«Системні технології» 3 (134) 2021 «System technologies»
DOI 10.34185/1562-9945-3-134-2021-14
УДК 004.912

О.В. Гавриленко, В.А. Дворник

ЗАСТОСУВАННЯ МЕТОДІВ КЛАСТЕРИЗАЦІЇ ДЛЯ ВИЗНАЧЕННЯ СФЕР ДІЯЛЬНОСТІ КАНДИДАТІВ ПРИ ПІДБОРІ КАДРІВ ДЛЯ ІТ-КОМПАНІЙ

Анотація. У даній статті розглянуто проблему підбору кадрів для ІТ-компаній. Розглянуто практичне застосування методів кластеризації на прикладі задачі визначення сфер діяльності кандидатів та підбору резюме на певну вакансію при підборі кадрів для ІТ-компаній. Приведено постановку задачі. Визначено, до яких моделей зводяться досліджувані проблемні ситуації, та, які методи можуть бути застосовані до розв'язання поставленої задачі. Наведено огляд відомих рішень, також висвітлено переваги та недоліки обраного методу. Приведено приклад визначення кількості оптимальних кластерів для заданого набору резюме.
Ключові слова: кластеризація, оптимізація, сфера діяльності, кандидат, рекрутер, підбір кадрів.

Постановка проблеми. На сьогоднішній день підбір підходящих кандидатів для найму з широкого кола кандидатів є основоположним питанням. Традиційними методами є проведення індивідуальних перевірок і різних технічних кваліфікаційних тестів, співбесід, а також групових обговорень. Виявлення здібностей кандидата за допомогою співбесід є традиційною практикою в процесі найму [1]. Проте ця традиційна практика займає дуже багато часу, а також може привести до несправедливого вибору кандидатів.

Сучасні менеджери з управління персоналом та кадровики повинні обробляти надзвичайно великі обсяги даних: дослідження портфоліо, скринінг соціальних медіа, ідентифікація наборів навичок, а також – дослідження резюме. Тому обрати правильного кандидата може бути важкою місією.

Отже, призначенням дослідження є спрощення процесу відбору кадрів для HR-менеджерів на прикладі ІТ-компаній з використанням резюме, в яких враховуються інформація про професійні та особисті якості претендента.

Аналіз останніх досліджень і публікацій. Автоматизовану класифікацію резюме за допомогою техніки кластеризації розглядали також професори Сагар Море, Бхамаре Приянка, Малі Пуджа та Качаве Каляні [2]. Вчені наголошують на тому, що на сьогоднішній день важким завданням для

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

менеджерів з підбору персоналу є завдання знайти найкращого кандидата який би відповідав усім побажанням та виправдовував усі очікування. Пропонується метод, що дозволяє відповідним визначити особливості та навички у кожному з надісланих резюме.

Цей підхід застосовує ідею кластеризації. На простому рівні кластеризація використовує один або кілька атрибутів в якості основи для ідентифікації кластера. Кластеризація корисна для ідентифікації різної інформації, так як вона корелює з іншими прикладами, так що можна побачити, де подібності та діапазони збігаються.

Пропоноване вченими рішення використовує методи інтелектуального аналізу даних. Метод кластеризації інтелектуального аналізу даних використовується для класифікації та розрахунку. Оскільки кластеризація корисна для ідентифікації різної інформації, так як вона корелює з іншими прикладами, так що можна бачити, де подібності та діапазони збігаються. Для кластеризації у системі застосовується алгоритм k-means. Цей алгоритм кластеризує резюме кандидатів у k кластерів [2].

Мета досліджень. Метою статті є дослідження методів кластеризації та перетворення задачі кластеризації на задачу оптимізації для підвищення ефективності та якості рекомендацій менеджерам з підбору персоналу.

Викладення основного матеріалу досліджень. Підбір персоналу – це процес, при якому HR-менеджер зазвичай визначає та залучає потенційних людей ззовні та зсередини організації, щоб оцінити їх та прийняти на певну посаду. У задачі визначення сфер діяльності працівників при підборі кадрів для IT-компаній в якості вхідної інформації розглядатимуться резюме у текстовому вигляді, в яких буде міститись уся інформація про професійну кар'єру працівника, а також мотиваційні листи, есе та тести з професійної орієнтації з вільними відкритими відповідями.

