УДК 004.85, 004.89

Бикластеризация объектно-признаковых данных на основе решеток замкнутых множеств

Д.И. Игнатов (*dignatov@hse.ru*)

С.О. Кузнецов (*skuznetsov@hse.ru*)

ГУ-ВШЭ, Москва, Покровский бульвар д.11, 109028

В работе предлагается новый метод бикластеризации объектно-признаковых данных, опирающийся на свойства решеток замкнутых множеств. Предложено определение плотного бикластера, эффективный алгоритм для поиска таких бикластеров, исследована его сложность, проведены вычислительные эксперименты на реальных данных. Исследована на практике возможность масштабирования (распараллеливания) алгоритма.

Введение

В настоящее время по сравнению с традиционным кластерным анализом методы бикластеризации завоевывают все большую популярность, особенно в рядах исследователей в области биоинформатики для изучения данных генной экспрессии [Madeira et al., 2004]. Это вызвано потребностью в сохранении объектно-признаковой структуры данных, 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 [Ignatov et al., 2008]. В данной работе мы представляем метод бикластеризации, основанный на решетках формальных понятий. Благодаря средствам анализа формальных понятий (АФП) [Ganter et al., 1999] 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 громоздкость, например, для объектно-признаковой таблицы размером 10x10 число бикластеров в худшем случае равно 210. 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 успешное устранение влияния шума на результаты [Besson et al., 2006]. 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. В дополнение к описываемой в статье версии алгоритма предлагается ее параллельная реализация. Проводятся эксперименты на реальных данных, иллюстрирующие улучшение производительности благодаря оптимизации и распараллеливанию при различных порогах плотности.

Основные определения

Контекстом в АФП называют тройку $K = (G, M, I)$, где G — множество объектов, M — множество признаков, а отношение $I \subseteq G \times M$ говорит о том, какие объекты обладают теми или иными признаками. Для произвольных $A \subseteq G$ и $B \subseteq M$ определены операторы Галуа:

$$A' = \{m \in M \mid \forall g \in A\ (g\ I\ m)\};$$
$$B' = \{g \in G \mid \forall m \in B\ (g\ I\ m)\}.$$

Оператор (.)'' (двукратное применение оператора (.)') является оператором замыкания: он идемпотентен ($A'''' = A''$), монотонен ($A \subseteq B$ влечет $A'' \subseteq B''$) и экстенсивен ($A \subseteq A''$). Множество объектов $A \subseteq G$, для которого $A'' = A$, называется замкнутым. Аналогично для замкнутых множеств признаков – подмножеств множества M. Пара множеств (A, B), таких, что $A \subseteq G$, $B \subseteq M$, $A' = B$ и $B' = A$, называется формальным понятием контекста K. Множества A и B замкнуты и называются объемом и содержанием формального понятия (A, B) соответственно. Для множества объектов A множество их общих признаков A' служит описанием сходства объектов из множества A, а замкнутое множество A'' является кластером сходных объектов (с множеством общих признаков A'). Множество объектных понятий образовано парами объектов (g'', g') для всех $g \in G$, а множество признаковых понятий – (m', m'') для всех $m \in M$. Термин бикластер предложен в работе [Миркин, 1996] и означает множество объектов, сходство которых задано посредством общих значений признаков. Другое определение бикластера предложено в работе [Barkov et al., 2006]. Алгоритм Bi-Max, описанный в этой работе строит *максимальные по вложению бикластеры*, определяемые следующим образом. Дано m генов, n ситуаций и бинарная таблица e такая, что $e_{ij}=1$ (ген i активен в ситуации j) или $e_{ij}=0$ (ген i не активен в ситуации j) для всех $i \in [1,m]$ и $j \in [1,n]$. Пара $(G, C) \in 2^{\{1,\ldots,n\}} \times 2^{\{1,\ldots,m\}}$ называется максимальным по вложению бикластером тогда и только тогда, когда

выполнено (1) для любых $i \in G$ и $j \in C$: $e_{ij}=1$ и (2) не существует $(G_1,C_1) \in 2^{\{1,\ldots,n\}} \times 2^{\{1,\ldots,m\}}$ таких, что (a) для любых $i_1 \in G_1$ и $j_1 \in C_1$: $e_{ij}=1$ и (b) $G \subseteq G_1$, $C \subseteq C_1$ и $(G_1, C_1) \neq (G, C)$. Пусть H – множество генов, S – множество ситуаций, а $E \subseteq H \times S$ – бинарное отношение задаваемое таблицей e, $|H|=m$, $|S|=n$. Верно следующее утверждение

Утверждение 1. Для любой пары (G, C), $G \subseteq H$, $C \subseteq S$ следующие два утверждения эквивалентны.

(G, C) максимальный по вложению бикластер таблицы e;

(G, C) формальное понятие контекста (H, S, E).

В следующем разделе мы дадим определение бикластера, основанное на АФП.

