УДК 004.8

МЕТОДЫ ГИБРИДНОЙ РЕЛЯЦИОННОЙ КЛАСТЕРИЗАЦИИ В АНАЛИЗЕ СРЕДНЕГО ПОТРЕБЛЕНИЯ ЭЛЕКТРОЭНЕРГИИ СТРАНАМИ БЫВШЕГО СОВЕТСКОГО СОЮЗА

А.С. КЛИМОВА

Казанский государственный технологический университет

В статье дается описание комплекса методов гибридной кластеризации на основе реляционных инвариантных алгоритмов. Рассматривается применение гибридной кластеризации к анализу временных рядов из базы данных энергопотребления.

Введение

Кластерные процедуры находят широкие приложения в анализе данных в различных областях человеческой деятельности [1, 2]. Кластеризация объектов позволяет разбить множество исследуемых объектов на классы сходных объектов и перейти от описания объектов к описанию классов объектов и отношений между ними, что является одним из важных этапов формирования знания об исследуемой области, ее моделирования и анализа. Кластерный анализ активно развивается последние десятилетия, что обусловлено, с одной стороны, появлением новых областей его приложения, таких как Интернет и генная инженерия, со специфическими объектами исследования и, с другой стороны, разрабатываются методы гибридной кластеризации, сочетающие результаты разных кластерных процедур или сочетающие кластерный анализ с другими методами анализа данных [3].

В работе [4] рассматривалась общая структура методов гибридной кластеризации данных на основе методов реляционной кластеризации и визуализации данных. В данной работе дается описание комплекса методов гибридной кластеризации на основе реляционных инвариантных алгоритмов. Методы иллюстрируются на примере анализа базы данных временных рядов энергопотребления.

Статья организована следующим образом. В разделе 1 дается общая схема гибридной реляционной кластеризации и визуализации данных. В разделе 2 описанный подход иллюстрируется на примере кластеризации временных рядов из области энергетики. В заключении отмечаются основные результаты статьи.

1. Общая схема гибридной реляционной кластеризации и визуализации

Пусть $M = \{x,y,\ldots\}$ – множество объектов и I – положительное число. Функции S,D:МЧМ$\rightarrow [0, I]$, удовлетворяющие на M условиям:

$$S(y, x) = S(x, y), \quad S(y, y) = I,$$

$$D(y, x) = D(x, y), \quad D(y, y) = 0,$$

будут называться, соответственно, взвешенным (нечетким) отношением сходства и отношением различия. Без ограничения общности будем полагать $I = 1$. Взвешенным (нечетким) отношением эквивалентности называется отношение сходства, удовлетворяющее условию $(\vee,\vee)$-транзитивности:

$$S(y, x) \geq \min\{(S(x, z), S(z, y)\}.$$

$(\wedge,\vee)$-транзитивность двойственна ультраметрическому неравенству:

© *А.С. Климова*

$D(x,y) \le \max\{(D(x, z), D(z, y)\}$.

Нечеткие отношения эквивалентности и ультраметрики определяют иерархию разбиений множества *M*. Они изучались в работах [5, 6]. Далее полагаем, что на *M* задано взвешенное отношение сходства *S*.

Пусть множество *RC* означает множество четких или нечетких отношений на *M*, а *PC* – множество описаний элементов из *M* в пространствах различной размерности. Гибридной реляционной кластеризацией и визуализацией *M* будет называться преобразование

$$HC(S) = <RC; PC>. \quad (1)$$

Рассмотрим частные случаи схемы (1).

Если $RC = \varnothing$, а *PC* задает описание элементов из *M* в пространстве размерности 2 или 3, то *HC* задает визуализацию данных в пространстве малой размерности. Схема (5) имеет вид

$$HC(S) = <\varnothing; P>,$$

где *P* задает координаты объектов в пространстве размерности два или три.

В работах [4, 7, 8] задача визуализации рассматривается как задача минимизации искажений исходных расстояний между объектами при их представлении в двух- или трехмерном пространстве, а сами процедуры визуализации основаны на использовании генетических алгоритмов.

Если $PC = \varnothing$ и $RC = \{E\}$, где *E* – четкое или взвешенное отношение эквивалентности на *M*, то (1) определяет, соответственно, обычную или иерархическую кластеризацию множества *M*.

Общая схема иерархических кластерных процедур, удовлетворяющих условиям инвариантности относительно исходной нумерации объектов и инвариантности относительно монотонных преобразований значений сходства, исследовалась в работах [7, 9]. Схема (1) имеет вид

$$HC(S) = <E; \varnothing>,$$

где *E* – взвешенное отношение эквивалентности. Преобразование *HC*(*S*) задается процедурой кластеризации $Q(S)=E$, которая определяется как:

$$Q(S) = TC(F(S)) = E,$$

где *TC* - процедура транзитивного замыкания взвешенных отношений [5, 6], а *F* - процедура "коррекции" заданного отношения сходства *S*, такая что $F(S) \subseteq S$.

