Обучение нейронной сети, моделирующей социально-экономическое развитие региона

УДК 517:316.4:37.0 DOI 10.26425/2658-347X-2019-2-34-40

Получено 18.07.2019 Одобрено 10.09.2019 Опубликовано 29.10.2019

Романчуков Сергей Викторович

Аспирант, ФГАОУ ВО «Национальный исследовательский Томский политехнический университет», г. Томск, Российская Федерация

E-mail: inoytomsk@yandex.ru

Берестнева Ольга Григорьевна

Д-р техн. наук, ФГАОУ ВО «Национальный исследовательский Томский политехнический университет», г. Томск, Российская Федерация

E-mail: ogb6@yandex.ru

Петрова Людмила Андреевна

Канд. пед. наук, ГОУ ВО МО «Государственный гуманитарно-технологический университет», г. Орехово-Зуево, Российская Федерация

E-mail: plandr50@mail.ru

АННОТАЦИЯ

Статья посвящена вопросам формирования массива данных для построения искусственной нейронной сети, предназначенной для поиска взаимосвязей между социальными и экономическими параметрами развития регионов Российской Федерации (далее – РФ). Актуальность исследований в этой области подтверждается большим количеством исследований в области региональной компаративистики, а также ограниченностью методик, применяемых в такого рода исследованиях, зачастую ограничивающихся описательными методами и базовыми техниками параметрической статистики. В этих условиях расширение математического аппарата и более активное внедрение информационных технологий, в том числе в области анализа больших данных (англ. big data) и построения прогностических моделей на основе искусственных нейронных сетей, представляется небезынтересным. При этом, однако, необходимо отметить, что ресурсов отдельного исследовательского коллектива может быть (и, вероятнее всего, будет) недостаточно для создания с нуля собственного программного решения для реализации алгоритмов машинного обучения. Использование сторонних программных платформ на основе облачных технологий (в первую очередь – инфраструктуры от IBM и Google) позволяет обойти проблему отсутствия у исследовательского коллектива дорогостоящей материально-технической базы, однако накладывают ряд ограничений, продиктованных как требованиями существующих алгоритмов машинного обучения, так и спецификой архитектуры предоставляемых платформ. Это ставит коллектив исследователей перед необходимостью подготовки накопленного массива данных к обработке: снижению размерности, проверке данных на соответствие требованиям платформы и исключения потенциальных проблемных зон: утечек данных, перекосов в обучении и иных. Работа была доложена на секции «Социология цифрового общества: структуры, процессы, управление» Международной конференц-сессии «Государственное управление и развитие России: национальные цели и институты».

Ключевые слова

Портрет региона, компаративистика, факторный анализ, снижение размерности, искусственная нейронная сеть, обучение нейронной сети.

Цитирование

Романчуков С.В., Берестнева О.Г., Петрова Л.А. Обучение нейронной сети, моделирующей социально-экономическое развитие региона//Цифровая социология. 2019. Т. 2. № 2. С. 34–40.

Благодарности. Исследование выполнено при частичной финансовой поддержке РФФИ в рамках научных проектов №18-37-00344 и № 18-07-00543.

Teaching a neural network modeling socio-economic development of the region

DOI 10.26425/2658-347X-2019-2-34-40

Received 18.07.2019 Approved 10.09.2019 Published 29.10.2019

Romanchukov Sergey

Graduate student, National Research Tomsk Polytechnic University, Tomsk, Russia

E-mail: inoytomsk@yandex.ru

Berestneva Olga

Doctor of Technical Sciences, National Research Tomsk Polytechnic University, Tomsk. Russia

E-mail: ogb6@yandex.ru

Petrova Lyudmila

Candidate of Pedagogical Sciences, University for Humanities and Technologies, Orekhovo-Zuyevo, Russia

