УДК 519.237.8

**М. А. Іванчук,** асистент

Буковинський державний медичний університет, м. Чернівці

# КЛАСИФІКАЦІЯ МНОЖИН МЕТОДОМ ЛІНІЙНОГО ВІДОКРЕМЛЕННЯ ЇХ ОПУКЛИХ ОБОЛОНОК

У статті представлений метод лінійного відокремлення опуклих оболонок (ЛВОО) для класифікації двох множин в евклідовому просторі $R^n$. Наводяться приклади для порівняння результатів класифікації методом ЛВОО, а також дискримінантним аналізом, наївним байєсівським класифікатором та методом опорних векторів.

**Ключові слова:** *лінійна класифікація, опуклі оболонки, відокремлююча гіперплощина.*

**Вступ.** Розглянемо наступну задачу класифікації. Нехай задано дві групи об'єктів $A$ та $B$. Множина $A$ складається з $m_A$ об'єктів, множина $B$ — з $m_B$ об'єктів. Для кожного з об'єктів цих множин відомо $n$ кількісних ознак, впорядкованих за інформативністю, наприклад, за інформативною мірою Кульбака [1, с. 93–120]. Задача полягає в знаходженні такого алгоритму класифікації, який би на заданому рівні значущості вірно розподіляв об'єкти за відповідними групами, використовуючи при цьому найменшу кількість ознак. Прикладом подібної задачі на практиці є задача медичного прогнозування: існує дві групи хворих ($A$ — хворі з ускладненнями та $B$ — хворі без ускладнень). Для кожного з хворих відомо $n$ ознак — факторів, що впливають на розвиток ускладнень. Необхідно, використовуючи найменшу кількість ознак, побудувати експертну систему, що з заданим рівнем значущості прогнозуватиме у хворих виникнення ускладнень.

**1. Постановка задачі.** Розглянемо математичну модель даної задачі. Будемо представляти об'єкти з множин $A$ та $B$ як точки у евклідовому просторі $R^n$. Кожна координата відповідатиме за одну з $n$ ознак, що описують дані об'єкти. Тоді множини $A$ та $B$ можна записати як дві множини точок $A = \left\{ a_i = (a_i^1, a_i^2, ..., a_i^n),\ i = \overline{1, m_A} \right\}$ та $B = \left\{ b_i = (b_i^1, b_i^2, ..., b_i^n),\ i = \overline{1, m_B} \right\}$. Виділимо з них випадковим чином множину точок контрольної групи $Z = \left\{ z_i = (z_i^1, z_i^2, ..., z_i^n), i = \overline{1, m_Z} \right\}$ для перевірки якості класифікації. У випадку малих значень $m_A$ та $m_B$ можна використати метод ковзаючого контролю [2, с. 266–268].

Позначимо через $\alpha$ наперед заданий допустимий рівень помилок при розподіленні елементів множин $A$ та $B$ за відповідними класами, $\theta$ — наперед заданий допустимий рівень помилок при перевірці класифікаційної моделі на контрольній групі $Z$.

Оскільки задача полягає в знаходженні класифікатора, що використовує найменшу кількість ознак, будемо проводити класифікацію множин в просторах розмірностей $k = \overline{1,n}$. Серед класифікаторів, що задовольняють наперед задані допустимі рівні помилок $\alpha$ та $\theta$, вибиратимемо той, що відповідає простору меншої розмірності.

**Означення.** Гіперплощину

$$L_{i^k} = \left\{ x \in R^k : \langle p, x \rangle = \gamma \right\},\ p \neq 0,\ k = \overline{1,n}$$

будемо називати *відокремлюючою гіперплощиною* для множин $A$ та $B$ в просторі $R^k$, якщо не менше ніж $(1-\alpha)(m_A + m_B)$ точок множин $A$ та $B$ можна помістити в різні півпростори $L^k_{i\,A} = \left\{ x \in R^k : \langle p, x \rangle > \gamma \right\}$ та $L^k_{i\,B} = \left\{ x \in R^k : \langle p, x \rangle < \gamma \right\}$.

