Динамические системы, вып. 28 (2010), 99–112

УДК 519.68 + 681.3

Экстремальная нумерация вершин гиперграфа и задача объектно-признаковой кластеризации

В. Б. Попов

Таврический национальный университет им. В. И. Вернадского,
Симферополь 95007. *E-mail: pvb55@mail.ru*

Аннотация. Рассматривается одна из задач интеллектуального анализа данных (Data Mining) – проблема структурирования данных, полученных в результате обработки множества транзакций. В общем случае проблему можно сформулировать следующим образом. Какой порядок следует задать на множестве строк и столбцов матрицы транзакционных данных, для того чтобы выявить скрытые кластеры данных, обладающие общими признаками и свойствами? Эта проблема является актуальной в случае визуализации транзакционных данных и решения задачи объектно–признаковой кластеризации в различных предметных областях — анализ генетической информации, решение задач анализа интернет–данных (Web–mining), моделирование маркетинговых задач в экономике (Web–marketing) и др.

Ключевые слова: интеллектуальный анализ данных, проблема визуализации данных, объектно-признаковая кластеризация, бикластер, гиперграф, вложение гиперграфа в целочисленную решетку, экстремальная нумерация вершин.

1. Введение

Данная работа касается вопросов интеллектуального анализа данных (Data Mining), таких как проблема визуализации данных и задача выявления кластеров объектов с общими признаками в базе данных, полученной в результате обработки множества транзакций. Обычно собранные транзакционные данные имеют сложную хаотичную структуру, которая мало пригодна для анализа поведения объектов предметной области. Все полученные данные представляются в виде матрицы. Проблему в общем виде можно сформулировать следующим образом.

Дана матрица данных, полученных в процессе наблюдения или изучения какого-либо процесса в конкретной предметной области. Каким образом нужно задать порядок следования строк и столбцов матрицы данных, для того чтобы новая структура была более релевантной для анализа исследуемых данных? Другими словами, необходимо выявить множества объектов, обладающих одинаковыми свойствами при определенных условиях.

Объекты, обладающие одинаковыми признаками в определенных условиях, образуют структуры, называемые *бикластерами*. Под термином *бикластеризация*

в настоящее время понимается довольно широкий круг проблем, методов и алгоритмов. Поэтому для бикластеризации в научной литературе по искусственному интеллекту существует целый ряд тождественных понятий. Можно выделить следующие: совместная кластеризация (simultaneous clustering, см., например, [33]), кокластеризация (co–clustering, см., например, [17]), двухвходовая кластеризация (two–way clustering, [29]), кластеризация подпространства (subspace clustering, [30]), двумерная кластеризация (bi–dimensional) и бокс–кластеризация (box–clustering). Исторически данная проблема впервые описывалась при изучении генной экспрессии (microarray data analysis, gene expression data) [10, 11, 26]. Данные генной экспрессии представляются в виде таблицы, в которой каждый ген соотносится со строкой и каждое экспериментальное условие соотносится со столбцом. Задача заключается в выявлении подмножества генов одинаково себя проявляющих в определенных условиях. В результате получаются структуры, называемые бикластерами. В дальнейшем рассматриваемая проблема нашла свое приложение в анализе интернет–данных (это направление исследований получило название Web–mining [8]). К задачам, которые интенсивно исследуются в данный период времени, относятся следующие проблемы. Поиск групп посетителей со схожими интересами для рекомендательных систем, выявление интернет–сообществ, научных сообществ, решение задач анализа социальных сетей, построение автоматических каталогов и рубрикаторов в информационных системах, поиск документов–дубликатов в корпоративных информационных системах. Актуальным остается применение изучаемых методов в задачах информационного поиска и анализа текстов (Text–mining). В данном случае объектно–признаковая бикластеризация применяется для обнаружения кластеров документов, обладающих сходными свойствами только по нескольким признакам, таким как слова и изображения. Такая информация очень важна для запросов и индексации поисковых интернет–систем. Например, в работе [17] используются методы бикластеризации для одновременного группирования документов и слов. Выявление бикластеров актуально и в анализе категориальных данных, см., например, [31]. Актуальным на данный момент является применение методов бикластеризации в экономических маркетинговых приложениях. Один из примеров — это анализ данных корзины покупателя. В данной работе исследуется проблема выделения в базе транзакционных данных подмножества близких по смыслу данных, т.е. ставится задача разбиения данных на кластеры, обладающих набором определенных свойств или признаков. Назовем такую задачу проблемой объектно–признаковой кластеризации и/или проблемой релевантной визуализации транзакционных данных.

2. Гиперграфовый подход к решению задачи

Обычно база данных DT транзакций представляется бинарной матрицей A такой, что в ячейке (i, j) матрицы A записана единица, если i-ая транзакция содержит j-ый элемент данных, ноль в противном случае. Иногда под базой

транзакционных данных понимают множество ячеек, которые содержат единицы. $\mathrm{DT} = \{(i,j) : a_{i,j} = 1 \wedge a_{i,j} \in \mathrm{A}\}$, $i = 1, 2, \ldots, n, j = 1, 2, \ldots, m$, где A бинарная матрица, соответствующая базе транзакций DT. Пусть T — множество всех транзакций и $|\mathrm{T}|{=}m$ и I — множество элементов данных, которые входят в транзакции, $|\mathrm{I}| = n$. Предполагается, что $B \subseteq \mathrm{A}$ подматрицы (называемые бикластерами), которые определяются как декартово произведение $B \subseteq \mathrm{T} \times \mathrm{I} = \{(\mathrm{i}, \mathrm{j}) : \mathrm{i} \in \mathrm{T} \wedge \mathrm{j} \in \mathrm{I}\}$. Задача двойной кластеризации или объектно–признаковой кластеризации заключается в выделении подматриц–бикластеров. В идеале нужно переставить столбцы и строки матрицы A таким образом, чтобы все элементы любого бикластера сгруппировать вместе, т.е. визуально выделить кластеры, содержащие объекты (строки), обладающие совокупностью свойств (столбцы).

