СРАВНИТЕЛЬНОЕ ИССЛЕДОВАНИЕ АЛГОРИТМОВ КЛАССИФИКАЦИИ БОЛЬШИХ ОБЪЕМОВ ДАННЫХ

Н.В. Ситникова [1], Р.А. Парингер [1,2], А.В. Куприянов [1,2]

[1] Самарский государственный аэрокосмический университет имени академика С.П. Королёва (национальный исследовательский университет) (СГАУ), Самара, Россия
[2] Институт систем обработки изображений РАН, Самара, Россия

Приводится краткий обзор двух методов классификации: метод ближайшего соседа и случайный лес. Оба метода реализованы с помощью технологии MapReduce, с целью применения к большим объемам данных. Приводится сравнительный анализ результатов, полученных на тестовых данных Heterogeneity Activity Recognition Data Set из репозитория UCI.

Ключевые слова: классификация, большие данные, параллельные вычисления, MapReduce, BigData.

Введение

Одним из наиболее крупных разделов, изучающих искусственный интеллект, является машинное обучение. Данный раздел решает задачи , для которых зачастую невозможно придумать явный алгоритм решения: распознавание речи, техническая диагностика, прогнозирование временных рядов, обнаружение спама, биржевой технический анализ, кредитный скоринг.

На данный момент известно достаточно много методов классификации: метод ближайшего соседа [1], метод стохастического градиента [2], машина опорных векторов [3], деревья решений [4], нейронные сети и др.

Решаемые в настоящее время практические задачи машинного обучения требуют обработки больших объёмов входных данных. Это связано и с количеством входных объектов и с тем, что каждый объект может быть описан вектором признаков, содержащим сотни или даже тысячи переменных. В связи с этим, на одно из первых мест встаёт вопрос производительности выбранного алгоритма. В данной статье рассматриваются и сравниваются два метода, реализованные с помощью парадигмы MapReduce.

Метод ближайших соседей

Метод ближайших соседей – один из простейших метрических классификаторов, основанный на оценивании сходства объектов. Объект относится к тому классу, которому принадлежат ближайшие к нему объекты из обучающей выборки. Метод ближайших соседей имеет несколько вариаций:

- Метод ближайшего соседа. Классифицируемый объект x относится к тому классу y_i, которому принадлежит первый ближайший объект обучающей выборки x_i.
- Метод k ближайших соседей. С целью повышения надёжности классификации объект относится к тому классу, которому принадлежит большинство из его соседей — k ближайших к нему объектов обучающей выборки x_i. В задачах, где число классов рано двум, число соседей берут нечётным, чтобы не возникало ситуа-

ций неоднозначности, когда одинаковое число соседей принадлежат разным классам.

- Метод взвешенных ближайших соседей. В задачах с числом классов 3 и более нечётность уже не помогает, и ситуации неоднозначности всё равно могут возникать. Тогда i-ому соседу приписывается вес w_i, как правило, убывающий с ростом ранга соседа i. Объект относится к тому классу, который набирает больший суммарный вес среди k ближайших соседей.

При последовательной реализации данного метода возникает проблема при сверхбольших выборках, так как каждый объект обучающей выборки нужно сравнивать с классифицируемым объектом. Однако метод ближайших соседей хорошо распараллеливается ввиду того, что расстояние между классифицируемым объектом и объектами из обучающей выборки считаются независимо друг от друга.

В рамках концепции MapReduce подсчёт расстояний между классифицируемым объектом и объектом из обучающей выборки может проходить на шаге map, а на шаге reduce выбор класса.

Случайный лес

В работе [4] был предложен алгоритм машинного обучения, в котором строился лес решающих деревьев. Данный алгоритм базируется на двух основных идеях: метод бэггинга Бреймана, и метода случайных подпространств, предложенный Tin Kam Ho [5]. Алгоритм находит свое применение в задачах классификации, регрессии и кластеризации.

Пусть обучающая выборка состоит из N прецедентов, M - размерность пространства признаков, и задан параметр m (в задачах классификации обычно $m \approx \sqrt{M}$).

Каждое дерево леса строится независимо друг от друга по следующей процедуре:

Из обучающей выборки сгенерируем случайную подвыборку с повторением размером N

Таким образом, получим, что некоторые прецеденты попадут в неё несколько раз, а в среднем $N \cdot \left(1+\frac{1}{N}\right)^N$, т.е. примерно $\frac{N}{e}$ прецедентов не войдут в неё вообще.

Построим решающее дерево, которое классифицирует прецеденты данной подвыборки, причём в ходе создания нового узла дерева будем выбирать признак, на основе которого производится разбиение, не из всех M признаков, а только из m случайно выбранных. Наилучший из этих m признаков может выбираться различными способами. В оригинальной статье Бреймана используется критерий Джини, который также применяется в алгоритме CART построения решающих деревьев. Существуют реализации алгоритма, где вместо него используется критерий прироста информации.

