УДК 004.62
DOI: 10.31673/2412-9070.2020.061719

Є. С. ТИХОНОВ, ст. викладач;
К. В. ТИХОНОВА, студентка,
Державний університет телекомунікацій, Київ

АНАЛІЗ ІСНУЮЧИХ АЛГОРИТМІВ КЛАСТЕРИЗАЦІЇ ДАНИХ. ПЕРЕВАГИ ТА НЕДОЛІКИ

Аналіз алгоритмів кластеризації даних все частіше стає популярною практикою, прийнятою багатьма організаціями з метою створення цінної інформації з великих обсягів даних. Чимала кількість досліджень ставлять собі за мету організацію отриманих даних у наочні структури. Фактично, кластерний аналіз є набором різноманітних алгоритмів класифікації. Техніка кластеризації застосовується в найрізноманітніших галузях, зокрема психології, біології, педагогіці, маркетингу, інформаційних технологіях. Кластеризація — це поділ даних на групи подібних об'єктів. Кластеризацію застосовують для розуміння отриманих даних, обсяг яких є проблематичним для аналізу людиною. Завдяки цьому алгоритми кластеризації стали інструментами мета-навчання для аналізу дослідницьких даних. Кожна група, що називається кластером, визначається як сукупність об'єктів, які мають більш високий ступінь схожості один з одним порівняно з об'єктами, що не належать до одного набору. Тип використовуваного алгоритму кластеризації залежить від програми та набору даних, що застосовуються в цьому полі. Числовий набір даних порівняно просто реалізувати, оскільки дані — це незмінно реальні числа і можуть використовуватися для статистичних застосувань. Важливо розуміти різницю між кластеризацією (непідконтрольную класифікацією) та дискримінаційним аналізом (контрольованою класифікацією). На першому етапі дослідники вдосконалювали деякі алгоритми кластеризації даних, на другому — впроваджували нові, а на третьому — вивчали та порівнювали різні алгоритми кластеризації даних. У статті проведено класифікацію та аналіз існуючих алгоритмів кластерного аналізу, також розглянуто переваги та недоліки цих алгоритмів.

Ключові слова: кластеризація; кластерний аналіз; ієрархічна кластеризіція; неієрархічна кластеризіція; алгоритм кластеризації за допомогою представників (CURE); алгоритм мінімального кістякового дерева (MST); алгоритм збалансованого ітеративного скорочення і кластеризації за допомогою ієрархій (BIRCH); алгоритм *k*-середніх (*k*-means); алгоритм розділення навколо медоїдів (PAM); алгоритм скупчення з нахилом (CLOPE).

ВСТУП

Постановка проблеми. Сьогодні стрімке зростання обсягу інформації в навколишньому світі спонукає сучасні технології до розв'язання актуального завдання підвищення ефективності пошуку необхідної інформації в глобальному інформаційному просторі. Це завдання вимагає дослідження та розроблення методів і алгоритмів розподілу інформаційних моделей об'єктів на певні групи і класи. Завдання такого роду постають у таких сучасних інформаційних технологіях, як Data Mining, розпізнавання образів, машинне навчання.

Щоденно світ виробляє майже 2,5 квантилійних байт даних (2,5 млрд гігабайт), при цьому 90% цих даних у світі неструктуровано [1].

Терміном кластерний аналіз прийнято позначати сукупність методів, підходів і процедур, розроблених для вирішення проблеми формування однорідних класів у довільній проблемній ділянці.

Методи аналізу даних, складовою частиною яких є методи кластерного аналізу, не використовують апріорних припущень про імовірнісну природу вихідної інформації і керуються лише евристичними міркуваннями щодо характеру і особливості досліджуваної сукупності об'єктів.

Задача кластеризації полягає в розбитті об'єктів з x на кілька кластерів, в яких об'єкти більш схожі між собою, ніж з об'єктами інших кластерів. У метричному просторі «схожість» звичайно визначають через відстань.

Аналіз останніх досліджень і публікацій. Аналізуючи великий перелік існуючих алгоритмів кластеризації даних можна сказати, що найбільший внесок у розвиток оброблення даних кластеризації зробили такі вчені, як М. Жамбю — в ієрархічному кластер-аналізі та відповідності, І. С. Єнюков — методи кластеризації об'єктів із категоризаційними ознаками, Л. Д. Мешалкін — класифікація та зниження розмірності, С. А. Айвазян — розроблення класифікації багатовимірних спостережень.

Формулювання мети статті. Метою статті є аналіз існуючих алгоритмів кластеризації даних для виокремлення переваг та недоліків кожного з алгоритмів.

ОСНОВНА ЧАСТИНА

Кластерний аналіз (автоматична класифікація, розпізнавання образів без учителя) посідає одне з центральних місць серед методів аналізу даних і є сукупністю підходів, методів і алгоритмів, призначених для відшукання деякого розбиття досліджуваної сукупності об'єктів на підмножини схожих між собою об'єктів.

