УДК 519.67 +004.852

Оценка информативности признаков в модели прогнозирования трудоустройства выпускников

Н.В. Смыкова, А.С. Авдеев, М.В. Гунер

Алтайский государственный технический университет им. И.И. Ползунова (Барнаул, Россия)

Selection of Informational Signs for Forecasting Employment of Graduates

N.V. Smykova, A.S. Avdeev, M.V. Guner

Polzunov Altai State Technical University (Barnaul, Russia)

Статья посвящена решению проблемы формирования пространства информационных признаков для модели прогнозирования трудоустройства выпускников по их успеваемости. Сформировано несколько вариантов признакового пространства с использованием разных методов автоматизированного отбора информационных признаков. Под оценкой информативности признаков в данной работе понимается определение степени их влияния на уровень потенциала трудовой активности выпускников вуза и соискателей работы. Эффективность методов автоматизированного отбора информационных признаков определялась эмпирическим путем: на сформированных выборках проводились эксперименты по обучению с использованием градиентного бустинга и случайных лесов.

Наибольшая точность прогноза была достигнута при формировании признакового пространства с помощью одномерного отбора и обучения методом случайных лесов. Полученные результаты могут использоваться при разработке автоматизированной информационной системы оценки потенциалов трудовой активности выпускников вузов и соискателей работы.

Ключевые слова: отбор информативных признаков, градиентный бустинг, случайный лес, метод главных компонент.

The article is devoted to solving the problem of forming the space of information signs for the model of forecasting the employment of graduates by their progress. Several variants of the recognition space were formed using different methods of automated selection of information signs. The evaluation of the informativeness of the signs is understood as determining the extent of their impact on the level of the potential of labor activity of graduates of the university and job applicants.The effectiveness of the methods of automated selection of information features was determined empirically: on the generated samples, training experiments were carried out using gradient boosting and random forests.The greatest accuracy of the forecast was achieved in the formation of the recognition space by means of one-dimensional selection and training by the method of random forests. The results obtained can be used in the development of an automated information system for assessing the potential of work activity of graduates and job applicants.

Key words: selection of informative features, gradient boosting, random forest, principal component method.

DOI 10.14258/izvasu(2020)1-22

Введение. В настоящее время рекрутерами применяются разные методы профессионального отбора персонала. Кандидатов, которые уже имеют опыт работы в других компаниях или выполнения каких-либо проектов, можно достаточно объективно оценить анализом резюме и портфолио. Задача усложняется, если трудоустраивается выпускник вуза без профессионального опыта. В таком случае возможно использовать сопоставление успеваемости выпускников с результатами их трудоустройства. При решении этой задачи важную роль играют вузовские информационные системы, которые позволяют формировать базы данных о студентах на протяжении всего времени обучения в университете [1–3].

Для получения высокого результата обучения модели важно качественно подготовить исходные дан-

ные, в том числе выделить наиболее информативные признаки. Под оценкой информативности признаков понимается определение степени их влияния на уровень потенциала трудовой активности выпускников вуза и соискателей работы.

В настоящее время существует ряд алгоритмов, которые позволяют провести анализ данных, а также сократить их размерность. Одни и те же алгоритмы на разных наборах данных могут иметь большие погрешности в результатах, поэтому предлагается формировать информационное пространство признаков для каждого конкретного случая экспериментальным путем.

В данной статье решается задача формирования пространства информационных признаков для модели прогнозирования трудоустройства выпускников на основе их успеваемости в учебном процессе и проводятся эксперименты по обучению уточненных моделей.

Описание данных и используемых информационных технологий

В качестве исходных данных используются сведения об успеваемости и трудоустройстве выпускников направления «Прикладная информатика» Алтайского государственного технического университета им. И.И. Ползунова за период 2014–2018 гг.

Входными переменными выступили данные об успеваемости студента по 40 профильным дисциплинам. Балльные оценки успеваемости общим числом 305 располагаются в диапазоне 0–100. Выходная переменная — результат трудоустройства. Множество значений результатов трудоустройства были приведены к нескольким классам, и использована шкала от 0 до 8: 0 — не работает; 1 — работает не по специальности; 2 — программист; 3 — программист 1С; 4 — веб-разработчик; 5 — тестировщик; 6 — аналитик; 7 — специалист техподдержки, консультант; 8 — менеджер, работа с клиентами.

В работе использовались язык программирования Python, библиотеки Pandas, Numpy, Scikit-learn, Mglearn.

Предварительный анализ данных

Для проведения анализа данных применялся метод главных компонент [4], позволяющий анализировать корреляционные связи в данных [5–8]. При анализе использовались две главных компоненты. На диаграмме рассеяния (рис. 1) по главным компонентам представлены данные о трудоустройстве выпускников. В рассматриваемом двумерном пространстве эти классы разделены недостаточно хорошо, и для решения задачи классификации недостаточно использования линейных классификаторов.

