Майнинг множеств признаков на основе сжатия: вероятностный подход

Tatiana Makhalova, Sergei O. Kuznetsov, Amedeo Napoli

УДК 519.716.5

МАЙНИНГ МНОЖЕСТВ ПРИЗНАКОВ НА ОСНОВЕ СЖАТИЯ: ВЕРОЯТНОСТНЫЙ ПОДХОД[1]

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

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

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

Майнинг паттернов (поиск множеств признаков в данных) является активно развивающимся направлением интеллектуального анализа данных. Большая часть современных подходов к поиску паттернов основана на принципе минимальной длины описания. Данный класс методов основан на оптимальном кодировании и подразумевает задание вероятностного пространства на множестве паттернов. В данной работе мы рассматриваем существующий подход к определению вероятностного пространства и предлагаем его модификацию, позволяющую уменьшить избыточность оптимального множества паттернов.

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

Введение

Майнинг паттернов (МП) занимает важное место среди методов обнаружения знаний и майнинга данных. Основной задачей МП является обнаружение небольшого набора интересных паттернов, обладающих в совокупности минимальной избыточностью, описывающих достаточно

[1] Работа выполнена при финансовой поддержке РНФ (проект № 17-11-01294).

полно анализируемые данные и предоставляющих возможность их интерпретации.

Все подходы к майнингу паттернов можно разделить на две группы: статические и динамические. Статические подходы подразумевают использование мер оценки интересности паттернов [Geng et al., 2006]. Поиск паттернов осуществляется при неизменных предположениях об интересности, например, поиск частых паттернов с частотой выше заданного порога. К основным недостаткам данного подхода относят следующие. Во-первых, вместо поиска “интересного” набора паттернов в совокупности, методы этой группы направлены на поиск отдельных интересных паттернов, не зависимо друг от друга. Такие паттерны, как правило, очень похожи и описывают избыточно анализируемые данные. Во-вторых, выбор мер интересности крайне субъективен. Зачастую, достаточно сложно обосновать выбор меры (в большинстве случаев, эксперт интуитивно выбирает ту или иную меру).

Динамические методы лишены перечисленных недостатков. Большинство динамических подходов основано на принципе минимальной длины описания (МДО) [Vreeken *et al.*, 2011; Siebes et al., 2011; Smets et al., 2012]. Он опирается на предположение о том, что оптимальная модель обеспечивает максимальное сжатие данных. Применительно к МП, задача состоит в поиске модели (набора паттернов), минимизирующей двухступенчатый код $L(M) + L(D|M)$, где $L(M)$ – длина модели M и $L(D|M)$– длина данных, закодированных с использованием данной модели. Минимизация данной целевой функции подразумевает два этапа: майнинг паттернов (редуцирование пространства поиска) и отбор МДО-оптимальных среди них. Модель M представлена как двухколонная кодовая таблица, где в первой колонке содержатся паттерны, а во второй – соответствующие им коды. Определение длин кодов паттернов, содержащихся в кодовой таблице, основано на введенном вероятностном распределении. Используемые вероятностные модели подвержены влиянию эвристик. Последнее делает модели плохо интерпретируемыми и толерантными к избыточности паттернов. В данной работе мы предлагаем новую вероятностную модель, основанную на частотных оценках. В экспериментах мы показываем, что данная модель менее подвержена влиянию эвристик и позволяет получить множества паттернов с меньшей избыточностью.

1. Основные понятия

Мы рассматриваем «транзакционные базы данных», представленные в виде бинарных таблиц. Фрагмент такой базы данных приведен на Рис. 1. Паттерны в таких данных представляют собой множества признаков. Мы исследуем замкнутые множества, поскольку последние (а) являются

максимальными множествами, которые включают паттерны одинаковой частоты, (б) позволяют представить любой из таких паттернов без потерь. Основные понятия, связанные с замкнутыми множествами признаков, приведены в терминах Анализа формальных понятий [Ganter et al., 1999].