Вычислительная модель

Определение 1. Если $(g,m) \in I$, то (m',g') называется *объектно-признаковым бикластером* или *оа-бикластером* с плотностью $\rho(m',g')=|I \cap (m' \times g')|/(|m'| \cdot |g'|)$.

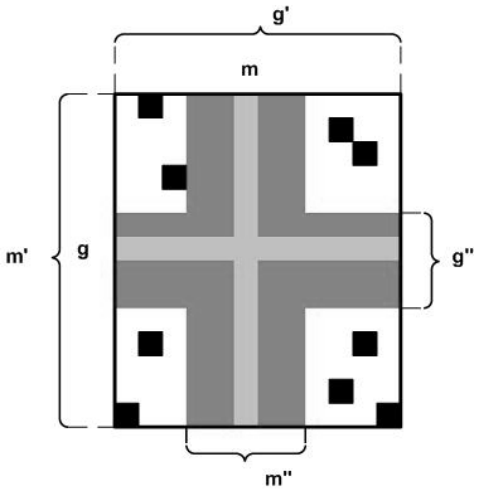

Рисунок 1. оа-бикластер

Приведем основные свойства оа-бикластеров.

Утверждение 2.

1. $0 \leq \rho \leq 1$.

2. оа-бикластер (m',g') является формальным понятием тогда и только тогда, когда $\rho=1$.

3. Если (m', g') – бикластер, то $(g'', g') \leq (m', m'')$.

Определение 2. Пусть $(A,B)\in 2^G\times 2^M$ будет бикластером и $\rho_{\min}$ неотрицательное действительное число такое, что $0\le\rho_{min}\le 1$, тогда (A, B) называется плотным, если он удовлетворяет ограничению $\rho(A,B)\ge\rho_{min}$.

Свойство монотонности (антимонотонности) некоторого ограничения часто используется в Data Mining, например, для поиска ассоциативных правил для улучшения эффективности алгоритма.

Отношение $\mid$ на множестве оа-бикластеров определяется по вложению соответствующих компонент: $(A, B) \mid (C, D)\Leftrightarrow A\subseteq C$ и $B\subseteq D$.

Утверждение 3. Ограничение $\rho(A,B)\ge \rho_{min}$ не является ни монотонным, ни антимонотонным в смысле отношения $\mid$.

Однако ограничение $\rho_{\min}$ обладает другими полезными свойствами. Если $\rho_{min}=0$, то мы получаем множество всех оа-бикластеров контекста K. Для $\rho_{min}=0$ каждое формальное понятие исходного контекста содержится в некотором его оа-бикластере, т.е. верно следующее утверждение.

Утверждение 4. Для любого $(A_c, B_c)\in \mathfrak{B}(G, M, I)$ существует бикластер $(A_b, B_b)\in \mathbf{B}$ такой что $(A_c, B_c) \mid (A_a, B_b)$.

Утверждение 5. Для данного формального контекста $K = (G, M, I)$ и $\rho_{min}=0$ наибольшее число оа-бикластеров равно $|I|$, а все оа-бикластеры порождаются за время $O(|I|\cdot(|G|+|M|))$.

Утверждение 6. Для данного формального контекста $K = (G, M, I)$ и $\rho_{min}>0$ наибольшее число оа-бикластеров равно $|I|$, а все оа-бикластеры порождаются за время $O(|I|\cdot|G|\cdot|M|)$.

Интересующее нас множество бикластеров является результатом работы Алгоритма 1.

Таблица 1. Алгоритм 1

Вход: $K = (G, M, I)$ – формальный контекст, ρmin – минимальное значение порога плотности.	
Выход: $\mathbf{B}=\{(A_k,B_k)\mid 0\le k\le\|\boldsymbol{B}\|\}$ – множество бикластеров	
1	*O*.size=|*G*|, *A*.size=|*M*|, ***B***={}
2	For *g* in *G*:
3	*O*[*g*]=*g*'
4	For *m* in *M*:
5	*A*[*m*]=*m*'
6	For *g* in *G*:
7	For *m* in *O*[*g*]:
8	*b*=(*O*[*g*], *A*[*m*])
9	if(ρ(*b*)≥ρ*min*):
10	***B***.Add(b)
11	return ***B***

Благодаря свойству антимонотонности оператора (.)' мы провели оптимизацию условия порождения бикластера в цикле 6-10 и избежали вычисления операции замыкания в циклах 2-3 и 4-5.

Результаты

Для экспериментальной оценки эффективности предложенного алгоритма бикластеризации была запрограммирована версия Алгоритма 1. 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 данных 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.

Таблица 2. Описание наборов данных

Название	$\|G\|\times\|M\|$	$\|I\|$	Плотность	$\mathsf{B}(G, M, I)$
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

Помимо производительности алгоритмов нас интересовала зависимость количества порождаемых оа-бикластеров от выбранного порога плотности, результаты экспериментов приведены в таблице 3.