Отбор признаков. Выделение из массива исходных данных наиболее информативных признаков велось по следующим методам: одномерный отбор признаков, рекурсивное исключение признаков, отбор с использованием ансамблевых алгоритмов на основе деревьев решений.

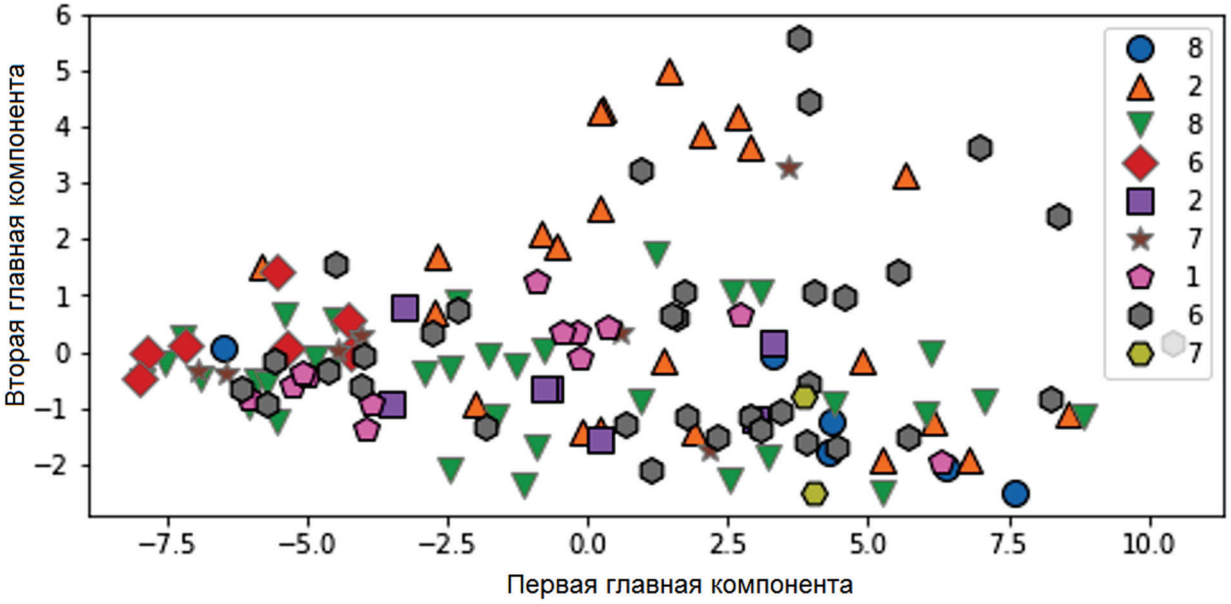

Рис. 1. Двумерная диаграмма рассеяния для набора данных

Одномерный отбор признаков заключается в том, что признаки, имеющие наиболее выраженную взаимосвязь с целевой переменной, могут быть отобраны с помощью статистических критериев [7–8]. Для каждого из входных параметров рассчитывается оценка, характеризующая информативность признака (табл. 1). Наиболее высокие оценки получили следующие дисциплины: основы инновационной и инвестиционной деятельности, объектно-ориентированное программирование, 1-я и 2-я производственные практики, информационные системы и технологии.

Таблица 1

Фрагмент результатов оценки информативности входных параметров

Шифр	Наименование дисциплины	Оценка
X1	1-я производственная практика	0,2930
X2	2-я производственная практика	0,2314
X3	WEB-программирование	0,0990
X4	Автоматизированный бухгалтерский учет	0,0492
X5	Базы данных	0,1714
X6	Бухгалтерский учет	0,0657
X7	Вычислительные системы, сети и телекоммуникации	0,0622
X8	Дискретная математика	0,1050
X9	Иностранный язык	0,1927

При рекурсивном исключении признаков происходит обучение модели на исходном наборе признаков и оценивается их значимость, затем исключается один или несколько наименее значимых признаков, модель обучается на оставшихся признаках и т.д. [9–11]. В ходе применения этого метода каждому признаку присваиваются оценка и значение True или False (табл. 2). На используемом наборе данных наиболее информативными признаками оказались сведения об успеваемости по дисциплинам: иностранный язык, объектно-ориентированное программирование, основы инновационной и инвестиционной деятельности, офисные информационные технологии, управление информационными системами в экономике.

