АДАПТАЦИЯ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ В ПРАКТИЧЕСКИХ ВОПРОСАХ АНАЛИЗА РЫНКА ТРУДА

Харичкин А.К.[1]

Ключевые слова: *человеческий капитал, HR-аналитика, машинное обучение, кластеризация, иерархическая кластеризация.*

Аннотация

Цель статьи: построение прикладной модели кластерного анализа в отношении вопросов оценки человеческого капитала, основанной на принципах глубокой аналитики из области машинного обучения.

Методы: общенаучные методы анализа и синтеза, математическое и компьютерное моделирование.

Результат: сформирована эффективная аналитическая структура для анализа рынка труда, в частности для вопросов, связанных с оценкой человеческого капитала. В ходе моделирования были адаптированы данные лонгитюдного опросника молодого населения ТрОП, проводимого в период 2011—2023 годов. На его базе сформирован и проанализирован алгоритм агломеративной иерархической кластеризации, выбор которого обусловлен категориальным характером большинства используемых показателей и малым размером набора данных. По результатам валидации модель показала высокое качество анализа и подтвердила возможность использования на первоначальных этапах изучения явлений рынка труда.

Практическая ценность: сформирован эффективный алгоритм, нацеленный на выведение эвристических зависимостей и оценку актуальных тенденций на рынке труда. В рамках этого выделяются несколько заинтересованных сторон. Предприятия реальной экономики получают возможность обоснованного анализа целевых групп при найме и управлении собственными ресурсами. Для исследовательских организаций и государственных институтов конкретизируются важные аспекты развития рынка труда, способствующие в формировании актуальных теоретических моделей и релевантных мер поддержки определенных социальных групп. По выделенным в итоге исследования классам приведены обобщающие характеристики и предложены рекомендации для дальнейшего практического анализа.

DOI: 10.21681/1994-1404-2024-2-102-107

Введение

Управление человеческими ресурсами в организации представляет один из наиболее важных процессов, обеспечивающих стабильное осуществление основной деятельности. В связи с этим работа сотрудников, связанных с HRM (от англ. human resource management — управление человеческими ресурсами), как и для всех подразделений отдельного предприятия, должна соответствовать тенденциям экономики и основываться на операционной структуре, гибкой в отношении значимых изменений рынка. В частности, особую роль эти положения играют в проведении аналитических процедур, направленных на исследование ключевых трендов и важных для бизнеса зависимостей. Тем не менее реальная ситуация отличается от этой модели. На настоящий момент бóльшая часть компаний продолжает полагаться на привычные методы анализа, делая при этом упор исключительно на статичные модели, дескриптивный анализ и обзор исторических данных без учета аспектов будущих явлений. По результатам обзора рынка в 2022 году, более 35% организаций не применяли средства автоматизации HR-процессов[2]. Остальная часть использовала их преимущественно в подборе персонала, расчете трудовых компенсаций и кадровом учете. Только 26% адаптировали такие инструменты для развития сотрудников на рабочем месте, 18% — для оценки эффективности. В целом можно отметить, что степень приложения низкая.

[2] Как компании используют HR-аналитику: исследование hh.ru // hh.ru: офиц. сайт. 22.07.2022. URL: https://clck.ru/39sSem (дата обращения: 30.04.2024).

[1] ***Харичкин Алексей Кириллович,*** *студент Инженерной академии Российского университета дружбы народов, г. Москва, Российская Федерация.*

E-mail: 1132223295@pfur.ru

Аналогичная ситуация также справедлива конкретно в области HR-аналитики. Что особенно актуально в рамках текущей работы, методы углубленного и предиктивного анализа находят применение лишь в каждой десятой организации. Такое положение может иметь ряд негативных последствий для бизнеса с точки зрения потенциала развития. Как отмечается в исследовании СберЗдоровья и Atsearch Group, на настоящий момент 51% предприятий осознают возможные провалы найма и удержания специалистов, поэтому ставят приоритетной задачей цифровизацию HR-процессов[3]. По заключению экспертов Gartner, от общего количества опрошенных менеджеров в области управления человеческими ресурсами 76% из них выделяют одну из ключевых ролей в трендах 2024 года внедрению подходов автоматизации на основе искусственного интеллекта [1, 2, 3].

В этой статье мы рассмотрим один из подходов к проведению аналитики в HRM — кластерный анализ. Более конкретно, внимание будет уделено алгоритмам машинного обучения, направленным на эту задачу, и тому, какие результаты могут получены на их основе.

