УДК 519.7:007.52

А.П. Чапланов, Е.Б. Чапланова

Харьковский национальный университет радиоэлектроники

КЛАСТЕРИЗАЦИЯ ОБЪЕКТОВ С ПОМОЩЬЮ АЛГОРИТМА DBSCAN

Рассмотрен алгоритм кластеризации по плотности DBSCAN. Он основывается на поиске некоторых областей, плотность объектов внутри которых превышает заданный порог.

алгоритм кластеризации по плотности, кластер, метрика Минковского

Введение

Сегодня наблюдается интенсивный рост объемов информации. Эта информация разнородна, требует упорядочивания для последующего анализа. Очевидно, что анализировать необходимо однородные массивы данных, следовательно, было бы удобно, если бы необходимая информация была распределена в соответствии с некоторыми заданными группами. Эти группы называют классами, и, используя определенные алгоритмы классификации, распределяют информацию в соответствии с заданными классами. Однако, если классы априорно не заданы, то возникает задача кластеризации. Решение такой задачи дает возможность использовать полученные результаты для корректного решения задачи классификации.

Данная работа посвящена рассмотрению алгоритма кластеризации, имеющего название DBSCAN (*Density Based Spatial Clustering of Application with Noise*) [1]. Этот метод кластеризации основан на соединении некоторых областей, плотность объектов внутри которых превышает некоторый заданный порог.

Постановка задачи. Основной целью проведения данного исследования является выявление достоинств и недостатков алгоритма DBSCAN при работе с кластерами различной природы (формы), а также программная реализация алгоритма.

Кластеризация по плотности

В данном случае, под кластеризацией понимается деление заданного множества точек данных (объектов) на подгруппы, каждая из которых, насколько это возможно, однородна. В основе метода кластеризации DBSCAN лежит объединение некоторых объектов в соответствии с их внутригрупповым «соединением». Для проведения корректной процедуры кластеризации необходимо указать критерии, по которым объекты будут объединены в кластеры. Прежде всего, необходимо сказать, что кластеры представляют собой плотные области некоторых объектов в пространстве данных, разделенных между собой объектами, плотность которых значительно ниже. Расположение точек в одном кластере обусловлено их соединением, т.е. некоторой связью между собой.

Плотность точек для данной точки X определяется двумя параметрами. Первым из них является α – радиус «соседства» (приближенности) точки X. Тогда множество $M_\alpha(X)$ будет включать в себя такие точки f_i, $\left(i=\overline{1,n}\right)$, для которых следующее неравенство будет истинно

$$\mathrm{dist}(X, f_i) \le \alpha, \ \left(i=\overline{1,n}\right). \qquad (1).$$

Функция $\mathrm{dist}(\mathrm{var}1, \mathrm{var}2)$ определяет расстояние между объектами выборки D. Это расстояние может вычисляться различными способами, например, как евклидово расстояние или с помощью метрики Минковского.

Вторым параметром определения плотности точек является MCP – это минимальное количество точек, которые расположены ближе всего к данной точке согласно определенному радиусу α.

Точка f_i, $\left(i=\overline{1,n}\right)$ будет являться окруженной точкой (согласно α и MCP) если

$$M_\alpha(X) \le \mathrm{MCP}. \qquad (2)$$

Это значит, что точка f_i, $\left(i=\overline{1,n}\right)$ окруженная, если количество «соседствующих» точек выборки D окажется бо́льшим, либо равным значению параметра MCP (рис. 1).

Точка X является прямо достижимой по плотности от точки f (при соответствующих α и MCP), если точка $X \in M(X)$, т.е. точка X – это одна из точек f для другого окружения (соседства), где f – окруженная точка (рис. 2).

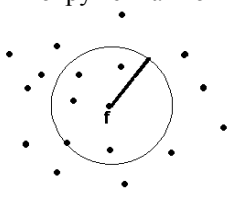

Рис. 1. Окруженная точка при MCP = 5

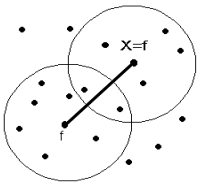

Рис. 2. Точка X прямо достижима по плотности от точки f

Достижимость по плотности – это транзитивное замыкание прямо достижимой по плотности точки. Точка f достижима по плотности из точки X, но точка X не достижима по плотности из точки f (рис. 3).

Точка X соединена (связана) по плотности с точкой f (согласно α и MCP) если существует точка e такая, что обе точки X и f являются достижимыми от точки e (согласно α и MCP) (рис. 4).

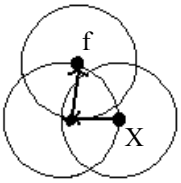

Рис. 3. Достижимость по плотности

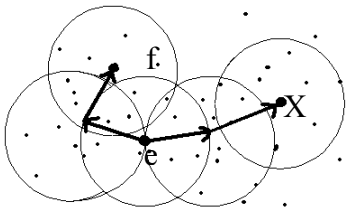

Рис. 4. Соединение по плотности

Кластер, сформированный на основе размещения объектов по плотности должен удовлетворять таким свойствам: максимальность; связность.