E-mail: plandr50@mail.ru

ABSTRACT

The article is devoted to the formation of an array of data for the construction of an artificial neural network, designed to search for relationships between social and economic parameters of the development of regions of the Russian Federation. The relevance of research in this area is confirmed both by a large number of studies in the field of regional comparativistics and by the limited methods used in this kind of research, often limited to descriptive methods and basic techniques of parametric statistics. Under these conditions, the expansion of the mathematical apparatus and the more active introduction of information technologies (including in the area of Big Data analysis and the construction of predictive models based on artificial neural networks) can be viable. At the same time, however, it should be noted that the resources of an individual research team may be (and most likely will be) insufficient to create their own software solution for the implementation of machine learning algorithms from scratch. The use of third-party cloud-based software platforms (primarily IBM and Google infrastructures) allows to bypass the problem of the research team's lack of expensive material and technical base, however they impose a number of limitations dictated by the requirements of the existing machine learning algorithms and the specific architecture provided platforms This puts the research team in front of the need to prepare the accumulated data set for processing: reducing the dimension, checking the data for compliance with the platform requirements and eliminating potential problem areas: "data leaks", "learning distortions" and others. The paper was reported to the section "Sociology of Digital Society: Structures, Processes, Governance" of the International Conference Session "Public Administration and Development of Russia: National Goals and Institutions".

Keywords

Portrait of the region, comparative studies, factor analysis, diminution of dimension, artificial neural network, teaching a neural network.

For citation

Romanchukov S.V., Berestneva O.G., Petrova L.A. Teaching a neural network modeling socio-economic development of the region (2019) Digital sociology, 2 (2), pp. 34–40. doi: 10.26425/2658-347X-2019-2-34-40

Acknowledgements. The study was carried out with partial financial support of the RFBR in the framework of research projects No. 18-37-00344 and No. 18-07-00543.

ВВЕДЕНИЕ

Эффективность осуществления органами государственной власти субъектов РФ и органами местного самоуправления их полномочий в определенной степени зависит от понимания роли и значимости взаимосвязей социально-экономических явлений и процессов. Без информации о взаимозависимости показателей выводы становятся поверхностными и необоснованными, а управленческий анализ приобретает формальный характер. Актуальность построения модели, описывающей механизм социально-экономического развития региона, объясняется обширностью территории РФ, многообразием природно-климатических зон, неоднородностью в распределении ее природных ресурсов, производств в количественном и национальном составе населения. Как следствие этих различий, в стране к началу 2009 г. более половины субъектов РФ имели статус депрессивных регионов. Кроме того, мировой экономический кризис, начавшийся в середине 2008 г., падение цен на энергоносители и начавшееся в 2014 г. санкционное давление со стороны западных партнеров осложнили положение и в наиболее удачных на то время регионах. В этой связи требуются новые подходы к изучению и мониторингу социально-экономической среды, а всю сложность и многообразие рассматриваемых признаков становится все сложнее свести к обобщающим показателям (таким как индикатор социального благополучия, валовой региональный продукт и т.д.).

РЕГИОНАЛЬНАЯ КОМПАРАТИВИСТИКА В РОССИЙСКОЙ ФЕДЕРАЦИИ

В настоящее время продолжаются работы по мониторингу состояния и динамики развития регионов России, а также поиска методик для улучшения точности прогнозирования социально-экономических процессов и эффективности управления ими. Так, например, с 2006 г. регулярно составляются так называемые социокультурные портреты регионов в рамках академической программы «Социокультурные проблемы эволюции России и ее регионов» и создания «Социокультурного атласа России». Мониторинг состояния социальной и экономической сферы регионов РФ осуществляется службами Росстата и целым рядом научных коллективов, выполняющих проекты регионального и местного уровня, в том числе подразделениями Института философии и Института социально-экономического развития территории при Российской академии наук[1].

Обработка результатов такого рода исследований представляет из себя массивную задачу, решаемую действующими научными коллективами преимущественно в рамках методов прикладной математической статистики и data mining (рус. добыча данных, интеллектуальный анализ данных, глубинный анализ данных). Наиболее частое упоминание в научных работах получают результаты, полученные с применением корреляционного, факторного и кластерного анализа. При этом наиболее популярными инструментами являются пакеты статистического анализа данных, такие как SPSS, Statistica, STATGRPHICS [Толстова, 2015], и языки обработки статистических данных, такие как R.