В большинстве задач оптимизации на перестановках размещаемые объекты обладают определенной совокупностью взаимосвязей, которую с большой достоверностью можно описать с помощью графа или гиперграфа. В этом случае задача будет заключаться в поиске такой нумерации вершин графа или гиперграфа, которая доставила бы экстремальное значение некоторой функции, определенной на множестве нумераций. Отметим также, что в большинстве случаев задачи оптимальной нумерации для графов и матриц являются NP–полными. Для них не существует эффективных алгоритмов точного решения. В такой ситуации представляется актуальным исследование частных случаев задач нумерации, поставленных на ограниченных классах графов, либо использование для решения существующей задачи эффективных эвристических алгоритмов. Можно ввести графовое представление задачи объектно-признаковой кластеризации следующим образом. Релевантным представлением такой проблемы является гиперграфовое описание, в котором множество вершин гиперграфа H соответствует множеству объектов, а гиперребра гиперграфа образуют подмножества взаимосвязанных объектов. Например, можно под множеством вершин понимать исследуемые объекты, а в качестве ребер рассматривать семейства объектов, обладающих определенными свойствами и/или признаками. Гиперграф — это такое обобщение неориентированного графа, когда ребрами служат произвольные, а не только одновершинные или двухвершинные, подмножества заданного множества вершин. Теория гиперграфов нашла применение в различных приложениях и современных IT–технологиях. На развитие общей теории гиперграфов повлияли следующие задачи. Выделение наименьшей системы ребер, содержащей все вершины, наибольшей системы ребер попарно без общих вершин, задачи на выбор системы различных вершин представителей для всех ребер, покрытия кликами и др. В случаях равенства некоторых ребер рассматриваются мультигиперграфы.

3. Основные понятия теории гиперграфов. Реализации гиперграфа

Основные понятия приводятся в работах [2, 3, 13, 14, 23, 28, 37].

Определение 1. Гиперграфом $H = (X, U; R)$ называется пара множеств $X =$

$\{x_i, i \in I\}$, $U = \{u_j, j \in J\}$ вместе с двуместным предикатом $R \Leftrightarrow R(x, u)$, определенном при всех $x \in X, u \in U$. Элементы $x \in X$ называются вершинами, элементы $u \in U$ - ребрами, а предикат R инцидентором гиперграфа H [4].

Вершина x и ребро u инцидентны или не инцидентны в H, смотря по тому, истинно или ложно для них высказывание $R = (x, u)$. Под элементом гиперграфа понимают вершину или ребро, т.е. любой элемент множества $X \bigcup U$. Вершина x и ребро u называются инцидентными, если $x \in u$. Для каждого $x \in X$ через $\deg(x)$ обозначается число ребер, инцидентных вершине x; $\deg(x)$ называется степенью вершины x. Степень ребра u — число вершин инцидентных этому ребру, которую будем обозначаеть через $r(u)$.

Определение 2. Гиперграфом, двойственным для $H = (X, U; R)$, называется такой $H^* = (X^*, U^*; R^*) = (U, X; R^*)$, в котором $X^* = U, U^* = X$ и $R^*(u, x) \Leftrightarrow R(x, u)(u \in X^*, x \in U^*)$.

Реберный граф $L(H)$ гиперграфа H определяется следующими условиями: вершины $L(H)$ биективно соответствуют ребрам H и две вершины смежны в $L(H)$ тогда и только тогда, когда соответствующие ребра пересекаются.

Эффективное решение проблемы кластеризации транзакционных данных с помощью графовых моделей предполагает анализ структуры гиперграфа. Для изучения структурных особенностей гиперграфов очень часто используются вспомогательные обыкновенные графы: $L(H)$ — реберный граф, $L^2(H)$ — граф полных реализаций ребер, $K(H)$ — кенигово представление (граф инциденций). Граф $L(H)$ отражает отношение смежности ребер, а граф $L^2(H)$ — смежности вершин гиперграфа H. При этом две вершины $x_1, x_2 \in X$ являются смежными в $H = (X, U)$, если существует ребро $u \in U$, которое содержит обе эти вершины, т.е. $x_1, x_2 \in X(u)$. Аналогичным образом, два ребра $u_1, u_2 \in U$ называются смежными, если $X(u_1) \bigcap X(u_2) \neq \emptyset$. Двойственный гиперграф H^* по определению сохраняет отношение смежности и инцидентности между элементами гиперграфа H. Оттого он наследует все свойства гиперграфа H, основанные на этих отношениях. Справедливо отношение равенства между графами смежности для H и H^* : $L(H) = L^2(H^*), L^2(H) = L(H^*)$. Кенигово представление гиперграфа $H = (X, U; R)$ — двудольный граф $K(H)$, отражающий отношение инцидентности элементов гиперграфа, с множеством вершин $X \bigcup U$ и долями X, U. Этот граф несет полную информацию о гиперграфе H и однозначно его определяет. Кроме того, $K(H) = K(H^*)$. Поэтому некоторые свойства гиперграфа устанавливаются через одноименные свойства графа $K(H)$. Существуют обыкновенные графы, которые представляет гиперграф, — это его реализации. Реализации гиперграфа могут быть получены следующим образом. Реализацией ребра $u \in U$ гиперграфа $H = (X, U; R)$ называется любой связный граф G_u, заданный на множестве вершин X_u. Реализация гиперграфа $H = (X, U; R)$ — граф $G(H) = (X, E)$, полученный объединением некоторых реализаций G_u всех его ребер.