1.1. Анализ формальных понятий. Основные понятия

Пусть задано множество объектов $G=\{g_1, g_2, \ldots, g_n\}$, множество признаков $M=\{m_1, m_2, \ldots, m_k\}$ 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 $X \subseteq G$ и $Y \subseteq M$ она принимает следующий вид:

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

X' представляет собой множество признаков, общих для всех объектов множества X, Y' – множество объектов, обладающих всеми признаками из Y. Формальным понятием называется пара (X,Y), где $X \subseteq G$, $Y \subseteq M$ и $X' = Y$, $Y' = X$. X и Y называют объемом и содержанием формального понятия, соответственно. Произвольное множество $Z \subseteq M$ называют паттерном [Pasquier *et al.*, 1999]. Размером паттерна Z называют мощность его содержания, т.е., $|Z|$, частотой паттерна называют размер объема соответствующего ему формального понятия, т.е. $|Z'|$.

Пример. Рассмотрим формальный контекст, представленный на Рис. 1. Множество формальных понятий, размер содержания и объема которых больше 1, составляют понятия $(g_1g_4g_5, AC)$, $(g_2g_3g_4, DE)$, (g_2g_4, CDE), (g_1g_2, BC). Данные понятия упорядочены по частоте (по убыванию, ↓), длине (↓) и лексикографически (↑).

	A	B	C	D	E
ABC	g_1	×	×	×		
BCDE	g_2		×	×	×	×
DE	g_3				×	×
ACDE	g_4	×		×	×	×
AC	g_5	×		×		

(a) (b)

A	B	C		
	B	C	D	E
			D	E
A		C	D	E
A		C		

(c)

I	U	$-log_2(\cdot)$
A	3	3/15
B	2	2/15
C	4	4/15
D	3	3/15
E	3	3/15
$\sum_{u \in U}$		$u = 15$

(d)

Рис.1. Транзакционная база данных (a), ее представление в виде формального контекста (b) и представление, используемое в данной статье (c). Стандартная кодовая таблица *ST* (d), *L(D, ST)* = 46.1, см. детали в Разделе 1.2.

1.2. Принцип минимальной длины описания в задаче отбора паттернов

В контексте МДО, под оптимальным набором паттернов понимается такое множество паттернов, кодирование на основе которых обеспечивает максимальное сжатие данных [Grünwald, 2007].

Основу данного подхода составляет кодовая таблица, в которую записаны некоторые паттерны и длины их кодов в битах. Процесс майнинга

паттернов состоит в минимизации длины $L(CT, D) = L(D|CT) + L(D|CT)$, где $L(D|CT)$ – длина набора данных D, закодированных с помощью кодовой таблицы CT, $L(CT|D)$ – длина кодовой таблицы CT, вычисленной на основе набора данных D. Далее, для краткости, мы будем обозначать множество паттернов из кодовой таблицы как CT. Кодирование объекта заключается в нахождении непересекающихся паттернов в CT, которые полностью покрывают описание данного объекта. Для нахождения покрытия паттерны таблицы просматриваются в фиксированном порядке. Изначально, за множество непокрытых признаков принимается полное описание объекта. Далее, для каждого последующего паттерна проверяется, содержит ли данный паттерн еще не покрытое описание. Если паттерн содержится в непокрытом описании, он добавляется в результирующее множество-покрытие а его признаки исключаются из множества непокрытых признаков описания. Как только множество непокрытых признаков оказывается пустым, функция возвращает множество паттернов – покрытие описания. Данная функция обозначается далее как $cover(\{g\}', CT)$. Количество использований паттерна X для покрытия объектов в D обозначается $usage(X) = |\{g \in G \mid X \in cover(\{g\}', CT)\}|$. Вероятность паттерна при фиксированной схеме кодирования вычисляется по следующей формуле: $P(X) = usage(X) / \sum_{X^* \in CT} usage(X^*)$.