Вместе с тем представляет интерес возможность использования преимуществ междисциплинарного подхода, основанного на более активном применении информационных и сетевых технологий и методик вычислительного эксперимента, реализованных с использованием облачных вычислений. Так, в рамках проекта Российского Фонда Фундаментальных Исследований № 18-37-00344 осуществляется поиск возможностей для обнаружения и моделирования предполагаемой взаимосвязи социальных и экономических параметров с помощью комплексного подхода, включающего не только классические статистические методы и экспертные оценки, но и применение искусственных нейронных сетей.

ИСКУССТВЕННЫЕ НЕЙРОННЫЕ СЕТИ КАК ИНСТРУМЕНТ АНАЛИЗА СОЦИОЛОГИЧЕСКИХ ДАННЫХ

Искусственные нейронные сети достаточно эффективны в решении задач классификации, распознавания образов, предсказания поведения сложных систем и подбора неизвестных параметров, связывающих характеристики сложных объектов, в том числе экономических систем[2]. Процесс создания и обучения нейронной сети происходит итеративно, что позволяет добиваться желаемой точности и достаточно гибко настраивать создаваемую модель. Обучение нейронной сети предполагает процесс, в котором параметры нейронной сети настраиваются через моделирование среды, в которую эта сеть встроена. Существует несколько способов так называемого обучения нейронной сети: обучение с учителем, обучение без учителя, обучение с подкреплением [Васенков, 2007].

Наиболее подходящим для решения задачи о взаимосвязи двух групп параметров представляется первый вариант – обучение с учителем (англ. supervised learning) – способ машинного обучения, в ходе которого

[1] Центр изучения социокультурных изменений//Институт философии Российской академии наук. Режим доступа: http://iphras.ru/soc_cult_changes.htm (дата обращения: 10.07.2019).

[2] *Кенин А.М., Мазуров В.Д., Первушин Д.Р.* Опыт применения нейронных сетей в экономических задачах. Режим доступа: http://masters.donntu.org/2009/kita/soloduha/library/article5.htm (дата обращения: 10.07.2019).

испытуемая система принудительно обучается с помощью примеров «стимул-реакция» [Mohri et al., 2012].

Способ обучения с учителем предполагает, что между входами и эталонными выходами может существовать некоторая зависимость, но она неизвестна. Известна только конечная совокупность прецедентов – пар «стимул – реакция», называемая обучающей выборкой. На основе этих данных происходит итеративный процесс подбора параметров с целью восстановления зависимости и построения модели отношений, пригодной для прогнозирования, способный для любого объекта выдать достаточно точный ответ. Для измерения точности ответов так же, как и в обучении на примерах, может вводиться функционал качества [James, 2003].

ИСТОЧНИКИ СТАТИСТИЧЕСКИХ ДАННЫХ

Применение вышеописанного подхода требует подготовки соответствующей обучающей выборки. В нашем случае (поиск взаимосвязей между социальными и экономическими параметрами регионов) это означает решение целого ряда вопросов, начиная от философских (вплоть до того, какие параметры считать входными, а какие – выходными (стимулом и реакцией соответственно)), пусть даже и воздействующими на вход в рамках обратной связи) до сугубо практических – о структуре обучающей выборки.

На этапе подготовки массива данных необходимо в первую очередь определиться с источниками информации, желательно с соблюдением следующих требований:

1) достоверность источников информации:

– степень доверия к источнику;

– наличие исторических данных за длительный период времени;

– регулярность обновления данных;

– сопоставимость данных с информацией из других источников;

2) доступность источников данных:

– наличие (или простота получения) разрешений на получение и использование данных;

– возможность автоматизированного получения данных (например, через API (англ. application programming interface – интерфейс прикладного программирования)).

Это требование, впрочем, на текущем уровне информатизации отечественных наук об обществе является чисто декларативным и практически не выполняется.

