УДК 004(9)

ОСОБЕННОСТИ И ПРИМЕРЫ КЛАСТЕРИЗАЦИИ В СИСТЕМЕ ГеоБазаДанных

В. Б. Таранчук

Белорусский государственный университет, Беларусь, Минск, taranchuk@bsu.by

Обсуждаются методические и технические вопросы развития программной системы ГеоБазаДанных (ГБД). Отмечены новые функциональные возможности, обеспеченные включением в ГБД исполняемых модулей интеллектуального анализа данных системы компьютерной алгебры Wolfram Mathematica. Примерами на представительных наборах данных иллюстрируются варианты настройки алгоритмов кластеризации.

Ключевые слова: система ГеоБазаДанных; интеллектуальная адаптация цифровых полей; кластеризация.

FEATURES AND EXAMPLES OF CLUSTERING IN THE SYSTEM GeoBazaDannych

V. B. Taranchuk

Belarussian state university, Belarus, Minsk, Taranchuk@bsu.by

Methodological and technical issues of the development of the GeoBazaDannych software system (GBD) are discussed. The new functionality provided by the inclusion of executable data mining modules of the Wolfram Mathematica computer algebra system into the GDB is noted. Examples on representative data sets illustrate options for configuring clustering algorithms.

Keywords: system GeoBazaDannych; intelligent adaptation of digital fields; clustering.

Введение

Создание компьютерных цифровых моделей физики, механики, химии, биологии, экологии, геологии, других предполагает сбор, предобработку, анализ исходной предметной информации, замеров и наблюдений – формирование геоданных [1]. Особенности решения задач разработки и сопровождения компьютерных моделей со средствами их адаптации и самонастройки, подходы обработки, анализа, интерпретации используемых и получаемых геоданных отмечены в [2, 3]. Акцентируется, что на

данном этапе к числу приоритетных направлений исследований и разработок относится интеллектуальный анализ данных, перечислены соответствующие классы систем для его реализации [4].

Ниже обсуждаются полученные в среде системы ГеоБазаДанных результаты и методические рекомендации кластерного анализа [5] геоданных.

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, контроля характеристик и последующего анализа конкретных кластеров. Также сегментация используется для обнаружения нетипичных объектов – выбросов, другими словами, 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.

О системе ГеоБазаДанных используемых исходных данных

ГеоБазаДанных – комплекс интеллектуальных компьютерных подсистем, математического, алгоритмического и программного обеспечения наполнения, сопровождения и визуализации входных данных для имитационных и математических моделей, средств выполнения вычислительных экспериментов (ВЭ), алгоритмических и программных средств создания постоянно действующих компьютерных моделей. Основные компоненты системы ГеоБазаДанных описаны в [6]. Инструментами ГБД можно формировать цифровые описания пространственных распределений данных; конструировать и реализовывать интерактивные сценарии обработки и визуализации результатов вычислительных экспериментов, рассчитывать локальные и интегральные характеристики моделируемых объектов, выводить на твердые копии тематические карты.

Для пояснения, в чем состоит дополнение предыдущих и новизна представляемых в данном исследовании результатов, отметим, что [3] изложены примеры интерактивного формирования в ВЭ отвечающих интуитивным требованиям эксперта цифровых моделей геологических объектов; примеры аппроксимации и восстановления цифрового поля, его интерактивной адаптации с оценками точности результатов в программ-

ном комплексе ГГМЗ. В [2] представлены и обсуждены результаты применения искусственных нейронных сетей при анализе и интерпретации геоданных, описаны возможности анализа погрешностей. В представляемой работе обсуждаются алгоритмы кластерного анализа геоданных, инструменты их реализации в среде ГБД. Сопоставлением на примерах специально подготовленных «конфликтных» данных приведены и обоснованы рекомендации выбора оптимальных параметров алгоритмов классификации в принимаемом при интерпретации приближении.

