РОЗДІЛ V. ІНФОРМАЦІЙНО-КОМП'ЮТЕРНІ ТЕХНОЛОГІЇ

УДК 519.85

А.И. Косолап, д-р физ.-мат. наук
А.А. Довгополая, аспирант
Украинский государственный химико-технологический университет, г. Днепропетровск, Украина

СФЕРИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ ДАННЫХ

А.І. Косолап, д-р фіз.-мат. наук
А.О. Довгопола, аспірант
Український державний хіміко-технологічний університет, м. Дніпропетровськ, Україна

СФЕРИЧНА КЛАСТЕРИЗАЦІЯ ДАНИХ

Anatoliy Kosolap, Doctor of Phisical and Mathematical Sciences
Alena Dovgopolaya, PhD student
Ukrainian State University of Chemical Technology, Dnepropetrovsk, Ukraine

SPHERICAL CLUSTERING OF DATA

Рассмотрена задача разбиения данных на сферические кластеры. Приведена новая оптимизационная постановка задачи, которая является многоэкстремальной. Предложен метод точной квадратичной регуляризации для решения оптимизационной задачи. Этот метод сравнивается с методом ближайшего соседа для решения задач кластеризации данных. Многочисленные примеры показывают более высокую эффективность метода точной квадратичной регуляризации.

Ключевые слова: *метод точной квадратичной регуляризации, метод ближайшего соседа, сферическая кластеризация данных, кластер, многоэкстремальные задачи.*

Розглянуто задачу розбиття даних на сферичні кластери. Наведено нову оптимізаційну постановку задачі, яка є багатоекстремальною. Запропоновано метод точної квадратичної регуляризації для розв'язання оптимізаційної задачі. Цей метод порівнюється з методом найближчого сусіда для вирішення задач кластеризації даних. Численні приклади показують більш високу ефективність методу точної квадратичної регуляризації.

Ключові слова*: метод точної квадратичної регуляризації, метод найближчого сусіда, сферична кластерізація даних, кластер, багатоекстремальні задачі.*

In paper we consider a problem of dividing of data on spherical clusters. New optimising statement of a problem which is multiextreme is resulted. We offer a method exact quadratic regularization for the solution of this optimising problem. This method is compared to a method of the nearest neighbour for the solution of problems clustering of data. Many computational examples are provided to show the effectiveness of the proposed method.

Key words: *exact quadratic regularization methods, nearest neighbour a method, spherical clustering of data, cluster, multiextremal problems.*

Введение. Одной из основных задач в искусственном интеллекте и теории распознавания образов является разбиение данных на кластеры [1–2]. Как правило, данные (объекты) представляются точкой в n-мерном пространстве. Характеристикам объектов, которые необходимо разбить на кластеры, соответствуют компоненты n-мерной точки. Если расстояние между двумя точками мало, то эти точки входят в один кластер. В качестве расстояния может выбираться различная метрика пространства. Существуют эффективные методы разбиения данных на два кластера. Однако эффективное разбиение множества точек на более, чем два кластера, представляет сложную задачу [1–2]. Это связано с тем, что оптимальное разбиение точек на кластеры является многоэкстремальной задачей. В настоящее время для решения таких задач чаще используют генетические или эволюционные методы, которые основаны на случайном поиске и позволяют находить оптимальное решение задач кластеризации только с некоторой вероятностью [3]. В работе использован новый метод точной квадратичной регуляризация для решения многоэкстремальных задач, который показал лучшие результаты, в сравнении с генетическими и эволюционными методами, при решении многих тестовых задач [4].

