Информационные процессы, Том 15, № 2, 2015, стр. 215–248.

О комбинаторной кластеризации: обзор литературы, методы, примеры

Марк Ш. Левин

Институт проблем передачи информации, Российская академия наук
Большой Каретный пер. 19, Москва 127994, Россия
email: mslevin@acm.org

Поступила в редколлегию 28.05.2015

Аннотация—В статье рассмотрены задачи кластеризации с комбинаторной точки зрения. Представлен системный обзор. Список рассмотренных опросов включает следующее: (1) анализ литературы о базовых комбинаторных методах и кластеризации данных/сетей очень большой размерности; (2) характеристики качества решений кластеризации; (3) многокритериальные модели кластеризации; (4) методы кластеризации на основе графов (на основе задачи минимального покрывающего дерева, модели кластеризации на основе выделения клик/квази-клик, корреляционная кластеризация, выделение сетевых сообществ): (5) методы "быстрой" кластеризации. В основном, материал направлен на сетевые приложения. Числовые примеры иллюстрируют модели, методы и приложения.

КЛЮЧЕВЫЕ СЛОВА: кластеризация, классификация, комбинаторная оптимизация, многокритериальное принятие решений, эвристики, сетевые приложения

1. ВВЕДЕНИЕ

Важность задач кластеризации/классификации существенно увеличилась в последние годы. Среди современных трендов можно указать следующие: (а) кластеризация данных очень большой размерности, (б) исследование и применение методов "быстрой" кластеризации, (в) модели и методы многокритериальной кластеризации. В области сетей связи и беспроводных сенсорных сетей основными типами прикладных задач базирующихся на кластеризации являются следующие: (1) кластеризация и позиционирование/размешение в сетях связи/сенсорных сетях (узлы, станции, рестрасляторы) [23, 36, 103, 211], (2) проектирование иерархических структур сетей, построение сетевых топологий [42, 43, 81, 104, 123], выделение/назначение головных узлов в сетевых кластерах [211]), (3) маршрутизация на основе кластеризации, включая построение многошаговых маршрутов [89, 93, 100, 103, 211], (4) построение иерархических методов доступа к каналам связи [31, 89, 105, 103, 116].

В данная стать посвящена следующим вопросам в области комбинаторной кластеризации: (1) анализ литературы по базовым комбинаторным методам/моделям, включая кластеризацию данных/сетей очень большой размерности; (2) характеристики качества решений кластеризации; (3) модели многокритериальной кластеризации; (4) модели кластеризации на основе графов (на основе минимального покрывающего дерева, на основе задачи о клике, модели корреляционной кластеризации, на основе сетевых сообществ): (5) методы "быстрой" кластеризации. Многие числовые примеры иллюстрируют задачи, модели и методы Кратко описаны две прикладные задачи на сетях: (1) кластеризация на основе использования узлов разного типа, (2) планирование системы связи с много-лучевой антенной. Статья основана на материале электронного препринта [124]. В таблице 1 представлен список основных современных подходов/направлний в области комбинаторной кластеризации.

Таблица 1. Комбинаторные подходы к кластеризации

Ном.	Подход, алгоритмическая схема	Источник
1.	Обзоры :	
1.1.	Общие вопросы	[67, 96, 134]
1.2.	Кластеризация на основе графов	[172]
1.3.	Приближенные методы разбиения графов	[58]
1.4.	Метод кросс-энтропии для кластеризации/разбиения	[112, 169, 184]
1.5.	Групповые технологии в машиностроении	[72, 175, 181]
1.6.	Алгоритмы интегрирования решений	[195]
1.7.	Многокритериальная классификация, сортировка	[168, 214]
2.	Основные задачи комбинаторной оптимизации:	
2.1.	Метод минимального поркывающего дерева	[75, 138, 154, 160, 186, 197, 201]
2.2.	Кластеризация в виде разбиения	[11, 25, 44, 52, 58, 180]
2.3.	Кластеризация на основе назначения/размещения	[71]
2.4.	Сопоставление графов	[173]
2.5.	Выделение доминирующего множества	[36, 81, 158, 211]
2.6.	Кластеризация на основе покрытия	[6, 137, 171]
2.7.	Кластеризация на основе клик	[6, 21, 30, 54, 70, 109, 176]
2.8.	Кластеризация на основе сетевых сообществ	[4, 144, 146, 163, 202]
3.	Корреляцинная кластеризация	[2, 18, 51, 111, 183]
4.	Кластеризация на основе графов с перекрытием	[62]
5.	Сегментация	[107]
6.	Модификация графа кластеров	[176, 177]
7.	Многокритериальное принятие решений, сортировка	[65, 165, 166, 214]
8.	Кластеризация на основе консенсуса:	
8.1.	Консенсус на основе методов голосования	[12, 170]
8.2.	Консенсус разбиений	[76]
9.	Алгоритмические схемы:	
9.1.	Переборные методы:	
9.1.1.	Методы типа ветвей и границ	[38]
9.1.2.	Динамическое программирование	[210]
9.2.	Локальная оптимизация:	
9.2.1.	Алгоритмы "отжига"	[29, 149, 174]
9.2.2.	Алгоритмы "tabu search"	[149, 182]
9.2.3.	"Муравьинные" алгоритмы	[99, 204]
9.2.4.	Методы "PSO"	[35, 185, 193]
9.2.5.	Поиск на основе изменяемых окрестностей (VNS)	[82, 83, 84, 85]
9.3.	Эволюционные методы	[14, 47, 90, 153, 190]
9.4.	Гипер-эвристики	[41, 114, 189]

2. О КЛАСТЕРИЗАЦИИ ДАННЫХ/СЕТЕЙ ОЧЕНЬ БОЛЬШОЙ РАЗМЕРНОСТИ

В последние годы значительно повысилась важность задач кластеризации в базах данных очень большой размерности и при анадизе/моделировании в сетях большой размерности: (i) кластеризация множеств данных большой размерности [21, 91, 94, 178]; (ii) выделение сообществ в сетях большой размерности [39, 73, 87, 88, 117, 162, 208]; (iii) выделение сообществ в свех-больших сатях [22, 196]; (iv) динамический анализ эволюции сообществ в больших сетях [88]. В таблице 2 приведена классификация систем данных/сетей по объему хранимых и обрабатываемых данных (т.е., по размерности).

Таблица 2. Уровни информационных систем/сетей

Ном.	Тип изучаемых множеств данных/сетей	Число объектов/ узлов сети	Примеры приложений	Источники
1.	Упрощенные множества данных/сети (малые группы)	$\sim$ 10...60	(i) группы студентов, (ii) спортивные клубы, (iii) отделы, (iv) Веб страницы, (v) ассортимент товаров	[212]
2.	Простые множества данных/сети	$\sim$ 100	(i) факультет университета, (ii) сеть животных, (iii) отдел фирмы, (iv) сеть книг/статей по тематике (v) сеть поставок, (vi) сеть программ и компонентов (vii) структура молекулы, (viii) структуры производств	[146]
3.	Обычные множества данных/сети	$\sim$ 1 k	(i) сеть ссылок, (ii) сеть университета, (iii) collaboration network, (iv) городские системы, (v) базы покупателей, (vi) компьютерная сеть,	[69]
4.	Множества даннных/сети большой размерности	$\sim$ 10 k	(i) сеть исследователей, (ii) сеть сенсоров sensor, (iii) сеть производственных технологий,	[146]
5.	Множества данных/сети очень большой размерности	$\sim$ 100 k	(i) базы клиентов, (ii) интегральные схемы (VLSI), (iii)медицинские базы пациентов	[39]
6.	Множества данных/ сети "мега"- размерности	$\sim$ 1 M	(i) библиотека университета, (ii) база издательства	[196]
7.	Множества данных/ сети "супер"- размерности	$\sim$ 10 M	(i) сети библиотек, (ii) Интернет-магазины, (iii) биологические базы	[22]
8.	Базы/сети на основе Веб	$\sim$ 100 M ...1 B	(i) Веб-системы, (ii) социальные (Twitter, Facebook)	

3. КАЧЕСТВО РЕШЕНИЙ КЛАСТЕРИЗАЦИИ

Здесь рассматриваются задачи кластеризации без пересечения кластеров ("hard" clustering). Пусть имеется исходное множество элементов $A = \{a_1, ..., a_j, ..., a_n\}$. Два типа исходных данных для кластеризации исследуются: *1.* имеется m параметров/критериев и измерение элемента $a \in A$ основано на векторе оценок $\overline{x} = (x_1, ..., x_1, ..., x_m)$; *2.* имеется бинарное отноше-

ние(я) над множеством элементов A (включая взвешенные бинарные отношения; это может быть основой для структуры над кластерами). Решение кластеризации состоит из двух частей:

1. Кластеры $\widehat{X} = \{X_1, ..., X_\iota, ...X_\lambda\}$, т.е., разбиение элементов множества A на кластеры: $X_\iota \subseteq A \ \ \forall \iota = \overline{1,\lambda}; \quad \eta_\iota = |X_\iota|$ - размер кластера (мощность кластера X_ι, $\iota = \overline{1,\lambda}$).

2. Структура над кластерами (если необходимо). Пусть $\Gamma(\widehat{X})$ будет структурой над кластерами решения кластеризации $\widehat{X}$, т.е., существует орграф $G = \widehat{X}, \Gamma(\widehat{X})$. Пусть $\Gamma(X_\iota)$ будет сруктура над элементами кластера $X_\iota \ \ (\forall X_\iota \in \widehat{X}\ )$.

Список основных характеритик качества приведен в таблице 3.

Таблица 3. Список характеристик качества

Ном.	Тип качества	Обозначение	Описание
I.	Кластер	X_ι	$1 \le \iota \le \lambda$
1.1.	Расстояние "интра-кластер"	$I^{intra}(X_\iota)$	Близость элементов кластера
1.2.	Размер кластера	$\|X_\iota\|$	Число элементов в кластере X_ι
1.3.	Качество формы кластера		Близость к заданной форме (шар, эллипсоид)
1.4.	Размер региона кластера		Различие между “max” и “min” координат/параметров
1.5.	Качество "состава" кластера		Конфигурация типов элементов
1.6.	Качество структуры кластера		Близость к заданной структуре
II.	Решение кластеизации	$\widehat{X}$	$\widehat{X} = \{X_1, ..., X_\iota, ..., X_\lambda\}$
2.1.	Общее "интер"-кластер качество	$Q^{intra}(\widehat{X})$	Интеграция "интра"-кластер параметров ($I^{intra}(X_\iota)$, $\iota = \overline{1,\lambda}$)
2.2.	Общее "интер"-кластер качество	$Q^{inter}(\widehat{X})$	Интеграция "интер"-кластер параметров ($I^{inter}(X_{\iota_1}, X_{\iota_1})$, ι_1, ι_2, $\iota_1 \ne \iota_2$)
2.3.	Число кластеров (λ)	$Q^{num}(\widehat{X})$	Число кластеров к решении
2.4.	Близость к размеру кластера	$Q^{bal}(\widehat{X})$	баланс по размерам кластеров, близость к заданному вектору
2.5.	Качество форм кластеров	$Q^{form}(\widehat{X})$	Интеграция параметров формы
2.6.	Параметр регионов кластеров	$Q^{reg}(\widehat{X})$	Интеграция размеров регионов кластеров
2.7.	Функционал корреляциной кластеризации	$Q^{corr}(\widehat{X})$	Интеграция показателя по каждому кластеру (agreement, max) и показателя между кластерами (disagreements, min) [17, 18]
2.8.	Качество модульности	$Q^{mod}(\widehat{X})$	Параметр модульности сети [69, 144]
III.	Качество структуры над кластерами	$Q^{struc}(\widehat{X})$	Близость к заданной структуре
IV.	Многокритериальное качество	$\overline{Q}(\widehat{X})$	Вектор, например: $(\overline{Q}(\widehat{X}) = (Q^{intra}(\widehat{X}), Q^{inter}(\widehat{X}), Q^{bal}(\widehat{X})\ )$

1. Качество кластеров:

1.1. "Интра"-кластер расстояние (обобщенная близость между элементами в внутри каждого кластера):

$I^{intra}(X_\iota)$ $(\iota = \overline{1,\lambda})$.

Версия 1. Количественный прараметр как интеграция количественных параметров близости (расстояний) в кластере.

Версия 2. Параметр на основе мультимножества порядковых оценок близости элементов [121, 123]. Этот подход иллюстрируется примером.

Пример 1. На Рис. 1 представлен пример для 3-х кластеров: $X_1 = \{1, 2, 3, 4\}$, $X_2 = \{5, 6, 7\}$, $X_3 = \{8, 9, 10, 11, 12\}$. Использована порядковая шкала $[1, 2, 3]$ для оценки близости элементов: 1 - очень близко (похожи), 2 - средний уровень близости, 3 - очень различаются (в этом случае дуга на Рис. 1 отсутствует.) Порядковые оценки дуг привдены в таблице 4. Результирующие оценки "интра"-кластер параметров имеют вид: $I^{intra}(X_1) = (2, 3, 1)$, $I^{intra}(X_2) = (1, 1, 1)$, $I^{intra}(X_3) = (4, 2, 4)$.

Рис. 1. Локальное "интра"-кластер качество

Таблица 4. Порядковые близости (интра-кластер, ребро (i_1, i_2))

i_1	i_2 :	2	3	4	6	7	9	10	11	12
1		2	1	2						
2			3	2						
3				1						
5					2	3				
6						1				
8							1	1	2	3
9								3	3	1
10									3	2
11										1

1.2. Число элементов в кластере (размер кластера), например: $\pi^- \le \eta_\iota = |X_\iota| \le \pi^+$ $(\pi^-, \pi^+$ - заданные границы для размера кластера) $(\forall X_\iota \in \widehat{X})$.

1.3. Качество формы кластера (близость к заданной форме).

1.4. Качество как ограничение на размер региона кластера (т.е., границы для интервалов координат элементов кластера). Рассмотрим кластер $X = \{x^1, ..., x^\xi, ..., x^\phi\}$, оценки для каждого элемента кластера x^ξ are (vector estimate, parameters $i = \overline{1, m}$ (вектор): $\overline{x^\xi} = (x_1^\xi, ..., x_i^\xi, ..., x_m^\xi)$. Ограничения $\forall i = \overline{1, \phi}$ (Рис. 2):

$$|\max_{\xi=\overline{1,\phi}} x_i^\xi - \min_{\xi=\overline{1,\phi}} x_i^\xi| \le d_i, \quad \forall i = \overline{1, m}.$$

1.5. Качество состава кластера (например, состав бригады): один элемент 1-го типа, три элементов 2-го типа, два элементов 3-го типа, один элемент 4-го типа.

1.6. Качество структуры кластера X_ι ($\forall X_\iota \in \widehat{X}$) как близость к заданной структуре $\delta(\Gamma(X_\iota), \Gamma^0(X_\iota))$, $\Gamma^0(X_\iota)$ - заданная структура над элементами кластера.

