

В. Ф. Фролов, І. Н. Срибная, Г. Н. Власенко

ВЛИЯНИЕ ИОНОСФЕРЫ НА КАЧЕСТВО РАБОТЫ СПУТНИКОВЫХ НАВИГАЦИОННЫХ СИСТЕМ

Исследованы возможности коррекции формы и качества навигационных сигналов с учетом комплексных показателей электромагнитной волны и комплексного показателя преломления при прохождении указанных сигналов через слои ионосферы, постоянно изменяющей свое физическое состояние в зависимости от изменений космической погоды.

Ключевые слова: навигационная система; точность навигационных определений; ионосфера; радиоволны; навигационные сообщения; показатель преломления.

V. F. Frolov, I. M. Sribna, G. M. Vlasenko

INFLUENCE OF THE IONOSPHERE ON THE QUALITY OF SATELLITE NAVIGATION SYSTEMS

The article considers the possibility to adjust the shape and quality of navigation signals, taking into account the complex parameters of the electromagnetic wave and the complex index of refraction at their passage through the thickness of the ionosphere, which constantly changes its physical states, depending on the change of space weather.

Keywords: navigation system; accuracy of navigation definitions; ionosphere; radio waves; navigation notifications; refractive index.

УДК 004.891.3

Г. І. ГАЙДУР, канд. техн. наук, доцент;

Є. В. ПРИЛЄПОВ, аспірант,

Державний університет телекомунікацій, Київ

Алгоритм визначення центроїдів масивів даних у парадигмі IoT

Розглянуто основні алгоритми кластеризації та підходи до розв'язання актуальних завдань кластерного аналізу. Докладно описано популярний алгоритм K-середніх і з'ясовано його переваги та недоліки. Запропоновано поліпшений алгоритм K-середніх і обґрунтовано його ефективність.

Ключові слова: алгоритм; аналіз; обробка; ідентифікація; центроїд; кластеризація.

Вступ

Останніми роками зростає потреба в обробці дедалі більших обсягів даних, а отже, і в інтелектуальному їх аналізі. Цей напрямок включає в себе методи, відмінні від методів класичного аналізу, передусім завдяки проникненню нових ідей, припущенням теорії штучного інтелекту.

Кластеризація, або кластерний аналіз, — це автоматичне розбиття деякої множини даних на групи подібних між собою елементів. Такі групи називають *кластерами*.

Головне завдання кластерного аналізу — виокремлення груп об'єктів, що мають певну спільну ознаку і максимально відмінні від об'єктів інших класів. Такий аналіз широко застосовується в інформаційних системах для відшукування закономірностей, якими характеризуються наявні дані.

Аналіз публікацій

Зауважимо, що нині застосовуються різні підходи до розв'язання завдань кластерного аналізу. В основу цих підходів покладено різні уявлення про завдання, спосіб використання специфічної для кожної предметної області додаткової інформації тощо. Розглянемо найчастіше використовувані підходи з таким застереженням. Пропонована класифікація не є канонічною, і деякі методи можуть бути розроблені на основі комбінації різних підходів [1–3].

Перелічимо шість основних підходів:

1. Імовірнісний підхід.
2. Підхід на базі теорії графів.
3. Ієрархічний підхід.
4. Підхід, що спирається на поняття найближчого сусіда.
5. Підхід, який використовує алгоритми кластерного аналізу.
6. Підхід на базі штучних нейронних мереж.

Постановка завдання

Незважаючи на численні дослідження в галузі кластерного аналізу, маємо тут чимало актуальних проблем [4].

1. Обґрунтування якості результатів.

Проблема полягає в тому, що один і той самий об'єкт може бути класифікований із включенням у різні групи незалежно від його внутрішніх властивостей, а згідно з різними експертними даними або різною побудовою системи. Для уникнення цього необхідно розробляти і вводити актуальні критерії якості.

2. Вибір метрики.

Аналіз великої кількості різнотипних і навіть однотипних даних породжує методологічну проблему вибору метрик, спричинявану можливою нерозрізнованістю відстаней.

3. Нелінійність взаємозв'язків.

Класичні методи зниження розмірності в кластерному аналізі мають на меті встановити лінійний взаємозв'язок між змінними. Для пошуку складніших залежностей необхідно переходити до ядерних методів.

4. Проблема пошуку глобального екстремуму функції, якою подається критерій якості.

Критерій якості — це функція, що, як правило, залежить від багатьох нелінійних факторів і має низку локальних екстремумів. Для побудови кластерів необхідно розв'язати складну комбінаторну задачу пошуку оптимального варіанта класифікації. Адже алгоритм повного перебору варіантів надто трудомісткий і експоненціально залежний від розмірності. Якщо кількість груп заздалегідь невідома, то завдання перебору додатково ускладнюється.

5. Нестійкість результатів кластеризації.

