УДК 004.853

О.І. Проніна
Державний вищий навчальний заклад «Приазовський державний технічний університет»
м. Маріуполь, Україна
pronina.lelka@gmail.com

Метод кластеризації для визначення результатів кредитування

В роботі розроблено метод використання кластеризації, а саме побудова кластерів згідно вибраних значущих показників, та на їх основі визначення згоди або відмови банку у видачі кредиту клієнту. У разі позитивного рішення стосовно видачі кредиту, визначення суми видачі.

Ключові слова: big data, data mining, кластеризація, карта Кохонена, відстань між кластерами, видача кредиту.

DOI: 10.31474/1996-1588-2019-1-28-67-72

Вступ

З розвитком інформаційних технологій розвиваються і сучасні обчислювальні системи і комп'ютерні мережі, які стали дозволяти накопичувати великі масиви даних для вирішення задач обробки та аналізу. Але інформація, яка необхідна людині, зберігається у прихованому вигляді, тобто для її отримання необхідно використовувати спеціальні алгоритми аналізу даних.

Великі дані (Big Data) – це набір методів та засобів опрацювання структурованих і неструктурованих різнотипних динамічних даних великих обсягів з метою їх аналізу та використання для підтримки прийняття рішень [1].

Визначальними характеристиками для Великих даних є обсяг (як величина фізичного обсягу), швидкість отримання результатів, різноманіття (як можливість одночасної обробки різних типів структурованих і слабо структурованих даних) [2].

Різноманіття визначається за допомогою [3]: реляційних даних (таблиці / транзакції); текстових даних (Web), напівструктурованих даних (XML); даних на основі графових моделей (соціальна мережа, Semantic Web, RDF); потокових даних; великих публічних даних (онлайн, погода, фінанси та ін.).

Швидкість Великих даних подана як: дані генеруються швидко і повинні бути опрацьовані швидко; онлайн аналіз даних; підтримка прийняття рішень з неповними даними.

Достовірність – поняття, зворотне до невизначеності, яка виникає через невідповідність даних, їх неповноту, латентність [4].

Опрацьовування даних за допомогою Data Mining включає в себе обробку, інтерпретацію та збір Великих даних. Застосовується технологія у різних сферах, таких, як бізнес, банківська справа, промисловість, некомерційні організації та інше. Для аналізу необхідно опрацьовувати величезні об'єми інформації, що надходять у різному вигляді, як то операції замовника, будь-який моніторинг, дані про клієнта, тощо. Все це можливо швидко інтерпретувати за допомогою інструментів аналізу даних.

При розгляді об'ємів інформації та можливих результатів, що будуються на цій інформації, можна усвідомити як саме зростає об'єм і де починають відігравати вирішальну роль великі данні.

Великі дані є терміном, який використовується для ідентифікації наборів даних, з якими не можливо впоратися з використанням існуючих методології та програмних засобів через їх великий розмір і складність. Багато дослідників намагаються розробити методики і програмні засоби для передачі даних або видобування інформаційних гранул з Великих даних [5, 6].

Методи видобутку даних (Data Mining) допомагають вирішити багато завдань, з якими стикається аналітик. З них основними є: класифікація, регресія, пошук асоціативних правил і кластеризація [7].

Кластеризація – об'єднання в групи схожих об'єктів, є однією з фундаментальних завдань в галузі аналізу даних і Data Mining [8]. Список прикладних областей, де вона застосовується, широкий: сегментація зображень, маркетинг, боротьба з шахрайством, прогнозування, аналіз текстів і багато інших. На сучасному етапі кластеризація часто виступає першим кроком при аналізі даних. Після виділення схожих груп застосовуються інші методи, для кожної групи будується окрема модель.

Завдання кластеризації в тому чи іншому вигляді формулювали в таких

наукових напрямках, як статистика, розпізнавання образів, оптимізація, машинне навчання, банківська справа [9]. На сьогоднішній момент число методів розбиття груп об'єктів на кластери досить великий - кілька десятків алгоритмів і ще більше їх модифікацій.

