АДАПТИВНАЯ КЛАСТЕРИЗАЦИЯ НА ОСНОВЕ ДИВИЗИМНЫХ И ИТЕРАЦИОННЫХ МЕТОДОВ

с.н.с., Нейский И.М.

МГТУ им. Н.Э. Баумана, г. Москва

Описывается методика адаптивной кластеризации фактографических данных на основе дивизимных и итерационных методов. Предложенная методика состоит из пяти этапов и позволяет выполнять докластеризацию дополнительного набора данных без перезапуска основного алгоритма.

Adaptive clustering based on partible and iterative methods

Neyskiy I.M.

Bauman Moscow State Technical University

The article describes principles of adaptive clustering in factual database based on partible and iterative methods. Offered principles include five points and allow dividing additional data without restarting of main process.

В настоящее время в вузах активно внедряются и разрабатываются информационно-аналитические системы, которые направлены на мониторинг образовательных процессов и анализ ключевых показателей качества. Одной из важнейших задач в этой области – анализ и агрегирование многочисленных фактографических данных, которое часто решаются с использованием методов кластеризации. На данный момент известно более 50 методов кластеризации, которые представлены в математической и алгоритмической форме, но при этом мало из них имеют реализацию и рекомендации по использованию в сфере образования. Знание того, какие методы дают наилучший результат, может подсказать направление движения тем, кто планирует применять кластерный анализ для решения практических задач, создаёт новые алгоритмы или совершенствует существующие.

По существующим методам кластерного анализа построена классификация, которая разделяет методы по способу обработки данных на иерархические и неиерархические [1]. Иерархические методы в соответствии с классификацией делятся на агломеративные и дивизимные методы [1]. Агломеративная группа методов характеризуется последовательным объединением исходных элементов и соответствующим уменьшением количества кластеров [2]. Дивизимная группа методов характеризуется последовательным разделением исходных элементов и соответствующим увеличением количества кластеров [2]. Самую значимую часть неиерархических методов представляют собой итеративные методы [1]. Данная группа методов основана на разделении набора данных на некоторое количество отдельных кластеров [2]. Существуют два подхода для разделения данных. Первый подход заключается в определении границ кластеров как наиболее плотных участков в многомерном пространстве характеристик объектов, т. е. определение кластера там, где имеется большое «сгущение» объектов [2]. Второй подход заключается в минимизации меры различия объектов [2].

Основной задачей кластеризации является получение множества кластеров на основе множества исходных объектов. Для существующих методов в литературных

источниках, как правило, приводятся практические рекомендации по использованию метода и описательные характеристики возможностей методов. Адаптивность кластеризации означает возможность применения метода к выбранной предметной области после применения соответствующих настроек метода, выполнения обучения метода. Следует отметить, что методы кластерного анализа являются контекстно-зависимыми. В данном направлении интеллектуального анализа данных выявлено две проблемы: 1) потеря значимых закономерностей при использовании одного инструмента анализа; 2) вычислительная сложность и большие временные затраты при применении инструментов на исходных данных.

Предлагаемая интеграция методов из двух разных классов: дивизимного и итеративного, направлена на устранение выявленных проблем. Практическая задача, для решения которой используется описанная методика, имеет следующие характеристики:

- Количество исходных данных – более 10 000 объектов;
- Количество значимых характеристик объектов – более 70 штук;
- Типы характеристик – числовые, текстовые;
- Форма получаемых кластеров – сложная, с пересечениями;
- Количество кластеров – результат анализа, а не входной параметр;
- Качество анализа – высокое.

В рамках данной интеграции предлагается использование базовых принципов двух методов кластеризации: MST [3] и Fuzzy C-Means [4]. В результате интеграции получается методика с двухэтапной кластеризацией (см. рис. 1). На первом этапе данной методики строится минимальное остовное дерево, образуя оптимизированную древовидную структуру из исходных элементов на основе характеристик кластеризуемых объектов. На втором этапе данной методики используется итеративный подход, с помощью которого сначала выделяются первичные кластерные центры на основе оптимизированной древовидной структуры, а потом центры кластеров и содержимое кластеров уточняются на основе вычисления степени принадлежности объекта кластеру и локального критерия остановки цикла.

