ОТЗЫВ

члена диссертационного совета на диссертацию Боярова Андрея Александровича на тему «Рандомизированный подход к обучению в условиях отсутствия разметки и малого количества данных», представленную на соискание ученой степени кандидата физико-математических наук по специальности 01.01.09 – Дискретная математика и математическая кибернетика.

Актуальность и практическая значимость темы диссертации. Такие научные направления, как машинное обучение и распознавание образов в последние несколько лет переживают бурный рост. В силу доступности для обучения баз данных большого объёма, современные подходы к интеллектуальному анализу данных показывают значительное увеличение точности своей работы. Обратной стороной такого развития событий является то, что для достижения высокого качества, стандартным алгоритмам машинного обучения с учителем необходимы размеченные тренировочные данные большого объёма, прочищенные от возможных ошибок разметки. Например, для обучения одной модели для распознавания изображений необходимо собрать и вручную разметить более миллиона изображений. Эта задача трудозатратна и требует усилий многих людей. При таких обстоятельствах особую актуальность приобретают методы машинного обучения, способные обучаться на данных без разметки или с малым количеством разметки. К первому классу методов относятся алгоритмы, решающие задачу кластеризации, а ко второму – алгоритмы, решающие задачу обучения по малому количеству данных. В обоих рассматриваемых в диссертационной работе задачах возникают неопределённости, связанные с отсутствием заранее известной структуры и разметки данных. Наличие таких неопределённостей актуализирует необходимость в разработке новых методов оптимизации для машинного обучения и распознавания образов, обладающих робастностью к возникающим неопределённостям.

Первая глава диссертации посвящена описанию проблем обучения в условиях отсутствия разметки и малого количества данных. Приводится обзор основных алгоритмов для решения задач такого типа. Рассмотрена задача кластеризации в контексте обучения без учителя, приводится формулировка проблемы кластеризации в виде оптимизационной задачи, рассматривается смесь гауссовых распределений, в том числе, с разреженными параметрами, приводятся описания алгоритмов машинного обучения, применяемых для решения сформулированных задач, описаны критерии оценки качества кластеризации. Более того, приведено описание глубоких нейронных сетей и основных понятий, связанных с ними, даётся описание методов, входящих в три основные группы алгоритмов классификации по малому количеству примеров: метрические подходы, оптимизационные, рекуррентные. Первую главу завершает описание алгоритмов оценивания неизвестных параметров системы на основе методов стохастической аппроксимации для решения поставленных задач.

Во второй главе формулируются и доказываются основные теоретические результаты диссертационного исследования. Рассматривается применение рандомизированного подхода в задаче кластеризации, описывается основной

рандомизированный алгоритм стохастической аппроксимации для кластеризации в модели смеси гауссовых распределений, сформулирована и доказана теорема о состоятельности оценок центров кластеров и ковариационных матриц, полученных с помощью этого алгоритма, проведён анализ основных свойств построенного рандомизированного алгоритма. В том числе рассматривается случай смеси гауссовых распределений с разреженными параметрами, для чего описывается модификация алгоритма стохастической аппроксимации с рандомизацией на входе для кластеризации, сформулирована и доказана теорема о состоятельности оценок параметров, получаемых с помощью описанной модификации. Приведено описание рандомизированного подхода для обучения в условиях малого количества данных, рассмотрена задача многозадачного обучения, описывается алгоритм стохастической аппроксимации с рандомизацией на входе для классификации в условиях малого количества примеров, а также формулируется и доказывается теорема о состоятельности получаемых с помощью этого алгоритма оценок параметров. Более того, предлагается метод для обучения по малому количеству данных, основанный на рандомизированном алгоритме стохастической аппроксимации для трекинга, формулируется и доказывается обосновывающая получаемые оценки теорема.