Метою дослідження є поліпшення інформаційної технології, що використовується у банківської справі стосовно кредитування.

Задачею дослідження є вибір значущих параметрів при побудові кластерів, визначення відстані між кластерами та виявлення проблемних місць в існуючому процесі кредитування.

Алгоритм кластеризації з точки зору його застосування в Data Mining

Кластеризація в Data Mining набуває цінність тоді, коли вона виступає одним з етапів аналізу даних, побудови закінченого аналітичного рішення. Аналітику часто легше виділити групи схожих об'єктів, вивчити їх особливості і побудувати для кожної групи окрему модель, ніж створювати одну загальну модель на всіх даних. Таким прийомом постійно користуються в маркетингу, виділяючи групи клієнтів, покупців, товарів і розробляючи для кожної з них окрему стратегію.

Саме цей метод було вирішено використовувати для банківської справи при прийнятті рішень стосовно видачі кредиту. Оскільки банк збирає дані не лише на кожного клієнта, але і на тих людей, що можуть стати клієнтами банку.

Дані, з якими стикається технологія DataMining, мають такі важливі особливості: висока розмірність (тисячі полів) і великий обсяг (сотні тисяч і мільйони записів) таблиць баз даних і сховищ даних (надвеликі бази даних); набори даних містять велику кількість числових і категорійних атрибутів. Що також присутня в банківській справі.

Більшість алгоритмів кластеризації припускають порівняння об'єктів між собою на основі певної міри близькості (подібності). Міра близькості має межу і зростає зі збільшенням близькості об'єктів. Заходи подібності «винаходяться» за спеціальними правилами, а вибір конкретних заходів залежить від завдання, а також від шкали вимірювань.

Після вибору міри близькості та її розрахунку, необхідно проаналізувати схожість об'єктів та проаналізувати побудовані кластери. Саме на цьому етапі аналітик визначає результат аналізу, та дає свою фінальну оцінку.

Основні характеристики вибору відстані між кластерами

Для об'єднання даних в кластери необхідно чітко визначити структуру майбутнього розбиття. Так, кластери можуть бути non-overlapping або exclusive, і overlapping. Окрім структури розбиття необхідно визначити, яким чином буде відбуватися розміщення в кластери нових даних, а саме за допомогою якої відстані буде визначатися приналежність вхідних даних до того чи іншого кластеру. У процесі визначення приналежності до кластерів можливо виявлення нових кластерів.

Основним методомдля об'єднання в кластери є «правило найближчого сусіда» (метод одиночного зв'язку) для визначення відстані між кластерами. В цьому методі відстань між двома кластерами визначається відстанню між двома найбільш близькими об'єктами (найближчими сусідами) в різних кластерах, тим самим формуючи об'єкти в найближчі кластери.

Також можливо використовувати «правило найдальшого сусіда» (метод повного зв'язку), при цьому об'єднання в кластери здійснюється наступним чином: новий об'єкт приєднується до того кластеру , у якого самий далекий елемент є найближчим до нового об'єкта, ніж до інших кластерів.

Для того, щоб розбити на кластери масив даних, необхідно для кожної вхідної ситуації визначити клас, до якого вона належить. Для цього слід використовувати одну з основних мір відповідності, а саме відстань між аналізованим об'єктом та існуючими класами.

Якщо змінні виміряні в різних одиницях виміру, то потрібне їх попереднє нормування, тобто перетворення вихідних даних, яке переводить їх у безрозмірні величини. Перехід від традиційних одиниць виміру до нормалізованих і назад з використанням методу лінійної нормалізації здійснюється з використанням наступних розрахункових співвідношень:

При нормалізації і денормалізації в межах [0, 1]:

$$\tilde{x}_{ik} = \frac{x_{ik} - x_{min_i}}{x_{max_i} - x_{min_i}},$$
$$y_{jk} = y_{min_j} + \tilde{y}_{jk}(y_{max_j} - y_{min_j}) \quad (1)$$

де x_{ik} та y_{jk} – - і-а вхідне та j-є вихідне значення k-го прикладу вихідної вибірки в традиційних одиницях виміру, прийнятих в розв'язуваної задачі;

$\tilde{x}_{ik}$, $\tilde{y}_{jk}$ – відповідні їм нормалізовані вхідні і вихідні значення;

N – кількість прикладів навчальної вибірки.