2. Качество решения кластерици (т.е., множества кластеров):

2.1. "Интра"-кластер качества для решения кластеризации $Q^{intra}(\widehat{X})$ является интегрированой характеристикой $(I^{intra}(X_\iota))$ на основе "интра"-кластер параметров всех кластеров решения $(\iota = \overline{1, \lambda})$.

Версия 1. Количественное качество на основе оценок для кластеров:

$$Q^{intra}(\widehat{X}) = \frac{1}{\lambda} \sum_{\iota=\overline{1,\lambda}} I^{intra}(X_\iota).$$

Отметим, интеграция может быть основана на суммировании или других операциях (максимизация, минимизация).

Версия 2. Качество решения как оценка на основе мульмножества (на примере).

Пример 2. рассматривается пример трех кластеров из Рис. 3: $X_1 = \{1,2,3\}$, $X_2 = \{4,5,6\}$, $X_3 = \{7,8,9\}$; решение кластеризации: $\widehat{X} = \{X_1, X_2, X_3\}$. аналогично оценены близость (порядковая шкала) $[1,2,3]$). Порядковые оценки (близость) представлены в таблице 5.

"Интра"-кластер параметры для кластеров в виде мультмножеств имеют вид: $I^{intra}(X_1) = (1,2,0)$, $I^{intra}(X_2) = (2,1,0)$, $I^{intra}(X_3) = (2,1,0)$.

Интеграция указанных оценок в виде мульимножеств может проводится двумя методами [121, 123]:

(а) суммирование по компонентам векторов: $Q^{intra}(\widehat{X}) = (5,4,0)$,

(б) поиск медиан для мультимножеств: $Q^{intra}(\widehat{X}) = (2,1,0)$.

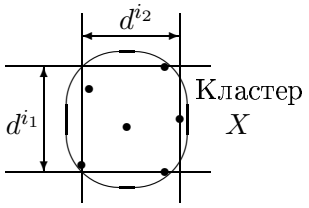


Рис. 2. Размер региона кластера

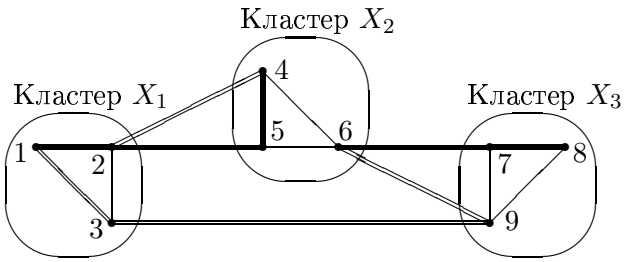


Рис. 3. Интра- и интер-кластер качества

Таблица 5. Порядковые близости (интра-кластер, ребро (i_1, i_2))

i_1	i_2 :	2	3	5	6	8	9
1		2	2				
2			1				
4				2	1		
5					1		
7						2	1
8							1

2.2. "Интер"-кластер качество для решения кластеризации $Q^{inter}(\widehat{X})$ является интегрированным показателем ($I^{intra}(X_{\iota_1}, X_{\iota_2})$) на основе оценок для пар кластеров в решении ($\iota_1 = \overline{1,\lambda}$, $\iota_2 = \overline{1,\lambda}$, $\iota_1 \neq \iota_1$).

Версия 1. Интеграция количесвенных "интер"-кластер показателей для пар кластеров:

$$Q^{inter}(\widehat{X}) = \frac{1}{\lambda(\lambda-1)} \sum_{\iota_1=\overline{1,\lambda},\iota_2=\overline{1,\lambda},\iota_1\neq\iota_2} I^{inter}(X_{\iota_1}, X_{\iota_2}).$$

Интеграция может основываться на различных операциях (суммирование, максимизация, минимизация). : Note, integration process can be based on summarization and some

Версия 2. Качество решения как оценка на основе мульмножества (на примере).

Приведем пример.

Пример 3. Использованы данные из примера 2 (Рис. 3). В таблице представлены порядковые близости для "интер"-кластер показателей.

Таблица 6. Порядковые близости (интер-кластер, ребро (i,j))

i	$j:$ 4	5	6	7	8	9
1	3	3	3	3	3	3
2	2	2	3	3	3	3
3	3	3	3	3	3	2
4				3	3	3
5				3	3	3
6				2	3	2

Показатели для пар кластеров в виде мультимножеств:

$I^{inter}(X_1,X_2)=(0,2,7)$, $I^{inter}(X_1,X_3)=(0,1,8)$, $I^{inter}(X_2,X_3)=(0,2,7)$.

Интеграция может быть основана на двух методах ([121, 123]):

(а) суммирование по компонентам векторов: $Q^{inter}(\widehat{X})=(0,5,22)$;

(б) поиск медианы мультимножеств: $Q^{inter}(\widehat{X})=(0,2,7)$.

2.3. Число кластеров в решении $Q^{num}(\widehat{X})$, например: $\Upsilon^- \le \lambda(\widehat{X}) \le \Upsilon^+$ (Υ^-,Υ^+ - заданные границы для числа кластеров в решении.

2.4. Близость размеров кластеров в решении к заданным границам, т.е., баланс по мощности кластеров, $Q^{bal}(\widehat{X})$, например: $\pi^- \le |X_\iota)| \le \pi^+$ (π^-,π^+ - границы для размера каждого кластера. Очевидно, баланс (или дисбаланс) решения может рассматриваться как число кластеров, Такая оценка может быть векторной или в виде мультимножества, например: число "хороших" кластеров (правильный размер), число "почти-хороших" кластеров (почти правильных размер), число остальных кластеров. Приведем векторные оценки. Обозначения: (а) $\pi^0(\widehat{X})$ - число "хороших" кластеров в $\widehat{X}$, т.е., размер кластера X_ι соответствует границам, (б) $\pi^{+l}(\widehat{X})$ - число кластеров $\widehat{X}$, где размер X_ι превышает $\widehat{\pi}^+$ (верхняя граница) на l элементов, (в) $\pi^{-l}(\widehat{X})$ - число кластеров в $\widehat{X}$, где размер кластера X_ι меньше, чем $\widehat{\pi}^-$ (нижняя граница) на l элементов. Векторная оценка имеет вид:

$$Q^{bal}(\widehat{X})=(\pi^{l^-}(\widehat{X}),...,\pi^{-1}(\widehat{X}),\pi^0(\widehat{X}),\pi^1(\widehat{X}),...,\pi^{l^+}(\widehat{X})).$$

Отметим, близкая по типу веторная оценка для сравнения ранжировок была предложена в [118]. Приведем пример.

Пример 4. Исходное множество: $A=\{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17\}$, решение кластеризации:: $\widehat{X}$: $X_1=\{1,5,7\}$, $X_2=\{2\}$, $X_3=\{3,6,10,13,17\}$, $X_4=\{11,12\}$, $X_5=\{4,12,14,15\}$, $X_6=\{8,16\}$. Ограничения для размера кластеров: $\widehat{\pi}_1=2$, $\widehat{\pi}_2=3$. Векторная оценка баланса решения: $Q^{bal}(\widehat{X})=(\pi^{-1}(\widehat{X}),\pi^0(\widehat{X}),\pi^1(\widehat{X}),...,\pi^2(\widehat{X}))=(1,3,1,1)$.

Данный подход близок к Υ-сбалансированному разбиению (решение $\widehat{X}$), когда размер каждого кластера $|X_\iota| \approx \frac{n}{\Upsilon(\widehat{X})}$ ($\forall X_\iota \in \widehat{X}$) где $\Upsilon(\widehat{X})$ (i.e., λ) - число полученных кластеров.

2.5. Качество баланса по формам кластеров в решении кластеризаци $Q^{form}(\widehat{X})$, например: большинство кластеров в решении имеют похожую заданную форму (сфера, эллипс). Очевидно, что аналогично можно рассматривать меру дисбаланса.

2.6. Характеристика баланса решения кластеризации по размерам регионов кластеров $Q^{reg}(\widehat{X})$. Рассмотрим кластер $X_\iota=\{x^{\iota,1},...,x^{\iota,\xi},...,x^{\iota,\phi_\iota}\}$. Оценки параметра по каждому элементу кластера $x^{\iota,\xi}$ $i=\overline{1,m}$ (кластеры: $\iota=\overline{1,\lambda}$): $\overline{x^{\iota,\xi}}=(x_1^{\iota,\xi},...,x_i^{\iota,\xi},...,x_m^{\iota,\xi})$. Ограничения по каждому параметру $\forall i=\overline{1,\phi_\iota}$) (Рис. 2):

$$|\max_{\xi=\overline{1,\phi_\iota}} x_i^\xi - \min_{\xi=\overline{1,\phi_\iota}} x_i^\xi| \le d_i, \quad \forall i=\overline{1,m}, \quad \forall \iota=\overline{1,\lambda}.$$

2.7. Функционал корреляционной кластеризации направлен на максимизацию "интра"-кластер "согласие" (agreement) и "интер"-кластер рассогласование (disagreement) $Q^{corr}(\widehat{X})$ [17,18]. Здесь рассматривается разбиение полностью связного графа с метками (labeled graph): метка "+" соответствует ребру между похожими вершинами, метка "-" соответствует ребру между различными вершинами. Оптимизируется функционал $Q^{corr}(\widehat{X})$ как сумма двух компонентов: (i) минимизация числа ребер типа "-" между кластерами (минимизация рассогласования), (b) максимизация чмсоа ребер типа "+" внутри кластеров (максимизация согласования) [2,15,17,18,51,111,183]. Также исследуется версия модели на взвешенном графе [32,33,51].

2.8. Модульность решения кластеризации $Q^{mod}(\widehat{X})$ определяется так [69,142,144,146] (Рис. 4). Пусть $G = (A, E)$ - исходный граф, где A - множество вершин, E - множество ребер; решение кластеризации: $\widehat{X} = \{X_1, ..., X_\iota, ..., X_\lambda\}$. Пусть A^ι - множество вершин в кластере X_ι ($\iota = \overline{1,\lambda}$). Пусть E^ι - множество внутренних ребер в кластере X_ι ($\iota = \overline{1,\lambda}$), т.е., все соответствующие вершина принадлежат A^ι. Пусть $\widetilde{E}^\iota$ - множество внешних ребер для кластера X_ι ($\iota = \overline{1,\lambda}$), т.е., только один конец ребра принадлежит A^ι. Определения иллюстрируются на Рис. 4 для решения, содержащем четыре кластера (кластер X_3).

Для каждого кластера X_ι используются параметры:

(а) $e_\iota = \frac{|E^\iota|}{|E|}$ (% число ребер в модуле ι),

(б) $a_\iota = \frac{|\widetilde{E}^\iota|+|E^\iota|}{|E|}$ (% ребра с одним концом в модуле ι).

Общая характеристика модульности решения кластеризации для графа G имеет вид:

$$Q^{mod}(\widehat{X}) = \sum_{\iota=1}^{\lambda} (e_\iota - (a_\iota)^2).$$

Задача кластеризации с целью максимизации характеристики модульности относится к классу NP-трудных задач [27]. Рассмотрим иллюстративный пример.

Пример 5. Рассмотрим характеристики модульности для решения на Рис. 4. Здесь $|E| = 26$, параметры для кластеров:

(1) $|E^1| = 6$, $|\widetilde{E}^1| = 4$, $e_1 = 0.23$, $a_1 = 0.38$; (2) $|E^2| = 4$, $|\widetilde{E}^2| = 4$, $e_2 = 0.15$, $a_2 = 0.3$;

(3) $|E^3| = 3$, $|\widetilde{E}^3| = 4$, $e_3 = 0.115$, $a_3 = 0.27$; (4) $|E^4| = 4$, $|\widetilde{E}^4| = 2$, $e_4 = 0.15$, $a_4 = 0.23$.

Результирующая характеристика модульности для решения: $Q^{mod}(\widehat{X}) = (0.23 - 0.14) + (0.15 - 0.09) + (0.115 - 0.073) + (0.15 - 0.053) = 0.09 + 0.06 + 0.42 + 0.097 = 0.667$.

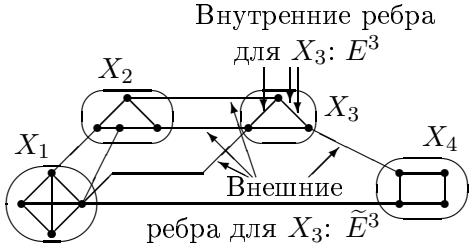


Рис. 4. Модульность в кластеризации графа

3. Качество структуры над кластерами (например, дерево, иерархия) ($Q^{struc}(\widehat{X})$). Здесь близость полученной структуры $\Gamma(\widehat{X})$ в решении кластеризации in clustering solution к заданной структуре Γ^0 исследуется: $Q^{struc}(\widehat{X}) = \delta(\Gamma(\widehat{X}), \Gamma^0)$. Для оценки данного показателя могут быть использованы различные шкалы (количественная, порядковая, векторная, в виде мультимножеств) [120, 123].

4. Многокритериальная (векторная) оценка качества решения кластеризации интегрирует указанные выше характеристики, на пример:
$Q(\widehat{X}) = (Q^{inter}(\widehat{X}), Q^{intra}(\widehat{X}), \pi(\widehat{X}))$.

Таким образом, задача кластеризации может формулироваться как многокритериальная оптимизационная задача (могут искаться Парето-эффективные решения), например:

$$\min\ Q^{intra}(\widehat{X}), \quad \max\ Q^{inter}(\widehat{X}), \qquad s.t. \quad Q^{bal}(\widehat{X}) \preceq \pi^0, \quad Q^{struc} = \delta(\Gamma(\widehat{X}), \Gamma^0) \leq \delta^0.$$

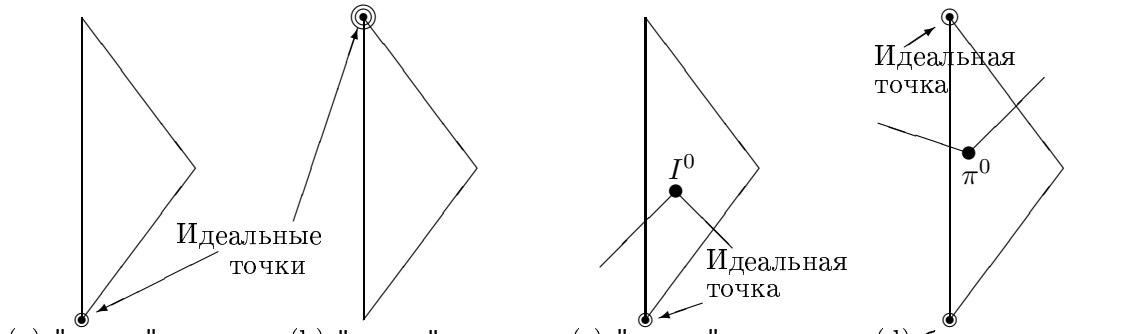
В случае использования оценок в виде мультимножеств, многокритериальная задача кластеризации может формулироваться на основе решеток качества (шкал в виде частично-упорядоченного множества) (при этом Парето-эффективные решения ищутся на основе указанных решеток):

$$\min\ Q^{intra}(\widehat{X}), \quad \max\ Q^{inter}(\widehat{X})$$

$$s.t. \quad I^{intra}(X_\iota) \succeq I^0,\ \forall \iota = \overline{1,\lambda}, \quad Q^{bal}(\widehat{X}) \preceq \pi^0, \quad Q^{struc} = \delta(\Gamma(\widehat{X}), \Gamma^0) \leq \delta^0.$$

Здесь: $I^0\ \ \forall X_\iota$ - опорная оценка в виде мультимножества (по решетке, Рис. 5c), π^0 - опорная оценка в виде мультимножества для баланса по размеру кластеров (по решетке, Рис. 5d).

