

В дальнейшем на основе созданных прототипов страниц будет осуществлена разработка дизайн-макета сайта.

Список литературы

1. Minimal wireframing tool. [Электронный ресурс]. Режим доступа: <http://wirefrime.cc/> (дата обращения: 04.05.2018).
2. Prototypes that feel real. [Электронный ресурс]. Режим доступа: <http://proto.io/> (дата обращения: 04.05.2018).
3. NinjaMock online wireframe and mockup tool. [Электронный ресурс]. Режим доступа: <http://ninjamock.com/> (дата обращения: 06.05.2018).

КЛАСТЕРНЫЙ АНАЛИЗ: БАЗОВЫЕ КОНЦЕПЦИИ И АЛГОРИТМЫ

Долгодворова Е.В.

Долгодворова Елена Валерьевна – студент,
кафедра информационных технологий и систем,
Дальневосточный государственный университет путей сообщения, г. Хабаровск

Аннотация: в статье рассмотрены базовые концепции и алгоритмы, применяемые при кластерном анализе данных. Выполнен сравнительный анализ алгоритмов кластеризации.

Ключевые слова: кластерный анализ, кластеризация, кластер, алгоритмы.

Кластерный анализ или кластеризация делит данные на группы (кластеры, классы), таким образом, чтобы каждый кластер состоял из похожих объектов, а объекты разных кластеров были отличны друг от друга. В кластеризации, в отличие от классификации, перечень групп не задан четко, а определяется в процессе работы алгоритма [1].

В сравнении с математико-статическими методами кластерный анализ не накладывает условий на тип рассматриваемых объектов, поэтому позволяет исследовать многообразие исходных данных произвольной природы [2].

Общая схема алгоритма кластеризации представлена на рисунке 1.

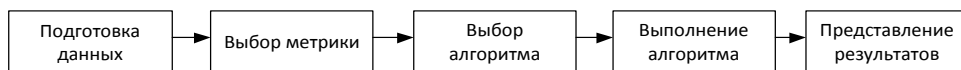


Рис. 1. Общая схема алгоритма кластеризации

Первый этап заключается в подготовке данных для кластерного анализа. В большинстве случаев, данные описывают в виде таблиц, где столбец является одним из атрибутов, а строка объектом данных.

На втором этапе происходит выбор метрики, с помощью которой определяем сходство объектов.

Так классической метрикой в двумерном пространстве является евклидово расстояние ($d(p, q) = \sqrt{\sum_{k=1}^n (p_k - q_k)^2}$). При этом существуют и другие метрики в евклидовых пространствах, например, параметрическое расстояние Минковского ($p(x, y) = (\sum_{i=1}^n |x_i - p_i|^p)^{1/p}$). При параметре со значением 2 оно обращается в евклидову метрику. Для нечисловых атрибутов выбираются метрики, позволяющие привести их к численному виду и сравнить. Популярным расстоянием для строковых данных является метрика Левенштейна (редакционное расстояние или дистанция

редактирования), определяет минимальное количество операций вставки одного символа, удаление одного символа и замены одного символа на другой, необходимых для превращения одной строки в другую.

Метрику подбирают для каждого конкретного типа данных индивидуально. Также бывают ситуации, когда не получается подобрать метрику для некоторого типа данных и приходится придумывать свою систему оценки расстояния между объектами. Результаты кластеризации могут значительно различаться при использовании разных мер близости.

На следующем этапе выбирается алгоритм, с помощью которого группируются объекты. Выбор алгоритма является сложной задачей, так как полученный результат во многом зависит от алгоритма. Чаще всего приходится использовать несколько алгоритмов, то есть их комбинировать для получения более точного результата.

На четвертом этапе происходит реализация выбранного алгоритма (или нескольких алгоритмов). Результатом данного этапа является группировка объектов по кластерам.

Пятый этап подразумевает представление полученной группировки в удобном для интерпретации виде. Представление результатов кластеризации призвано помочь наиболее точно интерпретировать результаты выполнения алгоритма [3].

В настоящее время существует достаточно много алгоритмов кластерного анализа. Принято все алгоритмы подразделять на две категории: иерархические и неиерархические. Деление алгоритмов обуславливается выходными данными. Классификация алгоритмов представлена на рисунке 2.



Рис. 2. Алгоритмы кластеризации

Результаты иерархических алгоритмов представляются в виде дендрограммы - древовидной диаграммы (схемы), которая показывает в какой последовательности происходит разбиение объектов по кластерам. Иерархические алгоритмы в свою очередь подразделяются на две группы: агломеративные и дивизимные.

Наиболее распространенным представителем агломеративных алгоритмов является алгоритм ближайшего соседа. Рассмотрим данный алгоритм.

На первом этапе составляется матрица расстояний между объектами и объектам назначается кластер. Второй этап заключается в поиске минимального расстояния между объектами (соседями). На третьем этапе происходит объединение кластеров. На заключительном этапе выполняется проверка на количество оставшихся кластеров: в случае если кластер 1, то алгоритм завершается, иначе возврат к первому шагу.

Иерархические дивизимные алгоритмы являются полной противоположностью агломеративным методам. То есть, изначально все объекты – это один кластер. На последующих этапах этот кластер подразделяется на меньшие кластеры [4].

Неиерархических методов гораздо больше, несмотря на это они работают на одних и тех же принципах.

Итеративные методы имеют такое название благодаря тому, что по итерациям распределяют объекты по кластерам. Таким является алгоритм k-means. Общая идея

этого алгоритма заключается в минимизации расстояний в кластерах между объектами. Алгоритм завершается, когда еще большая минимизация расстояний не возможна.