Третья глава отведена под результаты экспериментов, иллюстрирующих работоспособность предложенных подходов в реальной задаче. Приведены результаты, полученные для имитационного моделирования в задаче кластеризации смеси гауссовых распределений. Описывается случай построения оценок центроидов и ковариационных матриц. Помимо этого, проведены соответствующие численные эксперименты при наличии аддитивного внешнего шума, а также эксперименты в случае модели смеси гауссовых распределений с разреженными параметрами. Представлены результаты сравнения предложенных алгоритмов со стандартными аналогами. Тестирование разработанного подхода к обучению классификатора по малому количеству примеров проведено на стандартном наборе данных Омниглот, а апробация представленных методов произведена на примере верификации авторства средневековых арабских манускриптов.

Теоретическую значимость диссертационной работы составляют условия сходимости оценок, полученных с помощью общего рандомизированного алгоритма стохастической аппроксимации для кластеризации в моделях смеси гауссовских распределений и смеси гауссовских распределений с разреженными параметрами, а также адаптивного алгоритма стохастической аппроксимации для оценивания параметров в многозадачной функции потерь. **Научную новизну** представляют следующие результаты:

1. Обоснован общий рандомизированный алгоритм стохастической аппроксимации для кластеризации в модели данных, описываемой смесью гауссовских распределений, и способный качественно работать при неизвестных, но ограниченных помехах.
2. Предложена модификация общего рандомизированного алгоритма стохастической аппроксимации для кластеризации для условий смеси гауссовских распределений с разреженными параметрами.

3. Представлен подход к обучению адаптивного классификатора по малому количеству размеченных данных, использующий метод стохастической аппроксимации для оптимизации параметров в многозадачной функции потерь.

Результаты диссертационного исследования сформулированы в виде теорем, доказательство которых подтверждает обоснованность разработанных подходов. **Достоверность результатов** диссертации определяются строгими математическими доказательствами, публикацией представленных в работе результатов на международных научных конференциях и в рецензируемых изданиях, апробацией в рамках ряда прикладных научно-исследовательских работ.

К диссертационному исследованию можно сделать следующие **замечания**:

1. В разделах 1.3 и 2.1.1 в алгоритмах стохастической аппроксимации используется разная нотация для обозначения компонент векторов. В разделе 1.3 используется нижняя нотация, а в разделе 2.1.1 – верхняя.
2. В разделах 3.1.1 и 3.1.2 проводится сравнение качества работы различных методов кластеризации с представленным в диссертационной работе. Для полноты сравнения необходимо было бы сравнить и временные сложности этих алгоритмов.
3. В различных местах работы используется две версии задания диапазона значений: две точки или две запятые и три точки. В целях унификации лучше использовать только одну версию во всей диссертации.

Тем не менее, указанные замечания не влияют на общую положительную оценку, которую заслуживает диссертационное исследование А. А. Боярова. Работа является законченным научным исследованием и обладает теоретической и практической ценностью. Результаты диссертации представлены в 9 работах автора, 1 из которых опубликована в издании из списка ВАК (71 в Списке литературы), 3 опубликованы в изданиях, индексируемых в международных базах Web of Science и Scopus (69, 70, 71 в Списке литературы), а 1 (74 в Списке литературы) – свидетельство на программу для ЭВМ.

Диссертация Боярова Андрея Александровича на тему «Рандомизированный подход к обучению в условиях отсутствия разметки и малого количества данных» соответствует основным требованиям, установленным Приказом от 01.09.2016 № 6821/1 «О порядке присуждения ученых степеней в Санкт-Петербургском государственном университете», а ее автор Бояров Андрей Александрович заслуживает присуждения ученой степени кандидата физико-математических наук по специальности 01.01.09 – Дискретная математика и математическая кибернетика. Пункт 11 указанного Порядка диссертантом не нарушен.

Член диссертационного совета
Доктор физико-математических наук,
профессор Санкт-Петербургского
государственного университета  Крылатов Александр Юрьевич

18.08.2020