Результати групування можуть сильно змінюватися залежно від вибору початкових умов, порядку об'єктів, параметрів роботи алгоритмів. Чимало авторів пропонують способи підвищення стійкості розв'язків щодо групування, які спираються на застосування ансамблів алгоритмів. При цьому використовуються результати групування, отримані згідно з різними алгоритмами або одним алгоритмом, але з різними параметрами налаштування та різними підсистемами змінних, тощо. Після побудови відповідного ансамблю здійснюється пошук підсумкового колективного рішення.

6. Недостатність інформації про об'єкт.

Існує проблема щодо абстрактно формалізованих областей, в яких ускладнюється побудова моделі об'єкта. У такому разі застосування алгоритмів, що ґрунтуються на поданні класу як набору розподілених у просторі змінних, практично неприйнятне.

7. Труднощі, пов'язані з тлумаченням результатів.

Коли йдеться про будь-який алгоритм аналізу даних, важливо, наскільки зрозумілі його результати. Для уточнення інтерпретації рішень можна використовувати логічні моделі, такі як моделі розпізнавання образів і прогнозування кількісних показників, характерні для методів побудови дерев рішень або логічних функцій.

Основна частина

Популярний алгоритм K -середніх, розроблений понад чотири десятиріччя тому [5], являє собою надзвичайно простий, але й не досить точний спосіб кластеризації в класичній реалізації. Він передбачає розбиття масиву елементів векторного простору на заздалегідь відому кількість k кластерів, що має на меті мінімізувати середньоквадратичне відхилення в точках кожного кластера щодо центра його мас. Головна ідея — перерахунок на кожному кроці (для кожної ітерації) середньоква-

дратичного відхилення від центра мас кластера, отриманого на попередньому кроці. Далі вектори знову розбиваються на кластери відповідно до того, який із нових центрів дає найменше відхилення. Алгоритм завершується, коли на деякій ітерації не відбувається зміни кластера.

Нагадаємо, що метод кластеризації K -середніх (K -means) — це неієрархічний метод, який дозволяє розбити масив об'єктів на задану кількість кластерів згідно з доволі тонким критерієм, таким як статистика Фішера: відношення міжкластерної дисперсії до дисперсії внутрішньокластерної. Зрештою розбиття здійснюється так, щоб взаємовплив змінних усередині кластера був мінімальний, а між кластерами — максимальний.

Метод K -середніх, на відміну від ієрархічних методів, не передбачає розгляд всіх можливих варіантів кластеризації. Починається робота з якогось, як правило, довільного варіанта, а далі відбувається послідовна зміна розбиття, що триває доти, доки критерій оптимізації не припиняє змінюватися. Через це, якщо алгоритм на своєму шляху зустрівся з неглобальним максимумом критерію, метод може припинити свою роботу, не дійшовши до глобально оптимального розбиття. Тому при його використанні особливої важливості набувають початкові умови.

Алгоритм K -середніх включає в себе чотири кроки [5]:

1. Здійснити вибірки з k точок, що являє собою початковий центроїд кластера (випадкові k з n об'єктів або взагалі k випадкових точок).

2. Співвіднести кожний об'єкт кластера з найближчим центроїдом.

3. Обчислити загальну кількість центроїдів.

4. Якщо критерій зупинки алгоритму не задовольняється, повернутися до кроку 2.

Як критерій зупинки зазвичай вибирають одну з двох умов:

а) відсутність переміщення об'єктів із кластера в кластер на кроці 2.

б) мінімальна зміна середньоквадратичної похибки.

Якість алгоритму залежить від початкової позиції центроїдів.

Переваги методу K -середніх:

- простота та швидкість виконання;

- особлива зручність для кластеризації великої кількості спостережень, не притаманна методу ієрархічного кластерного аналізу.

Недоліки методу K -середніх:

- якість алгоритму залежить від початкової позиції центроїдів;

- алгоритм чутливий до будь-яких збурень;

- кількість кластерів має бути заздалегідь визначена дослідником.

Функцію продуктивності алгоритму K -середніх можна подати такою формулою:

$$J_{K_m} = \sum_{j=1}^N \sum_{i=1}^K \|x_i - m_j\|^2, \quad (1)$$

де K — кількість кластерів даних; N — кількість точок даних у кластері; x_i — точка, що відповідає i -му значенню даних; m_j — центроїд кластера; $\|x_i - m_j\|^2$ — евклідова відстань між x_i та m_j .

Спираючись на алгоритм K -середніх і беручи до уваги деякі інші алгоритми кластеризації, скажімо алгоритм нечіткої кластеризації C -середніх, K -середніх ++, зважений алгоритм K -середніх, обернено зважений алгоритм K -середніх тощо, вдалося розробити поліпшену версію алгоритму ідентифікації центроїдів у масивах даних.

