Как улучшить оценку множеств признаков с помощью принципа минимальной длины описания

Tatiana Makhalova, Sergei Kuznetsov, Amedeo Napoli

УДК 519.716.5

КАК УЛУЧШИТЬ ОЦЕНКУ МНОЖЕСТВ ПРИЗНАКОВ С ПОМОЩЬЮ ПРИНЦИПА МИНИМАЛЬНОЙ ДЛИНЫ ОПИСАНИЯ?

С.О. Кузнецов (*skuznetsov@hse.ru*)
Национальный исследовательский университет
Высшая школа экономики, Москва

Т.П. Махалова (*tpmakhalova@hse.ru*)
Национальный исследовательский университет
Высшая школа экономики, Москва
LORIA (CNRS – Inria – Университет Лотарингии),
Вандевр-ле-Нанси, Франция

А. Наполи (*amedeo.napoli@loria.fr*)
LORIA (CNRS – Inria – Университет Лотарингии),
Вандевр-ле-Нанси, Франция

Подходы к извлечению знаний, основанные на анализе формальных понятий, позволяют всесторонне исследовать имеющуюся информацию посредством обхода создаваемой решетки формальных понятий. Решетки могут содержать экспоненциальное число элементов. Одним из наиболее распространенных подходов к уменьшению их числа является отбор на основе индексов. В данной работе мы предлагаем использовать принцип минимальной длины описания и демонстрируем, как его применение может улучшить результаты отбора, основанного на индексах. Предлагаемый подход может использоваться не только для замкнутых, но и для произвольных множеств признаков.

Ключевые слова: формальные понятия, минимальная длина описания, отбор множеств признаков

Введение

Анализ формальных понятий (АФП) играет важную роль в майнинге данных и машинном обучении. На практике при применении АФП

сталкиваются с проблемой экспоненциальности числа формальных понятий.

В настоящее время данная проблема решается посредством пред- или постобработки, вычислением приближенных понятий (см. обзор [Kuznetsov et al., 2018]). Целью перечисленных методов является получение небольшого набора интересных, значимых понятий. Данное множество должно также обладать минимальной избыточностью [Buzmakov et al., 2015]. В данной статье мы исследуем такие множества понятий.

Наряду с индексами оценки понятий мы применяем принцип минимальной длины описания (МДО) [Grünwald, 2007], который позволяет отобрать небольшой набор разнообразных и хорошо интерпретируемых понятий. Обеспечивая сжатие данных, принцип МДО позволяет сохранить баланс между качеством «подстройки» под данные и сложностью модели и не требует предопределения дополнительных параметров [Aggarwal et al., 2014].

Как и методы, основанные на снижении размерности [Айвазян и др., 1989; Терехина, 1986; Глотов, 1984; Подиновский, 2007], предлагаемый подход позволяет получить описание объектов в пространстве меньшей размерности, однако обладает принципиальным отличием. Метод снижения размерности позволяет представить все объекты в одном, отличном от исходного, признаковом пространстве, тогда как предлагаемый подход можно рассматривать как метод отбора признаков в отдельных группах объектов, когда каждая из них представляется в разных подпространствах исходного признакового пространства.

Мы также изучаем вопрос о том, как представления об интересности отбираемых понятий могут быть использованы в рамках МДО. Мы предлагаем использовать индексы как априорные представления об интересности и применять МДО для улучшения «качества» отбираемых в соответствии с индексами понятий. Предлагаемый подход может быть рассмотрен как отбор множеств признаков (паттернов) на основе выбранных индексов оценки интересности с дальнейшей постобработкой, так и как метод минимизации длины описания с дополнительными ограничениями, где ограничениями являются индексы оценки интересности понятий. Стоит отметить, что предлагаемый подход может быть применен не только к понятиям и их содержанию, т.е. замкнутым множествам признаков, но и к произвольным множествам признаков.

1. Анализ формальных понятий. Основные понятия

Здесь мы приводим основные понятия теории анализа формальных понятий [Ganter et al., 1999]. Пусть заданы множество объектов G, множество признаков M и бинарное отношение между ними $I \subseteq G \times M$,

тогда формальным контекстом называется тройка (G, M, I). На множестве объектов и множестве признаков задана операция $(\cdot)'$. Для произвольных подмножеств $A \subseteq G$ и $B \subseteq M$ она принимает следующий вид:

$$A' = \{m \in M \mid \forall g \in A : gIm\},\ B' = \{g \in G \mid \forall m \in B : gIm\}.$$

A' представляет собой множество признаков, общих для всех объектов множества A, B' – множество объектов, обладающих всеми признаками из B. B также называют паттерном. Оператором замыкания $(\cdot)''$ на множестве G называется отображение $\varphi: X \to X$, где $X \subseteq G$. Данное отображение обладает следующими свойствами: идемпотентность ($\varphi\varphi X = \varphi X$), экстенсивность ($X \subseteq \varphi X$) и монотонность (из $X \subseteq Y$ следует $\varphi X \subseteq \varphi Y$).