Определение 3. Реализацией гиперграфа H называется любой граф G такой, что $V_G = V_H$, любое ребро графа G содержится в некотором гиперребре гиперграфа H и для любого гиперребра $e \in E_H$ порожденный подграф $G(e)$ является связным.

Из определения следует, что для гиперграфа может существовать несколько различных реализаций. Например, граф $L^2(H)$ – реализация гиперграфа $H = (X, U; R)$, где каждое ребро $u \in U$ представлено полным графом порядка $|X(u)|$. Необходимость построения реализаций гиперграфа с различными свойствами появляется при решении целого ряда практических задач. В [12] доказано, что если некоторое дерево является реализацией гиперграфа, то в этом гиперграфе пересечение любого множества попарно пересекающихся гиперребер не пусто (условие Хелли). С другой стороны, как показано в [15, 23], некоторый граф является графом пересечений некоторого семейства поддеревьев дерева тогда и только тогда, когда этот граф является триангулированным, т. е. из реализуемости гиперграфа деревом следует триангулированность его реберного графа. В [19, 36] доказывается, что триангулированность реберного графа и выполнение условия Хелли для гиперграфа являются также и достаточными для того, чтобы существовала реализация этого гиперграфа деревом. В [6] приведена достижимая верхняя оценка числа ребер в минимальной реализации гиперграфа. Достижимая нижняя оценка числа ребер в минимальной реализации найдена в [18]. В [6] найдены достижимые оценки минимального числа ребер реализации, подграф которой, индуцированный произвольным гиперребром, содержит гамильтонову цепь. В [5] рассматриваются различные виды гамильтоновых реализаций. Исследуются вопросы построения реализаций, имеющих гамильтонов цикл и определенным образом согласованных с реберным графом заданного гиперграфа. Доказана NP–полнота задачи построения таких реализаций в общем случае, выделены полиномиально разрешимые случаи. Для деревьев, хордовых графов имеются быстрые алгоритмы эффективного поиска, декомпозиции, элиминации и поэтому они широко используются в различных приложениях.

4. Проблемы экстремальной нумерации вершин гиперграфа

В большинстве задач комбинаторной оптимизации на перестановках размещаемые объекты обладают определенной совокупностью взаимосвязей, которую часто можно задать с помощью графа или гиперграфа. В этом случае задача состоит в поиске такой нумерации вершин графа или гиперграфа, которая доставила бы экстремальное значение некоторому функционалу, определенному на множестве нумераций. В общем случае большинство задач экстремальной нумерации для графов и матриц являются NP–полными, поэтому для них не существует эффективных полиномиальных алгоритмов точного решения [1]. В этой ситуации представляется актуальным исследование частных случаев задач нумерации, поставленных на ограниченных классах графов либо использование различных эвристик, например эволюционных стратегий.

Определение 4. Нумерацией n-вершинного графа $G = (X, E)$ называется взаимно однозначное отображение $\pi : X \to I$, где $I = \{i_1, i_2, \ldots, i_n\}$ — множество целых чисел, а π в общем случае принадлежит какому-то классу функций F.

Общая задача о нумерации ставится так: для заданного графа G найти такую нумерацию $\pi \in F$, чтобы функционал $\Phi(G, \pi)$ принимал наименьшее значение.

Среди задач нумерации можно выделить следующие наиболее известные.

Задача 1. Задача о ширине графа.

Задача о построении минимальной по ширине нумерации называется задачей о ширине графа.

Шириной графа $G = (X, E)$ при нумерации π называется число

$\Phi(G, \pi) = \max\{|\pi(x_i) - \pi(x_j)| : (x_i, x_j) \in E\}$, а шириной графа $G = (X, E)$ число $\Phi(G) = \min\limits_{f \in F}(\Phi(G, \pi))$.

Эта проблема в [1] описывается под номером ТГ40.

Условие. Заданы граф $G = (V, E)$ и положительное целое число $K \leq |V|$.

Вопрос. Существует ли линейное упорядочение множества V, такое, что его ширина не более K? Другими словами, существует ли взаимно однозначная функция $f : V \to \{1, 2, \ldots, |V|\}$, такая, что для всех $(u, v) \in E$ выполнено соотношение $|f(u) - f(v)| \leq K$? Задача соответствует минимизации ширины симметрической матрицы путем одновременных перестановок строк и столбцов.