Позначимо $Z^k_A$ — точки контрольної множини, що відносяться до групи $A^k$, відповідно $Z^k_B$ — точки контрольної множини, що відносяться до групи $B^k$, $Z^k = Z^k_{A} \cup Z^k_{B}$. Для кожної з знайдених відокремлюючих гіперплощин $L_{i^k}$, $i^k = \overline{1,l^k}$ будемо розпізнавати точки контрольної множини $Z^k$ відповідно до їх розташування відносно цієї гіперплощини. Позначимо $Z^{+}_{i^k A} = \left\{ z \in Z^k_A : z \in L^k_{i\,A} \right\}$ — множина вірно розпізнаних точок $Z^k_A$, $Z^{+}_{i^k B} = \left\{ z \in Z^k_B : z \in L^k_{i\,B} \right\}$ — множина вірно розпізнаних точок $Z^k_B$, $Z^{-}_{i^k A} = \left\{ z \in Z^k_A : z \in L^k_{i\,B} \right\}$ — множина невірно розпізнаних точок $Z^k_A$, $Z^{-}_{i^k B} = \left\{ z \in Z^k_B : z \in L^k_{i\,A} \right\}$ — множина невірно розпізнаних точок $Z^k_B$.

Тоді відносна помилка відокремлюючої гіперплощини $L_{i^k}$ буде

$$\theta_{i^k} = \frac{m_{Z^{k-}_{i\,A}} + m_{Z^{k-}_{i\,B}}}{m_{Z^k}}.$$

**Означення.** Нехай $\exists i^k : \theta_{i^k} < \theta$, тоді *оптимальною відокремлюючою гіперплощиною в просторі* $R^k$ називатимемо відокремлюючу гіперплощину $L_{i^k_{\min}}$ таку, що $i^k_{\min} = \arg\min\limits_{i^k} \left\{ \theta^k_i < \theta \right\}$.

**Означення.** *Оптимальним класифікатором* для множин $A$ та $B$ на заданому рівні помилок $\theta$ називатимемо оптимальну відокремлюючу гіперплощину, що відповідає простору найменшої розмірності.

**2. Пошук відокремлюючих гіперплощин.** Для знаходження відокремлюючих гіперплощин $L_{i^k}$ в просторі $R^k$, $k = \overline{1,n}$ пропонуємо наступну методику.

Побудуємо методом Джарвіса [3, с. 114–183] однозв'язні опуклі оболонки $conv_{A^k}$ та $conv_{B^k}$ множин $A^k$ та $B^k$. При цьому можливі наступні варіанти взаємного розташування множин та їх опуклих оболонок:

**1 випадок.** $conv_{A^k} \cap conv_{B^k} = \varnothing$. Тобто опуклі оболонки множин не перетинаються (рис. 1).

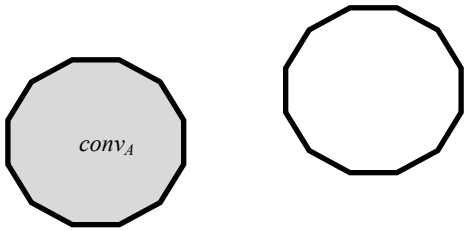



***Рис. 1.*** *1 випадок взаємного розташування множин*

Виходячи з наслідку з теореми Хана-Банаха [4, с. 139–150] та теореми про відокремлюючу вісь [5] сформулюємо наступне твердження [6]:

**Твердження.** Нехай задано два опуклих багатогранники $conv_A$ та $conv_B$, що не перетинаються та їх найближчі точки $a_{\min} \in conv_A$ та $b_{\min} \in conv_B$ такі, що

$$\left|\overrightarrow{a_{\min}b_{\min}}\right| = \min\left\{ \left|\overrightarrow{a_ib_j}\right| : a_i \in conv_A, b_j \in conv_B, i = \overline{1,m_A}, j = \overline{1,m_B} \right\}.$$

Серед гіперграней, що містять ці точки знайдеться хоча б одна така, паралельно якій через точку на відрізку $\overline{a_{\min}b_{\min}}$ можна провести відокремлюючу гіперплощину $L$, що розділяє $conv_A$ та $conv_B$.

Нехай $g_{i^k}$ — одна з гіперграней, паралельно якій, згідно твердження, можна провести відокремлюючу гіперплощину. Серед множини гіперплощин, паралельних гіперграні $g_{i^k}$ в якості відокремлюючої гіперплощини для множин $A$ та $B$ прийматимемо ту, що знаходиться на однаковій відстані від найближчих точок опуклих оболонок $a_{\min}$ та $b_{\min}$.