На виході отримаємо набір професійних сфер діяльності працівників з підібраними до них найкращими резюме, тобто, отримаємо групування вхідних даних до певних сфер діяльності та на певну вакансію.

Для групування та об'єднання вхідних даних пропонується залучити методи кластеризації текстових даних.

Таким чином, нехай $\mathbf{X}$ – множина об'єктів, тобто резюме, а $\mathbf{Y}$ – множина кластерів, тобто професійних сфер. Задана функція відстані між об'єктами $\rho(x, x')$. Маємо кінцеву навчальну вибірку об'єктів:

$$\mathbf{X}^{\mathbf{m}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{\mathbf{m}}\} \subset \mathbf{X} \quad (1)$$

Необхідно розбити вибірку на підмножини (кластери), тобто кожному об'єкту $x_i \in X^m$ поставити у відповідність $y_i \in Y$ таким чином щоб об'єкти всередині кожного кластера були близькі щодо метрики ρ , а об'єкти з різних кластерів істотно розрізнялися [3].

Розглянемо методи кластеризації. Кластеризація – це поділ множини вхідних векторів на групи (кластери) за ступенем «схожості» один на одного. Для того, щоб можна було порівнювати два об'єкти, потрібно мати критерій, на підставі котрого і буде відбуватися порівняння. Зазвичай, як правило, таким критерієм є відстань між об'єктами [3].

Найпопулярнішим алгоритмом кластеризації є алгоритм k-means. Проте, на жаль, алгоритм k-means не справляється із задачею, коли об'єкт не належить жодному кластеру або належить до різних кластерів у однаковій мірі.

З цією проблемою k-means чудово справляється алгоритм с-середніх (c-means). Замість точної відповіді на запитання до якого кластеру відноситься об'єкт, алгоритм визначає ймовірність належності об'єкту до того чи іншого кластеру. Таким чином, твердження вигляду «об'єкт В належить до кластеру 1 з імовірністю 90%, до кластеру 2 – 15%» вірне і набагато зручніше.

Алгоритм c -середніх (c -means) – це модифікація методу k -means. Далі наведено кроки роботи алгоритму [4]:

1. Вибір початкового нечіткого розбиття n об'єктів на k кластерів шляхом вибору матриці належності U розміром $n \times k$.

2. Визначення значення критерію нечіткої похибки алгоритму із застосуванням наступної матриці:

$$E^2(X,U) = \sum_{i=1}^{N} \sum_{k=1}^{K} U_{ik} \left\| x_i^{(k)} - \mu_k \right\|^2, \quad (2)$$

де μ_k – це «центр мас», тобто, центроїд нечіткого кластера k .

3. Перестановка (перегрупування) об'єктів із метою зменшення нечіткої помилки.

4. Перехід до п. 2 до тих пір, поки зміни матриці U не стануть незначними.

Застосовування алгоритму c-means може бути недоцільним, якщо число кластерів заздалегідь невідоме або є необхідність віднесення кожного об'єкту до певного кластеру однозначно.

Далі наведено переваги та недоліки методу c -means [5]:

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Переваги:

– можливість визначення ступеня приналежності елемента до кластеру;

– нечіткість при віднесення об'єкта до кластеру дозволяє включати об'єкти, які знаходяться на границі, в кластери.

Недоліки:

– число кластерів повинно бути відоме заздалегідь;

– метод зазвичай шукає кластери сферичної форми;

– комплікативність роботи з об'єктами;

– обчислювальна складність.

Так як одним із недоліків методу є необхідність знати кількість кластерів наперед, необхідно вказати точно визначене число. В даному випадку пропонується представити задачу кластеризації як оптимізації.

На відміну від завдання класифікації або регресії, в разі кластеризації складніше вибрати критерій, за допомогою якого було б просто уявити завдання кластеризації як задачу оптимізації.

У алгоритмах k-means та c-means поширений наступний критерій – сума квадратів відстаней від точок до центроїдів кластерів, до яких вони відносяться, повинна бути мінімальною.

$$J(C) = \sum_{i=1}^{K} \sum_{i \in C_k} \left\| x_i^{(k)} - \mu_k \right\|^2 \to \min_C, \tag{3}$$

де C – множина кластерів потужності K, μ_k - центроїд кластера C_k.

Зрозуміло, що у цьому є певний сенс: необхідно, щоб точки розташовувалися купчасто біля центрів своїх кластерів. Взагалі, мінімум такого функціоналу буде досягатися тоді, коли кластерів стільки ж, скільки і точок (тобто кожна точка - це кластер одного елемента).

