УДК 004.891.2
DOI 10.17513/snt.39611

МАШИННОЕ ОБУЧЕНИЕ В СИСТЕМАХ ОЦЕНИВАНИЯ ПРИГОДНОСТИ СОИСКАТЕЛЕЙ ДЛЯ ВАКАНСИЙ

Забержинский Б.Э., Золин А.Г., Козлов В.В.

ФГБОУ ВО «Самарский государственный технический университет», Самара,
e-mail: zab.borislav@gmail.com, zolin.a.g.@gmail.com, vco2005@mail.ru

Подбор подходящего кандидата на определенную должность – это ответственный и интенсивный процесс, с которым сталкиваются многие компании. Трудоустройство подходящего кандидата вызывает трудности у различных организаций, поскольку они предъявляют множество конкретных требований, упомянутых в должностной инструкции. Использование искусственного интеллекта позволяет измерить, предсказать и отобрать подходящего кандидата на основе базы данных резюме и требований должности. Данные разбиваются на четыре кластера, которые базируются на первичных и вторичных навыках, а также прилагательных и наречиях. Между этими кластерами измеряется сходство по Жаккару, и на основе параметров кластера предлагается мера пригодности. С помощью трех классификаторов линейной регрессии, дерева решений, Adaboost и XGBoost выполняется прогноз пригодности кандидата. Для оценки кандидата используются различные классификаторы, которые помогли выполнить прогнозирование метрик на очень высоком уровне. Исследовав проблему, мы можем смело утверждать, что в скором времени подобные системы будут использоваться нанимающим персоналом, что позволит компаниям нанимать на работу более квалифицированные кадры при минимизации периода поиска кандидатов. Эксперименты по прогнозированию проводятся и оцениваются с помощью пятикратной перекрестной проверки. Линейная регрессия дала среднюю степень классификации 85,60%, а максимальной точности достигает классификатор XGBoost с показателем 95,14%.

Ключевые слова: прогнозирование пригодности, приобретение талантов, профили, искусственный интеллект, качество, классификаторы

MACHINE LEARNING IN JOB SUITABILITY ASSESSMENT SYSTEMS

Zaberzhinskiy B.E., Zolin A.G., Kozlov V.V.

Samara State Technical University, Samara,
e-mail: zab.borislav@gmail.com, zolin.a.g.@gmail.com, vco2005@mail.ru

The selection of a suitable candidate for a certain position is a responsible and intensive process that many companies face. The employment of a suitable candidate causes difficulties for various organizations, since they make specific requirements mentioned in the job desciption. The use of artificial intelligence allows you to measure and predict a suitable candidate based on a resume database and job requirements. The data is divided into four clusters, which are based on primary and secondary skills, as well as adjectives and adverbs. The similarity in Jacquard is measured between these clusters, and a fitness measure is proposed based on the cluster parameters. With the help of three linear regression classifiers, a decision tree, Adaboost and XGBoost, the candidate's fitness prediction is performed. To evaluate the candidate, various classifiers are used, as well as the "bag of words" technique. Prediction experiments are conducted and evaluated using 5-fold cross-validation. Linear regression gave an average classification degree of 85.60%, and the maximum accuracy of the XGBoost classifier, which is 95.14%.

Keywords: suitability measurement, talent acquisition, profiles, artificial intelligence, quality, classification

Новейшие технологии внесли радикальные изменения в практику управления человеческими ресурсами (HR) [1]. Подключение к интернету открыло множество возможностей как для лиц, ищущих работу, так и для работодателей. Размещение вакансий на различных платформах, таких как порталы вакансий, социальные сети и веб-сайты собственных компаний, привлекает множество соискателей. Специалист по подбору персонала сталкивается с огромными трудностями при тщательном изучении соответствующих профилей среди множества претендентов. Этот процесс требует дополнительных затрат на HR-сотрудников и времени обработки на редкие и «труднодоступные» вакансии организации. В последние годы многие рутинные и сложно формализуемые проблемы стали доверять методам искусственного интеллекта, которые нашли множество применений в различных сферах человеческой жизни, и главным достоинством подобного применения является обработка большого количества однотипных данных и поиск в них нужных закономерностей.