**2 випадок.** $conv_{A^k} \cap conv_{B^k} \neq \varnothing$, але кількість точок множини $O_{A^k} = \left\{ o^k : o^k \in A^k \cap conv_{B^k} \right\}$ задовольняє нерівність $m_{O_{A^k}} \leq \alpha \cdot m_{A^k}$ (рис. 2)

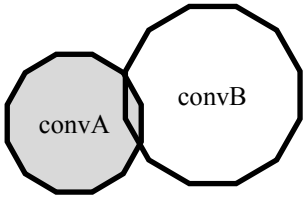



***Рис. 2.*** *2 випадок взаємного розташування множин*

**Означення**. Точку $a_i^k$ називатимемо *промахом* множини $A^k$, якщо $a_i^k \in conv_{B^k}$.

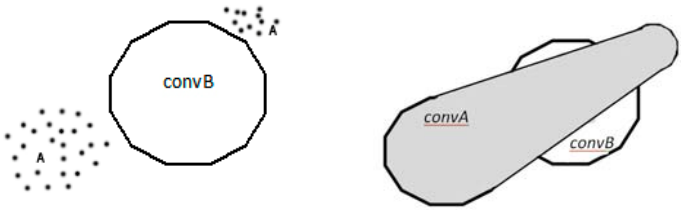
**Означення**. Множину $O_{A^k} = \left\{ o^k \in A^k : o^k \in conv_{B^k} \right\}$ будемо називати *множиною промахів* $A^k$.

Відкинемо промахи з тієї множини, у якої їх менше та побудуємо для неї нову опуклу оболонку. Будуємо нові опуклі оболонки та відкидаємо з них промахи до тих пір, поки не виконується умова

$$conv_{A^k} \cap conv_{B^k} = \varnothing$$

Коли ця умова виконується, $conv_{A^k}$ та $conv_{B^k}$ — це дві замкнені обмежені опуклі множини, що не перетинаються. Отримуємо випадок, аналогічний попередньому.

**3 випадок.** $A^k \notin conv_B$ (рис.3а), але $conv_{A^k} \cap conv_{B^k} \neq \varnothing$ (рис. 3б)




*а)* *б)*

***Рис. 3.*** *3 випадок взаємного розташування множин*

Тобто більшість точок множини $A^k$ розташовані по один бік множини $B^k$. При цьому деякі точки множини $A^k$ розташовані по інший бік множини $B^k$. Якщо їх кількість знаходиться в межах заданого рівня помилок $\alpha$, вважатимемо їх неявними промахами та відкидатимемо.

**Означення**. Пару точок $a_i$ та $a_j$ називатимемо *підозрілою на неявний промах* для множини $A^k$, якщо $a_i, a_j \notin conv_{B^k}$, але існує гіперграань $q \in conv_{B^k}$ така, що точки $a_i$ та $a_j$ лежать в різних півпросторах, утворених гіперплощиною $q$ (рис. 4).

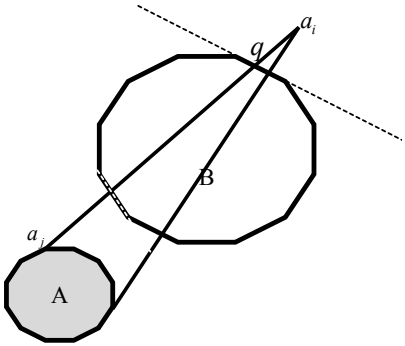



***Рис. 4.*** *Пара точок, підозрілих на неявний промах*

**Означення.** Точка $a_i$ є *неявним промахом* для множини $A^k$, якщо вона з не менше ніж $(1-\alpha)\cdot m_A$ точками множини $A^k$ складає пару, підозрілу на неявний промах.

Множину неявних промахів позначатимемо $P_{A^k}$.

Після відкидання неявних промахів та промахів множин приходимо до першого випадку.