Формальным понятием называется пара (A,B), где $A \subseteq G$, $B \subseteq M$ и $A'=B$, $B'=A$. A и B называют объемом и содержанием формального понятия, соответственно.

Пример. Рассмотрим формальный контекст, представленный в Таблице 1. Множество формальных понятий составляют 8 элементов. Например, понятие $(\{g_1, g_2, g_3\}, \{m_1, m_3\})$ соответствует животным.

Табл. 1. m_1: 4 ноги (лапы), m_2: шерсть, m_3: меняют размер, m_4: устойчивы к холоду, m_5: черно-белый, m_6: не выделяют CO_2, m_7: желто-коричневый, m_8: зеленый, m_9: серый

Объекты		m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9
g_1	собака	×	×	×			×			
g_2	кошка	×	×	×				×		
g_3	лягушка	×		×					×	
g_4	машина			×	×					×

2. Принцип минимальной длины описания в задаче отбора множеств признаков (паттернов)

Принцип минимальной длины описания в контексте отбора паттернов формулируется следующим образом: лучшими паттернами являются такие, которые обеспечивают максимальное сжатие данных [Luc, 2008].

Основу данного подхода составляет кодовая таблица, в которую записаны некоторые паттерны и длины их кодов в битах. Лучшая кодовая таблица обеспечивает максимальное сжатие минимальным набором паттернов, т.е. минимизирует длину $L(CT, D) = L(D \mid CT) + L(D \mid CT)$, где $L(D \mid CT)$ – длина набора данных D закодированных с помощью кодовой таблицы CT, $L(CT \mid D)$ – длина кодовой таблицы CT, вычисленной на основе набора данных D. Для кодирования объекта нужно выбрать набор непересекающихся паттернов, которые полностью покрывают множество признаков этого объекта. Мы обозначим как $u(B) = |\{t \in D \mid B \in \text{cover}(t)\}|$

количество использований паттерна B для покрытия объектов в D, $U = \sum_{B \in CT} u(B)$ – общее количество использований всех паттернов кодовой таблицы CT для покрытия набора данных D. Принципы построения таблицы будут рассмотрены ниже.

Для определения длины кодирования паттернов используется оптимальное префиксное кодирование Шеннона, т.е. $l(B) = -\log \Pr(B)$, где вероятность вычисляется следующим образом: $\Pr(B) = u(B)/U$. Таким образом, паттерны с высокой частотой использования имеют меньшую длину. Мы оставляем детали реализации схем кодирования за пределами данной статьи. Поскольку в данной работе мы не преследуем цель точного кодирования, далее будет рассматриваться упрощенная версия $L(CT, D)$, в которой учитываются исключительно слагаемые, характеризующие «специфичность» паттернов:

$$L(D \mid CT) = \sum_{g \in D} \sum_{B \in \text{cover(g)}} l(B) = -\sum_{B \in CT} u(B) \log \frac{u(B)}{U},$$

$$L(CT \mid D) = \sum_{B \in CT} \text{code}(B) + l(B),$$

где $\text{code}(B)$ – количество бит, необходимое для хранения множества признаков B.

На начальном этапе кодовая таблица состоит из одноэлементных паттернов, т.е. $\{\{m\} \mid m \in M\}$. Также имеется набор кандидатов – упорядоченное в соответствии выбранным критерием множество паттернов, которые могут быть добавлены в кодовую таблицу. Минимизация общей длины заключается в последовательном добавлении лучшего (по выбранному критерию) паттерна, перевычислении покрытия с обновленным набором признаков и вычислении новой длины. Если эта длина оказывается меньше предыдущей, то данный паттерн добавляется в кодовую таблицу. Процесс продолжается до тех пор, пока все паттерны из набора кандидатов не будут просмотрены.