Дескриптивный анализ данных

Для начала опишем информационную базу. При поддержке Программы фундаментальных исследований НИУ ВШЭ в работе адаптировались данные лонгитюдного опросника «Траектории в образовании и профессии» (далее — ТрОП), проводимого с 2011 по 2023 год[4]. В первой волне участниками являлись ученики школ восьмых классов. Реализация проекта продолжалась на протяжение всего периода получения ими образования и первых лет работы. Выборка для текущего исследования ограничена двумя волнами 2011 и 2019 годов. В этих временных промежутках индивиды детально описывали подробности воспитательных процессов, обучения в школе, качества образовательной среды, уровень полученного образования, имеющийся опыт работы и иные показатели накопленного капитала, повлиявшие на их личностное становление. Финальная структура данных выражена статичным набором, исключающим динамический компонент. Мы не рассматриваем изменения во времени ввиду особенностей самого опросника, так как в нем отсутствует связь между периодами по важным для нас признакам. Для задачи кластеризации размер выборки составил 1553 наблюдения. В процессе очищения набора были исключены опрошенные, не ответившие на вопросы по уровню образования матери или отца, а также с отсутствием данных по достатку семьи, в которой проходило воспитание.

В моделировании мы использовали следующие переменные:

1. Пол респондента, бинарная (0 — женщина, 1 — мужчина).
2. Наличие высшего образования у отца, бинарная (0 — нет, 1 — есть).
3. Наличие высшего образования у матери, бинарная (0 — нет, 1 — есть).
4. Достаток семьи, в которой проходило воспитание индивида, категориальная (1: менее 20 тыс. руб., 2: 20—29 тыс. руб., 3: 30—49 тыс. руб., 4: 50—79 тыс. руб., 5: более 80 тыс. руб. в месяц).
5. Наивысший полученный уровень образования респондента, категориальная, а именно: 1 — 9 классов школы, 2 — 11 классов школы, 3 — среднее профессиональное образование, 4 — бакалавриат, 5 — специалитет, 6 — магистратура).
6. Опыт работы, числовая дискретная (в годах).
7. Участие в мероприятиях по повышению профессиональной квалификации, бинарная (0 — нет, 1 — есть).
8. Среднее количество отработанных часов за неделю, числовая дискретная (в часах).

Следует также сделать оговорку, что мы заранее убираем из рассмотрения возраст человека, так как для текущей базы его вариация крайне незначительна. Большинство индивидов принимали участие в опросе в одном возрасте. Что касается иных признаков, опишем коротко каждый из них.

Сначала затронем общие характеристики. Так, в выборке — 44% респондентов мужского пола. В среднем для 34% из них родители имеют оконченное высшее образование, причем для матерей доля составляет 38%, для отцов — на 8 п. п. меньше, 30%. По медианному значению достатка семьи 50% выборки имели среднемесячный доход 20—29 тыс. руб. В четвертом квартиле распределены опрошенные с показателем от 50 тыс. руб. и более.

Теперь обратимся к профессиональным индикаторам. Средний и медианный уровни образования индивидов — среднее профессиональное. При этом большая часть опрошенных никогда не посещала профессиональные курсы: из общего количества только 26%. Тем не менее по изначальному предположению данный признак значимо воздействует на карьерный потенциал специалиста. По опыту работы, что вполне ожидаемо для выбранной социальной группы, средняя величина достаточно низка, 1 год. Однако вместе с тем стандартное отклонение варьируется около 1,2 года. Количество отработанных за неделю часов по медиане стандартно и соответствует норме занятости по работе по найму, 40 часов.

Спецификация алгоритма

Теперь перейдем к спецификации методологии. Говоря о самой модели кластеризации, можно отметить,

[3] Приоритеты, эффективность HR-процессов и забота о сотрудниках: HR 2023—2024 // ЭтСерч: офиц. сайт. 06.12.2023. URL: https://clck.ru/3Axc6h (дата обращения: 30.04.2024).

[4] Траектории в образовании и профессии // Институт образования НИУ ВШЭ: офиц. сайт. URL: https://clck.ru/3AtABY (дата обращения: 05.05.2024).

