ПРОГНОЗ УСПЕВАЕМОСТИ СТУДЕНТОВ СПО С ПОМОЩЬЮ ТЕХНОЛОГИИ БОЛЬШИХ ДАННЫХ

FORECAST OF ACADEMIC PERFORMANCE STUDENTS COLLEGE BY USING BIG DATA TECHNOLOGY

S. Suvorov
N. Tsarkova
V. Pereverzeva

Summary. The Article is devoted to the consideration of Big Data technology as a tool for assessing the effectiveness of secondary vocational education, which can determine the trends of student performance, the prerequisites for deductions and the change of one's specialty to another depending on various indicators. In order to solve the problem of assessing the effectiveness of SPO, the review of domestic and foreign literature is carried out, the task of creating a tool that can determine the trends of student performance using mathematical methods and information technologies is set. A Significant result of the study is the description of Big Data technology, identification of distinctive features of this technology, structuring of processes, definition of the system of data collection and processing in an educational institution and designation of the properties of the collected data Big Data. The considered technology of big data operation aimed at the forecast of progress will provide the analysis which can be used by educational Department for preventive measures and conversations with parents and children; for the reporting and for increase of efficiency of educational system.

Keywords: Big Data technology, data collection, education efficiency, students 'performance, forecasting.

Суворов Станислав Вадимович
К.э.н. профессор, ФГБОУ ВО «Московский политехнический университет»
ssw1168@mail.ru
Царькова Наталья Ивановна
К.п.н., доцент, ФГБОУ ВО «Московский политехнический университет»
tsarkovani@mail.ru
Переверзева Владислава Игоревна
ФГБОУ ВО «Московский политехнический университет»
pvlada737@gmail.com

Аннотация. Статья посвящена рассмотрению технологии Big Data в качестве создания инструмента оценки эффективности среднего профессионального образования, способного определит тренды успеваемости учащихся, предпосылки отчислений и смены своей специальности на другую в зависимости от различных показателей. Для решения проблемы оценки эффективности СПО производится обзор отечественной и зарубежной литературы, ставится задача создания инструмента, способного определит тренды успеваемости учащихся с использованием математических методов и информационных технологий. Значимым результатом исследования является описание технологии Big Data, выявление отличительных признаков этой технологии, структурирование процессов, определение системы сбора и обработки данных в учебном заведении и обозначение свойств собираемых данных Big Data. Рассмотренная технология оперирования большими данными предоставит анализ, который может использоваться воспитательным отделом для профилактических мер и бесед с родителями и детьми, а так же для отчетности и для повышения эффективности образовательной системы.

Ключевые слова: Технология Big Data, сбор данных, эффективность образования, успеваемость студентов, прогнозирование.

Введение

Уже продолжительное время одним из основных векторов развития России является совершенствование образования населения. Поэтому, тенденция пересмотра традиционных понятий в рамках современных информационных технологий все чаще касается области образовательных процессов. Тут увеличивается информация об учащихся категории «одаренные дети», о детях с особо развитым мышлением, о детях «золотые руки», о тех, кто уже состоялся как лидер, а также о тех, у кого двигательные или творческие способности на более высоком уровне. Опираясь на индивидуальные особенности каждого ученика и на стремительное развитие информационного обеспечения, происходит переосмысление понятий «эффективное обучение» и «предметное изучение», развиваются образовательные системы, пополняются реестры востребованных специальностей на основе потребностей современного рынка труда, а дети тем временем нуждаются в поддержке принятия подходящего выбора будущей профессии из такого большого списка решений.

Первыми сталкиваются с таким выбором те, кто решил получить среднее профессиональное образование. Одной из технологий, помогающей предпринимать действия, основываясь на анализе данных и статистики прошлых лет является технология оперирования большими данными или Big Data. Возникновение этой технологии тесно связано с ежегодно увеличивающимся объемом накопленных данных. Система образования собрала внушительные объемы информации и теперь стоит вопрос о методах ее обработки и об использовании информационно — коммуникационных технологий для решения поставленной задачи.

Big Data в образовании — это технология аналитики системы образования, включающая в себя измерение, сбор, анализ и представление структурированных и неструктурированных данных огромных объемов об обучающихся и об образовательной среде с целью понимания особенностей функционирования и развития образовательной системы.

