МЕТОДЫ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА В СИСТЕМАХ БОЛЬШИХ ДАННЫХ

MINING TECHNIQUES
IN BIG DATA SYSTEMS

A. Borisov

Annotation

The paper discusses statement of the problem, the urgency of the research and the main features of Big Data which complicating the process of it analysis by traditional methods. The paper explains the reasons for the growing popularity of mining methods. The article gives some methods of intellectual analysis and described algorithms of these methods. For each method gives the features that allow to achieve effective scaling algorithms and their efficiency at working with Big Data.

Keywords: Big Data, An intellectual analysis, Multivariate analysis, Regression, Classification, Association.

Борисов Александр Васильевич
Московский Государственный Технический Университет им. Н.Э. Баумана

Аннотация

Осуществляется постановка проблемы и обосновывается актуальность исследования, приводятся основные особенности Больших Данных, затрудняющие процесс их анализа традиционными методами. Объясняются причины роста популярности методов интеллектуального анализа. Рассматриваются некоторые методы интеллектуального анализа, описываются алгоритмы данных методов. Для каждого приведенного метода рассматриваются те особенности, которые позволяют добиться эффективного масштабирования алгоритмов и их эффективность при работе с Большими Данными.

Ключевые слова:

Большие Данные, Big Data, Интеллектуальный анализ, Многомерный анализ, Регрессия, Классификация, Ассоциация.

Введение

В эпоху стремительного развития информационных технологий постоянно растет скорость накопления информации. Исследование, проведенное в 2012 г., оценило объем сгенерированных данных в 2,8 зеттабайта и прогнозирует к 2020 г. увеличение объема до 40 зеттабайт, что превосходит прежние прогнозы на 14% [2]. Рост объемов информации и появление больших данных имеет место в самых различных областях, будь то научная деятельность, маркетинговые исследования или же анализ рисков при принятии решений. Всё большее число компаний, начинаю применять Большие Данные (Big Data) в своей деятельности. Согласно исследованиям агентства International Data Corporation (IDC) объем рынка решения Big Data достигнет $41 млрд к 2018 году.

Однако Большие Данные обладают рядом особенностей, которые делают крайне затруднительным процесс обработки их традиционными способами. К этим особенностям можно отнести не только огромный объем информации, но и её слабую структурированность, высокую изменчивость и слабую взаимосвязь данных. При этом постоянно увеличиваются требования к скорости анализа данных, что делает актуальным исследования в области методов и алгоритмов интеллектуального анализа данных.

Определение процесса интеллектуального анализа данных

По сути, интеллектуальный анализ данных – это обработка информации и выявление в ней моделей и тенденций, которые помогают принимать решения [5]. Интеллектуальный анализ применялся еще до появления технологий Больших Данных, но наибольшее развитие он получил именно в связи с технологиями Big Data.

Одной из основных причин роста популярности многих методов интеллектуального анализа, является увеличение объема информации и её растущее разнообразие. При анализе Больших Данных становится абсолютно недостаточным использование простой статистики, зачастую требуется такой анализ данных, который позволит выявить скрытые закономерности в данных, построить информационную модель и получить прогноз по полученной модели.

Основные методы интеллектуального анализа данных

Рассмотрим некоторые методы интеллектуального анализа и проанализируем возможность их применения в области Больших Данных.

Поиск ассоциативных правил

Ассоциация является наиболее распространенным методом интеллектуального анализа [5]. Построение моделей заключается в нахождении правил, которые описываю взаимозависимости между элементами данных, при этом правила имеют два параметра: вероятность срабатывания и поддержка. Вероятность срабатывания определяет, насколько часто происходит выполнение правила. Поддержка показывает, как часто применимо данное правило, то есть, как часто встречается сочетание определенных признаков. Целью метода является поиск правил с высокой вероятностью срабатывания и высокой поддержкой.

Существуют различные методы поиска ассоциативных правил, но не все из них могут быть применены для

Больших Данных, в связи с недостаточно эффективной масштабируемостью. Рассмотрим некоторые алгоритмы, которые вполне успешно применяются в Big Data и те методы повышения эффективности масштабирования, благодаря которым это возможно.