Задача 2. Задача о профиле графа. Задача о профиле чаще всего возникает в вычислительной математике при обработке разреженных матриц, например, когда надо решить профильным методом систему линейных уравнений $Ax = b$, где A — симметричная разряженная матрица порядка $n \times n$. Пусть в каждой строке i $a_{ii} \neq 0$ и j_i — позиция первого ненулевого элемента в строке. Число $\beta_i = i - j_i = i - \min\{j : a_{ij} \neq 0\}$ называют шириной строки i. Таким образом, первый ненулевой элемент строки i находится левее главной диагонали на β_i позиций. Оболочка матрицы A — это множество элементов $a_{ii} \neq 0$, для которых $0 < i - j < \beta_i$. В строке i оболочке принадлежат все элементы со столбцовыми индексами $j_{\min}(i)$ до $i - 1$, всего β_i элементов. Диагональные элементы не входят в оболочку. Профиль матрицы определяется как число элементов в ее оболочке: $profile\,(A) = \sum_i \beta_i$ [7]. При использовании профильной схемы хранения матрицы достигается значительная экономия памяти компьютера. Профиль меняется при перестановках строк и столбцов матрицы. Меньший профиль означает меньшую память и меньшее число операций в вычислениях, выполняемых с матрицей. Другими словами, решая задачу о профиле, необходимо так симметрично переставить строки и столбцы матрицы A, чтобы результирующая матрица имела минимальный профиль, т.е. необходимо найти матрицу перестановки Q такую, чтобы профиль матрицы $p(QAQ^T)$ был минимальным. Задача о профиле формулируется в графовой форме следующим образом. Симметричной матрице $A(n \times n)$ ставится в соответствие такой неориентированный граф $G = (X, E)$, что $X = \{x_1, x_2, \ldots, x_n\}$ и $E = \{(x_i, x_j) : i \neq j \wedge a_{ij} \neq 0\}$. Тогда $p(A) = \sum_i^n \beta_i = \sum_i^n (i - \min\limits_{x_j \in N[x_i]} j)$, где

$N[x_i]$ — закрытая окрестность вершины x_i графа $G = (X, E)$. Матрице перестановки Q соответствует такое взаимно однозначное отображение (нумерация вершин) $\pi : X \to \{1, 2, \ldots, n\}$, что $p(QAQ^T) = \sum_{x \in X}(\pi(x) - \min\limits_{y \in N[x]} \pi(y))$. При нумерации π профильная ширина вершины x графа определяется следующим образом: $\beta_\pi(x) = \pi(x) - \min\limits_{y \in N[x]} \pi(y)$.

Профилем графа $G = (X, E)$ при нумерации π называется число $p(G, \pi) = \sum_{x \in X} \beta_\pi(x) = \sum_{x \in X}(\pi(x) - \min\limits_{y \in N(x)} \pi(y))$, где $N[y]$ — закрытая окрестность вершины x графа $G = (X, E)$. Профилем графа $G = (X, E)$ называется число $p(G) = \min p(G, \pi)$, где π пробегает F. Нумерация π называется оптимальной профильной нумерацией, $p(G, \pi) = p(G)$.

Теорема 1. *Задача о профиле графа является NP–трудной.*

Очевидно, что задача о профиле, т.е. задача о нахождении профиля $p(G)$ графа G, эквивалентна задаче пополнения до графа интервалов, которая является NP-полной если ограничиться реберными графами (ТГ35 в [1]).

Условие. Заданы граф $G = (V, E)$ и неотрицательное целое число K.

Вопрос. Существует ли множество E', содержащее E, такое, что $|\mathrm{E}' - E| \leq K$ и граф $G = (V, \mathrm{E}')$ является графом интервалов? К этой проблеме сводится задача *Оптимальное линейное упорядочение*. Для $K = 0$ задача разрешима за полиномиальное время. Задача о нахождении профиля дерева может быть разрешена за полиномиальное время.

Нумерации вершин гиперграфа порождают следующие проблемы.

Задача 3. Задача о длине гиперграфа.

Пусть $H = (X, U; R)$ – некоторый гиперграф с множеством вершин $X = \{x_1, x_2, \ldots, x_n\}$ и множеством ребер $U = \{u_1, u_2, \ldots, u_m\}$.

Нумерацией n-вершинного гиперграфа $H = (X, U; R)$ называется взаимно однозначное отображение $\pi : X \to I$, где $I = \{i_1, i_2, \ldots, i_n\}$ — нумерующая последовательность из целых чисел. Длиной гиперграфа $H = (X, U; R)$ относительно нумерации π называется число $d(H, \pi) = \sum[\max\limits_{x_i \in u}(\pi(x_i)) - \min\limits_{x_j \in u}(\pi(x_j))]$, где $u \in U$, а длиной гиперграфа $H = (X, U; R)$ — число $d(H) = \min d(H, \pi)$, где π пробегает F.

Задача о длине гиперграфа возникает в ряде практических задач при автоматизации проектирования интегральных схем и компонент компьютеров (САПР), в биологии при решении задач о наследственности, а также в различных областях теории информации.

Задача 4. (Hypergraph Ordering Problem, HOP) Пусть задан гиперграф $H = (X, U; R)$. HOP проблема заключается в поиске порядка σ на множестве вершин X такого, что минимизируется величина $\arg\min_\sigma \sum_{u \in U}(\max\limits_{v_j \in u} \sigma(v_j) - \min\limits_{v_i \in u} \sigma(v_i))$.

Эта проблема сводится к проблеме минимизации линейного порядка [25], которая является NP–полной.

Задача 5. Hypergraph Optimal Linear Arrangement (HOLA). HOLA проблема формулируется следующим образом. Дан гиперграф $H = (X, U; R)$ и множество весов $W = \{w_1, w_2, \ldots, w_k\}$. Пусть $X = \{1, 2, \ldots, n\}$ — множество вершин, $U = \{U_1, U_2, \ldots, U_k\}$ — семейство подмножеств множества X. Линейное упорядочение вершин X представляется перестановкой $\pi_1, \pi_2, \ldots, \pi_n$ множества $1, 2, \ldots, n$. Вершина i связывается с $\pi(i)$ позицией в линейном порядке. Стоимостью перестановки π называется величина $C(\pi) = \sum_{i=1}^{k} w_i \max\limits_{q,l \in U_i} \{|\pi(q) - \pi(l)|\}$.