В этом случае, под кластером понимается непустое подмножество точек G из набора данных D, которое удовлетворяет вышеупомянутым свойствам, причем, максимальность интерпретируется таким образом: если $X \in G$ и f достижима по плотности от точки X, тогда и $f \in G$, это значит, что обе точки принадлежат одному кластеру.

Свойство связности гласит, что каждый объект в подмножестве G соединен по плотности со всеми объектами кластера (при заданных α и MCP).

Все объекты из набора данных D представляют собой совокупность подмножеств

$$D = \{G_1, G_2, ..., G_n, N\}, \qquad (3)$$

где G_1, G_2, …, G_n – кластеры, образованные по плотности; N – некоторое подмножество, объекты которого не принадлежат ни одному из подмножеств G_1, G_2, …, G_n.

Реализация метода

Реализация алгоритма DBSCAN может быть разделена на два этапа. В первую очередь из всего набора данных *D* необходимо выделить те точки, которые являются окруженными. Затем выполнять следующую процедуру: для каждого объекта *X* из набора данных *D* определить:

1) принадлежит ли текущий объект к какому-нибудь из кластеров;

2) является ли текущий объект окруженной точкой.

Если текущий объект – окруженная точка, то все объекты, достижимые по плотности от текущего объекта, соединяем в новый кластер. В противном случае, если объект не является окруженной точкой и не достижим по плотности ни от какого объекта, то текущий объект – выброс.

Псевдокод алгоритма DBSCAN можно представить следующим образом:

```
for ∀X ∈ D
  {
   if (X ∈ G_i, i = 1,n)
   {
    if (X_i ∈ M_α(X))
    { find X_i ∈ D достижимы по плотности
     from X_i ∈ M_α(X)}
   else if (X_i ∉ M_α(X) and X не достижим
   от любого другого объекта)
     X ∈ N
   }
  }
```

причем параметры α и MCP задаются пользователем.

Очевидно, что выбор параметров α и MCP весьма критичен для данного алгоритма, поэтому предлагается использовать некоторую процедуру, которая поможет избежать эвристического подбора параметров. В связи с этим предлагается рассмотреть некоторый обобщенный подход выбора параметров α и MCP.

Основное влияние выбора параметров α и MCP ощутимо при кластеризации так называемых «тонких» кластеров, т.е. кластеров с наименьшей плотностью объектов. Эта закономерность основывается на следующем наблюдении: пусть d – это расстояние от точки X к ее k-му ближайшему соседу. Тогда расстояние d, на котором находятся соседствующие с точкой X точки, включает в себя в точности (k + 1) точку практически для всех таких точек X. Расстояние может содержать более чем (k + 1) точку лишь в том случае, если все точки находятся от точки X в точности на расстоянии d. Вероятность такого распределения точек в выборке очень мала, поэтому данный вариант в рассмотрение не берется.

Для заданного значения k, $k = \overline{1,n}$, где n – число объектов в наборе данных D, определяется функция k_dist, с помощью которой будет вычисляться расстояние от каждой точки до ее k–го ближайшего соседа. Все точки из набора данных D упорядочиваются по убыванию согласно полученным значениям функции k_dist (так как каждой точке из D будет соответствовать одно значение k_dist). По полученным упорядоченным точкам строим график, который даст некоторое представление относительно плотности распределения объектов в выборке D.

Для произвольно взятой точки X устанавливаем параметр $\alpha = k\_dist(X)$ и $MCP = k$. В этом случае все точки, у которых значение функции k_dist меньше или равное значению $k\_dist(X)$, будут являться окруженными точками.

Для нахождения желаемых значений искомых параметров необходимо найти пороговую точку с максимальным значением k_dist в самом «тонком» кластере выборки D. Пороговая точка – это первая точка некоторой «впадины» на графике, который был построен в соответствии с отсортированными значениями k_dist, для примера опреде-

ляем k = 4 (рис. 5). Все точки, значение k_dist которых больше, чем значение k_dist пороговой точки (это точки, расположенные слева от пороговой точки (рис. 5)) рассматриваются как выбросы. Оставшиеся точки (справа от пороговой точки) будут относиться к определенному кластеру.

Рис. 5. Определение пороговой точки

В общем случае, автоматическое определение «первой впадины» и нахождение пороговой точки является очень трудной задачей. Значительно легче определять пороговую точку, визуализировав отсортированные значения k_dist.

С помощью данного графика также можно определить значение α, но лишь в том случае, если пользователю удастся оценить процент выбросов во всем наборе данных D (по графику). В этом случае можно описать некоторую функцию, в качестве аргумента которой будет выступать процент выбросов в данной выборке, и которая будет на основании этого процента определять пороговую точку. Тогда, соответствующее этой пороговой точке значение функции k_dist принимается в качестве параметра α.

