стратификацией по каждому критерию

Нормировки данных:

- Нормировка анализа данных $z_{ij} = \frac{x_{ij} - min(x_{.j})}{max(x_{.j}) - min(x_{.j})}$
- Статистическая нормировка $z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$

Верификация в задаче ранжирования научных журналов и стран по библиометрическим показателям.

Библиометрические показатели 118 научных журналов из раздела Artificial Intelligence за 2012 год. Источник scimagojr.com.

- Индекс SJR (Scientific Journal Ranking) [Gonzalez Pereira et al 2010].
- Индекс Хирша (H) [Hirsch 2005]
- Импакт-фактор журнала (I)

Данные о публикационной активности 102 стран за 2012 год в разделе Artificial intelligence

- Общее число документов опубликованных за 2012 (D);
- Число цитируемых документов, опубликованных в 2012 году (CD);
- Общее количество цитирований в 2012 году (C).
- Самоцитирование документов в 2012 году (country self-citations) (SC);
- Среднее число цитирований в 2012 году документов (CPD);
- H-индекс на уровне страны (H).

Результаты

Алгоритм ЛИНСТРАТ дал наиболее согласованное ранжирование

Таблица: Средние значения расстояний Кемени-Снелла для журналов

Нормировка	LSQ	LS	BC	LWO	AR	PS
Статистическая	**0.12**	**0.12**	0.17	0.13	**0.12**	0.23
Анализа данных	**0.12**	**0.12**	0.17	0.15	0.20	0.23

Таблица: Средние значения расстояний Кемени-Снелла для стран

Нормировка	LSQ	LS	BC	LWO	AR	PS
Статистическая	**0.10**	**0.10**	0.26	0.21	0.16	0.18
Анализа данных	**0.10**	**0.10**	0.26	0.17	0.12	0.18

Верификация алгоритма на синтетических данных

Параметры генерации синтетических данных:

- ориентация страт (а) w=(0.5, 0.5), (б) w=(0.8, 0.2), (в) w=(0.2, 0.8)
- толщина страт (г) σ=0.05, (в) σ=0.1, (г) σ=0.2
- интенсивность страт (ж) θ=(0.5, 0.3, 0.2), (з) θ=(0.7, 0.2, 0.1), (и) θ=(0.8, 0.15, 0.05)
- размах страт (к) φ=0.05, (л) φ=0.1, (м) φ=0.5
- размерность данных m
- размер выборки n

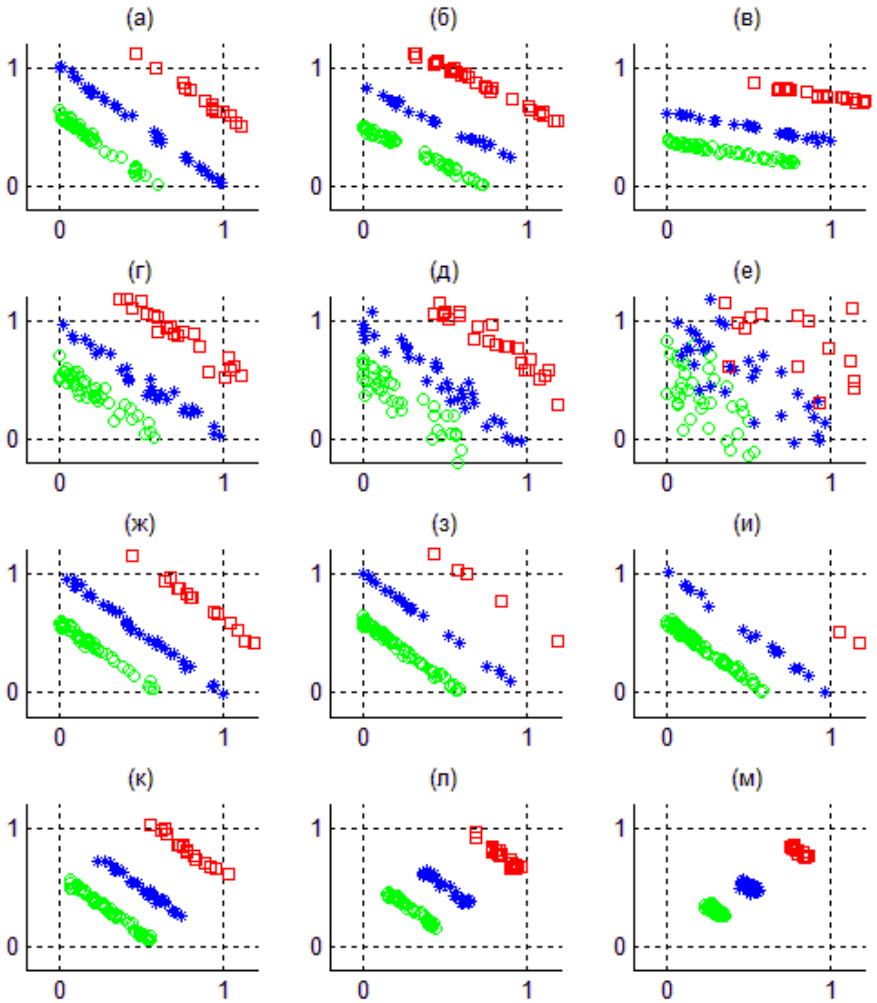


Рис.: Синтетические данные

Результаты

Линстрат:

- Превосходит по воспроизводству сгенерированных страт в большинстве случаев
- Более устойчив к изменениям размерности, размаха φи интенсивности θ
- При сильном увеличении толщины σуступает LWO