Дерево строится до тех пор, пока подвыборка не станет пустой, при этом дерево не подвергается процедуре прунинга (в отличие от решающих деревьев, которые построены по таким алгоритмам, как CART или C4.5).

Путём голосования проводится классификация объектов: каждое дерево леса относит классифицируемый объект к одному из классов, и побеждает класс, за который проголосовало наибольшее число деревьев.

Оптимальное число деревьев будем подбирать таким образом, чтобы число ошибок классификатора на тестовой выборке было минимальным. В случае отсутствия тестовой выборки, будем минимизировать оценку ошибки out-of-bag: доля примеров обучающей вы-

борки, классифицируемых лесом неправильно, если не учитывать голоса деревьев на примерах, входящих в их собственную обучающую подвыборку.
Так как каждое случайное дерево решений из комитета считается независимо от других, то весь комитет можно посчитать параллельно, а именно в маппере, а классификацию проводить на шаге reduce.

Сравнительный анализ

Автором были реализованы методы, описанные в предыдущих пунктах, с использованием технологии MapReduce. Набор данных Heterogeneity Activity Recognition Data Set для тестирования был взят из репозитория UCI. Данные содержат более четырех миллионов записей, пять процентов из которых использовались как тестовая выборка.
Тестирование проводилось на пяти компьютерах, объединенных в один кластер.
Из графика на рисунке 1 видно, что метод k ближайших соседей работает быстрее, что объясняется тем, что метод случайного леса тратит время на построение деревьев. С увеличением количества узлов, которые обрабатывают данные, время обработки сначала падает линейно, но затем видно, что график стремится к некоторой горизонтальной асимптоте. Это связано с тем, что на каждом узле кластера хранятся части выборок объемом 64 мегабайта. Поэтому при увеличении количество узлов в кластере, время работы уменьшаться не будет. Таким образом, максимальное количество используемых узлов можно вычислить по формуле $N_{used} = \left\lceil \frac{D}{64} \right\rceil$, где D – объем данных в мегабайтах.

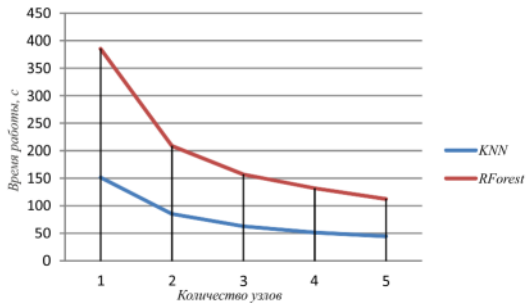

Рис. 1. График зависимости времени от количества узлов

На рисунке 2 приведена гистограмма ошибок, из гистограммы видно, что количество узлов никак не влияет на количество ложных срабатываний классификатора, реализующего метод ближайшего соседа, что нельзя сказать про другой классификатор. При увеличении количества узлов, увеличивается и количество сгенерированных случайных деревьев принятия решений.

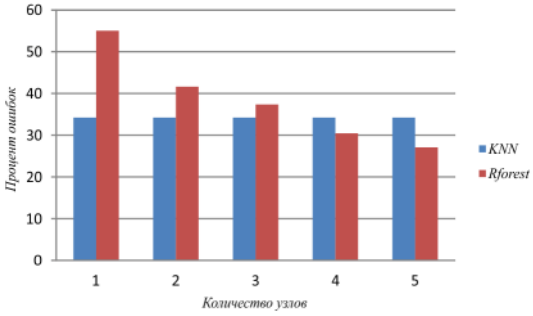

Рис. 2. Гистограмма ошибок

Заключение

В заключении хочется отметить, что большая часть методов классификации не ориентирована на параллельные вычисления с распределенной памятью, именно поэтому автором было рассмотрено только два метода.

Литература

1. Метод потенциальных функций в теории обучения машин / М. А. Айзерман, Э. М. Браверман, Л. И. Розоноэр — М.: Наука, 1970. — 320с.
2. Bottou, L. Stochastic Learning / Leon Bottou // Advanced Lectures on Machine Learning. – 2004. – P. 146-168.
3. Вапник, В. Н. Восстановление зависимостей по эмпирическим данным — М.: Наука, 1979. — 448 с.
4. Classification and Regression Trees / L. Breiman, J. Friedman, R. Olshen, C. Stone – Chapman and Hall/CRC, 1983. – P. 368
5. Ho, T. K. Random decision forests / Tin Kam Ho // Proceedings of the Third International Conference on, Montreal, Que. - 1995. - Vol.1. - P. 278-282.