Рис. 5 иллюстрирует интегрированное дискретное "пространство" качества для обобщенной векторной оценки качества в виде мультимножества для решения кластеризации.



(a) "интра"-кластер качество $Q^{intra}(\widehat{X})$ (b) "интер"-кластер качество $Q^{inter}(\widehat{X})$ (c) "интра"-кластер качество $I^{intra}(X_\iota), \forall \iota$ (d) баланс по размеру кластера $Q^{bal}(\widehat{X})$

Рис. 5. Иллюстрация для решеток качества

В случае "размытых" задач кластеризации (“soft” clustering), необходимо рассматривать меры "размытости" решений (например, как общая характеристика пересечения кластеров).

4. КЛАСТЕРИЗАЦИЯ НА ОСНОВЕ ГРАФА

4.1. Кластеризация на основе минимального покрывающего дерева

Предварительное построение минимальных покрывающих деревьев для исходных графов широко используется в схемах решения многих комбинаторных задач (e.g., [67]). Алгоритмическая сложность для такой задачи равна $O(n \log n)$ (n - число вершин графа). Алгоритмы на основе минимального покрывающего дерева исследовались и применялись многими исследователями [74,75,113,138,154,160,181,197,201]. Основные стадии такого алгоритма следующие:

Стадия 1. Вычисление матрицы расстояний/близости объектов Z.

Стадия 2. Построение соответствующего графа G.

Стадия 3. Построение миимального покрывающего дерева T для графа G.

Стадия 4. Кластеризация вершин дерева T (например, алгоритм удаление ветвей, алгоритм иерархической кластеризации).

Стадия 5. Останов.

Далее рассматривается применение иерархической кластеризации на стадии 4. Оценки сложности рассматриваемого алгоритма приведены в таблице 7.

Таблица 7. Сложность: кластеризация на основе минимального покрывающего дерева

Стадия	Описание	Оценка
Стадия 1.	Вычисление матрицы расстояний Z	$O(n^2)$
Стадия 2.	Построение соответствующего графа G	$O(n^2)$
Стадия 3.	Построение минимального покрывающего дерева	$O(n \log n)$
Стадия 4.	Кластеризация вершин дерева	$O(n \log n)$
Стадия 5.	Останов	$O(1)$

Стадии 3, 4, 5 соответствуют ситуации, когда в качестве исходных данных используется граф. $O(n \log n)$.

Пример 6. Рассматривается множество объектов $A = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$. Матрица близости представлена в таблице 8 (символ "$\star$" означает очень большое значение). Граф $G = (A, E)$ изображен на Рис. 6, покрывающее дерево $T = (A, E')$ - на Рис. 7 (включая решение кластеризации).

Представляется целесообразным использовать модификацию графа $G = (A, E)$ на основе удаление ребер по условию: вес ребра "$>$" порога. Уменьшение порога приводит к уменьшению мощности множества ребер E. Такой процесс может быть полезен при обработке исходных данных. Проиллюстрируем указанный процесс на примере данных из примера 6:

(i) порог 2.6: граф $G = (A, E)$ (Рис. 6);

(ii) порог 1.4: граф $G^1 = (A, E^1)$ (Рис. 8);

(iii) порог 0.5: граф $G^2 = (A, E^2) = T = (A, E')$ (Рис. 7, покрывающее дерево);

(iv) порог 0.3: граф $G^3 = (A, E^3)$ (Рис. 9).

Применение указанной процедуры позволяет найти "удобную" структуру.

Таблица 8. Близости для примера (ребро (i_1, i_2))

i_1	i_2 :	2	3	4	5	6	7	8	9	10	11	12
1		0.3	1.4	1.45	$\star$	$\star$	$\star$	$\star$	$\star$	$\star$	$\star$	$\star$
2			0.3	$\star$	$\star$	2.6	0.2	1.8	$\star$	$\star$	$\star$	$\star$
3				0.4	$\star$	$\star$	1.65	0.25	$\star$	$\star$	$\star$	$\star$
4					0.4	$\star$	$\star$	0.45	1.9	$\star$	$\star$	$\star$
5						$\star$	$\star$	$\star$	0.35	1.5	$\star$	$\star$
6							0.1	$\star$	$\star$	$\star$	1.4	$\star$
7								0.41	$\star$	$\star$	0.4	$\star$
8									0.9	$\star$	2.1	$\star$
9										0.15	$\star$	0.5
10											$\star$	2.0
11												2.5

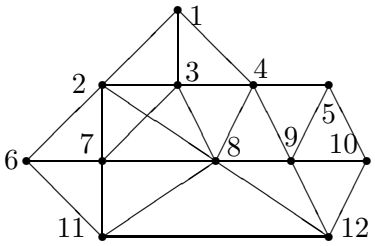


Рис. 6. Граф $G = (A, E)$

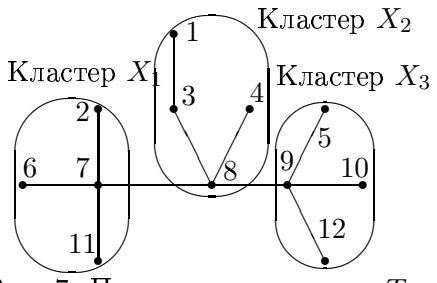


Рис. 7. Покрывающее дерево $T = (A, E')$

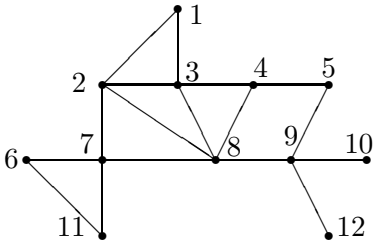


Рис. 8. Граф $G^1 = (A, E^1)$

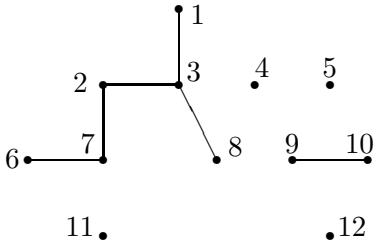


Рис. 9. Граф $G^3 = (A, E^3)$

Теперь рассмотрим адаптивной версию кластеризации на основе минимального покрывающего дерева. Исходные данные включают следующее: (а) множество объектов/альтернатив $A = \{A_1, ..., A_i, ..., A_n\}$, (б) множество параметров/критериев $\overline{C} = \{C_1, ..., C_j, ..., C_m\}$, (в) матрица оценок $X = \{x_{ij}\}$, $i = \overline{1,n}, j = \overline{1,m}$, x_{ij} - количественная оценка A_i по критерию C_j. Алгоритм включает стадии:

Стадия 1. Вычисление матрицы близости $Z = \{z_{ik}\}$, $i = \overline{1,n}$, $k = \overline{1,n}$, где z_{ik} - оценка близости (расстояние) между A_i и A_k (используется Эвклидова метрика). Очевидно, $z_{ii} = 0, \forall i = \overline{1,n}$.

Стадия 2. Преобразование матрицы Z в матрице с порядковыми значениями $Y = \{y_{ik}\}$. Пусть имеются максимальное и минимальное значения элементов матрицы Z:

$z^{min} = \min_{\forall i=\overline{1,n}, i=\overline{1,k}} \{z_{ik}\}$, $z^{max} = \max_{\forall i=\overline{1,n}, i=\overline{1,k}} \{z_{ik}\}$.

Тогда получается интервал $[z^{min}, z^{max}]$ и $d = z^{max} - z^{min}$. Используется дополнительный параметр δ (т.е., $3, 4, 5, 6$). Пусть $\delta = 5$. Элементы новой матрицы можно вычислить так:

$$y_{ik} = \begin{cases} 0, \text{if } \ 0.0 \le z_{ik} \le d/\delta, \\ 1, \text{if } \ d/\delta < z_{ik} \le 2d/\delta, \\ 2, \text{if } \ 2d/\delta < z_{ik} \le 3d/\delta, \\ 3, \text{if } \ 3d/\delta < z_{ik} \le 4/\delta, \\ 4, \text{if } \ 4d/\delta < z_{ik} \le d. \end{cases}$$

Стадия 3. Построение связного графа (итеративный процесс):

Пусть $\Delta = 1, 2, ...$ - целый параметр алгоритмы (индекс цикла).

Шаг 3.1. Исходное значение индекса $\Delta = 1$.

Шаг 3.2 Преобразование порядковой матрицы Y в Булеву матрицу $B = \{b_{ik}\}$:

$$b_{ik} = \begin{cases} 1, \text{if } \ y_{ik} < \Delta, \\ 0, \text{if } \ y_{ik} \ge \Delta. \end{cases}$$

Шаг 3.3. Построение графа $G^\Delta = (A, \Gamma^\Delta)$, где Γ^Δ - множество ребер, ребро (A_i, A_k) существует, если $b_{ik} = 1$.

Шаг 3.4. Анализ связности графа $G^{\Delta} = (A, \Gamma^{\Delta})$. Если граф связен, то GOTO *Шаг 3.6.*

Шаг 3.5. $\Delta = \Delta + 1$ и GOTO *Шаг 3.2*

Шаг 3.6. Построение минимального покрывающего дерева для графа $G^{\Delta} = (A, \Gamma^{\Delta})$: $T^{\Delta} = (A, \widehat{E}^{\Delta})$. Здесь могут использоваться известные алгоритмы, например: алгоритм Боровка, алгоритм Прима, алгоритм Крускала [7, 66, 67, 45, 161, 209]. Сложность этих алгоритмов: $O(p \log n)$ (или менее [209]) (p - число ребер, n - число вершин).

Step 3.7. Кластеризация множества объектов A на основе покрывающего дерева $T^{\Delta} = (A, \widehat{E}^{\Delta})$ с учетом алгоритмических параметров: в каждом кластере число элементов α удовлетворяет условию: $\alpha' \leq \alpha \leq \alpha''$, например: $\alpha' = 4$, $\alpha'' = 6$.

Стадия 4. Stop.

В таблице 9 указаны оценки сложности стадий алгоритма. Общая оценка сложность алгоритма равна $O(n^2)$. В общем случае задача k-сбалансированного разбиения дерева является NP-трудной (k - размер кластера в решении кластеризации) [61].

Таблица 9. Сложность: адаптивный алгоритм на основе покрывающего дерева

Стадия/шаг	Описание	Оценка
Стадия 1	Вычисление матрицы расстояний Z	$O(n^2)$
Стадия 2	Перобразование метрицы Z в порядковую матрицу Y	$O(n^2)$
Стадия 3	Построение связного графа над элементами A	$O(n^2)$
Шаг 3.1	Задание начала цикла	$O(1)$
Шаг 3.2	Преобразование метрицы Y в Булеву матрицу B	$O(n^2)$
Шаг 3.3	Построение графа G, соответствующего матрице B	$O(n^2)$
Шаг 3.4	Анализ связности графа G	$O(n)$
Шаг 3.5	Коррекция параметра цикла	$O(1)$
Шаг 3.6	Построение минимального покрывающего дерева T для графа G	$O(p \log n)$
Шаг 3.7	Кластеризация вершин дерева T (ограниченные размер кластера)	$O(n)$
Стадия 4.	Останов	$O(1)$

4.2. Кластеризация на основе выделения клик

Рассмотрим граф $G = (A, E)$ в качестве исходных данных. В клике (полный граф/подграф) каждая вершина связана со всеми остальными вершинами. Может рассматриваться квази-клика, например, как клика без одного ребра. Клики (или квази-клики) формиуют очень "насыщенный" ("сильный") кластер с точки зрения взаимосвязи элементов. Задача поиска максимальной клики в графе является известной NP-трудной проблемой [67,101]. Таким образом переборные методы или эвристики используются для этой задачи. Процесс кластеризации на основе клики организуется как последовательность задач поиска клик [67]:

Стадия 1. Поиск максимальной клики (или максимальной квази-клики) в графе $G = (A, E)$: подграф $H = (B, V)$ ($H \subseteq A$, $V \subseteq E$).

Стадия 2. Формирование кластера из подграфа H и сжатие исходного графа G: $G' = (A', E')$, ($A' = A \backslash H$, $E' = E \backslash \{V \bigcup W\}$, где W - множество внешних ребер клики, т.е., только один конец ребра принадлежит множеству H) (Рис. 10).

Стадия 3. Если G' - пустой, то GO TO Стадия 4 иначе GO TO Стадия 1.

Стадия 4. Останов.

Указанная схема решения основана на последовательности NP-трудных задач. Очевидно, что можно рассматривать параллельный поиск нескольких клик. В таблице 10 приведен список исследовательских направлений в области кластеризации на основе клики.

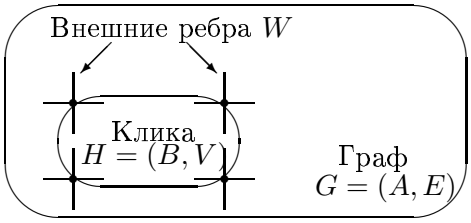


Рис. 10. Клика в графе

Таблица 10. Выделение клик/квази-клик, кластеризация

Ном.	Исследование	Источник
1.	Выделение, анализ клик в графах:	
1.1	Клики в графах	[98,136]
1.2.	Поиск клик/квази-клик в графах	[1,10,30,59,67,150,156]
1.3.	Клика максимального веса	[13,16,26,151,155]
1.4.	Поиск всех клик в графе	[28]
1.5.	Перечисление максимальных клик в большом графе	[9]
2.	Кластеризация на основе клик:	
2.1.	Кластеризация	[6,21,54,70,96,131,134,176,191]
2.2.	Разбиение графов на клики максимального веса	[49,109,148]
3.	Множественная (ансаблевая) кластеризация (построение результирующего решения на основе интеграции набора решений, полученных различными методами):	
3.1.	Множественная кластеризация с голосующими кластерами	[192]
3.2.	Использование клик для комбинирования решений	[133]
4.	Методы на основе выделения клик/квази-клик над потоками данных:	
4.1.	Кластеризация с К-кликами в динамических сетях	[54]
4.2.	Интеграция потоков графов на основе клик	[122]

Дополнительно, рассматривается задача поиска максимальной взвешенной клики, когда ребрам приписаны веса и ищется клика, в которой сумма весов ребер является максимальной. В случае разбиения графа на основе клик, в решении сумма весов во всех выделенных кликах должна быть максимальной (e.g., [109,148]).