Для определения длины кодов паттернов используется кодирование Шеннона, т.е. $L(code(X)) = -log(P(X))$, обеспечивающее оптимальные длины кодов в данной вероятностной модели [Grünwald, 2007]. Длина набора данных D, закодированных с использованием CT, вычисляется по формуле $L(D|CT) = \sum_{X \in CT} usage(X) L(code(X))$. Поскольку майнинг паттернов не преследует цели кодирования данных, в качестве целевой функции используется упрощенная длина данных и кодовой таблицы (допускается вещественное количество бит и упускается кодирование непосредственно паттернов – правого столбца кодовой таблицы).

Кодовая таблица, где все паттерны являются одиночными признаками, называется стандартной кодовой таблицей и обозначается ST (см. Рис.1, d).

Пространство поиска оптимального набора паттернов представляет собой множество всех возможных подмножеств различных паттернов. Все возможные паттерны, в свою очередь, представляют булеан множества признаков M. Таким образом, размер пространства поиска 2^N, где $N = 2^{|M|}$. На практике пространство поиска ограничивается подмножеством паттернов, которое называют множеством кандидатов. Для выбора множества кандидатов применяют различные эвристики [Поспелов и др. 1967; Гладун, 1977; Ройзензон, 2005].

1.2.1. Принципы вычисления кодовых таблиц

На начальном этапе кодовая таблица состоит из одноэлементных паттернов $\{\{m\} \mid m \in M\}$. Также имеется набор кандидатов – множество паттернов, упорядоченное в соответствии с выбранной мерой. Минимизация длины заключается в последовательном добавлении лучшего (по выбранной мере) паттерна, перевычислении покрытия с обновленным набором паттернов и вычислении новой длины. Если новая длина короче предыдущей, то данный паттерн добавляется в кодовую таблицу. Процесс продолжается до тех пор, пока все паттерны из набора кандидатов не будут просмотрены.

Стандартным порядком кандидатов называют упорядочивание по частоте (↓), длине (↓) и лексикографически (↑) [Vreeken et al., 2011]. Паттерны в самой кодовой таблице могут быть упорядочены в соответствии с другими мерами. Так упорядочивание паттернов по длине (↓), частоте (↓) и лексикографически (↑) называют стандартным порядком покрытия. Именно в этом порядке паттерны используются при жадном покрытии данных.

Одним из наиболее распространенных МДО-подходов к майнингу паттернов является Krimp [Vreeken et al., 2011].

Пример. Рассмотрим принцип работы Krimp на фрагменте данных, приведенном на Рис. 1. На начальном этапе рассматривается стандартная кодовая таблица (Рис. 1, d). Далее, кандидаты добавляются в таблицу в стандартном порядке кандидатов. Кандидаты в таблице упорядочиваются в соответствии со стандартным порядком покрытия. Если покрытие текущим набором паттернов обеспечивает меньшую длину, паттерн-кандидат принимается в таблицу.

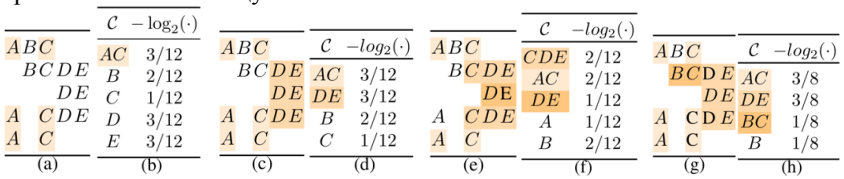

Рис. 2. Шаги Krimp. Кандидаты упорядочены по частоте (↓), длине (↓), лексикографически. Паттерны в таблице упорядочиваются по длине (↓), частоте (↓), лексикографически (↑). Шаг 1: добавление *AC* (a-b), *L(D,CT)=39.0 < L(D,ST),* паттерн принимается в таблицу. Шаг 2: добавление *DE* (c-d), *L(D,CT)*=25.5 < 39.0, паттерн принимается. Шаг 3: добавление *CDE* (e-f), *L(D,CT)* = 30.0 > 25.5, паттерн удаляется. Шаг 3: добавление *BC* (g-h), *L(D,CT)*=23.3 < 25.5, паттерн принимается.

