Метод бикластеризации на основе объектных и признаковых замыканий*

Игнатов Д. И., Каминская А. Ю., Кузнецов С. О., Магизов Р. А.
dignatov@hse.ru
Москва, Государственный Университет — Высшая Школа Экономики

Разработан новый метод бикластеризации объектно-признаковых данных, основанный на объектных и признаковых замыканиях из анализа формальных понятий (АФП). Предлагается опеределение бикластера, исследуются его свойства и связь с решетками формальных понятий. Приводится оценка ресурсной сложности двух версий алгоритмов для порождения бикластеров данного вида. Преимущества предлагаемого подхода перед алгоритмами поиска формальных понятий (ФП) заключается в меньшем размере выхода и более высокой производительности при условии «сохранения» ФП в смысле отношения покомпонентного вложения, определенного на ФП и бикластерах. Приведены результаты экспериментов на массивах данных из UCI Machine Learning Repository. Исследуются свойства масштабируемости алгоритмов на примере параллельной реализации.

A concept-based biclustering algorithm*

Ignatov D. I., Kaminskaya A. Y., Kuznetsov S. O., Magizov R. A.
State University Higher School of Economics, Moscow, Russia

A novel biclustering algorithm based on closure operator is proposed. Definition of a bicluster is given, its properties and relation to formal concept from Formal Concept Analysis (FCA) are studied. Complexities of two versions of the algorithm are compared. Experimental results on datasets from the UCI Machine Learning Repository are given. Scalability of the parallel implementations of the algorithm is studied.

В настоящее время методы бикластеризации завоевывают все большую популярность, особенно среди исследователей в области биоинформатики для изучения данных генной экспрессии [5]. Это вызвано потребностью в сохранении объектно-признаковой структуры кластеров, например, для адекватного понимания, в каких свойствах выражено сходство некоторой группы генов. Различные приложения этих методов востребованы в области анализа Интернет-данных, например, такие алгоритмы — основа некоторых рекомендательных Интернет-сервисов [6, 4]. В данной работе мы представляем метод бикластеризации, основанный на решетках формальных понятий. Благодаря средствам анализа формальных понятий (АФП) [3] для любых объектно-признаковых данных можно построить иерархическую структуру формальных понятий (бикластеров специального вида), позволяющую отобразить их таксономические свойства в удобном для аналитика виде. Основным недостатком решеток понятий является их большой размер, например, для объектно-признаковой таблицы размером 10×10 число таких бикластеров в худшем случае равно 2^{10}. Одной из задач, на решение которой направлены усилия ученых, работающих в данном сообществе, является разработка методов по отбору наиболее полезных, релевантных понятий, сокращению размеров порождаемого множества понятий. Один из подходов заключается в ослаблении требований к формальным понятиям; в его рамках возможно не только сокращение числа порождаемых бикластеров, но и успешное устранение влияния шума на результаты [2]. Нами предлагается метод бикластеризации, который использует только небольшую часть формальных понятий (объектные и признаковые) для порождения бикластеров особого вида. В качестве критерия отбора релевантных бикластеров применяется их плотность. Предлагаются два подхода к реализации метода — по определению и улучшенный, на основе свойства монотонности оператора Галуа. Проводятся эксперименты на реальных данных, иллюстрирующие улучшение производительности благодаря оптимизации и распараллеливанию при различных порогах плотности.

Основные определения

Определение 1. *Формальный контекст* $\mathbb{K}$ *есть тройка* (G, M, I), *где* G *— множество, называемое множеством объектов,* M *— множество, называемое множеством признаков,* $I \subseteq G \times M$ *— отношение.*

Отношение I интерпретируется следующим образом: для $g \in G$, $m \in M$ имеет место gIm, если объект g обладает признаком m.

Для формального контекста $\mathbb{K} = (G, M, I)$ и произвольных $A \subseteq G$ и $B \subseteq M$ определена пара отображений:

$$A' := \{m \in M \mid gIm \text{ для всех } g \in A\},$$
$$B' := \{g \in G \mid gIm \text{ для всех } m \in B\},$$

Работа выполнена при финансовой поддержке РФФИ, проект №08-07-92497-НЦНИЛ_а.