Искусственный интеллект можно использовать для прогнозной аналитики, которая предполагает принятие готового решения на основе существующих данных [2]. Методы искусственного интеллекта, такие как генетический алгоритм (GA) и искусственная нейронная сеть (ANN), используются для разработки комбинированной модели прогнозирования, также данные методы широко применяются в архитектуре безопасности интернета вещей с помощью блокчейна. Технологии искусственного интеллекта (AI) показывают высокую эффективность в подборе и управлении

персоналом. Контент-анализ показывает, что в организациях, где используются методы AI, эффективность процесса найма значительно возрастает, что видно в уменьшении «текучести» кадров и повторного размещения похожих вакансий.

1. Постановка задачи и анализ набора данных

В этом разделе проведем анализ набора данных, содержащего резюме соискателей и компании работодателей. В наборе данных насчитывается около 15 000 резюме соискателей и восемь вакансий от работодателей. На рис. 1 приведено несколько примеров вакансий (RV), таких как веб-разработчик, системный администратор Linux, разработчик на языке C и инженер облачных сервисов. Для работодателей в наборе данных представлены различные сведения, такие как название должности, сведения о компании и городе. В колонках «Описание» и «Обязанности» подробно описываются различные задачи, которые должны выполняться в рамках данной должности, а требуемые навыки указаны в разделе «Предпочтительные навыки». Столбец NaN говорит об отсутствии конкретных требований.

Например, обязанности веб-разработчика состоят в разработке и верстке веб-сайтов с использованием JavaScript, а от образования требуется окончание бакалавриата по любой специальности из области компьютерных наук.

На рис. 2 показано несколько примеров резюме (RC) для различных должностей в наборе данных. Этот набор данных включает различную информацию о кандидатах, такую как название резюме, город, описание работы, опыт работы на этой должности, сведения об образовании, навыках кандидатов и проведенной ими сертификации.

На рисунке приведены резюме на различные инженерные должности и профили в области компьютерных наук, когнитивной автоматизации и машинного обучения.

На рис. 3 показан график частоты встречаемости названий должностей, присутствующих в наборе данных резюме (CR). Есть около 2100 кандидатов с резюме на должность разработчика программного обеспечения. Второе место по количеству соискателей занимают 1310 веб-разработчиков, за которыми следуют специалисты в области машинного обучения в количестве 1148, что является частью набора данных CR.

В базе данных насчитывается 94 кандидата в специалисты по обработке данных, за которыми следует почти 50 кандидатов в системные администраторы. В наборе данных очень мало кандидатов с заголовком резюме «системные администраторы».

	Должность	Компания*	Город	Описание	Обязанности	Образование	Предпочтительные навыки
0	Веб-разработчик	A	Москва	NaN	Разработка веб-скриптов JavaScript	Степень бакалавра в области компьютерных наук	JavaScript, Python
1	Системный администратор Linux	B	Казань	NaN	Решение проблем, связанных с системой Linux	Степень бакалавра в области компьютерных наук	NaN
2	C-разработчик	C	Новосибирск	NaN	Проектирование на языке C++	Степень бакалавра в области компьютерных наук	Linux, Veritas Volume Manager
3	Инженер облачных сервисов	D	Санкт-Петербург	Хорошо владеет Microsoft, а также является экс...	Планирование и внедрение облачной инфраструкту...	Степень бакалавра в области компьютерных наук	AWS, DevOps

Рис. 1. Примеры RV из набора данных

	Резюме	Город	Описание	Опыт работы	Образование	Навыки	Сертификаты	Доп. информация
0	Devops инженер	Казань	Внутренние задачи организации	Разработка веб-сайтов с использованием HTML	Степень бакалавра в области компьютерных наук	HTML, CSS, PHP	NaN	CRUD API in PHP
1	Облачный архитектор	Москва	Сертификация AWS на протяжении 7 лет	Интеграция с AWS	Степень бакалавра в области компьютерных наук	Python	Решение AWS	Работал на производстве
2	Devops инженер	Санкт-Петербург	Работа в ИТ-секторе	Управление облаками Python	Степень бакалавра в области компьютерных наук	C++, HTML	NaN	NaN
3	Devops инженер	Новосибирск	NaN	Веб-разработчик	Степень бакалавра в области компьютерных наук	C, C++, PHP	Mahindra pride class	Языки: C++, веб-разработка, PHP
4	Когнитивная автоматизация	Москва	Датчики давления окружающей среды	Когнитивная автоматизация	Степень бакалавра в области компьютерных наук	Автоматизация	NaN	NaN
5	Машинное обучение	Москва	Уверенное программирование на Python	Проектирование	Степень бакалавра в области компьютерных наук	Анализ данных и машинное обучение	NaN	NaN