Отметим, что имеются задачи на графах, которые являются близкие к рассматриваемым. Среди таких задач можно указать следующие: задачи независимого множества, задачи доминирующего множества [36,42,43,46,81,92,158].

В последние годы повысилась значимость и актуальность задач обработки потоков данных (streams), включая задач выделения клик, квази-клик в потоках графов [5,40,77,122].

В тоже время, методы кластеризации на основе клик могут рассматриваться как следующие: (а) методы на основе решеток (grid-based clustering), (б) метода на основе выделения "плотных" областей (density-based clustering).

Также следует отметить, что клики и квази-клики рассматриваются как один из видов сетевых сообществ в методах кластеризации на основе сетевых сообществ [69,144,145,163].

В последние годы были предложены задачи кластеризации в многодольных графах/сетях [34,48,86,118,119,123,194]. Такая задача проиллюстрирована на Рис. 11. Направления исследований в этой области приведены в таблице 11.

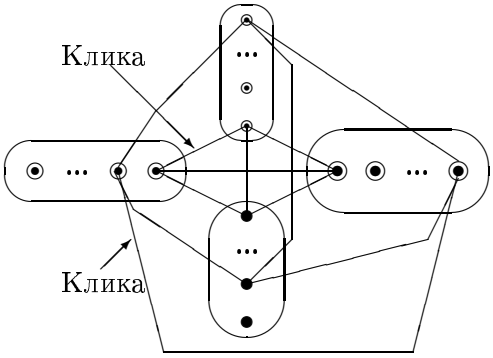


Рис. 11. Клики в 4-дольном графе

Таблица 11. Исследования многодольных графов

Ном.	Исследование	Источник
1.	Задача совместимых представителей	[108]
2.	Задача морфологической клики (порядковые оценки)	[118,119,123]
3.	Задача морфологической клики (оценки в виде мультимножеств)	[121,123]
4.	Кластеризация в многодольном графе	[34,194]
5.	Двух-дольная и многодольная клика	[48]
6.	Морфологическая клика над потоком графов	[122]
7.	Задачи множеств ядер (coreset problems)	[60,86]
8.	Задачи множеств ядер над динамическими потоками данных	[64]
9.	Выделение сообществ в k-дольных сетях	[127]

Корреляционная кластеризация

Корреляционная кластеризация направлена на разбиение полностью связного графа с метками ребер (метка "+" соответствует ребру, соединяющему близкие/похожие вершины, метка "-" соответствует ребру, соединяющему различающиеся вершины. При этом рассматриваются две целевые функции для решения (набора кластеров) [2,15,18,20,51,111,183]:

(i) минимизация "рассогласований" (disagreements) (минимизация числа ребер с меткой "-" в внутри кластеров ($Q^{disagr}(\widehat{X}) \rightarrow \min$) или максимизация числа меток "-" между кластерами),

(ii) максимизация "согласований" (agreements) (число меток "+" для ребер внутри кластеров) ($Q^{agr}(\widehat{X}) \rightarrow \max$).

В базовой формулировке указанные функции суммируются ("функционал" задачи). Другими словами, бинарная шкала $[-1,+1]$ используется для каждого ребра как вес (0 - не используется). В данном подходе не требуется задавать предварительное число кластеров (например, как это нужно в методе k-"means" кластеризации). Данная формальная остановка возникла в области кластеризации документов/Веб страниц. Эта задача является NP-полной [8,17,18].

Кроме того, различные версии данной задачи рассматриваются: (а) взвешенные версии дла функционала задачи [32,33,51], (б) задача с частичной информацией [50], (в) задача с шумом на входе [130]).

Рассмотрим взвешенную версию задачи. Пусть $A = \{A_1, ..., A_j, ..., A_n\}$ - исходное множество объектов/элементов. Можно рассмотреть $(n-1)^2$ пар элементов: $G = \{g_1, ..., g_{(n-1)^2}\}$. Каждый элемент множества G соответствует паре (A_{j_1}, A_{j_2}) и элементу матрицы близости $Z = ||z_{j_1,j_2}||$.

Далее заменяем шкалу $[-1,+1]$ для ребра на количественную шкалу для весов: отрицательная количественная шкала $[-w-,...,0)$ вместо "-1" и положительная количественная шкала $(0,...,w^+]$ вместо "$+1$". Множество пар элементов разделяется на два непересекающих подмножества $G = G^- \bigcup G^+$ ($|G^- \bigcap G^+| = 0$) где $\forall g^- \in G^-$ и $\forall g^+ \in G^+$. Решение кластеризации имеет вид: $\widehat{X} = \{X_1,...,X_\iota,...,X_\lambda\}$. Используются две функции:

(i) общее "согласование" (суммирование весов всех "интра"-кластер пар с позитивными весами ребер) $Q^{agr}(\widehat{X})$ (максимизация);

(ii) общее "рассогласование" (суммирование весов "интра"-кластер пар с негативными весами ребер) $Q^{disagr}(\widehat{X})$ (минимизация, по модулю).

Модель имеет вид (Рис. 12):

Найти решение кластеризации $\widehat{X}$ такое, что: (i) $Q^{agr}(\widehat{X}) \to \max$ и (ii) $|Q^{disagr}(\widehat{X})| \to \min$.

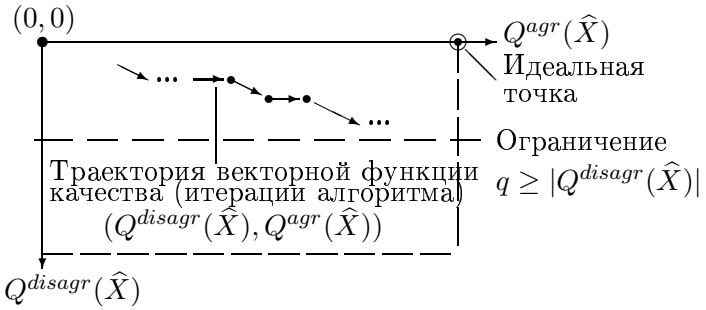


Рис. 12. "Пространство" качества решения

Для версий задачи корреляционной кластеризации были предложены эвристики и приближенные алгоритмы (включая полномиальные приближенные схемы PTAS) [15, 17, 18]. Известный иерархический (агломеративный) алгоритм может применяться здесь как "гриди" (greedy) эвристика (т.е., пошаговое улучшение решения задачи на основе выбор следующей пары объектов для объединения). Приведем пример такой схем решения (Рис. 12):

Стадия 1. Вычисление матрицы близости (расстояний) пар элементов $\forall(A(j_1), A(j_2))$, $A(j_1) \in A,\ A(j_2) \in A,\ j_1 \neq j_2$.

Стадия 2. Преобразование матрицы пар близости элементов в позитивные и негативные веса.

Стадия 3. Задание первоначального решения кластеризации $\widehat{X}^0$ как композиции исходных элементов, векторная целевая функция имеет вид: $\overline{f}^0 = (Q^{disagr}(\widehat{X}^0), Q^{agr}(\widehat{X}^0)) = (0,0)$ (задано исходное значение функции, начальный индекс - $\gamma = 0$).

Стадия 4. Поиск пары элементов, обеспечивающий наилучшее улучшение целевой функции $\overline{f}$ (т.е., поиск Парето-эффективных точек). Интеграция соответсвующих элементов (пары) в некий кластер или включение элемента в кластер с вторым элементом пары. Новок решение: $\widehat{X}^q$ (q - индекс итерации). Новое вычисление значения целевой функции: $\overline{f}^\gamma = (Q^{disagr}(\widehat{X}^\gamma), Q^{agr}(\widehat{X}^\gamma))$.

Стадия 5. Если все элементы обработаны, то GO TO Стадия 7.

Стадия 6. Увеличение индекса $\gamma = \gamma + 1$ пока выполняется $|Q^{disagr}(\widehat{X})| \leq q$ и *Go To* Стадия 4, иначе GO TO Стадия 7.

Стадия 7. Останов.

В таблице 12 указаны оценки сложности стадий алгоритма.

Table 12. Оценки сложности стадий агломеративного алгоритма

Стадия	Описание	Оценка
Стадия 1.	Вычисление матрицы расстояний (близости) Z	$O(n^2)$
Стадия 2.	Вычисление позитиыных/негативных весов ребер	$O(n^2)$
Стадия 3.	Задание начального решения	$O(1)$
Стадия 4.	Поиск наилучшей пары элементов для улучшения решения (улучшение по Парето)	$O(n^2)$
Стадия 5.	Анализ условия окончания работы алгоритма, вычисление нового значения целевой функции	$O(n)$ $O(n)$
Стадия 6.	переход вычислительного процесса	$O(1)$
Стадия 7.	Останов	$O(1)$

В таблице 13 указаны основные направления исследований в области корреляционной кластеризации.

С другой стороны, можно использовать для оценки целевой функции мультимножества. Заменим шкалу для ребер $[-1,+1]$ (или две количественных шкалы, указаных выше): (а) на отрицательную негативную шкалу $[-k^-,...,-1]$ вместо "-1", (б) на позитивную порядковую щкалу $[+1,...,k^+]$ вместо "$+1$". Заметим, вычисление новых весов по указанным порядковым шкалам выполняется достаточно просто. Две целевые функции для решения $\widehat{X} = \{X_1,...,X_\lambda\}$ определяются так: (i) общая оценка качества "согласований" на основе мультимножеств (суммирование по компонентам для всех "интра"-кластер пар с позитивными весами) $Q^{agr}(\widehat{X})$ (максимизация); (ii) общая оценка качества "рассогласований" на основе мультимножеств (суммирование по компонентам для всех "интра"-кластер пар с негативными весами) $Q^{disagr}(\widehat{X})$ (минимизация). В результате получается задача корреляционной кластеризации на основе мультимножеств (Рис. 13):

Найти решение кластеризации $\widehat{X}$ такое, что $Q^{agr}(\widehat{X}) \to \max$ и $|Q^{disagr}(\widehat{X})| \to \min$.

Таблица 13. Корреляционная кластеризация

Ном.	направление исследований	Источник
1.	Основная формулировка задачи, сложность	[8, 15, 17, 18, 20, 111]
2.	Обзоры	[8, 18, 111]
3.	Сравнение методов	[56]
4.	Приближенные алгоритмы (включая PTAS)	[15, 17, 18, 68]
5.	Взвешенные версии задачи	[32, 33, 51]
6.	Корреляционная кластеризация с фиксированным размером кластеров	[68]
7.	Максимизацмя "согласований" посредством полуопределенного (semidefinite) программирования	[183]
8.	Минимизация "рассогласований" на взвешенном графе	[57]
9.	Глобальная корреляционная кластеризация	[3]
10.	Корреляционная кластеризация с частичной информацией	[50]
11.	Корреляционная кластеризация с шумом на входе	[130]
12.	Корреляционная кластеризация с ограничением по ошибкам	[97]
13.	Робастная корреляционная кластеризация	[2, 110]
14.	Корреляционная кластеризация в сегментации изображений	[106]

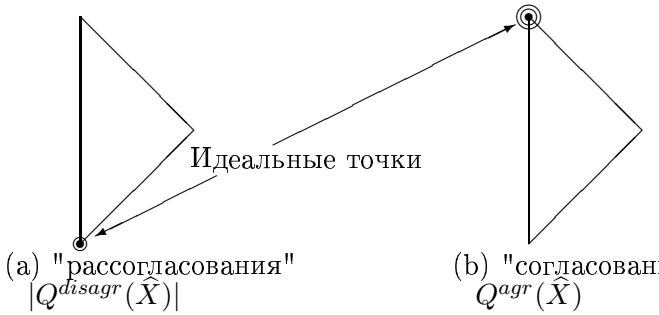


Рис. 13. Решетка качества на основе мультимножеств

4.3. Кластеризация на основе сетевых сообществ

В последнее десятилетие активно исследуются и применяются методы кластеризации на основе выделения сетевых сообществ [63, 69, 87, 117, 142, 143, 144, 145, 146, 163]. В качестве сетевых сообществ рассматривают подструктуры сети: клика, квази-клика, клика с листом (висячей вершиной), квази-клика с листом (висячей вершиной), цепь клик/квази-клик, интегрированные группы клик/квази-клик (Рис. 14).

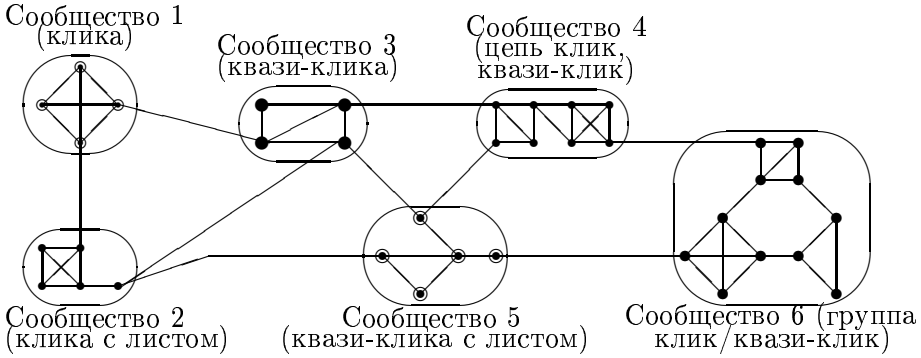
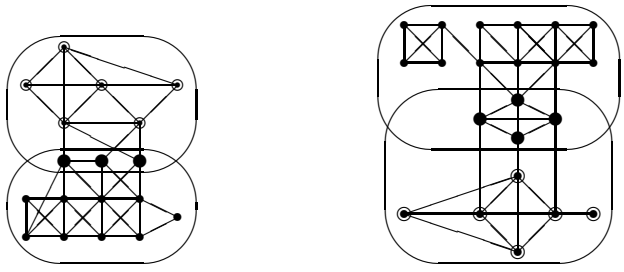


Рис. 14. Иллюстрация для сетевых сообществ

Иллюстративный пример сети на Рис. 14 не содержит перекрытий (т.е., без пересечений сетевых сообществ). Рис. 15 иллюстрирует перекрытия.



(a) Редкие перекрытия (b) плотные перекрытия

Рис. 15. Иллюстрация перекрытий структур

Выделение сетевых сообществ соответствует сложным моделям комбинаторной оптимизации (линейное/нелинейное целочисленное программирование, смешанное целочисленное программирование). Эти модели относятся к классу NP-трудных задач [27, 39, 63, 145]. Таблица 14 содержит список основных направлений исследований с данной области.