Методи за способом оброблення даних [1] можна поділити на ієрархічні методи та неієрархічні методи.

Ієрархічна кластеризація у здобутті даних та статистиці — метод кластерного аналізу, який намагається побудувати ієрархію кластерів. Відомі такі стратегії побудови ієрархічної кластеризації [2]:

• агломератовий (об'єднувальний). Це підхід «знизу-вгору». Спочатку кожна точка має власний кластер, а далі пари кластерів об'єднуються під час сходження по ієрархії;

• розділювальний. Це підхід «згори-вниз». Спочатку всі точки містяться в єдиному кластері, потім відбувається рекурсивне розбиття під час руху вниз по ієрархії.

У процесі ієрархічної кластеризації виконується послідовне об'єднання менших кластерів у великі або поділ великих кластерів на менші [1].

Огляд алгоритмів ієрархічної кластеризації

Алгоритм кластеризації за допомогою представників (CURE — *Clustering Using REpresentatives*). Виконує ієрархічну кластеризацію з використанням набору визначальних точок для визначення об'єкта в кластер [3].

Призначення: кластеризація дуже великих наборів числових даних.

Обмеження: ефективний для даних низької розмірності, працює тільки на числових даних.

Переваги: виконує кластеризацію на високому рівні навіть за наявності викидів, виокремлює кластери складної форми і різних розмірів, має лінійно залежні вимоги до місця зберігання даних і тимчасову складність для даних високої розмірності.

Недоліки: є потреба в заданні порогових значень і кількості кластерів.

Алгоритм мінімального кістякового дерева (MST — *Algorithm based on Minimum Spanning Trees*). У цьому алгоритмі використовуються неорієнтовані графи та орієнтовані графи.

Призначення: кластеризація великих наборів довільних даних.

Обмеження: обмеження щодо пам'яті на GPU не дають можливості обчислювати великі графи.

Переваги: виокремлює кластери довільної форми, зокрема кластери опуклої і запалої форм, вибирає з кількох оптимальних вирішень найоптимальніше.

Недоліки: необхідно заздалегідь встановлювати кількість кластерів, алгоритм чутливий до первісного вибору центрів кластерів, можлива збіжність до локальних оптимумів.

Алгоритм збалансованого ітеративного скорочення і кластеризації за допомогою ієрархій (BIRCH — *Balanced Iterative Reducing and Clustering using Hierarchies*). У цьому алгоритмі передбачено двоетапний процес кластеризації [4].

Призначення: кластеризація дуже великих наборів числових даних.

Обмеження: робота з тільки числовими даними.

Переваги: двоступенева кластеризація, кластеризація великих обсягів даних, працює на обмеженому обсязі пам'яті, є локальним алгоритмом, може працювати при одному скануванні вхідного набору даних, використовуючи той факт, що дані неоднаково розподілено по простору, і обробляє ділянки з великою щільністю як єдиний кластер.

Недоліки: робота з тільки числовими даними, добре виокремлює тільки кластери сферичної форми, є необхідність у заданні порогових значень.

Огляд алгоритмів неієрархічної кластеризації

Алгоритм *k*-середніх (*k*-means). Алгоритм *k*-середніх будує *k*-кластери, розташовані на якможливо великих відстанях один від одного.

Призначення: основний тип задач, які розв'язує алгоритм *k*-середніх, це наявність припущень (гіпотез) щодо кількості кластерів, при цьому вони мають бути різні настільки, наскільки це можливо. Вибір числа *k* може базуватися на результатах попередніх досліджень, теоретичних міркуваннях або інтуїції.

Обмеження: невеликий обсяг даних.

Переваги: простота використання, швидкість використання, зрозумілість і прозорість алгоритму.

Недоліки: алгоритм занадто чутливий до викидів, повільна робота з великими обсягами даних, необхідно задавати кількість кластерів.

Алгоритм розділення навколо медоїдів (PAM — *partitioning around medoids*). Цей алгоритм досить схожий на *k*-means алгоритм, але PAM працює з медоїдами — об'єктами, котрі є частиною вихідної безлічі і представляють групу, до якої вони належать, тоді як *k*-means працює з центроїдами — штучно створеними об'єктами, що представляють кластер.

Обмеження: невеликий обсяг даних.

Переваги: простота використання, швидкість використання, зрозумілість і прозорість алгоритму, алгоритм менш чутливий до викидів порівняно з *k*-means.

Недоліки: необхідно задавати кількість кластерів, повільна робота з великими обсягами даних.

Алгоритм скупчення з нахилом (CLOPE — *Clustering with sLOPE*). Алгоритм CLOPE відбиває кластеризацію транзакційних даних.

