Отзыв

официального оппонента к. ф.-м.н. Бурнаева Евгения Владимировича

о диссертационной работе Трофимова Ильи Егоровича

«Разработка и обоснование методов параллельного покоординатного спуска для обучения обобщенных линейных моделей с регуляризацией», представленной на соискание ученой степени кандидата физико-математических наук по специальности 05.13.17 - «теоретические основы информатики»

Актуальность темы

Обобщенные линейные модели с регуляризацией применяются для решения большого числа задач машинного обучения и анализа данных. Во многих практических задачах возникают большие обучающие выборки. В качестве примера можно привести задачи поиска в интернете, онлайн рекламы, обработки текстов, анализа показателей датчиков, генетики и т.д. Такие задачи характеризуются большим числом обучающих примеров, высокой размерностью, или и тем и другим одновременно. Желательным свойством является разреженность полученного решения. Если в этих задачах использовать для обучения только часть имеющихся данных, то качество прогноза, как правило, падает. Поэтому важным направлением исследований является разработка методов машинного обучения, специально предназначенных для больших выборок, а также разработка алгоритмов, позволяющих обучать обобщенные линейные модели с регуляризацией на больших выборках.

Содержание

Диссертационная работа состоит из введения, 5 глав, заключения, приложения и списка литературы.

Во введении обосновывается актуальность выбранной темы исследования - обучение на больших выборках (big data) обобщенных линейных моделей с регуляризацией, с использованием вычислительного кластера. В главе 1 вводятся основные понятия и определения, которые будут использоваться в работе. В главе 2 описываются наиболее популярные архитектуры вычислительных систем для распределенного машинного обучения (Map/Reduce, MPI, сервера параметров, Spark, GraphLab).

В главе 3 описывается предложенный автором метод параллельного покоординатного спуска - d-GLMNET. В методе d-GLMNET выполняются параллельные шаги по блокам переменных. Доказывается, что независимая оптимизация по блокам переменных эквивалентна оптимизации квадратичного приближения к целевой функции с блочно-диагональным приближением Гессиана. Обосновывается модификация минимизируемого локального квадратичного приближения к целевой функции. Получены достаточные условия сходимости и линейной скорости сходимости метода d-GLMNET. В разделах 3.5-3.6 описывается программная реализация метода d-GLMNET. Программа запускается на кластере Map/Reduce, дополнительно используя стандарт MPI для передачи данных между узлами кластера. В разделе 3.7 описываются модификация метода d-GLMNET - ``асинхронная балансировка нагрузки'' (ALB) для ускорения в случае, если узлы кластера имеют различающуюся производительность.

В главе 4 описываются численные эксперименты, проведенные для сравнения предложенного метода d-GLMNET с общепринятыми методами для обучения обобщенных линейных моделей с регуляризацией на вычислительном кластере - параллельное онлайн

обучение, L-BFGS, ADMM. Численные эксперименты доказывают превосходство разработанного метода, особенно для наборов данных с большим количеством переменных. Доказывается эффективность метода асинхронной балансировки нагрузки.

В главе 5 описывается применение методов обучения обобщенных линейных моделей с регуляризацией на больших выборках для задачи прогнозирования вероятности клика в онлайн рекламе.

Подробный вывод шага метода d-GLMNET приведен в Приложении 1.

Основные результаты и их новизна

В своей диссертационной работе Трофимовым И.Е. были получены следующие новые результаты:

1. Разработан метод d-GLMNET, предназначенный для параллельного покоординатного спуска для минимизации функций риска обобщенных линейных моделей с регуляризацией "elastic net";
2. Получены достаточные результаты сходимости и линейной скорости сходимости метода d-GLMNET;
3. Доказана возможность получать разреженные решения с помощью метода d-GLMNET при использовании L1-регуляризации;
4. Предложен метод ``асинхронной балансировки нагрузки'' для обеспечения эффективного выполнения метода d-GLMNET при наличии медленных узлов кластера;
5. Проведены численные эксперименты, доказывающие, что метод d-GLMNET более эффективен, чем общепринятые методы при работе с разреженными обучающими выборками с высокой размерностью признакового пространства.
6. Разработана общедоступная программная реализация метода d-GLMNET: https://github.com/IlyaTrofimov/dlr.

Достоверность результатов

Достоверность результатов обеспечивается доказательствами теорем и описаниями проведенных экспериментов, допускающими их воспроизводимость. Исходный код программ и выборки, использовавшиеся для численных экспериментов общедоступны.

Значимость результатов

Теоретическая значимость состоит в установлении достаточных условий сходимости и линейной скорости сходимости разработанного метода d-GLMNET, в том числе, для модификации метода d-GLMNET, использующей технику асинхронной балансировки нагрузки.

Практическая значимость определяется тем, что разработанный Трофимовым И.Е. метод d-GLMNET позволяет проводить обучение обобщенных линейных моделей быстрее, чем при использовании общепринятых методов, что позволяет экономить вычислительные ресурсы и получать более точные решения при ограниченном бюджете вычислительных ресурсов.

Замечания

1. В тексте диссертации используется термин "распределенный покоординатный спуск" (см. Алгоритм 15) без предварительного определения.
2. В описании входных данных алгоритма на стр. 24 имеется опечатка, затрудняющая понимание алгоритма.
3. В разделе 2.1.6 сказано, что наиболее масштабируемой и производительной системой для распределенного машинного обучения является архитектура "сервера параметров". В тоже время, в диссертации предлагается реализации метода d-

GLMNET с использованием стандарта MPI (раздел 3.6). Не ясно, почему был сделан именно такой выбор и можно ли реализовать метод d-GLMNET для системы “сервера параметров”.

Заключение. Целью диссертационной работы Трофимова И.Е. является разработка и обоснование методов параллельного покоординатного спуска для обучения обобщенных линейных моделей с регуляризацией. Разработанные методы применимы для больших обучающих выборок, обладающих высокой размерностью признакового пространства. Для проведения численных экспериментов Трофимовым И.Е. был разработан комплекс программ с открытым исходным кодом. Трофимовым И.Е. проделан большой объем работы – с одной стороны, теоретическое исследование и обоснование разработанного им метода параллельного покоординатного спуска; с другой стороны, выполнение численных экспериментов с разработанным методом.

Диссертационная работа Трофимова И.Е. «Разработка и обоснование методов параллельного покоординатного спуска для обучения обобщенных линейных моделей с регуляризацией» является законченной самостоятельной научно-исследовательской работой, содержащей новые научные результаты.

Работа удовлетворяет всем требованиям, предъявляемым ВАК к диссертациям, представленным на соискание ученой степени кандидата физико-математических наук по специальности 05.13.17 - «теоретические основы информатики». Диссертант Трофимов И.Е. заслуживает присвоения ему ученой степени кандидата физико-математических наук.

Официальный оппонент,
кандидат физико-математических наук,
доцент “Центра по научным и инженерным
вычислительным технологиям для задач
с большими массивами данных”
АНОО ВО «Сколковский институт науки и технологий»
Адрес: г. Москва, Территория Инновационного Центра “Сколково”,
ул. Нобеля, д.3
Телефон: +7-926-562-33-55
E-mail: e.burnaev@skoltech.ru

Бурнаев Евгений Владимирович

дата 21.01.2019

Подпись Бурнаева Е.В.
подтверждаю

РУКОВОДИТЕЛЬ ОТДЕЛА
КАДРОВОГО АДМИНИСТРИРОВАНИЯ
БУРДЕНКО Н.Г.

Skoltech