Набор кандидатов может быть составлен из самых разнообразных паттернов (множеств признаков): произвольных, замкнутых, δ-множеств и др. В качестве критерия, в соответствии с которым упорядочиваются множества признаков, может быть выбран любой индекс, в частности, в алгоритме Krimp [Vreeken et al., 2011] паттерны упорядочиваются по размеру, а затем по частоте их встречаемости в наборе данных.

Поскольку проблема выбора оптимального набора паттернов подразумевает полный перебор всех возможных комбинаций, на практике применяют различные эвристики [Гладун, 1977; Ройзензон, 2005].

Пример. Рассмотрим, как вычисляется кодовая таблица, используя формальный контекст, представленный в Таблице 1. В качестве

кандидатов выберем все формальные понятия (*A*,*B*), упорядоченные по площади, т.е. $|A|\times|B|$. В Таблице 2 представлен пошаговый процесс заполнения кодовой таблицы. ЧИ – частота использования паттерна из кодовой таблицы. На начальном этапе таблица состоит из одноэлементных паттернов m_i. На первом шаге добавляется первый кандидат, перевычисляется покрытие. Поскольку длина *L(D, CT)* для новой таблицы меньше предыдущей, кандидат $m_1m_2m_3$ включается в кодовую таблицу. Покрытие этим множеством фиксируется и больше не изменяется (см. колонку «Данные с покрытием»). Процесс завершается, когда все кандидаты будут просмотрены.

Табл. 2

	Кодовая таблица (КТ)		Данные и их покрытием элементами КТ	Набор кандидатов, площадь
	МП	ЧИ		
Шаг 0	m_3 m_1 m_2 m_4 m_6-m_9 m_5	4 3 2 1 1 0	$(m_1)(m_2)(m_3)(m_6)$ $(m_1)(m_2)(m_3)(m_7)$ $(m_1)(m_3)(m_8)$ $(m_3)(m_4)(m_9)$	$m_1m_2m_3$, 6 m_1m_3, 6 $m_1m_2m_3m_6$, 4 $m_1m_2m_3m_7$, 4 $m_1m_3m_8$, 3 $m_3m_4m_9$, 3
Шаг 1	$m_1m_2m_3$ m_3 m_1, m_4, m_6-m_9 m_2, m_5	2 2 1 0	$(m_1m_2m_3)(m_6)$ $(m_1m_2m_3)(m_7)$ $(m_1)(m_3)(m_8)$ $(m_3)(m_4)(m_9)$	$m_1m_3m_8$, 3 $m_3m_4m_9$, 3 m_1m_3, 2
Шаг 2	$m_1m_2m_3$ $m_1m_3m_8$ m_3, m_4, m_6, m_7, m_9 m_1, m_2, m_5, m_8	2 1 1 0	$(m_1m_2m_3)(m_6)$ $(m_1m_2m_3)(m_7)$ $(m_1m_3m_8)$ $(m_3)(m_4)(m_9)$	$m_3m_4m_9$, 3
Шаг 3	$m_1m_2m_3$ $m_1m_3m_8$ $m_3m_4m_9$ m_6, m_7 m_1 - m_5, m_8-m_9	2 1 1 1 0	$(m_1m_2m_3)(m_6)$ $(m_1m_2m_3)(m_7)$ $(m_1m_3m_8)$ $(m_3m_4m_9)$	

3. Принцип минимального описания в задаче отбора формальных понятий (замкнутых множеств признаков)

Применение принципа МДО позволяет получить ряд преимуществ по сравнению с отбором лучших *n* понятий по выбранным индексам. К основным преимуществам относятся следующие: не требуется явно фиксировать, сколько элементов должно быть выбрано; отобранное

множество содержит разнообразные (непохожие попарно) понятия; множество отобранных понятий небольшое по размеру, при этом число элементов не растет линейно с увеличением числа понятий в исходном множестве.

Для демонстрации указанных преимуществ была проведена серия экспериментов. Мы использовали данные репозитория LUCS-KDD [Coenen, 2003]. В качестве мер интересности понятия (A,B) рассматривались следующие индексы: мощность содержания $\text{len}(B) = |B|$, мощность объема $\text{fr}(B) = |A|$, лифт $\text{lift}(B) = \Pr(B) / \prod_{b \in B} \Pr(\text{b})$, где $\Pr(\cdot)$ – относительная частота множества признаков в наборе данных, отделимость $\text{sep}(A,B) = (|A||B|) / \left(\sum_{g \in A} |g'| + \sum_{m \in B} |m'| - |A||B| \right)$. Отметим, что отбор понятий в нашем случае основан на оценке их содержаний и рассматриваемая задача может быть отнесена к более широкому классу: отбор произвольных множеств признаков.

