

Отзыв официального оппонента

д. ф.-м. н., член-корр. РАН Аветисяна Арутюна Ишхановича

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. Во многих практических задачах машинного обучения и анализа данных возникают большие обучающие выборки. В качестве примера можно привести задачи поиска в интернете, онлайн рекламы, обработки текстов, анализа показателей датчиков и т.д. Такие задачи характеризуются большим числом обучающих примеров, высокой размерностью, или и тем и другим одновременно. 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. Кроме того, учитывая увеличивающийся с каждым годом объем доступных для анализа данных, желаемым свойством методов обучения является возможность эффективного выполнения алгоритма в многопроцессорных и распределенных системах. Один из часто применяемых на практике методов машинного обучения является обобщенная линейная модель с регуляризацией.

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 (Map/Reduce, MPI, сервера параметров, Spark, GraphLab).

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 Гессияна. 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. Исследуется ускорение сходимости в зависимости от числа узлов кластера.

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 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. Печать диссертации выполнена в черно-белой палитре, что делает тяжело читаемыми некоторые рисунки (рис. 1 на стр. 11, рис. 8 на стр. 72, и т.д.).
2. В программной реализации разработанного метода (разделы 3.5, 3.6) используется способ хранения разреженных матриц "список списков" (List of lists. LIL). Другие способы хранения разреженных матриц могут быть более вычислительно эффективны, данный аспект не исследован в работе.
3. Как показывают экспериментальные результаты в разделе 4.5, ускорение метода d-GLMNET сублинейное от числа узлов. Представляет интерес анализ причин этого явления – является ли это следствием увеличения времени на передачу данных или более частого конфликта между шагами по переменным.

Указанные замечания не носят принципиального характера и не снижают общей положительной характеристики диссертационной работы.

Закключение. В своей диссертационной работе Трофимов И.Е. разработал и выполнил теоретическое обоснование метода параллельного покоординатного спуска для обучения обобщенных линейных моделей с регуляризацией. Данный метод применим для больших обучающих выборок, обладающих высокой размерностью признакового пространства. Для проведения численных экспериментов Трофимовым И.Е. был разработан комплекс программ с открытым исходным кодом. Кроме того, были выполнены численные эксперименты, доказывающие эффективность разработанного метода.

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 научные результаты, обладающие новизной и практической ценностью.

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», а ее автор, Трофимов И.Е. заслуживает присуждения ему ученой степени кандидата физико-математических наук.

Диссертация была заслушана на научном семинаре «Системное программирование» под руководством А.И. Аветисяна 18.01.2019 и получила положительную оценку.

Директор Федерального государственного бюджетного учреждения науки
Институт системного программирования им. В.П. Иванникова Российской академии наук,
доктор физико-математических наук, член-корр. РАН

Адрес: 109004 г. Москва, ул. Александра Солженицына, д. 25

Телефон: +7-495-912-44-25

E-mail: arut@ispras.ru

дата 24.01.2019



Аветисян А.И.