В случае, когда вместо гиперграфа рассматривается обыкновенный граф, проблема называется Graph Optimal Linear Arrangement (GOLA). Дан граф $G = (X, E)$ с весом $w(i, j)$ ребра $(i, j) \in E$. Требуется найти перестановку, доставляющую минимум величине $C(\pi) = \sum_{(i,j) \in E} w(i, j)\,|\pi(i) - \pi(j)|$. Граф всегда рассматривается как частный случай гиперграфа, мощность каждого ребра которого равна 2. Другими словами, граф 2-униформный гиперграф. В случае, когда все $w(i, j) = 1$ проблема известна как Optimal Linear Arrangement (OLA) problem. OLA проблема является широко известной NP–трудной задачей. HOLA проблема сводится к OLA проблеме и также является NP–трудной. Для GOLA проблемы, если граф корневое дерево, то решение можно найти за время $O(|X| \log |X|)$ [9]. Для проблемы OLA, если граф неориентированное дерево, то оценка решения равна $O(|X|^{2.2})$ [35]. GOLA проблема изучалась в [9]. Эвристики для HOLA проблем рассматривались в [34, 24]. HOLA проблема с единичными весами изучалась в работе [27]. Одной из первых работ, связанных к экстремальными нумерациями на вершинах графов, является [25]. В этой работе проблема формулируется следующим образом. Given a graph $G = (V, E)$, the minimum linear arrangement problem finds the order σ of V that can minimize the graph_cost(G, σ)=$\sum_{(i,j) \in E} |\sigma(v_i) - \sigma(v_j)|$. Проблема описывается в работах [21, 22]. В [20] проблема упоминается под номером GT42. В [1] проблема описывается под номером [ТГ 42] — Оптимальное линейное упорядочение.

Условие. Заданы граф $G = (V, E)$ и положительное целое число k.

Вопрос. Существует ли взаимно-однозначная функция $f : V \to \{1, 2, \ldots, |V|\}$, такая, что $\sum_{(u,v) \in E} |f(u) - f(v)| \le k$? Задача NP–полна для двудольных графов и разрешима за полиномиальное время, если G — дерево. В [16] проблема описана в GT44. Тут же отмечается, что в [32] приводится алгоритм с оценкой $O(\log |V|)$.

Надо отметить, что в настоящее время достаточно активно изучаются инварианты графов и гиперграфов, определяемые через экстремальные по различным критериям нумерации вершин. Интерес к подобным инвариантам связан, в частности, с тем, что они естественным образом возникают в таких приложениях, как искусственный интеллект (в задачах Data Mining, Web–mining), экономика, архитектура компьютеров, находят применение в теории программирования.

5. Спецификация модели объектно-признаковой кластеризации

Опишем теперь проблему объектно-признаковой кластеризации в виде оптимальной линейной нумерации вершин гиперграфа.

В идеальном случае подматрицы, соответствующие кластерам, должны быть полностью заполнены элементами (в случае бинарной матрицы единицами). На практике в исходных $\{0,1\}$–данных, имеющих объектно-признаковую природу, могут присутствовать ошибочные значения — пропущенные объекты/признаки или напротив лишние, либо могут появляться различные шумы. В практических приложениях можно использовать понятие плотности бикластера. Это может быть количество единиц или отношение количества единиц к общему числу элементов бикластера.

Здесь под бикластером понимается подмножество вершин гиперграфа, имеющих минимальную оценку при оптимальной нумерации вершин гиперграфа.

Вложение гиперграфа в целочисленную решетку и оценка множества ребер гиперграфа. Пусть $H = (X, U; R)$ — некоторый гиперграф с множеством вершин $X = \{x_1, x_2, \ldots, x_n\}$ и семейством подмножеств ребер $U = \{u_1, u_2, \ldots, u_m\}$ гиперграфа. Обозначим через N^2 множество упорядоченных пар $\{(x, y) : x, y \in N\}$, образующих двумерную целочисленную решетку. Вложением вершин гиперграфа в решетку будем называть инъективное отображение $\pi : X \to \mathrm{N}^2$. Вершины гиперграфа отображаются в узлы решетки. Если $v \in \mathrm{N}^2$, то назовем $x(v), y(v)$ координатами узла решетки.

Пусть q множество бикластеров. Бикластер представляет собой прямоугольную подматрицу $Z \times Y \subseteq X \times U, Z \subseteq X, Y \subseteq U$. Количество бикластеров q зависит от их структуры и количества элементов, входящих в них.

Дадим оценку бикластера B_l гиперграфа при отображении (нумерации) π следующим образом:

$$C(B_l) = (\max\{x(\pi(u_j)) : u_j \in B\} - \min\{x(\pi(u_k)) : u_k \in B\}) +$$
$$+(\max\{y(\pi(u_i)) : u_i \in B\} - \min\{y(\pi(u_l)) : u_l \in B\}).$$

Оценку гиперграфа относительно отображения π обозначим как $S(H, \pi)$.

$$S(H, \pi) = \sum_{l \in \{1,2,\ldots,s\}} C(B_l) =$$

$$= \sum_{u \in B}((\max\{x(\pi(u_j)) : u_j \in B\} - \min\{x(\pi(u_k)) : u_k \in B\}) +$$
$$+(\max\{y(\pi(u_i)) : u_i \in B\} - \min\{y(\pi(u_l)) : u_l \in B\})).$$

Оценка подматрицы – это сумма разностей наибольших и наименьших координат вершин бикластера B_l по длине и ширине.

