КЛАСТЕРИЗАЦИЯ ПОКАЗАТЕЛЕЙ ЭНЕРГЕТИЧЕСКОЙ АКТИВНОСТИ ЗВЕЗД НА БАЗЕ АЛГОРИТМОВ ГУСТАФСОНА-КЕССЕЛЯ И DBSCAN С ПОМОЩЬЮ СУПЕРКОМПЬЮТЕРНОГО КЛАСТЕРА

С.В. Аксёнов, Д.Н. Лайком

Национальный исследовательский Томский политехнический университет
E-mail: axoenowsw@tpu.ru, wedun@tpu.ru

Статья посвящена решению задачи кластеризации звезд на основе их активности и содержит описание используемых методов кластеризации, описание реализации и вывод по проделанному эксперименту. Показано, что решение задачи кластеризации требует высоких вычислительных затрат, и предложено решение.

Введение

Анализ колоссальных объёмов астрофизической информации о физических параметрах и энергии объектов во внеземном пространстве предполагает нахождение групп объектов со схожими характеристиками. Для анализируемой предметной области сложность решения задач кластеризации возрастает многократно как из-за большой размерности анализируемых факторов, так и из-за огромного пространства поиска. Количество объектов в фрагменте базы данных OGLE Европейского космического агентства (ESA), предоставленном Новым университетом г. Лиссабона, [1] превышает 100000 строк. Решение проблемы разделения на кластеры позволяет выявлять скрытые зависимости между анализируемыми особенностями (различными физическими характеристиками), недоступные для восприятия из-за большой размерности задачи. В работе представлен алгоритм распределения вычислений для этой задачи, позволяющий значительно ускорить обработку астрофизических данных.

Кластеризация

Одной из особенностей задачи кластеризации является требование, накладываемое на вид кластера: группы могут принимать любую форму. Подобное ограничение исключает возможность использовать без дополнительных техник алгоритмы как четкого (K-средних [2], самоорганизующиеся карты Кохонена [3]), так и нечеткого (Fuzzy K-means, алгоритм Густафсона-Кесселя [4]) разбиения. Таким образом, возможное решение проблемы – формирование кластера путем предварительного анализа объектов методом Густафсона-Кесселя и включения близких из них в кластер методом DBSCAN (*Density Based Spatial Clustering of Application with Noise*) [5].

Используемые методы

В основе работы метода Густафсона-Кесселя лежит понятие близости объектов к центру предполагаемого кластера и предварительной сортировки объектов относительно этого центра. Формирование разрастающихся кластеров происходит на основе понятия плотности объектов. Для проведения сортировки объектов необходимо задать начальные условия: центры кластеров – и сформировать матрицу нечеткого разбиения, а также точность и экспоненциальный вес. Предварительная сортировка объектов происходит в нескольких потоках. Отсортированные точки обрабатываются на узлах су-

перкомпьютерного кластера. Формирование конечных кластеров происходит методом DBSCAN. Согласно этому методу для формирования кластера необходимо превысить порог плотности minPts. Предварительная сортировка позволяет выделить группы объектов, которые впоследствии образуют кластеры. Необходимость использования именно этого метода, основанного на некоторой заданной плотности объектов, обусловлена неравномерным распределением объектов и необходимостью поиска кластеров сложной формы. На рис. 1 представлен вариант формирования двух кластеров методом DBSCAN. При объединении оба этих кластера будут суммированы в один.

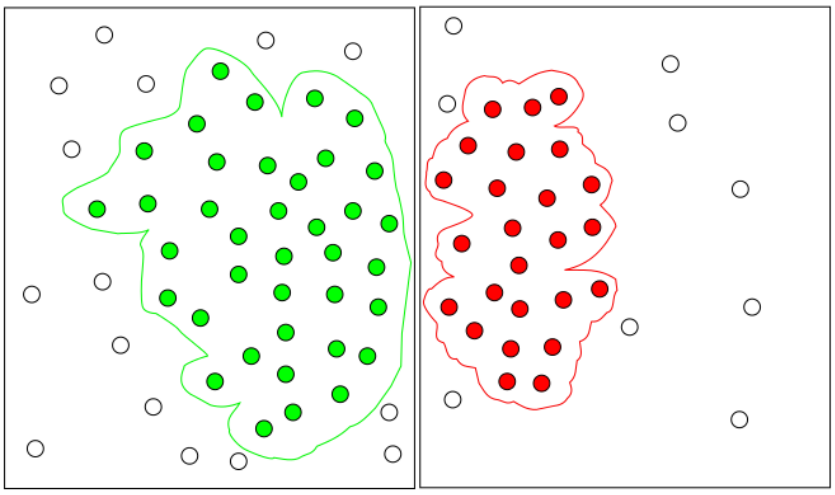

Рис. 1. Формирование кластеров методом DBSCAN на нескольких итерациях процесса кластеризации

