УДК 510.63

Кривша Н.С., Кривша В.В., Бутенков С.А.

ЭФФЕКТИВНЫЕ АЛГОРИТМЫ ГРАНУЛЯЦИИ ДАННЫХ НА ОСНОВЕ ПРОСТРАНСТВЕННОЙ ГРАНУЛЯЦИИ

Инженерно технологическая академия Южного Федерального Университета,
Россия, Таганрог, пер. Некрасовский, 44, 347922
ООО "НИЦ супер-ЭВМ и нейрокомпьютеров",
Россия, Таганрог, пер. Итальянский, 106, 347900

Krivsha N.S., Krivsha V.V., Butenkov S.A.

EFFICIENT DATA GRANULATION ALGORITHMS BASED ON SPATIAL GRANULATION

Institute of Computer Technologies of Southern Federal University,
Russia, Taganrog, Nekrasovskij st., 44, 347922
Scientific Research Center of Super-computer and Neuro-computer,
Russia, Taganrog, Italyanskij st., 106, 347900

Аннотация. В работе рассматривается жадные алгоритмы, с высокой эффективностью выполняющие обработку данных измерений путем перевода в гранулированное представление. Такая форма представления содержит ту же информацию, что и исходные данные, но требует меньше памяти для хранения и меньше вычислительных затрат для обработки.

Ключевые слова: мягкие вычисления, информационная грануляция, гранулированные вычисления, жадные алгоритмы, анализ данных.

Abstract. In this paper we describe the use of kind of greedy algorithms for the efficient data processing by mean of conversion to the granular space. The granular representation contains as soon information as original data, but require the smaller memory size and computer resources.

Key words: soft computing, information granulation, granular computing,

greedy algorithms, data mining.

Вступление.

Широкое использование вычислительной техники для извлечения, обработки и накопления различных видов данных требует разработке все более эффективных методов и алгоритмов решения задач, связанных с извлечением знаний из больших объемов разнотипных данных в соответствии с *информационными* критериями. Одним из разделов методологии искусственного интеллекта, занимающимся подобным интеллектуальным анализом данных является Data Mining, DM. Термин Data Mining является «зонтичным» и включает в себя задачи: грануляции (кластеризации), классификации, регрессии, поиска ассоциативных правил и обработки данных.

Описательная задача грануляции является задачей начального этапа и предназначена для улучшения понимания и наглядного представления неупорядоченных и неформализованных данных. Как правило, грануляция применяется, когда о данных ничего не известно, либо объем информации огромен. В связи с этим, на первый план выходит разработка высокоэффективных алгоритмов грануляции многомерных данных в форме, пригодной для дальнейшей обработки.

Обзор литературы.

Классические подходы в задачах грануляции данных изложены в [1] и основываются, в основном, на методах кластер-анализа [2]. Важнейшей особенностью таких подходов является требование к полной определенности измерений данных. Методы более широкого класса, использующие модели неточности в виде *нечетких множеств* (*Fuzzy Sets*, FS) описаны в [3], [4]. На практике приходится иметь дело с данными в условиях неопределённости типа *грануляции* данных [5], являющейся обобщением понятия кластеризации [2]. Одним из возможных подходов к обобщению данных и их информационному сжатию является метод пространственной грануляции [6], вводящий методы гранулированных вычислений для обработки многомерных данных (например, изображений) [7]. В работе [8] в развитие методологии [7] были введены

нечеткие топологические отношения на элементах гранулированного представления исходных данных [9]. На этой основе становится возможным построение эффективных алгоритмов грануляции из класса жадных алгоритмов.

Входные данные и методы.

В теории информационной грануляции (ТИГ) L.A. Zadeh *информационной гранулой* называется подмножество универсума, на котором определено отношение сходства, неразличимости и т.п. [4]. Множество гранул, которое содержит все объекты универсума, называется *гранулированием* универсума. Приведем необходимые определения для формализации метода грануляции данных согласно [6].

