НЕЙРОІНФОРМАТИКА ТА ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ

NEUROINFORMATICS AND INTELLIGENT SYSTEMS

НЕЙРОИНФОРМАТИКА И ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

УДК 004.8:004.032.26

НЕЧІТКА ДОВІРЧА КЛАСТЕРИЗАЦІЯ ДАНИХ НА ОСНОВІ АНАЛІЗУ ЩІЛЬНОСТІ РОЗПОДІЛУ ДАНИХ ТА ЇХ ПІКІВ

Бодянський Є. В. – д-р техн. наук, професор, професор кафедри штучного інтелекту, Харківський національний університет радіоелектроніки, Харків, Україна.

Плісс І. П. – канд. техн. наук, провідний науковий співробітник ПНДЛ АСУ, Харківський національний університет радіоелектроніки, Харків, Україна.

Шафроненко А. Ю. – канд. техн. наук, доцент, доцент кафедри інформатики, Харківський національний університет радіоелектроніки, Харків, Україна.

Калиниченко О. В. – канд. техн. наук, доцент, доцент кафедри програмної інженерії, Харківський національний університет радіоелектроніки, Харків, Україна.

АНОТАЦІЯ

Актуальність. Задача кластеризації – класифікації без вчителя масивів даних займає важливе місце в інтелектуальному аналізі даних. Для вирішення цієї задачі на цей час запропоновано безліч підходів, що відрізняються між собою як апріорними припущеннями що до характеру даних у масивах, що досліджуються та аналізуються, так і математичним апаратом, що полягає в основі тих або інших методів, однак вирішення задач кластеризації ускладнюють велика розмірність векторів спостережень, що аналізуються, їх збуреність та забрудненість різного типу завадами та пропусками, можливою складною формою кластерів, тощо.

Мета. Мета роботи полягає у запровадженні процедури нечіткої кластеризації, що об'єднує в собі переваги методів, заснованих на аналізі щільностей розподілу даних та їх піків, характеризуються високою швидкодією та може ефективно працювати за умов перетинних класів.

Метод. Введено метод нечіткої кластеризації масивів даних, що базується на ідеях аналізу щільностей розподілу цих даних, їх піків та довірчого нечіткого підходу. Перевагою запропонованого підходу є скорочення часу вирішення оптимізаційних задач, пов'язаних з відшуканням атракторів функцій щільностей, оскільки кількість звернень до блоку оптимізації визначається не обсягом аналізованого масиву, а кількістю піків щільностей цього ж масиву.

Результати. Метод є досить простим у чисельній реалізації і не критичним до вибору оптимізаційної процедури. Результати експериментів підтверджують ефективність пропонованого підходу в задачах кластеризації за умов перетинних кластерів та дозволяють рекомендувати запропонований метод для використання на практиці для вирішення проблем автоматичної кластеризації великих даних.

Висновки. Введено метод нечіткої кластеризації масивів даних, що базується на ідеях аналізу щільностей розподілу цих даних, їх піків та довірчого нечіткого підходу. Перевагою запропонованого підходу є скорочення часу вирішення оптимізаційних задач, пов'язаних з відшуканням атракторів функцій щільностей, оскільки кількість звернень до блоку оптимізації визначається не обсягом аналізованого масива, а кількістю піків щільностей цього ж масиву. Метод є досить простим у чисельній реалізації і не критичним до вибору оптимізаційної процедури. Результати експериментів підтверджують ефективність запропонованого підходу в задачах кластеризації за умов перетинних кластерів.

КЛЮЧОВІ СЛОВА: нечітка кластеризація, правдоподібна кластеризація, піки щільності розподілу даних.

АБРЕВІАТУРА

DI індекс Дана;
DBI індекс Девіса-Болдіна;
CA кластерна точність;
GA генетичний алгоритм;
SA імітований відпал.

НОМЕНКЛАТУРА

X – матриця набору даних;
k – номер вектору-спостереження;
i – номер атрибуту вектора-спостереження;
j – номер класу;

DOI 10.15588/1607-3274-2022-3-6

$x(k)$ – вектор-спостереження;

x_j^* – атрактор;

l, q – номери кластерів;

m – кількість неперетинних класів;

μ_j – рівень нечіткої належності j-го кластеру;

D – матриця відстаней між спостереженнями;

d – відстань між спостереженнями;

ρ – вектор локальної щільності;

c – центроїд кластера;

δ_k^* – точка з максимальною щільністю;

Cr – рівень правдоподібності;

$\delta(k)$ – відстань від спостереження $x(k)$ до точки з більш високою щільністю;

σ – параметр ширини – відстань зрізу в прийнятій метриці функції впливу;

$f_G^{\tilde{x}}(x)$ – гаусівська функція впливу;

$f_E^{\tilde{x}}(x)$ – функція Єпанечнікова;

$f_C^{\tilde{x}}(x)$ – функція Коши;

x_j^P – піки-центроїди кластерів.

ВСТУП

Задача кластеризації – класифікації без вчителя масивів даних займає важливе місце у інтелектуальному аналізі даних (Data Mining, Data Stream Mining, Big Data Mining), а для її вирішення на цей час запропоновано безліч підходів, що відрізняються між собою як апріорними припущеннями що до характеру даних у масивах, що досліджуються та аналізуються, так і математичним апаратом, що полягає в основі тих або інших методів [1–4]. Дещо особливе місце тут займають методи нечіткої кластеризації [5, 6], що призначені для роботи за умов, коли кластери, що формуються можуть довільним чином перетинатися у просторі ознак. Зрозуміло також, що вирішення задач кластеризації ускладнюють велика розмірність векторів спостережень, що аналізуються, їх збуреність та забрудненість різного типу завадами та пропусками, можливою складною формою кластерів, тощо.

