УДК 519.237.8

ПРИМЕНЕНИЕ ВРЕМЕННЫХ РЯДОВ И МАШИННОГО ОБУЧЕНИЯ ДЛЯ АНАЛИЗА ДАННЫХ ОБУЧАЮЩИХСЯ

Киселёв А.И.

Белорусский государственный университет информатики и радиоэлектроники, г. Минск, Беларусь, and.kis135@gmail.com

Аннотация. Рассмотрен процесс извлечения полезной информации из данных об успеваемости обучающихся, применения полученной информации для извлечения признаков и проведения дальнейшего анализа средствами машинного обучения.

Ключевые слова. Временные ряды, визуальный анализ, извлечение признаков, машинное обучение.

В современном образовательном процессе применение инновационных подходов играет ключевую роль в повышении качества обучения. В данной работе рассматривается использование машинного обучения для анализа данных обучающихся, особое внимание уделяется процессу извлечения признаков из данных об успеваемости, также оценивается возможность использования полученной информации для проведения визуального анализа.

Наиболее частой структурой данных, находящейся в электронных системах образовательных учреждений, является временной ряд оценок обучающихся в группах по предметам. Далее в работе будет рассматриваться возможность применения данной информации для улучшения образовательного процесса посредством его анализа.

Для дальнейшего удобства первым шагом в процессе анализа является обобщение ряда оценок путём их нормализации. Одним из возможных путей достижения этого является отображение временного ряда из исходного диапазона возможных оценок в диапазон [0, 1]. Пример нормализованных данных оценок обучающихся в группах по предмету представлен на рисунке 1.

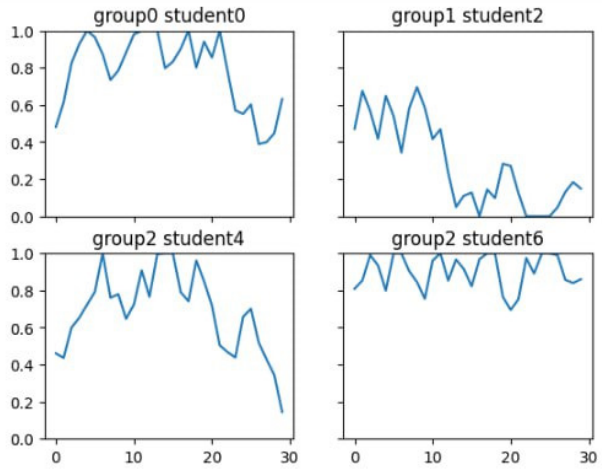

Рисунок 1 – Пример нормализованных данных

Для получение релевантных признаков из данных, необходимых для успешного анализа, необходимо понимать их внутреннюю структуру и потенциальные скрытые закономерности.

Так, например, полезным является временной ряд средних оценок в группе в определённый момент времени (за конкретное занятие или работу) , где – номер группы. Такой временной ряд может содержать информацию об условной сложности отдельно взятого задания. Элементы данного ряда могут быть рассчитаны по формуле (1).

$$C_{ij} = \frac{1}{N_i} \cdot \sum_{k=0}^{N_i} V_{ikj};$$

где N_i – количество студентов в группе i; V_{ikj} – нормализованная оценка студента k в группе i в момент времени j.

Полученные временные ряды C_i, для раннее представленных временных рядов нормализованных оценок, изображены на рисунке 2.

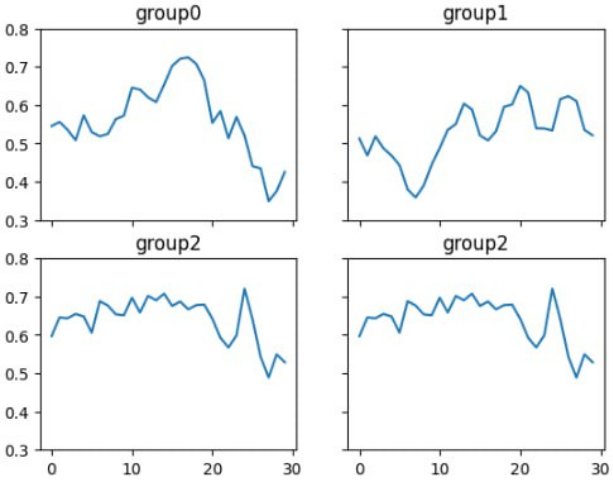

Рисунок 2 – Пример временных рядов

Временной ряд можно использовать для получения ряда , условного показателя успеваемости студента j в группе i, освобождённого от корреляции со сложностью заданий выдаваемых группе. Элементы данного ряда могут быть рассчитаны по формуле (2).

$$R_{ijk} = V_{ijk} - C_{ik};$$

где V_{ijk} – нормализованная оценка студента j в группе i в момент времени k; C_{ik} – средняя нормализованная оценка в группе i в момент времени k.

