УДК 004.021

Ву Вьет Тханг[1], *Д. В. Пантюхин*[1], *А. И. Галушкин*[1,2]

[1]Московский физико-технический институт (государственный университет)
[2]Центр информационных технологий и систем органов исполнительной власти (ЦИТиС)

Гибридный алгоритм кластеризации FastDBSCAN

Кластеризация – это одна из самых важных задач интеллектуального анализа данных (DataMining). Хотя существует много исследованных способов кластеризации, таких как K-Means, Fuzzy C-Means и др., но существует проблема повышения точности и ускорения алгоритмов кластеризации, из-за того, что в течение 10 последних лет количество обрабатываемых данных существенно выросло. В данной работе представлен новый подход для ускорения алгоритма кластеризации на основе плотности DBSCAN (Density Based Spatial Clustering of Applications with Noise) [1]. Практические исследования показывают, что скорость кластеризации предложенного алгоритма выше при сохранении точности.

Ключевые слова: кластеризация, DBSCAN, K-means.

Vu Viet Thang[1], *D. V. Pantiukhin*[1], *A. I. Galushkin*[1,2]

[1]Moscow Institute of Physics and Technology (State University)
[2]Center of Information Technologies and Systems of Executive power

Hybrid clustering FastDBSCAN algorithm

Clustering is one of the most important tasks of data mining.. Although there are a lot of ways to explore clustering such such as K-Means, Fuzzy C-Means et al., there is a problem of increasing the accuracy and acceleration of algorithms for clustering, because during the past 10 years the amount of data to be processed increases substantially. This paper presents a new approach to speed up the clustering algorithm based on DBSCAN density (Density Based Spatial Clustering of Applications with Noise). The practical studies show that the speed of clustering algorithm proposed is higher, while maintaining accuracy.

Key words: clustering, DBSCAN, K-Means.

1. Введение

Кластеризация – это процесс разбиения множества с N элементами x_1,x_2, $\ldots$, x_n (x_i имеет размерность m) на K кластеров, так, чтобы в каждом кластере все элементы были схожи в каком-то смысле. x_i могут быть числовыми, категориальными или смешанными данными. В данной работе сделан акцент на ускорении алгоритма DBSCAN. Предложен новый алгоритм FDBSCAN (Fast Density based Clustering), основанный на алгоритме кластеризацib, K-means и выборе примеров так, чтобы пропорция плотности между кластерами не изменялась.

2. Метод кластерного анализа DBSCAN

Алгоритм DBSCAN – это плотностный алгоритм для кластеризации пространственных данных с присутствием шума, был предложен Мартином Эстер, Гансом-Питером Кригель и их коллегами в 1996 году как решение проблемы разбиения (изначально пространственных) данных на кластеры произвольной формы [1]. Большинство алгоритмов, производящих плоское разбиение, создают кластеры по форме близкие к сферическим, так как минимизируют расстояние точки до центра кластера. Авторы DBSCAN экспериментально показали, что их алгоритм способен распознавать кластеры различной формы, например, как на рис. 1.

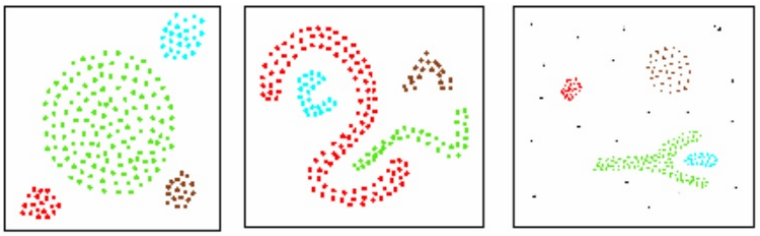

Рис. 1. Примеры кластеров произвольной формы, распознанных DBSCAN

Идея, положенная в основу алгоритма, заключается в том, что внутри каждого кластера плотность точек (объектов) заметно выше, чем плотность снаружи кластера, а также плотность в областях с шумом ниже плотности любого из кластеров. Еще точнее, для каждой точки кластера ее окрестность в диапазоне заданного радиуса должна содержать не менее некоторого числа точек, которое задается пороговым значением. В общем случае алгоритм DBSCAN имеет квадратичную вычислительную сложность из-за поиска Eps-соседства- $O(N^2)$. Однако авторы алгоритма использовали для этой цели специальную структуру данных – R*-деревья, в результате поиск Eps-соседства для одной точки – O(logn). Общая вычислительная сложность DBSCAN – O(n*logn).

