М. М. Биков

РОЗРОБКА ЕФЕКТИВНОГО МЕТОДУ ІДЕНТИФІКАЦІЇ ОБ'ЄКТІВ В СИСТЕМАХ УПРАВЛІННЯ

Вінницький національний технічний університет

Анотація. *В роботі представлено теоретичне обґрунтування і програмне забезпечення ефективного методу ідентифікації даних в інтелектуальних системах управління, який базується на використанні етапів кластеризації і класифікації об'єктів, представлених їх параметричним описом. Автором запропоновано метод кластеризації даних з фіксацією граничних точок кластера за принципом знаходження їх поверхневого натягу. Застосування даного методу кластерного аналізу дозволяє підвищити ефективність ідентифікації об'єктів в системах управління з використанням класифікатора типу SVM. Класифікатор типу SVM (Support Vector Machine – метод опорних векторів) на разі є найбільш досконалим і за рахунок використання узагальнених вирішувальних функцій здатний оперувати як з лінійними, так і з нелінійно розподіленими зображеннями об'єктів в просторі параметрів. На сьогодні є відомою значна кількість методів і алгоритмів кластерного аналізу. Їх можна поділити на класичні, що ґрунтуються на загальному алгоритмі Фу К. С., і новітні, засновані на врахуванні природи і структури даних та мети їх використання. За приклади таких методів можна навести такі: k- внутрішньо групових середніх (k-means), на основі щільності (DBSCAN), середнього зсуву (Mean-Shift), з використанням гаусових моделей сумішей (GMM), методи ієрархічної агломеративної кластеризації (HAC). Спільним їхнім недоліком є те, що вони формують кластери точок у просторі параметрів, проте не фіксують їхні граничні точки. Остання характеристика (наявність маркованих граничних точок кластера) є дуже корисною при побудові класифікаторів, що реалізують задачу ідентифікації об'єктів. В даній роботі автор запропонував і теоретично обґрунтував метод кластеризації з визначенням поверхневих точок по аналогії з фізичним принципом поверхневого натягу рідини.*

В роботі розроблено алгоритм для реалізації даного методу, а також програмне забезпечення на мові Python у вигляді додатку ClusterBorderFinder. На розроблену програму отримано авторське свідоцтво.

Ключові слова: системи управління, ідентифікація, кластерний аналіз, класифікація, принцип поверхневого натягу рідини, градієнтні методи навчання.

Вступ

Ідентифікація об'єктів за їх параметричним описом є одним із основних етапів прийняття управлінських рішень. Розв'язання даної задачі в сучасних комп'ютерних системах управління переважним чином реалізується за рахунок впровадження методів машинного навчання (Machine Learning - ML), таких, як кластеризація і класифікація даних [1,2,3,4]. Ці методи широко використовуються в вирішенні завдань прийняття маркетингових рішень і розпізнавання образів, що посідають значний обсяг в бізнес та технологічних процесах сучасних автоматизованих виробництв.

На разі найбільш поширеними методами кластерного аналізу даних є [5,6,7,8]: DBSCAN - кластеризація на основі щільності; *K*-means - *k*-внутрішньо групових середніх; Mean-Shift - метод середнього зсуву; GMM - кластеризація з застосуванням гаусових моделей сумішей; HAC - методи ієрархічної агломеративної кластеризації. Найбільш вживаним і досконалим методом класифікації є SVM – машина (метод) опорних векторів [5]. Його перевага полягає в універсальності, тобто здатності оперувати як з лінійно, так і нелінійно розподіленими зображеннями об'єктів в описовому просторі параметрів за рахунок використання вирішувальних функцій узагальненого типу.

Як показав аналіз практичного використання вказаних методів кластеризації і класифікації даних для ідентифікації об'єктів за їхнім параметричним описом, їхнім недоліком є відсутність між ними взаємозв'язку, що полягає у можливості використання класифікатором SVM поверхневих точок кластерів в якості опорних для підвищення ефективності процедури класифікації [9]. Тому дослідження, присвячені розробці методів кластеризації даних в просторі параметрів, які б

дозволяли виконувати фіксацію поверхневих точок кластерів для подальшого їх використання SVM класифікатором, є на разі актуальними.

Метою даної роботи є вирішення актуальної проблеми підвищення ефективності процесу ідентифікації об'єктів в системах управління за рахунок розробки покращених методів кластеризації і класифікації.