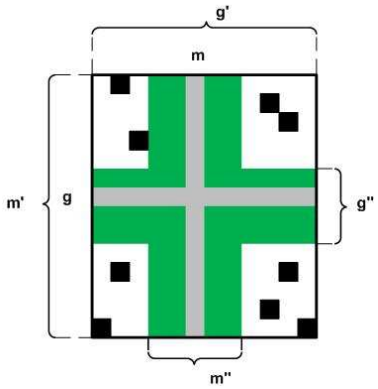

Рис. 1. Бикластер на основе объектных и признаковых замыканий.

которые задают *соответствие Галуа* между частично упорядоченными множествами $(2^G, \subseteq)$ и $(2^M, \subseteq)$, а оператор $(\cdot)''$ является *оператором замыкания* на $G \dot\cup M$ — дизъюнктном объединении G и M, т. е. для произвольного $A \subseteq G$ или $A \subseteq M$ имеют место следующие соотношения [1]:

1) $A \subseteq A''$ (экстенсивность);
2) $A'''' = A''$ (идемпотентность);
3) если $A \subseteq C$, то $A'' \subseteq C''$ (изотонность).

Множество A называется *замкнутым*, если $A'' = A$ [1].

Определение 2. *Формальное понятие формального контекста $\mathbb{K} = (G, M, I)$ есть пара (A, B), где $A \subseteq G$, $B \subseteq M$, $A' = B$ и $B' = A$. Множество A называется объёмом, а B — содержанием понятия (A, B).*

Очевидно, что объем и содержание произвольного формального понятия являются замкнутыми множествами.

Множество формальных понятий контекста $\mathbb{K}$, которое мы будем обозначать через $\mathfrak{B}(G, M, I)$, частично упорядочено по вложению объёмов: формальное понятие $X = (A, B)$ является *менее общим* (*более частным*), чем понятие $Y = (C, D)$, $(A, B) \leqslant (C, D)$, если $A \subseteq C$, что эквивалентно $D \subseteq B$ (Y — *обобщение* X).

Описание модели бикластеризации и алгоритмов

Рассмотрим алгоритм для предложенного нами определения бикластера. Предварительно введем определения объектного и признакового понятия.

Определение 3. *Объектным понятием формального контекста $\mathbb{K} = (G, M, I)$ называется формальное понятие вида (g'', g'), где $g \in G$, а признаковым понятием называется формальное понятие вида (m', m''), где $m \in M$.*

Обозначим множество всех объектных понятий формального контекста $\mathbb{K} = (G, M, I)$ как $Obj = \{(g'', g') \mid \forall g \in G\}$, а множество всех признаковых понятий как $Attr = \{(m', m'') \mid \forall m \in M\}$.

Определение 4. *Для формального контекста $\mathbb{K} = (G, M, I)$ и любой пары объектных и признаковых понятий $(g'', g') \in Obj$ и $(m', m'') \in Attr$, связанных отношением вложения $(g'', g') \preceq (m', m'')$, назовем бикластером пару вида (m', g'), см. рис. 1.*

Требование вложения объектного понятия (g'', g') в признаковое (m', m'') объясняется тем фактом, что объектные понятия имеют самые большие по размеру содержания, а признаковые — самые большие объемы.

Определение 5. *Плотностью бикластера (A, B) формального контекста $\mathbb{K} = (G, M, I)$ назовем величину $\rho(A, B) = \frac{|I \cap A \times B|}{|A||B|}$.*

Очевидно, что для произвольного бикластера (A, B) справедливо соотношение $0 \leqslant \rho \leqslant 1$.