Об'єкт дослідження швидка нечітка кластеризація даних на основі піків щільності розподілу даних.

Предмет дослідження процедура аналізу піків щільності розподілу даних.

Мета роботи полягає у запровадженні процедури нечіткої кластеризації, що об'єднує в собі переваги методів, заснованих на аналізі щільностей розподілу даних та їх піків, характеризуються високою швидкодією та можна ефективно працювати за умов перетинних класів.

1 ПОСТАНОВКА ЗАВДАННЯ

Вихідною інформацією для вирішення задачі кластеризації є масив даних

$$X = \{x(1), x(2), ..., x(k), ..., x(N)\}, x(k) = \{x_i(k)\} \in R^n,$$

при цьому компоненти цих даних попередньо передоброблені так, щоб вони належали деякому обмеженому інтервалу, наприклад $-1 \le x_i(k) \le 1 \forall i,k$. На основі цих даних формується $(n \times N)$ матриця «об'єкт-властивість», елементи якої опрацьовуються прийнятим алгоритмом.

2 ОГЛЯД ЛІТЕРАТУРИ

Для вирішення цієї задачі одними з найбільш ефективних є алгоритми, що базуються на аналізі щільностей розподілу даних у вихідних масивах, серед яких можна відзначити DENCLUE [7] та його модифікації [8–10], призначені для вирішення задач кластеризації великих масивів медіаданих, спотворених збуреннями різної природи, та класів складної форми. В той же час DENCLUE характеризується досить низькою швидкодією, оскільки його використання пов'язане з необхідністю багатократного вирішення задачі оптимізації за допомогою градієнтних процедур. Більш швидкими є алгоритми, засновані на аналізі піків щільностей [11], однак тут в якості центроїдів-прототипів кластерів виступають спостереження вихідного масиві даних. Якщо кластери мають складну неопуклу форму, то їх центоїди можуть не співпадати із спостереженнями і, більш того, розташовуватися на значній відстані від них. Крім того, як DENCLUE, так і піковий алгоритм є чіткими процедурами, тобто призначені для роботи за умов коли кластери не перетинаються, а кожна точка – векторне спостередення може належати лише одному класу. Тому є доцільним ввести у розгляд процедуру нечіткої кластеризації, що об'єднує в собі переваги методів, заснованих на аналізі щільностей розподілу даних та їх піків, характеризуються високою швидкодією та можна ефективно працювати за умов перетинних класів.

3 МАТЕРІАЛИ І МЕТОДИ

Базовими поняттями, що використовуються у подальшому, є функція впливу, функція щільності даних у вибірці, атрактори щільності, що відповідають екстермумам – максимумам функції щільності, що досягаються у точках вибірки в околі атракторів щільності.

Для будь якої точки $\tilde{x}$ з масиву X її базова функція впливу $f_B^{\tilde{x}}(x) = f(x, \tilde{x})$ є деякою ядерною дзвонуватою функцією функцією, серед яких автори методу [7–9] відзначають, так звану, прямокутну хвилеву функцію впливу

$$f_S^{\tilde{x}}(x) = \begin{cases} 0, \text{ якщо } d(x, \tilde{x}) > \sigma, \\ 1 \text{ інакше} \end{cases}$$

(тут d – відстань у прийнятій метриці, зазвичай евклідовій, σ – параметр ширини – відстань зрізу в прийнятій метриці функції впливу) та гаусівську функцію впливу

DOI 10.15588/1607-3274-2022-3-6

$$f_G^{\tilde{x}}(x)=\exp\left(-\frac{d^2(x,\tilde{x})}{2\sigma^2}\right)=\exp\left(-\frac{\|x-\tilde{x}\|^2}{2\sigma^2}\right),$$

що є найбільш популярною, завдяки зручності обчислення її градієнта.

Нескладно бачити, що в якості функцій впливу можуть також бути використана функція Коші, що часто виникає у задачах нечіткої кластеризації [12]

$$f_C^{\tilde{x}}(x)=\left(1+\frac{d^2(x,\tilde{x})}{\sigma^2}\right)^{-1}=\left(1+\frac{\|x-\tilde{x}\|^2}{\sigma^2}\right)^{-1},$$

та функція Єпанечнікова [13]

$$f_E^{\tilde{x}}(x)=\left[1-\frac{d^2(x,\tilde{x})}{2\sigma^2}\right]_+=\left[1-\frac{\|x-\tilde{x}\|^2}{2\sigma^2}\right]_+$$

(тут $[\bullet]_+=\max\{0,\bullet\}$), цей градієнт має просту форму

$$\nabla_x f_E^{\tilde{x}}(x)=\left[\frac{\tilde{x}-x}{\sigma^2}\right]_+,$$

де операція проєктування не достатній ортант $[\bullet]_+$ реалізується покомпонентно.

