Y. Tykhonov, K. Tykhonova

ANALYSIS OF EXISTING DATA CLUSTERING ALGORITHMS. ADVANTAGES AND DISADVANTAGES

Analyzing data clustering algorithms is increasingly becoming a popular practice adopted by many organizations to create valuable information from large amounts of data. A great deal of research aims to organize the data obtained into supervisory structures. In fact, cluster analysis is a set of different classification algorithms. The clustering technique is used in various fields, such as psychology, biology, pedagogy, marketing, information technology. Clustering is the division of data into groups of similar objects. Clustering is performed to understand the data obtained, the volume of which is problematic for human analysis. Thanks to this, clustering algorithms have become a meta-learning tool for analyzing research data. Each group, called a cluster, is defined as a set of objects that have a higher degree of similarity to each other than objects that are not in the same set. The type of clustering algorithm used depends on the application and the data set used in this field. The numerical data set is relatively simple to implement since the data is invariably real numbers and can be used for statistical applications. It is important to understand the difference between clustering (uncontrolled classification) and discriminant analysis (controlled classification). At one stage, the researchers were refining some of the data clustering algorithms, the second was implementing new ones, and at the third, they were studying and comparing different data clustering algorithms. This article provides a classification and analysis of existing cluster analysis algorithms, as well as the advantages and disadvantages of these algorithms.

Keywords: clustering; cluster analysis; hierarchical clustering; non-hierarchical clustering; Clustering Using REpresentatives (CURE) algorithm; Minimum Spanning Tree (MST) algorithm; Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) algorithm; *k*-means algorithm; Partition Around Medoids (PAM) algorithm; Clustering with sLOPE (CLOPE) algorithm.

УДК 004.852

DOI: 10.31673/2412-9070.2020.062023

М. В. ТИМОШИК, студент;
В. І. СТРЕЛЬНИКОВ, аспірант,
Державний університет телекомунікацій, Київ

ЗАСТОСУВАННЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ КОРИСТУВАЧІВ ЗА ДОПОМОГОЮ АЛГОРИТМІВ КЛАСТЕРИЗАЦІЇ

Розглянуто застосування інтелектуального аналізу даних користувачів за допомогою алгоритмів кластеризації на прикладі алгоритму агломеративної ієрархічної кластеризації. Показано сутність агломеративної ієрархічної кластеризації. Визначено сфери застосування підходу оброблення даних із мережі.

Роботу проведено на основі аналізу алгоритмів кластеризації, що дає змогу оцінити застосування технології для даних користувачів, а також дає розуміння, який вплив має використання інтелектуального аналізу в цілому.

Ключові слова: інтелектуальний аналіз даних; кластеризація; масштабованість; класифікація; аналітичне оброблення в реальному часі; використання кластеризації; закономірності; збір даних; інтеграція даних; аналіз даних; кластерний аналіз; ієрархічні агломеративні методи.

ВСТУП

Постановка проблеми. У процесі розвитку інформаційних технологій, а також систем збору і зберігання даних все гостріше постає проблема аналізу великих обсягів інформації. Іншим, не менш важливим завданням є завдання наочного і компактного подання даних. Ці проблеми вирішуються в рамках міждисциплінарної галузі знань — інтелектуального аналізу даних (Data Mining) [1].

Сьогодні все більшої актуальності набуває аналіз даних, отриманих із інтернету, так званий ***Web Mining***. Основна мета Web Mining — це збір даних (парсинг) з подальшим збереженням у потрібному форматі [2]. При цьому необхідно зважати на те, що інформацію в інтернеті подано у вигляді спеціальних форматів, таких як мова позначки HTML, RSS, Atom, SOAP тощо. Веб-сторінки можуть мати додаткову метаінформацію, а також інформацію про структуру документа.

У Web Mining можна виокремити два основних напрямки: *Web Content Mining* і *Web Usage Mining* і, відповідно, два види завдань, які висуваються перед системами Web Mining [3]. Web Content Mining означає автоматизований пошук інформації з різних джерел в інтернеті. Другий напрямок, більш пристосований, Web Usage Mining має на меті виявлення закономірностей у діях відвідувача сайтів, а також збір статистики і подальший її аналіз.