Следствие 1. *Пусть (A, B) — формальное понятие, тогда $\rho(A, B) = 1$*

Определение 6. *Пусть дан бикластер $(A, B) \in \in 2^G \times 2^M$ и положительное целое число $\rho_{\min}$ такое, что $0 \leqslant \rho_{\min} \leqslant 1$. Тогда (A, B) называется плотным в том и только том случае, когда он удовлетворяет ограничению*

$$C_{\text{dense}}(\rho_{\min}, (A, B)) \equiv (\rho(A, B) \geqslant \rho_{\min}).$$

Определение 7. *Отношение вложения на бикластерах определим покомпонентно: $(X_1, Y_1) \sqsubseteq (X_2, Y_2) \Leftrightarrow X1 \subseteq X2$ и $Y1 \subseteq Y2$. Ограничение называется антимонотонным по отношению $\sqsubseteq$ тогда и только тогда, когда для пары бикластеров b_1 и b_2, таких что $b1 \subseteq b2$ выполняется $C(b_2) \Rightarrow C(b_1)$. Двойственным образом определяется монотоннсть ограничения. Запись $C(b)$ означает, что ограничение C выполняется для бикластера b.*

Утверждение 1. *Данное ограничение не является ни монотонным, ни антимонотонным.*

Пусть $\rho_{\min} = 0$, тогда для любого формального понятия некоторого контекста $\mathbb{K}$ существует бикластер из множества всех бикластеров данного контекста $\mathbf{B}$, в который оно вкладывается.

Утверждение 2. *Для любого $(A_c, B_c) \in \mathfrak{B}(G, M, I)$ найдётся $(A_b, B_b) \in \mathbf{B}$ такой, что $(A_c, B_c) \sqsubseteq (A_b, B_b)$.*

Тот факт, что бикластер представим в виде (m', g') наводит на мысль о том, что возможно избежать дорогостоящего (в худшем случае $O(|G||M|)$) вычисления оператора замыкания $(\cdot)''$. Алгоритм 2 является улучшенной версией алгоритма 1 и использует свойство антимонотонности оператора $(.)'$ для оптимизации.

Утверждение 3. *Выходы алгоритмов 1 и 2 равны.*

Алгоритм 1. Алгоритм поиска бикластеров.

Вход: $K = (G, M, I)$ – формальный контекст;
$\rho_{\min}$ – порог на значение плотности бикластера;
Выход: $B = \{(A_k, B_k) \mid (A_k, B_k) — \text{бикластер}\}$.

1: $Obj = \varnothing$; $Attr = \varnothing$; $B = \varnothing$;
2: **для всех** $g \in G$
3: $Obj.Add((g'', g'))$;
4: **для всех** $m \in M$
5: $Attr.Add((m', m''))$;
6: **для всех** $(A, B) \in Attr$
7: **для всех** $(C, D) \in Obj$
8: **если** $C \subseteq A$ и $\rho(A, D) \geqslant \rho_{\min}$ **то**
9: $B.Add((A, D))$;

При достаточно больших значениях $\rho_{\min}$ не все формальные понятия могут оказаться вложенными в некоторый бикластер, построенный по формальному контексту $\mathbb{K} = (G, M, I)$.

Утверждение 4. *Для некоторого $\rho_{\min} > 0$ существует $\mathbb{K} = (G, M, I)$ такой, что для некоторого формального понятия $(A_c, B_c) \in \mathfrak{B}(G, M, I)$ верно, что не существует $(A_b, B_b) \in \mathbf{B}$ такого, что $(A_c, B_c) \sqsubseteq (A_b, B_b)$.*

Оценим теоретически ресурсную временную и емкостную сложность алгоритмов 1 и 2:

Как уже отмечалось, вычисление операции замыкания занимает время $O(|G||M|)$, поэтому в циклах 2–3 и 4–5 алгоритма 1 эта величина умножается на число итераций циклов, т. е. на число объектов и признаков соответственно. Модифицированная версия алгоритма избавлена от такого дорогостоящего вычисления, для каждого объекта (признака). В каждой итерации циклов 2–3 и 4–5 алгоритма 2 вычисляются только содержание для объектов и объем для признаков соответственно, поэтому временная сложность каждого из этих циклов равна $O(|G||M|)$. Следует отдельно рассмотреть вычисления в цикле 6–9: т. к. при значении минимального порога плотности $\rho = 0$ проверку выполнять не стоит, то фактически необходимо только «склеить» пары объектных и признаковых понятий в согласии с определением 4. Для наивной версии алгоритма это потребует времени $O(|G|^2|M|)$, а для улучшенной $O(|I|)$, т. к. факт вложения проверять явно не нужно. В случае проверки ограничения, наложенного на значение плотности бикластера, ее вычисление займет $O(|G||M|)$, что увеличит оценку времени выполнения цикла 6–9 до $O(|G|^2|M|^2)$ в первом и до $O(|I||G||M|)$ во втором алгоритмах.

