УДК 681.142.2 (075.5)

**М. М. Биков**
**А.С. Задачін**

# РОЗРОБКА ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ КЛАСТЕРИЗАЦІЇ ДАНИХ В ІНТЕЛЕКТУАЛЬНИХ СИСТЕМАХ ПРИЙНЯТТЯ РІШЕНЬ

Вінницький національний технічний університет;

**Анотація**

В роботі розроблено алгоритм і програмне забезпечення кластерного аналізу даних в інтелектуальних системах управління, орієнтованого на його подальше застосування для оптимізації процедури навчання класифікатора SVM. Алгоритм реалізує запропонований авторами метод пошуку поверхневих точок кластера за принципом знаходження їх поверхневого натягу.

**Ключові слова:** кластеризація, машина опорних векторів (SVM), принцип поверхневого натягу, критерій якості пошуку поверхневих точок кластера, оптимізаційна процедура.

**Abstract**

The algorithm and software of cluster data analysis in intelligent control systems, focused on its further application for optimization of the SVM classifier training procedure, are developed in the work. The algorithm implements the method proposed by the authors to find the surface points of the cluster on the principle of finding their surface tension.

**Keywords:** clustering, support vector machine (SVM), surface tension principle, cluster surface point search quality criterion, optimization procedure.

## Вступ

Задача кластерного аналізу даних в інтелектуальних комп'ютерних системах системах управління пов'язана з необхідністю групування об'єктів різного типу в групи похожих між собою за параметрами, що представляють собою інтерес для дослідника [1]. На сьогодні існує велика кількість методів і алгоритмів кластеризації даних (більше 100). До них відносяться як класичні методи, що ґрунтуються на узагальненому алгоритмі К. С. Фу [2], так і новітні, що ґрунтуються на урахуванні структури і природи даних і мети їх використання [3,4]. Всі ці методи і алгоритми кластеризації мають свої недоліки і переваги, зумовлені орієнтацією на вирішення вибраного класу задач. Загальним їхнім недоліком є те, що вони формують кластери точок у просторі ознак, але не фіксують їхні границі. Остання характеристика (наявність маркованих граничних точок кластера) є дуже корисною при побудові класифікаторів. Автори запропонували в роботі [5] метод пошуку опорних точок для класифікатора типу SVM шляхом визначення поверхневих точок кластера по аналогії з фізичним принципом поверхневого натягу рідини. Аналіз результатів машинного експерименту, проведений авторами в роботі [5] показав, що запропонований підхід можна використовувати не тільки для оптимізації процедури побудови класифікатора, а й для кластеризації даних шляхом фіксації поверхневих точок кластерів, які описують їх границі.

В даній доповіді наводиться модифікація запропонованих у вказаній роботі алгоритму і програми з метою їх використання для побудови кластерів шляхом визначення їх поверхнвих точок за методом поверхневого натягу.

## Результати дослідження

В загальному випадку задачу кластерного аналізу шляхом фіксації поверхневих (граничних) точок кластера можна сформулювати в такий спосіб:

Відома множина об'єктів $X$ і множина номерів (міток) кластерів $Y$. Також задана вибірка зображень об'єктів у вигляді підмножини точок $X^m = \{x_1,\ldots,x_m\} \subset X$ у просторі ознак $\Re^n$, де $m$ – кількість точок у вибірці, $n$ – розмірність простору ознак. Крім того, вибрано міру відстані між

точками у вигляді метрики $D(x_1, x_2)$ і задана функці натягу (стрес) між деякою точкою $x_k$ і групою найближчих до неї точок $\Delta(x_k, z_k)$, де $z_k$ - центр вибраної поточної групи точок. По заданій інформації потрібно розподілити задану вибірку точок на непересічні підмножини (кластери) таким чином, щоб кожен кластер містив тільки близькі за метрикою *D* точки, крім того точки, які лежать на поверхні кластера (є найбільш віддаленими від його центру), повинні фіксуватися в масиві $S_{Y_i}$, де $Y_i$ - кластер з номером *i*.

Алгоритм класифікації будується у вигляді функції $f: X \to Y$, яка будь-якому об'єкту $x \in X$ ставить у відповідність номер кластера $y \in Y$, причому для кожного кластера фіксується підмножина поверхневих точок $S_{Y_i} \in X$. Здебільшого множина $Y$ наперед невідома і ставиться задача пошуку оптимального числа кластерів.

В загальному випадку побудова такого алгоритму вимагає покрокової реалізації процедури градієнтного спуску знаходження для кожного кластеру множини поверхневих точок, оптимальної з точки зору критерія максимізації поверхневого натягу (стресу). Розгляд задачі в такій узагальненій постановці виходить за межі даної доповіді. В даній роботі ми розглянемо алгоритм і програму знаходження поверхневих точок тільки для одного із знайдених кластерів, що не виключає можливості розширити отримані результати на випадок розв'язання задачі кластеризації для декількох кластерів.

