УДК 519.237.8

М.А. Іванчук (Буковинський держ. мед. ун-т)

ВИКОРИСТАННЯ КЛАСТЕРНОГО АНАЛІЗУ ТА ДІАГРАМ ВОРОНОГО ДЛЯ НЕЛІНІЙНОЇ КЛАСИФІКАЦІЇ

Two ways for solution the task of non-linear classification of two sets in the Euclidian space R^n are proposed. The first way is based on the using k-means cluster analysis, the second one uses Voronoy diagram. There are examples for comparing classification results by the own method and by using Support Vector Machines in the manuscript

В роботі представлені два підходи до розв'язання задачі нелінійної класифікації двох множин в евклідовому просторі R^n. Перший метод базується на використанні кластерного аналізу методом k-середніх з подальшим лінійним розділенням опуклих оболонок отриманих кластерів. В другому методі для класифікації множин використовуються діаграми Вороного. Наводяться приклади для порівняння результатів класифікації запропонованими методами та методом опорних векторів.

Вступ. Для розв'язання задач лінійної класифікації існує багато добре розроблених методик, з яких найвідомішими є лінійний дискримінантний аналіз [1] та байєсівський класифікатор [2]. Задачі нелінійної класифікації є складнішими для розв'язання, оскільки кожна конкретна задача вимагає свого підходу. Так, наприклад, при використанні методу опорних векторів [3] якість класифікації залежить від правильно вибраного ядра. Ми пропонуємо власний підхід до розв'язання задач нелінійної класифікації.

1. Постановка задачі. Нехай задано дві множини точок $A = \{a_i = (a_i^1, a_i^2, \ldots, a_i^n),\ i = \overline{1, m_A}\}$ та $B = \left\{b_i = (b_i^1, b_i^2, \ldots, b_i^n),\ i = \overline{1, m_B}\right\}$ у евклідовому просторі R^n. Необхідно створити класифікатор, який з наперед заданим рівнем помилок α розділить множини A та B. Нехай крім цього задані множина точок контрольної групи $Z = \left\{z_i = (z_i^1, z_i^2, \ldots, z_i^n),\ i = \overline{1, m_Z}\right\}$, для кожної з яких відомо, до якої групи (A чи B) вона відноситься, та допустимий рівень помилок на контрольній групі θ.

В даній роботі ми розглянемо випадок взаємного розташування множин, коли множини A та B не перетинаються, але опукла оболонка множини A лежить всередині опуклої оболонки множини B. Тобто $conv_A \bigcap conv_B \neq \emptyset$, але $convA \bigcap B = \emptyset$.

Для розв'язання поставленої задачі пропонуємо 2 методи.

1. Нелінійна класифікація з використанням кластерного аналізу. Проведемо кластерний аналіз методом k-середніх [4] для множин A та B. Розіб'ємо на k кластерів окремо кожну з множин A та B. Одержимо $A = \bigcup_{i=1}^{k} A_i$ та $B = \bigcup_{i=1}^{k} B_i$. Розглянемо всі можливі k^2 пар підмножин A_i,B_j, $i = \overline{1, k},\ j = \overline{1, k}$. Для кожної з пар підмножин шукатимемо відокремлюючі гіперплощини $L_{i,j}$ методом лінійного відокремлення опуклих оболонок [5]. Опишемо суть даного методу.

Означення. Гіперплощину $L_{ij} = \{x \in R^n : \langle p_{ij}, x\rangle = \gamma_{ij}\}$, $p_{ij} \neq 0$ будемо називати *відокремлюючою гіперплощиною* для підмножин A_i та B_j, якщо не менше ніж $(1-\alpha)(m_A + m_B)$ точок множин A_i та B_j можна помістити в різні півпростори $L_{ij}^A = \{x \in R^n : \langle p_{ij}, x\rangle > \gamma_{ij}\}$ та $L_{ij}^B = \{x \in R^n : \langle p_{ij}, x\rangle < \gamma_{ij}\}$.