Розглянемо функцію продуктивності виду

$$J_I = \sum_{i=1}^N \sum_{k=1}^K \frac{1}{\|x_i - m_k\|^P}, \quad (2)$$

для якої справджуються такі співвідношення:

$$\frac{\partial J_I}{\partial m_k} = \sum_{i=1}^N P(x_i - m_k) \frac{1}{\|x_i - m_k\|^{P+2}}, \quad (3)$$

$$\frac{\partial J_I}{\partial m_k} = 0 \rightarrow m_k = \frac{\sum_{i=1}^N b_{ik} x_i}{\sum_{i=1}^N b_{ik}}, \quad (4)$$

де b_{ik} — відстань між точками значень даних та центроїда кластера,

$$b_{ik} = \frac{1}{\|x_i - m_k\|^{P+2}} \quad (5)$$

P — випадкове додатне число, $P > 0$.

Частинна похідна від J_I за m_k може бути використана для максимізації функціональності J_I .

Отже, реалізація (5) завжди переміщатиме m_k до найближчої точки даних, спрямовуючи J_I до нескінченності. Проте реалізація (5) не буде ідентифікувати кластери, оскільки центроїди завжди прямують до найближчої точки даних. Але перевага цієї функції полягає в тому, що вона не залишає жодного центроїда віддаленим від даних: усі центроїди наближені до даних.

Рецензент: доктор техн. наук, професор Л. Н. Беркман, Державний університет телекомунікацій, Київ.

Г. І. Гайдур, Е. В. Прилепов

АЛГОРИТМ ОПРЕДЕЛЕНИЯ ЦЕНТРОИДОВ МАССИВОВ ДАННЫХ В ПАРАДИГМЕ IoT

Рассмотрены основные алгоритмы кластеризации и определены подходы к решению задач кластерного анализа. Проанализированы актуальные проблемы кластерного анализа. Подробно описан популярный алгоритм K -средних и отмечены его преимущества и недостатки. Предложен улучшенный алгоритм K -средних и дано обоснование его эффективности.

Ключевые слова: алгоритм; анализ; обработка; идентификация; центроид; кластеризация.

H. I. Haidur, Ye. V. Pryliepov

CENTROID DETECTION ALGORITHM FOR DATA ARRAYS IN IoT PARADIGM

The main algorithms of clustering and approaches to solving cluster analysis problems are considered. The analysis of actual problems of cluster analysis is carried out. Disassembled the popular algorithm of K -means and its main advantages and disadvantages. The use of the improved K -means algorithm is proposed and the effectiveness of this method is substantiated.

Keywords: algorithm; analysis; processing; identification; clustering.

Щоб підвищити здатність цього алгоритму ідентифікувати кластери, не втрачаючи властивості переміщення центроїдів усередину кластерів даних, змінимо b_{ik} у (5) на такий вираз:

$$b_{ik} = \frac{|x_i - m_{k^*}|^{P+2}}{|x_i - m_k|^{P+2}}, \quad (6)$$

де m_{k^*} — найближчий до x_i центроїд.

У результаті цієї заміни поліпшилась максимізація J_I завдяки переміщенню центроїдів до вільних точок даних (або кластерів) замість переміщення до найближчої точки даних (або локального кластера).

Зауважимо, що формула (6) впливає з такої функції ефективності:

$$J_{IWC} = \sum_{i=1}^N \left[\sum_{j \neq k^*}^K \frac{1}{|x_i - m_j|^P} \right] \left(|x_i - m_k|^{P+2} \right), \quad (7)$$

де $k^* = \arg \min_{k=1}^K (|x_i - m_j|)$.

Висновки

У запропонованій версії алгоритму K -середніх вдалося досягти істотних переваг, передусім значного зменшення чутливості до початкової ініціалізації центроїдів і здатності алгоритму ідентифікувати кластери. При цьому не втрачається властивість щодо переміщення центроїдів усередину кластерів даних, а отже, усі центроїди наближені до даних.

Список використаної літератури

1. **Gowda, K. C.** Agglomerative clustering using the concept of mutual nearest neighborhood / K. C. Gowda, G. Krishna // *Pattern Recognition*.— 1977. — V. 10.— P. 105–112.
2. **Jain, A. K.** Artificial neural networks: A tutorial / A. K. Jain // *IEEE Computer*.— 1996.— V. 29.— P. 31–44.
3. **Zahn, C. T.** Graph-theoretical methods for detecting and describing gestalt clusters / C. T. Zahn // *IEEE Trans. Comput.*— 1971.— C-20.— P. 68–86.
4. **Загоруйко, Н. Г.** Прикладные методы анализа данных и знаний / Н. Г. Загоруйко.— М., 1999.
5. **Котов, А.** Кластеризация даних / А. Котов, Н. Красильников.— К., 2006.