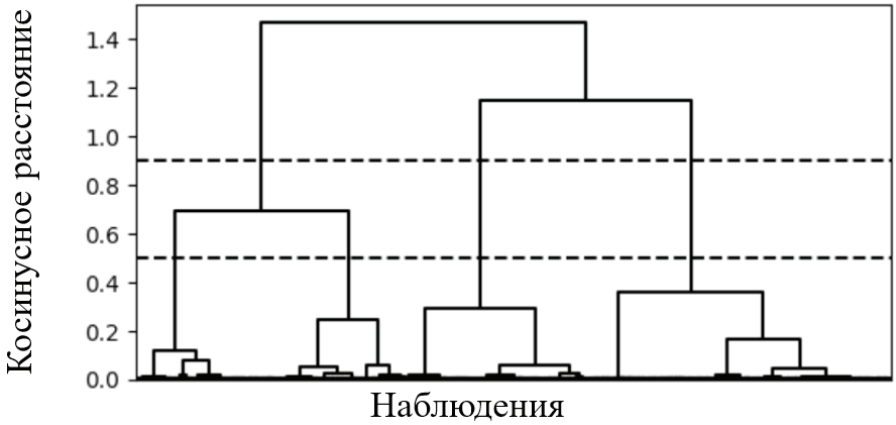

Рис. 1. Дендрограмма модели

что данный тип методов характеризуется процессом обучения без заранее заданных ответов, то есть без учителя. В большинстве случаев их использование нацелено на промежуточное исследование данных, после которого адаптируются конкретизирующие алгоритмы регрессионного или иных типов. В нашем случае будет разобран пример такой спецификации и возможные рекомендации по применению полученных результатов. При этом основной целью мы ставим формирование эффективного метода кластерного моделирования в отношении рынка труда и выявление при этом самих классов наблюдений с выработкой первоначальных рекомендаций для заинтересованных сторон государства, предприятий реальной экономики и самого человека.

В статье мы сконцентрируем внимание на агломеративной иерархической кластеризации, являющейся одним из популярных подходов к данной задаче [4]. Она обладает рядом преимуществ относительно иных видов моделей и эффективно адаптируется к наборам данных малого и среднего размера, что релевантно для данной статьи. Отличительной особенностью этого подхода является восходящий процесс образования классов, при котором каждое наблюдение изначально считается отдельной группой. В процессе объединения вычисляются метрики близости объектов, на основе которых формируются общие кластеры с ближайшими соседями. Модель заканчивает действие при достижении заданного количества кластеров или при объединении всей выборки. В нашем случае этот параметр будет задан конечным числом. Сто́ит также заметить, что алгомеративная кластеризация работает исключительно на основе переданной ей матрицы расстояний без опоры на внутренние зависимости между признаками.

Говоря о процессе агломерации, затронем выбор метрики для вычисления расстояний в пространстве выборки. Наиболее распространенными вариантами расстояний на практике можно назвать следующие: евклидово, манхеттенское, Минковского, косинусное подобие и др. [5]. На основе них применяется сам алгоритм расчета близости. Среди способов расчета также есть различные критерии, разница между которыми заключается в выборе объектов кластеров для попарного вычисления и виде связей между ними. Некоторые из них: метод Уорда, полная связь, средняя связь, единичная связь [6].

Мы не будем углубляться в подробности каждой метрики и метода и коротко опишем только те, которые используем в модели. Мы будем опираться на косинусное расстояние и метод средней связи. Мера косинусного сходства часто применяется при анализе данных в формате изображений и текстов. Она оценивает расстояние между двумя ненулевыми векторами на базе расчета косинуса угла между ними. Значения при этом могут варьироваться от -1 до 1. 1 соответствует максимальной степени сходства, -1 — минимальной. Что важно, ввиду методики вычисления косинусное расстояние эффективно рассчитывается в отношении категориальных признаков, преобладающих в нашей выборке. Что касается вида связи, метод средней основывается на усреднении расстояний между всеми парами элементов кластеров. Дополнительно отметим, что для нивелирования мультиколлинеарности данных размерность пространства признаков была уменьшена до двух с помощью метода главных компонент [7].

Теперь перейдем к определению количества классов для модели. В иерархической кластеризации, как и в некоторых других подходах, этот гиперпараметр является важным элементом, который задает сам аналитик. Оценим его, используя дендрограмму (рис. 1).

Эта диаграмма показывает, каким образом происходит объединение объектов в кластеры по мере увеличения порога метрики расстояния. Выбор числа классов основывается на вертикальном отрезке (т. е. величине расстояния) с наибольшей длиной. При пересечении горизонтальной линией его и параллельных ему других отрезков мы можем наглядно вывести потенциальные варианты для искомого параметра. В данном случае проведены две горизонтальные линии для отрезка в середине диаграммы. Отсюда видно, что выбор стоит между тремя и четырьмя кластерами. Далее будем работать именно с четырьмя, так как даже при таком решении можно заметить, что деление на подвыборки происходит достаточно четко с большими

Таблица 1

Результаты кластеризации по четырем группам

Кластер	Объем	Общие параметры	Специальные параметры
0	621	- Низкий семейный достаток - Низкая образованность родителей	- Низкая образованность - Много опыта
1	255	- Семейный достаток выше среднего - Средняя образованность родителей	- Низкая образованность - Много опыта
2	375	- Низкий семейный достаток - Средняя образованность родителей	- Высокая образованность - Мало опыта
3	302	- Высокий семейный достаток - Высокая образованность родителей	- Высокая образованность - Мало опыта

средними расстояниями между ними. Далее перейдем к самому моделированию и представим его результаты (табл. 1).