3. Вероятностные модели для кодирования паттернов

В предыдущем разделе было отмечено, что определение кодов на основе вероятностного распределения обеспечивает оптимальную длину кодирования. Однако в теории сжатия информации, в частности

применительно к МП, вопросу определения вероятностных пространств не было уделено должного внимания. В данном разделе мы рассматриваем вероятностную модель, используемую в методах майнинга паттернов на основе МДО, и ее недостатки, а также вводим новую модель, позволяющую минимизировать эти недостатки.

3.1. Модель на основе функции покрытия

Во всех существующих подходах к МП на основе МДО, вероятностное пространство определяется в соответствии с алгоритмом жадного покрытия. Другими словами, вероятность паттерна зависит от вероятности других паттернов, которые ранее использовались при покрытии данных.

Пусть $X_1, X_2, \ldots, X_n \in CT$ множество упорядоченных паттернов кодовой таблицы, $n(X_i)$ – количество объектов, описание которых содержит паттерн X_i; $n(X_i\overline{X}_j)$ – количество объектов, описание которых содержит паттерн X_i и не содержит паттерн X_j. Тогда $usage(X_i) = n(X_i\overline{X}_{i_1} \ldots \overline{X}_{i_k})$, где $i_1, \ldots, i_k < i, \overline{X}_{i_j} \cap X_i \neq \emptyset, usage(X_i) \leq frequency(X_i)$, что следует из рассмотренной стратегии жадного покрытия. Определенное таким образом вероятностное пространство подразумевает учет достаточно сложной зависимости между паттернами, а также покрытие данных без пересечений. В экспериментах мы покажем, что это приводит к оптимистичным оценкам избыточного описания.

3.2. Модель на основе независимых паттернов

В данной работе мы предлагаем вероятностную модель, основанную на предположении о независимости паттернов в кодовой таблице, т.е. вероятность наблюдения паттерна X в данных вычисляется независимо от других паттернов из CT. В результате, оценки вероятности оцениваются не на основе частоты встречаемости в покрытии, а на основе частоты встречаемости в данных, т.е. $P(X) = frequency(X) / \sum_{X^* \in CT} frequency(X^*)$. Предположение о независимости паттернов облегчает интерпретацию модели и снижает зависимость оценок длин от способа покрытия данных (порядка паттернов в кодовой таблице).

4. Эксперименты

В данном разделе мы приводим результаты экспериментов, показывающие, что замена оценок позволяет существенно уменьшить избыточность описания и размер МДО-оптимального набора паттернов. Мы использовали 20 наборов данных из репозитория LUCS-KDD [Coenen, 2003].

Существует большое количество различных мер оценки качества множества признаков [Ignatov et al., 2015]. Условно их можно разделить на

3 группы: те, что оценивают простоту анализа (интерпретации), избыточность (разнообразие) и описательную способность.

В данной работе мы выбрали указанные ниже меры.

– *Количество паттернов (|CT|)* в таблице. Меньшее количество предпочтительнее, поскольку облегчает последующий анализ паттернов.

– *Коэффициент перекрытия (O),* среднее число паттернов, приходящихся на покрытые отношения «объект-признак». Оценивает избыточность набора паттернов. Наборам паттернов с отсутствием избыточности соответствует коэффициент перекрытия 1.

– *Коэффициент непокрытия (U),* доля непокрытых отношений «объект-признак». Оценивает описательную способность паттернов. Значения близкие к 0 соответствуют множествам паттернов с наилучшей описательной способностью.