В качестве наиболее очевидного, авторитетного и объемного источника статистических данных может выступать Росстат. Кроме того, в рамках исследования интерес представляют данные проекта «Социокультурные проблемы эволюции России и ее регионов», представленные, в том числе, посредством отдельной информационной системы [Айвазян и др., 1989] в качестве источника как сырых, так и прошедших статистическую обработку массивов данных по социальному самочувствию в регионах. В ряду параметров экономической сферы интерес представляют также маркеры инновационного развития – рейтинги и материалы Ассоциации инновационных регионов России[3].

В то же время примечательна возможность включения материалов исследовательских групп на местах, а также неофициальных и зарубежных источников. С одной стороны, это позволит придать получаемым результатам больший вес с точки зрения организаций, ставящих под сомнение материалы официальной статистики (в последние годы ряд источников в зарубежных и официальных средствах массовой информации все чаще обвиняют Федеральную службу статистики в политической ангажированности), и уточнить данные там, где масштабы работы всероссийских исследовательских институтов могут уступать в точности локальным измерениям. С другой стороны, работая с подобными источниками необходимо отдавать себе отчет в их потенциальной ненадежности.

Для решения этой задачи был выбран критерий качества, опирающийся на три числовых оценки: «авторитет источника», «новизна» и «серийность» наблюдений. Эти параметры перекликаются с названными ранее и тем самым позволяют использовать вышеназванные официальные источники в качестве основы, в которую могут привноситься дополнительные данные из источников с меньшим авторитетом и меньшим периодом ведения наблюдений. Результаты одиночных исследований могут использоваться для разового импорта статистических данных путем загрузки CSV-файлов, содержащих таблицы с результатами. Однако ценность таких разово добавленных таблиц изначально невелика (в силу отсутствия многолетних наблюдений и/или недостаточной авторитетности источника) и со временем будет и далее сокращаться по причине отсутствия в них обновлений и, соответственно, устаревания данных и невозможности убедиться, что они сохраняют свою актуальность.

ОБЛАЧНЫЕ СЕРВИСЫ ДЛЯ ГЛУБОКОГО АНАЛИЗА ДАННЫХ И ИХ ОГРАНИЧЕНИЯ

В процессе обработки больших массивов данных и обучения нейросетевых моделей на первый план выходит несколько проблем, связанных с вычислительными мощностями имеющегося аппаратного обеспечения, сложностей реализации и тестирования алгоритмов обучения и последующего применения нейронной сети. Этот набор проблем может быть серьезным и даже непреодолимым препятствием для

[3] Рейтинг инновационных регионов. Режим доступа: http://www.i-regions.org/reiting/rejting-innovatsionnogo-razvitiya (дата обращения: 10.07.2019).

небольшого исследовательского коллектива, не располагающего штатом разработчиков и необходимой материально-технической базой. Одним из возможных решений может быть применение SaaS (англ. software-as-a-sevice – программное обеспечение как услуга) продуктов для обработки данных, таких как IBM Watson или Google AutoML. Эти продукты предполагают предоставление доступа к необходимой инфраструктуре и программному обеспечению как бесплатно (в рамках пробной версии, ограниченной по объему и функциональности), так и за плату (как правило, это подписка, стоимость которой пропорциональна объему данных и сложности дополнительных сервисов), значительно уступающую стоимости разработки собственного ML-решения.

Компания IBM (также разработавшая SPSS (англ. statistical package for the social sciences – статистический пакет для общественных наук) и ряд других приложений для анализа данных) представляет Watson Machine Learning – в рамках среды Watson Studio и собственной облачной платформы для коллективной работы с данными, включая анализ больших данных и разработку элементов искусственного интеллекта. API для разработанных моделей создаются автоматически, позволяя разработчикам создавать приложения, обращающиеся к инфраструктуре IBM извне для загрузки данных из внешних источников и получения результатов их обработки автоматически. Удобные панели мониторинга Watson Machine Learning упрощают управление моделями, а оптимизированные процессы позволяют организовать непрерывное обучение моделей для повышения их точности[4].