Постановка задачи и метод ее решения. Будем рассматривать множество m точек $\{x^1,...,x^m\}$ в n-мерном евклидовом пространстве. Необходимо разбить это множество на

k кластеров таким образом, чтобы каждая точка попала только в один кластер и точки с близкими расстояниями образовывали кластер. Будем покрывать множество заданных точек *n*-мерными шарами с разными радиусами r_i. Необходимо определить центры $\{z^1,...,z^k\}$ шаров $B_i = \{x \mid\mid\mid x - z^i \mid\mid^2 \le r^2\}$, $i = 1,...,k$ и их радиусы так, чтобы $B_i \bigcap B_j = \Theta, \forall i \ne j$. Это условие непересечения шаров равносильно системе неравенств

$$\| z^j - z^i \|^2 \ge (r_i + r_j)^2, i, j = 1,...k, i \ne j. \quad (1)$$

Теперь достаточно потребовать, чтобы каждая точка $x^i \in B_j$, что равносильно следующим ограничениям, при выполнении условий (1)

$$\prod_{i=1}^{k} (\| x^j - z^i \|^2 - r_i^2) \le 0, j = 1,...,m. \quad (2)$$

Если точка x^i попадает в *i*-й кластер, то соответствующее выражение в круглых скобках (2) будет отрицательным, а все другие – положительными.

В качестве критерия оптимальности покрытия множества точек $\{x^1,...,x^m\}$ непересекающимися шарами выберем минимизацию суммы квадратов радиусов шаров

$$\sum_{i=1}^{k} r_i^2. \quad (3)$$

Если данную последовательность точек эффективнее покрыть меньшим числом кластеров, то некоторые радиусы шаров будут равны нулю. Эффективность разбиения данных на кластеры будем определять суммарным среднеквадратичным отклонением точек кластера от их средней точки.

Решение задачи (1–3) определит центры шаров, которые разобьют множество точек $\{x^1,...,x^m\}$ на кластеры. Задача (1–3) имеет $(n + 1)k$ искомых переменных $\{z^1,...,z^k\}$, $\{r_1,...,r_k\}$ и $m+k(k+1)/2$ ограничений (1–2). Целевая функция (3) является выпуклой, а допустимое множество будет невыпуклым, поэтому задача (1–3) будет многоэкстремальной. Классические методы ее решения, такие, например, как методы внутренней точки, позволяют найти только локальное решение, при этом могут возникнуть проблемы поиска допустимого решения. Поэтому используем метод точной квадратичной регуляризации для решения (1–3), который показал значительное преимущество в нахождении точек глобального минимума при решении многих тестовых задач [4].

Преобразуем задачу (1–3) к виду

$$\min\{z \mid \sum_{i=1}^{k} r_i^2 + s \le z, \prod_{i=1}^{k} (\| x^j - z^i \|^2 - r_i^2) \le 0, j = 1,...,m, \| z^j - z^i \|^2 \ge (r_i + r_j)^2, \forall i \ne j\}, \quad (4)$$

где параметр *s* удовлетворяет условию

$$s \ge \sum_{i=1}^{k} \| z^i \|^2.$$

Далее, задачу (4) преобразуем к следующей

$$\min\{\|y\|^2 \mid \sum_{i=1}^{k} r_i^2 + s \le \|y\|^2, \prod_{i=1}^{k} (\| x^j - z^i \|^2 - r_i^2) \le 0, j = 1,...,m, \| y^j - y^i \|^2 \ge (r_i + r_j)^2, \forall i \ne j\}, \quad (5)$$

где вектор *y* равен

$$y = (z^1,..., z^k, z, r_1,..., r_k).$$

Добавим к ограничениям задачи (5) квадратичное слагаемое так, чтобы функции, определяющие допустимую область задачи, стали выпуклыми. Существует такое значения параметра $q > 0$, при котором допустимая область задачи

$$\min\{\| y \|^2 | \sum_{i=1}^{k} r_i^2 + s + (q-1)\| y \|^2 \le d, \prod_{i=1}^{k} (\| x^j - y^i \|^2 - r_i^2) + q\| y \|^2 \le d,$$
$$j = 1,...,m, q\| y \|^2 - \| y^j - y^i \|^2 + (r_i + r_j)^2 \le d, \forall i \ne j, q\| y \|^2 = d\} \quad (6)$$