Функція щільності розподілу даних у масиві X, що містить N спостережень, формується на основі N функцій впливу у вигляді

$$f^x(x)=\sum_{k=1}^{N} f\left(x,x(k)\right) \qquad (1)$$

і є близькою за суттю до Парзенівських вікон [14] та оцінок Надарая-Ватсона [15, 16].

Власне процедура кластеризації полягає у відшуканні максимумів функції $f^x(x)$, що задовільняють умові

$$f^x\left(x,x^*\right)>\xi, \qquad (2)$$

де ξ – деякий поріг, що визначає, який із відшуканих атракторів є значущим, тобто «фільтрує» окремі аномальні викиди у вибірці X та виключає із розгляду «міні-кластери», що містять занадто мало спостережень. Зрозуміло, що чим більше значення ξ, тим менша кількість значущих кластерів буде сформована.

Для відшукання атракторів – екстремумів-максимумів функції щільності розподілу даних $f^x(x)$ зазвичай використовується градієнтна процедура оптимізації [17], що може бути записана у вигляді

$$x(k)=x_0;$$

$$x^l=x^{l-1}+\eta^l\frac{\nabla f^x\left(x,x^{l-1}\right)}{\left\|\nabla f^x\left(x,x^{l-1}\right)\right\|},$$

$$l=0,1,2,...;\ \forall k=1,2,...,N, \qquad (3)$$

де η^l – параметр кроку пошуку, що визначає швидкість збіжності алгоритму. Різні модифікації DENCLUE пов'язані саме з намаганнями пришвидшити процес оптимізації прийнятої функції щільності [8, 10]. Тут же слід відзначити, що використання гаусіанів в якості функцій впливу пов'язане з простою формою їх градієнтів, оскільки

$$f_G^x\left(x,x^l\right)=\sum_{k=1}^{N}\left(x^l-x\right)f_G^{x^l}(x)=$$

$$=\sum_{k=1}^{N}\left(x^l-x\right)\exp\left(-\frac{\left\|x-x^l\right\|^2}{2\sigma^2}\right).$$

Помітимо також, що функції Єпанечнікова мають ще більш просту форму градієнта

$$\nabla f_E^x\left(x,x^l\right)=\sum_{k=1}^{N}\left[\frac{x^l-x}{2\sigma^2}\right]_+.$$

Процес оптимізації починається з кожної точки $x(k)$ масива даних X і закінчується відшуканням всіх екстремумів – максимумів функції щільності (1), що задовольняють нерівності (2).

Зрозуміло, що чим більше обсяг вибірки X, тим більше разів N повинна запускатися процедура оптимізації – пошуку атракторів. Пришвидшити процес кластеризації можна, скоротивши кількість запусків цієї процедури.

Тому пропонується починати процес пошуку атракторів не з кожної точки масиву даних X, а з, так званих, піків щільності [11] цього масиву. Для знаходження цих піків у розгляд вводиться два параметри: $\rho(k)$ – локальна щільність та $\delta(k)$ – відстань від спостереження $x(k)$ до точки з більш високою щільністю. Крім того, аналогічно DENCLUE використовується відстань зрізу σ, що зазвичай задається та варіюється користувачем для отримання потрібної точності вирішення задачі.

Процес пошуку піків щільності починається з того, що на основі вихідної $(n\times N)$ – матриці «об'єкт-властивість» формується $(N\times N)$ матриця відстаней між спостереженнями

$$D=\left\{d\left(x(k),x(q)\right)\right\},$$

$$d\left(x(k),x(q)\right)=\left\|x(k)-x(q)\right\|\forall k,q.$$

DOI 10.15588/1607-3274-2022-3-6

На основі цієї матриці формується $(N\times 1)$-вектор локальних щільностей $\rho=\{\rho(k)\}\in R^N$:

$$\rho(k)=\sum_{k=1}^{N}\chi\left(d\left(x(k),x(q)\right)-\sigma\right),$$

де

$$\chi(d)=\begin{cases}1, \text{ if } d<0,\\ 0, \text{ else.}\end{cases}$$

Тут відстань зрізу σ є найбільш впливовим параметром, що визначає якість кластеризації та обирається з суто емпіричних міркувань. Тут слід відмітити, що автори пікового алгоритму [11] радять обирати цю відстань так, щоб вона «накривала» $0,01N-0,02N$ спостережень з масиву, що аналізується.

Після цього розраховується вектор мінімальних відстаней

$$\delta(k)=\min_{\forall q,\rho(q)>\rho(k)}\left\{d\left(x(k),x(q)\right)\right\},$$

а для спостереження з мінімальною щільністю $\delta^*(k)$ покладається

$$\delta^*(k)=\max\left\{d\left(x(k),x(q)\right)\right\}.$$

На базі цієї інформації формуються піки-центроїди кластерів $x_j^P, j=1,2,...,m$, при цьому в якості цих піків-центроїдів обираються спостереження з найбільш високою щільністю, тобто центроїди згідно з цим підходом є деякі із спостережень вихідної вибірки. В той же час в ситуаціях, коли кластери мають досить складну форму, центроїд може не співпадати з жодною із точок $x(k)$. Тому пропонується після знаходження всіх піків $x_j^P, j=1,2,...,m$ запускати процедуру оптимізації (3) не з точок $x(k), k=1,2,...,N$, а тільки з піків $x_j^P, j=1,2,...,m$, кількість яких є значно меншою ніж обсяг вибірки X, тобто

$$m<<N.$$

Процедури DENCLUE та пікові кластеризації є алгоритмами чіткої кластеризації, тобто апріорно припускається, що кластери, які формуються за їх допомогою, не перетинаються у просторі ознак. Якщо ж ці класи «накривають» один одного, що досить часто зустрічається у реальних задачах, то мусять бути застосовані алгоритми нечіткої (фаззі) кластеризації [5, 6], що базуються на двох основних підходах: імовірнісному та можливісному. Кожен з цих підходів має свої переваги та недоліки, яких позбавлений, так званий, довірчий підхід до нечіткої кластеризації [18, 19].