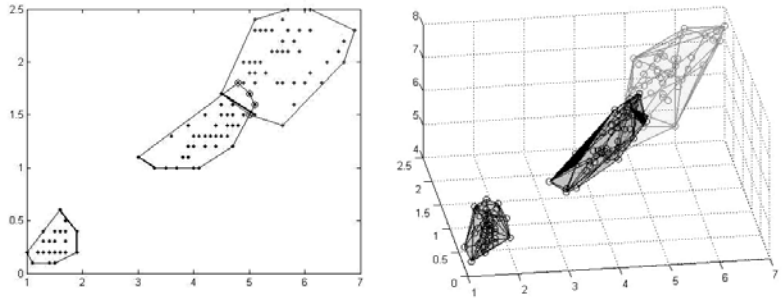
**3. Складність алгоритму методу ЛВОО.** В усіх описаних випадках взаємного розташування множин найбільшу складність має процедура побудови опуклих оболонок. Якщо опуклі оболонки будувати методом Джарвіса, складність алгоритму для кожного $k$ в найгіршому випадку дорівнює $O(m^2)$ [3, с. 135]. Отже, загальну складність алгоритму методу ЛВОО можна оцінити як $O(nm^2)$, якщо використані всі ознаки. Якщо оптимальний класифікатор отримується при використанні невеликої кількості ознак ($k << m$), складність алгоритму методу ЛВОО можна оцінити як $O(m^2)$.

Зауважимо, що складність алгоритму лінійного дискримінантного аналізу оцінюється як $O(mnt + t^3)$, де $t = \min(m, n)$ [7]; наївного байєсівського класифікатору — $O(mn)$ [8]; методу опорних векторів — $O(m^3)$ [9].

**Приклади застосування алгоритму ЛВОО.** Описаний вище алгоритм ЛВОО був реалізований нами в програмі MATLAB R2013b. Наведемо приклади його застосування.

**Приклад 1.** Розглянемо відому базу даних іриси Фішера [10]. Є дані про 150 рослин трьох видів іриса (*setosa, versicolor, virginica*). Для кожної рослини відомо 4 ознаки — довжина і ширина чашолистика та довжина і ширина пелюстки. Необхідно за цими даними побудувати алгоритм класифікації. Ця база даних є цікавою тим, що перший клас (*setosa*) лінійно відокремлюється від двох інших, а класи *versicolor* та *virginica* між собою лінійно не відокремлюються.

Класифікацію проводили за допомогою дискримінантного аналізу [10], наївного байєсівського класифікатору [11, с. 47-48], методом опорних векторів [12, с.156-163] та за допомогою методу ЛВОО (рис. 5). Оскільки об'єми груп невеликі — кожна група складається з 50 рослин, контроль за якістю класифікації проводили методом ковзаючого контролю [2, с. 266-268].



*а)* *б)*

***Рис. 5.*** *Застосування методу ЛВОО для класифікації ірисів Фішера а)* $n = 2$*, б)* $n = 3$

Сумарна кількість помилок в кожному класі, отриманих при застосування різних методів класифікації представлені в таблиці 1.

Таблиця 1

*Сумарна кількість помилок класифікації*

| **Метод** | **Setosa** | **Versicolor** | **Virginica** |
|---|---|---|---|
| Дискримінантний аналіз | 0 | 2 | 1 |
| Наївний байєсівський класифікатор | 0 | 3 | 3 |
| Метод опорних векторів (*лінійне ядро*) | 0 | 4 | 0 |
| Метод опорних векторів (*ядро — гаусівська радіальна базисна функція*) | 0 | 1 | 0 |
| Метод ЛВОО | 0 | 1 | 0 |

**Приклад 2.** Для хворих на гострий деструктивний панкреатит проводили класифікацію між групами хворих з післяопераційними

ускладненнями та без них. Для класифікації використовували дані про 89 хворих (група $A$ — 28 хворих з ускладненнями, група $B$ — 61 хворий без ускладнень).

При класифікації методом ЛВОО контроль за якістю класифікації проводили методом ковзаючого контролю. Оскільки практична мета даної класифікаційної моделі — не пропустити хворих, у яких можуть виникнути післяопераційні ускладнення, тобто зменшити кількість помилок гіподіагностики, то за допустимий рівень помилок прийняли відсоток помилок класифікації множини $A$: $\theta_A = 0,05$. Заданий рівень помилок був досягнений при використанні $n = 3$ ознак (рис.6). Класифікацію іншими методами відповідно проводили також за трьома ознаками.