Таблица 1. Теоретическая временная сложность основных блоков алгоритмов бикластеризации.

Шаги алгоритма	Алгоритм 1	Алгоритм 2
цикл $2-3$: порождение объектных понятий	$O(\|G\|^2\|M\|)$	$O(\|G\|\|M\|)$
цикл $4-5$: порождение признаковых понятий	$O(\|G\|\|M\|^2)$	$O(\|G\|\|M\|)$
цикл $6-9$: построение бикластеров ($\rho_{\min} = 0$)	$O(\|G\|\|M\|^2)$	$O(\|I\|)$
цикл $6-9$: построение бикластеров($\rho_{\min} > 0$)	$O(\|G\|^2\|M\|^2)$	$O(\|I\|\|G\|\|M\|)$

Алгоритм 2. Улучшенный алгоритм поиска бикластеров.

Вход: $K = (G, M, I)$ – формальный контекст;
$\rho_{\min}$ – порог на значение плотности бикластера;
Выход: $B = \{(A_k, B_k) \mid (A_k, B_k) — \text{бикластер}\}$.

1: $Obj.Size = |G|$; $Attr.Size = |M|$; $B = \varnothing$;
2: **для всех** $g \in G$
3: $Obj[g] = g'$;
4: **для всех** $m \in M$
5: $Attr[m] = m'$;
6: **для** $g = 1, \ldots, |G|$
7: **для всех** $m \in Obj[g]$
8: **если** $\rho\big(Attr[m], Obj[g]\big) \geqslant \rho_{\min}$ **то**
9: $B.Add\big((Attr[m], Obj[g])\big)$;

Оценим нижнюю и верхнюю границы размера выхода алгоритма для случая, когда все объектные и признаковые формальные понятия участвуют в формировании бикластеров, т. е. $\rho = 0$. Получим соотношение $1 \leqslant |\mathcal{B}| \leqslant |I|$. Число бикластеров не может быть равным нулю, потому что в формировании бикластеров для любого контекста участвуют хотя бы одно признаковое и одно объектное понятия, которые вкладываются друг в друга. Очевидно, что количество бикластеров не может быть больше числа пар, принадлежащих отношению I. Такие оценки сложности вычислений и размера выхода, а также возможность «сохранения» формальных понятий в множестве порожденных бикластеров при $\rho = 0$, позволяют говорить о предложенном алгоритме как о хорошей альтернативе формальным понятиям в тех случаях, когда допустимы приближенные описания извлекаемых из данных знаний.

Эмпирический анализ эффективности алгоритма

Для экспериментальной оценки эффективности предложенного алгоритма бикластерзации были запрограммированы версии алгоритма 1 и 2. В качестве языка реализации был выбран C# из среды разработки Microsoft Visual Studio 2008. Дополнительно была исследована возможность распараллеливания (масштабирования) алгоритма с помощью средств библиотеки Task Parallel Library, входящей в состав Microsoft .NET Framework 4.0.

Эксперименты были проведены на данных из UCI Machine Learning Repository и на массиве дан-

Таблица 2. Описание наборов данных.

Набор данных	$\lvert G\rvert\times\lvert M\rvert$ ×	$\lvert I\rvert$	Плотность	$\lvert\mathfrak{B}(G,M,I)\rvert$
advertising	2000×3000	92 345	0,015	8 950 740
breast-cancer	286×43	2851	0,232	9918
flare	1389×49	18057	0,265	28742
postoperative	90×26	807	0,345	2378
SPECT	267×23	2042	0,333	21550
vote	435×18	3856	0,492	10644
zoo	101×28	862	0,305	379

Таблица 3. Зависимость количества бикластеров от величины $\rho_{\min}$.