Определение 1. Разбиение конечного универсума **G** – это конечное множество подмножеств ${}^{i}G \in \mathbf{G}$, $i = 1,...,n$ (*атомарных гранул*), удовлетворяющих следующим аксиомам:

1. ${}^{i}G \neq \varnothing, \quad i = 1,...,n;$

2. ${}^{i}G \cap {}^{j}G = \varnothing$ при $i \neq j; \quad i = 1,...,n; \; j = 1,...,n;;$

3. $\bigcup_{i} {}^{i}G = \mathbf{G}, \quad i = 1,...,n.$

Каждое подмножество разбиения называется гранулой эквивалентности. Подмножество ${}^{i}G \subseteq \mathbf{G}$ называется составной гранулой (не элементарной) если оно представляет собой объединение атомарных гранул определения 4 [9].

Математической основой построения массовых эффективных (линейной сложности) алгоритмов (жадных алгоритмов) является преобразование исходной задачи к задаче на матроиде. Известна теорема, показывающая, что на конечном множестве с неотрицательной весовой функцией можно построить жадный алгоритм поиска независимого множества с максимальным весом. К такой задаче можно свести многие классические численные задачи, в том числе, и задачу кластеризации [2]. Покажем, что пространство гранулированных данных является матроидом. Рассмотрим аксиоматический подход к построению алгебраических структур на множествах (пространствах)

гранул данных.

Алгебраические основы грануляции данных.

Определение 2. Матроидом $M = \langle E, \varepsilon \rangle$ называется конечное множество E с нормой $|E| = n$ и семейство его подмножеств $\varepsilon \subset 2^E$, для которых выполняются следующие аксиомы:

$$\begin{cases} M_1: \ \varnothing \in \varepsilon; \\ M_2: \ A \in \varepsilon \ \& \ B \subset A \Rightarrow B \in \varepsilon; \\ M_3: \ A, B \in \varepsilon \ \& \ |B| = |A| + 1 \Rightarrow \exists e \in B \setminus A A \cup \{e\} \in \varepsilon. \end{cases}$$

Так, семейство линейно независимых множеств элементов любого векторного пространства является матроидом [10]. Рассмотрим применение теории матроидов в задачах укрупнения (грануляции, кластеризации) многомерных данных. Покажем, что множество гранул $\mathbf{G}$, покрывающих заданные наборы данных, является матроидом.

Поскольку все элементы грассмановского покрытия ${}^{i}G, {}^{j}G \in \mathbf{G}$ линейно независимы в силу определения 1, то среди множества точек, принадлежащих элементу ${}^{i}G$ есть по крайней мере одна точка, не входящая в элемент ${}^{j}G$ и не выражающаяся в виде линейно независимой комбинации точек, принадлежащих ${}^{j}G$. Тогда, добавляя такую точку к ${}^{j}G$, мы получим новый элемент покрытия ${}^{k}G$, удовлетворяющий аксиоме M_3 определения 2. Элементы покрытия также удовлетворяют аксиомам M_1 и M_2 определения 2 [7]. Следовательно, покрытие множества $\mathbf{G}$ является матроидом.

Оптимальные алгоритмы на матроидах.

Рассмотрим теперь матроид M и весовую функцию на нем $w: E \to R_+$. Типовая задача оптимизации формулируется в виде: необходимо найти $X \in \varepsilon$, для которого

$$w(X) = \max_{Y \in \varepsilon} w(Y), \text{ где } w(Z) := \sum_{e \in Z \subset E} w(e), \qquad (1)$$

т.е. выбрать в указанном семействе подмножество наибольшего веса [10]. Придавая различный смысл весовой функции, можно привести к формулировке

в виде (1) множество известных задач оптимизации, кластеризации и т.д. [9]. В [11] доказывается следующая теорема о том, что для произвольного матроида и произвольной весовой функции существует жадный алгоритм, который находит независимое множество X с наибольшим весом. Следовательно, покрытие по методу [7] позволяет строить жадные алгоритмы обработки гранулированных данных с линейной сложностью.

Алгоритмы гранулирования исходных данных.