Ряд R_{ij} может быть далее разложен на тренд T_{ij}, который и будет наиболее точно представлять успеваемость обучающегося, и шум E_{ij}, выраженный в отклонении от тренда.

Для нахождение тренда можно использовать сглаживание методом экспоненциально взвешенно-

го скользящего среднего [1], при этом коэффициент α (сглаживающая константа) становится гиперпараметром модели. Шум E_{ij} является разностью между рядами R_{ij} и T_{ij} и, в контексте доменной области, представляет условное значение, насколько лучше (или хуже) обучающийся справился с заданием чем обычно.

На рисунке 3 синей сплошной линией представлен исходный ряд R_{ij} и его составные части: ряд T_{ij} оранжевой пунктирной линией и E_{ij} зелёной штрихпунктирной линией.

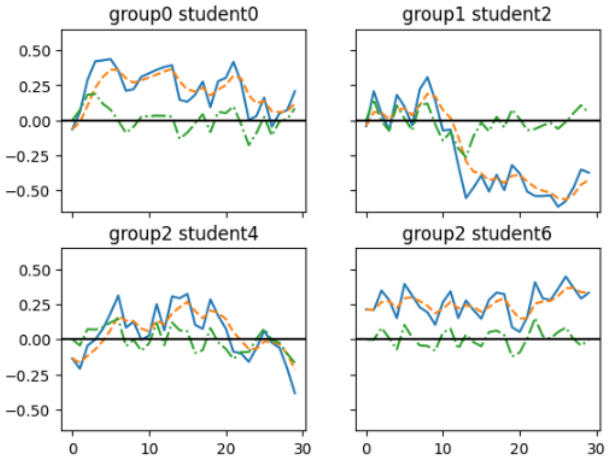

Рисунок 3 – Разложение ряда

На текущий момент из данных об оценках обучающихся в группах по предметам получены ряды C_i, R_{ij}, T_{ij} и E_{ij}, смысловая интерпретация которых была рассмотрена ранее. Визуальный анализ диаграмм данных временных рядов может быть использован для оценки качества образовательного процесса, изучения эффективности вносимых в него изменений, а также разработке индивидуального подхода к обучающимся.

В качестве примера возможностей использования полученных рядов для анализа данных о студентах средствами машинного обучения рассмотрим решение задачи классификации.

Для верификации, посредством модели случайного блуждания, были созданы временные ряды оценок для 12 групп из 20 обучающихся, содержащие 30 значений в каждом.

Каждый из обучающихся в группе, исходя из комплексной зависимости, включающей среднюю оценку обучающегося, медианное значение оценки по отношению к группе, изменение тренда оценок к концу занятий, был отнесён к одному из двух классов: условно «сдача» и «несдача» итогового экзамена.

Из каждого из рядов C_i, T_{ij} и E_{ij} для созданных данных, как из рядов, содержащих полную информацию об исходных данных и имеющих наименьшую корреляцию между друг другом, были извлечены следующие признаки: среднее значение, медианное значение, автокорреляция с лагом 1, среднеквадратическое отклонение, изменение среднего [2].

Полученные 240 маркированных наборов (15 признаков в каждом) данных об обучающихся были разделены на тренировочную и тестовую выборку.

На тренировочной выборке был обучен классификатор метода опорных векторов [3]. Характеристики качества классификации, полученные путём верификации модели на тестовой выборке представлены в таблице 1.

Таблица 1 – Характеристик качества классификации

Характеристика	Значение
Accuracy	0.87
Precision	0.77
Recall	0.96
F1 Score	0.85

Полученные характеристики качества модели показывают возможность применения описанных методик разложения временных рядов успеваемости обучающихся с последующем извлечением признаков для осуществления анализа данных с использованием инструментов машинного обучения. Описанный метод может быть применён для мониторинга и корректировки образовательного процесса с целью его улучшения посредством анализа и выявления скрытых закономерностей в данных обучающихся.

Литература

1. EWMA Control Charts [Электронный ресурс] – Режим доступа: https://www.itl.nist.gov/div898/handbook/pmc/section3/pmc324.htm.

2. Extracted features [Электронный ресурс] – Режим доступа: https://tsfresh.readthedocs.io/en/latest/text/list_of_features.html.

3. SVM [Электронный ресурс] – Режим доступа: https://scikit-learn.org/stable/modules/svm.html.

APPLICATION OF TIME SERIES AND MACHINE LEARNING FOR STUDENT DATA ANALYSIS

A.I. Kiselev

Belarusian State University of Informatics and Radioelectronics, Minsk, Belarus, and.kis135@gmail.com

Abstract. The process of extracting useful information from student performance data, applying the obtained information to generate features, and conducting further analysis using machine learning tools is examined.

Keywords. Time series, visual analysis, feature extraction, machine learning.