Рис. 2. Примеры RC из набора данных

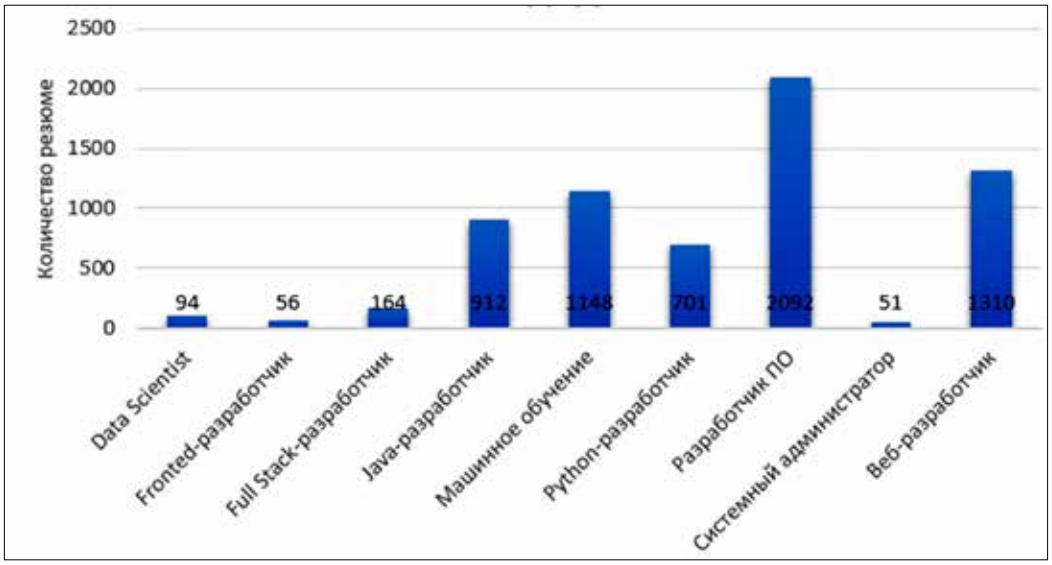

Рис. 3. График частоты упоминания должностей у разных кандидатов

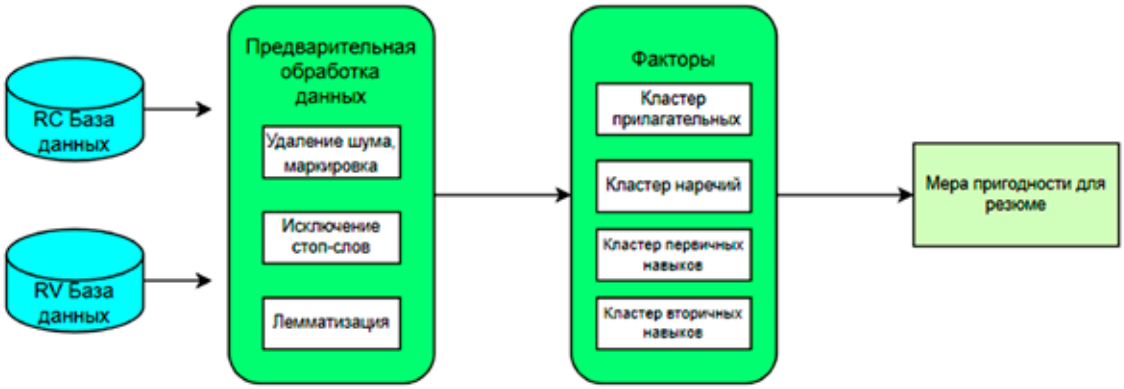

Рис. 4. Архитектура измерения пригодности резюме

2. Измерение и прогнозирование пригодности

В рамках исследования была разработана модель для поиска наиболее подходящего кандидата в отношении работодателя (JD). Для вычисления показателя пригодности резюме используются различные методы искусственного интеллекта, такие как NLP, кластеры и измерение расстояния [3]. Разработанная архитектура для измерения пригодности резюме показана на рис. 4, где на первом шаге считываются RC и RV, присутствующие в наборе данных.

В RC и RV выполняются различные этапы предварительной обработки. Процесс токенизации проводится на основе сбора должностей и полученного списка слов из резюме. Затем к каждому CR и JD применяется «шумоподавление». Выполняется исключение стоп-слов, и собираются важные слова, после чего применяется лемматизация. Лемматизация определяется на базе коренных слов из словаря [4]. Части речи, предложения, лексемы CR определяются с помощью набора инструментов естественного языка NLTK.