Модульность графа определяется как нормализованный компромисс между ребрами, которые покрываются кластерами, и суммами в квадрате размеров кластеров (squared cluster degree sums) [27, 146]. Зачача формулируется как модель комбинаторной оптимизации (максимизация модульности). Несколько основных алгоритмов предложены [27]:

(a) "гриди" (greedy) агломерация [39, 142],

(b) спектральное разделение [144, 199],

(c) метод "отжига" [79, 164],

(d) экстремальная оптимизация [55].

Приведем пример "гриди" алгоритма (агломерация) [142]:

Стадия 1. Тривиальная кластеризация: каждый узел - кластер.

Стадия 2. Цикл по парам кластерам:

Стадия 2.1. Вычисление возможных улучшений модульности при слиянии каждой пары кластеров.

Стадия 2.2. Слияние пары кластеров с максимальным улучшением модульности.

Стадия 2.3. Если увеличение модульности на основе слияния пар кластеров невозможно, то GO TO стадия 3.

Стадия 2.4. Go To Стадия 2.2.

Стадия 3. Останов.

Оценка сложности данного алгоритма: $O((p+n)n)$ or $O(n^2)$ [142].

Общая схема GN-алгоритма на основе ребер "между" (betweenness) имеет вид [69]:

Стадия 1. Вычисление показателя "между" (betweenness) для каждого ребра.

Стадия 2. Удаление ребер с высоким показателем "между".

Стадия 3. Анализ качества сетевых компонентов.

Стадия 4. Если все ребра удалены и система сеть разбита на N несвязных узлов, то Go TO Стадия 5. Иначе GO TO Step 1.

Стадия 5. Stop.

Оценка сложности данного алгоритма равна $O(p^2n)$ (p - число) [69].

5. О "БЫСТРОЙ" КЛАСТЕРИЗАЦИИ

Многие современные приложения базируются на множеств данных/сетей очень большой размерности. Это требует применения "быстрых" методов кластеризации [39, 142, 179, 188, 196]. В таблице 15 приведен перечень основных подходов в области построения схем "быстрой" кластеризации.

Следует отметить, что такие многие методы часто основаны на двух-уровневом походе: глобальный уровень и локальный уровень:

(а) разбиение исходной задачи на локальные задачи меньшей размерности (уменьшение размерности, ограничение типов объектов) (верхний уровень),

(б) решение локальных задач кластеризации (локальный уровень),

(в) композиция/интеграция локальных решений в финальное решение (верхний уровень).

В таблице 16 приведен список основных "быстрых" методов "локальной" кластеризации.

Таблица 14. Кластеризация на основе сетевых сообществ

Ном.	Направление	Источник
1.	Базовые вопросы:	
1.1.	Основные постановки	[27, 63, 69, 117, 146, 144, 145, 163, 198, 207, 208]
1.2.	Обзоры	[27, 24, 63, 117, 135, 146, 144, 145, 163]
1.3.	Сложность	[27, 39, 63, 145]
1.4.	Покрывающие сетевые структуры	[73, 198, 200, 205, 206, 207]
1.5.	Анализ/оценивание сообществ	[117, 146, 198, 208]
2.	Основные схемы решения:	
2.1.	Алгоритм на основе ребра "между" (betweeness)	[69]
2.2.	Алгоритм на основе модульности, (greedy) эвристика	[142, 152]
2.3.	Алгоритм типа "клуб карате"	[146]
2.4.	Метод Kernighan-Lin его варианты	[102]
2.5.	Покрывающие сообщества кликовая перколяция, расширение, динамические методы	[73, 200, 205, 206, 207]
2.6.	Спектральные алгоритмы	[208]
2.7.	Генетические алгоритмы	[126]
2.8.	Много-агентные алгоритмы	[80]
3.	Максимизация модульности:	
3.1.	Обзоры	[27, 144, 213, 208]
3.2.	3-дольные сети	[139, 140]
3.3.	k-дольные сети	[127]
3.4.	Агломеративные алгоритмы	[39, 142]
3.5.	Спектральные алгоритмы (division)	[144, 199]
3.6.	Алгоритмы "отжига"	[79, 164]
3.7.	Выделение сообществ (слияние клик)	[203]
3.8.	Математическое программирование	[4, 55]
3.9.	Глобальная оптимизация	[132]
3.10.	"Меметик" (Memetic) алгоритмы	[141]
3.11.	Алгоритмы случайного блуждания	[162]
3.12.	Много-уровневые алгоритмы	[53, 147]
4.	Большие сети:	
4.1.	Сообщества в болших сетях	[22, 39, 73, 87, 88, 117, 162, 208]
4.2.	Сообщества в мега-сетях	[196]
4.3.	Сообщества в супер-сетях	[22]
4.4.	Эволюции сообществ в больших сетях	[88]
5.	Приложения:	
5.1.	Веб	[53, 117, 139]
5.2.	Сети ссылок, сети публикаций	[37, 63, 167]
5.3.	Социальные сети	[63, 69, 78, 142, 145, 146, 196, 208]
5.4.	Биологические сети	[63, 69, 146]
5.5.	Сети снабжения	[39]
5.6.	Проектирование	[146]
5.7.	Мобильные сети связи	[23, 128]

Таблица 15. Основные подходы к "быстрой"

Ном.	Подход	Схема решения	Источник
1.	Агрегация объектов/ узлов сети	Иерархическая кластеризация (агрегация объектов)	[95, 96, 179]
2.	Разбиение множества объектов/ узлов сети (декомпозиция):	Схема "Сверху-Вниз"	
2.1.	Удаление объектов (Рис. 16)	1.Выбор ребер "между" (betweenness) в графе, разделение графа (две части)) 2.Кластеризация каждой части	[69, 179]
2.2.	Многоуровневая схема:	1.Разбиение множества объектов 2.Локальная кластеризация 3.Композиция общего решения	[187, 188]
2.2.1.	Схема на основе выделения основных ("ключевых") объектов (Рис. 17)	1.Выбор осовных объектов (фильтрация) 2.Кластеризация основных объектов 3.Присоединение других объектов к полученному решению	
2.2.2.	Кластеризация на основе решетки (grid)	Разбиение области на ячейки (cells)	[111, 116, 125]
2.2.3.	Кластеризация на основе решетки (grid) в потоках данных (streams)	Кластеризация в режиме реального времени	[129, 157]
2.2.4.	Кластеризация (композиция): "решетка" - декомпозиция на основе параллельных "подпространств" (Рис. 18)	1.Решетка над "пространством" объектов 2.Анализ частей (регионов) "решетки" 3.Выбор "непустых" регионов 4.Кластеризация "густых" (dense) регионов 5.Кластеризация "редких" (sparse) регионов 6.Объединение решений	
2.2.5.	Кластеризация (расширение): "решетка" - декомпозиция на основе параллельных "подпространств" (Рис. 18) (аналог схемы динамического программирования)	1.Решетка над "пространством" объектов 2.Анализ частей (регионов) "решетки" 3.Выбор "непустых" регионов 4.Кластеризация для "густых" (dense) регионов 5.Расширение "густых" регионов и их решений	
2.2.6.	Декомпозиция "пространства" объектов по типам объектов (k-дольная сеть)	1.Выделение объектов по типам 2.Кластеризация для каждого типа объектов 3.Композиция решений	[127, 139, 140]
3.	Составные стратегии	Комбинация различных методов	[115]

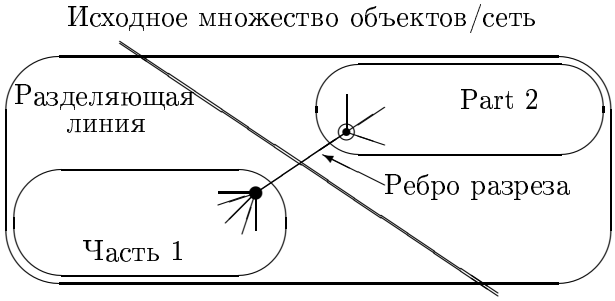


Рис. 16. Ребро "между" для разбиения

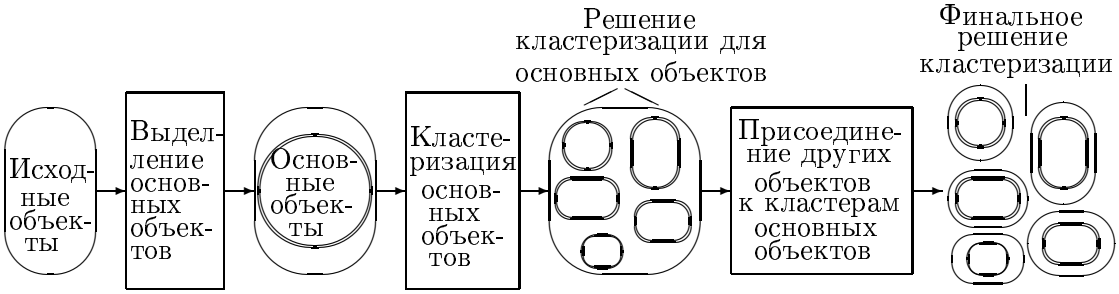


Рис. 17. Кластеризация на основе "основных" объектов

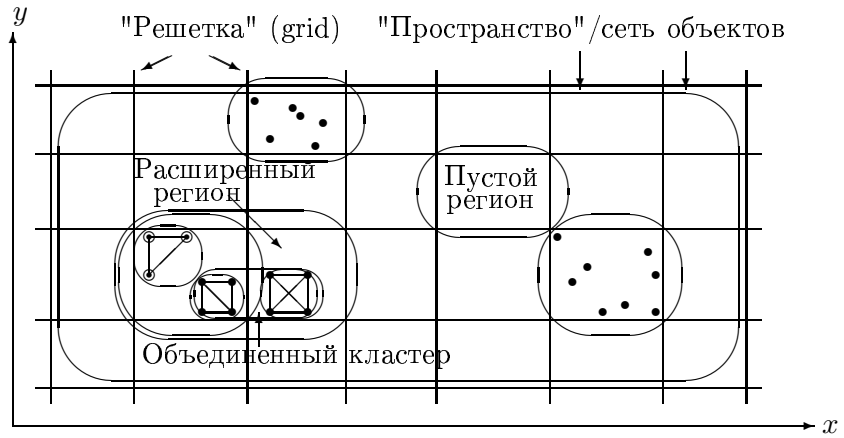


Рис. 18. “Решетка (grid) над "пространством"/сетью объектов

6. ПРИКЛАДНЫЕ ПРИМЕРЫ

6.1. Анализ сети

Анализ сети может быть основан на предварительном разбиении узлов сети на классы-типы по их структурным свойствам. На Рис. 19 показаны основные типы узлов сети.

В результате можно получить стратегию кластеризации на основе многотипных объектов. Типы узлов сети могут быть получены на основе анализа из связей (например, число и типы соседей). Можно выделить типы узлов сети: (1) многосвязанные узлы (тип 1), (2) связанные узлы (тип 2) (например, несколько связей), (3) висячие узлы (outliers) (тип 3), (4) изолированные узлы (тип 4). Пусть имеется граф/сеть $G = (A, E)$, где $A = \{A_1, ..., A_i, ..., A_n\}$ - множество узлов/вершин, E - множество ребер ($|E| = h$).

Таблица 16. Перечень локальных алгоритмов "быстрой" кластеризации

Ном.	"Быстрая" схема	Описание	Оценка	Источник
1.	Иерархический (агломеративный) алгоритм	Объединение ближайших пар объектов ("Снизу-Вверх")	$O(n^3)$	[96]
2.	Сбалансированный по размеру кластера иерархический алгоритм	Объединение ближайших пар объектов ("Снизу-Вверх") с ограничением на размер кластера	$O(n^3)$	
3.	Алгоритм минимального покрывающего дерева	Кластеризация узлов покрывающего дерева	$O(n \log n)$	[74, 75, 138] [154, 160, 181] [197, 201]
4.	Сбалансированный по размеру кластера алгоритм на основе минимального покрывающего дерева	Кластеризация узлов покрывающего дерева с ограничением на размер кластера	$O(n \log n)$	
5.	Кластеризация на основе графа	Выделение сетевых сообществ на основе ребра "между" (betweeness) в графе	$O(p^2 n)$	[69]
6.	Графовый алгоритм на основе модульности	Выделение сетевых сообществ на основе модульности	$O((p+n)n)$ или $O(n^2)$	[142]
7.	Алгоритм на основе решетки (grid) над "пространством координат/параметров объектов"	Назначение объектов по по ячейкам/регионам решетки	$O(n + n' \times n'')$ $(n' \ll n, n'' \ll n)$	[188]
8.	Кластеризация на основе декомпозиции ядер (cores) сети	Предварительная декомпозиция ядер (cores) покрывающего графа	$O(n^2) + O(h)$	[19]

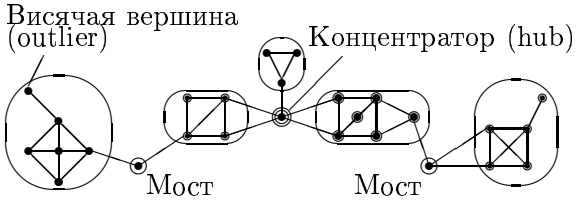


Рис. 19. Иллюстрация типов узлов сети

Можно рассмотреть схему кластеризации:

Стадия 1. Построение списка узлов с информацией о каждом узле (данные о связях/соседях) (оценка сложности - $O(n)$.

Стадия 2. Выбор подмножества узлов с многими связями (тип 1) ($O(n)$). Результат: $B_1 \subset A$.

Стадия 3. Выбор "висячих" узлов (узлы с одной связью - "лист" (outlier)) (тип 3) ($O(n)$). Результат: $B_3 \subset \{A \backslash B_1\}$.

Стадия 4. Выбор других узлов (тип 2) ($O(n)$). Результат: $B_2 \subset A$, $B_2 = \{A \backslash \{B_1 \bigcup B_3\}$.

Стадия 5. Кластеризация элементов множества B_1 (оценка сложности - $O(|B_1|^2)$ (примерно $O((n/3)^2)$). Таким образом получается предвартельное решение кластеризации: $\widehat{X}^1 = \{X_1^1, ..., X_l^1, ..., X_{q_1}^1\}$. Теперь строится макро-сеть: $G^1 = (\widehat{X}^1, E^1)$, где $\widehat{X}^1$ - сеть узлов, которая соответствует полученному решению кластеризации (т.е., множество кластеров) E^1 - построенное множество ребер. (Заметим, полученные кластеры можно использовать как центроиды для метода кластеризации "k-means" и применить это метод).

Стадия 6. Кластеризация узлов типа 2 (множество B_2 (если требуется). Полученное решение: $\widehat{X}^2 = \{X_1^2, ..., X_l^2, ..., X_{q_2}^2\}$. Теперь можно построить "макро-сеть": $G^2 = (\widehat{X}^2, E^2)$, где $\widehat{X}^2$ -сеть узлов, которые соответствуют полученному решению кластеризации (множество кластеров) E^2 - построенное множество ребер.