В приведенной таблице отражены агрегированные характеристики по каждому найденному кластеру. Отсюда видно, что наблюдения распределились со смещением в отношении 0 кластера: в него отнесена бóльшая часть объектов выборки. По остальным группам группировка сравнительно равномерна. Качество алгоритма оценим по двум метрикам: коэффициенту силуэта и статистике Дэвиса-Булдина. Значение первого составило 0,452. Это свидетельствует о высокой способности модели к выделению отдельных классов в нашем наборе данных и показывает возможность его использования для анализа. Статистика Дэвиса-Булдина отражает меньшее качество со значением 0,858. Однако сама статистика не является достаточной для заключения о потенциале использования и должна быть рассмотрена совместно с иными критериями. Таким образом, мы делаем вывод о хорошем качестве модели. Далее опишем результаты оценки.

Обсуждение

По итогу моделирования мы получили качественное разделение изначальной выборки из 1553 молодых специалистов на 4 отдельных кластера, имеющих значимые различия между собой по отдельным показателям собственного капитала. Для начала обратимся к общим характеристикам. По каждой из них 1 и 3 кластеры имеют более высокие показатели относительно 0 и 2. По уровню ежемесячного достатка семьи средние значения для них выше на 30 тыс. руб. Высшее образование имеют отцы 39% респондентов 1 кластера и 52% респондентов 3 кластера. Во 2 кластере — 33%, в 0 — только 14%. Аналогичное соотношение между группами по образованию матери.

Профессиональные признаки имеют иное распределение. Наиболее выделяющимися по параметру опыта работы являются 0 и 1 кластеры. В них индивиды имеют на 0,5 года больше опыта, чем во 2 и 3, с соответствующими значениями 1,2 и 1,3 года. В отношении образования ситуация противоположная: каждый индивид из 2 и 3 групп имеет оконченное высшее образование, в то время как для 0 и 1 можно говорить только об 11 классах школы или среднем профессиональном образовании. Курсы по развитию профессиональных компетенций посещают 30% опрошенных из каждого кластера за исключением 0, в нем доля на 8 п. п. меньше и составляет 22%.

В целом, несмотря на то что мы не проводим дальнейших детальных исследований в данной работе, по результатам алгоритма можно сделать логичные выводы.

3 кластер представляет специалистов с наибольшим уровнем накопленного капитала. В отсутствие значимого профессионального опыта они обладают высоким уровнем квалификации и имеют потенциал для ее эффективного приложения в выбранной отрасли. С точки зрения предприятий данную группу стоит рассматривать в первую очередь на позиции с высокими требованиями по имеющимся компетенциям. Сами индивиды должны стремиться к участию в реальной экономической деятельности для совершенствования и актуализации своих возможностей.

1 кластер имеет более низкий потенциал как по общим характеристикам, так и в отношении образования. Однако его представители более опытны на рынке труда, из-за чего в некоторых случаях могут быть более конкурентоспособны для определенных позиций и компаний. Тем не менее повышение уровня образования, включая дополнительные курсы, может улучшить капитал кластера.

2 кластер, как и 3, имеет высокий уровень образования, но его эффект понижается в связи с низкими общими характеристиками и опытом работы. Последним по уровню развития человеческого капитала является 0 кластер — по данным модели, единственным преимуществом опрошенных является наличие профессионального опыта. В отношении таких групп населения государству стóит осуществлять активную поддержку

в повышении профессиональных компетенций, включая широкодоступные инициативы вне основного образования, предоставлять возможности для получения реального опыта и, вероятно, вводить меры дополнительной материальной помощи.

Заключение

В заключение статьи еще раз отметим, что проведение глубокой аналитики рынка труда представляет собой крайне важную задачу для современного рынка. Учитывая волатильность большинства тенденций, исследования необходимо поддерживать и актуализировать на регулярной основе. В этой работе мы показали, как возможно адаптировать при этом один из методов кластерного анализа, а именно — агломеративную иерархическую кластеризацию. При этом мы подтвердили адекватность ее применения и получили результаты, для которых предложили потенциальные рекомендации.