При нормалізації і денормалізації у межах [-1, 1]:

$$\tilde{x}_{ik} = 2 \cdot \frac{x_{ik} - x_{min_i}}{x_{max_i} - x_{min_i}} - 1,$$
$$y_{jk} = y_{min_j} + \frac{(\tilde{y}_{jk}+1)(y_{max_j} - y_{min_j})}{2}, \quad (2)$$

де $x_{min_i} = \min_{k=1,N}(x_{ik})$; $x_{max_i} = \max_{k=1,N}(x_{ik})$;
$y_{min_i} = \min_{k=1,N}(y_{ik})$; $y_{max_i} = \max_{k=1,N}(y_{ik})$.

Це традиційній спосіб обчислення нормалізації, який має достеменні переваги, так то введення нового елементу, при цьому відстань невідворотна оскільки елемент може виявитися викидом. Також ще використовується нелінійна нормалізація, для якої використовується сигмоїдноподібна логістична функція, а також гіперболічний тангенс. При порівнянні способів нормалізації співпадає лише центр нормалізованого інтервалу.

Для поставленої задачі нормалізація проводилась для ряду показників згідно формули (1).

Кластерний аналіз дозволяє провести об'єктивну класифікацію будь-яких об'єктів, які охарактеризовані рядом ознак. З цього можна витягти ряд переваг:

1) Отримані кластери можна інтерпретувати, тобто описувати, які ж власне групи існують.

2) Окремі кластери можна вибраковувати. Це корисно в тих випадках, коли при наборі даних допущені певні помилки, в результаті яких значення показників у окремих об'єктів різко відхиляються. При застосуванні кластерного аналізу такі об'єкти потрапляють в окремий кластер.

3) Для подальшого аналізу можуть бути обрані лише ті кластери, які мають важливими характеристиками. Кластерний аналіз має деякі неточність і локалізація, а саме зміст і чисельність кластерів залежить від обраних критеріїв розбиття.

5) При створенні вихідного масиву даних та переходу до компактного вигляду можливим є поява певних викривлень, а також можливо будуть губитися індивідуальні деталі окремих об'єктів за рахунок заміни їх характеристик узагальненими значеннями параметрів кластера..

Кластерний аналіз для визначення об'ємів кредитування

У ролі проведеного дослідження виступали дані отримані при зборі інформації про видачу кредиту, а саме: сума кредиту, вартість кредиту, термін кредиту, дата кредитування, мета кредитування, кількість, вік, стать, освіта, приватна власність, квартира, площа квартири, спосіб придбання власності, розташування, машина, термін експлуатації машини, заміський будинок, земельна ділянка, прописка в даному районі, гараж, клас підприємства, час роботи підприємства та інше. Кількість показників, що описує клієнта може достигати сімдесят показників.

Оператору банку при вирішенні стосовно кредитування клієнта необхідно проаналізувати усі показники, та дати рішення. При цьому може виникати ситуації, коли клієнту можливо видати кредит, але оператор сумнівався, тому у видачі відмовлено, та навпаки, коли кредит видано, але погашення кредиту у подальшому неможливо. У роботі було використано кластерний аналіз, для цього було обрано вибірку даних на який проводився експеримент. У подальшому даний метод використовувався для великих даних.

Для побудови кластерного аналізу було виділено значущі характеристики, стосовно кожного з варіантів розбиття. Всього було сформовано дванадцять різноманітних варіантів, в яких деяка кількість наборів показників є незмінною.

На наступному етапі було визначено схожості, та побудована карта Кохонена, для цього було встановлено помилки, для навчальної множини: максимальна помилка 5,94E-01, та середня помилка 1,40E-01, та для тестової множини: максимальна помилка 7,20E-01, та середня помилка 5,17E-01.