Стадия 7. Сопоставление двух графов $G^1 = (\widehat{X}^1, E^1)$ и $G^2 = (\widehat{X}^2, E^2)$. Этот процесс может быть основан на анализе ребер исходной сети и/или на использовании дополнительной информации (например, координаты узлов) Получается интегрированное решение кластеризации $\widehat{X}^{12} = \{X_1^{12}, ..., X_l^{12}, ..., X_{q12}^{12}\}$.

Стадия 8. Подсоединение "висячих" вершин (outliers) (B_3) к решению кластеризации $\widehat{X}^{12}$. В результате получается интегрированное решение кластеризации $\widehat{X}^{123}$.

6.2. Планирование работы системы связи с много-лучевой антенной

Имеется описание системы связи с много-лучевой антенной (Рис. 20): (а) система связи с много-лучевой антенной (расположение, объем ресурсов), (б) число лучей антенны: μ, (в) множество узлов связи (пользователей) $A = \{A_1, ..., A_i, ..., A_n\}$ (и их расположение), (г) объем передаваемых данных для каждого узла A_i (т.е., требуемый ресурс, для упрощения предполагается, что узлы примерно одинаковы). Задача заключается в следующем:

Найти расписание связи (доступа) с учетом следующего: (i) минимизация общего времени связи, (ii) обеспечение наилучшего качества связи (посредством наименьшей интерференции между соседними соединениями, по углу):

$$\max_{i \in A} \min_{i_1, i_2 \in A} D^{angular}(A_{i_1}, A_{i_2}),$$

где $D^{angular}(A_{i_1}, A_{i_2})$ угловое разделение (параметр, соответствующий углу между лучами к узлам).

Указанная мера близости определяется так [159]. Имеются элементы $x = (x_1, ..., x_\iota, ..., x_m)$ и $y = (y_1, ..., y_\iota, ..., y_m)$ (т.е., даны векторы координат параметров). Угловая близость для x и y имеет вид: $D^{angular}(x, y) = \frac{\sum_{i=\iota}^{m} x_\iota y_\iota}{[\sum_{i=\iota}^{m} x_\iota^2 \sum_{\iota=1}^{m} y_\iota^2]^{1/2}}$. Мера близости соответствует углу между векторами элементов x и y (т.е., лучами, направленными к элементам из антенны) Схема решения (эвристика) имеет вид:

Стадия 1. Linear ordering of communication nodes by their angle (Рис. 20, узел 1 - 1-й).

Стадия 2. Разбиение полученного списка узлов на μ равных по размеру групп (последняя группа может содержать меньше элементов) и перенумерация узлов следующим образом:

группа 1: $\{(1,1), (1,2), ..., (1,k)\}$,

группа 2: $\{(2,1), (2,2), ..., (2,k)\}$,

...

группа μ: $\{(\mu,1), (\mu,2), ..., (\mu,k)\}$.

Здесь $k = \lceil \frac{n}{\mu} \rceil$.

Стадия 3. Построение расписания по правилам: Слот j ($j = \overline{1,k}$): j-й элемент из каждой группы ($\zeta = 1, 2...\mu$), т.е, элемент $\{\zeta, j)\}$ (Рис. 21).

Стадия 4. Стоп.

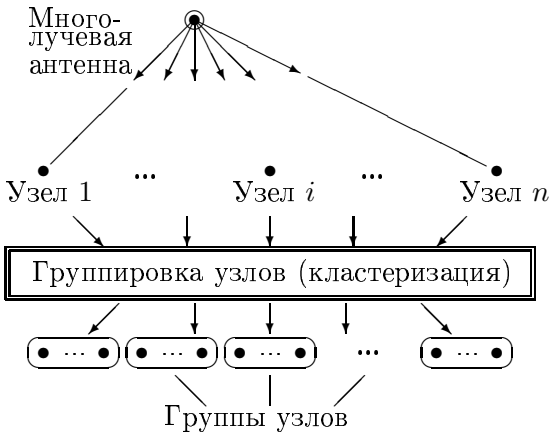


Рис. 20. Система с много-лучевой антенной

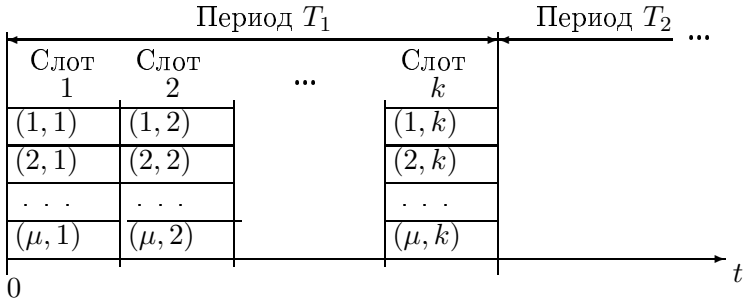


Рис. 21. Схема расписания работы

7. ЗАКЛЮЧЕНИЕ

Статья базируется на "инженерном" взгляде автора на комбинаторную кластеризацию. Представлен анализ литературы: (а) по комбинаторным моделям/методам, (б) по кластеризации данных/сетей большой размерности. Внимание уделено характеристикам качества решений кластеризации, многокритериальной кластеризации. В работе описаны идеи схем решения методов кластеризации на графах: (i) на основе минимального покрывающего дерева, (ii) на основе выделения клик, (iii) на основе оптимизационного разбиения графа (корреляционная кластеризация), (iv) на основе выделения сетевых сообществ. Кратко представлены подходы к построению методов "быстрой" кластеризации. Два прикладных примера ориентированы на сети связи (кластеризация узлов сети, планирование доступа в системе с много-лучевой антенной). Направления будущих исследований могут включать следующее: *1.* построение новых составных комбинаторных схем решения; *2.* проектирование специальных модульных средств поддержки комбинаторной кластеризации; *3.* исследование моделей/ методов динамической кластеризации; *4.* применение комбинаторной кластеризации в сетевых системах связи (проектирование, маршрутизация, управление и мониторинг).

Исследование выполнено в ИППИ РАН при финансовой поддержке РНФ в рамках научного проекта 14-50-00150 “Цифровые технологии и их применения”. Автор благодарит проф. А.И. Ляхова за предварительное инженерное описание задачи планирования работы системы связи с много-лучевой антенной.

СПИСОК ЛИТЕРАТУРЫ

1. Abello J., Resende M.G.C., Sudarsky S., Massive quasi-clique detection. In: S. Rajsbaum (ed), *5th Latin American Symp. on Theor. Inform. LATIN 2002*, LNCS 2286, Springer, 598–612, 2002.
2. Achtert E., Bohm C., Kriegel H.-P., Kroger P., Zimek A., Robust, complete, and efficient correlation clustering. In: *SIAM Int. Conf. on Data Mining (SDM)*, 413–418, 2007.
3. Achtert E., Bohm C., David J., Kroger P., Zimek A., Global correlation clustering based on the hough transform. *Statistical Analysis and Data Mining* 1, 111–127, 2008.
4. Agarwal G., Kempe D., Modularity maximizing network communitites using mathematical programming. *The Eur. Physical J. B* 66, 4009–418, 2008.
5. Aggarwal C.C. (ed.), *Data Streams: Models and Algorithms*. New York: Springer, 2007.
6. Agrawal R., Gehrke J., Gunopulos D., P. Raghavan, Automatic subspace clustering of high dimensional data. *Data Mining & Knowl. Discov.* 11(5), 5–33, 2005.
7. Aho A.V., Hopcroft J.E., Ullman J.D., *The Design and Analysis of Computer Algorihtms*. Reading, MA: Addison Wesley, 1974.
8. Ailon N., Charikar M., Newman A., Aggregating inconsistent information: Ranking and clustering. *J. of the ACM* 55(5), art. No. 23, 2008.
9. Akkoyunlu E., The enumeration of maximal cliques of large graph. *SIAM J. on Comput.* 2(1), 1-6, 1973.
10. Alon N., Krivelevich M., Sudakov B., Finding a large hidden clique in a random graph. In: *Ninth Ann. ACM-SIAM Symp. on Discr. Alg.*, SIAM, 594–598, 1998.
11. Augeri C.J., Ali H.H., New graph-based algorithms for partitioning VLSI circuits. In: *ISDAS'04*, vol. 4, 521–524, 2004.
12. Ayad H., Kamel M.S., On voting-based consensus of cluster ensembles. *Pattern Recogn.* 43(5), 1943–1953, 2010.
13. Babel L., A fast algorithm for the maximum weight clique problem. *Computing* 52, 31–38, 1994.
14. Babu G., Nurty M., Clustering with evolution strategy. *Pattern Recogn. Lett.* 14(10), 763–769, 1993.
15. Bagon S., Galun M., Optimizing large scale correlation clustering. Electr. prepr., 9 p., Dec. 13, 2011. http://arxiv.org/abs/1112.2903 [cs.CV]
16. Balas E., Chvatal V., Nesetril J., On teh maximum weight clique problem. *Math. Oper. Res.* 12(3), 522–535, 1987.
17. Bansal N., Blum A., Chawla S., Correlation clustering. In: *FOCS*, 2002, 238–250, 2002.
18. Bansal N., Blum A., Chawla S., Correlation clustering. *Machine Learning* 56(1-3), 89–113, 2004.
19. Batagelj V., Zavershik M., An O(m) algorithm for cores decomposition of networks. Electr. prepr., 10 p., Oct. 25, 2003. http://arxiv.org/abs/0310.0049 [cs.DS]
20. Ben-Dor A., Shamir R., Yakhini Z., Clustering gene expression patterns. *J. of Computational Biology* 6(3-4), 281–292, 1999.
21. Berkhin P., A survey of clustering data mining techniques. In: *Grouping Multidimensional Data*, Springer, 25–71, 2006.
22. Blondel V.D., Guillaume J.-L., Lambiotte R., Lefebvre E., Fast unfolding of communities in large networks. Electr. prepr., 12 p., July 25, 2008. http://arxiv.org/abs/0803.0476 [physics.soc-ph]
23. Blondel V.D., Esch M., Chan C., Clerot F., Deville P., Huens E., Morlot F., Smoreda Z., Ziemlicki C., Data for development the d4d challenge on mobile phone data. Electr. prepr., 10 p., Jan. 28, 2012. http://arxiv.org/abs/1210.0137 [cs.CY]

24. Boccaletti S., Latora V., Moreno Y., Chavez M., Hwang D.-U., Complex netwoprks: Structure and dynamics. *Physics Reports* 424, 175—208, 2006.

25. Boley D., Gini M., Gross R., Han S., Hastings K., Kapyris G., Kumar V., Mobasher B., Moor J., Partitioning-based clustering of web document categorization. *DSS* 25(3), 329–341, 1999.

26. Bomze I.M., Budinich M., Pardalos P.M., Pelillo M., The maximum clique problem. In: D.-Z. Du, P.M. Pardalos (eds.), *Handbook of Comb. Optim.*. (Supp. vol. A), Springer, New York, 659-729, 1999.

27. Brandes U., Delling D., Gaertler M., Gorke R., Hoefer M., Nikolosk Z., Wagner D., On modularity clustering. *IEEE Trans KDE* 20, 172–188, 2008.

28. C. Bron, J. Kerbosch, Algorithm 457: Finding all cliques of an undirected graph. *Commun. of the ACM* 16(9), 575–577, 1973.

29. Brown D., Huntley C., A practical application of simulated annealing to clustering. *Pattern Recogn.* 25(4), 401–412, 1992.

30. Butenko S., Wilhelm W., Clique-detection models in computational biochemistry and genomics. *EJOR* 173(1), 1–17, 2006.

31. Cai Z., Lu M., Wang X., Channel access-based self-organized clustering in ad hoc networks. *IEEE Trans. Mobile Comput.* 2(2), 102–113, 2003.

32. Charikar M., Guruswami V., Wirth A., Clustering with quantitative information. In: *FOCS 2003*,524–533, 2003.

33. Charikar M., Guruswami V., Wirth A., Clustering with quantitative information. *J. of Comput. Syst. Sci.* 71(3), 360–383, 2005.

34. Charon I., Hundry O., Optimal clustering in multipartite graph. *Disc. Appl. Math.* 156(8), 1330–1347, 2008.

35. Chen C.-Y., Ye F., Particle swam optimization algorithm and its application to cluster analysis. In: *2004 IEEE Int. Conf. on Netw., Sens. & Contr.*, vol. 2, 789-794, 2004.

36. Chen Y.P., Liestman A.L., Maintaining weakly-connected dominating sets for clustering ad hoc networks. *Ad Hoc Netw.* 3, 629–642, 2005.

37. Chen P., Redner S., Community structure of the physical review citation network. *J. of Informetrics* 4(3), 278–290, 2010.

38. Cheng C.H., A branch-and-bound clustering algorithm. *IEEE Trans. SMC* 25, 895–898, 1995.

39. Clauset A., Newman M.E.J., Moore C., Finding community structure in very large networks. *Physical Review E* 70, no. 066111, 2004.

40. Coble J., Cook D.J., Holder L.B., Structure discovery in sequentially-connected data streams. *Int. J. on Artif. Intell. Tools* 15(6), 917–944, 2006.

41. Cobos C., Mendoza M., Leon E., A hyper-heuristic approach to design and tuning heuristic methods for web document clustering. In: *2011 IEEE Cong. on Evol. Comput. (CEC)*, 1350–1358, 2011.

42. Cokuslu D., Erciyes K., Dagdeviren O., A dominating set based clustering algorithm for mobile ad hoc networks. In: V.N. Alexandrov et al. (eds), *ICCS 2006*, LNCS 3991, Springer, 571–578, 2006.

43. Cokuslu D., Erciyes K., A hierarchical connected dominating set based clustering algorithm for mobile ad hoc networks. In: *15th Int. Symp. MASCOTS'07*, 60–66, 2007.

44. Condon A., Karp R.M., Algorithms for graph partitining on the planted partition model. *Random Struct. & Alg.* 18, 116–140, 2001.

45. Cormen T.H., Leiserson C.E., Rivest R.L., *Introduction to Algorithms*. 3rd ed., MIT Press and McGraw-Hill, 2009.

46. Corneil D.G., Perl Y., Clustering and domination in perfect graphs. *Disc. Appl. Math.* 9(1), 27–39, 1984.

47. Cowgill M.C., Harvey R.J., Watson L.T., A genetic algorithm approach to cluster analysis. *Comput. Math. Appl.* 37(7), 99–108, 1999.

48. Dawande M., Keskinocak P., Swaminathan J.M., Tayur S., On bipartite and multipartite clique problems. *J. of Algorithms* 41(2) (2001) 388-403.