Сам метод основан на следующих шагах. Алгоритм начинает работу с произвольной точки, которая еще не была посещена. Если ε-окрестность выбранной точки содержит достаточное количество объектов, то создается кластер. В противном случае точка обозначается как шум для ближайшего возможного кластера. Если точка оказывается частью кластера, то ее ε-окрестность также оказывается частью кластера. Таким образом все точки, найденные в ε-окрестности, добавляются вместе со своими ε-окрестностями.

Реализация

Использование технологий распределенных вычислений на суперкомпьютерном кластере предполагает распределение данных между вычислительными узлами кластера. Для решения поставленной задачи были использованы технологии MPI.NET и Microsoft.Threading.Tasks.

Процесс кластеризации можно описать следующими шагами:

- сформировать матрицу нечеткого разбиения;
- рассчитать центры кластеров;
- рассчитать матрицу ковариации;
- рассчитать расстояние между объектами и центрами кластеров;
- пересчитать элементы матрицы нечеткого разбиения;
- передать сформированные группы объектов на узлы кластера;
- проверить все объекты на возможность формирования кластеров методом DBSCAN.

Необходимость предварительной сортировки

Необходимость предварительной сортировки объясняется повышением производительности алгоритма и отсутствием необходимости поиска неучтенных кластеров.

Алгоритм нечеткой кластеризации DBSCAN показывает лучшие результаты на отсортированном наборе входных данных. Объекты, попавшие на обработку в узлы кластера, имеют большую вероятность формирования кластера с меньшим количеством шума.

Предварительно из огромного массива характеристик выбирается случайным образом по N*10000 строк, где N – количество узлов, используемых для поиска кластеров классическим методом DBSCAN, и каждый фрагмент из 10000 строк рассылается по узлам. Основная цель последующей обработки заключается в том, чтобы получить грубый набросок картины кластеров методом Густафсона-Кесселя. Отметим, что каждый узел запускает кластеризацию на своем наборе независимо от других узлов. Внутри каждого из узлов используется распараллеливание на уровне ядер, т.к. элементы матрицы нечеткого разбиения, матрицы ковариации характеристик рассчитываются независимо друг от друга и код обработки лишен рекуррентных процедур.

Получение набора независимых кластеризаций, каждая из которых выражена центрами кластеров и матрицей ковариации, ставит вопрос о нахождении наиболее подходящей для всего набора. Решение этой задачи заключается в широковещательной рассылке тестового набора 1000 векторов, выбранных случайным образом из первоначальной выборки, и определении их принадлежности к найденным кластерам внутри каждого из узлов. Для этого мы соотносим проценты объектов, попавших в каждый кластер на разных узлах. Понятно, что если большинство объектов (в наших экспериментах мы брали 80%), принадлежащих первому кластеру в первом узле, принадлежат также третьему кластеру во втором узле, это означает, что данные кластеры одинаковые. Проверяя таким образом принадлежности объектов к кластерам в разных узлах, происходит выделение похожих кластеризаций. Из набора найденных похожих кластеризаций случайным образом (т.к. их центры и матрицы ковариации приблизительно одинаковы) выбирается одна, которая признается за базовую, она рассылается на все узлы.

Теперь все узлы получают приблизительно равные фрагменты начальной выборки и кластеризуют её. Далее все точки, которые попали в первый кластер, отправляются на обработку на первый узел, которые попали во второй кластер – идут на обработку на второй узел и т.д. Таким образом, предварительная обработка позволяет разбросать объекты между узлами так, чтобы в узле располагались только близкие объекты.