Задано кількість епох, та час навчання навчальної множини, на цьому етапі було проведено експеримент, для визначення оптимальної кількості епох навчання.

На наступному кроці було побудовано зв'язок між кластерами. Для наочності результатів було побудовано самоорганізована карта Кохонена, для кожної значущої характеристики, інтерпретація результатів представлена на рис. 1.

Після побудови кластерів, було задано вид відстані, для цього було проведено експеримент, який показав, що для обраної предметної галузи, та значущих показників доцільно

використання манхетенскьої відстані або відстані до найближчого сусіда.

Розраховано відстань між кластерами, та кількість змішаних крапок, окрім цього визначено вкидання, та межи кластерів, що представлено на рис.2. Кольором продемонстровано створені кластери, у кількості дев'яти.

Розроблена самоорганізована карта Кохонена демонструє результати по видачі кредиту, спираючись на чотирнадцять значущих характеристик. Інші показники були убрані з аналізу, оскільки мали найнижчу силу зав'язків кластерів, та добавляли вкидання, що зменшувала достовірність отриманих результатів. Крім цього деякі пари показників були дуже близькі по своїй природі, тому також були видалені з аналізу.

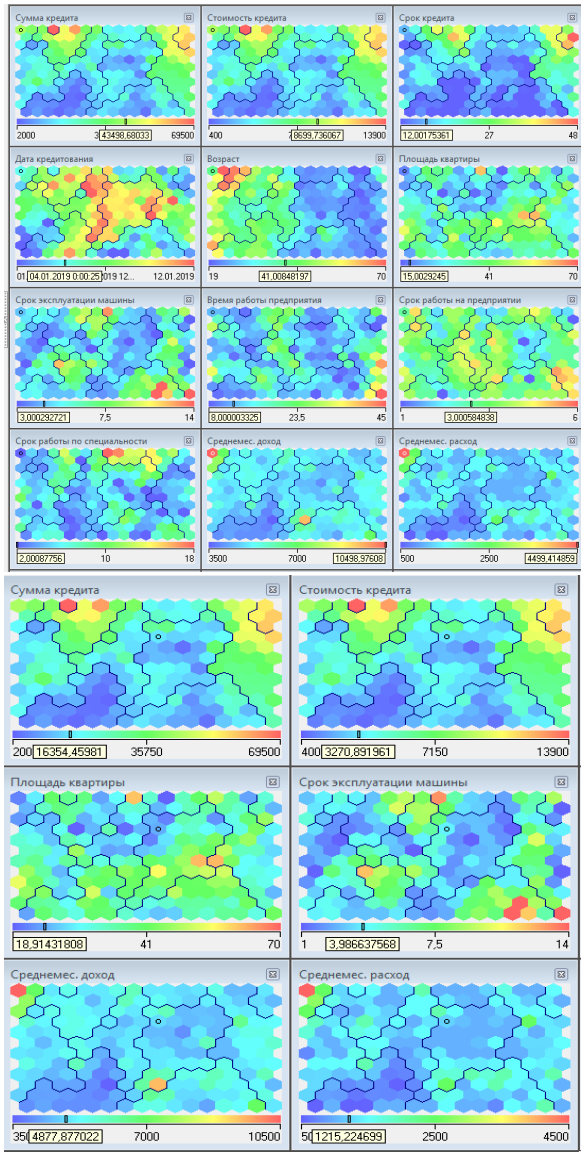

Рисунок 1 – Побудовані кластери для обраної предметної галузі

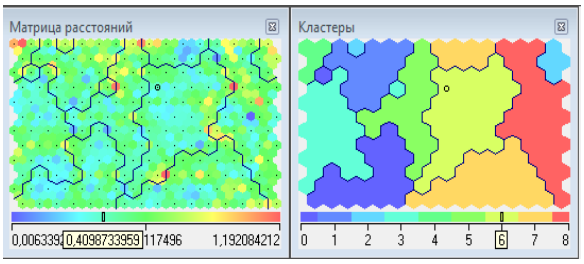

Рисунок 2 – Загальний вигляд кластерів, та матриця відстані між кластерами

На етапі експерименту були обрані різні значущі показники, але при використанні кластерного аналізу, було забагато «вкидань» та кількість фінального розбиття була забагато, що ускладнювала подальший аналіз.