49. de Amorim S.G., Barthelemy J.-P., Ribeiro C.C., Clustering and clique partitioning: Simulated anealing and tabu search approaches. *J. of Classif.* 9(1), 17–41, 1992.

50. Demaine E.D., Immorlica N., Correlation clusteirng with partial information. In: *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, Springer, 1–13. 2003.

51. Demaine E.D., Emanuel D., Fiat A., N. Immorlica, Correlation clustering in general weighted graphs. *Theor. Comp. Sci.* 361(2), 172–187, 2006.

52. Ding C.H.Q., He X., Zha H., Gu M., Simon H.D., A min-max algorithm for graph partitioning and data clustering. In: *ICDM 2001*, 107–114, 2001.

53. Djidjev H.N., A scalable multilevel algorithm for graph clustering and community structure detection. In: W. Aiello et. al. (eds), *WAW 2006*, LNCS 4936, Springer, 117–128, 2008.

54. Duan D., Li Y., Li R., Lu Z., Incremantal K-clique clustering in dynamic social networks. *Artif. Intell. Rev.* 38(2), 129–147, 2012.

55. Duch J., Arenas A., Community detection in complex networks using extremal optimization. *Physical Review E* 72, no. 027104, 2005.

56. Elsner M., Schudy W., Bounding and comparing methods for correlation clustering beyond ILP. In: *NAACL HLT Workshop on Integ. Lin. Progr. for Nat. Lang. Proc.*, 19–27, 2009.

57. Emanuel D., Fiat A., Correlation clustering - minimizing disagreements on arbitrary weighted graphs. In: *Algorihtms-ESA 2003*, Springer, 208–220, 2003.

58. Even G., Naor J., Rao S., Schieber B., Fast approximate graph partitioning algorithms. *SIAM J. on Comput.* 28(6), 2187–2214, 1999.

59. Feige U., Krauthgamer R., Finding and certifying a large clique in a semi-random graph. *Random Struc. Alg.* 16(2), 195–208, 2000.

60. Feldman D., Langberg M., A unified framework for approximating and clustering data. In: *STOC 2011*, 569–578, 2011.

61. Feldman A.E., Foschini L., Balanced partitions of trees and applications. *Algorithmica* 71(2), 354–376, 2015.

62. Fellows M.R., Guob J., Komusiewicz C., Niedermeier R., Uhlmann J., Graph-based data clustering with overlaps. *Disc. Optim.* 8(1), 2–17, 2011.

63. Fortunato S., Community detection in graphs. Electr. prepr., 103 p., Jan. 25, 2010. http://arxiv.org/abs/0906.0612v2 [physics.soc-ph]

64. Frahling G., Sohler C., Coresets in dynamic geometric data streams. In: *STOC 2005*, 209–217, 2005.

65. Furems E.M., Dominance-based extension of STEPCLASS for multiattribute nominal classification. *Int. J. of Inform. Technol. & Dec. Making* 12(5), 905–925, 2013.

66. Gabow H.N., Galil Z., Spencer T., Tarjan R.E., Efficient algorithms for finding minimum spanning trees in undirected and directed graphs. *Combinatorica* 6(2), 109–122, 1986.

67. Garey M.R., Johnson D.S., *Computers and intractability. The Guide to the theory of NP-completeness.* San Francisco: W.H. Freeman and Company, 1979.

68. Giotis I., Guruswami V., Correlation clustering with a fixed number of clusters. In: *Seventeenth Ann. ACM-SIAM Symp. on Disc. Alg.*, SIAM, 1167–1176, 2006.

69. Girvan M., Newman M.E.J., Community structure in social an biological networks. Community structure in social and biological networks. PNAS 99(12), 7821–7826, 2002.

70. Gramm J., Guo J., Huffner F., Niedermeier R., Graph-modeled data clustering: Fixed-parameter algorithm for clique generation. *Theory of Comput. Syst.* 38(4), 373–392, 2005.

71. Goldberger J., Tassa T., A hierarchical clustering algorithm based on the Hungarian method. *Pattern Recogn. Lett.* 29(11), 1632–1638, 2008.

72. Goldengorin B., Krushinsky D., Pardalos P.M., *Cell Formation in Industrial Engineering: Theory, Algorithms and Experiments.* Springer, 2013.

73. Gopalan P.K., Blei D.M., Efficient discovery of overlapping communites in massive networks. *PNAS* 110(36), 14534–14539, 2013.

74. Gower J., Ross G., Minimum spanning trees and single linkage cluster analysis. *J. of the Royal Statistical Society, Series C (Applied Statistics)* 18(1), 54–64, 1969.

75. Grygorash O., Zhou Y., Jorgensen Z., Minimum spanning tree based clustering algorithms. In: *ICTAI'06*, 73–81, 2006.

76. Guenoche A., Consensus partitions: a constructive approach. *Adv. Data Anal. and Classif.* 5(3), 215–229, 2011.

77. Guha S., Mishra N., Motwani R., O'Callagham L., Clustering data streams. In: *FOCS 2000*, 359–366, 2000.

78. Guimera R., Dadon L., Diaz-Guilera A., Giralt F., Arenas A., Self-similar community structure in a network of human interactions. *Physical Review E* 68, no. 065103, 2003.

79. Guimera R., Sales-Pardo M., Amaral L.A.N., Modularity from fluctuations in random graphs and complex networks. *Physical Review E* 70, no. 025101, 2004.

80. Gunes I., Bingol H., Coomunity detection in complex networks using agents. Electr. prepr., 5 p., Oct. 23, 2006, arXiv:cs/0610129 [cs.MA]

81. Han B., Jia W., Clustering wireless ad hoc networks with weakly connected dominating set. *J. of Parall. and Distr. Comput.* 67(6), 727–737, 2007.

82. Hansen P., Mladenivic N., Variable neighborhood search for the p-median. *Location Science* 5(4), 207–226, 1997.

83. Hansen P., Brimberg J., Urosevic D., Mladenovic N., *Data Clustering using Large p-Median Models and Primal-Dual Variable Neighborhood Search.* TR G-2007-41, 2007 GERAD, June 2007.

84. Hansen P., Brimberg J., Urosevic D., Mladenovic N., Primal-Dual Variable Neighborhood Search for the simple plant-location problem. *INFORMS J. on Computing* 19, 552-564, 2007.

85. Hansen P., Brimberg J., Urosevic D., Mladenovic N., Solving large p-median clustering problems by primal-dual variable neighborhood search. *Data Min. and Knowl. Discov.* 19(3), 351–375, 2009

86. Har-Peled S., Mazumdar S., On coresets for k-mean and k-median clustering. In: *STOCS 2004*, 291–300, 2004.

87. Hopcroft J., Khan O., Kulis B., Selman B., Natural communities in large linked networks. In: *Ninth ACM SIGKDD Int. Conf. on Knowl. Discov. & Data Min. KDD'03*, 541–546, 2003.

88. Hopcroft J., Khan O., Kulis B., Selman B., Tracking evolving communities in large linked networks. *PNAS* 101(Suppl 1), 5249–5353, 2004.

89. Hou T.C., Tsai T.-J., An access-based clustering protocol for multihop wireless ad hoc networks. *IEEE J. on Selec. Areas in Commun.* 19(7), 1201–1210, 2001.

90. Hruschka E.R., Campello R.G.B., Freitas A.A., Carvalho A.P.L., A survey of evolutionary algorithms for clustering. *IEEE Trans. SMC, Part C* 39(2), 133–155, 2009.

91. Huang Z., Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining Knowl. Discov* 2(3), 283–304, 1998.

92. Indukuri R.K.R., Penumathsa S.V., Dominating sets and spanning tree based clustering algorithms for mobile ad hoc networks. *Int. J. of Adv. Comp. Sci. & Appl.* 2(2), 75–81, 2011.

93. Ivanov A.S., Lyakhov A.I., Khorov E.M., Analytical model of batch flow multihop transmission in wireless networks with channel reservation. *Autom. and Rem. Cont.*, July 2015 (in press)

94. Jain A.K., Data clustering: 50 years beyond k-means. *Pattern Recogn. Lett.* 31(8), 651–666, 2010.

95. Jain A.K., Dubes R.C., *Algorithms for Clustering Data.* Upper Saddle River, NJ: Prentice Hall, 1988.

96. Jain A.K., Murty M.N., Flynn P.J., Data clustering: a review. *ACM Comput. Surv.* 31(3) (1999) 264–323.

97. Joachims T., Hopcroft J., Error bounds for correlation clustering. In: *22rd Int. conf. on Mach. Learn. ICML'05*, ACM, 385–392, 2005.

98. Johnson D.S., Trick M.A. (eds), *Cliques, Coloring, and Satisfiability.* DIMACS Ser. in Disc. Math. and Theor. Comp. Sci., vol. 26, AMS, Providence, 1996.

99. Jovanovic R., Tuba M., Voss S., An ant colony optimization algorithm for partitioning graphs with supply and demand. Electr. prep. 21 p., March 3, 2015. http://arxiv.org/abs/1503.00899 [cs.AI]

100. Kargin I., Khorov E., Lyakhov A., A mathematical method to estimate packet loss ratio for a multipath route with error correlation. *Probl. of Inform. Transmission*, 2015 (in press)

101. Karp R.M., Reducibility among combinatorial problems. In: R.E. Miller, J.W. Thatcher (eds), *Complexity of Computer Computations*, Plenum, pp. 85–103, 1972.

102. Kernigham B., Lin S., An efficient heuristic procedure for partitioning graphs. *Bell Syst. Techn. J.* 49, 291–307, 1970.

103. Khorov E., Lyakhov A, Krotov A., Guschin A., A survey on IEEE 802.11ah: an enabling networking technology for smart cities. *Comput. Commun.* 58, 53–69, 2015

104. Khorov E.M., Kiruanov A.G., Kureev A.A., Lyakhov A., Study of mechanism for building a logical network topology in MANET. *J. of Commun. Technol. & Electr.* 60(12), 2015.

105. Khorov E., Krotov A., Lyakhov A., Modeling machine type communication in IEEE 802.11ah network. In: *IEEE Int. Conf. on Commun.-Workshop on 5G & Beyond Enabl. Technol. & Appl.*, 2015.

106. Kim S., Nowozin S., Kohli P., Yoo C.D., Higher-order correlation clustering for image segmentation. In: *Advances in Neural Information Processing Systems*, 1530–1538, 2011.

107. Kleinberg J.M., Papadimitriou C., Raghavan P., Segmentation problems. In: *STOC'98*, 473–482, 1998.

108. Knuth D.E., Raghunathan A., The problem of compatible representatives. *SIAM J. on Disc. Math.* 5(3) (1992) 422–427.

109. Kochenberg G., Glover F., Alidaee B., Wang H., Clustering of microarray data via clique partitioning. *J. of Combin. Optim.* 10(1), 77–92, 2005.

110. Kriegel H.-P., Kroger P., Schubert E., Zimek A., A general framework for increasing the robustness of PCA-based correlation clustering algorithms. In: *20th Int. Conf. SSDBM*, Springer, 418–435, 2008.

111. Kriegel H.-P., Kroger P., Zimek A., Clustering high dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Trans. KDD* 3(1), 1–58, 2009.

112. Kroese D.P., Rubinstein R.Y., Taimre T., Application of the cross-entropy method for clustering and vector quantization. *J. of Global Optim.* 37(1), 137–157, 2007.

113. Kumar V., Steinbach M., Tan P.-N., *Introduction to Data Mining.* Addison-Wesley, 2005.

114. Kumari A.C., Srinivas K., Software module clustering using a fast multi-objective hyper-heuristic evolutionary algorithm. *Int. J. of Appl. Inform. Syst.* 5(6), 12–18, 2012.

115. Kyperountas M., Tefas A., Pitas I., Dynamic training using multistage clustering for face recognition. *Pattern Recogn.* 41(3), 894–905, 2008.

116. Lai Y.C., Lin P., Liao W., Chen C.M., A region-based clustering mechanism for channel access in vehicular ad hoc networks. *IEEE J. on Selec. Areas in Commun.* 29(1), 83–93, 2011.

117. Leskovec J., Lang K.J., Dasgupta A., Mahoney M.W., Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Math.* 6, 29–123, 2009.

118. Levin M.Sh., *Combinatorial Engineering of Decomposable Systems.* Dordrecht: Kluwer, 1998.

119. Levin M.Sh., *Composite Systems Decisions.* New York: Springer, 2006.

120. Levin M.Sh., Aggregation of composite solutions: strategies, models, examples. Electr. prepr., 72 p., Nov. 29, 2011. http://arxiv.org/abs/1111.6983 [cs.SE]

121. Levin M.Sh., Multiset estimates and combinatorial synthesis. Electr. prep., 30 p., May 9, 2012. http://arxiv.org/abs/1205.2046 [cs.SY]

122. Levin M.Sh., Clique-based fusion of graph streams in multi-function system testing. *Informatica* 23(3), 391–404, 2012.

123. Levin M.Sh., *Modular System Design and Evaluation.* Springer, 2015.

124. Levin M.Sh., Towards combinatorial clustering: preliminary research survey. Electr. prepr., 102 pp., May 28, 2015. http://arxiv.org/abs/1505.07872 [cs.AI]

125. Lin N.P., Chang C.-I., Chueh H.-E., Chen H.-J., Hao W.-H., A deflected grid-based algorithm for clustering analysis. *WSEAS Trans. on Computers* 3(7), 125–132, 2007.

126. Liu X., Li D., Wang S., Tao Z., Effective algorithm for detecting community structure in complex networks based on GA and clustering. In: Y. Shi et. al. (eds), *7th Int. Conf. on Comput. Sci. ICCS'07*, Part II, LNCS 4488, Springer, 657-664, 2007.

127. Liu X., Murata T., Detecting communities in k-partite k-uniform (hyper)networks. *J. of Comp. Sci. and Technol.* 26(5), 778–791, 2011.

128. Liu X., Murata T., Wakita K., Extending modularity by capturing the similarity attraction feature in the null model. Electr. prepr., 10 p., Feb. 12, 2013. http://arxiv.org/abs/1210.4007 [cs.SI]

129. Lu Y., Sun Y., Xu G., Liu G., A grid-based clustering algorithm for high-dimensional data streams. In: *Advanced Data Mining and Applications*, Springer, 824–831, 2005.

130. Mathieu C., Schudy W., Correlation clustering with noisy input. In: *Twenty-first Ann. ACM-SIAM Symp. on Disc. Alg.*, SIAM, 712–728, 2010.

131. Mehrotra A., Trick M.A., Cliques and clustering: A combinatorial approach. *Oper. Res. Lett.* 22(1), 1–12, 1998.

132. Medius A., Acuna G., Dorso C.O., Detection of community strcuture in networks via global optimization. *Physica A* 358, 396–405, 2005.