Название	advertising	breast-cancer	flare	SPECT	vote	zoo
$\lvert\mathfrak{B}(G,M,I)\rvert$	8950740	9918	28742	21550	10644	379
$\rho = 0{,}0$	92345	2851	18057	807	3856	862
$\rho = 0{,}1$	89735	2851	18057	2042	3856	862
$\rho = 0{,}2$	80893	2851	18057	2042	3856	862
$\rho = 0{,}3$	65881	2849	18050	2042	3855	862
$\rho = 0{,}4$	45665	2678	17988	2029	3829	853
$\rho = 0{,}5$	25921	1908	17720	1753	3527	776
$\rho = 0{,}6$	10066	310	16459	835	2575	521
$\rho = 0{,}7$	2081	17	9353	262	1458	341
$\rho = 0{,}8$	165	2	1450	85	382	225
$\rho = 0{,}9$	3	2	293	32	33	63
$\rho = 1{,}0$	0	2	3	12	1	7

ных о покупке рекламных словосочетаний компании Yahoo. В таблице 2 приведено описание этих наборов данных, для каждого из них указано количество объектов, признаков, число пар, принадлежащих отношению I, плотность контекста и количество его формальных понятий.

Помимо производительности алгоритмов нас интересовала зависимость количества порождаемых бкиластеров от выбранного порога плотности. Результаты экспериментов приведены в таблице 3.

В шести экспериментах из семи мы наблюдаем заметно меньшее число бикластеров по сравнению с количество формальных понятий. Для седьмого эксперимента такой результат объясняется тем, что имеется большое количество объектов, содержания которых представимы в виде пересечения содержаний других объектов (операция редуцирования контекста по объектам оставляет только 59 объектов из 100).

Зависимость количества бикластеров от величины порога носит довольно плавный характер, хотя для некоторых данных UCI Machine Learning Repository это количество остается постоянным при уменьшении значения порога на 0,1, начиная с 1, для нескольких первых точек.

Результаты экспериментов по оценке временной эффективности алгоритмов на самом крупном из имеющихся массивов реальных данных — advertising, говорят о том, что в среднем время работы «наивной» версии алгоритма превосходит

Таблица 4. Отношение числа порожденных понятий к количеству бикластеров.

Набор данных	Сокращение
advertising	96,9
breast-cancer	3,5
flare	1,6
postoperative	2,9
SPECT	10,6
vote	2,8
zoo	0,4

время работы оптимизированной версии на порядок, более точно в 28 раз. Параллельная версия оптимизированного алгоритма работает быстрее последовательной реализации на 45%. Между тем заметного выигрыша от применения техник параллельного программирования для случая «наивной» версии алгоритма не получено. Дополнительные эксперименты проведены на массивах данных из репозитория UCI, результаты носят аналогичный характер.

В качестве направлений дальнейших исследований можно предложить разработку жадных версий разработанного алгоритма, например, по покрытию множества признаков M или отношения I, а также исследование их предсказательных свойств для рекомендательных систем и анализа данных генной экспрессии.

Литература

[1] *Биркгоф Г.* Теория решеток. — М.:Наука, 1989.

[2] *Besson J., Robardet C., Boulicaut J-F.* Constraint-based mining of formal concepts in transactional data // In proceedings of the 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2004.

[3] *Ganter B., Wille R.* Formal Concept Analysis: Mathematical Foundations // Springer, 1999.

[4] *Ignatov D. I., Kuznetsov S. O.* Concept-based Recommendations for Internet Advertisement // In proceedings of The Sixth International Conference Concept Lattices and Their Applications (CLA'08), Olomouc, Czech Republic, 2008.

[5] *Madeira S. K., Oliveira A. L.* Biclustering Algorithms for Biological Data Analysis: A Survey // IEEE/ACM Transactions on Computational Biology and Bioinformatics, VOL 1, NO. 1, pp. 24-45 January-March 2004.

[6] *Symeonidis P., Nanopoulos A., Papadopoulos A. N., Manolopoulos Y.* Nearest-biclusters collaborative filtering based on constant and coherent values // Inf. Retr. 11(1): 51–75 (2008).

[7] *Wille R.* Restructuring Lattice Theory: an Approach Based on Hierarchies of Concepts // Ordered Sets / Ed. by I. Rival. — Dordrecht; Boston: Reidel, 1982.— P. 445–470.