Призначення: кластеризація величезних наборів категорійних даних.

Переваги: високі масштабованість і швидкість роботи, а також якість кластеризації, що досягається використанням глобального критерію оптимізації на основі максимізації градієнта висоти гістограми кластера. Він легко розраховується і інтерпретується. Під час роботи алгоритм зберігає в RAM невелику кількість інформації щодо кожного кластера і потребує мінімальної кількості сканувань набору даних. CLOPE автоматично підбирає кількість кластерів, причому це регулюється одним-єдиним параметром — коефіцієнтом відштовхування.

Недоліки: незначним недоліком може бути потреба у нормалізації даних для будь-яких категорійних даних.

ВИСНОВКИ

Розв'язання задач кластеризації може здійснюватися за допомогою різноманітних методів та алгоритмів, які мають свої переваги, недоліки і специфічні особливості реалізації. Вибір найкращого алгоритму за умов конкретного завдання має обґрунтовано здійснюватися особою, яка приймає рішення.

Окрім безпосереднього етапу кластеризації вкрай важливим є процеси інтерпретації та оцінювання здобутих результатів. На даній стадії головну роль відіграє експерт у досліджуваній предметній галузі, який на основі апріорних уявлень і знання ключових цільових показників може здійснити додаткову верифікацію результатів окрім використання формальних критеріїв, які відслідковуються в процесі реалізації алгоритмів.

Кластеризація даних є ефективним методом підготовки даних для подальшого їх використання експертною групою. Для досягнення максимального результату потрібен комплексний підхід до аналізу даних, що включає в себе як використання апріорних знань фахівців для попереднього оброблення даних і інтерпретації результатів, так і застосування спеціалізованих алгоритмів кластеризації.

Список використаної літератури

1. ***Чубукова І. А.*** *Data Mining: навч. посіб.: Інтернет-університет інформаційних технологій. БІНОМ: Лабораторія знань, 2006. 382 с.*

2. ***Rokach Lior, Oded Maimon.*** *«Clustering methods» Data mining and knowledge discovery handbook. Springer US, 2005. P. 321–352.*

3. ***Sudipto Guha, Rajeev Rastogi, Kyuseok Shim.*** *CURE: An Efficient Clustering Algorithm for Large Databases.*

4. ***Tian Zhang, Raghu Ramakrishnan, Miron Livny.*** *BIRCH: An Efficient Data Clustering Method for Very Large Databases.*

5. ***Akerkar R.*** *Big data computing. CRC Press, Taylor & Francis Group, Florida, USA, 2014.*

Е. С. Тихонов, К. В. Тихонова

АНАЛИЗ СУЩЕСТВУЮЩИХ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ ДАННЫХ. ПРЕИМУЩЕСТВА И НЕДОСТАТКИ

Анализ алгоритмов кластеризации данных все чаще становится популярной практикой, принятой многими организациям с целью создания ценной информации из больших объемов данных. Достаточно много исследований ставят себе целью организацию полученных данных в наглядные структуры. Фактически, кластерный анализ является набором различных алгоритмов классификации. Техника кластеризации применяется в самых различных областях, таких как психология, биология, педагогика, маркетинг, информационные технологии.

Кластеризация — это разделение данных на группы подобных объектов. Кластеризация проводится для понимания полученных данных, объем которых проблематичен для анализа человеком. Благодаря этому алгоритмы кластеризации стали инструментами мета-обучение для анализа исследовательских данных. Каждая группа, которая называется кластером, определяется как совокупность объектов, имеющих более высокую степень сходства друг с другом по сравнению с объектами, не принадлежащих одному набору. Тип используемого алгоритма кластеризации зависит от программы и набора данных, используемых в этом поле. Числовой набор данных сравнительно просто реализовать, поскольку данные неизменно реальные числа и могут использоваться для статистических приложений. Важно понимать разницу между кластеризацией (неподконтрольной классификацией) и дискриминационным анализом (контролируемой классификацией). На первом этапе исследователи совершенствовали некоторые алгоритмы кластеризации данных, на втором — внедряли новые, а на третьем — изучали и сравнивали различные алгоритмы кластеризации данных.

В статье проведена классификация и анализ существующих алгоритмов кластерного анализа, также рассмотрены преимущества и недостатки этих алгоритмов.

Ключевые слова: кластеризация; кластерный анализ; иерархическая кластеризиция; неиерархическая кластеризиция; алгоритм кластеризации с помощью представителей (CURE); алгоритм минимального скелетного дерева (MST); алгоритм сбалансированного итеративного сокращения и кластеризации с помощью иерархий (BIRCH); алгоритм *k*-средних (*k*-means); алгоритм разделения вокруг медоедов (PAM); алгоритм скопления с наклоном (CLOPE).

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