

Заключение

В ходе выполнения работы была решена проблема описания структуры резервирования технических систем путём создания матрицы смежности графа и создания структурированного текстового файла по заданным правилам. Результаты работы используются для повышения характеристик надёжности технических систем.

Список литературы

1. К. Берж. Теория графов и ее применения / Издательство ИНОСРАННОЙ ЛИТЕРАТУРЫ, Москва 1960. – 142 с.
2. Шкляр, Виктор Николаевич. Надёжность систем управления: учебное пособие / В.Н. Шкляр; Национальный исследовательский Томский политехнический университет (ТПУ). – Томск: Изд-во ТПУ, 2011. – 126 с.

УДК 004

КЛАСТЕРИЗАЦИЯ ОБЪЕКТОВ НА ОСНОВЕ НЕЧЕТКОЙ ЛОГИКИ С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМА DBSCAN

К.С. Арышева

Научный руководитель: С.В. Аксенов, доцент каф. ОСУ ИК ТПУ

Томский политехнический университет

E-mail: varks@tpu.ru

This article deal with methods of clusterization DBScan based on fuzzy logic. This type of clustering allows solving native tasks and provides flexible results. This article spoken in detail about DBScan clustering. Text gives information about center-orientating and density-orienting clustering technics and describes implementation. The article is of interest to students, who deal with scientific project.

Keywords: fuzzy logic, clustering, mountain clustering, DBScan algorithm.

Ключевые слова: нечеткая логика, кластеризация, горный метод, алгоритм DBScan.

Развитие науки в настоящее время, наблюдения и эксперименты, рост количества получаемой информации и необходимость ее обработки требуют создания высокопроизводительных вычислительных систем для кластеризации или классификации огромного неструктурированного множества данных. В данной работе рассматривается задача кластеризации астрономических объектов с использованием методов нечеткой логики.

Нечеткая логика представляет собой обобщение традиционной логики и теории множеств, базирующееся на понятии нечеткого множества, которое расширяет определение классического множества, допуская значение функции принадлежности множеству в интервале $[0;1]$. Это означает, что объект может принадлежать множеству с некоторой степенью. Такой тип принадлежности позволяет описывать более естественные задачи кластеризации объектов. Для определения кластера используются логические выражения вида: если $x_1 = a$, $x_2 = b$, ... $x_n = n$, то $y \in I$, где y – объект кластера I , имеющий $\{x_1; x_n\}$ параметров.

В дополнение к алгоритмам нечеткой кластеризации большое внимание привлекает алгоритм кластеризации на основе плотности *DBScan*. Данный алгоритм позволяет оценить точки кластеризуемого множества как основные, пограничные или шумовые, что также можно использовать при построении правил кластеризации и оценке расположения кластера в пространстве. Алгоритм *DBScan* основан на центр-ориентированном подходе.

В центр-ориентированном подходе плотность рассчитывается для отдельной точки в наборе данных путем пересчета точек внутри определенного радиуса, включая саму точку. Этот метод

прост в реализации, но плотность любой момент будет зависеть от заданного радиуса. Например, если радиус достаточно велик, то все точки будут иметь плотность m – число точек в наборе данных. Точно так же, как если радиус слишком мал, то все точки будут иметь плотность 1.

Центр-ориентированный подход к плотности позволяет классифицировать точку внутри плотной области (основная точка или ядро), по краю плотной области (пограничная точка или граница), или в скудно оккупированном регионе (шум или фоновая точка). Основные точки – точки, которые находятся внутри кластера основанного на плотности. Точка является основной в том случае если количество точек внутри данного радиуса, определяемое функцией дистанции и определяемого пользователем радиуса Eps , превышает определенное пороговое значение, $MinPts$, который также является параметром указанный пользователем. Граница – точки не являющиеся основными, но попадающими в окрестность радиуса основной точки. Граница может оказаться внутри радиуса нескольких основных точек. Шум – любая точка, которая не является ни границей ни основной.

Определение $DBScan$ можно дать следующим образом: любые две основные точки, находящиеся на достаточно близком расстоянии друг от друга, находятся в одном кластере. Точно так же, любая пограничная точка, находящаяся достаточно близко к основной точке, находится с ней в одном кластере. Точки шума отбрасываются.

Алгоритм $DBScan$ можно разбить на следующие этапы:

1. Пометить все точки, как основные, пограничные или шумовые,
2. Отбросить точки шума,
3. Соединить все основные точки, находящиеся на расстоянии Eps радиуса одна от другой,
4. Объединить каждую группу соединенных основных точек в отдельный кластер,
5. Назначить каждую пограничную точку одному из кластеров, ассоциированных с ней основных точек.

Базовая сложность $DBScan$ алгоритма – $O(m \times \text{время, затраченное на поиск, чтобы найти точки в } EPS\text{-окрестности})$, где m – число точек. В худшем случае, это сложность $O(m^2)$. Перед реализацией алгоритма необходимо выделить параметры: радиус окрестности Eps и пороговое значение плотности $MinPts$. Основной способ – оценка поведения расстояния от точки до ее k -го ближайшего соседа – расстояние k . Для точек, принадлежащих некоторому кластеру, значение k будет мало, если k не больше, чем размер кластера. Тем не менее, для точек, которые не принадлежат кластеру, таких как точки шума, k -расстояние будет относительно большим. Поэтому, если вычислять k -расстояние для всех точек данных для некоторого k и сортировать их в порядке возрастания, ожидается увидеть резкое изменение в значениях k -расстояния что соответствует подходящему значению Eps . Если выбрать это расстояние в качестве параметра Eps и значение k в качестве параметра $MinPts$, то точки, для которых k -расстояние меньше Eps будут помечены как основные точки, в то время как другие точки будут помечены как шум или пограничные точки.

Так как $DBScan$ использует определение кластера, основанное на плотности, он относительно устойчив к шуму и может обрабатывать кластеры произвольных форм и размеров. $DBScan$ можете найти кластеры, которые не могут быть найдены с помощью алгоритма k -средних, однако, его эффективность снижается, когда кластеры различны по плотности.

Список литературы

1. Joseph M. Barone, Dimitar P. Filev, Ronald R. Yager, “Moimtain method-based fuzzy clustering: methodological considerations”, International Journal of General Systems, vol 23:4, 281–305, (1995).
2. Алгоритм DBSCAN, Blogpost [2015] URL: <http://matchast.blogspot.ru/2014/11/dbscan.html>.
3. Обзор алгоритмов кластеризации числовых пространств данных, SavePearlHarbor [2015] URL: <http://savepearlharbor.com/?p=164417>.