будет выпуклой, за исключением условия $q \| y \|^2 = d$. В задаче (6) значение новой переменной *d* необходимо определить. Если решение (y^0, d_0) выпуклой задачи

$$\min\{d | \sum_{i=1}^{k} r_i^2 + s + (q-1)\| y \|^2 \le d, \prod_{i=1}^{k} (\| x^j - y^i \|^2 - r_i^2) + q\| y \|^2 \le d,$$
$$j = 1,...,m, q\| y \|^2 - \| y^j - y^i \|^2 + (r_i + r_j)^2 \le d, \forall i \ne j, q\| y \|^2 \le d\} \quad (7)$$

удовлетворяет условию $q\| y^0 \|^2 = d_0$, то y^0 определяет решение задачи (1–3). В противном случае, необходимо решать задачу

$$\max\{\| y \|^2 | \sum_{i=1}^{k} r_i^2 + s + (q-1)\| y \|^2 \le d, \prod_{i=1}^{k} (\| x^j - y^i \|^2 - r_i^2) + q\| y \|^2 \le d,$$
$$j = 1,...,m, q\| y \|^2 - \| y^j - y^i \|^2 + (r_i + r_j)^2 \le d, \forall i \ne j\} \quad (8)$$

и искать минимальное значение d^*, при котором будет выполняться условие $q\| y^* \|^2 = d^*$, где y^* – решение задачи (8) при фиксированном значении d^*. При увеличении переменной *d* значение целевой функции задачи (8) монотонно возрастает, пока не выполнится условие $q\| y^* \|^2 = d^*$. Рассмотренная последовательность преобразований задачи (1–3) к эквивалентной задаче (8), при условии $q\| y \|^2 = d$, есть метод точной квадратичной регуляризации [4].

Вычисления. Рассмотренный метод точной квадратичной регуляризации реализован в виде компьютерной программы. Были проведены многочисленные эксперименты по разбиению данных на кластеры. Результаты расчетов сравнивались с решениями, полученными методом ближайшего соседа. Этот метод разбивает данные на произвольные кластеры. Суть его заключается в том, что ближайшие точки объединяются в один кластер. Эта пара точек заменяется одной средней точкой. Следующая пара ближайших точек образовывает новый кластер, если ни одна из этих точек не содержит точку существующего кластера. Иначе новая точка попадает в существующий кластер. Таким образом, на каждой итерации число точек, подлежащих кластеризации, сокращается на единицу (каждый кластер представляется средней точкой). Процесс разбиения точек на кластеры продолжается до тех пор, пока минимальное расстояние между двумя точками не превысит заданный порог.

Рассмотрим разбиение точек трехмерного пространства (2, 1, 3; 3, 2, 2; 1, 3, 3; 4, 1, 2; 5, 1, 3; 3, 1, 2; 1, 1, 3; 2, 2, 2; 1, 4, 3; 2, 3, 3; 3, 5, 4; 4, 4, 4; 5, 3, 5; 5, 4, 5; 3, 3, 5; 4, 3, 3; 1, 2, 1; 1, 1, 2)

на три кластера. Для этих данных задача (8) решалась при значениях параметров *s* = 60, *q* = 90. Найдены центры шаров в точках (1.75, 1.5, 2; 1.19, 3.35, 2.74; 3.85, 2.96, 3.26) с радиусами соответственно (1.35; 0.72; 2.33).

Методом ближайшего соседа данные точки были разбиты на три кластера с центрами кластеров в точках (2.93, 2.36, 3.07; 1.33, 3.33, 3; 5, 1, 3).

Полученные результаты сравнивались по суммарному среднеквадратичному отклонению точек кластера от их среднего. Для рассмотренного примера метод точной квадратичной регуляризации дал в 1,6 раз меньшее значение среднеквадратичного отклонения, чем метод ближайшего соседа.

Выводы. В работе приведена новая постановка задачи разбиения многомерных данных на сферические кластеры. Для решения полученной многоэкстремальной зада-

чи используется метод точной квадратичной регуляризации, эффективность которого подтверждена многочисленными экспериментами.