Для вычисления покрытия мы использовали жадную стратегию (см. алгоритм Krimp [Vreeken et al., 2011]). При этом наборы кандидатов упорядочивались в соответствии со следующими индексами: $\text{area\_lf}(B) = \text{len}(B) \times \text{fr}(B)$, $\text{area\_ll}(B) = \text{len}(B) \times \text{lift}(B)$, $\text{area\_ls}(B) = \text{len}(B) \times \text{sep}(B)$, и последовательным упорядочиванием по парам индексов: $\text{len}(B)$ и $\text{fr}(B)$, $\text{len}(B)$ и $\text{lift}(B)$, $\text{len}(B)$ и $\text{sep}(B)$.

Кроме того, мы отобрали лучшие n понятий (исключая те из них, содержание которых состоит из единственного признака) по каждому из индексов, число n определялось размером соответствующего множества МДО-оптимальных понятий. Подобный подход позволяет проанализировать преимущества принципа МДО в отборе по заданным индексам интересности.

Табл. 3

	$\|G\|$	$\|M\|$	Кол-во замкнутых понятий	area len fr	area len lift	area len	len fr	len lift	len sep
breast	699	16	702	32	20	26	37	37	37
ecoli	336	29	690	56	16	25	64	66	64
iris	150	19	183	29	13	14	35	34	35
led7	3 200	24	3808	118	64	98	109	109	108
pima	768	38	2769	106	36	65	121	112	121

В Таблице 3 показаны среднее количество понятий, отобранных по выбранным индексам с применением МДО-подхода для пяти наборам

данных. Подход на основе МДО позволяет отобрать 1.3-19.1% понятий (в среднем 6.4%).

Существенным преимуществом данного подхода является отбор понятий с разнообразными и «типичными» (характеризующими набор данных в целом, а не отдельные его объекты) содержаниями. Для демонстрации данного факта для каждого множества МДО-оптимальных понятий мощности *n* (обозначаемого далее S_{MDL}) мы рассмотрели множество первых *n* упорядоченных понятий (обозначаемого далее S_{top}) по выбранным индексам и сравнили их содержания.

В данном разделе мы приведем результаты по следующим характеристикам: среднее число понятий, имеющих в этом же множестве более общее понятие (т.е. содержание которых покрывает какое-либо другое содержание) и доля признаков в наборе данных, покрываемых множествами содержаний отобранных понятий. Чем меньше первый показатель, тем более разнообразны содержания отобранных понятий, чем больше второй показатель, тем более содержательными они являются.

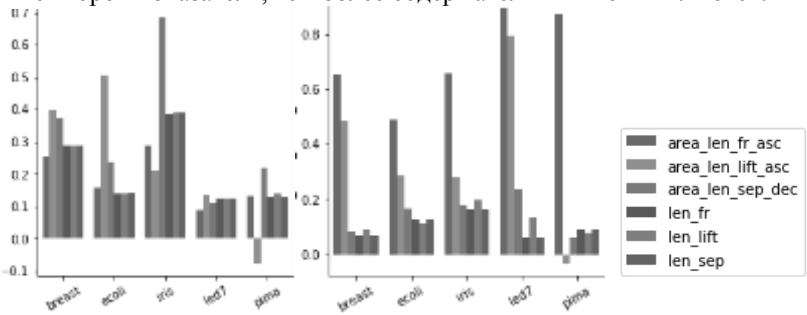

Рис. 2 Средняя разница в числе более общих содержаний понятий между S_{top} и S_{MDL} (слева) и средняя разница в доле покрытых признаков в наборе данных между S_{MDL} и S_{top} (справа). Положительные значения указывают на то, что S_{MDL} демонстрирует лучшие результаты, чем S_{top}. Чем больше значения, тем больше преимущества S_{MDL}.

Представленные результаты показывают, что S_{MDL} в среднем позволяет получить на 20% больше разнообразных содержаний (не имеющих более общих в том же множестве), при этом наиболее разнообразными оказываются содержания, отобранные по area_ls (они в среднем на 32% более разнообразнее, чем в S_{top}).

S_{MDL} покрывают в среднем на 21.5% больше признаков в исходном множестве, при этом наибольшая разница с S_{top} в среднем составляет 71.4% для area_lf, что означает следующее: для индекса area_lf с использованием МДО-подхода удается получить *n* понятий, содержания которых покрывают в среднем на 71.4% больше признаков в наборе

данных, чем содержания лучших n понятий, выбранных по этой же метрике.