При создании математических моделей физики, механики, химии, биологии, геологии, экологии одной из ключевых является задача цифрового описания пространственных объектов, их структуры и свойств. Решения задач компьютерного моделирования объектов многими реализуются, следуя концепции, в соответствии с которой ядром и теоретической основой для построения моделей является цифровое описание ограничивающих объем поверхностей, построение генерализованных поверхностей, описывающих геометрию объекта, последовательность слоев – своеобразная структурная «этажерка». Для структурированной по слоям трехмерной модели применим подход построения в режиме «конструктор», когда сборка и редактирование модели производятся по частям, которыми служат отдельные элементы. Для принимаемых слоев в описание включаются распределения изучаемых параметров на каждом из них.

В практике, как правило, исходные данные – значения наблюдаемого параметра в точках с известными геометрическими координатами, а сами точки с замерами чаще размещены на площади в плане нерегулярно – например, данные наблюдений сети станций, пунктов постоянного учета, трасс наблюдения, точек замеров, контроля. В цифровом описании для моделей значения параметра восстанавливаются на равномерной прямоугольной (правильной регулярной) сетке. Подобные цифровые поля не что иное, как сеточные функции, а с ними можно эффективно работать средствами численного анализа. В системе ГБД реализованы обоснованные, апробированные в разных приложениях методические решения, математические методы расчета аппроксимирующих цифровых полей, высоко производительные алгоритмы составной сплайн-аппроксимации.

Примеры (имитирующих практику сбора) представительных наборов данных для ВЭ, иллюстрирующие результаты и выводы в настоящей работе, подготовлены с частично измененными данными [2]. Изменения сделаны, чтобы не было повторяющихся элементов, а также исключены типовые фигуры (пирамиды, траншеи), потому что такие объекты в ГБД идентифицируются соответствующим модулем системы, которая вос-

производит их точно. Также, и это важно отметить, добавлено возмущение «усеченный по вертикали холм». Оно в отличие от других не является непрерывно помещенным на базовую поверхность – имитируется разлом. Как и в упомянутых публикациях, моделируемая поверхность (эталон) имеет полное математическое описание.

При обсуждении результатов в докладе будут показаны иллюстрации формы эталонной поверхности в вариантах Plot3D (виды поверхностей) и RegionPlot3D (виды объемов), изолиний и карт плотностей уровней высот.

Инструментарий, примеры кластерного анализа геоданных

Эффекты числа кластеров. Одной из важнейших проблем сегментации является определение количества кластеров. Проведены расчеты, подготовлены и будут демонстрироваться и поясняться серии результатов, полученных с установками по умолчанию, используя функцию Wolfram Mathematica FindClusters с числом кластеров 3, 4, 5, 6. Из сопоставления приведенных вариантов следует, что необходимы дополнительные уточнения метода, метрики и других параметров алгоритмов кластеризации.

Эффекты принятого метода кластеризации. Кластерный анализ допускает много различных типов методов/алгоритмов кластеризации [5] для определения конечного результата. В рассмотренных примерах по выбору метода и метрики использована априорная информация, задается число кластеров – 5. Почему столько – принято во внимание, что в исходных данных замеры проведены для поверхности, которая включала 5 разных её искажений с индивидуальным позиционированием возмущений. Эффекты принятого метода кластеризации (Possible settings for Method) иллюстрируются в примерах серии расчетов только для пар координат, когда учитывается только относительное положение точек рассеянного множества. В программном модуле использована функция FindClusters с разными критериями CriterionFunction, норма в примерах основной серии вычислялась по метрике DistanceFunction EuclideanDistance.