Для перехода от исходного векторного пространства к гранулированному представлению (покрытию или разбиению) нами предложены два жадных алгоритма, основанных на различных структурных подходах к множеству гранул $\mathbf{G}$. Первый алгоритм относится к классу *агломеративных алгоритмов* [11] и позволяет эффективно работать со сравнительно разреженными множествами данных

Алгоритм 1. Гранулирование данных с помощью деления осей декартового произведения на интервалы.

Входные данные: Множество точек данных $DtSet$.

Выходные данные: множество гранулированных данных $GrSet$.

С1. Разбить каждую из осей декартового произведения на интервалы.

С2. Выделить множество $TmpDtSet$ всех подмножеств $PartSet_i$ полученного разбиения.

С3. Инициализировать множество гранулированных данных $GrSet = TmpDtSet$.

С4. Просмотреть все подмножества $PartSet_i$ множества $GrSet$.

С41. Если $PartSet_i \neq \varnothing$, то оптимизировать размеры подмножества путем применения операции сжатия, иначе удалить подмножество из множества $GrSet$.

С5. Вывести множество гранулированных данных $GrSet$.

Данный алгоритм использует разбиение исходного множества на ряд искусственных кластеров, которые затем оптимизируются по одному из предложенных в наших работах критериев [8].

Следующий алгоритм относится к классу *дивизимных алгоритмов* [11] и основан на делении исходной гранулы **G** на подмножества.

Алгоритм 2. Гранулирование данных с помощью задания меры информативности.

Входные данные: Множество точек данных $DtSet$, порог меры информативности β.

Выходные данные: множество гранулированных данных $GrSet$.

С1. Инициализировать множество гранулированных данных $GrSet = DtSet$.

С2. Просмотреть все подмножества $GrSet_i$ множества $GrSet$.

С2.1. Вычислить меру информативности $InfMeg(\mathrm{GrSet}_i)$ для подмножества $GrSet_i$.

С2.2. Если $InfMeg(\mathrm{GrSet}) < \beta$:

С2.2.1. Просмотреть каждую i-ую ось декартового произведения.

С2.2.2. Разбить интервал i-ой оси подмножества $GrSet_i$ на два интервала.

С2.2.3. Выделить для каждой i-ой оси два новых подмножества исходного множества $GrSet_i$.

С2.2.4. Определить среди всех возможных вариантов оптимальное разбиение, используя меру информативности.

С2.2.5. Заменить подмножество $GrSet_i$ исходного множества $GrSet$ на два новых, полученных в результате оптимального разбиения.

С3. Вывести множество гранулированных данных $GrSet$.

Результатом работы алгоритмов 1 и 2 является множество гранул ${}^iG \in \mathbf{G}$, $i = 1,2,...,\mathrm{n}$ на которых можно дополнительно выполнить объединение (кластеризацию) для оптимизации общего критерия качества по (1).

Алгоритм инкапсуляции гранулированных данных.

Предлагаемый алгоритм отличается от алгоритмов класса k-means, ISODATA и других, широко используемых в гранулированных вычислениях [9], тем, что он автоматически определяет число кластеров в гранулированном

представлении. Это обеспечивает его универсальность [8].

Алгоритм 3. Инкапсуляция в пространстве гранул.

Входные данные: Множество гранул данных $GrSet$, порог инкапсуляции гранул α.

Выходные данные: Множество кластеров $ClSet$.

С1. Инициализировать множество $ClSet = GrSet$.

С2. Определить мощность множества гранул $CardSet = \left|ClSet\right|$.

С3. Просмотреть все гранулы из множества $ClSet$

С3.1. Выделить текущую гранулу ${}^{i}Gr$ из множества гранул $ClSet$.

С3.2. Просмотреть оставшиеся гранулы из множества $ClSet$.

С3.2.1. Выделить текущую гранулу ${}^{j}Gr$ из оставшихся гранул множества $ClSet$.

С3.2.2. Вычислить меру близости $Meg({}^{i}\mathrm{Gr},{}^{j}\mathrm{Gr})$ гранул ${}^{i}Gr$ и ${}^{j}Gr$.