Корпорация Google, один из признанных лидеров в области сбора и анализа больших данных, предлагает AutoML Tables – сервис, интегрированный с облачными сервисами Google Cloud, продуктами Google, такими как GSheet, и мощностями Google для обработки больших данных[5].

Оба эти решения позволяют вашей команде исследователей, аналитиков и разработчиков данных автоматически создавать и развертывать современные модели машинного обучения на структурированных (табличных) данных, пользуясь удобным графическим интерфейсом, библиотекой шаблонных моделей и гибкой системой настройки параметров. Обе среды позволяют реализовать алгоритмы обучения с учителем, используя табличные данные для обучения модели машинного обучения прогнозированию новых данных. На текущий момент наша команда экспериментирует с возможностями обеих платформ, так как их основная функциональность в значительной степени схожа: в зависимости от потребностей пользователя, оба упомянутых решения (как и ряд их менее известных аналогов) предлагают несколько наборов алгоритмов для обучения нейросетей, поставляемых в качестве шаблонных моделей:

– набор моделей для задач бинарной классификации;

– набор моделей для задач мультиклассовой классификации;

– набор регрессионных моделей.

Принято полагать, что с ростом количества наблюдений в выборке, точность получаемых результатов растет. Кроме того, количество наблюдений, необходимых для обучения нейронной сети прямо связано со сложностью задачи: построение корректного бинарного классификатора, как правило, требует меньше данных, чем построение мультиклассовго классификатора. Регрессионные модели, напротив, более требовательны к количеству и качеству собранных данных. Полагаясь в большей мере на практический опыт, чем на аналитические формулы, разработчики обеих сред предлагают одинаковые рекомендации по размеру обучающей выборки:

– для обучения классификатора: количество признаков, умноженное на 50;

– для обучения регрессионной модели: количество признаков, умноженное на 200[6].

Это требование естественным образом принуждает нас, с одной стороны, к сбору максимально возможного количества наблюдений, а с другой – к максимально возможному сужению пространства признаков. Кроме того, использование этих продуктов налагает еще ряд требований к массиву данных, вследствие чего накопленные сырые данные требуют предварительной подготовки. Сокращение пространства признаков, однако, является лишь первым шагом. Импорт данных в систему также предполагает их анализ и подготовку. Прежде чем датасет (от англ. data set – набор данных) можно будет использовать в обучении модели, необходимы:

– очистка массива исходных данных от ошибок и шума, вызванных ошибками, опечатками, сбоями кодировки или форматирования, повреждением CSV-файлов с данными и т.д.;

– проверка формата импортированных переменных целочисленных, вещественных, текстовых и т.д. (оба вышеупомянутых программных продукта пытаются определить формат полученных входных переменных автоматически, но в обоих случаях этот процесс не отличается точностью и требует корректировок со стороны оператора);

– проверка на наличие в колонках отсутствующих или неопределенных (NULL) значений (на практике неполнота входных данных достаточно частое явление,

[4] Официальный сайт IBM Watson. Режим доступа: https://www.ibm.com/ru-ru/cloud/machine-learning (дата обращения: 10.07.2019).

[5] AutoML Tables documentation. Режим доступа: https://cloud.google.com/automl-tables/docs (дата обращения: 10.07.2019).

[6] AutoML Tables documentation. Режим доступа: https://cloud.google.com/automl-tables/docs (дата обращения: 10.07.2019).

но оно негативно сказывается на эффективности и точности обученной на таких данных нейросети);

– в случае большого количества пропущенных и неопределенных значений необходима рефлексия о причинах такого положения дел (пропуски данных случайны или следуют какому-то паттерну? Если речь идет о систематическом пропуске данных, как это повлияет на погрешность?);

– проверка на корреляции и ревью наборов коррелирующих переменных;

– удаление отсутствующих, неполных и некорректных наблюдений (по возможности).