Якщо опуклі оболонки $conv_{A_i}$ та $conv_{B_j}$ не перетинаються, то згідно наслідку з теореми Хана-Банаха [6] та теореми про відокремлюючу вісь [7] відокремлюючу гіперплощину шукаємо паралельно до гіпергранеї цих опуклих оболонок так, щоб вона знаходилася на однаковій відстані від найближчих точок опуклих оболонок $a_{\min\, i}$ та $b_{\min\, j}$. Отримаємо скінчену кількість відокремлюючих гіперплощин $L_{ij}^s,\ s = \overline{1, l_{ij}}$. Для визначення оптимальної відокремлюючої гіперплощини для підмножин A_i та B_j, перевіряємо результат на контрольній множині Z.

Позначимо Z_A – точки контрольної множини, що відносяться до групи A, відповідно Z_B – точки контрольної множини, що відносяться до групи B, $Z = Z_A \bigcup Z_B$. Для кожної з знайдених відокремлюючих гіперплощин $L_{ij}^s,\ s = \overline{1, l_{ij}}$. будемо розпізнавати точки контрольної множини Z відповідно до їх розташування відносно цієї гіперплощини. Позначимо

$Z_{ij,A}^{s^+} = \left\{z \in Z_A : z \in L_{ij,A}^s\right\}$ — множина вірно розпізнаних точок множини Z_A,

$Z_{ij,B}^{s^+} = \left\{z \in Z_B : z \in L_{ij,B}^s\right\}$ — множина вірно розпізнаних точок множини Z_B,

$Z_{ij,A}^{s^-} = \left\{z \in Z_A : z \in L_{ij,B}^s\right\}$ — множина невірно розпізнаних точок множини Z_A,

$Z_{ij,B}^{s^-} = \left\{z \in Z_B : z \in L_{ij,A}^s\right\}$ — множина невірно розпізнаних точок множини Z_B.

Тоді відносна помилка відокремлюючої гіперплощини L_{ij}^s буде $\theta_{ij}^s = \frac{m_{Z_{ij,A}^{s^-}} + m_{Z_{ij,B}^{s^-}}}{m_Z}$.

Означення. *Оптимальною відокремлюючою гіперплощиною* для підмножин A_i та B_j називатимемо відокремлюючу гіперплощину $L_{ij}^{s_{\min}}$ таку, що $s_{\min} = \arg\min\limits_s \left\{\theta_{ij}^s\right\}$.

Якщо опуклі оболонки підмножин A_i та B_j перетинаються, знаходимо для них множини промахів $O_{A_{ij}}$ та неявних промахів $P_{A_{ij}}$

Означення. Точку $o \in A_i$ називатимемо *промахом* множини A_i відносно множини B_j, якщо $o \in conv_{B_j}$.

Означення. Множину $O_{A_{ij}} = \left\{o : o \in A_i \bigcap conv_{B_j}\right\}$ будемо називати *множиною промахів* A_i відносно B_j.

Означення. Пару точок $a_1, a_2 \in A_i$ називатимемо *підозрілою на неявний промах* для підмножини A_i відносно підмножини B_j, якщо $a_1, a_2 \notin conv_{B_j}$, але існує гіперграань $q \in conv_{B_j}$ така, що точки a_1 та a_2 лежать в різних півпросторах, утворених гіперплощиною q.

Означення. Точка $a \in A_i$ є *неявним промахом* для підмножини A_i відносно підмножини B_j, якщо вона з не менше ніж $(1-\alpha)\cdot m_A$ точками підмножини A_i складає пару, підозрілу на неявний промах.

Множину неявних промахів A_i відносно B_j позначатимемо $P_{A_{ij}}$.

Відкинувши промахи та неявні промахи з підмножини A_i, можна побудувати відокремлюючі гіперплощини для підмножин A_i та B_j та знайти серед них оптимальну.

Позначимо $O = \bigcup_{i,j=1}^k O_{ij}$ — множина всіх точок-промахів, $P = \bigcup_{i,j=1}^k P_{ij}$ — множина всіх точок — неявних промахів.