Результати дослідження

В роботі [9] автори запропонували метод пошуку опорних точок для класифікатора типу SVM шляхом визначення поверхневих точок кластерів за принципом, аналогічним до фізичного принципу поверхневого натягу молекул в рідині. Під час реалізації запропонованого підходу були прийняті такі припущення:

а) модулі сил взаємодії між точками простору ознак обернено пропорціональні відстані між ними;

б) на деяку вибрану точку діють сили притягання тільки найближчих до неї точок.

В узагальненому вигляді математичну постановку задачі розробки методу кластеризації з фіксацією поверхневих точок кластера сформулюємо таким чином:

Задано: множину X об'єктів і множину номерів кластерів Y, в які потрібно згрупувати об'єкти з множини. Також задана вибірка зображень об'єктів у n –вимірному просторі параметрів у вигляді підмножини точок $\{\vec{x}_1,..\vec{x}_k,..,\vec{x}_M\} \subset X$, де M – кількість точок у вибірці. Для вибраної за визначеною метрикою D міри відстані між точками задана функція натягу $\Delta(\vec{x}_k, \vec{z}_k)$ між деякою точкою $\vec{x}_k$ та набором з p найближчих до неї точок, де $\vec{z}_k$ - центр заданої плинної групи точок.

Необхідно: за заданими вхідними даними розподілити вибірку точок на непересічні кластери таким чином, щоб в кожному кластері містилися тільки близькі за метрикою D точки, а поверхневі (граничні) точки кластера фіксувалися в масиві S_{Y_i}, де Y_i - кластер з номером і.

Часто множина Y наперед невідома, і постає задача пошуку оптимальної кількості кластерів. В даному дослідженні передбачається, що кількість кластерів була попередньо визначена, наприклад, максимінною процедурою кластеризації [1,5], і алгоритм кластеризації з фіксацією поверхневих точок розробляється для множини зображень об'єктів, що відносяться до одного кластера.

В загальному випадку побудова такого алгоритму повинна передбачати покрокову реалізацію процедури градієнтного спуску, здатної під час знаходження множини поверхневих точок кластера мінімізувати величину загального стресу (поверхневого натягу), що складається з поверхневих натягів окремих точок,

Оскільки кожна точка в просторі ознак описується вектором значень її n координат, то в початковому представленні функцію (силу) поверхневого натягу можна описати як функцію багатьох змінних, процедура мінімізації якої за градієнтним методом приведе до складного математичного опису. Для приведення цієї процедури до більш простого математичного опису вибрано такий спосіб представлення функції поверхневого натягу, який привів її до скалярного виду. Таким способом є представлення сили натягу, яка діє на вибрану точку, у вигляді величини абсолютного зміщення координати деякої точки під впливом рівнодійної сил притягання найближчих навколишніх точок. Це зміщення може бути визначене як нормована відстань між координатою дано] точки і центром координат найближчих навколишніх точок. Тому приймемо наступне значення виразу для сили натягу, яка діє на вибрану точку кластера:

$$\Delta_k(\vec{x}_k, \vec{z}_k) = \frac{\sqrt{\sum_{i=1}^{n}(x_{ki} - z_{ki})^2}}{d_k}, \qquad (1)$$

де d_k - середня відстань для вибраної групи точок, найближчих до точки $\vec{x}_k$;

$\vec{z}_k$ - центр координат (середнє значення координат) цих точок.

З такого визначення є зрозумілим, що величина сили натягу буде залежати від кількості точок, які створюють силу натягу на вибрану точку. Мала кількість точок буде приводити до великої чутливості сили натягу до зміни координат навколишніх точок, а велика кількість точок – до згладжування цього впливу, що і в першому і в другому випадку приведе до помилок у визначенні поверхневих (граничних) точок кластера. Також на точність результату фіксації

поверхневих точок кластера буде впливати величина порогу сили натягу, при якому точка повинна вважатися поверхневою. З наведених міркувань випливає, що величина сили натягу, яка діє в кластері на поверхневу точку, є фунцкією двох змінних – кількості навколишніх точок p і визначеного порогу сили натягу σ. Математичний опис процедури оптимізації цих величин тягне за собою значні математичні труднощі, для уникнення яких в роботі для пошуку оптимального значення цих величин вибрано метод навчання на прецедентах (навчання з учителем). Для його реалізації генерувались в двовимірному просторі параметрів випадкові вибірки зображень об'єктів, визначались їх поверхневі точки, і до них застосовувався нижче описаний алгоритм кластеризації при різних значеннях стресу σ і кількості точок p. Остаточні значення σ і p отримувались усередненням значень, отриманих у всіх циклах навчання.