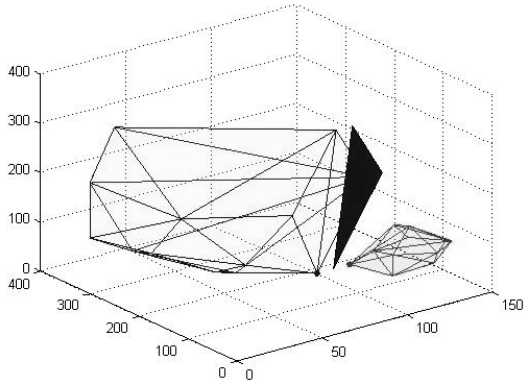



***Рис. 6.*** *Застосування методу ЛВВО для прогнозування виникнення ускладнень у хворих на гострий деструктивний панкреатит ( $n = 3$ )*

Сумарна кількість помилок в кожному класі, отриманих при застосування різних методів класифікації представлені в таблиці 2.

Таблиця 2

*Сумарна кількість помилок класифікації*

| **Метод** | **Хворі з ускладненнями** | **Хворі без ускладнень** |
|---|---|---|
| Дискримінантний аналіз | 1 | 4 |
| Наївний байєсівський класифікатор | 2 | 2 |
| Метод опорних векторів (*лінійне ядро*) | 1 | 4 |
| Метод опорних векторів (*ядро — гаусівська радіальна базисна функція*) | 1 | 3 |
| Метод ЛВОО | 1 | 4 |

**Висновок.** В роботі запропонований метод лінійної класифікації двох однозв'язних множин. Суть методу полягає в знаходженні відокре-

млюючої гіперплощини, що паралельна одній з гіперграней опуклих оболонок множин, що класифікуються. За оптимальну відокремлюючу гіперплощину вибирається гіперплощина з найменшою кількістю помилок в контрольній групі. В якості класифікатора вибирається оптимальна відокремлююча гіперплощина, що відповідає простору меншої розмірності. Складність методу ЛВОО оцінюється як $O(nm^2)$.

**Список використаних джерел:**


1. Кульбак С. Теория информации и статистика / С. Кульбак. — М. : Наука, 1967. — 408 с.
2. Вапник В. Н. Восстановление зависимостей по эмпирическим данным / В. Н. Вапник. — М. : Наука, 1979. — 448 с.
3. Препарата Ф. Вычислительная геометрия: Введение / Ф. Препарата, М. Шеймос. — М. : Мир, 1989. — С. 478.
4. Колмогоров А. Н. Элементы теории функций и функционального анализа / А. Н. Колмогоров, С. В. Фомин. — 7-е изд. — М. : ФИЗМАТЛИТ, 2004. — 572 с.
5. SAT (Separating Axis Theorem) [Електронний ресурс]. — Режим доступу: http://www.codezealot.org/archives/55.
6. Ivanchuk M. Mathematical Modeling of the Expert System Predicting the Severity of Acute Pancreatitis / M. Ivanchuk, V. Maksimyuk, I. Malyk // Journal of Computational Medicine. — Vol. 2014. — Article ID 532453.
7. Deng Cai, Xiaofei He, Jiawei Han Training Linear Discriminant Analysis in Linear Time [Електронний ресурс]. — Режим доступу: http://researchweb.iiit.ac.in/~nataraj.j/poseSearchReports/icde08_dengcai.pdf
8. Chris Fleizach, Satoru Fukushima A naive Bayes classifier on 1998 KDD Cup [Електронний ресурс]. — Режим доступу: http://sysnet.ucsd.edu/~cfleizac/-cse250b/project1.pdf.
9. Vector Machines: Fast SVM Training on Very Large Data Sets / I. W. Tsang, J. T. Kwok, Pak-Ming Cheung // Journal of Machine Learning. — 2005. — Research 6. — P. 363–392. Submitted 12/04; Published 4/05
10. Fisher R. A. The Use of Multiple Measurements in Taxonomic Problems / R. A. Fisher // Annals of Eugenics 7. — 1936. — P. 179–188.
11. Прикладная статистика: классификация и снижение размерности / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин. — М. : Финансы и статистика, 1989. — 608 с.
12. Vapnik V. N. The nature of statistical learning theory / V. N. Vapnik. — 2nd ed. — New York, 2000. — 314 p.



The method of convex hulls linear separation for classification of two sets in Euclidian space in $R^n$ is proposed. Examples for comparing the results of using the method of convex hulls linear separation and discriminant analysis, naive Bayes classifier and SVM are described in the manuscript.



**Key words:** *linear classification, convex hulls, separating hyperplane.*

Отримано: 14.04.2015