Оценкой гиперграфа называется величина равная $S(H) = \min S(H, \pi)$ для всех возможных отображений π. Будем говорить, что π^* оптимальное отображение, если $S(H) = S(H, \pi^*)$.

Определение 5. Для данного гиперграфа $H = (X, U)$ под проблемой оптимального вложения гиперграфа в решетку будем понимать нахождение такого отображения $\pi : X \to \mathrm{N}^2$, которое минимизирует величину $S(H, \pi)$:

$$S(H,\pi) = \arg\min_{\pi} \sum_{u \in B} ((\max\{x(\pi(u_j)) : u_j \in B\} - \min\{x(\pi(u_k)) : u_k \in B\}) +$$

$$+ (\max\{y(\pi(u_i)) : u_i \in B\} - \min\{y(\pi(u_l)) : u_l \in B\})).$$

Выявляемые структуры бикластеров подматриц. Среди современных наиболее известных алгоритмов существуют не только такие, которые находят один бикластер, но и порождающие множество бикластеров. Бикластеры, входящие в такое множество, могут иметь различную структуру. Наиболее известные в практических приложениях.

1. Бикластеры, исключающие пересечения по строкам и столбцам (прямоугольные диагональные блоки после переупорядочивания строк и столбцов). В этом случае матрица A базы транзакционных данных имеет блочно–диагональную структуру.
2. Неперекрывающиеся бикластеры со структурой шахматной доски.
3. Бикластеры с перекрытием. Блоки матрицы имеют зацепление по строкам и столбцам.
4. Неперекрывающиеся бикластеры с древесной структурой.
5. Перекрывающиеся бикластеры с иерархической структурой.
6. Произвольно расположенные перекрывающиеся бикластеры.
7. Неперекрывающиеся не исключающие пересечения бикластеры.
8. Одиночный бикластер.
9. Бикластеры, имеющие зацепление по строкам матрицы данных (только соседние блоки).
10. Бикластеры, имеющие зацепление по столбцам матрицы данных (только соседние блоки).

В последних двух случаях матрица имеет квазиблочную структуру с зацеплением блоков (бикластеров) по строкам или столбцам.

Разбиение транзакционных данных на блоки. Для приложений интересны случаи, когда матрица обладает блочно–диагональной структурой или квазиблочной структурой, а также бикластеры с перекрытием по строкам и столбцам. Последняя структура имеет на практике наибольшее применение.

Пусть DT = (T, I) — база данных транзакций, где T — множество транзакций, I — множество элементов данных входящих в транзакции. T = X, I = U.

Пусть множество бикластеров B = $\{B_1 = (Z_1 \times Y_1), B_2 = (Z_2 \times Y_2), \ldots, B_q = (Z_q \times Y_q)\}$ покрывает матрицу A, которая соответствует базе транзакционных данных. Решение проблемы получения кластеров сводится к оптимизации упорядочения вершин двух гиперграфов. Гиперграф $H_1 = (V_1, E_1; R_1)$ и гиперграф $H_2 = (V_2, E_2; R_2)$, где $V_1 = T$, $E_1 = Z$, $Z = \{Z_1, Z_2, \ldots, Z_q\}$, $V_2 = I$, $E_2 = Y, Y = \{Y_1, Y_2, \ldots, Y_q\}$. Отметим, что в данном случае происходит огрубление исходного гиперграфа, некоторые из вершин могут не входит ни в одно из ребер гиперграфов H_1, H_2. Оптимальное решение проблемы соответствует минимизации величины

$$C(\mathrm{B}, \pi_Z, \pi_Y) = \sum_{k=1}^{q} (\max_{z_u \in Z_k} \pi_Z(z_u) - \min_{z_v \in Z_k} \pi_Z(z_v)) + \sum_{k=1}^{q} (\max_{y_u \in Y_k} \pi_Y(y_u) - \min_{y_v \in Y_k} \pi_Y(y_v)) =$$

$$= \sum_{Z_k \in E^1} (\max_{z_u \in Y_k} \pi_Z(z_u) - \min_{z_v \in Z_k} \pi_Z(z_v)) + \sum_{Y_k \in E^2} (\max_{y_u \in Y_k} \pi_Y(y_u) - \min_{y_v \in Y_k} \pi_Y(y_v)) =$$

