Проблемы классификации успеваемости обучающегося на основе данных из московской электронной школы

Кондратьев С. А., магистрант кафедры «Прикладная информатика»,
Московский политехнический университет, Москва, Россия
Рабинович А. Е., к.э.н., доцент кафедры «Прикладная информатика»,
Московский политехнический университет, Москва, Россия

Аннотация. Интенсивное развитие информационных технологий в 21 веке позволило сделать большой шаг к внедрению средств электронного обучения с применением алгоритмов интеллектуального анализа данных. В данной статье предлагается модель глубокой нейронной сети (DNN) для определения уровня успеваемости учащихся. Выводы, полученные в статье, дают образовательным учреждениям направление для корректировки образовательных программ. Проводится сравнение с уже применяемыми алгоритмами машинного обучения, которые используют тот же набор данных, что и предложенная модель.

Ключевые слова: классификация данных, глубокие нейронные сети, машинное обучение

Problems of classifying student performance based on data from the Moscow electronic school

Kondratev S.A., magistrate of the «Applied Informatics» Department at the Moscow polytechnic University, Moscow, Russia
Rabinovich A.E., candidate of economic sciences, Associate Professor of the «Applied Informatics» Department at the Moscow polytechnic University, Moscow, Russia

Annotation. Intensive development of information technologies in the 21st century has made it possible to take a big step towards the introduction of e-learning

tools using data mining algorithms. This article proposes a deep neural network (DNN) model for determining student performance. The conclusions obtained in the article give educational institutions a direction for adjusting educational programs. A comparison is made with already used machine learning algorithms that use the same data set as the proposed model.

Keywords: data classification, deep neural networks, machine learning

Актуальность проблемы

В рамках реализации государственной программы «Цифровая экономика РФ» доля представителей, использующих стандарты безопасного информационного взаимодействия государственных и общественных институтов, должна достигнуть 75% в период до 2021 года. Доля населения, обладающего цифровыми навыками, должна к 2021 г. составлять 40%. Также в рамках реализации программы планируется успешное функционирование не менее 10 отраслевых (индустриальных) цифровых платформ для основных предметных областей экономики (в том числе для цифрового здравоохранения, цифрового образования и «умного города») [8].

В реализующемся городском проекте «Московская электронная школа» консолидируются функции цифрового оборудования аудитории и образовательного учреждения с устройствами для решения основных образовательных задач. В рамках «Московской электронной школы» проводится формирование электронного рабочего места педагога с основными функциями классного журнала/дневника. Учителя и обучающиеся обеспечиваются устройствами для работы на уроках.

Содержательное ядро МЭШ [9] представлено на рис. 1.

При реализации программ обучения на базе МЭШ необходимы: эффективное распределение ресурсов, учёт интересов всех обучающихся, более гибкая система контроля знаний учащихся, эффективное использование ресурсов школы.

Рисунок 1 – Содержательное ядро МЭШ

В связи с вышесказанным актуальным становится вопрос разработки информационной системы, мотивирующей всех участников образовательных отношений к достижению высоких образовательных результатов и развитию компетенций, необходимых для экономики города.

Постановка задачи исследования

На основании вышеизложенного ставится задача классификации данных об успеваемости на основе данных из информационной системы «Московская электронная школа» с целью повышения эффективности обучения школьников.

Решение поставленной задачи сводится к выполнению ряда действий:

1. Определение наборов данных для анализа.
2. Сбор исходных данных.
3. Исследование исходной выборки данных для дальнейшего анализа.
4. Предварительная обработка исходных данных. Эти данные подготавливаются для последующего применения. Используя технологию машинного обучения, сеть обучается на основе данных, полученных на предыдущих этапах.
5. Сбор информации об успеваемости обучающегося по итогам выполнения контрольных и домашних работ, результатов экзамена, диагностик.
6. Сбор данных о посещении и пропусках аудиторных занятий.
7. Классификация обучающихся на основе вышеизложенных данных.

Методы решения проблемы

Использование методов машинного обучения позволяет своевременно выявить учащихся, у которых есть высокая вероятность получения низких результатов, с тем, чтобы учитель мог предоставить учащемуся возможность повысить свою эффективность обучения.

В настоящее время в образовательном интеллектуальном анализе широко используются алгоритмы машинного обучения, такие как Decision Tree [4] и Naive Bayes [3], данных. Существует ограничение для таких алгоритмов, как заявил Хаван Агравал [1], когда ввод данных в непрерывном диапазоне, согласно байесовской классификации, снижает точность моделей. Такая классификация лучше работает с дискретными данными. Также заявлено, что нейронная сеть выигрывает, если ей предоставляются непрерывные данные.