С3.2.3. Если $Meg({}^{i}\mathrm{Gr},{}^{j}\mathrm{Gr}) > \alpha$:

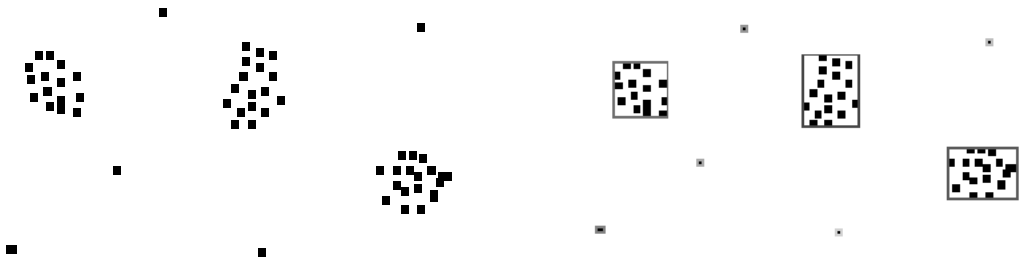
С3.2.3.1. Произвести инкапсуляцию гранул ${}^{i}Gr$ и ${}^{j}Gr$.

С3.2.3.2.Удалить из множества $ClSet$ гранулу ${}^{i}Gr$.

С3.2.3.3. Заменить в множестве $ClSet$ гранулу ${}^{j}Gr$ на инкапсулирующую гранулу Gr^{+}.

С4. Вывести множество кластеров инкапсулированных данных $ClSet$.

Модельный пример применения алгоритмов 1–3 в задаче кластеризации точек на плоскости приведен на следующем рисунке.



a) b)

Рис. 1. Пример исходных двумерных данных a) и их инкапсуляции гранулами b).

В результате выполнения алгоритмов выделения и инкапсуляции гранул мы автоматически получили структуру кластеров в исходных данных для решения задачи Data Mining [1].

Результаты. Обсуждение и анализ.

Разработанный комплекс алгоритмов тестировался на различных классах черно-белых изображений (для наглядности интерпретации результатов). Использовались различные критерии оптимизации гранулированного представления, предложенные в наших работах [6] – [8]. Целью изучения являлась сравнительная оценка параметров алгоритмов с разными целевыми функциями. Применялись функции суммарной плотности и суммарной энтропии гранулированного представления [6].

Ниже приведены две таблицы с различными характеристиками инкапсуляции для данных изображенных на рис.1. Для получения результатов табл.1 применялись алгоритмы 1 и 3, а для табл.2 алгоритмы 2 и 3 соответственно.

Таблица 1

Параметры кластеризации при использовании алгоритмов гранулирования данных с помощью деления осей декартового произведения на интервалы

Кол. разбиений n	Кол. гр. до инк-ции	Плотн. гр. изобр.	Энтропия гр. изобр.	Порог инкапс. α	Кол. класт.	Плотн класт. изобр.	Энтропия класт. изобр.
25[1]	9[1]	0,095[1]	0,451[1]	0,42	1	0,681	0,904
				0,87[1]	8[1]	0,107[1]	0,489[1]
				0,92	9	0,103	0,480
400[2]	50[2]	0,053[2]	0,299[2]	0,42	8	0,089	0,434
				0,79[2]	8[2]	0,089[2]	0,434[2]
				0,85	18	0,074	0,381
1600	104	0,03	0,195	0,42	1	0,681	0,904
				0,62	8	0,089	0,434
				0,98	54	0,03	0,195
10000	330	0,025	0,168	0,42	1	0,681	0,904
				0,64	8	0,089	0,434
				0,83	48	0,026	0,174

Таблица 2

Параметры кластеризации при использовании алгоритмов гранулирования данных с помощью задания меры информативности

Пар. плотн. β	Кол. гранул до инкапс.	Плотн. гран. изобр.	Энтропия гран. изобр.	Порог инкапс. α	Кол. класт.	Плотн. класт. изобр.	Энтропия класт. изобр.
0,11	11	0,08	0,403	0,42	8	0,089	0,434
				0,87	9	0,085	0,419
				0,92	11	0,08	0,403
0,35	29	0,053	0,298	0,42	8	0,089	0,434
				0,79	12	0,076	0,389
				0,85	16	0,065	0,347
$0,61^3$	65^3	$0,03^3$	$0,192^3$	$0,42^3$	8^3	$0,089^3$	$0,434^3$
				0,62	18	0,074	0,380
				0,98	48	0,03	0,192
0,91	96	0,025	0,168	0,42	8	0,089	0,434
				0,64	20	0,065	0,346
				0,83	50	0,026	0,171