Литературный обзор

Сегодня наблюдается достаточная разобщенность исследований в отечественной и зарубежной литературе по вопросу использования Big Data в системе образования, однако, машинное обучение становится языком общения для образовательных организаций, стремящихся улучшить свои стратегические и тактические технологии принятия решений. Так, анализ названий более трех тысяч научных статей в отрасли «Компьютеры и образование», проведенный О. Заваки-Рихтером и С. Латчемом, позволяет в течение последних 40 лет выделить четыре хронологических этапа компьютеризации образования: развитие и рост компьютерного обучения (1976–1986 годы); мультимедийное обучение (1987–1996 годы); сетевые технологии для организации совместного обучения (1997–2006 годы); онлайн — обучение (2007–2016 годы). Из этого анализа можно заметить большое внимание научного сообщества к вопросам онлайн-обучения в последние годы, а способам повышения его эффективности не может быть найден без разностороннего анализа показателей студентов, собранных по результатам обучения [1].

Итоги исследования С. Виейра, П. Парсонс, В. Бердпо интеллектуальному анализу данных в образовании [2] показывают, что весомым параметром анализа данных является демография и предыдущие учебные успехи, а Н. Бунийамин, У. Б. Мет, П. М. Аршад провели анализ наиболее часто используемых методов классификации в области интеллектуального анализа образовательных данных для прогнозирования академических успехов учащихся [3]. Группа ученых под руководством Дж. Окумпау при анализе результатов обучения учеников с помощью Big Data делает вывод, что выявленные закономерности одной демографической группы, не обобщают результаты, взятые преимущественно из других демографических групп, однако, все эти группы населения могут являться частью одной и той же региональной или национальной культуры [4].

Исследования Big Data в образовании выделяется еще одна область, которой является инфраструктура собираемых данных. Так, в работах А. Кастро и Ф. А. Де Алмейда Нето [5], решением стали онлайн-платформы, с размещающейся в них информацией. На таких платформах данные хранятся в локальных базах данных, и пополняются путем взаимодействия пользователей и самой платформы. Над данными производятся выборки, они собираются и хранятся, а затем локальные БД собираются, группируясь в глобальную базу.

Следующей областью стали исследования сравнивания инструментов прогнозирования качественных показателей обучения. Так, Д. Буэненьо-Фернандес и С. Луан-Мора произвели анализ трех инструментов, используемых в интеллектуальном образовании, исходный код которых является открытым: RapidMiner, Knime и Weka, [6].

Еще одним важным направлением исследований являются вопросы, связанные с внутренним взаимодействием. Прогнозирование академической успеваемости — одна из ключевых тем исследований в области Big Data в образовании. Б. Го делает вывод, что оценка успеваемости является сложной задачей, поскольку успеваемость учащихся зависит от различных факторов. Взаимосвязь между параметрами успеваемости и факторами для прогнозирования производительности участвует в сложных нелинейных связях, поэтому направления сбора данных должны быть всеохватывающими [7]. Так, для охвата направлений О. Москосо-Зеа в своих трудах описывает структуру управления большими данными. Управление дает возможность обработки информации для анализа ключевых показателей учебной эффективности [8]. Таким образом, в настоящее время развитие технологии Big Data в образовании описывается через множество подходов и моделей, что мешает систематическому накоплению данных о Big Data для развития системы образования.

Постановка задачи

В отличие от высшего образования, среднее профессиональное образование отличается акцентированием внимания на воспитательных функциях обучения и социализацию, которые обусловлены подготовкой детей несовершеннолетнего возраста.

Аудитория СПО состоит из следующих категорий детей [9]:

- поступившие после 9 лет обучения в школе по собственному желанию на специальность, которую выбрали сами (40%);
- по наставлению родителей, советам друзей, из рекламы в СМИ (50%);
- поступившие на 2 курс после 11 классов, не прошедшие проходной балл ЕГЭ (10%).

Для контроля социально-культурной сферы жизни студента в колледжах организованы воспитательные отделы, которые активно контактируют с родителями и детьми от момента подачи документов в образовательное учреждение и до момента выпуска, помогая обозна-

чить позицию учащегося в образовательном процессе. Если студент не справляется с освоением учебного материала и из-за неуспеваемости близок к отчислению, то перед ним встает выбор отчислиться, уйти в академический отпуск или перевестись на более легкую специальность, что чаще всего практикуется.