В словесному виді розроблений алгоритм кластеризації навчальної вибірки даних у вигляді масиву даних $MAS[M]=\{\vec{x}_1,\vec{x}_2,...,\vec{x}_k,...,\vec{x}_M\}$ N-вимірного простору ознак можна сформулювати таким чином:

1) обчислюємо відстані $d(\vec{x}_i,\vec{x}_j)$ між всіма парами точок (для визначеності обрана Евклідова метрика простору параметрів) і формуємо з них матрицю відстаней розміром [M x M].

2) переглядаємо послідовно всі точки масиву $MAS[M]$ шляхом зміни індексу k, і для кожної з них виконуємо наступні дії:

а) вибираємо p найближчих точок до чергової k-ої – передбачена можливість зміни кількості найближчих точок на вході алгоритму, досліджувалися випадки при 3, 4, 5 і 6 найближчих точках;

в) визначаємо середню відстань для вибраної групи точок з центром у точці з номером k:

$$d_k=\frac{\sum_{i=1}^{p} d(\vec{x}_k,\vec{x}_i)}{p}, \qquad (2)$$

г) визначємо координати (z_{k1}, z_{k2}, …, z_{kn},…, z_{kN}) центра $\vec{z}_k$ вибраної сукупності p точок як їх середнє арифметичне;

д) за формулою (1) знаходим зміщення Δ_k плинної точки $\vec{x}_k$ відносно центра $\vec{z}_k$:

е) перевіряємо виконання умови $\Delta_k>\sigma$ і заносимо точку $\vec{x}_k$ до масиву поверхневих точок кластера, якщо умова виконується.

Процедура повторюється ітеративно для різних значень порогу σ поверхневого натягу $\Delta(x_k,z_k)$ для визначення максимального натягу поверхневих точок, який є умовою припинення пошуку поверхневих точок кластера.

Алгоритм реалізовано мовою Python у вигляді програмного додатку ClusterBorderFinder у складі трьох модулів: Point, ClusterBoundPointFinder і MainWindow (рис.1). Ці програмні модулі використовуються на етапі навчання системи ідентифікації для пошуку опорних точок для класифікатора типу SVM. На розроблену програму отримано авторське право на твір №128906. На етапі ідентифікації об'єктів застосовується модуль класифікатора SVM, покращеного за рахунок використання в якості опорних точок знайдених на етапі кластеризації поверхневих точок сусідніх кластерів, найближчих між собою.

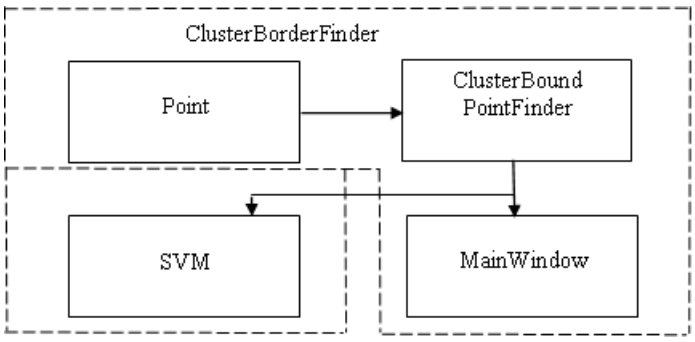

Рис.1. Склад програмного пролукту ClusterBorderFinder і SVM

Результати роботи програми проілюстровані у вікні графічного інтерфейсу розробленого програмного додатку на рис. 2. Наведені дані відповідають знайденим експериментально оптимальним значенням кількості точок в ядрі p=5 і порогу поверхневого натягу σ=0,35÷0,45.