Для вирішення цього питання (вибору числа кластерів) необхідно скористатися такою евристикою: обирають саме те число кластерів, починаючи з якого описаний функціонал падає «вже не так швидко». Або більш формально [6]:

$$D(k) = \frac{|J(C_k) - J(C_{k+1})|}{|J(C_{k-1}) - J(C_k)|} \to \min_k$$

На рис. 1 показано експерименти для k = 25, 50 та 80 кластерів. По осі OY – залежність суми квадратів відстаней від точок до центроїдів кластерів, по осі OX – кількість кластерів.

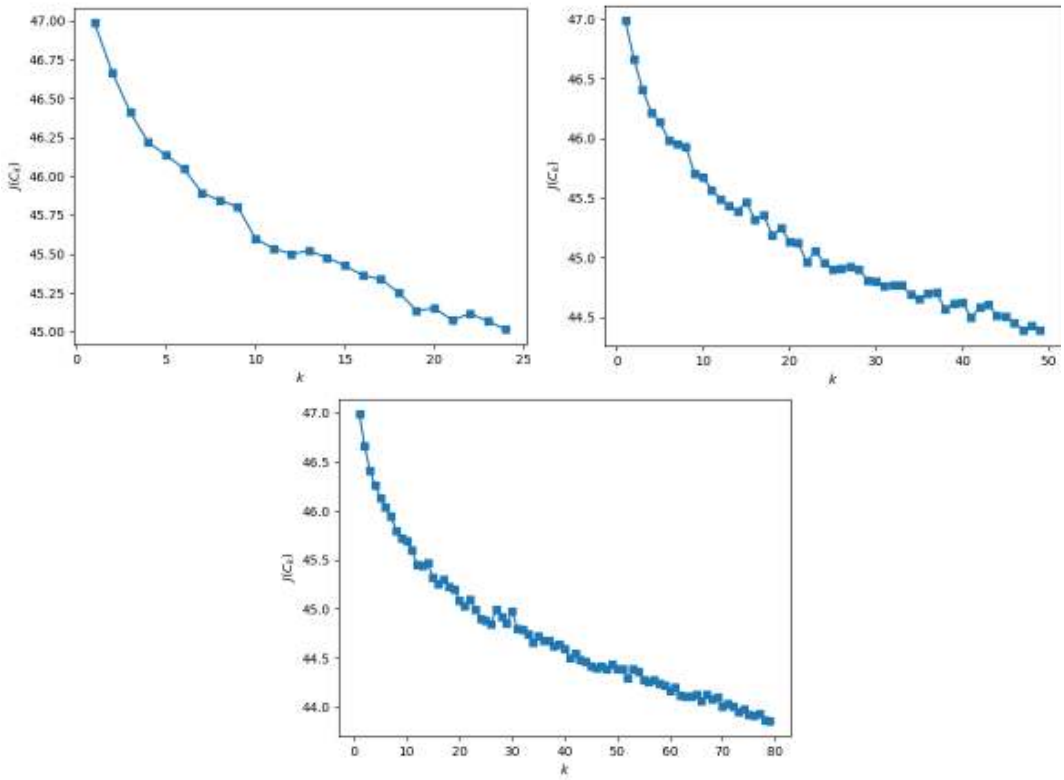

Рисунок 1 – Залежність суми квадратів відстаней від точок до центроїдів кластерів до кількості кластерів (k = 25, 50 та 80)

При невеликій кількості кластерів оптимальне значення можна визначити дивлячись на графік. Але при великих k, оптимальне значення кластерів визначити достатньо важко.

Для визначення оптимальної кількості кластерів можна застосувати метод «ліктя» або метод «коліна».

У кластерному аналізі метод ліктя використовується для визначення кількості кластерів в наборі даних. Даний метод розглядає характер змін $J(C_k)$ із збільшенням числа груп k (кластерів). Об'єднавши усі n спостережень в одній групі, на певному етапі дійсно можна придивитись, що $J(C_k)$ падає вже не так сильно - на графіку це відбувається в точці, яка і називається «ліктем».

Для визначення ліктя необхідно провести пряму лінію від кінцевих точок дослідження, і після цього обчислити відстань від кожної точки до цієї лінії. Точкою з найбільшою відстанню повинна бути точка, яка містить лікоть.