В настоящее время в отрасли СПО отсутствует инструмент поддержки принятия решения, который подскажет и определит тренды успеваемости учащихся и предпосылки смены своей специальности на другую на ранних этапах обучения. Цель состоит в создании такого инструмента, который будет анализировать не только качественные и количественные показатели студента, но и его индивидуальные особенности.

Для определения эффективности обучения при помощи технологии больших данных и достижения цели выделим следующие этапы:

- определить исходные данные — триггеры, влияющие на комплексное решение и по возможности преобразовать их к нейронному стандарту;
- определить архитектуру ИНС;
- произвести обучение сети.

Реализация

Определение триггеров — первый, базовый этап решения задачи. Это определение кратких и понятных коэффициентов, метрик-сигналов, оцененных в конкретных величинах. Здесь стоит задача в поиске усредненных показателей, которые окажутся важными для работников воспитательного отдела, для студентов и их родителей.

Вся информация о студентах хранится в трех отдельных СУБД. Первая представляет собой учетную карточку студента. Карточки организованы в таблицах Microsoft Access, которые содержат личную информацию студентов, средний балл аттестата среднего общего образования, даты и номера приказов о поступлении, переводах на другие специальности, о выпусках, отчислениях или о взятии академического отпуска и др. Вторая система содержит информацию об оценках и задолженностях по дисциплинам, а третья — некоторые данные о психологическом портрете ученика и о его увлечениях, полученные из обязательного анкетирования при поступлении.

При помощи SQL запросов на объединение собирается сводная таблица из основных триггеров, которые включаются в обучающую выборку. Значимость триггеров подтверждается проведением корреляционного и регрессионного анализа. На вход в ИНС будут подаваться данные:

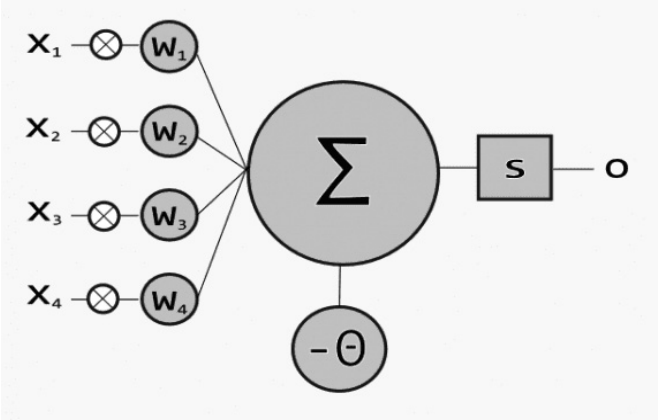

Рис. 1.Искусственная нейронная сеть с 4 входами

- средний балл аттестата;
- средний балл за каждый семестр;
- процент прогулов;
- специальность;
- хобби;
- увлечение спотом;
- наличие высшего образования у родителей;
- отношение к выполнению задач;
- город.

На выходе получаем следующую информацию в булевом типе данных:

- продолжает обучение;
- академический отпуск;
- перевод на другую специальность;
- отчисление.

Для анализа и прогнозирования данных будет проектироваться искусственная нейронная сеть или ИНС. ИНС — это вычислительная нелинейная модель, способная обучаться только благодаря рассматриваемым примерам и решать задачи принятия решений, предсказания, классификации, визуализации.

Архитектура ИНС представляет собой структуру из трех слоев, состоящих из искусственных нейронов или элементов обработки. Все слои нейронной сети связаны друг с другом и классифицируются как входной, выходной и скрытый (состоит из одного или более слоев).

Нейронные сети включают в себя следующее:

- входной слой, x;
- выходной слой, ŷ;
- скрытые слои (произвольное количество);
- выбор функции активации для каждого скрытого слоя σ;
- набор весов, W;