Якщо $m_O + m_P > \alpha(m_A + m_B)$, тобто загальна кількість відкинутих точок промахів та неявних промахів не відповідає заданому рівню помилок α, збільшуємо кількість кластерів на 1. Якщо при цьому знайдеться хоча б один кластер, кількість точок якого менша за $\alpha(m_A + m_B)$, робимо висновок про неможливість класифікації множин на заданому рівні помилок.

Якщо заданий рівень помилок α досягнуто, то розв'язок задачі класифікації — це сукупність оптимальних відокремлюючих гіперплощин всіх пар підмножин. Для визначення, до якої множини відносити точку контрольної вибірки $z \in Z$, необхідно знайти найближчі до неї кластерні середні $i' = \arg\min\{\rho(z, a_i), i = \overline{1,k}\}$, $j' = \arg\min\{\rho(z, b_j), j = \overline{1,k}\}$, де ρ — евклідова метрика, a_i — середнє i-го кластеру множини A, b_j — середнє j-го кластеру множини B. Точку z відносимо до множини A, якщо вона належить півпростору $L^{s_{\min}}_{i'j',A}$ або до множини B, якщо вона належить півпростору $L^{s_{\min}}_{i'j',B}$.

Алгоритм методу

1. $k = 2$.
2. Розділити множини A та B на k кластерів: $A = \bigcup_{i=1}^{k} A_i$ та $B = \bigcup_{i=1}^{k} B_i$.
3. Розглянути пари підмножин A_i,B_j $(\forall i = \overline{1,k},\ j = \overline{1,k})$.
   - 3.1. Знайти множини промахів $O_{A_{ij}}$ та неявних промахів $P_{A_{ij}}$. Відкинути промахи та неявні промахи.
   - 3.2. Знайти всі можливі відокремлюючі гіперплощини L^s_{ij}, $s = \overline{1, l_{ij}}$.
   - 3.3. Знайти оптимальну відокремлюючу гіперплощину $L^{s_{\min}}_{ij}$
4. Знайти загальну кількість промахів m_O та неявних промахів m_P.
5. *Якщо* $m_O + m_P > \alpha(m_A + m_B)$, то

   якщо існує хоча б один кластер, кількість точок якого менша за $\alpha(m_A + m_B)$, то

   висновок про неможливість класифікації на заданому рівні помилок α. Вихід з алгоритму.

   інакше

   $k = k + 1$. Перейти до кроку 2.

   інакше

   знайдена сукупність оптимальних відокремлюючих гіперплощин.
6. Для довільної точки $z \in Z$:
   - 6.1. Знайти найближчі кластерні середні $i' = \arg\min\left\{\rho(z, a_i), i = \overline{1,k}\right\}$, $j' = \arg\min\left\{\rho(z, b_j), j = \overline{1,k}\right\}$, де ρ — евклідова метрика, a_i — середнє i-го кластеру множини A, b_j — середнє j-го кластеру множини B.
   - 6.2. *Якщо* $z \in L^{s_{\min}}_{i'j',A}$, то відносимо z до множини A, *інакше* відносимо z до множини B.

Визначимо **складність** даного алгоритму. Процедура кластерного аналізу методом k-середніх має складність $O(mkl)$, де $m = m_A + m_B$, l – число ітерацій [8]. При побудові опуклих оболонок методом Джарвіса складність алгоритму оцінюється як $O(m^2)$ [8]. Процедуру знаходження найближчого кластера можна реалізувати з складністю $O(\log m)$ [9]. Пошук опуклих оболонок на кожному з k

кроків повторюється k^2 разів, отже, в загальному складність алгоритму можна оцінити як $O(k^3m^2)$.

3. Нелінійна класифікація з використанням діаграм Вороного. Для розв'язання поставленої задачі будемо використовувати діаграми Вороного [8]. Об'єднаємо множини A та B, побудуємо для множини $D = A \bigcup B$ діаграму Вороного.