Моделирование метода DBSCAN было произведено в среде MatLab на m–языке.

Алгоритм был протестирован на искусственно сгенерированных выборках, одна из которых представляет собой спираль, объекты другой расположены на окружностях различного диаметра (рис. 6).

При проведении эксперимента были получены следующие результаты: очевидно, что в зависимости от выбора параметров α и MCP зависит и плотность объектов кластера, и количество кластеров. Например, если использовать эвристический подход выбора параметров α и MCP, то при $\alpha = 0.5$ и $MCP = 2$, получаем, что первый объект выборки (внутренняя точка) и 44 последних объектов были вынесены в отдельный кластер – это выбросы (рис. 7).

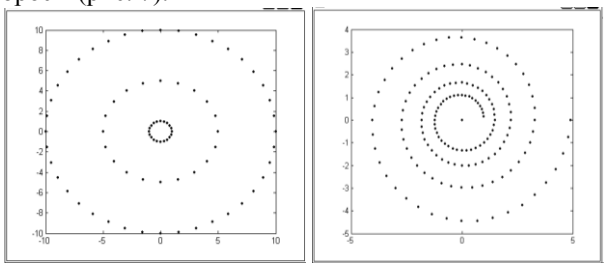

Рис. 6. Выборки для тестирования

При выборе значений параметров на основании описанных процедур определения α и MCP ($\alpha = 1$ и $MCP = 2$), получаем, что лишь один объект был отнесен к выбросам – это первый объект выборки (на рисунке – внутренняя точка «спирали») (рис. 8).

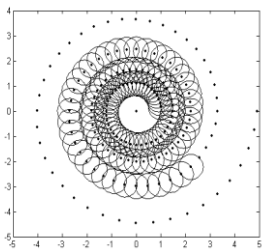

Рис. 7. Разбиение на два кластера

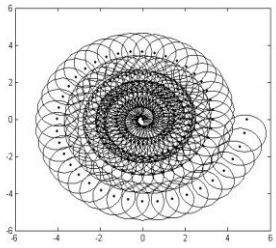

Рис. 8. Определение единственного выброса

Что касается второй выборки, то были получены такие результаты: при использовании процедур выбора параметров α и MCP, получили $\alpha = 3{,}7$ и $MCP = 3$, при этом все объекты искусственной выборки были разделены на 3 кластера без выбросов. (рис. 9). Однако, при эвристическом выборе параметров (на $\alpha = 4{,}3$ и $MCP = 5$), получаем объекты, разделенные на три кластера, один из которых кластер, содержащий выбросы (рис. 10).

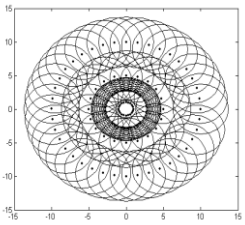

Рис. 9. Разделение на три кластера

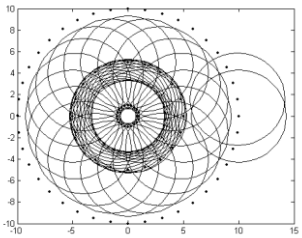

Рис. 10. Определение кластера с выбросами

Выводы

Рассмотренный алгоритм кластеризации обладает рядом преимуществ, а именно: алгоритм не чувствителен к выбросам, то есть в процессе кластеризации все выбросы выносятся в отдельный кластер с заранее заданной меткой; данный метод не требует априорного задания количества кластеров; использование данного метода позволяет работать с кластерами различной природы (формы); применение данного алгоритма позволяет работать с выборками большого объема. Кроме того, использование вышеуказанных процедур определения параметров α и MCP, дает возможность работать с n-мерными объектами (это объекты, количество атрибутов которых более 3) при условии адекватного выбора функции для расчета расстояния (в общем случае можно использовать метрику Минковского).

Однако существенным недостатком является достаточно трудоемкая процедура определения необхо-

димых параметров для корректной работы алгоритма. Тем не менее, использование алгоритмов кластеризации по плотности на сегодняшний день является достаточно эффективным и перспективным. Алгоритм DBSCAN, в свою очередь, лег в основу целого ряда иерархических методов кластеризации.

Список литературы

1. Martin Ester, Hans-Peter Kriegel, J&g Sander, Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, KDD'96. – P. 226-231.

2. Jiawei Han, Micheline Kamber. Data Mining: Concepts and Techniques: – Academic Press, 2001. – P. 363-370.

3. Haykin S. Neural Networks. A Comprehensive Foundation. – Upper Saddle River, N.1: Prentice Hall, Inc., 1999. – 1104 c.

Поступила в редколлегию 28.09.2006

Рецензент: д-р техн. наук, проф. Е.В. Бодянский, Харьковский национальный университет радиоэлектроники, Харьков.