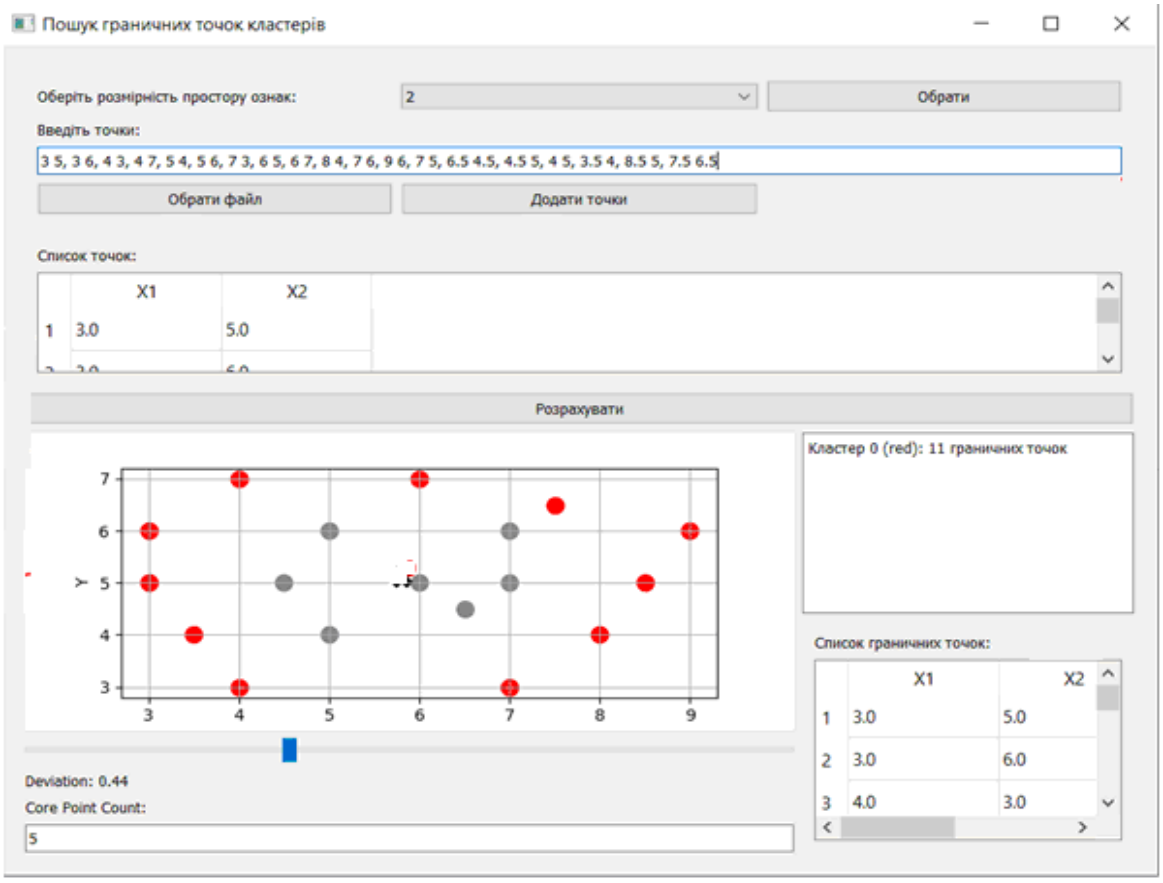

Рис.2. Результати кластерного аналізу даних у двовимірному просторі при значеннях σ =Deviation=0,44 і p=Core Point Count=5

У вікні графічного інтерфейсу сірими точками показані внутрішні точки кластера, а червоними точками показані поверхневі точки кластера. Проведені за допомогою даного програмного додатку на різних навчальних вибірках дослідження показали, що розроблений алгоритм кластеризації правильно фіксує поверхневі точки кластера при заданні порогу натягу σ в межах 0,35÷0,45 і кількості точок в ядрі натягу p=Closest point=5.

Дослідження, проведені на класифікації даних еталонного файла “iris.dat” показали, що модифікований за рахунок використання опорних точок кластерів класифікатор SVM, дає найбільшу точність класифікації за умови використання радіально-базисної функції ядра RBF і параметра Gamma=0.44 [10]. При цьому отримана точність є порівняною з точністю стандартного SVM класифікатора, але використання знайдених на етапі кластеризації опорних точок дозволило підвищити швидкість навчання в 3 рази. Даний результат підтверджує ефективність застосування запропонованого підходу до ідентифікації об’єктів в системах управління за рахунок розроблених покращених алгоритмів кластеризації та класифікації даних.

Висновки

Поставлена в роботі задача підвищення ефективності процесу ідентифікації об’єктів в системах управління за рахунок розробки покращених методів кластеризації і класифікації вирішена за рахунок використання запропонованого автором нового методу кластеризації даних за принципом, аналогічним фізичному принципу поверхневого натягу молекул в рідині. Наведене математичне обґрунтування методу дозволило розробити алгоритм і програмне забезпечення кластерного аналізу даних про об’єкти ідентифікації з фіксацією поверхневих точок кластерів з подальшим їх використанням в якості опорних точок для класифікатора типу SVM. В результаті проведених машинних експериментів були визначені оптимальні параметри для алгоритму

кластеризації. Даний підхід дозволив підвищити швидкість роботи класифікатора при збереженні точності його роботи, і тим самим підвищити ефективність процесу ідентифікації об'єктів в системах управління. Для реалізації запропонованого підходу розроблено програмне забезпечення у вигляді додатку ClusterBorderFinder, яке захищено авторським правом на твір.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] Биков М. М. Основи інтелектуальних технологій. Частина 2. Технології машинного навчання : електронний навчальний посібник комбінованого (локального та мережного) використання [Електронний ресурс] / Биков М. М., Ковтун В. В., Грищук Т. В., Вінниця, ВНТУ, 2024, 153 с.