Используя ссылки на части речи и детали, приведенные в наборе данных, формируются четыре кластера слов как в RC, так и в RV. RC и RV содержат четыре важных информационных источника (первичные и вторичные навыки, а также прилагательные и наречия), которые наиболее важны и чувствительны при проверке резюме. Первичные и вторичные навыки выражают набор навыков, необходимых для работы, в то время как их функциональные свойства описываются прилагательными и наречиями в профилях. Кластеры первичных

и вторичных навыков создаются на основе сведений, доступных в RC и RV.

Пригодность измеряется между RV и RC с использованием сходства Жаккара между четырьмя кластерами. В общем случае сходство Жаккара для двух документов Doc_A и Doc_B, содержащих предложения и слова, определяется как

$$J\left(Doc_A, Doc_B\right)=\frac{Doc_A \cap Doc_B}{Doc_A Doc_B}. \quad (1)$$

Мы определяем меру подобия Жаккара между кластером RC и RV, как указано в (2). Для кластера RC_C и RV_C коэффициент подобия Жаккара задается как

$$J\left(RC_C, RV_c\right)=\frac{RC_C \cap RV_C}{RC_C RV_C}. \quad (2)$$

Сходство кластеров по Жаккару – это отношение количества общих слов к общему количеству слов в этих кластерах [5]. Сходство по Жаккару между кластером первичных навыков и кластером вторичных навыков составляет $J(RC_{PS}, RV_{PS})$ и $J(RC_{SS}, RV_{SS})$ соответственно. Затем та же методика используется для вычисления показателя пригодности между CR_S и RV_S. Предлагается следующее уравнение для вычисления показателя пригодности:

$$\text{Suitability} = J\left(RC_{PS}, RV_{PS}\right)+J\left(RC_{SS}, RV_{SS}\right)+ + J\left(RC_{Adj}, RV_{Adj}\right)*\left|RC_{Adj}\right|. \quad (3)$$

Здесь RC_{PS} и RC_{SS} – это совокупность первичных и вторичных навыков для RCs. RV_{PS} и RV_{SS} – это кластеры начальных и вторичных навыков для RV. RC_{Adj} – это группа прилагательных в резюме кандидата.

JD_{Adj} представляет собой группу прилагательных в RV_S. $|RC_{Adj}|$ – количество слов, присутствующих в группе прилагательных в RC_S. Третье слагаемое в (3) умножается на $|RC_{Adj}|$, чтобы пропорционально увеличить его вес на количество прилагательных в RC_S.

Измерение пригодности вычисляется между RV_S и RC_S с использованием уравнения (2) для всего набора данных. На выходе мы имеем прогнозную модель измерения пригодности соискателя на конкретную должность.

3. Экспериментальные исследования

Как было показано выше, в этом исследовании эксперименты проводились на репрезентативной выборке в 15 тыс. резюме. Как описано в разделе 2, четыре кластера сформированы из первичных навыков, вторичных навыков, прилагательных и наречий из RC и RV. В табл. 1 показаны кластеры прилагательных, подготовленные с использованием нескольких RC_S и RV_S. В строке 1 показана группа прилагательных для RV:1. В строках 2, 3 и 4 приведены группы прилагательных из трех резюме RC:1604, RC:1667 и RC:1721. Сходство по Жаккару между RV:1 и RC:1604 составляет 0,2857, что не является эталонным результатом.

Таблица 1

Группы прилагательных для резюме

RC/RV	Кластер прилагательных
RV:1	аналитический, соответствующий, глубокий, точная настройка, фреймворки, необходимый, плюс, решение проблем, статистический
RC:1604	построение, клиническое, соразмерное, полное, клиентское, основанное на данных, лишенное, эффективное, хорошее, в магазине, интеллектуальное, крупномасштабное, логичное, крупное, много, среднее, основанное на ml, потенциальное, прогнозирующее, первичное, реальное, находчивое, ответственное, розничная торговля, несколько, общительный, подходящий, широкий
RC:1667	модифицированный, новый, оперативный, Python, тщательный
RC:1721	приложения, серверная часть, сборка, разные, эффективные, инновационные, внутренние, основные, медицинские, новые, онлайн, общие, частные, программы, прогрессивные, проверенные, масштабируемые, оговоренные, технические, полезные, ориентированные на пользователя, визуализация, специализированные
RV:2	алгоритмический, аномальный, ранний, извлечение, генерация, человеческий, нейронный, основанный на паттернах, потенциальный, реальный
RC:1603	клиентский, глубокий, динамичный, вовлекающий, обширный, будущий, индивидуальный, необходимый, организационный, конкретный, реальный, статистический, достаточный, контролировать
RC:1609	текущий, далекий, межличностный, ярлык, механический, потребности, научный, технический, сквозной, различный
RC:1820	ежедневный, конечный, передний, полный, полный спектр, жизненный цикл, множественный, ответственный, богатый, древовидный, разнообразный, веб