Вербальний опис розробленого алгоритму для вибраної вибірки з 12 точок і двовимірного простору ознак має такий вигляд.

1 )Визначаємо масив відстаней між різними парами точок:

$$D(\vec{X}_i, \vec{X}_j) = \sqrt{\sum_{k=1}^{2} (x_{ik} - x_{jk})^2},$$

кількість різних пар відстаней для вибраного прикладу дорівнює числу сполучень $C_{12}^2 = \frac{12 \cdot 11}{2} = 66$.

2) Змінюючи індекс масиву точок *k* від 1 до 12, переглядаємо послідовно всі точки масиву $M_X(12)$ і для кожної з них виконуємо наступні дії:

а) Вибираємо 4 найближчих точки до чергової *k*-ої (використовуємо масив відстаней) – в алгоритмі передбачена можливість зміни кількості найближчих точок, тобто можливості введення кількості точок на вході алгоритму, досліджувалися випадки при 3, 4, 5 і 6 найближчих точках, що відповідало разом з взятою *k*-ою точкою 4, 5, 6 і 7 точкам в ядрі;

б) Визначаємо середню відстань для вибраної групи точок:

$$d_k = \frac{\sum_{p=1}^{5} D(\vec{X}_k, \vec{X}_p)}{5},$$

де *p* – індекси чотирьох найближчих точок до *k*-ої .

в) Визначаємо координати *t* і *l* центра $\vec{Z}_k$ даної сукупності точок:

$$t = \frac{\sum_{n=1}^{5} x_n}{5}, \qquad l = \frac{\sum_{m=1}^{5} x_m}{5}$$

де $x_n$ і $x_m$ - перша і друга координати сукупності точок.

За умови зміни розмірності масиву кількість координат центра повинна відповідно збільшуватися, тобто у даному випадку для двомірного масиву дві координати, для трьохмірного – три координати, для *q*-вимірного – *q* координат.

г) Визначаємо відносне зміщення точки $\vec{X}_k$ відносно центру $\vec{Z}_k$:

$$\Delta_k = d(\vec{X}_k, \vec{Z}_k) = \frac{\sqrt{\sum_{n=1}^{2} (x_n - z_n)^2}}{d_k},$$

де $n$ – номери координат точок $\vec{X}_k$ і $\vec{Z}_k$, для заданоо двовимірного масиву $n$=1,2.

д) Перевіряємо умову:

якщо $\Delta_k > \sigma$, то точка $\vec{X}_k$ є поверхневою в кластері.

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

Результати роботи програми, що реалізує даний алгоритм на заданій вибірці точок, представлені на рис.1 і рис.2. Програма написана на мові Python і може бути надана авторами за запитом.

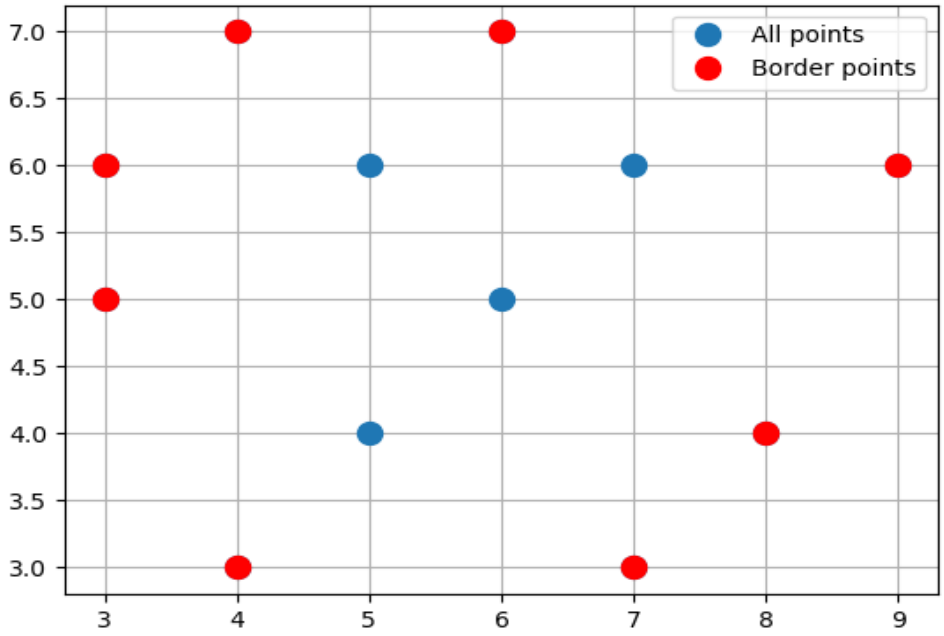


Рис. 1 – Результат пошуку поверхневих точок в заданому кластері при при заданному порозі стресу σ=0.35