Розглянемо деяку точку $d_i \in D$, $i = \overline{1, m_A + m_B}$. Не порушуючи загальності міркувань, припустимо, що $d_i \in A$. Позначимо V_i — багатокутник Вороного для точки d_i. Точки, для яких багатокутники Вороного мають суміжні грані з багатокутником V_i будемо називати найближчими сусідами точки d_i. Множину найближчих сусідів точки d_i позначатимемо NV_i.

Для багатокутника Вороного V_i існують наступні можливості:

1) Всі найближчі сусіди точки d_i, належать точкам з множини A, тобто $\forall n \in \in NV_i : \ n \in A$. В цьому випадку точка d_iє ***внутрішньою для множини*** A.

2) Всі найближчі сусіди точки d_i, належать точкам з множини B, тобто $\forall n \in \in NV_i : \ n \in B$. В цьому випадку точка d_i є ***промахом для множини*** A. Сукупність промахів для множини A позначатимемо O_A.

3) Серед найближчих сусідів точки d_i є такі, що належать множині A та такі, що належать множині B. При цьому точка d_i може бути крайньою точкою множини A або разом з одним (або декількома) своїми сусідами бути промахом множини A.

Будемо вважати точку d_i ***крайньою точкою множини*** A, якщо від точки d_i до будь-якої внутрішньої точки множини A існує шлях, що проходить лише через точки множини A. В противному випадку вважатимемо точку d_i промахом множини A.

Відкинемо всі точки-промахи множин A та B. Якщо кількість відкинутих точок $(m_{O_A} + m_{O_B}) > \alpha(m_A + m_B)$, робимо висновок про неможливість класифікації на заданому рівні значущості. В противному випадку будуємо діаграму Вороного для множини $D' = A' \bigcup B'$, де$A' = A/O_A$, $B' = B/O_B$.

Одержана діаграма Вороного розбиває простір R^n на два півпростори: $R^n_{A'}$ та $R^n_{B'}$. Точку контрольної вибірки $z \in Z$ відносимо до множини A, якщо вона попадає в багатокутник Вороного, що відповідає деякій точці $a \in A'$, тобто $z \in \in V_i : d_i \in A'$. Або до множини B, якщо вона попадає в багатокутник Вороного, що відповідає деякій точці $b \in B'$, тобто $z \in V_i : d_i \in B'$.

Алгоритм методу

1. Побудувати діаграму Вороного для множини $D = A \bigcup B$.

2. Відкинути всі точки-промахи множин A та B.

3. *Якщо* $(m_{O_A} + m_{O_B}) > \alpha(m_A + m_B)$, *то* висновок про неможливість класифікації на заданому рівні значущості. Вихід з алгоритму *інакше* побудувати діаграму Вороного для множини $D' = A' \bigcup B'$, де $A' = A/O_A$, $B' = B/O_B$.