Заключение

В данной статье был предложен метод отбора множеств признаков (содержаний формальных понятий), основанный на принципе минимальной длины описания. Предложенный метод позволяет с легкостью использовать дополнительные экспертные знания об интересности отбираемых элементов (выраженные посредством индексов оценки множеств признаков). Проведенные эксперименты показывают, что подход на основе МДО, не требуя введения дополнительных параметров, позволяет получить небольшой набор множеств признаков, которые являются разнообразными и обладают хорошей описательной способностью.

Список литературы

[Kuznetsov et al., 2018] Kuznetsov S.O., Makhalova T. On interestingness measures of formal concepts // Information Sciences. 2018. №443-443.

[Buzmakov et al., 2015] Buzmakov A., Kuznetsov S.O., Napoli A. Fast generation of best interval patterns for nonmonotonic constraints // Joint European Conference on Machine Learning and Knowledge Discovery in Databases. 2015.

[Grünwald, 2007] The minimum description length principle // MIT press. 2007

[Aggarwal., et al., 84] Aggarwal C.C., Han J. Frequent pattern mining // Springer. 2014.

[Ganter et al., 1999] Ganter B., Wille R., Formal concept analysis: Logical foundations // 1999.

[Ganter et al., 2000] Ganter B., Kuznetsov S.O. Formalizing hypotheses with concepts // International Conference on Conceptual Structures. – Springer. 2000.

[Kuznetsov, 2004] Kuznetsov S.O. Machine learning and formal concept analysis // Concept Lattices. – Springer Berlin Heidelberg, Berlin, Heidelberg. 2004.

[Baldi et al., 2000] Baldi P., Brunak S., Chauvin Y., Andersen C.A., Nielsen H. Assessing the accuracy of prediction algorithms for classification // Bioinformatics. 2000. №16 (5).

[Vreeken et al., 2011] Vreeken J., Van Leeuwen M., Siebes A. Krimp: mining itemsets that compress //Data Mining and Knowledge Discovery. 2011. №23 (1).

[Coenen, 2003] Coenen F. The lucs-kdd discretised/normalised arm and carm data library – http://www.csc.liv.ac.uk/~frans/KDD/Software/LUCS KDD DN

[Айвазян и др., 1989] Прикладная статистика. Классификация и снижение размерности / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин; Под ред. С. А. Айвазяна. — М.: Финансы и статистика, 1989. — 607 с.

[Гладун, 1977] Гладун В. П. Эвристический поиск в сложных средах. Киев: Наукова думка, 1977.

[Глотов, Павельев, 1984] Глотов В. А., Павельев В. В. Векторная стратификация. — М.: Наука, 1984.— 94 с.

[Подиновский, 2007] Подиновский В. В. Введение в теорию важности критериев в многокритериальных задачах принятия решений. — М.: Физматлит, 2007. — 64 с.

[Ройзензон, 2005] Ройзензон Г. В. Способы снижения размерности признакового пространства для описания сложных систем в задачах принятия решений // Новости искусственного интеллекта. — 2005. — № 1. — С. 18–28.

[Терехина, 1986] Терехина А. Ю. Анализ данных методами многомерного шкалирования. — М.: Наука, 1986. — 168 с.

HOW TO IMPROVE ITEMSET ASSESSMENT USING MINIMUM DESCRIPTION LENGTH PRINCIPLE

S.O. Kuznetsov (*skuznetsov@hse.ru*)
National Research University Higher School of Economics, Moscow

T.P. Makhalova (*tpmakhalova@hse.ru*)
National Research University Higher School of Economics, Moscow
LORIA (CNRS – Inria – U. Of Lorraine), Vandoeuvre-les-Nancy, France

A. Napoli (*amedeo.napoli@loria.fr*)
National Research University Higher School of Economics, Moscow

LORIA (CNRS – Inria – U. Of Lorraine), Vandoeuvre-les-Nancy, France

Formal Concept Analysis plays an important role in Knowledge Extraction. It provides powerful tools for the thorough data analysis. The main drawback that hampers its practical application is the exponential number of generated concepts. The common way to tackle this issue is application of indices to them. In this paper we propose to use the Minimal Description Length principle for this purpose and discuss how it can improve the conventional index-based selection. The proposed approach can be applied to any type of itemsets, not only to the closed ones.

Keywords: formal concepts, MDL, itemset assessmet