Глубокое обучение рассматривается как современный инструмент для исследования искусственного интеллекта, который применяется в различных приложениях [5]. Глубокое обучение представляет собой различные методы: глубокая нейронная сеть (DNN), рекуррентная нейронная сеть (RNN), Q-обучение. Также глубокое обучение в последнее время используется для распознавания голоса / звука [7], обработки естественного языка [3], компьютерного зрения [2].

В настоящей статье рассматривается модель классификатора Deep Neural Network (DNN) для прогнозирования успеваемости учащихся. Предложенная модель DNN направлена на то, чтобы с помощью анализа логистической классификации спрогнозировать, какие учащиеся потенциально могут попасть под категорию отказов. В сети реализовано два скрытых слоя, первый скрытый имеет функцию активации Relu, второй скрытый слой с функцией активации Soft-Max. Предложенная модель эффективна в прогнозировании и выявлении неэффективных учащихся с оценкой точности 85%.

Проблема может быть решена с помощью глубокой нейронной сети, где все элементы извлекаются и подаются на слои одновременно. После подачи нейрона функция активации проверит условие критерия и активирует нейрон,

как только он пройдет функцию. Правильная функция активации должна использоваться в каждом слое для активации правильного нейрона.

Методики и расчеты

Источник данных для построения предлагаемой глубокой нейронной сети для прогнозирования успеваемости учащихся МЭШ был получен из набора образовательных данных, собранных из системы управления обучением, называемой kalboard 360. Набор данных состоит из 500 записей учащихся. Он имеет 16 различных атрибутов (таблица 1).

Таблица 1

Структура данных обучающихся

Наименование	**Тип данных**	**Значения**
Пол	номинальный	2
Национальность	номинальный	14
Место рождения	номинальный	14
Этапы	номинальный	3
Оценки	номинальный	12
SectionID	номинальный	3
Тема	номинальный	12
ParentResponsible	номинальный	2
Семестр	номинальный	2
Поднятая рука	числовой	0-100
Посещенный ресурс	числовой	0-100
Просмотр объявления	числовой	0-100
Дискуссионная группа	числовой	0-100
Родительский ответ	номинальный	2
Удовлетворенность родителей	номинальный	2
Пропуски занятий	номинальный	2

Методы глубокого обучения нацелены на изучение взаимосвязи иерархий атрибутов с атрибутами более высоких уровней, которые формируются путем объединения других низших признаков. Модель с входными слоями,

произвольным количеством скрытых слоев и выходным слоем. Слои состоят из нейронов, которые имеют сходство с нейронами человеческого мозга.

Нейрон - это нелинейная функция, которая отображает входные векторы {11 … In) к выходу Y через взвешенный вектор {w1,, ..., w „) и к функции f, также известный как прямая связь [5, с. 110].

$$Y = f(\sum w_i I_I) = f(w^t I) \quad (1)$$

Цель модели - оптимизировать веса w таким образом, чтобы минимизировать квадратичную ошибку потерь. Это может быть достигнуто с помощью стохастического градиентного спуска (SGD). SGD итеративно обновляет весовой вектор, конечной целью которого является направление к минимальному градиенту функции потерь, чтобы получить уравнение обновления SGD (2):

$$w^{new} = w^{old-} n.(Y - t).Y(1 - Y).I \quad (2)$$

Глубокое обучение обладает отличной способностью к самообучению и самоадаптированию, что позволяет его всесторонне изучить и успешно использовать для решения сложных реальных проблем.

Набор данных, полученных в ходе проведенных вычислений, целесообразно разделить на три класса (таблица 2).

Таблица 2

Классы данных анализа результатов

Интервал	**Класс**
0-69	Низкий
70-89	Средний
90-100	Высокий

Предварительная обработка данных после сбора данных необходима для улучшения качества набора данных. Выбор атрибутов данных, очистка данных, преобразование данных и сокращение данных - все это часть предварительной обработки данных. Это часть процесса открытия знаний. Набор данных содержит 20 пропущенных значений в различных функциях, которые в результате очистки были удалены, и в итоге количество записей составило 480.

К набору данных также применяются различные преобразования. Атрибуты номинального типа данных Gender, Relation, Semester, ParentAnsweringSurvey, ParentSchoolSatisfacation и StudentAbsenceDays преобразуются в двоичные данные «0» и «1». Другие номинальные атрибуты типа данных Nationality, PlaceofBirth, StageID, GradeID, SectionID и Topic преобразуются в числовой тип данных.