Згідно з цим підходом для кожного з атракторів (або піків) $x_j^*, j=1,2,...,m$ та спостережень $x(k), k=1,2,...,N$ розраховуються рівні нечіткої належності або у загальній формі [5]

$$\mu_j(k)=\frac{d^{-2}\left(x_j^*,x(k)\right)}{\sum_{r=1}^{m}d^{-2}\left(x_j^*,x(k)\right)}=\frac{\left\|x_j^*-x(k)\right\|^{-2}}{\sum_{r=1}^{m}\left\|x_j^*-x(k)\right\|^{-2}},\qquad(4)$$

або після деяких перетворень [20, 21]

$$\mu_j(k)=\left(1+\frac{d^{-2}\left(x_j^*,x(k)\right)}{\sigma_j^2}\right),\qquad(5)$$

де

$$\sigma_j^2=\left(\sum_{\substack{r=1\\ r\neq j}}^{m}d^{-2}\left(x_j^*,x(k)\right)\right)^{-1},$$

тобто знов-таки виникає функція щільності розподілу Коші, що може бути використана в якості функції впливу у DENCLUE.

Оцінки (4), (5) пов'язані з так званою, ймовірнісною нечікою кластеризацією. На основі оцінок можуть бути розраховані рівні довіри отриманих результатів за допомогою співвідношень [18, 19]:

$$\begin{cases}Cr_j(k)=\frac{1}{2}\left(\mu_j^*(k)+1-\sup\mu_r^*(k)\right),\\ \mu_j^*(k)=\frac{\mu_j(k)}{\sup\mu_r(k)}.\end{cases}$$

Таким чином, введена процедура нечіткої кластеризації, що базується на аналізі щільностей розподілу даних та їх піків, дозволяє скоротити час вирішення задачі за рахунок зменшення кількості звернень до блоку оптимізації, що відшукує екстремуми-атрактори прийнятої функції щільності.

4 ЕКСПЕРИМЕНТИ

Дослідження методу нечіткої довірчої кластеризації даних на основі аналізу щільності розподілу даних та їх піків (NCrCP) проводились на двох навчальних вибірках UCI репозиторію Page Blocks та Spambase. В Таблиці 1 продемонстровані основні характеристики наборів даних.

DOI 10.15588/1607-3274-2022-3-6

Таблиця 1 – Зразки даних

Назва вибірки	Кількість спостережень	Кількість атрибутів	Кількість кластерів
Page Blocks	5472	10	5
Spambase	4601	57	2

Page Blocks – набір даних, який містить інформацію про класифіковані блоки макету сторінки в документі, який було виявлено процесом сегментації.

Spambase – ілюструє класифіковану електронну пошту як спам та не спам.

На рисунках 1 та 2 продемонстровані набори вибірок даних, що аналізуються.

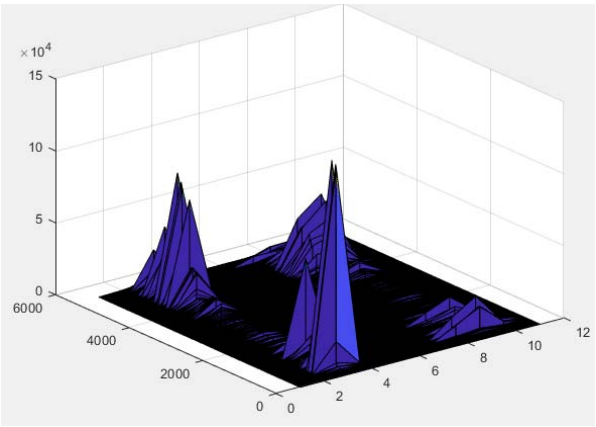

Рисунок 1 – Навчальна вибірка Page Blocks

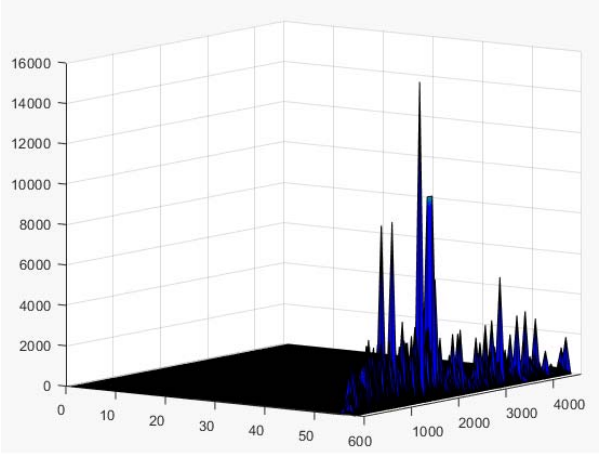

Рисунок 2 – Навчальна вибірка Spambase

Робота запропонованого методу перевірялась за допомогою декількох класичних показників якості кластеризації, а саме індекс Дана (DI), індекс Девіса-Болдіна (DBI) та кластерна точність (CA).