133. Mimaroglu S., Yagci M., CLICOM: Cliques for combining multiple clusterings. *ESwA* 39(2), 1889–1901, 2012.

134. Mirkin B., Muchnik I., Combinatorial optimization in clustering. In: D.-Z. Du, P.M. Pardalos (Eds.), *Handbook of Combinatorial Optimization*, vol. 2, Springer, New York, 261–329, 1999.

135. Mitchell M., Complex systems: Network thinking. *Artif. Intell.* 179, 1194–1212, 2006.

136. Moon J.W., Moser L., On cliques in graphs. *Israel J. of Math.* 3(1), 23–28, 1965.

137. Muller E., Assent I., Gunnemann S., Krieger R., Seidl T., Relevant subspace clustering: Mining the most interesting non-redundent concepts in high dimensional data. In: *ICDM'09*, 377-386, 2009.

138. Muller A.C., Nowozin S., Lampert C.H., Information theoretic clustering using minimum spanning trees. In: A. Pinz et al. (eds.), *Joint 34th DAGM and 36th OAGM Symp. Pattern Recognition*, LNCS 7476, Springer, 205–2015, 2012.

139. Murata T., Detecting communities from tripartite networks. In: *World Wide Web Conf. (WWW'2010)*, 1159–1160, 2010.

140. Murata T., Modularity for heterogeneous networks. In: *21th ACM Conf. on Hypertext and Hypermedia HyperText2010*, 129–134, 2010.

141. Naeni L.M., Berretta R., Moscano P., MA-Net: A reliable memetic algorithm for community detection by modularity optimization. In: H. Handa et. al. (eds), *18th Asia Pac. Symp. on Intell. & Evol. Syst.*, Springer vol. 1, 311-323, 2015.

142. Newman M.E.J., Fast algorithm for detecting community structure in networks. Electr. prepr., 5 p., Sep. 22, 2003. http://arxiv.org/abs/0309508 [cond-mat.stat-mech]

143. Newman M.E.J., Detecting community structure in networks. *Eur. Phys. J. B* 38(2), 321–330, 2004.

144. Newman M.E.J., Modularity and community structure in networks. *PNAS* 103(23), 8577–8582, 2006.

145. Newman M.E.J., *Networks: an Introduction.* Oxford: Oxford Univ. Press, 2010.

146. Newman M.E.J., Girvan M., Finding and evaluating community structure in networks. Electr. prepr., 16 p., Aug. 11, 2003. http://arxiv.org/abs/0308217 [cond-mat.stat-mech]

147. Noack A., Rotta R., Multi-level algorithms for modularity clustering. Electr. prepr., 12 p., Dec. 22, 2008. http://arxiv.org/abs/0812.4073 [cs.DC]

148. Oosten M., Rutten J.G.C., Spieksma F.C.R., The clique partioning problem: Facets and patching facets. *Networks* 38(4), 209–226, 2001.

149. Osman I.H., Christofides N., Capacitated clustering problems by hybrid simulated annealing and tabu search. *Int. Trans. on Oper. Res.* 1(3), 317–336, 1994.

150. Osteen R.E., Tou J.T., A clique-detection algorithm based on neighborhoods in graphs. *Int. J. of Comp.&Inform. Sci.* 2(4), 257–268, 1973.

151. Ostergard P.R.J., A new algorithm for the maximum-weight clique problem. In: *Electr. Notes in Disc. Math., 6th Twente Workshop on Graphs&Comb. Optim.*, vol. 3, 153–156, 1999.

152. Ovelgonne M., Geyer-Schulz A., A comparison of agglomerative hierarchical algorithms for modularity clustering. In: *Conf. on Challenges at the Interface of Data Anal., Comp. Sci., and Optim.*, Springer, 225–232, 2012.

153. Ozyer T., Alhajj R., Parallel clustering of high dimensional data by integrating multi-objective genetic algorithm with divide and conquer. *Appl. Intell.* 31(3), 318–331, 2009.

154. Paivinen N., Clustering with a minimum spanning tree of scale-free-like structure. *Pattern Recogn. Lett.* 26(7), 921–930, 2005.

155. Pardalos P.M., Xue J., The maximum clique problem. *J. of Global Optimization*, 4(3), 301–328, 1994.

156. Pardalos P., Batzyn M., Maslov E., Cliques and quasi-cliques in large graphs: theory and applications. In: *Int. Conf. on Disc. Optim.&Oper. Res. DOOR-2013*, Novosib., Sobolev Inst. of Math., 24–28, 2013.

157. Park N.H., Lee W.S., Statistical grid-based clustering over data streams. *ACM SIGMOD Record* 33(1), 32-37, 2004.

158. Pavan M., Pelillo M., Dominant sets and pairwise clustering. *IEEE Trans. PAMI* 29(1), 167–172, 2007.

159. Pedrycz W., *Knowledge-Based Clustering: From Data to Information Granules.* Hoboken, NJ: Wiley, 2005.

160. Peter S.J., Victor S.P., A novel algorithm for dual similarity clusters using minimum spanning tree. *J. of Theor. & Appl. Inform. Technol.* 14(1), 60–66, 2010.

161. Pettie S., Ramashandran V., An optimal minimum spanning tree algorithm. *J. of the ACM* 49(1), 16–34, 2002.

162. Pons P., Latapy M., Computing communities in large networks using random works. *J. of Graph Alg. & Appl.* 10, 191–218, 2006.

163. Porter M.A., Onnela J.-P., Mucha P.J., Communities in networks. *Notices of the AMS* 56(9), 1082–1097,1164, 2009.

164. Reichardt J., Bornholdt S., Statistical mechanics of community detection. *Physical Review E* 74, no. 016110, 2006.

165. Rocha C., Dias L.C., Dimas I., Multicriteria classification with unknown categories: A clustering-sorting approach and an application to conflict management. *J. of Multi-Cri. Dec. Anal.* 20(1-2), 13–27, 2013.

166. Rocha C., Dias L.C., MPOC - an agglomerative algorithm for multicriteria partially ordered clustering. *4OR* 11(3), 253–273, 2013.

167. Rosvall M., Bergstrom C.T., An information-theoretic framework for resolving community structure in complex networks. *PNAS* 104(18), 7327–7331, 2007.

168. Roy B., *Multicriteria methodology for decision aiding.* Dordrecht: Kluwer, 1996.

169. Rubinstein R.Y., Cross-entropy and rare-events for maximal cut and partition problems. *ACM Trans. on Model. & Comp. Simul.* 12(1), 27–53, 2002.

170. Saeed F., Salim N., Abdo A., Voting-based consensus clustering for combining multiple clusterings of chemical structures. *J. of Cheminformatics* 4(37), 1–8, 2012.

171. Salzmann J., Behnke R., Gag M., Timmermann D., 4-MASCLE - improved coverage aware clustering with self healing abilities. In: *IEEE UIC-ATC'09*, 537–543, 2009.

172. Schaeffer S.E., Graph clustering. *Computer Sci. Rev.* 1(1), 27–64, 2007.

173. Schenker A., Last M., Bunke H., Kandel A., Classification of web documents using graph matching. *IJPRAI* 18(3), 475–496, 2004.

174. Selim S., Alsultan K., A simulated annealing algorithm for the clustering problems. *Pattern Recogn.* 24(10), 1003–1008, 1991.

175. Selim H.M., Askin R.G., Vakharia A.J., Cell formation in group technology: review, evaluation and direction for future research. *Comp. & Ind. Eng.* 34(1), 3–20, 1998.

176. Shamir R., Sharan R., Tsur D., Cluster graph modification problems. In: *Proc. of 28th WG*, LNCS 2573, Springer, 379–316, 2002.

177. Shamir R., Sharan R., Tsur D., Cluster graph modification problems. *Disc. Appl. Math.* 144(1-2), 173–182, 2004.

178. Sheikholeslami G., Chattterjee C., Zhang A., WaveCluster: a wavelet-based clustering approach for spatial data in very large databases. *The VLDB* J. 8(3-4), 289–304, 2000.

179. Shiokawa H., Fujiwara Y., Onizuka M., Fast algorithm for modularity-based graph clustering. In: *Twenty-Seventh AAAI Conf. on Artif. Intell.*, 1170–1176, 2013.

180. Spielman D.A., Teng S.-H., A local clustering algorithm for massive graphs and its application to nearly linear time graph partitioning. *SIAM J. Comput.* 42(1), 1–26, 2013.

181. Srinivasan G., A clustering algorithm for machine cell formation in group technology using minimum spanning tree. *The Int. J. of Prod. Res.* 32(9), 2149–2158, 1994.

182. Sung C.S., Jin H.W., A Tabu-search-based heuristic for clustering. *Pattern Recogn.* 33(5), 849–858, 2000.

183. Swamy C., Correlation clustering: maximizing agreements via semidifinite programming. In: *Fifteenth Ann. ACM-SIAM Symp. on Disc. Alg.*, SIAM, 526–527, 2004.

184. Tabor J., Spurek P., Cross-entropy clustering. *Pattern Recogn.* 47(9), 3046–3059, 2014.

185. Tillet J., Rao R., Sahin F., Cluster-head identification in ad hoc sensor networks using particle swam optimization. In: *2002 IEEE Int. Conf. on Personal Wireless Commun.*, 201–205, 2002.

186. Trifunovic A., Knottenbelt W.J., Parallel multilevel algorithms for hypergraph partitioning. *J. of Paral. & Distr. Comput.* 68(5), 563–581, 2008.

187. Tsai C.-F., Yen C.-C., ANGEL: a new effective and efficient hybrid clustering techniques for large databases. In: Z.-H. Zhou et. al., *PAKDD 2007*, LNCS 4426, Springer, 817–824, 2007.

188. Tsai C.-F., Yeh H.-F., Chang J.-F., Liu N.-H., PHD: an efficient data clustering scheme using partition space technique for knowledge discovery in large databases. *Appl. Intell.* 33(1), 39–53, 2010.

189. Tsai C.-W., Song H.-J., Chiang M.-C., A hyper-heuristic clustering algorithm. In: *2012 IEEE Int. Conf. on SMC*, 2839–2844, 2012.

190. Tseng L.Y., Yang S.B., A genetic approach to the automatic clustering problem. *Pattern Recogn.* 34(2), 415–424, 2001.

191. Tsuda K., Kudo T., Clustering graphs by wieghted substructure mining. In: *23rd Int. Conf. on Mach. Learn.*, 953–960, 2006.

192. Tumer K., Agogino A.K., Ensemble clustering with voting active clusters. *Pattern Recogn. Lett.* 29(14), 1947–1953, 2008.

193. Van der Merwe D.W., Engelbrecht A.P., Data clustering using particle swam optimization. In: *The 2003 Congr. on Evol. Comput. CEC'03*, vol. 1, 215–220, 2003.

194. Vashist A., Kulikowsky C.A., Muchnik I., Orthlog clustering on a multipartite graph. *IEEE/ACM Trans. Comput. Biology and Bioinformatics* 4(1), 17–27, 2007.

195. Vega-Pons S., Ruiz-Schulcloper J., A survey of clustering ensemble algorithms. *Int. J. of Pattern Recogn. Artif. Intell.* 25(11), 337–372, 2011.

196. Wakita K., Tsusumi T., Finding community structure in mega-scale social networks. Electr. prepr., 9 p., Fev. 8, 2007. http://arxiv.org/abs/0702.2048 [cs.CY]

197. Wang X., Wang X., Wikes D.M., A divide-and-conquer approach for minimum spanning tree-based clustering. *IEEE Trans. KDE* 21(7), 945–958, 2009.

198. Wang Q., Fleury E., Overlapping community structure and modular overlaps in complex networks. In: T. Ozyer et. al., *Mining Soc. Netw. & Sec. Inform.*, Lect. Not. in Soc. Netw., Springer, 15–40, 2013.

199. White S., Smyth P., A spectral clustering approach to finding communities in graph. In: *SIAM Data Mining Conf.*, 76–84, 2005.

200. Xie J., Kelley S., Szymanski B.K., Overlapping community detection in networks: The state-of-the-art and comparative study. *ACM Comp. Surv.* 45(4), art. 443, 2013

201. Xu Y., Olman V., Xu D., Minimum spanning trees for gene expression data clustering. *Genome Informatics* 12, 24–33, 2001.

202. Xu X., Yuruk N., Feng Z., Schweiger T.A.J., SCAN: a structural clustering algorithm for networks. In: *SIGKDD'07*, 824–833, 2007.

203. Yan B., Gregory S., Detecting communities in networks by merging cliques. In: *ICIS 2009*, 832–836, 2009.

204. Yang Y., Kamel M.S., An aggregated clustering approach using multi-ant colonies algorithms. *Pattern Recogn.* 39(7), 1278–1289, 2006.

205. Yang J., Leskovec J., Overlapping community detection at scale: A nonnegative matrix factorization approach. In: *WSDM 2013*, 587–596, 2013.

206. Yang J., Leskovec J., Overlapping communities explain core-periphery organization of networks. *Proc. of the IEEE* 102(12), 1892–1902, 2014.

207. Yang J., Leskovec J., Structure and overlaps of ground-truth communities in networks. *ACM Trans. on Intell. Syst. & Technol. (TIST)* 15(2), art. 26, 2014.

208. Yang J., Leskovec J., Designing and evaluation network communities based on ground-truth. *Knowl. Inf. Syst.* 42(1), 181–213, 2015.

209. Yao A.C., An $O(|E|loglog|V|)$ algorithm for finding minimum spanning trees. *Inform. Process. Lett.* 4(1), 21–23, 1975.

210. Yeh D.Y., A dynamic programming approach to the complete set partitioning problem. *BIT Numerical Mathematics* 26(4), 467–474, 1986.

211. Younis O., Krunz M., Ramasubramanian S., Node clustering in wireless sensor networks: Recent developments and deployment challenges. *IEEE Networks* 20(3), 20–25, 2006.

212. Zachary W.W., An information flow model for conflict and fission in small groups. *J. of Anthropological Research* 33, 452–473, 1977.

213. Ziv E., Middendorf M., Wiggins C., Information-theoretic approach to network modularity. *Physical Review* E 71, no. 046117, 2005.

214. Zopounidis C., Doumpos M., Multicriteria classification and sorting methods: a literature review. *EJOR* 138(2), 229–246, 2002.

Towards Combinatorial Clustering: literature review, methods, examples

Levin M.Sh.

The paper addresses clustering problems from combinatorial viewpoints. A systemic survey is presented. The list of considered issues involves the following: (1) literature analysis of basic combinatorial methods and clustering of very large data sets/networks; (2) quality characteristics of clustering solutions; (3) multicriteria clustering models; (4) graph based clustering methods (minimum spanning tree based clustering methods, clique based clustering as detection of cliques/quasi-cliques, correlation clustering, detection of network communities); and (5) fast clustering approaches. Mainly, the presented material is targeted to networking. Numerical examples illustrate models, methods and applications.

KEYWORDS: clustering, classification, combinatorial optimization, multicriteria decision making, heuristics, network applications