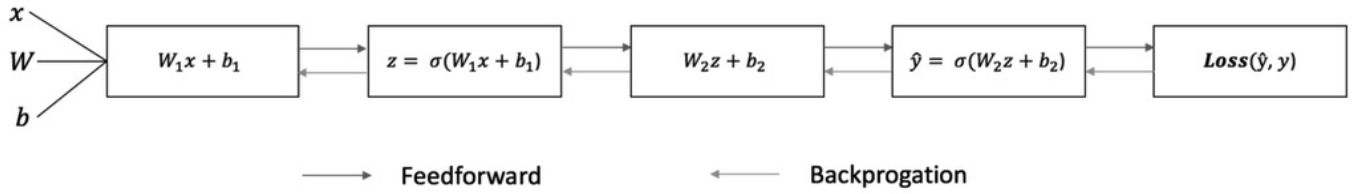

Рис. 2.График последовательности

- набор смещений между каждым слоем, b;

Обучением — процесс оптимизации весов, с целью минимизации ошибок предсказания, при которой сеть достигает требуемого уровня точности. Наиболее подходящий метод для данной задачи, определяющий вклад каждого нейрона в ошибку — обратное распространение ошибки, который помогает вычислитьт градиент, являющийся одной из модификаций градиентного спуска.

Метод обучения, который будет использоваться — обучение с учителем, так как выполняется условие наличия полного набора ответов, который должен вывести алгоритм.

Используемая архитектура — многослойный перцептрон, Такая архитектура позволяет классифицировать линейно неразделимые данные, что достигается за счет полностью связной сети, каждый узел в слое которой соединен с каждый узлом в последующем слое. Для решения данной задачи будем использовать два скрытых слоя и функцию активации — Sigmoid.

Выход ŷ двухслойной нейронной сети:

$$\hat{y} = \sigma(W_2\sigma(W_1x + b_1) + b_2)$$

В уравнении веса W и смещения b являются единственными переменными, имеющими влияние на результатный выход ŷ, следовательно, правильность их значений определяют точность предсказаний системы.

Обучающий процесс состоит из набора итераций, которые в свою очередь разбиваются на шаги, представленные на рисунке 2:

- вычисление прогнозируемого выхода ŷ (прямое распространение);
- обновление весов и смещений (обратное распространение).

Изначально предполагаем, что смещения b равны 0, тогда возникает необходимость в расчете ошибок, чтобы узнать отклонение от правильного ответа. В качестве функции потери будем использовать сумму квадратов ошибок, которая является: средним значением разницы между каждым фактическим и прогнозируемым значением.

$$Sum - of - Squares\ Error = \sum_{i=1}^{n}(y - \hat{y})^2$$

Задача обучения в данном случае заключается в поиске таких набор весов и смещений, которые приведут к минимуму значение функции потери.

После измерения ошибки прогноза, необходимо найти способ, применяемый к распространению ошибки полярно для обновления показателей смещений и весов. Таким способом является поиск производной функции потери по отношению к весам и смещениям, учитывая, что производная функции является тангенсом угла наклона функции, рисунок 3.

При нахождении производной становится возможным обновлять смещения и веса путем их уменьшения или увеличения. Такой метод называется градиентным спуском. Однако нет возможности непосредственно производить вычисления производной функции потерь по отношению к смещениям и весам, так как уравнение функции потерь их не содержит. Поэтому возникает необходимость в применении правила цепи для помощи в вычислении.

$$Loss(y, \hat{y}) = \sum_{i=1}^{n}(y - \hat{y})^2$$

$$\frac{\partial\, Loss(y,\hat{y})}{\partial W} = \frac{\partial Loss(y,\hat{y})}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial z} * \frac{\partial z}{\partial W} \quad \text{where } z = Wx + b$$

$$= 2(y - \hat{y}) * \text{derivative of sigmoid function} * x$$

$$= 2(y - \hat{y}) * z(1\text{-}z) * x$$

Таким образом, находится производная(наклон) функции потерь по отношению к весам, которые теперь можно регулировать. Далее переходим к самому обучению.

После определения данных стоит задача переноса их в массив и прогона через нейросеть, для этого будем ис-