Мета статті — описати використання інтелектуального аналізу даних на прикладі алгоритму агломеративної ієрархічної кластеризації, застосування варіанту Ланса-Вільямса з оптимізацією продуктивності (редуктивний алгоритм) під час роботи з даними з мережі, дослідити

сферу застосування підходу у процесі приведення до подібного формату даних.

ОСНОВНА ЧАСТИНА

Кластерний аналіз, принципи якого використовувалися на етапі побудови алгоритму, не потребує апріорних припущень про вихідні дані, як і не накладає обмежень на подання досліджуваних об'єктів, дає можливість аналізувати показники різних типів даних.

На відміну від завдань класифікації розв'язок задач кластеризації ґрунтується на порівнянні самих об'єктів і встановленні їх схожості за певними характеристиками. Розроблений алгоритм було застосовано до реального прикладу об'єднання блогів, що містять семантично близькі записи. Алгоритми кластеризації поділяють сукупність даних на підмножини, або кластери. Мета цих алгоритмів — створити кластери, однорідні всередині, але такі, що чітко різняться один від одного. Дані або вектори характеристик, які є елементами безлічі, всередині кластера мають бути максимально схожими один на одного, але водночас максимально відрізнятися від елементів іншого кластера.

Із усіх методів кластерного аналізу найпоширенішими є ієрархічні агломеративні методи. Сутність цих методів полягає в тому, що на першому етапі кожний об'єкт вибірки розглядається як окремий кластер. Процес об'єднання кластерів відбувається послідовно: на підставі матриці відстаней або матриці подібності. Якщо матриця подібності спочатку має розмірність $m \times m$, то повністю процес кластеризації завершується за $m - 1$ кроків, у результаті всі об'єкти буде об'єднано в один кластер. Послідовність об'єднання легко піддається геометричній інтерпретації та може бути подано у вигляді графа-дерева (дендрограми). На дендрограмі зазначаються номери об'єктів, що об'єднуються, і відстань (або інший ступінь подібності), за якого відбулося об'єднання (рис. 1).

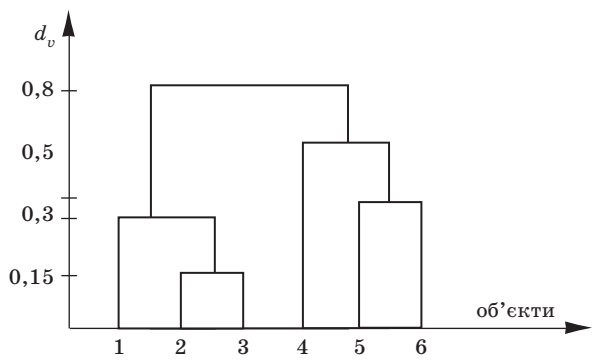

Рис. 1. Приклад дендрограми ієрархічного агломеративного кластерного аналізу

Дендрограма на рис. 1 показує, що в даному разі на першому кроці було об'єднано в один кластер об'єкти n_2 і n_3. Відстань між ними 0,15. На другому кроці до них приєднався об'єкт n_1. Відстань від першого об'єкта до кластера, що містить об'єкти n_2 і n_3, було 0,3 і т. д.

Кластеризація здійснюється на основі входження знаків в об'єкти, що кластеризуються. Для кожного елемента будується вектор лічильників знаків розмірності $1 \times M$, де M — об'єднання всіх можливих знаків в елементах. Якщо знак не входить в елемент, лічильник дорівнює нулю, інакше — кількості входжень. Після цього всі вектори об'єднуються в матрицю розмірності $N \times M$, де N — кількість елементів, що кластеризуються. Далі обчислюється матриця подібності, а на її основі об'єднуються найбільш схожі один на одного в поточний момент кластери. На кожній новій ітерації найбільш схожі кластери об'єднуються, а рядки і стовпці, відповідні об'єднаному кластеру, обчислюються заново. Утворені у такий спосіб розбиття зберігаються у вигляді списку об'єднань.

Після цього розраховується схожість кластера з об'єднанням кластерів.