3. Ускорение алгоритма DBSCAN за счет использования алгоритма K-means

Одно из достоинств алгоритма DBSCAN – он хорошо работает с множествами данных произвольной формы. Примеры указаны на рис. 2.

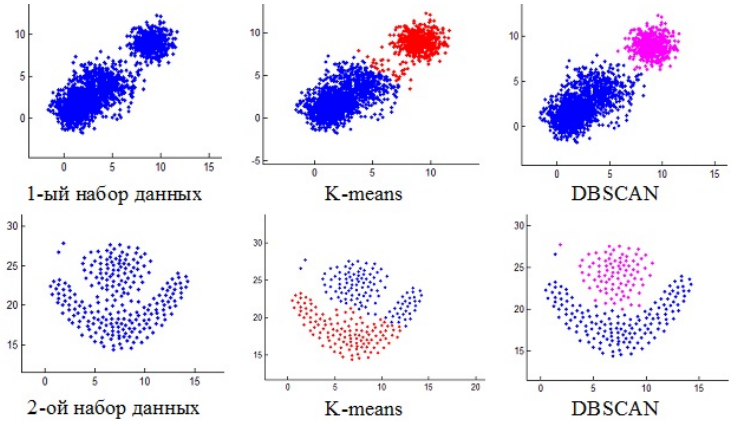

Рис. 2. Примеры кластеризации данных алгоритмом DBSCAN и K-means

Видно, что для первого набора данных оба алгоритма, K-means и DBSCAN, хорошо проводят кластеризацию, так как множество данных имеет круглую форму. Но для второго набора данных K-means работает хуже, чем DBSCAN, так как у второго набора данных форма на много сложнее, чем у первого. Алгоритм K-means имеет низкую вычислительную сложность O(K*N) – это основное достоинство K-means и он хорошо работает с большим числом данных. DBSCAN довольно медленно работает с большим количеством данных. Поэтому для ускорения DBSCAN будем использовать алгоритм K-means. K-means применяется для разбиения множества данных D на K кластеров (K достаточно большое, чтобы покрыть все множество данных). После этого выбирается случайно t% данных из каждого кластера и получается новое множество E. Так делается для того, чтобы относительная плотность между регионами множества D не изменилась. После применения алгоритма K-means для поиска промежуточных кластеров, используется алгоритм DBSCAN на множестве E.

Алгоритм FDBSCAN

Вход: множество D, количество промежуточных кластеров для K-Means K, доля t;

Выход: кластеры и аномалии.

Шаг 1: инициализируем K центров случайно,

Шаг 2: реализуем алгоритм K-Means с K центрами,

Шаг 3: взяв $t \cdot 100$ процентов каждого полученного кластера, получаем новое множество E,

Шаг 4: выполняем алгоритм DBSCAN с множеством E,

Шаг 5: отображаем обратно результаты, чтобы получить кластеры и аномалии для множества D.

Вычислительная сложность FDBSCAN определяется вычислительной сложностью K-means и DBSCAN. В общем случае вычислительная сложность K-means- O(K.N), где K – количество кластеров, N – количество данных. Сложность DBSCAN – $O((t \cdot N)^2)$, где t – заданная доля. Общая вычислительная сложность FDBSCAN – $O(K.N+(t \cdot N)^2)$.

4. Экспериментальная оценка результатов кластеризации с DBSCAN и FDBSCAN

Выполним алгоритмы DBSCAN и FDBSCAN с шестью популярными множествами D31, t1.2k, t4.8k, t5.8k, t8.8k, t7.10k с произвольной формой, разным количеством данных [2]. Для измерения точности мы используем ранд-статистики (Randstatistic) [3], которые измеряют аналогичность между двумя наборами кластеров X и Y, имеющих одно и тоже количество точек n (объектов), как: $R = (a + b)/(n/2)$, где a – это количество пар объектов, отнесенных к одному и тому же кластеру в обоих X и Y, а b – количество пар объектов, отнесенных к различным кластерам в X и в Y. Результаты показаны на рис. 3 и 4.