Індекс Дана (DI) – цей індекс оцінює ступінь поділу між спостереженнями одного кластера, тобто внутрішню схожість спостережень в кластері. Чим вище значення, тим краща кластеризація.

Індекс Девіса-Болдіна (DBI) – цей індекс, як DI, також оцінює ступінь поділу між кластерами (межкластерна несхожість), найменше значення вказує на кращу кластеризацію.

Кластерна точність (CA) – вимірює відсоток правильно класифікованих об'єктів у кластері на основі попередньо визначених міток класів. Цей індекс не працює з немаркованою базою даних, високе значення вказує на найкращу якість кластеризації.

Порівняльний аналіз проводився з більш відомими методами кластеризації даних, такими як DENCLUE-SA (імітований відпал), DENCLUE та DENCLUE-GA (генетичний алгоритм).

5 РЕЗУЛЬТАТИ

Результати кластеризації тестових даних різними методами кластеризації представлено на рисунках 3 і 4, які демонструють якісні характеристики кластеризації.

Як видно із гістограм, можна зробити висновки, що запропонований метод нечіткої довірчої кластеризації даних на основі аналізу щільності розподілу даних та їх піків (NCrCP) кластеризує дані якісніше за більшістю якісних характеристик кластеризації.

Зокрема, якість методу кластеризації повинна відповідати вимогам не тільки якості кластеризації, а й швидкості і простоти з точки зору математичних розрахунків. Тому був проведений аналіз швидкості розрахунків кластеризації методу нечіткої довірчої кластеризації даних на основі аналізу щільності розподілу даних та їх піків та вище згаданих методів кластеризації. В табл. 2 наведений порівняльний результат швидкості роботи методів кластеризації.

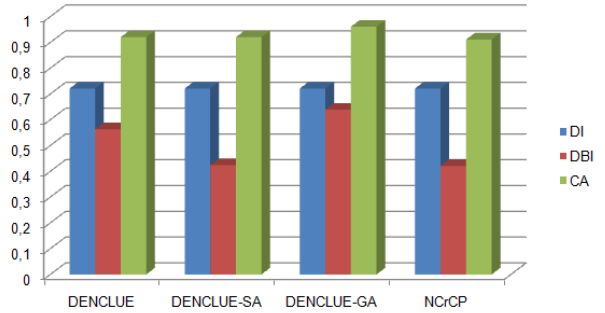

Рисунок 3 – Показники якості кластеризації Page Blocks

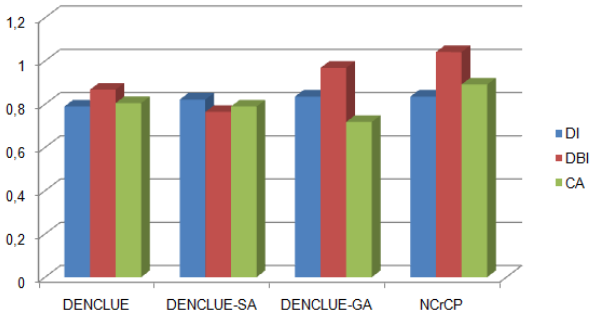

Рисунок 4 – Показники якості кластеризації Spambase

DOI 10.15588/1607-3274-2022-3-6

Таблиця 2 – Порівняння роботи алгоритмів за часом (с.)

Data	DENCLUE	DENCLUE-SA	DENCLUE-GA	NCrCP
Page Blocks	71	107	158	**7**
Spambase	1285.9	1347	574	**29**

6 ОБГОВОРЕННЯ

За результатами експериментальних досліджень та аналізу отриманих результатів, можна зробити висновок, що запропонований метод нечіткої довірчої кластеризації даних на основі аналізу щільності розподілу даних та їх піків (NCrCP) кластеризує дані якісніше за більшістю якісних характеристик кластеризації, демонструє гарні результати роботи порівняно із класичними методами кластеризації: DENCLUE-SA, DENCLUE та DENCLUE-GA.

Можна також зауважити, що метод не поступається кластерною точністю та швидкістю, що само за себе говорить про якість отриманих результатів кластеризації.

ВИСНОВКИ

Введено метод нечіткої кластеризації масивів даних, що базується на ідеях аналізу щільностей розподілу цих даних, їх піків та довірчого нечіткого підходу.

Перевагою запропонованого підходу є скорочення часу вирішення оптимізаційних задач, пов'язаних з відшуканням атракторів функцій щільностей, оскільки кількість звернень до блоку оптимізації визначається не обсягом аналізованого масива, а кількістю піків щільностей цього ж масиву. Метод є досить простим у чисельній реалізації і не критичним до вибору оптимізаційної процедури. Результати експериментів підтверджують ефективність пропонованого підходу в задачах кластеризації за умов перетинних кластерів.

Наукова новизна: вперше запропонований метод нечіткої кластеризації даних, що базується на ідеях аналізу щільностей розподілу цих даних, їх піків та довірчого нечіткого підходу.

Практичне значення: результати експерименту дозволяють рекомендувати запропонований метод для використання на практиці для вирішення проблем автоматичної кластеризації великих даних.

Перспективи подальших досліджень методи нечіткої кластеризації даних для широкого класу практичних проблем.