Оценка научного вклада: три аспекта

Цитирование

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

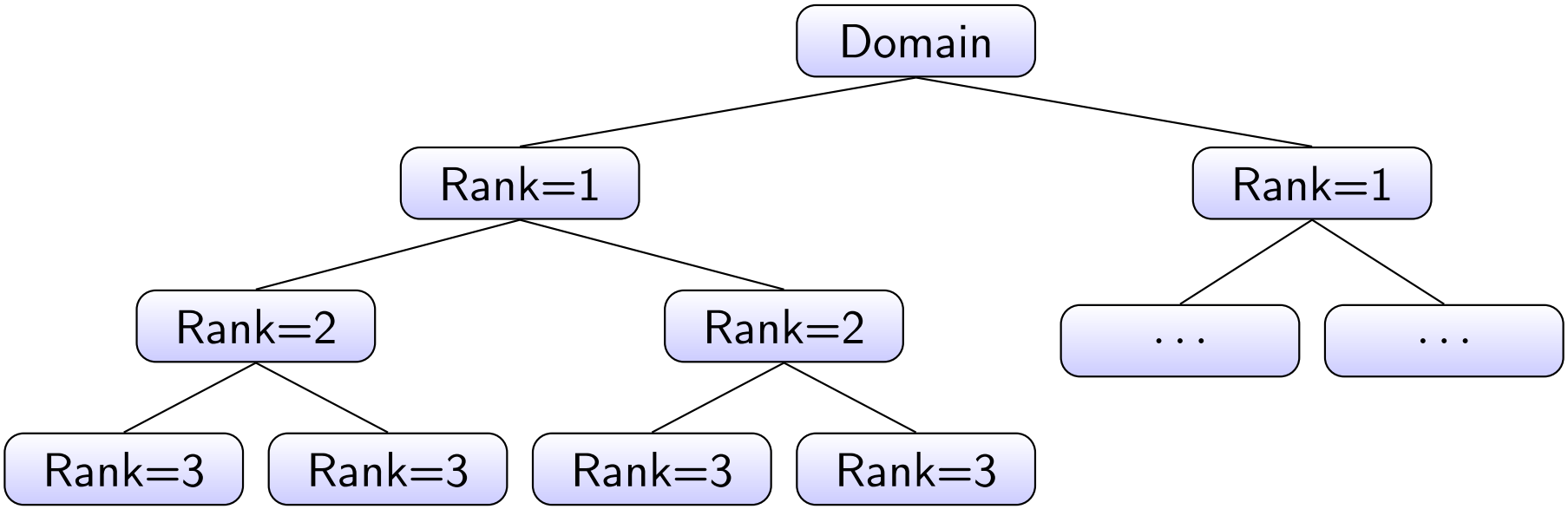
Заслуги

- Число организованных 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 таксономии предметной области [Миркин 2013]

Таксономический ранг



Таксономический ранг

Рассматриваются результаты, которые создали новый элемент таксономии или существенно повлияли на существующий.

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. Число уровней на которые отображаются результаты задает число страт.

Порядок эмпирического исследования.

1. Определить предметную область
2. Построить таксономию предметной области
3. Собрать репрезентативную выборку ученых с известными результатами
4. Отобразить результаты каждого ученого на таксономию
5. Вычислить таксономический ранг каждого ученого
6. Сформировать стратификацию по уровню цитирования и стратификацию по уровню заслуг
7. Сравнить полученные стратификации с таксономическим рангом

Используемые данные. Характеристики выборки.

Выборка 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.

Используемые данные. Таксономия анализа данных (всего 5-6 уровней)

Таблица: ACM CCS 2012 верхние уровни таксономии

Subject index	Subject name
1.	Theory of computation
1.1.	Theory and algorithms for application domains
2.	Mathematics of computing
2.1.	Probability and statistics
3.	Information systems
3.1.	Data management systems
3.2.	Information systems applications
3.3.	World Wide Web
3.4.	Information retrieval
4.	Human-centered computing
4.1.	Visualization
5.	Computing methodologies
5.1.	Artificial intelligence
5.2.	Machine learning

Результаты анализа данных и стратификации 1

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

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

Таблица: Сравнение полученных стратификаций

		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

Уровень результатов не коррелирован с заслугами и цитированием. Положительная корреляция между заслугами и цитированием: популярность. Для всесторонней оценки научного вклада следует учитывать все три аспекта

Дальнейшая работа.

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

Основные результаты исследования

- Предложена оптимизационная модель многокритериальной линейной стратификации
- Разработан алгоритм решения оптимизационной задачи линейной стратификации
- Разработан параметрический алгоритм генерации линейных стратифицированных данных
- Разработан комплекс программ для численного решения задачи стратификации, генерации синтетиических данных и проведения вычислительных экспериментов
- Алгоритм и программное обеспечение верифицированы на синтетических и реальных данных
- Принято участие в разработке методики оценки уровня научных результатов с использованием таксономии предметной области; проведены экспериментальные расчеты
- Алгоритм и программное обеспечение, примененные для оценки разных аспектов научного вклада на выборке ведущих специалистов в области машинного обучения и анализа данных, привели к согласованным результатам.

Публикации по теме диссертации в изданиях из списка ВАК

Orlov M., Mirkin B. (2014) A concept of multicriteria stratification: A definition and solution *Procedia Computer Science, 31, 273–280.*

Орлов М. (2014) Алгоритм формирования многокритериальной стратификации *Бизнес-информатика, № 4 (30), 24–35*

Mirkin B., Orlov M. (2015) Three aspects of the research impact by a scientist: measurement methods and an empirical evaluation *A. Migdalas, A. Karakitsiou, Eds., Optimization, Control, and Applications in the Information Age, Springer Proceedings in Mathematics and Statistics. 130. 233–260.*

Спасибо за внимание.