Провівши досліждення для 50-ти кластерів було визначено оптимальне значення – 15 кластерів, тобто 15 сфер діяльності кандидатів, до яких можуть

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

бути віднесені вхідні дані, тобто резюме. На рис. 2 показано визначений лікоть для експерименту з 50-ти кластерів.

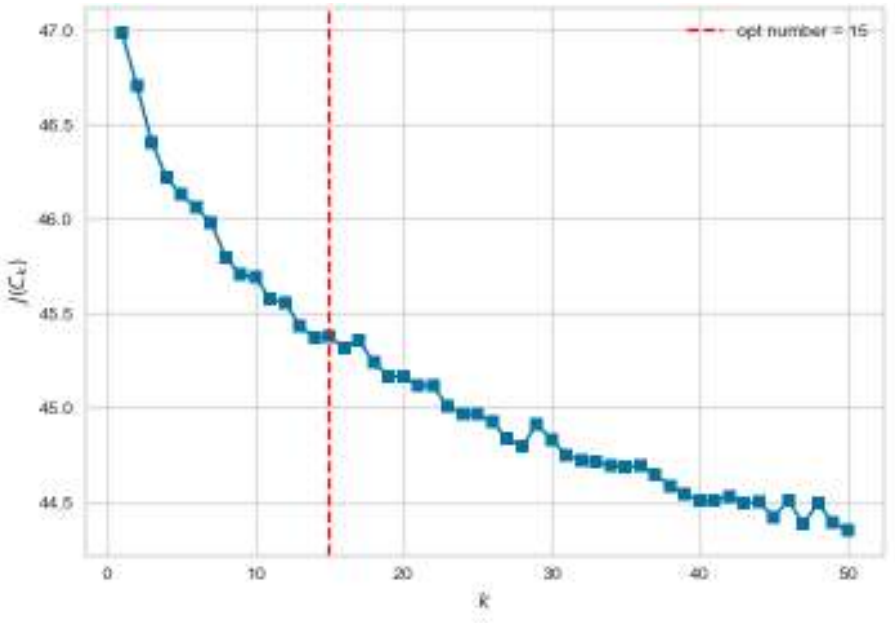

Рисунок 2 – Визначення оптимальної кількості кластерів
за домопогою методу «ліктя»

Висновки. Аналіз результатів показав, що використання методу c-means має важливу перевагу: можливість визначення ступеня приналежності елемента до кластеру. А визначення кількості кластерів для розбиття в даному випадку пропонується знайти, представивши задачу кластеризації як задачу оптимізації і застосувати ліктьовий метод, вказавши ту кількість кластерів, де залежність суми квадратів відстаней від точок до центроїдів кластерів падає «вже не так швидко».

ЛІТЕРАТУРА / ЛИТЕРАТУРА

1. Rout, Jayashree & Bagade, Sudhir & Yede, Pooja & Patil, Nirmiti. (2019). Personality Evaluation and CV Analysis using Machine Learning Algorithm. International Journal of Computer Sciences and Engineering. 7. 1852-1857. 10.26438/ijcse/v7i5.18521857.

2. Prof. Sagar More, Bhamare Priyanka, Mali Puja, Kachave Kalyani. (2019). Automated CV Classification using Clustering Technique. International Research Journal of Engineering and Technology (IRJET). Volume 6, Issue 6, Page No 302-305.

3. Klasternyiy analiz [Elektronniy resurs] — Rezhim dostupa k state: http://www.machinelearning.ru/wiki/index.php?title=Кластеризация

4. Dunn J.C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters // Journal of Cybernetics. — 1973. — 17 сентября (т. 3, № 3). — С. 32–57. — ISSN 0022-0280. — doi:10.1080/01969727308546046.

5. Chasovskih A. Obzor algoritmov klasterizatsii dannyih [Elektronniy resurs] — Rezhim dostupa k state: https://habr.com/ru/post/101338/

6. KorolYov S., Kashnitskiy Yu. Otkryityiy kurs mashinnogo obucheniya. Obuchenie bez uchitelya: PCA i klasterizatsiya [Elektronniy resurs] — Rezhim dostupa k state: https://habr.com/ru/company/ods/blog/325654/

REFERENCES

1. Rout, Jayashree & Bagade, Sudhir & Yede, Pooja & Patil, Nirmiti. (2019). Personality Evaluation and CV Analysis using Machine Learning Algorithm. International Journal of Computer Sciences and Engineering. 7. 1852-1857. 10.26438/ijcse/v7i5.18521857.