В полостных алгоритмах за кластер принимается некоторая группа объектов, которые расположены очень плотно. Примером таких алгоритмов выступает алгоритм DBSCAN, ключевая идея которого заключается в том, что для всех объектов одного кластера их k -ый сосед находится приблизительно на одном и том же расстоянии.

Идея модельных алгоритмов состоит в том, что существует некоторая модель кластера и необходимо найти максимальное сходство между данной моделью и данными. Для представления таких моделей часто используется аппарат математической статистики. Например, алгоритм EM использует метод максимизации ожиданий Expectation Maximization. Суть алгоритма заключается в том, что помимо известных данных, есть неизвестные, которые относятся к распределению по кластерам. Эти неизвестные стараемся максимально точно оценить.

Концептуальные алгоритмы предполагают создание концепции описания каждого кластера, обычно генерируют иерархию классов, но не относятся к иерархической кластеризации. Один из таких алгоритмов COWEB. Основная идея создание последовательно строит дерево кластеров, рассматривая полезность отнесения объекта к какому-либо кластеру.

Одним из классических примеров сетевых алгоритмов является метод WaveCluster. Он рассматривает все данные как сигнал в N -мерном пространстве атрибутов и пытается выделить в этом сигнале поддиапазоны, в которых связанные компоненты и будут кластерами.

В таблице представлено сравнение алгоритмов кластеризации.

Таблица 1. Сравнение алгоритмов

Алгоритм	Достоинства	Недостатки
Алгоритм ближайшего соседа	• строит оптимальное разбиение • получение всех разбиений в виде дендрограммы	• вычислительная сходимость • кластеры не пересекаются • вычислительная сложность
Дивизмные алгоритмы	• построение всех разбиений • результат в виде дендрограммы	• высокая вычислительная сложность
Алгоритм k -means	• простота и быстрота использования • понятность и прозрачность алгоритма	• необходимо знать количество кластеров • вычислительная сложность
Алгоритм DBSCAN	• хорошо поддается модифицированию • малое время вычисления	• неспособность соединять кластеры через проём • высокая сложность
Алгоритм EM	• линейное увеличение ложности при росте объема данных • возможность построения желаемого числа кластеров • быстрая сходимость при удачной инициализации	• алгоритм может остановиться на локальном минимуме • при неудачной инициализации сходимость алгоритма может быть низкой
Алгоритм COWEB	• масштабируемость • работа при неизвестном количестве объектов	• относительно сложный математический аппарат
Алгоритм WaveCluster	• может обнаруживать кластеры произвольных форм	• сложность реализации • применим только к данным низкой размерности

Кластерный анализ уже давно сыграл важную роль в самых различных областях: психология и другие социальные науки, биология, статистика, распознавание образов, поиск информации, машинное обучение и сбор данных.

Таким образом, кластеризация является неотъемлемой частью современной обработки больших массивов информации, заметно облегчая работу с ней и охватывая множество областей применения. Выбор различных методов кластеризации зависит от конкретного случая и не может быть определен однозначно. Для достижения наилучшего результата необходимо экспериментировать с выбором мер расстояний, а иногда даже менять алгоритм. Никакого единого решения не существует.

Список литературы

1. *Воронцов К.В.* Алгоритмы кластеризации и многомерного шкалирования. Курс лекций. МГУ, 2007.
2. *Заде Л. А.* Кластеризация и кластер / Л.А. Заде, С. Рао и др. Москва: Мир, 1980. 383 с.
3. Bourabai Research: методы и средства анализа данных. [Электронный ресурс]. Режим доступа: <https://bourabai.ru/tpoi/analysis6.htm/> (дата обращения: 14.05.18).
4. *Чубукова И.А.* Курс лекций «Data Mining , Методы кластерного анализа // Интернет-университет информационных технологий. [Электронный ресурс]. Режим доступа: <https://www.intuit.ru/studies/courses/6/6/lecture/182?page=2/> (дата обращения: 14.05.18).

РАЗРАБОТКА ВЕБ-ГИС ПО ПРЕДОСТАВЛЕНИЮ УСЛУГИ ВЫСОКОТОЧНОГО ПОЗИЦИОНИРОВАНИЯ В РЕСПУБЛИКЕ БАШКОРТОСТАН

Антонова А.С.

*Антонова Арина Сергеевна – студент,
кафедра геоинформационных систем,*

Уфимский государственный авиационный технический университет, г. Уфа

Аннотация: в данной статье рассматривается актуальность использования высокоточного позиционирования, а также обосновывается выбор инструмента, с помощью которого заинтересованные лица смогут воспользоваться данной услугой.

Ключевые слова: высокоточное позиционирование, высокоточная навигация, СПТН, разработка веб-ГИС.

Благодаря развитию GPS и глобальных навигационных спутниковых систем, технологии высокоточного позиционирования (высокоточной навигации), которые включают в себя совокупность приемов и методов определения местоположения объектов с точностью до нескольких сантиметров, все чаще используются в геодезии и картографии, промышленном и гражданском строительстве, архитектурном и градостроительном проектировании, при мониторинге смещений грунтов и деформаций крупных инженерных сооружений и конструкций, в коммунальном хозяйстве, навигационном обеспечении городского транспортного комплекса, при создании геоинформационных систем и в других областях.

Услуга высокоточного позиционирования предоставляется с использованием системы повышения точности и надежности навигации Государственной