МЕТОДЫ И СРЕДСТВА СИСТЕМЫ ГЕОБАЗАДАННЫХ ДЛЯ АДАПТАЦИИ КОМПЬЮТЕРНЫХ МОДЕЛЕЙ. ИНСТРУМЕНТЫ КЛАСТЕРИЗАЦИИ

Таранчук В. Б.
Кафедра компьютерных технологий и систем, Факультет прикладной математики и информатики,
Белорусский государственный университет
Минск, Республика Беларусь
E-mail: taranchuk@bsu.by

Обсуждаются методические и технические вопросы развития программной системы ГеоБазаДанных (ГБД). Отмечены новые функциональные возможности, обеспеченные включением в ГБД исполняемых модулей интеллектуальноого анализа данных системы компьютерной алгебры Wolfram Mathematica. В частности, подготовлены и предполагается обсудить на представительных наборах геоданных варианты выбора наилучших алгоритмов кластеризации.

Введение

Решением задачи кластерного анализа (сегментации) являются разбиения, удовлетворяющие принимаемому критерию. Критерий обычно представляет собой формализованный функционально набор правил для определения уровней различий при разбиениях и группировках (целевая функция). Кластерный анализ широко применяется во многих областях, в частности, в компьютерных системах при распознавании образов, анализе изображений, поиске информации, сжатии данных, в компьютерной графике, биоинформатике, машинном обучении. При интеллектуальном анализе данных сегментация может использоваться как самостоятельный инструмент для принятия решения о распределении данных, для контроля характеристик и последующего анализа наборов данных, конфигураций и содержания конкретных кластеров. В качестве альтернативы, кластерный анализ может служить этапом предварительной обработки для других алгоритмов. Также сегментация используется для обнаружения нетипичных объектов – выбросов (значения, которые находятся «далеко» от любого кластера), это – обнаружение новизны, такие объекты могут быть более интересными, чем включенные в кластеры.

Важное достоинство кластерного анализа в том, что при его выполнении можно производить разбиение объектов не только по одному параметру, а по набору признаков. Кроме того, кластерный анализ в отличие от большинства математико-статистических методов не накладывает никаких ограничений на вид рассматриваемых исходных данных.

I. Кратко о ГеоБазаДанных

ГеоБазаДанных – комплекс интеллектуальных компьютерных подсистем, математического, алгоритмического и программного обеспечения наполнения, сопровождения и визуализации баз данных, входных данных для имитационных и математических моделей, средств проведения вычислительных экспериментов, алгоритмических и программных средств создания постоянно действующих компьютерных моделей [1]. Средствами ГБД можно формировать и визуализировать цифровые описания пространственных распределений данных об источниках загрязнения, о геологическом строении изучаемых объектов; графически иллюстрировать решения задач, описывающих динамические процессы многофазной фильтрации, миграции флюидов, переноса тепла, влаги, минеральных водорастворимых соединений в толщах пород; конструировать и реализовывать интерактивные сценарии визуализации и обработки результатов вычислительных экспериментов. Подсистемы ГБД позволяют рассчитывать и выполнять в разных приближениях экспертные оценки локальных и интегральных характеристик экосистем, выполнять расчет распределений концентраций и массовых балансов загрязняющих веществ; создавать постоянно действующие модели объектов нефтедобычи; формировать и выводить на твердые копии тематические карты.

II. Подготовка исходных данных

Примеры, включаемые для обсуждения в докладе, рассчитаны с данными [2]. Напомним, что моделируемая поверхность (рис. 1 [2]) имеет полное математическое описание. Данные для демонстраций методов и алгоритмов интеллектуального анализа получены имитацией замеров, соответствующий набор данных – пункты замеров уровня восстанавливаемой поверхности, представляющие (по факту) рассеянное множество точек, интерпретируются, как данные на профилях наблюдений. Соответствующая схема их размещения показана на рис. 2, 3 [2].

III. Инструментарий, примеры кластерного анализа геоданных

Эффекты числа кластеров. Одной из важнейших проблем сегментации является определение количества кластеров. Определение ко-

личества кластеров – это одна из важнейших проблем сегментации. В более широком смысле – это проблема инициализации алгоритма: выбора оптимальных значений управляющих параметров, используемых оценочных функций, метрики, условий остановки и т.п. По результатам выполненной первой серии расчетов в докладе будут приведены иллюстрации, рассчитанные с установками по умолчанию, используя функцию Wolfram Mathematica FindClusters. Сопоставление вариантов дает основание утверждать, что необходимы дополнительные действия по выбору метода, метрики и других параметров алгоритмов кластеризации.