2. Prof. Sagar More, Bhamare Priyanka, Mali Puja, Kachave Kalyani. (2019). Automated CV Classification using Clustering Technique. International Research Journal of Engineering and Technology (IRJET). Volume 6, Issue 6, Page No 302-305.

3. Кластерный анализ [Електронний ресурс] — Режим доступу до статті: http://www.machinelearning.ru/wiki/index.php?title=Кластеризация

4. Dunn J.C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters // Journal of Cybernetics. — 1973. — 17 сентября (т. 3, № 3). — С. 32–57. — ISSN 0022-0280. — doi:10.1080/01969727308546046.

5. Часовских А. Обзор алгоритмов кластеризации данных [Электронный ресурс] — Режим доступа к статье: https://habr.com/ru/post/101338/

6. Королёв С., Кашницкий Ю. Открытый курс машинного обучения. Обучение без учителя: PCA и кластеризация [Электронный ресурс] — Режим доступа к статье: https://habr.com/ru/company/ods/blog/325654/

Received 19.03.2021.
Accepted 25.03.2021.

Применение методов кластеризации для определения сфер деятельности кандидатов при подборе кадров для ИТ-компаний

В данной статье рассмотрена проблема подбора кадров для ИТ-компаний. Рассмотрено практическое применение методов кластеризации на примере задачи определения сфер деятельности кандидатов и подбора резюме на определенную вакансию при подборе кадров для ИТ-компаний. Приведены постановку задачи. Определено, к которым моделей сводятся исследуемые проблемные ситуации, а также, какие методы могут быть применены к решению поставленной задачи. Приведен обзор известных решений, также освещены преимущества и недостатки выбранного метода. Приведены пример определения количества оптимальных кластеров для заданного набора резюме.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Application of clustering methods to determine the areas of activity of candidates in recruitment for IT-companies

Nowadays the selection of candidates for recruitment from a wide range of candidates is a fundamental issue. Today's HR managers have to handle extremely large amounts of data: portfolio research, social media screening, skill set identification, and, of course, resume research.

Professors Sagar More, Bhamara Priyanka, Mali Puja and Kachave Kalyani were considering the automated classification of resumes using clustering techniques. The solution proposed by scientists uses methods of data mining. The method of data mining clustering is used for classification and calculation.

The aim of the article is to study the methods of clustering and the transformation of the clustering problem into an optimization problem to improve the efficiency and quality of recommendations to recruitment managers.

In the task of determining the areas of activity of employees in recruitment for IT-companies an input information will be summarized in text form, which will contain all the information about the professional career of the employee, as well as cover letters, essays and career guidance tests with free open answers.

At the output we get a set of professional areas of activity of employees with the best resumes selected for them, that is, we get a grouping of input data to certain areas of activity.

It is suggested to use text clustering methods to group and combine input data. For clustering can be used c-means algorithm – a modification of the k-means method.

There is one disadvantage of the method: the need to know the number of clusters in advance. In this case, it is proposed to present the problem of clustering as optimization. The «elbow» method or the «knee» method can be used to determine the optimal number of clusters.

Analysis of the results showed that the use of the c-means method has an important advantage: the ability to determine the degree of belonging of the element to the cluster. And, also with usage of «elbow» method optimal number of clusters can be chosen.

Гавриленко Олена Валеріївна – доцент кафедри автоматизованих систем обработки інформації та управління Національного технічного університету України «Київський політехнічний інститут ім. І. Сікорського».
Дворник Вікторія Анатоліївна – студентка кафедри автоматизованих систем обработки інформації та управління Національного технічного університету України «Київський політехнічний інститут ім. І. Сікорського».

Гавриленко Елена Валерьевна - доцент кафедры автоматизированных систем обработки информации и управления Национального технического университета Украины «Киевский политехнический институт им. И. Сикорского ».
Дворник Виктория Анатольевна - студентка кафедры автоматизированных систем обработки информации и управления Национального технического университета Украины «Киевский политехнический институт им. И. Сикорского ».

Gavrylenko Olena – Associate Professor of the Department of Computer-Aided Management and Data Processing Systems of the National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute».
Dvornyk Viktoriia – student of the Department of Computer-Aided Management and Data Processing Systems of the National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute».

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)