Заметим также, что выполненное моделирование является лишь промежуточной мерой. На последующих шагах аналитику необходимо формировать иные модели и прибегать к более детальным подходам, основываясь на полученных ранее предварительных выводах. В частности, такими вариантами в нашем случае могут быть индивидуальные регрессионные модели для каждого класса с прогнозированием целевой метрики потенциала. Такой метрикой может выступать размер заработной платы или другой показатель. Сто́ит также адаптировать иные виды кластеризации, такие как алгоритмы нечеткой логики или вероятностные вариации, предназначенные для выявления пороговых значений и диапазонов вероятности, при которых объект может быть отнесен к кластеру.

Литература

1. Whittle L. M. What must HR leaders prioritize in 2024? // UNLEASH: official site. 15.12.2023. URL: https://clck.ru/3AxcP7 (дата обращения: 05.05.2024).
2. Карцхия А. А., Макаренко Г. И. Правовые проблемы применения искусственного интеллекта в России // Правовая информатика. 2024. № 1. С. 4—19.
3. Карцхия А. А., Макаренко Г. И. Правовые горизонты технологий искусственного интеллекта // Вопросы кибербезопасности. 2024. № 1 (59). С. 2—14.
4. Abualigah L., Agushaka J.O., Akinyelu A.A. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects // Engineering Applications of Artificial Intelligence. 2022. Vol. 110. Pp. 165–173.
5. Mercioni M. A., Holban S. A survey of distance metrics in clustering data mining techniques // Proceedings of the 3rd International Conference on Graphics and Signal Processing. 2019. Pp. 44–47.
6. Murtagh F., Contreras P. Algorithms for hierarchical clustering: an overview // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2012. Vol. 2. № 1. P. 86–97.
7. Abdi H., Williams L. J. Principal component analysis // Wiley interdisciplinary reviews: computational statistics. 2010. Vol. 2. № 4. P. 433–459.

ADAPTING CLUSTERING ALGORITHMS IN PRACTICAL QUESTIONS OF LABOUR MARKET ANALYSIS

Aleksei Kharichkin, *student at the Academy of Engineering of the Peoples' Friendship University of Russia, Moscow, Russian Federation.*
E-mail: 1132223295@pfur.ru

Keywords: *human capital, HR analytics, machine learning, clustering, hierarchical clustering.*

Abstract
Purpose of the paper: building an applied cluster analysis model for assessing human capital based on principles of deep analytics from the machine learning field.

Methods used in the study: general scientific methods of analysis and synthesis, mathematical and computer modelling.

Study findings: an efficient structure for analysing the labour market was set up, in particular for questions of assessing human capital. In the course of modelling, data from the TrEP longitudinal questionnaire of the young population carried out in 2011 to 2023 were adapted. Based on it, the agglomerative hierarchical clustering algorithm was set up and analysed, it was chosen due to the categorial nature of the majority of used indicators and a small amount of the available dataset. The validation results showed a high quality of analysis performed by the model and confirmed the feasibility of using it at the initial stages of studying the labour market phenomena.

Practical value: an efficient algorithm was set up aimed at determining heuristic dependencies and assessing current tendencies at the labour market. Several interested parties can be identified here. Real economy enterprises get the possibility for a well-grounded analysis of target groups in hiring personnel and managing their own resources. For research organisations and government institutions, important aspects of labour market development are outlined which help forming topical theoretical models and relevant support measures for certain social groups. For classes identified in the study, generalised characteristics and recommendations for further practical analysis are given.

References

1. Whittle L. M. What must HR leaders prioritize in 2024? UNLEASH: official site. 15.12.2023. URL: https://clck.ru/3AxcP7 (data obrashcheniia: 05.05.2024).
2. Kartskhiia A. A., Makarenko G. I. Pravovye problemy primeneniia iskusstvennogo intellekta v Rossii. Pravovaia informatika, 2024, No. 1, pp. 4–19.
3. Kartskhiia A. A., Makarenko G. I. Pravovye gorizonty tekhnologii iskusstvennogo intellekta. Voprosy kiberbezopasnosti, 2024, No. 1 (59), pp. 2–14.
4. Abualigah L., Agushaka J.O., Akinyelu A.A. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. Engineering Applications of Artificial Intelligence. 2022. Vol. 110. Pp. 165–173.
5. Mercioni M. A., Holban S. A survey of distance metrics in clustering data mining techniques. Proceedings of the 3rd International Conference on Graphics and Signal Processing. 2019. Pp. 44–47.
6. Murtagh F., Contreras P. Algorithms for hierarchical clustering: an overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2012. Vol. 2. No. 1. P. 86–97.
7. Abdi H., Williams L. J. Principal component analysis. Wiley interdisciplinary reviews: computational statistics. 2010. Vol. 2. No. 4. P. 433–459.