ОЦЕНКА ВЫЧИСЛИТЕЛЬНОЙ СЛОЖНОСТИ ОБОБЩЕННОГО АЛГОРИТМА КЛАСТЕРИЗАЦИИ

Сасин Е. А., Сидорович А. С.
Кафедра электронных вычислительных машин, Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: a.s.sidorovich@gmail.com

В работе рассмотрен обобщенный алгоритм кластеризации данных. Приведены алгоритмы последовательной и паралельной работы алгоритма. Оценена их временная сложность

Введение

Развитие методов работы с информацией привело к ее увеличению до колоссальных объемов. Для удобства работы с информацией проектируются и используются базы данных. Но в большинстве своем они таких размеров, что получение и анализ данных из них является проблемой. Часто для решения таких задач используется DataMining. Data Mining — собирательное название, используемое для обозначения совокупности методов обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

I. Кластерный анализ

Одним из методов DataMining является кластеризация(кластерный анализ). Кластеризация занимается разбиением данных на группы(кластеры) на основе схожести определенных признаков. Методы кластерного анализа позволяют решать следующие задачи:

- проведение классификации объектов с учетом признаков, отражающих сущность объектов;
- проверка выдвигаемых предположений о наличии некоторой структуры в изучаемой совокупности объектов;
- построение новых классификаций для установления наличия связей внутри совокупности;
- сжатие данных – если исходная выборка избыточно большая, можно сократить ее, оставив по одному наиболее типичному представителю от каждого кластера.

Опишем обобщенный алгоритм всех используемых методов кластеризации:

1. Случайным образом выбираются центры кластеров;
2. Вычисляются расстояния от каждого объекта до центра каждого кластера;
3. Объекты причисляются к кластерам;
4. Производится пересчет центров каждого кластера;
5. Если центры не изменились, то работа алгоритма заканчивается, иначе продолжаем работу алгоритма с вычисления расстояния.

Для вычисления расстояний между центрами кластеров используются различные метрики. Тремя наиболее распространенными метриками расстояний являются классическая Евклидова метрика и метрики Чебышева и Манхеттена.

II. Оценка сложности

Особое место при построении систем и алгоритмов является оценка их сложности. Оценка сложности помогает реализовывать более эффективное решение в каждой конкретной ситуации. Сложность алгоритма бывает временная и пространственная. Пространственная – показывает затраты памяти на реализацию алгоритма, а временная описывает количество и время выполнения элементарных опреаций алгоритма. При работе с большими объемами данных подразумевается использование достаточных физических ресурсов данных, а вот временные затраты на реализацию алгоритмов можно эффективно сокращать (без потери точности).Точные затраты времени рассчитать практически не возможно, т.к. это зависит от физических характеристик вычислительной системы: архитектуры и количества процессоров, особенностей компилятора и многих других. Для оценки временной сложности можно использовать O - нотацию, которая использует математическую функцию f(n), которая зависит от количества операций n. Использование данной нотации позволяет описать характер функции f(n) с изменением n: насколько быстро растет эта функция. Данная нотация использует принцип «худшего случая». Стандартные функции используемые в O - нотации в зависимости от степени их роста:

- остоянные функции, которые с ростом n не изменяются, O(1);
- функции с логарифмической скоростью роста O(log_2 n);
- функции с линейным ростом O(n);
- функции с линейно-логарифмической скоростью роста O(nlog_2n);
- функции с квадратичной скоростью роста O(n^2);

– функции сэкспоненциальной скоростью роста O(2n);
– функции с факториальной степенью роста O(n!).

Стоит заметить, что сложность алгоритма по данной нотации не обязательно должна принадлежать одной из этих групп.

III. Сравнение сложности выполнения алгоритма

При выполнении алгоритма кластеризации на многояпроцессорах появляется возможность сокращения временных затрат. Алгоритмы кластеризации получают в качестве входных данных n векторов, каждый размерностью m. Количество разбиений на кластеры предопределено в начале и равно k. Пусть t – количество итераций, необходимых для завершения кластеризации.

Сравнении временной сложности последовательного и параллельного выполнения алгоритмов может проводиться без учета сложности вычислений метрик расстояний и нахождения центров кластеров. Т.к. эти алгоритмы усложняют и параллельную, и последовательную реализацию одинаково.

При обработке больших объемов данных последовательно большую часть времени вычисляются расстояния между кластерами и их центры. Таким образом, последовательно делается n одинаковых операций, которые не затрагивают другие вектора. Обобщенная схема последовательной реализации алгоритма представлена на рисунке 1.

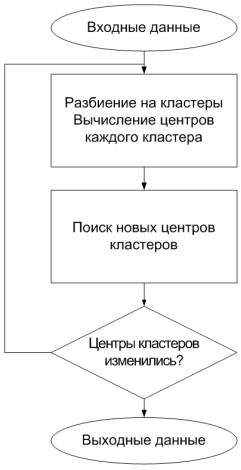

Рис. 1 – Последовательная реализация алгоритма кластеризации

Вычислительная сложность последовательного алгоритма кластеризации O(nmkt)

Для организации распараллеливания необходимо выполнить за один такт сразу n операций вычисления центров кластеров, а затем столько же для вычисления расстояний . Обобщенный алгоритм параллельной кластеризации можно представить следующим образом:

1. Генерируется k объектов-центров классов2.
2. Производится разбивка на кластеры, путем выяснения какой либо меры близости вектора с центром кластера.Для каждого вектора мера близости вычисляется отдельным процессом, параллельно с остальными
3. Для каждого кластера выполняется поиск нового центра. Т.к. в каждом кластере данные не связаны с векторами из других кластеров, то и эту опреацию можно выполнять одновременно для k разных кластеров.
4. Выполняется проверка, изменился ли центр класса.Если да, то выполняется переход к шагу 2. Если нет, то кластеризация завершена. .

Обобщенная схема параллельной реализации алгоритма представлена на рисунке 2.

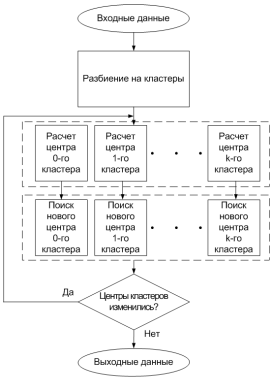

Рис. 2 – Параллельная реализация алгоритма кластеризации

Используя алгоритм нахождения частичной суммы элементов массива путем сдваивания временная сложность вышеприведенного алгоритма составит O(log(nmkt)).

Заключение

В ходе выполнения данной работы были получены следующие выводы:

– алгоритмы кластеризации могут быть распараллелены независимо от методов вычисления центров и расстояний.
– алгоритм можно выполнять параллельно, если есть минимум две операции с независимыми данными
– при увеличении количества векторов – производительность параллельных вычислений увеличится

1. Методы и модели анализа данных: OLAP и DataMining. / А. А. Барсегян [и др.]//Спб.: БХВ-Петербург/ –2004.
2. Крупский, В. Н Введение в сложность вычислений. / В. Н. Крупский//–М.: Факториал Пресс. –2006.
3. Arabie, P., Hubert, L. J. Clustering and Classification. / P. Arabie, L. J. Hubert, // Singapore: World Scientific/ –1996