5. Подход к обнаружению компьютерных атак за счет применения алгоритмов кластеризации данных

Компьютерные сети за несколько последних десятилетий из чисто технического решения превратились в глобальное явление, развитие которого оказывает влияние на большинство сфер экономической деятельности. Параллельно с развитием компьютерных сетей стало наблюдаться все больше попыток несанкционированного доступа (атак). Поэтому необходимо решение для обнаружения компьютерных атак. Признаки атак обнаруживаются в большом количестве измеряемой информации (логи, данные мониторинга и др.), что требует повышения скорости обработки информации. Одним из подходов к обнаружению атак является метод обнаружения аномалий, основанный на кластеризации измеряемых данных (признаков атак). Объединяя данные в компактные кластеры, проводят анализ типичных представителей каждого кластера и принимают решение о том, являются ли такие данные признаком атаки или нет. Затем это решение переносится на всех представителей исследуемого кластера. Такой подход существенно сокращает объемы необходимой для успешной классификации атаки информации (обучающего множества). Поскольку кластеры могут принимать в многомерном пространств сложные формы, оправдано применение алгоритма кластеризации DBSCAN. С другой стороны, как уже отмечалось, DBSCAN имеет высокую вычислительную сложность, поэтому предложенное решение – алгоритм кластеризации FastDBSCAN, позволит существенно сократить время обнаружения. Поэтому нашей ближайшей задачей станет исследование работы алгоритма FastDBSCAN на нескольких базах данных компьютерных атак, таких как KDD'cup 99 [4] или ADFA [5], для оценки производительности этого алгоритма.

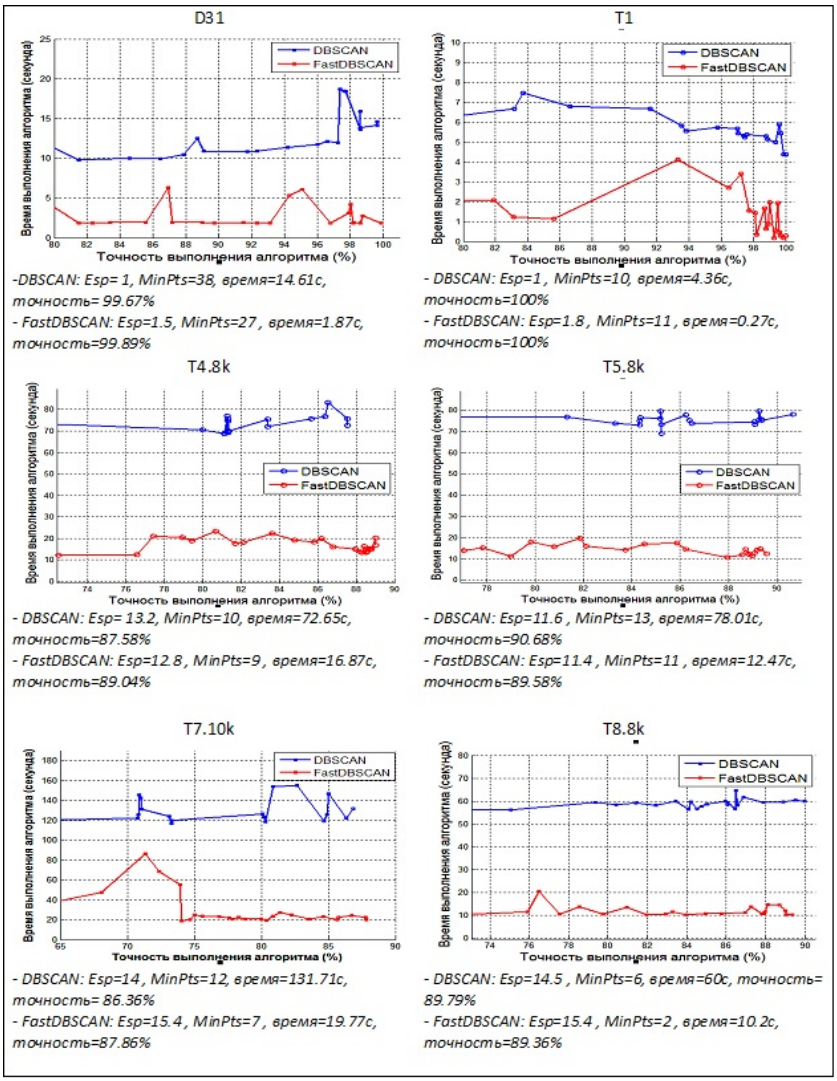

Рис. 3. Сравнение времени выполнения алгоритмов DBSCAN и FDBSCAN (в подписи приведены значения параметров для лучших по точности результатов)