ПОДЯКА

Робота виконана в рамках науково-дослідного проєкту державного бюджету Харківського національного університету радіоелектроніки «Розробка методів та алгоритмів комбінованого навчання глибинних нейро-нео-фаззі систем за умов короткої навчальної вибірки» (номер державної реєстрації 0122U001701).

ЛІТЕРАТУРА/ЛИТЕРАТУРА

1. Gan G. Data Clustering: Theory, Algorithms and Applications / G. Gan, Ch. Ma, J. Wu. – Philadiphia, Pensilvania : SIAM, 2007. – 455 p.
2. Abonyi J., Feil D. Cluster Analisis for Data Mining and System Identification / J. Abonyi, D. Feil. – Basel : Birlhause, 2007. – 303 p.
3. Xu R. Clustering / R. Xu, D. C. Wunsch. – Hoboken N.J. : John Wiley & Sons, Inc., 2009. – 398 p.
4. Aggarwal C. C. Data Mining / C. C. Aggarwal. – Switzerland : Springer, 2015. – 727 p. DOI https://doi.org/ 10.1007 / 978-3-319-14142-8.
5. Fuzzy Clustering Analysis: Methods for Classification, Data Analisys and Image Recognition / [Höppner F., Klawonn F., Kruse R., Runkler T.] – Chichester : John Wiley &Sons, 1999. – 300 p.
6. Bezdek J. C. et al. Fuzzy models and algorithms for pattern recognition and image processing [Bezdek J. C. et al.]. – Springer Science & Business Media, 1999. – Vol. 4.
7. Hinneburg A. An efficient approach to clustering in large multimedia databases with noise / A. Hinneburg, D. Klein // Proc. 4th Int. Conf. in Knowledge Discovering and Data Mining – KDD98, N.Y.: AAAI Press, Aug. 27, 1998. – Hinneburg, 1998. – P. 58–65.
8. Hinneburg, A., DENCLUE 2.0: Fast Clustering Based on Kernel Density Estimation / A. Hinneburg, HH. Gabriel In: R. Berthold, M., Shawe-Taylor, J., Lavrač, N. (eds) // Advances in Intelligent Data Analysis VII. IDA. – 2007. – Lecture Notes in Computer Science. – Vol. 4723. – Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-74825-0_7
9. Hinneburg A. A general approach to clustering in large databases with noise / A. Hinneburg, D. A. Keim // Knowledge and Identification Systems. – 2003. – 5 (4). – P. 387–415. https://doi.org/10.1007 /s10115-003-0086-9
10. DENCLUE-IM: A new approachfor big data clustering / [H. Rehhioni, A. Idrissi, M. Abourezq, F. Zegrary] // Procedia Computer Science. – 2016. – 83. – P. 560–567.
11. Rodriguez A. Clustering by fast seach and find of density peaks / A. Rodriguez, A. Laio // Science. – 2014. – № 34. – P. 1492–1496. https://doi.org/10.1126/science.124207
12. Online Credibilistic Fuzzy Clustering Method Based on Cauchy Density Distribution Function / [A. Shafronenko, Ye. Bodyanskiy, I. Pliss, I. Klymova] // 2021 11th International Conference on Advanced Computer Information Technologies (ACIT): proceedings. – Deggendorf, Germany: IEEE, 2021. – P. 704–707. DOI: 10.1109/ ACIT52158.2021.9548572
13. Epanechnikov V. A. Nonparametric estimation of multivariate probability density / V. A. Epanechnikov // Probability theory and its Application – 1968 – 14, № 2 – P. 156–161.
14. Parzen E. On estimation of a probably density function and mode / E. Parzen // The Annals of Math Statistics. – 1962. – 33, № 3. – P. 1065–1076. http://dx.doi.org/10.1214/aoms/1177704472
15. Nadaraya E. A. On nonparametric estimates of density function and regression curves / E. A. Nadaraya // Theory of Probabilistic Application. – 1965. – № 10 – P. 186–190.
16. Watson G. S. Smoth regression analisys / G. S. Watson // The Indian Journal of Statistics. Sankhya. – 1964. – Ser. A. – 26, № 4. – P. 359–372.
17. Fukunaga K. The estimation of the gradient of a density function with application in pattern recognition / K. Fukunaga, L. D. Hostler // IEEE Trans. on Inf. Theory, Jan.,

DOI 10.15588/1607-3274-2022-3-6

1975 – IEEE. – 1975. – № 21 – P. 32–40. https://doi.org/10.1109/TIT.1975.10 55330.
18. Credibilistic clustering: the model and algorithms. / [J. Zhou, Q. Wang, C.-C. Hung, X. Yi] // International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. – 2015. – Vol. 23, №4. – P. 545–564. https://doi.org/10.1142/S0218488515500245
19. Zhou J. Credibilistic clustering algorithms via alternating cluster estimation / J. Zhou, Q. Wang, C. C. Hung // Journal of Intelligent Manufacturing. – 2017. – Vol. 28. – P. 727–738. DOI: https://doi.org/10.1007/s10845-014-1004-6.
20. Online credibilistic fuzzy clustering of data using membership functions of special type[Electronic resource]/ [Shafronenko A., Bodyanskiy Ye., Klymova I., Holovin O.] // Proceedings of The Third International Workshop on Computer Modeling and Intelligent Systems (CMIS-2020), April 27–1 May 2020. – Zaporizhzhia, 2020. – Access mode: http://ceur-ws.org/Vol-2608/paper56.pdf.