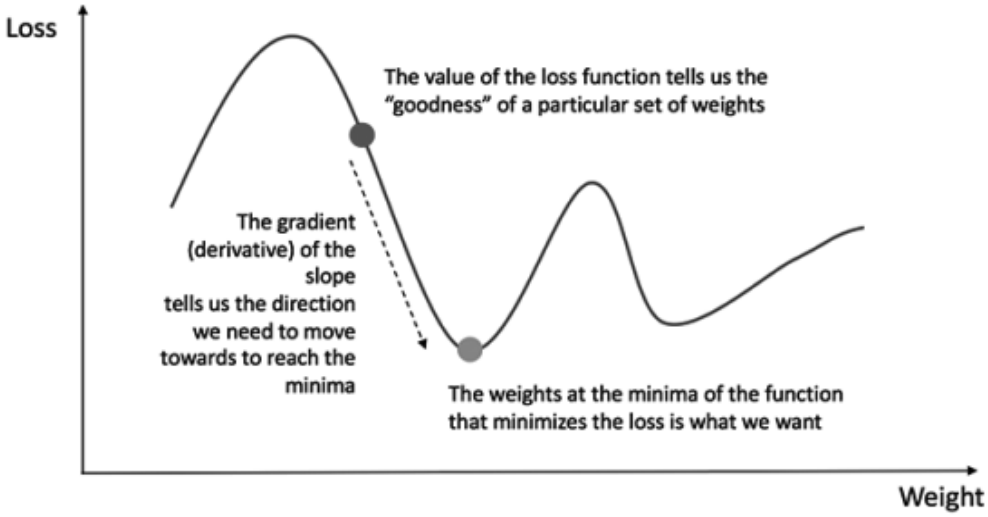

Рис. 3.График функции потерь

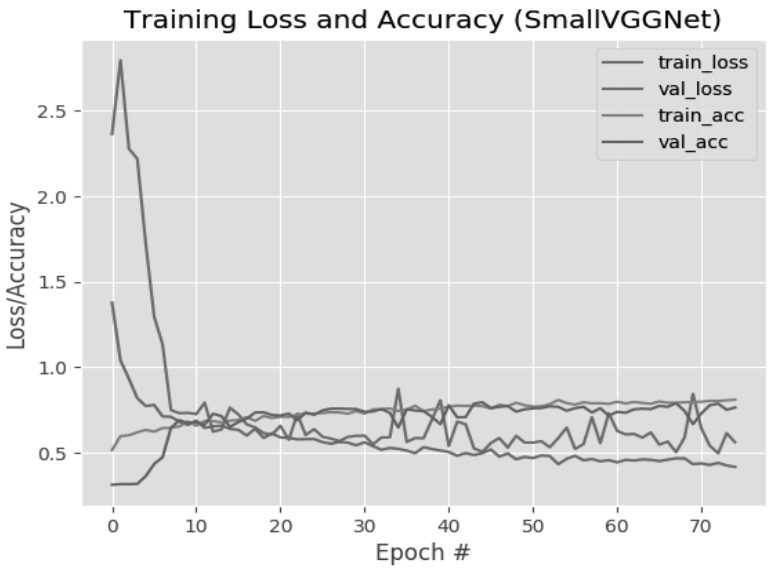

Рис. 4.График степени обучения сети

пользовать язык программирования Python 3.7 совместно с фреймворками Keras, TensorFlow и OpenCV, которые предоставят доступ к нужным пакетам и библиотекам.

После выгрузки данных их необходимо разделить на тестовую и обучающую выборку. Большую часть данных выделим для обучения, и около 15–30% для тестирования. Это необходимо для понимания степени обучения сети.

На следующем этапе с использованием Keras укажем количество нейронов на слоях нейронной сети. При необходимости, число нейронов на слоях будет корректироваться. Количество узлов на выходном слое — 4 для каждой категории:

- продолжает обучение(0001);
- академический отпуск(0100);
- перевод на другую специальность(0010);
- отчисление(0001).

После задания параметров архитектуры происходит компилирование, при котором указывается скорость обучения и общее число эпох (прогонов по выборке данных) и применяется метод «стохастический градиентный спуск». При завершении компилирования при необходимости произвести подгонку.

Следующий этап заключается в оценке модели с помощью тестовой выборки, позволяющей понять, как именно нейросеть будет работать с данными, на которых она не обучалась. Для этого можно использовать комбинацию методов.predict и classification_report из библиотеки scikit-learn. После запуска программы сеть начнет обучаться и в итоге покажет, на сколько были точны ее ответы.

Также можно вывести график, содержащий информацию о потерях при обучении и ценке, а так же о точности обучения и оценивания.

С его помощью мы можем отследить недообучение или переобучение модели, рисунок 4.

Просматривая данный график, можно увидеть, высокую степень обучения. После выполнение всех шагов нейросеть сохраняется с возможностью дальнейшего использования.