На основі зазначеного опису об'єктів із використанням матриці подібності було запропоновано алгоритм ієрархічної кластеризації даних, що відбиває семантичну близькість елементів.

1. *Аналіз каналів для отримання лічильників знаків.* У процесі використання алгоритму ієрархічної кластеризації передбачається отримувати дані безпосередньо або за допомогою RSS-каналів.

RSS-канал — це простий XML-документ, що містить інформацію про канал і всі записи в ньому. Потрібно сформувати список каналів, з якими можна працювати.

2. *Реалізація функції, яка буде витягувати окремі знаки.* У каналах формату RSS і Atom завжди є заголовок і список записів. У кожного запису зазвичай є тег summary або description, всередині якого розміщено тему запису.

3. *Генерація списку знаків, які будуть враховані в лічильниках для кожного каналу.*

4. *Створення матриці лічильників знаків для кожного каналу на основі списку знаків і списку каналів.* На завантаження всіх каналів часто витрачається досить багато часу, але врешті-решт маємо таблицю, в якій стовпці відповідають знакам, а рядки — каналам. Такий формат забезпечує можливість застосування отриманих результатів в інших алгоритмах інтелектуального аналізу даних.

Розглянемо основні сфери застосування підходу.

Телекомунікації

Телекомунікаційна індустрія була однією з перших, хто застосував інтелектуальний аналіз здобутих даних і розгорнув численні програми для їх отримання. Приклади таких програм стосуються маркетингу, виявлення шахрайства та моніторингу мережі. Обмін даними в галузі телекомунікацій стикається з проблемами через розмір наборів даних, послідовний та часовий характер даних та вимоги багатьох додатків у режимі реального часу. Розроблено нові методи та вдосконалено наявні для відповіді на ці виклики. Конкурентний та мінливий характер галузі у поєднанні з тим, що галузь генерує величезні обсяги даних, гарантує, що здобуття даних відіграватиме важливу роль у майбутньому галузі телекомунікацій.

Страхування

Головний чинник набуття конкурентної переваги в страховій галузі — це визнання того, що бази даних клієнтів у разі належного керування, аналізу та експлуатації є унікальними, цінними корпоративними активами. Страхові компанії можуть розблокувати цінність, що міститься у базах даних клієнта, за допомогою сучасної технології інтелектуального аналізу даних. Для аналізу використовується прогнозоване моделювання, сегментація бази даних, аналіз ринкових кошиків та їх комбінації для більш швидкого реагування на важливі для бізнеса питання з більшою точністю. Нові продукти і маркетингові стратегії дають можливість страховій фірмі перевести цінність невикористаної наразі інформації в «багатство» передбачуваності, стабільності та прибутку.

Прикладна хімія

Великі бази даних набувають все більшого значення в хімії. Для всеосяжного використання цих баз даних необхідні автоматичні обчислювальні методи. У багатьох актуальних питаннях у всьому світі, зокрема дизайну каталітичних матеріалів для збору парникових газів, оптимізації та дослідження відновлюваної енергії ІАД показав прогнозовану потужність для побудови взаємозв'язків між різними внутрішніми і зовнішніми властивостями. Зазвичай, місія процесу ІАД — передбачити (або вивести) ті змінні, яких важко дістати через експерименти чи моделювання за допомогою змінних, котрі можна легко підставити як вхідні дані. Завдяки добре підігнаної нелінійної форми прогнозовані змінні можуть бути швидко виведені за допомогою входів цих незалежних змінних. Інакше кажучи, машинне навчання сприяє ІАД для прискорення, оптимізації інженерних процесів, відкриття нових функціональних матеріалів та розуміння хімічних процесів.

Генна інженерія

Біомедицина є сферою, багатою на знання і яка має численні стимули кодувати її в електронному форматі та ділитися нею за допомогою відкритих та підтримуваних у спільнотах баз даних та знань. Вона містить інформацію про послідовність та структуру послідовностей, взаємодії генів та білків, анотацію функцій та онтологій або генетичні та метаболічні шляхи. Ця інформація може істотно доповнити будь-який аналіз даних та покращити його результати. Залучення додаткових джерел знань у процес аналізу даних може запобігти виявленню очевидних речей, доповнити, використовуючи ці дані, гіпотезу посиланнями на вже запропоновані взаємозв'язки, допоможе уникнути надто впевнених прогнозів і, нарешті, дасть змогу співвідносити результати аналізу для систематизації знань.