В работе [4] рассматривается совместное использование методов иерархической кластеризации с методами двух- и трехмерной визуализации данных. В результате иерархической агломеративной кластеризации все объекты и кластеры поэтапно будут объединяться в один кластер. Рассматриваемый подход к гибридной визуализации данных пытается уменьшить изменения расстояний между объектами внутри маленьких классов за счет увеличения изменений расстояний между объектами из разных классов на высших уровнях иерархии.

Схема гибридной кластеризации (1) имеет вид

$$HC(S) = <E;P>,$$

где *E* – взвешенное отношение эквивалентности, получаемое реляционной процедурой иерархической кластеризации, а *P* – описание объектов множества *M* в двух- или трехмерном пространстве.

Метод гибридной кластеризации с визуализацией сильных связей [4] исходит из того, что структура данных может не содержать «естественной» кластеризации, и визуализация сильных связей может быть полезна для анализа отклонения полученной кластеризации от структуры данных. Схема (1) гибридной кластеризации имеет вид

$$HC(S) = <\{E, S_\alpha\};\varnothing>,$$

где E - четкое отношение эквивалентности на M, задаваемое реляционным кластерным алгоритмом, а S_α - оптимальный α-срез исходного взвешенного отношения сходства S на уровне сходства α.

2. Гибридная кластеризация временных рядов в анализе временных рядов из базы данных энергопотребления

Описанные методы были использованы в задаче анализа уровня потребления электроэнергии странами бывшего Советского Союза за 1992-2004 годы (http://www.eia.doe.gov/iea/wec.html). В качестве примера рассмотрим применение метода гибридной кластеризации для анализа данных по потреблению электроэнергии.

На рис. 1 показаны результаты иерархической кластеризации этих временных рядов методом, описанным в работах [5, 6]. В качестве расстояния между временными рядами использовалось Евклидово расстояние между ними. Для каждого кластера приведено значение среднего потребления электроэнергии странами из этого кластера за рассматриваемый период. Как видно из построенной дендрограммы, страны сгруппированы в кластеры не только по сходству расстояний между соответствующими временными рядами, но также и по величине среднего потребления электроэнергии.



Выводы

В работе рассмотрены общая постановка гибридной реляционной кластеризации и визуализации данных и различные частные случаи общей схемы. Гибридная реляционная кластеризация позволяет сочетать кластеризацию

данных с их пространственной визуализацией, что дает более полную информацию о структуре исследуемых данных. Разработанные методы гибридной кластеризации реализованы в виде комплекса программ на языке МАТЛАБ.

Summary

There are description of the battery of methods of hybrid clustering based on relational invariant algorithms consider in this paper. Propose application of methods of hybrid clustering for the analysis of time series from data base of power consumption.

Литература

1. Дюран Б., Оделл П. Кластерный анализ. – М.: Статистика, 1977. – 128с.

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

3. Strehl A., Ghosh J. Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions. - Journal of Machine Learning Research 3, 2002. – pp. 583-617.

4. Батыршин И.З., Климова А.С. Гибридная реляционная кластеризация и визуализация данных: Труды Всероссийской научной конференции по нечетким системам и мягким вычислениям НСМВ-2006 (20 – 22 сентября 2006 г., Тверь). – М.: Физматлит, 2006. – 193-209.

5. Tamura S., Higuchi S., Tanaka K. Pattern classification based on fuzzy relations // IEEE Trans. Systems, Man, Cybernetics, SMC-1. 1971. pp. 61- 66.

6. Zadeh L.A. Similarity relations and fuzzy orderings // Information Sciences. 1973. – № 3. – pp. 177-200.

7. Батыршин И. З, Климова А. С. Инвариантные кластерные процедуры, основанные на нечетком отношении сходства. Интегрированные модели и мягкие вычисления в искусственном интеллекте: Сборник научных трудов III-го Международного научно-практического семинара, 2005. – Москва: Физматлит, 2005. – С. 119 - 125.

8. Klimova A. Evolutionary procedures of visualization of multidimensional data // Proc. Int. Conf. Fuzzy Sets and Soft Computing in Economics and Finance, St. Petersburg, Russia. 2004. vol. I. – pp. 130–139.

9. Batyrshin I.Z., Rudas T., Klimova A. On general scheme of invariant clustering procedures based on fuzzy similarity relation // Proc. Int. Conf. Fuzzy Sets and Soft Computing in Economics and Finance, FSSCEF 2004, St. Petersburg, Russia. 2004. – vol. I. – pp. 122-129.

Поступила 04.07.2007