[2] M. Abadi, P. Barham, J. Chen, et al., "TensorFlow: A system for large-scale machine learning", Communications of the ACM, vol. 62, no. 10, pp. 22-35, Oct. 2019. doi: 10.1145/3360324.

[3] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 8, pp. 1798-1828, Aug. 2020. doi: 10.1109/TPAMI.2019.2903152.

[4] І. Гаврилюк і М. Дубчак, "Алгоритми і методи машинного навчання для класифікації даних у сучасних інформаційних системах", Вісник Київського національного університету імені Тараса Шевченка, вип. 12, с. 24-31, червень 2022.

[5] Биков, М. М. Основи інтелектуальних технологій. Частина 1. Технології розпізнавання : електронний навчальний посібник комбінованого (локального та мережного) використання [Електронний ресурс] / Биков М. М., Ковтун В. В., Гаврилюк В. О., Вінниця, ВНТУ, 2023, 229 с.

[6] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD)*, San Francisco, CA, USA, 2020, pp. 226-231.

[7] С. Іванченко, О. Ткаченко, і Ю. Петренко, "Методи кластеризації даних у великих масивах інформації", Науковий вісник Національного університету біоресурсів і природокористування України, вип. 23, с. 98-104, квітень 2021.

[8] О. Мельник і В. Карпенко, "Моделі класифікації даних у задачах аналітики великих даних," Інформаційні технології та комп'ютерна інженерія", вип. 32, с. 17-23, жовтень 2022.

[9] Биков М.М., Волоський Б.О. Розробка ефективного класифікатора даних в інтелектуальних системах управління [Електронний ресурс] / М.М. Биков, Б.О. Волоський // Матеіали XLIX науково-технічної конференції підрозділів ВНТУ, Вінниця, 27-28 квітня 2020 р.,Електр. текст. дані., 2020. Режим доступу: https://conferences.vntu.edu.ua/index.php/all-fksa/all-fksa-2020/paper/view/9730 ,

[10] Биков М.М., Бушин Д.О. Розробка ефективного класифікатора для автоматизованої системи допуску персоналу на виробництві. Матеріали Міжнародної наук.-техн. Конференції "Молодь в науці: дослідження, проблеми, перспективи (МН-2024)", [Електронний ресурс]. Режим доступу: https://conferences.vntu.edu.ua/index.php/mn/mn2024/paper/view/21460атеріали

Биков Микола Максимович — канд.. техн. наук, професор кафедри комп'ютерних систем управління, e-mail: nkbykov@vntu/edu.ua; mbykov123@ukr.net

Вінницький національний технічний університет, Вінниця.

M. M. Bykov

Development of an efficient method of objects identifying in control systems

Vinnytsia National Technical University

Annotation. *The paper presents the theoretical justification and software for an effective method of data identification in intelligent control systems, which is based on the use of the stages of clustering and classification of objects represented by their parametric description. The author proposed a method of data clustering with fixation of cluster boundary points based on the principle of finding their surface tension. The application of this method of cluster analysis allows to increase the efficiency of identification of objects in control systems using the SVM type classifier. The SVM (Support Vector Machine) type classifier is currently the most advanced and, due to the use of generalized decision functions, is able to operate with both linear and non-linearly distributed images of objects in the parameter space. Today, a significant number of cluster analysis methods and algorithms are known. They can be divided into classical, based on the general algorithm of Fu K.S., and the latest, based on taking into account the nature and structure of data and the purpose of their use. Examples of such methods include: k-means, density-based (DBSCAN), Mean-Shift, using Gaussian mixture models (GMM), methods of hierarchical agglomerative clustering (HAC) . Their common drawback is that they form clusters of points in the parameter space, but do not fix their boundary points. The last characteristic (the presence of marked cluster boundary points) is very useful when building classifiers that implement the task of object identification. In this paper, the author proposed and theoretically substantiated the method of clustering with the definition of surface points by analogy with the physical principle of surface tension of a liquid.*

The work developed an algorithm for the implementation of this method, as well as software in the Python language in the form of the ClusterBorderFinder application. The author's certificate was obtained for the developed program.

Key words: control systems, identification, cluster analysis, classification, principle of surface tension of liquid, gradient learning methods.

Bykov Mykola Maksymovych - PhD, professor of the Department of Computer Control Systems,, e-mail: nkbykov@vntu/edu.ua; mbykov123@ukr.net