Табл. 1

Порядок кандидатов	*|X|*|X'|↓, |X'|↓, лексикограф.↑*				*|X'|↓, |X|↓, лексикограф. ↑*			
Порядок покрытия	*|X|*|X'|↑, лексикогр.↑*		*|X|↓, |X'|↓, лексикогр.↑*		*|X|*|X'|↑, лексикогр.↑*		*|X|↓, |X'|↓, лексикогр.↑*	
Оценки	*us.*	*freq.*	*us.*	*freq.*	*us.*	*freq.*	*us.*	*freq.*
L(D,CT)/ L(D, ST)	0,59	0,69	0,53	0,70	0,66	0,78	0,53	0,78
|CT|	29,23	20,91	123,82	21,41	50,14	20,05	114,82	20,14
O	2,75	2,12	12,58	2,26	3,74	1,69	11,37	1,71
U	0,32	0,26	0,20	0,26	0,22	0,25	0,19	0,25

В результате серии экспериментов, проведённых с разным упорядочиванием кандидатов и паттернов (покрытия), было выявлено, что предложенные оценки вероятности (столбец «*freq.*»), позволяют получить кодовые таблицы меньшего размера (строка «*|CT|*»). В случае, когда используется стандартный порядок покрытия (столбец «*|X|↓, |X'|↓, лексикограф.↑*»), размер таблиц уменьшается более чем в 5 раз. Кроме того, значительно уменьшается избыточность наборов паттернов (строка «*O*»), при этом описательная способность сокращается лишь на 4.5% (строка «*U*»). В случае упорядочивания кандидатов по «*|X|*|X'|↓, |X'|↓, лексикограф.↑*» и покрытия по «*|X|*|X'|↑, лексикограф. ↑*», описательная способность также улучшается вместе с уменьшением избыточности и размера кодовых таблиц.

Стоит отметить, что коэффициент компрессии *L(D,CT)/L(D,ST)* повышается лишь незначительно. Данный параметр не свидетельствует об ухудшении модели. Экспериментально было показано, что для одних и тех же наборов паттернов, коэффициент компрессии в рамках предложенной вероятностной модели всегда выше.

Заключение

В данной статье была предложена новая вероятностная модель для методов майнинга паттернов на основе принципа минимальной длины описания. В результате экспериментов было показано, что введенные оценки позволяют получить более компактные наборы признаков, обладающих меньшей избыточностью и сохраняющих при этом описательную способность (в сравнении с множествами признаков большего размера).

Список литературы

[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

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

[Ignatov et al., 2015] Ignatov D.I., Gnatyshak D.V., Kuznetsov S.O., Mirkin B.G. Triadic formal concept analysis and triclustering: searching for optimal patterns // Machine Learning – T. 101, №. 1-3, 2015..

[Geng et al., 2006] Geng L., Hamilton H. J. Interestingness measures for data mining: A survey //ACM Computing Surveys (CSUR), 38(3):9, 2006.

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

[**Pasquier et al., 1999]** Pasquier N., Bastide Y., Taouil R., and Lakhal L. Efficient mining of association rules using closed itemset lattices // Information systems, 24(1):25–46, 1999.

[Siebes et al., 2011] Siebes A., Kersten R. A structure function for transaction data // In Proceedings of SIAM, pages 558–569, 2011.

[Smets et al, 2012] Smets K., Vreeken J. Slim: Directly mining descriptive patterns // In Proceedings of SIAM, pages 236–247, 2012.

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

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

[Поспелов и др., 1967] Поспелов Д. А., Пушкин В. Н., Садовский В. Н. Эвристическое программирование и эвристика как наука // Вопросы философии. 1967. №7. С. 45–56.

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

PATTERN MINING THROUHG COMPRESSION : TOWARDS TO PROBABISTIC MODELS

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

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

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

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

In this paper we introduce a probabilistic model of MDL-based approaches to Pattern Mining. We show in experiments that the proposed model allows for computing pattern sets of a smaller size, reducing redundancy, while retaining their descriptiveness.

Keywords: minimal description length principle, pattern mining, closed itemsets