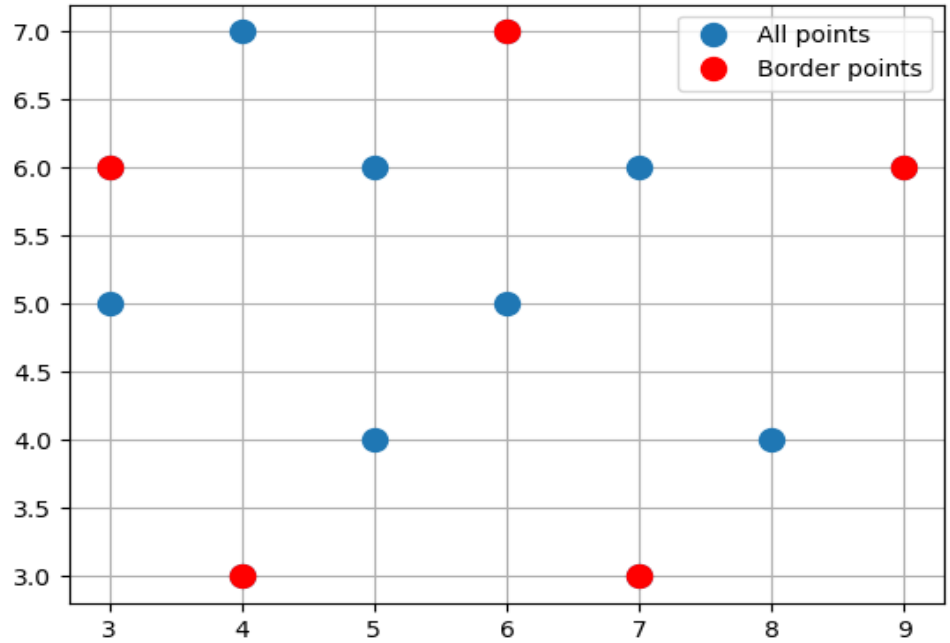


Рис. 2 – Результат пошуку поверхневих точок в заданому кластері при при заданному порозі стресу σ=0.6

Як показали результати досліджень роботи алгоритму, оптимальним для пошуку поверхневих точок кластера є значення максимального стресу $0.35 \le \sigma \le 0.45$.

## Висновки

В роботі розроблено алгоритм і програму кластеризації даних з пошуком границь кластерів у вигляді його поверхневих точок, принцип роботи яких ґрунтується на фізичному принципі поверхневого натягу рідини. В результаті машинного експерименту знайдено оптимальне значення порогу поверхневого натягу для визначення умови припинення процедури пошуку поверхневих точок. Перевагою запропонованого підходу до кластерного аналізу є те, що він дозволяє визначити граничні точки кластерів, що спрощує в подальшому процедуру побудови класифікаторів даних.

## Список використаної літератури


1. Биков М.М. Лекція №2 з дисципліни “Новітні технології машинного навчання та штучного інтелекту”. – [Електронний ресурс]. Режим доступу: https://iq.vntu.edu.ua/fm/fdb/632Лекція№2_НТМНШІ.pdf

2. Looney C.G. Pattern Recognition Using Neural Networks / C. G. Looney –New York: Oxford University Press, 2004. – 449 p.

3. Кластерний аналіз. – [Електронний ресурс]. Режим доступу: https://uk.wikipedia.org/wiki/Кластерний аналіз .

4. Маннинг К.Д., Рагхаван П., Шютце Х. Введение в информационный поиск.— М.: Изд-во “Вильямс”, 2011. – 528 с.

5. 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 квітня2020 р. – Електр. текст. дані. – 2020. – Режим доступу: https://conferences.vntu.edu.ua/index.php/all-fksa/all-fksa-2020/paper/view/9730 .



***Андрій Сергійович Задачін —*** студент групи 2 АКІТ-17б, факультет комп’ютерних систем та автоматики, Вінницький національний технічний університет, м.Вінниця, e-mail: market11102015@gmail.com

***Микола Максимович Биков*** — професор кафедри комп’ютерних систем управління, Вінницький національний технічний університет, м. Вінниця, e-mail: nkbykov@vntu.edu.ua .

***Andriy S. Zadachin*** — student of Computer System and Automation Department, 2 AKIT-17b group, Vinnytsia National Technical University, Vinnytsia, e-mail: bogdan.volosky@gmail.com.

***Mykola M. Bykov*** — professor of Computer Control System Department, Vinnytsia National Technical University, Vinnytsia, e-mail: nkbykov@vntu.edu.ua.