Apriori – масштабируемый алгоритм поиска ассоциативных правил. Данный алгоритм работает в два этапа. На первом шаге находятся наиболее часто встречающиеся группы элементов, на втором происходит извлечение из них правил. Для сокращения размерности пространства поиска на первом этапе используется свойство анти–монотонности, гласящее: любой набор элементов не может иметь поддержку больше минимальной поддержки его подмножеств. Из данного свойства можно сделать вывод, что любой набор из n элементов будет часто встречающимся лишь тогда, когда все его (n–1) подмножества являются часто встречающимися [4]. Данное свойство позволяет отбросить все вышестоящие множества, если нижестоящее множество встречается редко.

Алгоритм Frequent Pattern–Growth Strategy (FPG). Основой данного алгоритма является предобработка базы данных, в результате которых она преобразуется в компактное дерево популярных предметных наборов [1]. Данный алгоритм позволяет произвести декомпозицию сложной задачи на несколько простых и избежать процедуры генерации кандидатов.

Многомерный анализ

Данный метод основывается на построении многомерных кубов и получении их различных срезов. Результатом является таблица, содержащая агрегированные показатели. Системы многомерного анализа делятся на MOLAP, ROLAP и HOLAP. ROLAP является простой реляционной базой, данные хранятся в плоских таблицах, агрегаты сохраняются в дополнительных реляционных таблицах. Данная технология отличается высокой масштабируемостью, но имеет низкое быстродействие. В технологии MOLAP детальные и агрегированные данные содержатся в многомерной базе в виде явного, физически хранимого многомерного куба, с выполнением аналитических запросов только над ними. Данная технология позволяет добиться большей скорости анализа, но при этом генерируются огромные объемы данных. HOLAP использует реляционные таблицы для хранения базовых данных и многомерные таблицы для агрегатов, обеспечивая средние значения масштабирования и быстродействия, и является оптимальным выбором для больших данных [3].

Регрессия

Метод регрессии основан на построении параметрической функции, которая описывает изменение некоторого числового значения в определенный временной промежуток. На основе имеющихся данных, по полученной функции, прогнозируются дальнейшие значения этой величины. Для этого рассчитывается суммарная разница между наблюдаемыми значениями и значениями, выдаваемыми функцией при текущих параметрах. Следующим шагом подбираются новые параметры (веса), которые позволяют уменьшить текущую разницу. Операции повторяются, пока разница не уменьшится до приемлемого значения.

Метод регрессии вполне применим для Больших Данных, так как он сводится к операциям по вычислению взвешенных сумм, которые достаточно легко распараллеливаются, и могут выполняться на нескольких серверах.

Классификация

Метод классификации основан на построения зависимости одной переменной от нескольких других. В отличие от регрессии входные значения не упорядочены по периоду. Несмотря на большое количество существующих методов классификации, они работают по одному принципу. На первом этапе производится обучение алгоритма на небольшой выборке, на втором – применение полученных правил к остальным данным. Первый этап как правило проходит без распараллеливания работы. На втором же этапе данные можно обрабатывать независимо, так как правила, полученные на первом этапе, можно копировать на все сервера и использовать для анализа находящихся там данных. Можно сделать вывод, что методы классификации подходят для работы с Большими Данными.

Выводы

Большие Данные являются относительно новым и быстро развивающимся направлением развития информационных технологий. Как и любая новая технология, она обладает своей спецификой, которая приводит к необходимости изменения привычных методов анализа данных, для адаптации их к новым условиям. Для эффективной работы с Большими Данными необходимо активно использовать распараллеливание процессов, применяемых в интеллектуальном анализе, задействовать большее количество аппаратных ресурсов и улучшать возможности масштабирования применяемых алгоритмов.

ЛИТЕРАТУРА

1. Орешков В., FPG – альтернативный алгоритм поиска ассоциативных правил – URL: https://basegroup.ru/community/ articles/fpg (Дата обращения: 25.10.15).
2. Рост объема информации – реалии цифровой вселенной // Технологии и средства связи – 2013. – № 1. – С. 24.
3. Селезнев К. Проблемы анализа Больших Данных // Открытые системы – 2012. – № 7. – С. 25 –30.
4. Шахиди А., Apriori – масштабируемый алгоритм поиска ассоциативных правил – URL: https://basegroup.ru/community/ articles /apriori (Дата обращения: 20.10.15).
5. Data mining techniques – URL: http://www.ibm.com/developerworks /opensource/library/ba–data–mining–techniques/index.html?S_TACT= 105AGX99&S_CMP=CP (Дата обращения: 17.10.15).

© А.В. Борисов, (brsff@mail.ru), Журнал «Современная наука: актуальные проблемы теории и практики»,