Отримані результати підтверджують зв'язок між характеристиками, які впливають на результати «видачі кредиту», крім цього можна визначити суму кредиту для видачі. Оскільки сума кредиту також є значущим показником, та при побудові кластерів дозволяє встановити границі стосовно того чи іншого варіанту кредитування.

Заключення

Використання кластерного аналізу дозволяє класифікувати об'єкти за їх ознаками, при цьому необхідно, щоб результати класифікації мали змістовну інтерпретацію. Результати, отримані методами кластерного аналізу, застосовують в самих різних областях, при цьому дають точні результати.

Для визначення рішення стосовно видачі кредиту, та його об'єму було виконано кластерний аналіз та побудовано карту Кохонена. У процесі експерименті виявлені труднощі з інтерпретацією отриманих результатів, які виникають внаслідок того, що подібності між різними кластерами можуть бути причиною деякої відмінності підмножин змінних.

Проведенні дослідження в галузі банківської справи, при великої кількості даних показали, що визначення суми кредиту та рішення стосовна «видачі кредиту», або «відмови у видачі кредиту» є актуальним практичним завданням оскільки дозволяє зменшити ризики як для банку, що видає кредит, так і для людини, яка можливо не зможе його виплатити, та буде зазнавати збитків, а можливо і більш суттєві наслідки своїх необдуманих вчинків.

Наукова новизна запропонованого методу полягає у використанні кластерізації, в основі системи, що радить яка забезпечить видачу кредиту та його об'єм, аналізуючи множину показників.

Практична значимість метода полягає в зменшенні впливу людини при вирішенні видачі кредиту та його обсягів, що в свою чергу дозволить з часом повністю усунути людський фактор та автоматизувати систему банку по видачі кредиту. За рахунок автоматизації процесу обробки запиту на видачу кредиту збільшиться пропускна спосібність оператору банку та збільшиться комфорт клієнту.

Список літератури

1. Laney D. The Importance of «Big Data»: A Definition [Text] [Electronic Resours] / Mark A. Beyer, Douglas Laney. – Access mode: https://www.gartner.com/doc/2057415/importance-big-data-definition.
2. Jacobs A. The Pathologies of Big Data [Text] / Jacobs A. // Databases. – 2009. – Vol. 7, issue 6. – P.1 – 12.
3. Oracle and FSN: Mastering Big Data: CFO Strategies to Transform Insight into Opportunity [Electronic Resours]. Access mode: http://www.oracle.com/us/solutions/ent-performance-bi/business-intelligence/mastering-bigdata-cfo-strategies-1853061.pdf.
4. SnijdersC. «Big Data»: Big gaps of knowledge in the field of Internet [Electronic Resours] / C.Snijders, U.Matzat, and U.-D.Reips // International Journal of Internet Science. – Vol.7. – P. 1-5. – Access mode: http://www.ijis.net/ijis7_1/ijis7_1_editorial.html.
5. Information Granularity, Big Data, and Computational Intelligence [Електроннийресурс] / W. Pedrycz and S.-M. Chen (eds.). – Access mode: https://books.google.com.ua/books?id.
6. MapReduce and Parallel DBMSs: Friends or Foes [Text] / Stonebraker M., Abadi D., DeWitt D. J., Madden S., Pavlo A., Rasin, A. // Communications of the ACM. – 2012. – Vol. 53, №1. – P. 64-71.
7. Верес О. М. Класифікація методів аналізу Великих даних / О. М. Верес, Р. М. Оливко // Вісник Національного університету «Львівська політехніка». Серія: Інформаційні системи та мережі. — Львів : Видавництво Львівської політехніки, 2017. – № 872. – С. 84 – 92.
8. Ершов К. С. Анализ и классификация алгоритмов кластеризации / К. С. Ершов, Т. Н. Романова // Новые информационные технологии в автоматизированных системах, 2016. – № 19. – С 274 – 279.
9. Abbas O.A. (2008). Comparisons Between Data Clustering Algorithms, The Interna onal Arab Journal of Informa on Technology Vol. 5.