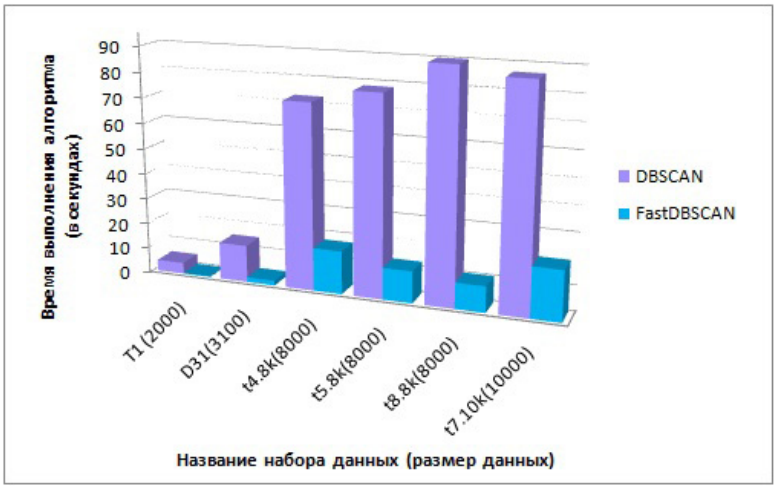

Рис. 4. Зависимость времени выполнения алгоритма от размера набора данных (для лучших по точности результатов)

6. Заключение

Представлен новый способ ускорения алгоритма кластеризации DBSCAN, основанный на применении алгоритма K-means, названный FDBSCAN, который позволяет существен-

но, до 3–4-х раз, ускорить кластеризацию данных при сохранении точности, достигаемой оригинальным DBSCAN. Полученные результаты подтверждены интенсивными экспериментальными исследованиями на ряде тестовых множеств. Предложенный способ может быть использован для решения многих классов задач кластеризации, требующих сокращения времени решения, например, таких задач, как: распознавание лиц, обнаружение компьютерных атак [4], обработка изображений, документов и др. Дальнейшее совершенствование метода FDBSCAN необходимо вести в следующих направлениях:

1) кластеризация данных, имеющих различную плотность;
2) распараллеливание метода на современные параллельные аппаратные средства, такие как суперЭВМ с графическими процессорами и др.;
3) разработка подхода к ускорению гибридного метода обучения с учителем и самообучения (Semi-supervised FDBSCAN);
4) исследование влияния параметров метода на скорость и точность выполнения;
5) исследование и тестирование алгоритма в многомерном случае.

Литература

1. *Ester M., Kriegel H.P., Sander J., XiaoweiXu* A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise // Published in Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96).
2. *Karypis G.* Chameleon data set available from http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download,2008.
3. *Rand W.M.* Objective Criteria for the Evaluation of Clustering Methods // Journal of the American Statistical Association Volume 66, Issue 336, 1971.
4. *Пантюхин Д.В., Нгуен Данг Тао, Ву Вьет Тханг* Применение нейронной сети типа многослойный персептрон для распознавания типа атаки на информационную систему на примере базы KDD'99 // XI Всероссийская научная конференция «Нейрокомпьютеры и их применение», 17 Марта 2015, МГППУ.
5. *Creech G.* Developing a high-accuracy cross platform Host-Based Intrusion Detection System capable of reliably detecting zero-day attacks, 2014.

References

1. *Ester M., Kriegel H.P., Sander J., XiaoweiXu* A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Published in Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96).
2. *Karypis G.* Chameleon data set available from http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download,2008.
3. *William M. Rand.* Objective Criteria for the Evaluation of Clustering Methods. Journal of the American Statistical Association Volume 66, Issue 336, 1971.
4. *Pantiukhin D.V., Vu Viet Thang, Nguyen Dang Tao.* Application of neural network type multilayer perceptron to recognize the type of attack on information system on the example database KDD'99. XI Russian scientific conference Neurocomputers and their application, 17.3.2015, MGPPU.
5. *Creech G.* Developing a high-accuracy cross platform Host-Based Intrusion Detection System capable of reliably detecting zero-day attacks, 2014.

Поступила в редакцию 10.09.2015.