Эффекты принятого метода кластеризации. Во второй серии будут представлены результаты, которые иллюстрируют особенности наиболее часто используемых алгоритмов кластеризации. Кластеризация в примерах этой серии рассматривалась только для пар координат, т.е. учитывалось относительное положение точек рассеянного множества, причем, в программном модуле использована функция FindClusters с разными критериями CriterionFunction, норма в примерах серии 2 вычислялась по метрике DistanceFunction EuclideanDistance. Вообще говоря, включенное в ГеоБазаДанных из системы Wolfram Mathematica соответствующее программное приложение допускает варианты метода кластеризации (CriterionFunction): Automatic, find clustering hierarchically, find clustering by local optimization, density-based spatial clustering of applications with noise, variational Gaussian mixture algorithm, Patrick clustering algorithm, k-means clustering algorithm, partitioning around medoids, mean-shift clustering algorithm, displace examples toward high-density region, minimum spanning tree-based clustering algorithm, spectral clustering algorithm.

Влияние метрики. В рассмотренных примерах серий 1 и 2, а также в третьей серии результатов сходство или различие между классифицируемыми объектами устанавливается в зависимости от метрического расстояния между ними. Вопросы измерения близости объектов приходится решать при любых трактовках кластеров и различных методах классификации, причем, имеют место неоднозначность выбора способа нормировки и определения расстояния между объектами. Влияние метрики (DistanceFunction) будет проиллюстрировано схемами серии 3. Приведенные в этой серии результаты получены средствами включенного в ГеоБазаДанных из системы Wolfram Mathematica соответствующего программного приложения, которое допускает разные варианты задания DistanceFunction (Possible settings for Method). В системе Wolfram Mathematica различные меры расстояния или сходства удобны для различных типов анализа. Язык Wolfram предоставляет встроенные функции для многих стандартных измерений расстояния, а также возможность давать символьное определение произвольной меры. В частности, для анализа цифровых данных доступны следующие варианты метрик: Euclidean Distance, Squared Euclidean Distance, Normalized Squared Euclidean Distance, Manhattan Distance, Chessboard Distance, Bray Curtis Distance, Canberra Distance, Cosine Distance, Correlation Distance, Binary Distance, Warping Distance, Canonical Warping Distance.

Резюмируя серии 1–3 можно отметить, что для рассматриваемой конфигурации точек с данными, учитывая цифровое поле оригинала, однозначно назвать какой-то из рассчитанных вариантов предпочтительным трудно.

IV. Влияние учета значений в точках

В рассмотренных примерах серий 1–3, в результатах сходство или различие между классифицируемыми объектами устанавливается в зависимости от метрического расстояния между ними. Отдельно выполнен анализ, получены варианты классификации с использованием функции Wolfram Mathematica Cluster Classify, которая позволяет выполнять кластеризацию не только, принимая во внимание координаты точек рассеянного множества, но и значения в них. Другими словами – в рассмотренных в четвертой серии результатах в алгоритмах учитываются не пары координат, а тройки – координаты и значение в каждой точке.

Из полученных результатов следует, что для рассматриваемого набора данных учет значений в точках дополнительного явно положительного эффекта в реализации кластеризации не дает. Но подобные результаты полезны и важны, так как из сопоставления ясно, где необходимы дополнительные исходные данные.

Заключение

Рассматриваются вопросы инструментального наполнения и использования интерактивной компьютерной системы ГеоБазаДанных. Представлены и обсуждаются результаты кластеризации представительного набора данных типичной цифровой модели пространственного объекта.

V. Список литературы

1. Таранчук, В. Б. Компьютерные модели подземной гидродинамики / В. Б. Таранчук // Минск : БГУ, 2020. – 235 с.
2. Таранчук, В. Б. Методы и средства системы ГеоБазаДанных для адаптации компьютерных моделей. Примеры адаптации / В. Б. Таранчук // Информационные технологии и системы 2020 (ИТС 2020) : материалы междунар. науч. конф. (Республика Беларусь, Минск, 18 ноября 2020 года) = Information Technologies and Systems 2020 (ITS 2020). –Minsk, – 2020. – С. 15–17.