Список использованных источников

1. *Хант Э.* Искусственный интеллект / Э. Хант ; пер. с англ. Д. А. Белова и Ю. И. Крюкова ; под ред. В. Л. Стефанюка. – М. : Мир, 1978. – 560 с.

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

3. *Kenneth V. P.* Differential Evolution. A Practical Approach to Global Optimization / V. P. Kenneth, R. M. Storn, J. A. Lampinen. – Berlin Heidelberg : Springer-Verlag, 2005. – 542 p.

4. *Косолап А. И.* Методы глобальной оптимизации / А. И. Косолап. – Днепропетровск : Наука и образование, 2013. – 316 с.

УДК 004.912:004.632

В.В. Литвинов, д-р техн. наук
О.П. Мойсеенко, ассистент
Черниговский национальный технологический университет, г. Чернигов, Украина

ФОРМИРОВАНИЕ КЛАСТЕРОВ ПРИ РАБОТЕ С НЕИЕРАРХИЧЕСКИМИ МЕТОДАМИ КЛАСТЕРНОГО АНАЛИЗА

В.В. Литвинов, д-р техн. наук
О.П. Мойсеєнко, асистент
Чернігівський національний технологічний університет, м. Чернігів, Україна

ФОРМУВАННЯ КЛАСТЕРІВ ПРИ РОБОТІ З НЕІЄРАРХІЧНИМИ МЕТОДАМИ КЛАСТЕРНОГО АНАЛІЗУ

Vitaliy Litvinov, Doctor of Technical Sciences
Oleg Moyseenko, assistant
Chernigov National Technological University, Chernigov, Ukraine

THE FORMATION OF CLUSTERS AT WORK WITH NON-HIERARCHICAL METHODS OF CLUSTER ANALYSIS

Рассматриваются принципы объединения похожих текстовых документов в тематические кластеры, механизмы формирования центров кластеров и правила остановки процесса автоматической кластеризации. Подробно описаны основные характеристики кластера и виды кластеризации. Сделан отступ в сторону итеративных методов как таких, что пригодны для работы с большими коллекциями документов и потому, что именно на них основана разрабатываемая система автоматизированной обработки больших объемов динамической текстовой информации. Разрабатываемая система нацелена на выполнение функций поиска, классификации и кластеризации текстовых документов согласно пользовательским запросам.

***Ключевые слова**: текстовая коллекция, центроид, неиерархическая кластеризация, обработка текстовых документов.*

Розглянуто принципи об'єднання схожих текстових документів у тематичні кластери, механізми формування центрів кластерів та правила зупинки процесу автоматичної кластеризації. Детально описані основні характеристики кластера та види кластеризації. Зроблений відступ у бік ітеративних методів як таких, що придатні для роботи з великими колекціями документів і тому, що саме на них ґрунтується розроблювана система автоматизованої обробки великих об'ємів динамічної текстової інформації. Розроблювана система націлена на виконання функцій пошуку, класифікації та кластеризації текстових документів за запитами користувача.

***Ключові слова:** колекція документів, центр кластера, неієрархічна кластеризація, обробка текстових документів.*

Discusses the principles of association or similar text documents into thematic clusters, mechanisms of formation of the cluster centers and the stopping rule of the automatic clustering. Described in detail the main characteristics of the cluster and the kinds of clustering. Indented toward iterative methods such as that are suitable to work with large collections of documents s because it is based on their developed system of automated processing of large volumes of dynamic textual information. The developed system aims to perform search functions, classification and clustering text documents according to user requests. The system itself is described in more detail in other works of the author.

***Key words**: text collection, centroid, non-hierarchical clustering, the processing of text documents.*

Введение. Кластеризация текстовых данных является многоэтапной процедурой, на каждом шаге которой должна решаться отдельная задача выбора наиболее адекватного способа реализации, влияющего на последующие этапы.