$$= C(H_1, \pi_Z^*) + C(H_2, \pi_Y^*),$$

где $C(H, \pi^*)$ - оценка гиперграфа при оптимальной нумерации π^*, где π_Z^*, π_Y^* - соответствующие нумерации гиперграфов. Таким образом, задача сводится к поиску минимальных нумераций вершин двух гиперграфов H_1 и H_2. Очевидно, что эти проблемы можно рассматривать независимо друг от друга. Покажем, что минимизация величины $C(\mathrm{B}, \pi_Z, \pi_Y)$ является NP–трудной задачей. Эта задача сводится к проблеме минимизации линейного упорядочения вершин графа $G = (V, E)$. Пусть $S = (S_1, S_2, \ldots, S_q)$ какое-либо разбиение матрицы A на q блоков, образующих прямоугольные подматрицы. В дальнейшем под блоками понимаются блоки $S_j, j = 1, 2, \ldots, q$, состоящие из последовательных строк матрицы A. Такие блоки будут соответствовать бикластерам матрицы. В результате этого любой блок $S_j, j = 1, 2, \ldots, q$ можно однозначно определить с помощью индекса i_j последней строки, входящей в этот блок, и с помощью индекса последней строки i_{j-1} предшествующего блока S_{j-1}. Данное разбиение $S = (S_1, S_2, \ldots, S_q)$ матрицы A на бикластеры можно записать в виде последовательности последних строк, входящих в блоки — $\{i_1, i_2, \ldots, i_q\}$. Построим неориентированный граф $G = (V, E)$, соответствующий разбиению матрицы A на блоки, следующим образом. Поставим в соответствие каждой строке i матрицы A некоторую вершину v_i графа G. Вершины v_i и $v_j, (i < j)$ соединяются ребром $e = (v_i, v_j) \in E$, если строки i, j могут быть последними в двух соседних блоках. Если каждая из строк матрицы может быть последней строкой первого блока кластера, то можно ввести для удобства доминирующую вершину v_0. Вершина v_0 не всегда будет соединяться ребром с каждой вершиной, на практике не каждая строка может представлять первый блок. Разбиению $S = (S_1, S_2, \ldots, S_q)$, которое можно записать с помощью индексов последних строк блоков $\{i_1, i_2, \ldots, i_q\}$, соответствует в графе G путь $\{v_0, v_{i_1}, v_{i_2}, \ldots, v_{i_q}\}$. Очевидно, что задача разбиения матрицы на блочные структуры является в общем

случае NP–полной проблемой. Таким образом, проблема гиперграфа сводится к проблеме минимальной линейной нумерации вершин графа $G = (V, E)$. Сформулируем проблему следующим образом. Пусть дан граф $G = (V, E)$ (в общем случае граф взвешенный), найти порядок π, который бы минимизировал следующую величину $G_C(G, \pi) = |\pi(v_i) - \pi(v_j)|$. Очевидно, что эта проблема является частным случаем проблемы гиперграфа, когда гиперграф представляет из себя 2-униформный гиперграф. В общем случае разбиение данных на блоки не единственно. Поэтому возникает задача поиска оптимального разбиения информации на бикластеры.

6. Результаты и дальнейшие исследования

В работе предлагается сведение проблемы выявления бикластеров в базе транзакционных данных к оптимизационной задаче оптимальной нумерации вершин гиперграфа. Экстремальные задачи на графах являются NP–трудными, а значит, актуальным является поиск новых алгоритмов решения цитируемых проблем. Перспективным является анализ некоторых реализаций гиперграфа (древовидная реализация, гамильтонов путь и др.), для которых существуют полиномиальные алгоритмы решения рассматриваемых проблем. В качестве возможных направлений дальнейших исследований можно выделить следующие. Актуальным остается изучение свойств топологической структуры бикластеров, в частности, определение степени перекрытия бикластеров. Для различных структур актуальным является оценка максимального и минимального числа элементов внутри q блоков матрицы данных, а также элементов оказавшихся вне блоков бикластеров, что в свою очередь позволит дать верхнюю и нижнюю оценку числа бикластеров матрицы данных. Надо отметить, что задача оценки числа всевозможных разбиений матрицы на блоки бикластера является NP–трудной, поэтому актуальным остается разработка эффективных алгоритмов, в частности, эвристических эволюционных алгоритмов. Очевидно, что выявление блоков бикластеров матрицы транзакционных данных в общем случае не единственно. Естественным образом возникает проблема оптимального разбиения матрицы на подматрицы, соответствующие искомым бикластерам. Важным остается оценка возможности определения порядка на бикластерах, а также анализ их алгебраической структуры.

Список цитируемых источников

1. *Гэри М.* Вычислительные машины и труднорешаемые задачи: Пер. с англ. / М. Гэри, Д. Джонсон — М.: Мир, 1982. — 416 с.
2. *Евстигнеев В. А.* Толковый словарь по теории графов / В. А. Евстигнеев, В. Н. Касьянов — Новосибирск: Наука, 1999. — 187 с.
3. *Емеличев В.А.* Лекции по теории графов / В.А. Емеличев и др. — М. : Наука, 1990. — 384 с.
4. *Зыков А.А.* Гиперграфы / А.А Зыков // Успехи математических наук. — 1974. т. 29. вып. 6. — С. 89-154.