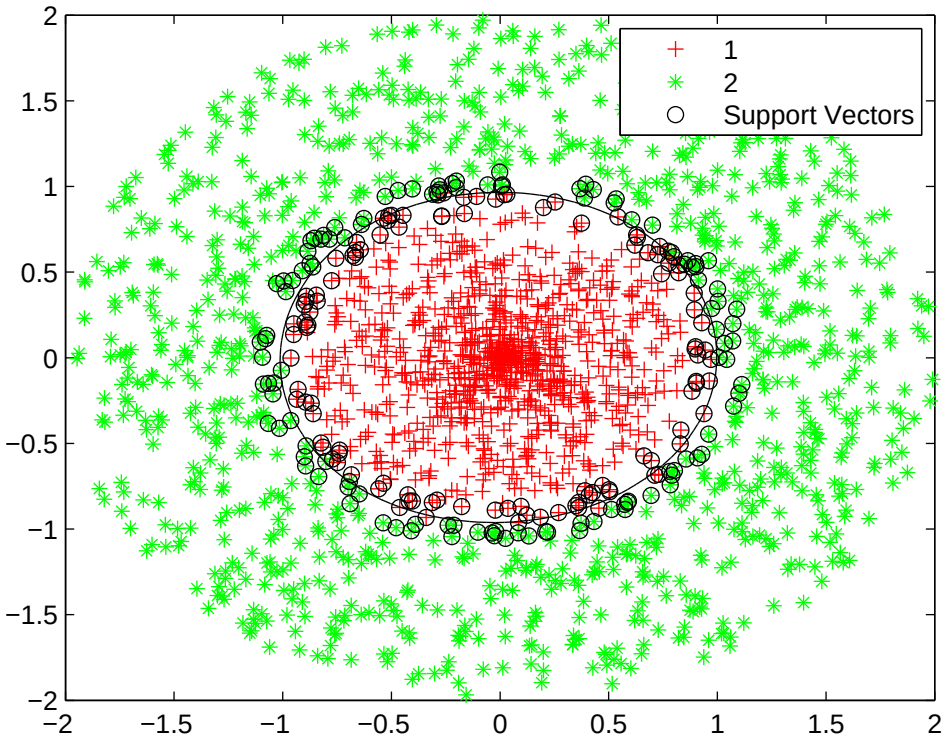
4. Для точки контрольної вибірки $z \in Z$ *якщо* $z \in V_i : d_i \in A'$, то відносимо z до множини A, *інакше* відносимо z до множини B.

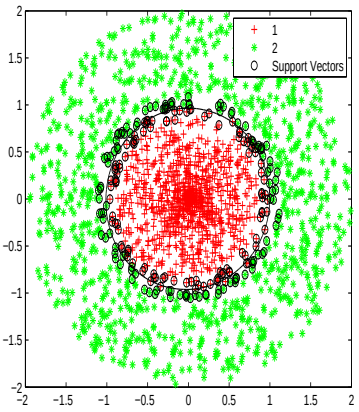
Складність алгоритму відповідає складності побудови діаграми Вороного, що оцінюється як $O(m \log m)$ [8].

Приклади застосування алгоритму НВОО

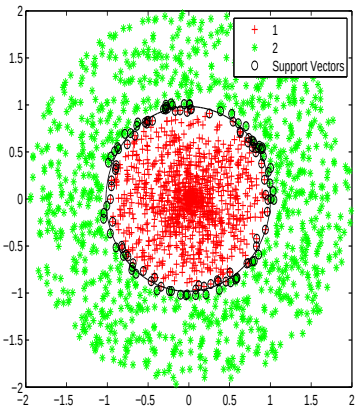
Проводитимемо класифікацію методом опорних векторів та за допомогою методу НВОО. Зауважимо, що складність методу опорних векторів оцінюється як $O(m^3)$ [10]. Для перевірки якості класифікації скористаємося методом Монте-Карло, який полягає в отриманні великого числа реалізацій стохастичного процесу [11]. Для цього випадковим чином згенеруємо множини Aта B по 1000 точок в кожній, а також контрольну множину Z, що складається з 1000 точок, що відповідають множині A та 1000 точок, що відповідають множині B. Для кожного методу класифікації визначимо кількість невірно розпізнаних точок контрольної множини. Повторимо проведення класифікації 500 разів та знайдемо середню кількість невірно розпізнаних точок контрольної множини для кожного методу класифікації

Приклад 1. Розглянемо дві множини точок $A = \{(x, y) : x^2 + y^2 \leq 1\}$ та $B = \{(x, y) : x^2 + y^2 > 1,\ x^2 + y^2 \leq 2\}$, які не відокремлюються лінійно (рис.1). Результати класифікації представлені в таблиці 1.