Анализ данных таблиц результатов эксперимента показывает, что при применении дивизимного алгоритма гранулирования данных мы получаем более компактное гранулированное представление данных по сравнению с агломеративным алгоритмом. Наиболее это ощутимо при увеличении объема данных представляющих интерес, когда необходимо увеличивать число разбиений или повышать значимость меры информативности [8].

Другим преимуществом использования дивизимного алгоритма является более удобное использование нормированного параметра отвечающего за робастность представление данных, по сравнению с количеством разбиений, которое в общем случае может стремиться к бесконечности.

В качестве меры информативности в данной работе использовалось понятие плотности изображения, подробно описанное в [8]. В качестве альтернативы может использоваться понятие энтропии представления данных, описанное в [7].

К недостаткам дивизимного алгоритма можно отнести увеличение времени обработки данных при увеличении объема исходных данных, т. е. значительно

увеличивается время получения гранулированных данных. Эту проблему можно решить, используя агломеративный алгоритм. Можно существенно снизить вычислительные затраты, но при этом количество гранул будет избыточным. Следующие рисунки показывают результат применения каждого алгоритма для выделенных строк табл.1, 2,3.

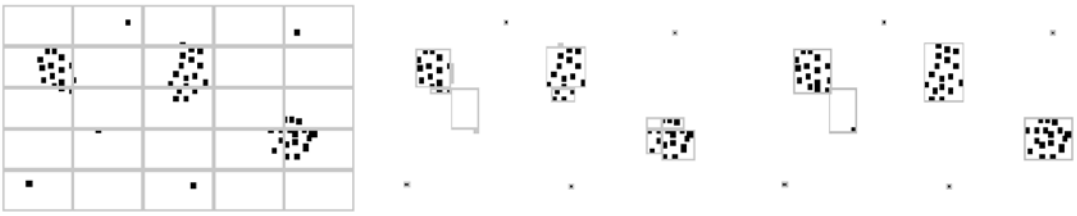


Рис. 2. Пример выделения кластеров с помощью агломеративного алгоритма с малым числом разбиений ([1]).

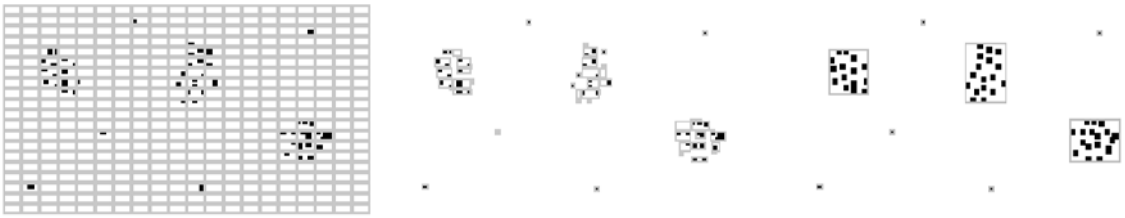


Рис. 3. Пример выделения кластеров с помощью агломеративного алгоритма с большим числом разбиений ([2]).

Рис. 4. Пример гранулирования данных с помощью дивизимного алгоритма ([3]).

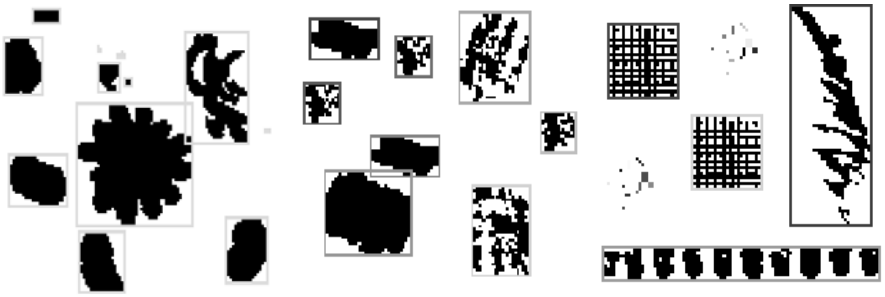


Рис. 5. Примеры грануляции для различных типов кластеров.