ВИСНОВКИ

З огляду на зростання темпів нагромадження інформації постає нагальна потреба в технологіях аналізу даних, які також стрімко розвиваються. Розвиток цих технологій в останні роки уможливив перехід від сегментування клієнтів на групи з аналогічними перевагами до побудови моделей у режимі реального часу, ґрунтуючись, зокрема, на запитах в інтернеті і відвідуванні тих чи інших сторінок. Стає реальним виводити конкретні пропозиції і рекламу на основі аналізу інтересів споживача, роблячи ці пропозиції набагато більш цільовими. Також можливі коригування та переналаштування моделі в режимі реального часу. Кластерний аналіз можна назвати найзручнішим і найоптимальнішим інструментом виокремлення сегментів ринку.

Використання даних методів набуло особливої актуальності в століття високих технологій, коли так важливо прискорити трудомісткі і тривалі процеси за допомогою технологій.

Список використаної літератури

1. ***Henzinger M., Raghavan P., Rajagopalan S.*** *Computing on Data Streams // Digital Equipment Corporation. SRC TN-1998-011, August 1998.*

2. ***Тиндова М. Г.*** *Предварительная кластеризация многомерных объектов в интеллектуальном анализе данных // Вестник Саратов. гос. соц.-эконом. ун-та. 2008. №. 4. С. 137–138.*

3. ***Murphy S. A.*** *Data visualization and rapid analytics: applying tableau desktop to support library decision-making // Journal of Web Librarianship. 2013. Vol. 7, № 4. P. 465–476.*

М. В. Тимошик, В. И. Стрельников

ПРИМЕНЕНИЕ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ ПОЛЬЗОВАТЕЛЕЙ С ПОМОЩЬЮ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ

Рассмотрено применение интеллектуального анализа данных пользователей с помощью алгоритмов кластеризации на примере алгоритма агломеративной иерархической кластеризации. Показана сущность агломеративной иерархической кластеризации. Обозначены области применения подхода обработки данных из сети.

Работа проведена на основе анализа алгоритмов кластеризации, что позволяет оценить применение технологии для данных пользователей, а также дает понимание того, какое влияние имеет интеллектуальный анализ в целом.

Ключевые слова: интеллектуальный анализ данных; кластеризация; масштабируемость; классификация; аналитическая обработка в реальном времени; использование кластеризации; закономерности; сбор данных; интеграция данных; анализ данных; кластерный анализ; иерархические агломеративные методы.

M. Tymoshyk, V. Strelnikov

APPLYING USER DATA MINING USING CLUSTERING ALGORITHMS

In this article we are talking about the applying of user data mining using clustering algorithms.

In the process of information technology development, as well as data collection and storage systems, the problem of analyzing large amounts of information is becoming increasingly acute. Another equally important task is the visual and compact presentation of data. These problems are solved within the framework of an interdisciplinary area of knowledge - Data Mining.

Today, the analysis of data obtained from the Internet, the so-called Web Mining, is becoming increasingly relevant. The main purpose of Web Mining is to collect data (Parsing) and then save it in the desired format.

The information on the Internet is presented in the form of special formats, such as markup language HTML, RSS, Atom, SOAP and others. Web pages may have additional meta information as well as document structure information.

In Web Mining, there are two main areas of focus: Web Content Mining and Web Usage Mining, and, accordingly, two types of tasks that Web Mining systems are facing. Web Content Mining means the automated search for information from various sources on the Internet. The second direction is more adapted, Web Usage Mining implies the detection of patterns in the actions of the site visitor, as well as the collection of statistics and its subsequent analysis.

This work is based on the analysis of clustering algorithms, allows to evaluate the use of technology for user data, gives an understanding of the impact the use of intellectual analysis has in general.

Main applications fields are also being shown describing the benefits of integrating this analyze approach.

Keywords: data mining; clustering; scalability; classification; real-time analytical processing; the use of clustering; patterns; data collection; data integration; data analysis; cluster analysis.