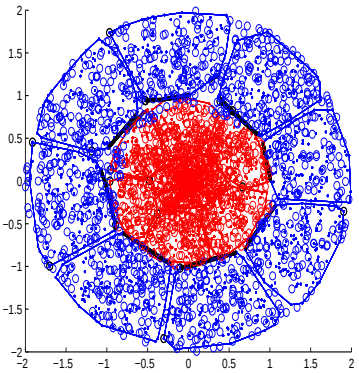




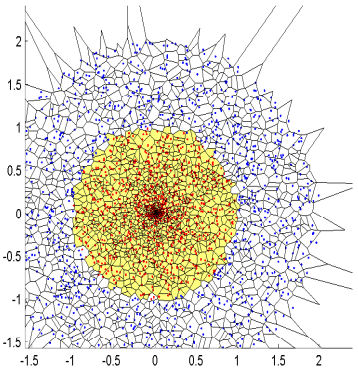
Метод опорних векторів (*ядро – гаусівська радіальна базисна функція*)



Метод опорних векторів (*квадратичне ядро*)



Метод НВОО 1 спосіб (7 кластерів)



Метод НВОО 2 спосіб

Рис.1

Таблиця 1. Середня кількість невірно розпізнаних точок

Метод	**Множина** Z_A	**Множина** Z_B
Метод опорних векторів (*ядро – гаусівська радіальна базисна функція*)	8,86	6,25
Метод опорних векторів (*квадратичне ядро*)	3,83	6,47
Метод НВОО 1 спосіб	69,98	7,87
Метод НВОО 2 спосіб	11,73	12,6

Метод	**Множина** Z_A	**Множина** Z_B
Метод опорних векторів (*ядро – гаусівська радіальна базисна функція*)	2,776	0,014
Метод опорних векторів (*квадратичне ядро*)	2,667	0,456
Метод HBOO (*кластерний аналіз*)	1,102	0,434
Метод HBOO (*діаграми Вороного*)	1,790	4,938

Висновок

Задачі нелінійної класифікації можна розв'язувати за запропонованою методикою з використанням кластерного аналізу або діаграм Вороного. Складність алгоритму класифікації з використанням кластерного аналізу оцінюється як $O(k^3m^2)$, з використанням діаграм Вороного – $O(m \log m)$, що є меншим, ніж складність методу опорних векторів $O(m^3)$. При цьому результати класифікації запропонованими методами не гірші, ніж при застосуванні методу опорних векторів.

Список використаної літератури

1. *Fisher, R.A.* The Use of Multiple Measurements in Taxonomic Problems // Annals of Eugenics 7. – 1936. – P. 179–188.
2. *Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д.* Прикладная статистика: классификация и снижение размерности. – М.: Финансы и статистика, 1989. – 608 с.
3. *Vapnik V.N.* The nature of statistical learning theory. – 2^{nd} ed. – New York. – 2000. – 314 p.
4. *Hoppner Frank et all.* Fuzzy Cluster Analysis. Methods for classification, data analysis and image recognition. – John Wiley & Sons, LTD. – 289 p.
5. *Ivanchuk M., Maksimyuk V., Malyk I.* Mathematical Modeling of the Expert System Predicting the Severity of Acute Pancreatitis // Journal of Computational Medicine. – vol. 2014. – Article ID 532453
6. *Колмогоров А. Н., Фомин С.В.* Элементы теории функций и функционального анализа. — 7-е изд. — М.: ФИЗМАТЛИТ, 2004. – 572 с.
7. SAT (Separating Axis Theorem) [Электрон. ресурс]. – Режим доступу: `http://www.codezealot.org/archives/55`.
8. *Препарата Ф.* Вычислительная геометрия: Введение / Франко Препарата, Майкл Шеймос. – М.: Мир, 1989. – С. 478.
9. *Мандель И.Д.* Кластерный анализ . – М.: Финансы и статистика. – 1998. – 176 с.: ил.
10. *Ермаков С. М.* Метод Монте-Карло и смежные вопросы. – М.: Наука, 1971. – 328 с.
11. *Tsang I. W., Kwok J. T.* Pak-Ming Cheung Core Vector Machines: Fast SVM Training on Very Large Data Sets // Journal of Machine Learning Research. – 6 (2005). – P. 363–392

Одержано 08.06.2015