

УТВЕРЖДАЮ

Директор

ФГБУН «Институт проблем управления
им. В.А. Трапезникова» РАН, член-корр. РАН



Новиков Д.А.

«29» января 2019 г.

ОТЗЫВ

ведущей организации - Федерального государственного бюджетного учреждения науки
«Институт проблем управления им. В.А. Трапезникова» Российской академии наук
на диссертационную работу Трофимова Ильи Егоровича
«Разработка и обоснование методов параллельного покоординатного спуска для обучения
обобщенных линейных моделей с регуляризацией», представленную на соискание ученой
степени кандидата физико-математических наук по специальности
05.13.17 - «теоретические основы информатики»

Актуальность темы. Во многих практических задачах прогнозирования и анализа данных необходимо работать с большими обучающими выборками. Например, такие выборки возникают в задачах обработки текстов, анализа данных телеметрии, поиска в интернете, онлайн рекламе и т.д. Эти 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. Одним из часто используемых для решения этих задач методов является обучение обобщенных линейных моделей с регуляризацией. Поэтому важным направлением исследований является разработка методов численной оптимизации, позволяющих обучать обобщенные линейные модели с регуляризацией на больших выборках.

Содержание. Диссертация 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.

Основные теоретические результаты приведены 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 Гесса. 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.

Основные результаты диссертации. 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.

Достоверность результатов обеспечивается доказательствами теорем и описаниями проведенных экспериментов, допускающими их воспроизводимость. Исходный код программ и выборки, использовавшиеся для численных экспериментов общедоступны. Научные положения и выводы, сформулированные в диссертации, подкреплены реальными данными, наглядно представленными в виде таблиц и рисунков.

Значимость полученных результатов для науки и производства. Научная 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.

Рекомендации по использованию результатов и выводов диссертации. Результаты диссертации могут быть использованы для решения широкого класса задач машинного обучения и анализа данных. Разработанный метод хорошо масштабируется и применим к большим обучающим выборкам, может эффективно выполняться на вычислительном кластере и в

многопроцессорных системах с общей памятью. Поэтому метод d-GLMNET может применяться для решения задач регрессии и классификации, использующих данные онлайн приложений, телекоммуникации, больших массивов текстов, показателей телеметрии и т.п.

Замечания

1. В обзорной части работы (раздел 1) не упомянуты статьи Ю. Е. Нестерова и некоторые другие современные работы по методам покоординатного спуска.
2. В разработанном алгоритме оптимизации "d-GLMNET" используется линейный поиск (Алгоритм 12, стр. 49) с условием останова (п. 5). Данное условие в работе называется "условие Армихо", хотя, строго говоря, не является им, т.к. минимизируемая функция дополнительно содержит негладкое слагаемое (L1-регуляризатор).
3. В диссертации имеются опечатки (например, в алгоритме 5, алгоритме 13).

Отмеченные недостатки не оказывают влияния на общую положительную характеристику данной работы.

Закключение. Диссертационная работа Трофимова И.Е. «Разработка и обоснование методов параллельного покоординатного спуска для обучения обобщенных линейных моделей с регуляризацией» является законченной самостоятельной научно-квалификационной работой, содержащей решение научной задачи, имеющей важное значения.

Данную работу можно квалифицировать как вклад в научное направление, связанное с разработкой и исследованием моделей и алгоритмов анализа данных, обнаружением закономерностей в данных, что согласуется с п. 5 паспорта специальности 05.13.17 - «теоретические основы информатики».

Автореферат диссертации адекватно отражает ее содержание.

Работа удовлетворяет всем требованиям, предъявляемым ВАК к диссертациям, представленным на соискание ученой степени кандидата физико-математических наук, а ее автор, Трофимов И.Е. заслуживает присвоения ему ученой степени кандидата физико-математических наук.

Диссертация Трофимова И.Е. была заслушана на научном семинаре «Теория автоматического управления» 23.01.2019, протокол №1, и получила положительную оценку.

Заведующий лабораторией №7

ФГБУН Института проблем управления им. В.А. Трапезникова
Российской академии наук
доктор физико-математических наук , профессор РАН



Хлебников М.В.

Главный научный сотрудник

ФГБУН Института проблем управления им. В.А. Трапезникова
Российской академии наук
доктор технических наук



Поляк Б.Т.

117997, г. Москва, ул. Профсоюзная, 65

Телефон: +7-495-334-89-10

E-mail: dan@ipu.ru