УТВЕРЖДАЮ
Директор ФИЦ ИУ РАН,
академик РАН
_____Соколов И.А.
«9» марта 2018 г.

ЗАКЛЮЧЕНИЕ

Федерального государственного учреждения «Федеральный исследовательский центр «Информатика и управление» Российской академии наук»

Диссертация «Разработка и обоснование методов параллельного покоординатного спуска для обучения обобщенных линейных моделей с регуляризацией» выполнена в отделе интеллектуальных систем Федерального государственного учреждения «Федеральный исследовательский центр «Информатика и управление» Российской академии наук».

В период подготовки диссертации с 01.02.2017 по 31.01.2018 Трофимов И.Е.- экстерн ФИЦ ИУ РАН.

В 2008 г. Трофимов И.Е. окончил с отличием физический факультет Московского государственного университета им. М.В. Ломоносова.

Научный руководитель – д. ф.-м. н., профессор Воронцов Константин Вячеславович. Основное место работы - Федеральное государственное автономное образовательное учреждение высшего образования «Московский физико-технический институт (государственный университет)», заведующий лабораторией машинного интеллекта, профессор.

По итогам обсуждения диссертации «Разработка и обоснование методов параллельного покоординатного спуска для обучения обобщенных линейных моделей с регуляризацией» **принято следующее заключение**:

Тема диссертации является актуальной, поскольку, с одной стороны, обобщенные линейные модели с регуляризацией применяются для решения большого числа задач машинного обучения и анализа данных, с другой стороны, 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. Обучающие выборки, как правило, разреженные. Желательным свойством также является разреженность полученного решения. Если в этих задачах использовать для обучения только часть имеющихся данных, то качество предсказания, как правило, падает.

Поэтому важным направлением исследований является разработка методов машинного обучения, специально предназначенных для больших выборок, а также

разработка алгоритмов, позволяющих применять существующие методы на больших выборках.

Цель диссертационной работы - разработка и обоснование методов параллельного покоординатного спуска для обучения обобщенных линейных моделей с регуляризацией. Разработка методов, применимых для больших обучающих выборок и выполнения на вычислительном кластере.

Основные положения, выносимые на защиту:

1. Метод d-GLMNET, предназначенный для параллельного покоординатного спуска для минимизации функций риска обобщенных линейных моделей с регуляризацией ``elastic net'';
2. 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. Метод ``асинхронной балансировки нагрузки'' для обеспечения эффективного выполнения метода d-GLMNET при наличии медленных узлов кластера;
5. Численные эксперименты, доказывающие, что метод d-GLMNET более эффективен, чем общепринятые методы при работе с разреженными обучающими выборками с высокой размерностью признакового пространства.
6. Общедоступная программная реализация метода d-GLMNET: https://github.com/IlyaTrofimov/dlr.

Научная новизна данной работы заключается в разработке нового метода минимизации функций риска обобщенных линейных моделей с регуляризацией ``elastic net''. Метод эффективно работает с большими обучающими выборками, обладающими высокой размерностью признакового пространства. Научной новизной обладают теоретические результаты относительно сходимости метода, а также модификация метода (асинхронная балансировка нагрузки), обеспечивающая эффективное выполнение при неравномерности скорости работы узлов кластера.

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 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.

Теоретические и экспериментальные результаты данной работы используются в курсах ``Машинное обучение и большие данные'', которые автор читал в 2015-2018 гг. на факультете инноваций и высоких технологий (ФИВТ) МФТИ и Школе анализа данных (ШАД) Яндекса.

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

Апробация работы. Основные положения и результаты работы докладывались автором на конференциях:

1. Sixth International Workshop on Data Mining for Online Advertising and Internet Economy, 2012, Пекин;
2. 22nd International Conference on World Wide Web, 2013, Рио-де-Жанейро;
3. Machine learning and Very Large Data Sets, 2013, Москва
4. ;Seventeenth International Conference on Artificial Intelligence and Statistics, 2014, Рейкьявик;
5. The 4-th International Conference on Analysis of Images, Social Networks and Texts, 2015, Екатеринбург;
6. Machine Learning: Prospects and Applications, 2015, Берлин.

Список работ, опубликованных автором по теме диссертации:

1. Trofimov I., Kornetova A., Topinskiy V. Using boosted trees for click-through rate prediction for sponsored search // Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy. – ACM, 2012. – С. 2.
2. Trofimov I. New features for query dependent sponsored search click prediction // Proceedings of the 22nd International Conference on World Wide Web. – ACM, 2013. – С. 117-118.
3. Trofimov I., Genkin A. Distributed coordinate descent for l1-regularized logistic regression // International Conference on Analysis of Images, Social Networks and Texts. – Springer, Cham, 2015. – С. 243-254.
4. Trofimov I., Genkin A. Distributed coordinate descent for generalized linear models with regularization // Pattern Recognition and Image Analysis. – 2017. – Т. 27. – №. 2. – С. 349-364.
5. Трофимов И. Е. Распределенные вычислительные системы для машинного обучения // Информационные технологии и вычислительные системы. – 2017. – №. 3. – С. 56-69.

Личный вклад диссертанта является решающим во всех результатах, выносимых на защиту.

Содержание диссертации соответствует паспорту специальности **05.13.17 – теоретические основы информатики.**

Диссертация «Разработка и обоснование методов параллельного покоординатного спуска для обучения обобщенных линейных моделей с регуляризацией» рекомендуется к защите на соискание ученой степени кандидата физико-математических наук по специальности **05.13.17 – теоретические основы информатики**.

Заключение принято единогласно на заседании семинара отдела интеллектуальных систем Федерального государственного учреждения «Федеральный исследовательский центр «Информатика и управление» Российской академии наук». Присутствовало на

заседании 16 человек. Результаты голосования «за» - 16 человек, «против» - нет, «воздержалось» - нет.

Протокол № 1 от 08 февраля 2018 г.

Председатель семинара: _____ академик РАН, Рудаков К.В.

Секретарь: _____ д.ф.-м.н. Воронцов К.В.