Аналогичным образом также показано сходство Жаккара между RV:1 и RC:1667 и RC:1721. Наибольшее сходство по Жаккару, равное 0,5608, получено между RV:2 и RC:1609. Предлагаемая мера пригодности вычисляется с использованием сходства Жаккара между четырьмя кластерами RV_S и RC_S из уравнения (3).

Резюме классифицируются на три класса: наиболее подходящий ($НБП_S$), умеренно подходящий ($УМП_S$) и неподходящий ($НП_S$) – на основе оценки пригодности, чтобы облегчить менеджерам быстрое принятие решения в процессе отбора резюме. Значение пригодности выше 0,6 считается классом $НБП_S$. RC:1721 – это НБП для RV:1 со значением пригодности 1,881. Значение пригодности менее 0,1 рассматривается как $НП_S$ для соответствующего RV_S. RC:2907 – это $НП_S$ для RV:4.

Процент RC_S, имеющих класс $НБП_S$, в нашем наборе данных составляет 23,5 %, в то время как процент профилей $УМП_S$ составляет 23,4 %. В нашем наборе данных 53,2 % резюме классифицируются как НП.

В этом исследовании прогнозирование RC на три подходящих класса осуществляется с использованием классификаторов на основе искусственного интеллекта, а именно линейной регрессии, дерева решений, классификаторов Adaboost и XGBoost [6]. Эти классификаторы обучаются на основе набора слов, собранного из каждого RC, для выполнения классификации по трем классам. Производительность классификатора проверяется при пятикратной перекрестной проверке (табл. 2).

Таблица 2

Точность классификаторов

Классификатор	Точность классификатора, %
Линейная регрессия	85,60
Дерево решений	94,47
Adaboost	94,78
XGBoost	95,14

Минимальный средний коэффициент классификации 85,60 % наблюдается для линейных методов Adaboost и XGBoost. Улучшение качества классификации наблюдается для таких классификаторов, как дерево решений, методы Adaboost и XGBoost. Максимальный средний коэффициент классификации для XGBoost составляет 95,14 %.

Заключение

Для успешного привлечения наиболее подходящих кадров необходимо определить и выбрать из множества резюме того кандидата, который наиболее точно вписывается в видение HR для данной должности, что является сложной задачей из-за огромного количества данных, связанных с ними. В данной работе предлагается система на основе методов искусственного интеллекта, которая сгруппировала кандидатов в четыре кластера на основе их первичных навыков, вторичных навыков, прилагательных и наречий. Также было разработано измерение пригодности, основанное на подобии Жаккара, для оценки соответствия кандидатов требованиям вакансии. Результаты исследования показывают возможность добиться уровня классификации более 95 %. В будущем предлагается задействовать функции социальных сетей для формирования дополнительных кластеров кандидатов и улучшения классификации.

Список литературы

1. Устинова Л.Н., Аракелова А.О. Технологии управления человеческими ресурсами на основе цифрового подхода // π-Economy. 2021. № 14 (6). С. 40–52.

2. Белых Т.И., Бурдуковская А.В. Использование способа реализации искусственного интеллекта в прогнозировании // Известия Байкальского государственного университета. 2018. № 28 (3). С. 500–507.

3. Долгодворова Е.В. Кластерный анализ: базовые концепции и алгоритмы // Вопросы науки и образования. 2018. № 7 (19). С. 73–76.

4. Жердева М.В., Артюшенко В.М. Стемминг и лемматизация в lucene. Net // Лесной вестник. 2016. № 20 (3). С. 131–134.

5. Смирнов А.А., Салып Б.Ю. Анализ программных моделей для определения меры смысловой близости предложений естественного языка // StudNet. 2022. № 5 (5). С. 3498–3508.

6. Моршин А.В. Глубинное машинное обучение // Известия Тульского государственного университета. Технические науки. 2019. № 3. С. 270–273.