Вообще говоря, включенное в ГеоБазаДанных из системы Wolfram Mathematica соответствующее программное приложение допускает варианты метода кластеризации (CriterionFunction): Automatic, Agglomerate (find clustering hierarchically), Optimize (find clustering by local optimization), DBSCAN (density-based spatial clustering of applications with noise), GaussianMixture (variational Gaussian mixture algorithm), JarvisPatrick (Patrick clustering algorithm), KMeans (k-means clustering algorithm), KMedoids

(partitioning around medoids), MeanShift (mean-shift clustering algorithm), NeighborhoodContraction (displace examples toward high-density region), SpanningTree (minimum spanning tree-based clustering algorithm), Spectral (spectral clustering algorithm). Все перечисленные методы проанализированы, результаты иллюстрируются интегрированными картами, на которых показаны изолинии эталонного цифрового поля, профили и точки замеров, окрашенные индивидуальными цветами примитивы узлов в конкретных кластерах, которые дополнительно окаймлены. Какие методы сегментации применены в расчетах, записано в заголовках.

Влияние метрики. В рассмотренных примерах основной 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) иллюстрируют специальные расчеты, в которых показаны результаты для вариантов задания DistanceFunction (Possible settings for Method): EuclideanDistance, SquaredEuclideanDistance, ManhattanDistance, NormalizedSquaredEuclideanDistance, ChessboardDistance, BrayCurtisDistance, CanberraDistance, CosineDistance, CorrelationDistance, BinaryDistance, WarpingDistance, CanonicalWarpingDistance. Алгоритмические особенности перечисленных метрик можно уточнить в [7]).

Итоги сопоставления – из полученных результатов для рассматриваемой конфигурации точек с данными, учитывая цифровое поле эталона, однозначно назвать какой-то из показанных вариантов предпочтительным трудно.

Влияние учета значений в точках. В рассмотренных и приведенных результатах заключительной серии представлены варианты классификации с использованием функции Mathematica ClusterClassify, которая позволяет выполнять кластеризацию не только, принимая во внимание координаты точек рассеянного множества, но и значения в них.

Другими словами – в представленных в такой серии результатах в алгоритмах учитываются не пары (X_i, Y_i), а тройки – (X_i, Y_i, Z_i). Из сопоставления результатов следует, что для рассматриваемого набора данных дополнительный учет значений в точках явно положительного эффекта в кластеризации с целью выявлений возмущений не дает, но получаемые результаты полезны и важны, так как на участках отличия ясно, где необходимы дополнительные исходные данные (уплотнение точек замеров).

Заключение

В тезисах рассматриваются вопросы инструментального наполнения и использования интерактивной компьютерной системы ГеоБазаДанных. Изложены и обсуждаются результаты кластеризации для представительного набора данных типичной цифровой модели пространственного объекта.

Библиографические ссылки

1. *Савиных В. П., Цветков В. Я.* 2014. Геоданные как системный информационный ресурс. Вестник Российской академии наук. 84(9): 826–829.

2. *Taranchuk, V.* Tools and examples of intelligent processing, visualization and interpretation of GEODATA / V. Taranchuk // Modelling and Methods of Structural Analysis. IOP Publishing. IOP Conf. Series: Journal of Physics: Vol. 1425 (2020) 012160. – P. 9.

3. *Taranchuk, V.* Interactive Adaptation of Digital Fields in the System GeoBaza-Dannych / V.B. Taranchuk // Communications in Computer and Information Science. Book series Springer (CCIS, volume 1282): 2020. – P. 222-233.

4. *Шайтура, С. В.* Интеллектуальный анализ данных геоданных /С.В. Шайтура // Перспективы науки и образования. – 2015. – № 6 (18). С. 24–30.

5. *Everitt B. S, Landau S, Leese M, Stahl D.* Cluster Analysis. 5th Edition (John Wiley & Sons). 2011. –360 p.

6. *Таранчук, В. Б.* Компьютерные модели подземной гидродинамики / В.Б. Таранчук. – Минск : БГУ, 2020. – 235 с.

7. *Amigó E, Gonzalo J, Artiles J et al.* A comparison of extrinsic clustering evaluation metrics based on formal constraints. Inf Retrieval vol. 12, 2009 pp 461–486 https://doi.org/10.1007/s10791-008-9066-8