5. *Крикун В. С.* О гамильтоновой реализуемости гиперграфов / В. С. Крикун — Минск, 1986. — (Препринт / Институт математики АН БССР; № 8).
6. *Пилипосян Т. Э.* Оценки числа реализаций гиперграфа / Т. Э. Пилипосян // Математические вопросы кибернетики и вычислительной техники. — 1985. — т.14. — С. 26 - 33.
7. *Писсанецки С.* Технологии разреженных матриц: Пер. с англ. / С. Писсанецки — М.: Мир. 1988. — 410 с.
8. *Попов В.Б.* Эволюционные алгоритмы в задачах персонализации контента / В.Б. Попов // Теория и практика экономики и предпринимательства.: Материалы 5 межд. научной конференции. — Симферополь: ТНУ, 2008. — С. 59.
9. *Adolphson D.* Optimal Linear Ordering / D. Adolphson, T.C. Hu // SIAM J. Appl. Math. — 1973. — N25. — P. 403-423.
10. *Aguilar-Rui J. S.* Biclastering of gene expression data with Evolutionary Computation / J. S. Aguilar-Ruiz, F. Divina // IEEE Transactions on Knowledge and Data Engineering — 2006. — Volume 18, Issue 5 (May 2006). — P. 590-602.
11. *Ben-Dor A.* Discovering local structure in gene expression data: the order-preserving submatrix problem / A. Ben-Dor et al. // In Proceedings of the 6th Annual International Conference on Computational Biology.: ACM Press NY, USA, 2002. — P. 49-57.
12. *Berge C.* Graphs and Hypergraphs / C. Berge — Amsterdam: North-Holland, 1973. — 450 p.
13. *Bodlaender H. L.* Discovering Treewidth / H. L. Bodlaender // Proceedings of SOFSEM 2005. — Springer-Verlag, Lecture Notes in Computer Science, 2006. — vol. 3381 — P. 1-16.
14. *Bodlaender H.L.* Discovering Treewidth / H. L. Bodlaender // Institut of information and computing science. Utrecht university technical report UU-CS-2005-018. — [http://www.cs.uu.nl/research/techreps/repo/CS-2005/2005-18.pdf]
15. *Buneman P. A* characterization of rigid circuit graphs / P. A. Buneman // Discrete Mathematics. — 1974. — V. 9. № 3. — P. 10-18.
16. *Crescenzi P.* Approximation on the web: A compendium of NP optimization problems / P. Crescenzi, V. A. Kann // Lecture Notes in Computer Science. — 1997. — Volume 1269/1997. — P. 111-118.
17. *Dhillon I. S.* Co-clustering documents and words using Bipartite Spectral Graph Partitioning / I. S. Dhillon // Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining — NY: ACM, 2001. — P. 269-274.
18. *Du D. Z.* An optimization problem on graphs / D. Z. Du // Discrete Applied Mathematics. — 1986. — V. 14. — P. 101-104.
19. *Flament C.* Hypergraphs arbores / C. Flament // Discrete Mathematics. — 1978. — V. 21, № 3. — P. 223-227.
20. *Garey M.R.* Computers and Intractability. A guide to the theory of NP-completeness / M.R. Garey, D.S. Jonhson — NY: Freeman and Company, 1978. — 260 p.
21. *Garey M.R.* Some simplified NP-problems / M.R. Garey, D.S. Johnson, I. Stockmeyer // In STOC-74; Proceedings of the sixth annual ACM symposium on Theory of computing. — NY: ACM, 1974. — P. 47-63
22. *Garey M.R.* Some simplified NP-complete graph problems / M.R. Garey, D.S. Johnson, I. Stockmeyer // Theor. Computing Science. — 1976. — №1. — P. 237-267.

23. *Gavril F.* The intersection graphs of subtrees in trees are exactly the chordal graphs / F. Gavril // Journal Combinatorial Theory. — 1974. — V. B-16. — P. 47-56.
24. *Hanan M.* Placement Techniques / M. Hanan, J.M. Kurtzberg // Design Automation of Digital Systems, Theory and Techniques. — 1972. — Vol 1, Prentice Hall, Englewood Cliffs, New Jersey. — p. 23-45.
25. *Harper L.H.* Optimal assignments of numbers to vertices / L.H. Harper // Journal of the Society for Industrial and Applied Mathematics. — 1964. — №12. — P. 131-135.
26. *Jiang D.* Cluster Analysis for Gene Expression Data: A Survey / D. Jiang, C. Tang , A. Zhang — NY: Department of Computer Science and Engineering State University of New York at Buffalo, 2008. — 380 p.
27. *Kang S.* Linear Ordering and Application to Placement / S. Kang // Proceedings 20th DAC. — NY: Plenum Press,1983. — P. 457-464.
28. *Karp R.* Reducibility Among Combinatorial Problems / R. Karp // Proceedings of a Symposium on the Complexity of Computer Computations. — NY: Plenum Press,1972. — P. 85-103.
29. *Pascual-Montano A.* Two-way clustering of gene expression profiles by sparse matrix factorization / A. Pascual-Montano et al. // Computational Systems Bioinformatics Conference. Workshops and Poster Abstracts. — Stanford, California: IEEE. 2005. — P. 103-104.
30. *Parsons L.* Subspace Clustering for High Dimensional Data: A Review / L.Parsons, E. Haque, H. Liu // SIGKDD Explorations. 2004. — Volume 6, Issue 1 — P. 90-105.
31. *Pensa R. G.* A bi-clustering framework for categorical data / R. G. Pensa, C. Robardet, J.F. Boulicaut // PKDD 2005. LNCS (LNAI). — Heidelberg: Springer, 2005. — vol. 3721, — P. 643-650.
32. *Raos W.* New approximation techniques for some ordering problems / W. Raos, A.W. Richa // Proc. 9th Ann. ACM-SIAM Symp. On Discrete Algorithm. — NY: ACM-SIAM, 2006. — P. 211-218.
33. *Saka E.* Simultaneous Clustering and Visualization of Web Usage Data Using Swarm-Based Intelligence / E. Saka, O. Nasraoui // Tools with Artificial Intelligence.: ICTAI '08. 20th IEEE International Conference on. 2008 — NY: IEEE, 2008. — P. 539 – 546.
34. *Schuler D. M.* Clustering and Linear Placement / D.M. Schuler, E.G. Ulrich // Proceedings 9th DAC. — NY: IEEE Press, 1972. — P. 57-62.
35. *Shiloach Y. A.* Minimum Linear Arrangement Algorithm for Undirected Trees / Y. A. Shiloach // SIAM J. Computing. — 1979. — Vol 8, No 1, Feb 1979, — P. 15-32.
36. *Slater P.* A characterization of SOFT hypergraphs / P. Slater // Canadian Math. Bull. — 1978. — V. 21, № 3. — P. 335 - 337.
37. *Yannakakis M.* Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs / M. Yannakakis // SIAM J. Comput. — 1984. — IS, № 3. — P. 566 - 579.

Получена 01.08.2010