Рис. 1. Методика адаптивной кластеризации.

Сравнение достоинств и недостатков разработанной методики (Adakl) представлено в Таблице 1.

Метод	Достоинства	Недостатки
MST	простота использования; высокая скорость; понятность и прозрачность алгоритма.	алгоритм слишком чувствителен к выбросам; требуется задание количества кластеров.
Fuzzy C-Means	возможность частичного отнесения объекта к нескольким кластерам	высокая вычислительная сложность; требуется задание количества кластеров; неопределенность с объектами, которые удалены от центров всех кластеров.
Adakl	понятность и прозрачность алгоритма; двухэтапная кластеризация; нечеткость	нелинейная зависимость времени анализа от количества исходных объектов;

	при определении объекта в кластер; возможность использования объектов с разными типами атрибутов (числовые и текстовые); количество кластеров определяется в результате анализа; приемлемое время работы и конечность результата.	чувствительность к выбросам.

Табл. 1. Сравнение методов кластеризации MST и Fuzzy C-Means.

Количество объектов исследований с целью разбиения на кластеры постоянно растет, в том числе во время проведения анализа сформированного массива фактографических данных, увеличивая время проведения исследований, поэтому возникает задача оптимизации времени анализа исходных данных в виде докластеризации «новых» объектов. В общем случае, при увеличении количества исследуемых объектов, требуется повторный запуск исследования на всем массиве данных, который потребует соответствующих затрат времени, технических и человеческих ресурсов. Задачей докластеризации (см. рис. 2) является определение необходимости повторного запуска исследования полного массива данных и, в случае отсутствия признаков появления новых значимых групп объектов, распределение «новых» объектов по имеющимся кластерам на основе оценки близости распределяемых объектов к распределенным объектам.

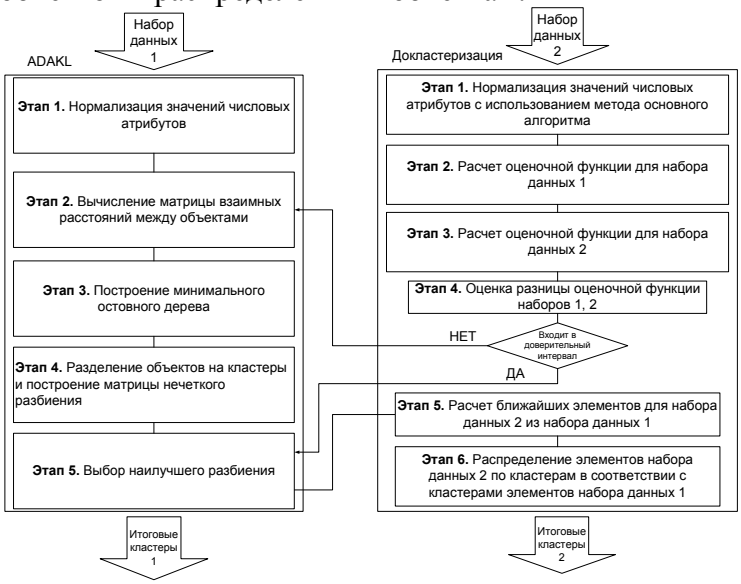

Рис. 2. Докластеризация расширяющего набора фактографических данных.

Для устранения чувствительности к выборам в дополнение к методике Adakl предлагается использование предобработки исходных данных в виде фильтрации незначащих компонентов, нормализация данных и т.п. Более подробная информация о методике представлена по адресу http://philippovich.ru/Persons/Neyskiy/Neyskiy.htm.

Литература.

1. Баргесян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP, 2-е изд., перераб. и доп.. - СПб.: БХВ-Петербург, 2008.

2. Чубукова И.А. Data Mining: Учебное пособие. - М.: Интернет-Университет Информационных Технологий; БИНОМ. Лаборатория знаний, 2006.

3. He H., Singh A. Efficient Algorithms for Mining Significant Substructures in Graphs with Quality Guarantees. - Department of Computer Science University of California, Santa Barbara, 2004.

4. Штовба С. Д. Введение в теорию нечетких множеств и нечеткую логику. Источник: http://matlab.exponenta.ru/fuzzylogic/book1/index.php.