References

1. Beyer, M., Laney, D., The Importance of «Big Data»: A Definition. Available at https://www.gartner.com/doc/2057415/importance-big-data-definition.
2. Jacobs, A. (2009). The Pathologies of Big Data. Databases. Vol. 7, issue 6. p. 1 – 12.
3. Oracle and FSN: Mastering Big Data: CFO Strategies to Transform Insight into Opportunity.
4. Available at http://www.oracle.com/us/solutions/ent-performance-bi/business-intelligence/mastering-bigdata-cfo-strategies-1853061.pdf.
5. Snijders, C., Matzat, U., Reips, U.-D. «Big Data»: Big gaps of knowledge in the field of Internet. International Journal of Internet Science. Vol.7. p. 1-5. Available at: http://www.ijis.net/ijis7_1/ijis7_1_editorial.html.
6. Pedrycz, W., Chen, S.-M. Information Granularity, Big Data, and Computational Intelligence. Available at https://books.google.com.ua/books?id.
7. Stonebraker, M., Abadi, D., DeWitt, D. J., Madden, S., Pavlo, A., Rasin, A. (2012). MapReduce and Parallel DBMSs: Friends or Foes. Communications of the ACM. Vol. 53, №1. P. 64-71.
8. Veres, O. M., Olivko, R.M. (2017). Classification of Big Data Analysis Methods [*Klasyfikatsiya metodiv analizu velykykh danykh*]. Bulletin of the National University "Lviv Polytechnic". № 872. p. 84 – 92.
9. Ershov, K.S., Romanova, T.N. (2016). Analysis and classification of clustering algorithms [*Analiz i klassifikatsiya algoritmov klasterizatsii*]. New information technologies in automated systems. № 19. P. 274 - 279.
10. Abbas, O.A. (2008). Comparisons Between Data Clustering Algorithms. The International Arab Journal of Information Technology. Vol. 5.

Надійшла до редакції 10.06.2019

О. И. ПРОНИНА
Государственное высшее учебное заведение «Приазовский государственный технический университет»

МЕТОД КЛАСТЕРИЗАЦИИ ДЛЯ ОПРЕДЕЛЕНИЯ РЕЗУЛЬТАТОВ КРЕДИТОВАНИЯ

В работе применен метод кластеризации, а именно построены кластеры, согласно выбранным значащим показателям, и на их основе определено согласие или отказ банка в выдаче кредита. В случае позитивного решения относительно выдачи кредита, определение суммы выдачи.

Ключевые слова: big data, data mining, кластеризация, карта Кохонена, расстояние между кластерами, выдача кредита.

O. I. PRONINA
State Higher Educational Institution "Priazovsky State Technical University"

CLUSTERING METHOD FOR DETERMINING CREDIT RESULTS

The lending function is by far one of the most popular features provided by banks. The issue of issuing a loan is decided by a representative of the bank. The result obtained on the basis of the decision of the bank employee is not always beneficial, both to the bank and people. In some cases, the user cannot repay the loan and the bank suffers losses, and the user receives heavy fines. In order to minimize losses, a methodology for assessing creditworthiness and loan volumes has been proposed. For this purpose, a clustering method was applied for a large amount of data.

Significant parameters were chosen, clusters were built on their basis. An experiment was conducted on the choice of indicators that affect the user's creditworthiness. Taking into account the chosen significant indicators, the consent or refusal of the bank in granting a loan to the user is determined. In case of a positive decision regarding the issue of a loan, the amount of the issue is determined.

The application of the method will reduce the influence of a person when deciding on the issuance of a loan and its volume, which in turn will eventually completely eliminate the human factor and automate the bank's system for issuing a loan. Due to the automation of the processing of the request for a loan, the throughput capacity of the bank operator will increase and the customer's comfort will increase.

Keywords: big data, data mining, clustering, Kohonen map, distance between clusters, loan disbursement.