Рис.5 демонстрирует сравнительное качество алгоритмов при кластеризации (выделении) объектов сложной формы для последующего распознавания. В наших работах были предложены алгоритмы распознавания изображений в гранулированной форме, использующие нечеткие отношения на гранулах [8].

Заключение и выводы.

Были рассмотрены подходы к кластеризации плохо формализованных данных на основе предложенных в работе дивизимного и агломеративного алгоритмов гранулирования данных В результате определены ключевые параметры обоих алгоритмов и выявлены их основные характеристики по качеству кластеризации и вычислительной эффективности. В результате сравнения были получены следующие результаты:

1. При разреженных, и относительно небольших размеров кластеров достаточно использование агломеративного алгоритма.

2. Для более точного решения и для боле сложных по структуре кластеров более эффективен метод, базирующийся на дивизимном алгоритме гранулирования.

Большую гибкость обоим алгоритмам может придать использование разных мер информативности или их комбинации [8]. Разработанные жадные алгоритмы кластеризации допускают реализацию на параллельных системах.

Литература:

1. Финн В.К. Об интеллектуальном анализе данных // Новости искусственного интеллекта, 2004. -№3.- С. 3-18.

2. Мандель И. Д. Кластерный анализ. – М.: Финансы и статистика, 1988. - 176 с.

3. Батыршин И.З., Недосекин А.О., Стецко А.А., Тарасов В.Б., Язенин А.В., Ярушкина Н.Г. Нечеткие гибридные системы. Теория и практика / Под ред. Н.Г. Ярушкиной. – М. ФИЗМАТЛИТ, 2007. – 208 с.

4. Zadeh L.A. Fuzzy sets and Information Granularity / Advances in Fuzzy Set

Theory and Applications. M. Gupta, R. Ragade, and R. Yager, Eds. Amsterdam, The Netherlands: North-Holland. 1979, pp. 3–18.

4. Zadeh L.A. Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets System. vol. 90. 1997., pp. 111–127.

5. Zadeh L.A. From Computing with Numbers to Computing with Words – From Manipulation of Measurements to Manipulation of Perceptions. IEEE Trans. // Circuits and Systems – Fundamental Theory and Applications. vol. 45. №1. 1999. pp. 105-119.

6. Butenkov S. Granular Computing in Image Processing and Understanding // Proc. of IASTED Conf. In Artificial Intelligence and applications “AIA 2004”. Innsbruk, Austria. February 16-18 2004. Pp. 811-816.

7. Бутенков С.А. Развитие парадигмы интеллектуального анализа многомерной информации применительно к теории информационной грануляции // Сборник трудов IV Международного научно-практического семинара “Интегрированные модели и мягкие вычисления в искусственном интеллекте”. Коломна. 28-30 мая 2007 г. т.1. 188-194 с.

8. Бутенков С.А., Кривша В.В., Аль-Доуяни С.Х.С. Построение системы нечетких отношений взаимного положения на декартовых гранулах // Сборник трудов международной научно-технической конференции „Искусственные интеллектуальные системы” (IEEE AIS’06).- М:. ФИЗМАТЛИТ, 2006. т.2. с. 99-105.

9. Pedrysz W. Granular Computing – the emerging paradigm // Journal of Uncertain Systems, Vol.1, No.1, pp.38-61, 2007

10. Общая алгебра. Т. 1, 2. / В.А. Артамонов, В.Н. Салий, Л.А. Скорняков и др. Под общ. ред. Л.А. Скорнякова. - М.: Наука, 1991.

11. Хаггарти Р. Дискретная математика для программистов. - М.: Техносфера, 2003. - 320 с.

Статья отправлена: 12.12.2015 г.