После предварительной обработки записи делятся на две части: набор данных для обучения (обычно 90 – 95%) и тестирования (соответственно 5 – 10%). В обучающем наборе данных объекты и классы разделяются и хранятся в заполнителе тензорного потока. Обе записи классов обучающих наборов данных имеют горячее кодирование (One-Hot кодирование), это процесс, в котором переменные класса преобразуются в числовую форму, которая будет предоставлена модели глубокой нейронной сети для эффективного прогнозирования.

Таблица 3

Результаты One-Hot кодирования

Классы	One-Hot кодирование
Низкий	[1, 0 , 0]
средний	[0, 1 , 0]
Высоко	[0, 0 , 1]
средний	[0, 1 , 0]
Низкий	[1, 0 , 0]

Затем выходные данные передаются в функцию стоимости, где они сравниваются с фактическими результатами. Функция стоимости возвращает ошибку, которая передается в функцию оптимизации. Функция оптимизации обновляет вес слоев, так чтобы функция стоимости возвращала наименьшее значение ошибки. После построения графика потока данных для запуска должен быть активирован график статистических вычислений. Вычислительный график можно активировать с помощью сеанса тензорного потока. Далее осуществляется создание сеанса и передача входных данных в функцию run. В

данной модели задано 50 эпох, где вычислительный граф повторяется 50 раз, чтобы обеспечить более высокую точность.

Эксперимент проводился на операционной системе Ubuntu 16.04 с конфигурацией 8 ГБ ОЗУ и 4 ядрами Intel. Для запуска глубокой нейронной сети использовались такие инструменты, как python3 и тензор потока. Tensorboard и библиотека matplotlib использовались для визуализации внутренней работы модели. На выполнение программы ушло 5 минут.

В экспериментах использовались две меры для оценки качества классификатора: функция стоимости и точность. Целью использования точности является достижение более высокого значения, а целью функции стоимости является уменьшение значения.

С предложенной глубокой нейронной сетью удалось достичь максимальной точности 84,3%. Начальная точность модели составляла 29,8%. В первую эпоху наблюдалось повышение точности на 20%. Как только целевой результат не совпадает с выходным значением модели, вычисляется величина ошибки, с применением функции стоимости Soft-Max кросс-энтропии. Затем используется функция оптимизатора для обновления весов нейронов таким образом, что функция стоимости уменьшается. Хотя набор данных очень ограничен для глубокой нейронной сети, он все же превосходит другие алгоритмы машинного обучения. Поскольку набор данных ограничен, модель должна быть точно настроена для обеспечения лучшей производительности. Первоначально четыре скрытых слоя с каждым из 300 нейронов были установлены в графе потока данных, что не улучшило модель.

Благодаря оптимизации эта ошибка уменьшилась. Первая оптимизация резко снизила функцию стоимости. Для определения функции стоимости использовалась кросс-энтропия softmax. На 17-й итерации происходит увеличение стоимости из-за неэффективных обновлений градиента, но затем стоимость начинает уменьшаться с 34-й итерации. Всегда должно быть ограничение на количество эпох, поскольку из-за определенного предела оптимизатор начинает инвертировать веса, что приводит к увеличению

погрешности затрат. В предлагаемой модели после 50-й итерации происходит постоянное увеличение стоимости.

График используется для того, чтобы показать разброс показателей точности каждой перекрестной проверки в 10 раз для каждого алгоритма на рис. 2.

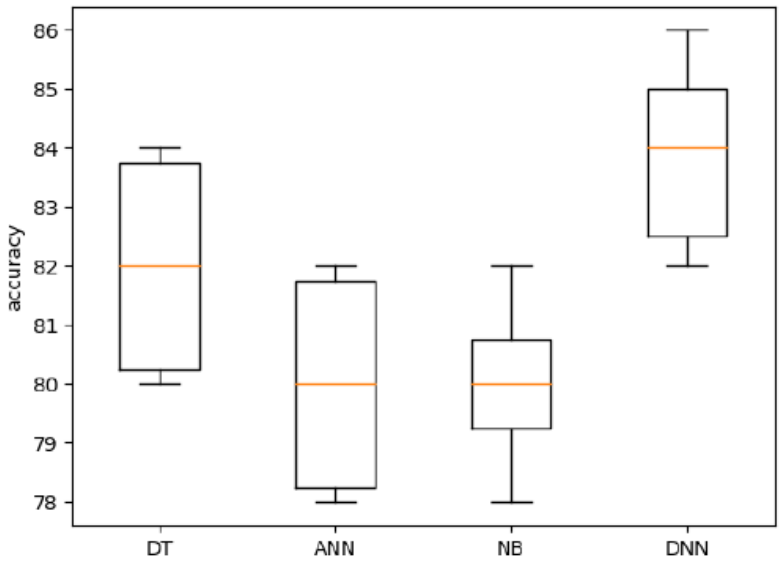

Рисунок 2 – Сравнение алгоритмов

Как видно из таблицы 4, Deep Neural Network может превзойти другие методы классификации машинного обучения даже с меньшим количеством данных, но необходимо точно настроить модель для повышения производительности.