Таблица 2

Фрагмент оценки информативности методом исключения признаков

Шифр	Наименование дисциплины	Оценка	
X1	1-я производственная практика	16	False
X2	2-я производственная практика	7	False
X3	WEB-программирование	15	False
X4	Автоматизированный бухгалтерский учет	6	False
X5	Базы данных	19	False
X6	Бухгалтерский учет	28	False
X7	Вычислительные системы, сети и телекоммуникации	30	False
X8	Дискретная математика	33	False
X9	Иностранный язык	1	True

Отбор с использованием ансамблевого алгоритма на основе деревьев решений [3, 10] показал наиболее информативными признаками данные об успеваемости по дисциплинам: основы инновационной и инвестиционной деятельности, информационная безопасность, объектно-ориентированное программирование, математические методы в экономике, банковские информационные системы. В таблице 3 приведен фрагмент результата применения данного метода.

Таблица 3

Фрагмент оценки информативности на основе важности признаков

Шифр	Наименование дисциплины	Оценка
X1	1-я производственная практика	0,02431
X2	2-я производственная практика	0,03219
X3	WEB-программирование	0,0296
X4	Автоматизированный бухгалтерский учет	0,01986
X5	Базы данных	0,02628
X6	Бухгалтерский учет	0,02776
X7	Вычислительные системы, сети и телекоммуникации	0,02757
X8	Дискретная математика	0,01863
X9	Иностранный язык	0,02914

Эксперименты по обучению модели прогнозирования трудоустройства

На основании данных отбора информативных признаков были сформированы новые обучающие и тестовые выборки. Далее производилось обучение вариантов моделей прогнозирования трудоустройства полученных выборок с использованием разных методов. Наиболее эффективными показали себя методы случайного леса и градиентного бустинга [3, 4, 10, 12]. На тестовых выборках большее количество правильных ответов было получено при использовании метода случайного леса (рис. 2).

Рис. 2. Результаты обучения выборок с применением метода случайного леса

Стоит отметить, что стабильный положительный результат достигался на выборке, для которой информативные признаки были выбраны с использованием одномерного отбора. Выборка, сформированная данным способом, обучалась лучше и при использовании градиентного бустинга (рис. 3).

Рис. 3. Результаты обучения выборок с применением градиентного бустинга

Исследование показало, что метод рекурсивного исключения признаков является наименее эффективным при отборе информативных признаков в данном примере.

Заключение. В настоящей работе рассмотрена задача формирования пространства информационных признаков для модели прогнозирования трудоустройства выпускников вуза на основании данных об их успеваемости по профильным дисциплинам. Сформировано несколько вариантов признакового пространства с использованием разных методов автоматизированного отбора информационных признаков.

Эффективность методов определялась эмпирическим путем: сформированные выборки обучались с использованием градиентного бустинга и случайных лесов. Наибольшая точность прогноза была достигнута при формировании признакового пространства с помощью одномерного отбора и обучения методом случайных лесов.

Библиографический список

1. Мутанов Г.М., Мамыкова Ж.Д., Бобров Л.К. Роль и место дата-центра в ИТ-инфраструктуре университета // Вестник НГУ. Серия : Информационные технологии. 2014. Т. 12. Вып. 2.

2. Федотов А.М. и др. Концептуальная модель научно-образовательной информационной системы // Вестник НГУ. Серия : Информационные технологии. 2015. Т. 13. Вып. 3.

3. Смыкова Н.В., Авдеев А.С., Томашев М.В. Отбор информативных признаков для прогнозирования трудоустройства выпускников вуза // Наука и бизнес: пути развития. 2018. № 12 (90).

4. Паклин Н.Б., Орешков В.И. Бизнес-аналитика: от данных к знаниям. СПб., 2013.

5. Мюллер А. Введение в машинное обучение с помощью Python : руководство для специалистов по работе с данными. М., 2017.

6. Николенко С., Кадурин А., Архангельская Е. Глубокое обучение. СПб., 2018.

7. Таскин А.С. Тест обобщающей способности линейных методов прогнозирования // Вестник НГУ. Серия : Информационные технологии. 2013. Т. 11. Вып. 2.

8. Анисимов Д.С., Рязанов М.А., Шаповал А.И. Подход к обработке многомерных данных пептидных микрочипов // Известия Алт. гос. ун-та. 2015. № 1/2(85).

9. Ketkar N. Deep Learning with Python: A Hands-on Introduction. Bangalore, Karnataka, India. 2017.

10. Scikit-learn. Machine Learning in Python. URL: http://scikit-learn.org/stable.

11. Пальчунов Д.Е., Яхъяева Г.Э. Нечеткие логики и теория нечетких моделей // Алгебра и логика. 2015. Т. 54. № 1.

12. Большакова Е.И. и др. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика. М., 2011.