Тестирование и результаты

Для проведения тестирования были использованы астрофизические данные, предоставленные Новым университетом г. Лиссабона [1]. Характеристики тестовых данных представлены в табл. 1.

Таблица 1. Характеристики наборов тестовых данных Dataset1 и Dataset2

Набор данных	Точность	Число объектов	eps	minPts	Число узлов кластера
Dataset1	0,001	10965	0,2	20	10
Dataset2	0,001	100787	0,2	20	10

Результаты тестирования подтверждают эффективность используемого метода. Предварительная сортировка входных данных позволила увеличить производительность алгоритма DBSCAN.

Выводы

В результате проделанной работы было предложено решение задачи кластеризации данных методом DBSCAN с предварительной сортировкой объектов методом Гу-

стафсона-Кесселя. Кластеризация объектов является сложной и ресурсоемкой задачей. Предложенное решение позволяет повысить производительность метода и решить проблему некорректного формирования кластеров и может быть рекомендовано к использованию при решении задачь кластеризации данных.

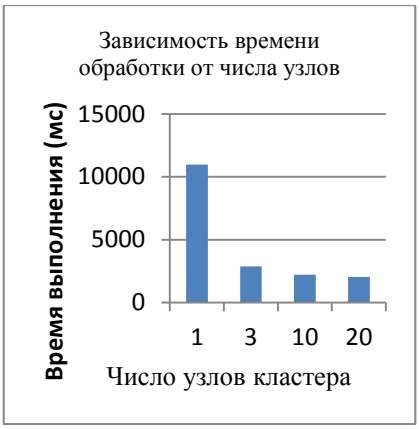

Рис. 2. Время обработки Dataset1

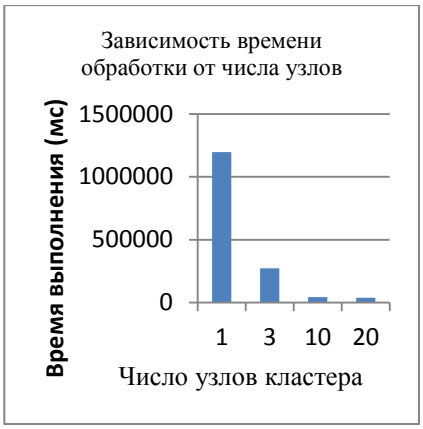

Рис. 3. Время обработки Dataset2

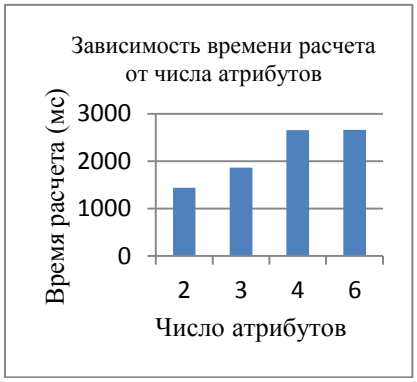

Рис. 4. Время обработки Dataset1

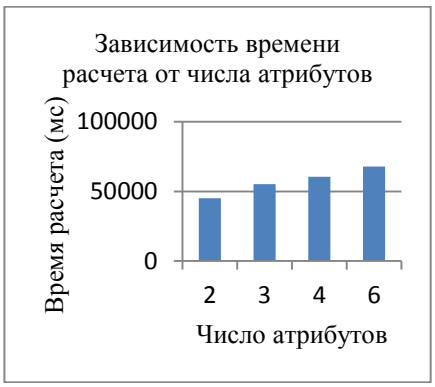

Рис. 5. Время обработки Dataset2

Литература

1. База данных характеристик звезд. URL – [http://sirius.astrouw.edu.pl/ogle]
2. MacQueen, J. B. Some Methods for classification and Analysis of Multivariate Observations // Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, 1967. P. 281–297.
3. Haykin S. Self-organizing maps. Neural networks – A comprehensive foundation (2nd ed.). Prentice-Hall, 1999.
4. Hoeppner F., Klawonn F., Kruse R., Runkler T. Fuzzy Cluster Analysis, Methods for classification, data analysis and image recognition. Wiley, 2000.
5. Sander J., Ester M., Kriegel H.-P., Xu X. Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications // Data Mining and Knowledge Discovery. Berlin: Springer-Verlag, 1998. №2 (2). P.169–194.