Стаття надійшла до редакції 15.05.2022.
Після доробки 23.08.2022.

УДК 004.8:004.032.26

НЕЧЕТКАЯ ДОВЕРИТЕЛЬНАЯ КЛАССИЗАЦИЯ ДАННЫХ НА ОСНОВЕ АНАЛИЗА ПЛОТНОСТИ РАСПРЕДЕЛЕНИЯ ДАННЫХ И ИХ ПИКОВ

Бодянский Е. В. – д-р техн. наук, профессор, профессор кафедры искусственного интеллекта Харьковского национального университета радиоэлектроники, Харьков, Украина.

Плисс И. П. – канд. техн. наук, ведущий научный сотрудник ПНДЛ АСУ Харьковского национального университета радиоэлектроники, Харьков, Украина.

Шафроненко А. Ю. – канд. техн. наук, доцент, доцент кафедры информатики Харьковского национального университета радиоэлектроники, Харьков, Украина.

Калиниченко О. В. – канд. техн. наук, доцент, доцент кафедры програмной инженерии Харьковского национального университета радиоэлектроники, Харьков, Украина.

АННОТАЦИЯ

Актуальность. Задача кластеризации – классификации без учителя массивов данных занимает достаточно важное место в интеллектуальном анализе данных. Для решения этой задачи на данный момент предложено множество подходов, отличающихся друг от друга априорными предположениями в исследуемых и анализируемых массивах, а так же математическим аппаратом, заключающимся в основе тех или иных методов, однако решение задач кластеризации усложняет большая размерность векторов анализируемых наблюдений, их искаженность разного типа.

Цель. Цель работы заключается во внедрении процедуры нечеткой кластеризации, объединяющей преимущества методов, основанных на анализе плотностей распределения данных и их пиков, которые характеризуются высоким быстродействием и может эффективно работать в условиях классов, которые пересекаются.

Метод. Введен метод нечеткой кластеризации массивов данных, основанный на идеях анализа плотностей распределения этих данных, их пиков и доверительного нечеткого подхода. Преимуществом предлагаемого подхода является сокращение времени решения оптимизационных задач, связанных с отысканием аттракторов функций плотностей, поскольку количество обращений в блок оптимизации определяется не объемом анализируемого массива, а количеством пиков плотностей этого же массива.

Результаты. Метод достаточно прост в численной реализации и не критичен к выбору оптимизационной процедуры. Результаты экспериментов подтверждают эффективность предлагаемого подхода в задачах кластеризации при условии пересечения кластеров и позволяют рекомендовать предложенный метод для использования на практике для решения проблем автоматической кластеризации больших объемов данных.

Выводы. Введен метод нечеткой кластеризации массивов данных, основанный на идеях анализа плотностей распределения этих данных, их пиков и доверительного нечеткого подхода. Преимуществом предлагаемого подхода является сокращение времени решения оптимизационных задач, связанных с отысканием аттракторов функций плотностей, поскольку количество обращений в блок оптимизации определяется не объемом анализируемого массива, а количеством пиков плотностей этого же массива. Метод достаточно прост в численной реализации и не критичен к выбору оптимизационной процедуры. Результаты экспериментов подтверждают эффективность предлагаемого подхода в задачах кластеризации в условиях пересекающихся кластеров.

КЛЮЧЕВЫЕ СЛОВА: нечеткая кластеризация, правдоподобная кластеризация, пики плотности распределения данных.

UDC 004.8:004.032.26

CREDIBILISTIC FUZZY CLUSTERING BASED ON ANALYSIS OF DATA DISTRIBUTION DENSITY AND THEIR PEAKS

Bodyanskiy Ye. V. – Dr. Sc., Professor at the Department of Artificial Intelligence, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

Pliss I. P. – PhD, Leading Researcher at Control Systems Research Laboratory, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

Shafronenko A. Yu. – PhD, Associate Professor Professor at the Department of Informatics, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

DOI 10.15588/1607-3274-2022-3-6

Kalynychenko O. V. – PhD, Associate Professor Professor at the Department of Software Engineering, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

ABSTRACT

Context. The task of clustering – classification without a teacher of data arrays occupies a rather important place in Data Mining. To solve this problem, many approaches have been proposed at the moment, differing from each other in a priori assumptions in the studied and analyzed arrays, in the mathematical apparatus that is the basis of certain methods. The solution of clustering problems is complicated by the large dimension of the vectors of the analyzed observations, their distortion of various types.

Objective. The purpose of the work is to introduce a fuzzy clustering procedure that combines the advantages of methods based on the analysis of data distribution densities and their peaks, which are characterized by high speed and can work effectively in conditions of classes that overlapping.

Method. The method of fuzzy clustering of data arrays, based on the ideas of analyzing the distribution densities of these data, their peaks, and a confidence fuzzy approach has been introduced. The advantage of the proposed approach is to reduce the time for solving optimization problems related to finding attractors of density functions, since the number of calls to the optimization block is determined not by the volume of the analyzed array, but by the number of density peaks of the same array.

Results. The method is quite simple in numerical implementation and is not critical to the choice of the optimization procedure. The experimental results confirm the effectiveness of the proposed approach in clustering problems under the condition of cluster intersection and allow us to recommend the proposed method for practical use in solving problems of automatic clustering of large data volumes.