ПОДГОТОВКА МАССИВА ДАННЫХ

Формируемый массив данных изначально представляет собой таблицу (набор таблиц), строки в которых формируются парой ключей «регион» – «год». Количество столбцов (переменных) в данном массиве и ограниченность доступного количества наблюдений делает очевидной необходимость в реорганизации данных и в первую очередь снижении размерности пространства признаков (об этом же нам говорит желаемое соотношение числа признаков, рассматриваемых в модели, и количества необходимых для обучения модели наблюдений).

В качестве решения этой задачи выступают методы факторного анализа, поиск в выборке скрытых (латентных) переменных или факторов, объясняющих поведение выборки в пространстве меньшей размерности[7]. Природа имеющихся переменных, позволяет применить к ним категориальный метод главных компонент (англ. categorical principal components analysis; CATPCA). Суть этого метода сводится к тому, что каждому уровню категориальной переменной назначаются значения масштаба, оптимальные для решения задачи поиска главных компонент (невозможных в исходном пространстве категориальных переменных). Решение категориального анализа главных компонент максимизирует корреляции оценок объектов с каждой из квантифицированных переменных для числа компонент (измерений)[8].

Другой важной задачей является отсеивание в импортированном наборе данных части наблюдений, целью которого является обеспечение более равномерного распределения целевых признаков в выборке. Проще всего объяснить этот этап на примере задачи классификации: если подавляющее большинство объектов в выборке принадлежит к одному классу, то нейронная сеть, следуя самому простому из возможных путей, будет относить любые новые объекты к этому же классу с высокой степенью вероятности. Модель, обучаемая на результатах теста, в котором в 80 % случаев правильным является ответ «С», быстро придет к тому, что ответ «С» превосходно подходит к большинству вопросов, и дальнейший процесс обучения ни к чему не приведет. В случае, если количество наблюдений в выборке не позволяет уравнять численность разных классов (например, удаление большого числа записей приведет к тому, что общий размер выборки станет недостаточным для проведения анализа), разработчики AutoML рекомендуют следовать соотношению 1 к 10: класс с наименьшим количеством наблюдений должен составлять не меньше чем 10 %, от численности наибольшего класса. Таким образом, если в наибольшем классе имеется порядка 10 000 наблюдений, наименьший должен описываться как минимум 1 000 примеров[9].

Другим важным риском, которого необходимо избежать, является так называемая утечка данных – ситуация, когда во время обучения используются входные данные, в которых содержится информация о переменных, которые нейронная сеть пытается предсказать, недоступные на этапе фактического применения обученной модели. Это может быть обнаружено, когда функция, которая тесно связана с целевым столбцом, включена как одна из входных переменных. Например, модель обучается задаче прогнозирования размера доходной части бюджета в течении года, и в качестве входных переменных получает статистические данные по размеру налоговых поступлений за этот же год, которые в реальности не будут доступны до завершения года и предоставления итоговой отчетности.

Следующий риск, который требуется предотвратить в процессе подготовки массива входных данных – перекос при обучении (англ. training-serving skew) – ситуация, когда входные переменные, используемые во время обучения, отличаются от тех, которые будут предоставляться модели во время ее эксплуатации, что приведет к низкому качеству прогнозов, предоставляемых подготовленной моделью. Частой причиной такой ситуации может служить несовпадение временных отрезков во время обучения и эксплуатации, например, подготовка модели, обученной на среднегодовых данных, в то время как декларируемая цель обучения модели – прогноз среднемесячных показателей.

На текущий момент завершены процедуры по формированию пространства факторов. Частично выполнены перечисленные операции. В качестве следующих шагов предполагается попытка выделения

[7] Информационная система «Модернизация» ЦИСИ ИФРАН. Режим доступа: http://mod.vscc.ac.ru/ (дата обращения: 10.07.2019).

[8] Categorial Principal Components Analysis // IBM KnowledgeCenter. Режим доступа: http://www.ibm.com/support/knowledgecenter/ru/SSLVMB_23.0.0/spss/categories/choosing_catpca.html (дата обращения: 10.07.2019).

[9] AutoML Tables documentation. Режим доступа: https://cloud.google.com/automl-tables/docs (дата обращения: 10.07.2019).