Результаты

Значимым результатом исследования является описание технологии Big Data. В статье были выявлены отличительные признаки этой технологии, структурированы процессы, управляющие системой, направления сбора и обработки данных в учебном заведении и определены свойства собираемой базы Big Data.

Заключение

Рассмотренная технология оперирования большими данными, направленная на прогноз успеваемости предоставит анализ, который может использоваться воспитательным отделом для профилактических мер и бесед с родителями и детьми; для отчетности и для повышения эффективности образовательной системы.

Таким образом, Машинное обучение на основании заданных алгоритмов при построении математических моделей процессов в образовательной системы СПО позволит с помощью самообучаемых нейронных сетей автоматизировать процесс поддержки принятия решения, который подскажет и определит тренды успеваемости учащихся, а так же позволит после определения человеком основных параметров и алгоритмов самостоятельно выстраивать структурные связи между показателями успеваемости обучающихся в СПО.

ЛИТЕРАТУРА

1. Zawacki-Richter O., Latchem C. Exploring four decades of research in computers & education // Computers and Education.— 2018.— № 122.— P. 136–152. doi: 09.2019/j.compedu.2018.04.001.
2. Vieira C., Parsons P., Byrd V. Visual learning analytics of educational data: A systematic literature review and re-search agenda // Computers and Education.— 2018.— № 122.— P. 119–135. doi: 09.2019/j.compedu.2018.03.018.
3. Buniyamin N., Mat U. B., Arshad P. M. Educational data mining for prediction and classification of engineering students achievement // Paper presented at the 2015 IEEE7th International Conference on Engineering Educa-tion, ICEED2015.— 2018.— P. 49–53. doi: 09.2019/ICEED.2015.7451491
4. Population validity for educational data mining models: A case study in affect detection / J. Ocumpaugh, R. Baker, S. Gowda, N. Heffernan, C. Heffernan // British Journal of Educational Technology.— 2014.— № 45(3).— P. 487–501. doi: 09.2019/bjet.12156
5. De Almeida Neto F. A., Castro A. A reference architecture for educational data mining // Paper presented at the Proceedings Frontiers in Education Conference, FIE.— 2017.— October.— P. 1–8. doi: 09.2019/FIE.2017.8190728.
6. Buenaño-Fernández D. B., Luján-Mora S. Comparison of applications for educational data mining in engineering education // Paper presented at the EDUNINE2017 — IEEE World Engineering Education Conference: Engineering Education — Balancing Generalist and Specialist Formation in Technological Carriers: A Current Challenge, Pro-ceedings.— 2017.— P. 81–85. doi: 09.2019/EDUNINE.2017.7918187.
7. Predicting students performance in educational data mining / B. Guo, R. Zhang, G. Xu, C. Shi, L. Yang // Paper presented at the Proceedings — 2015 International Symposium on Educational Technology, ISET 2015.— 2016.— P. 125–128. doi: 09.2019/ISET.2015.33.
8. Moscoso-Zea O., Andres-Sampedro, Luján-Mora S. Datawarehouse design for educational data mining // Paper presented at the 2016 15th International Conference on Information Technology Based Higher Education and Training, ITHET 2016. doi: 09.2019/ITHET.2016.7760754.
9. Организация профориентационной URL: https://presentacii.ru/ presentation/organizaciya-proforienta-cionnoy-raboty-v-gbou-sosh-301-frunzenskogo — rayona (дата обращения 01.09.2019).
10. Царькова Н.И., Ерисов В. Д., Пекова Е. А. ТЕХНОЛОГИЯ BIG DATA КАК ИНСТРУМЕНТ УПРАВЛЕНИЯ В МЕЖКУЛЬТУРНОЙ КОММУНИКАЦИИ // Управление экономическими системами: электронный научный журнал, 2019. № 7.
11. Суворов С.В., Царькова Н. И., Спиридонова А. К. Анализ больших данных компании Uber Technologies Inc с помощью технологии Data Mining// Управление экономическими системами: электронный научный журнал, 2019. № 7.

© Суворов Станислав Вадимович (ssw1168@mail.ru),
Царькова Наталья Ивановна (tsarkovani@mail.ru), Переверзева Владислава Игоревна (pvlada737@gmail.com).
Журнал «Современная наука: актуальные проблемы теории и практики»