Таблица 4

Сравнение методов классификации

Классификатор	Точность модели
Дерево решений (J48)	82.2
Искусственная нейронная сеть (ANN)	80.0
Байес (NB)	80.0
Предлагаемая модель (DNN)	84.3

Из 120 студентов из тестового набора 19 студентов были ошибочно классифицированы. Эта модель может быть достаточно надежной для прогнозирования успеваемости студентов.

Заключение

В этой статье предлагается модель глубокой нейронной сети для прогнозирования успеваемости учащихся на базе МЭШ. В ходе эксперимента стало понятно, что DNN может работать лучше даже при меньшем количестве данных, обладая глубокими знаниями о наборе данных и настройке качества модели. Предложенная модель достигла точности 84,3%. С большими записями и функциями набора данных DNN может достичь более высокой точности и превзойти другой алгоритм машинного обучения. Эта модель является надежной и может помочь в прогнозировании успеваемости учащегося и выявлении учеников, у которых больше шансов потерпеть неудачу в поиске решения проблемы.

Библиографический список

1. Агравал, Х. и Мавани, Х., 2015. В Прогнозирование успеваемости учащихся с использованием машинного обучения. Международный журнал инженерных исследований и технологий.

2. Амри Е.А., Хамтини Т., Алярах И., 2016. Использование образовательных данных для прогнозирования успеваемости учащихся с использованием ансамблевых методов. Международный журнал теории баз данных и приложений, – 9 (8), – с.119-136.

3. Джанг Х., Баэ С., 2017, февраль. Глубокие нейронные сети для прогнозирования транспортных потоков. Big Data и Smart Computing (BigComp), 2017 IEEE Международная конференция, с. 328-331.

4. ЛеКун Ю., Бенжио Ю., Хинтон Г., 2015. Глубокое обучение. Nature, – 521 (7553), – с. 436-444.

5. Ливирез и др. (2012): Прогнозирование успеваемости студентов с использованием искусственных нейронных сетей. 8-я Пан-эллинская конференция с международным участием Информационные и коммуникационные технологии, с.321-328.

6. Моукари С.Е., Хэйер М., Закхем В., 2016. Повышение успеваемости учащихся с помощью кластеризации данных и нейронных сетей в высшем образовании на иностранном языке. Исследовательский вестник Иордании ACM, – 2 (3), – с. 27-34.

7. Стэфпл М., Женг З., Пинкворт Н., 2016. Метод ансамбля для прогнозирования успеваемости учащихся в онлайн-среде обучения математике. – с. 231-238.

8. Сайт Государственной программы «Цифровая экономика РФ» http://government.ru/rugovclassifier/614/events/

9. Сайт Московской электронной школы (МЭШ) https://www.mos.ru/city/projects/mesh/

References

1. Agrawal, H. and Mawani, H., 2015. In Predicting student performance using machine learning. International Journal of Engineering Research and Technology.

2. Amri E.A, Khamtini T., Aljarah I., 2016. Using educational data to predict student performance using ensemble methods. International Journal of Database and Application Theory, – 9 (8), – pp. 119-136.

3. Jang H., Bae S., 2017, February. Deep neural networks for predicting traffic flows. Big Data and Smart Computing (BigComp), 2017 IEEE International Conference, – p. 328-331.

4. LeKun U., Benjio U., Hinton G., 2015. Deep learning. Nature, – 521 (7553), – p. 436-444.

5. Livirez et al. (2012): Predicting student performance using artificial neural networks. 8th Pan-Hellenic Conference with International Participation Information and Communication Technologies, – p. 321-328.

6. Moukari S.E., Heyer M., Zackham V., 2016. Improving student performance by clustering data and neural networks in higher education in a foreign language. Jordan Research Bulletin ACM, – 2 (3), – p. 27-34.

7. Staffl M., Zheng Z., Pinkworth N., 2016. Ensemble method for predicting student performance in an online math learning environment. from. 231-238.

8. The site of the State program «Digital Economy of the Russian Federation» http://government.ru/rugovclassifier/614/events/

9. Site of the Moscow Electronic School (MES) https://www.mos.ru/city/projects/mesh/