Таблица 3. Зависимость количества бикластеров от величины ρ_{min}

Название	advertising	breast-cancer	post-operative	flare	SPECT	vote	zoo
$\mathsf{B}(G,M,I)$	8950740	9918	2378	28742	21550	10644	379
$\rho_{min}=0$	92345	2851	807	18057	807	3856	862
$\rho_{min}=0,1$	89735	2851	807	18057	2042	3856	862
$\rho_{min}=0,2$	80893	2851	807	18057	2042	3856	862
$\rho_{min}=0,3$	65881	2849	807	18050	2042	3855	862
$\rho_{min}=0,4$	45665	2678	807	17988	2029	3829	853

ρ_{min}=0,5	25921	1908	725	17720	1753	3527	776
ρ_{min}=0,6	10066	310	402	16459	835	2575	521
ρ_{min}=0,7	2081	17	18	9353	262	1458	341
ρ_{min}=0,8	165	2	2	1450	85	382	225
ρ_{min}=0,9	3	2	2	293	32	33	63
ρ_{min}=1	0	2	2	3	12	1	7

В 6 экспериментах из 7 мы наблюдаем заметно меньшее число бикластеров по сравнению с количеством формальных понятий. Для 7 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).

Таблица 4. Отношение числа порожденных понятий к количеству бикластеров

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

Приведем результаты экспериментов по оценке временной эффективности алгоритмов на самом крупном из имеющихся массивов реальных данных – advertising (см. рис. 2).

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.

По рисунку 2 можно судить о том, что параллельная версия оптимизированного алгоритма работает быстрее последовательной реализации на 45%. Дополнительные эксперименты проведены на массивах данных из UCI Machine Learning Repository, результаты носят аналогичный характер.

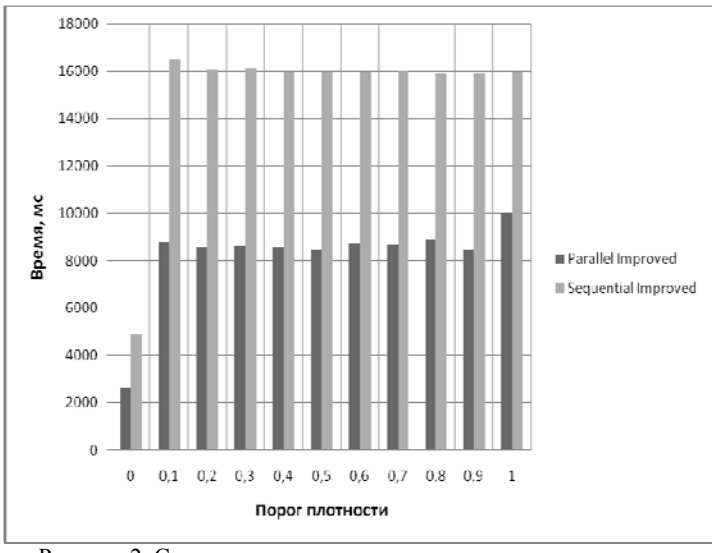

Рисунок 2. Сравнение производительности алгоритмов

Благодарности. Работа выполнена при финансовой поддержке РФФИ (проект № 08-07-92497-НЦНИЛ_а).

Выводы и дальнейшие исследования

Получена эффективная реализация алгоритма бикластеризации на основе решеток замкнутых множеств для случая бинарных объектно-признаковых данных. Для некоторых задач кластеризации возможно применение жадных стратегий, которые позволяют добиться решения близкого к оптимальному за меньшее число итераций. Для предложенного алгоритма это позволит избежать применения параметра ρ_{min}, возможны следующие жадные стратегии: покрытие (некоторой доли) отношения I или множества M (актуально для рекомендательных систем). Необходимо исследование шумоустойчивых свойств алгоритма.

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.

Список литературы

[Barkov et al., 2006] Barkow, S., Bleuler, S., Prelic, A., Zimmermann, P., and E. Zitzler. BicAT: a biclustering analysis toolbox, Bioinformatics, 22(10), 2006.

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

[Besson et al., 2006] Besson, J., Robardet, C., Boulicaut, J.F.: Mining a New Fault-Tolerant Pattern Type as an Alternative to Formal Concept Discovery, In: Schärfe, H., Hitzler, P., Ohhrstrom, P. (eds.) ICCS 2006. LNCS (LNAI), vol. 4068, Springer, 2006.

[Ganter et al., 1999] B. Ganter and R. Wille, Formal Concept Analysis: Mathematical Foundations, Springer, 1999.

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

[Madeira et al., 2004] Sara C. Madeira and Arlindo L. Oliveira, "Biclustering Algorithms for Biological Data Analysis: A Survey IEEE/ACM Transactions on Computational Biology and Bioinformatics, VOL 1, NO. 1, 2004.

[Mirkin, 1996] Mirkin B.G. Mathematical Classification and Clustering. Kluwer Academic Publishers, 1996.