Conclusions. The method is quite simple in numerical implementation and is not critical to the choice of the optimization procedure. The advantage of the proposed approach is to reduce the time for solving optimization problems related to finding attractors of density functions, since the number of calls to the optimization block is determined not by the volume of the analyzed array, but by the number of density peaks of the same array. The method is quite simple in numerical implementation and is not critical to the choice of the optimization procedure. The experimental results confirm the effectiveness of the proposed approach in clustering problems under conditions of overlapping clusters.

KEYWORDS: fuzzy clustering, credibilistic clustering, density peak of dataset.

REFERENCES

1. Gan G., Ma Ch., Wu J. Data Clustering: Theory, Algorithms and Applications. Philadiphia, Pensilvania, SIAM, 2007, 455 p.
2. Abonyi J., Feil D. Cluster Analisis for Data Mining and System Identification. Basel, Birlhause, 2007, 303 p.
3. Xu R., Wunsch D. C. Clustering. Hoboken N.J., John Wiley & Sons, Inc., 2009, 398 p.
4. Aggarwal C. C. Data Mining. Switzerland, Springer, 2015, 727 p. DOI https://doi.org/ 10.1007 / 978-3-319-14142-8.
5. Höppner F., Klawonn F., Kruse R., Runkler T. Fuzzy Clustering Analysis: Methods for Classification, Data Analisys and Image Recognition. Chichester, John Wiley &Sons, 1999, 300 p.
6. Bezdek J. C. et al. Fuzzy models and algorithms for pattern recognition and image processing. Springer Science & Business Media, 1999, Vol. 4.
7. Hinneburg A., Klein D. An efficient approach to clustering in large multimedia databases with noise, *Proc. 4th Int. Conf. in Knowledge Discovering and Data Mining, KDD98, N.Y.: AAAI Press, Aug. 27, 1998*. Hinneburg, 1998, pp. 58–65.
8. Hinneburg A., Gabriel HH. In: R. Berthold, M., Shawe-Taylor, J., Lavrač, N. (eds) DENCLUE 2.0: Fast Clustering Based on Kernel Density Estimation. Advances in Intelligent Data Analysis VII. IDA 2007. Lecture Notes in Computer Science, Vol. 4723. Springer. Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-74825-0_7
9. Hinneburg A., Keim D. A. A general approach to clustering in large databases with noise, *Knowledge and Identification Systems,* 2003, 5 (4), pp. 387–415. https://doi.org/10.1007 /s10115-003-0086-9
10. Rehhioni H., Idrissi A., Abourezq M., Zegrary F. DENCLUE-IM: A new approachfor big data clustering, *Procedia Computer Science*, 2016, 83, pp. 560–567.
11. Rodriguez A., Laio A. Clustering by fast seach and find of density peaks, *Science*, 2014, No. 34, pp. 1492–1496. https://doi.org/10.1126/science.124207
12. Shafronenko A., Bodyanskiy Ye., Pliss I., Klymova I. Online Credibilistic Fuzzy Clustering Method Based on Cauchy Density Distribution Function, *2021 11th International Conference on Advanced Computer Information Technologies (ACIT): proceedings.* Deggendorf, Germany, IEEE, 2021, pp. 704–707. DOI: 10.1109/ ACIT52158.2021.9548572
13. Epanechnikov V. A. Nonparametric estimation of multivariate probability density, *Probability theory and its Application,* 1968, 14, No. 2, pp. 156–161.
14. Parzen E. On estimation of a probably density function and mode, *The Annals of Math Statistics,* 1962, 33, No. 3, pp. 1065–1076. http://dx.doi.org/10.1214/aoms/1177704472
15. Nadaraya E. A. On nonparametric estimates of density function and regression curves, *Theory of Probabilistic Application,* 1965, No. 10, pp. 186–190.
16. Watson G. S. Smoth regression analisys, *The Indian Journal of Statistics*. Sankhya, 1964, Ser. A, 26, No. 4, pp. 359–372.
17. Fukunaga K., Hostler L. D.// The estimation of the gradient of a density function with application in pattern recognition, *IEEE Trans. on Inf. Theory,* Jan., 1975, IEEE, 1975, No. 21 pp. 32–40. https://doi.org/ 10.1109/TIT.1975.10 55330.
18. Zhou J., Wang Q., Hung C.-C., Yi X. Credibilistic clustering: the model and algorithms, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* 2015, Vol. 23, No. 4, pp. 545–564. https://doi.org/ 10.1142/S0218488515500245
19. Zhou J., Wang Q., Hung C. C. Credibilistic clustering algorithms via alternating cluster estimation, *Journal of Intelligent Manufacturing,* 2017, Vol. 28, pp. 727–738. DOI: https://doi.org/10.1007/s10845-014-1004-6.
20. Shafronenko A., Bodyanskiy Ye., Klymova I., Holovin O.] Online credibilistic fuzzy clustering of data using membership functions of special type[Electronic resource, *Proceedings of The Third International Workshop on Computer Modeling and Intelligent Systems (CMIS-2020), April 27-1 May 2020*. Zaporizhzhia, 2020. Access mode: http://ceur-ws.org/Vol-2608/paper56.pdf.

DOI 10.15588/1607-3274-2022-3-6