в полученном пространстве наиболее информативных признаков для определения целевых переменных модели, но применение информационных метрик потребует значительной реорганизации массива данных.

ЗАКЛЮЧЕНИЕ

Предполагается, что проведение вышеназванных процедур позволит сократить массив данных в достаточной мере, чтобы упростить их дальнейшую обработку, но в то же время сохранить информацию о закономерностях, для поиска которых пришлось аккумулировать столь значительный (и неудобный для непосредственной обработки) массив информации. Последующие вычисления покажут правоту или ошибочность этого предположения. Текущее исследование далеко от своего завершения, однако авторы полагают необходимой дискуссию о расширении пула методов обработки результатов социально-экономических исследований и более активном вовлечении в него сетевых технологий, включая искусственные нейронные сети.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. (1989). Прикладная статистика. Классификация и снижение размерности: справочное издание / Под ред. С.А. Айвазяна. М.: Финансы и статистика.

Васенков Д.В. (2007). Методы обучения искусственных нейронных сетей//Компьютерные инструменты в образовании. № 1. С. 20–29.

Толстова Ю.Н. (2015). Социология и компьютерные технологии//Социологические исследования. № 8. С. 3–13.

James G. (2003). Variance and Bias for General Loss Functions//Machine Learning. Режим доступа: http://www-bcf.usc.edu/~gareth/research/bv.pdf (дата обращения: 10.07.2019).

Mohri M., Rostamizadeh A., Talwalkar A. (2012). Foundations of Machine Learning. The MIT Press, Cambridge, USA.

REFERENCES

Aivazyan S.A., Bukhshtaber V.M., Enyukov I.S., Meshalkin L.D. (1989), Applied statistics: classification and dimension reduction [*Prikladnaya statistika. Klassifikatsiya i snizhenie razmernosti*], in Aivazyan S.A. (ed.), Finansy i statistika, Moscow, Russia. [in Russian].

Vasenkov D.V. (2007), “Methods of teaching artificial neural networks” [“Metody obucheniya iskusstvennykh neironnykh setei”], *Computer tools in education*, no. 1, pp. 20–29.

Tolstova Yu.N. (2015), “Sociology and Computer Technologies” [“Sotsiologiya i komp’yuternye tekhnologii”], *Sotsiologicheskie issledovaniya*, no. 8, pp. 3–13.

James G. (2003), “Variance and Bias for General Loss Functions”, *Machine Learning*. Available at: http://www-bcf.usc.edu/~gareth/research/bv.pdf (accessed 10.07.2019).

Mohri M., Rostamizadeh A., Talwalkar A. (2012), *Foundations of Machine Learning*, The MIT Press, Cambridge, USA.

TRANSLATION OF FRONT REFERENCES

[1] Centre for the Study of Sociocultural Change. Available at: http://iphras.ru/soc_cult_changes.htm (accessed 10.07.2019).

[2] Kenin A.M., Mazurov V.D. The experience of using neural networks in economic tasks. Available at: http://www.uralstars.com/Docs/Editor/Neuro.htm (accessed 10.07.2019).

[3] TOP Innovation regions. Available at: http://www.i-regions.org/reiting/rejting-innovatsionnogo-razvitiya (accessed 10.07.2019).

[4] Official site IBM Watson. Available at: https://www.ibm.com/ru-ru/cloud/machine-learning (accessed 10.07.2019).

[5] AutoML Tables documentation. Available at: https://cloud.google.com/automl-tables/docs (accessed 10.07.2019).

[6] Ibid**.**

[7] Information system “Modernisation”. Available at: http://mod.vscc.ac.ru/ (accessed 10.07.2019).

[8] Categorical Principal Components Analysis, IBM KnowledgeCenter. Available at: http://www.ibm.com/support/knowledgecenter/ru/SSLVMB_23.0.0/spss/categories/choosing_catpca.html (accessed 10.07.2019).

[9] AutoML Tables documentation. Available at: https://cloud.google.com/automl-tables/docs (accessed 10.07.2019).