УДК 519.7

Галкін О. А., к.ф.-м.н., м.н.с.

Непараметричний метод класифікації для задач з нееліптичним розподілом даних на основі глибиннозалежної Σ-схеми

Київський національний університет імені Тараса Шевченка, 03680, м. Київ, пр-т. Академіка Глушкова 4д,
e-mail: galkin.o.a@gmail.com

O. A. Galkin, candidate of physical and mathematical sciences, junior researcher

Nonparametric classification method for problems with non-elliptical distribution of data based on depth dependent Σ-diagram

Taras Shevchenko National University of Kyiv, 03680, Kyiv, Glushkova st., 4d,
e-mail: galkin.o.a@gmail.com

Стаття присвячена розробці непараметричного методу класифікації, що не вимагає попередньої інформації про розподіл або форму розділової кривої. Запропоновано новий Σ-класифікатор для розв'язання багатокласових задач класифікації з нееліптичним розподілом даних на основі методу мажоритарного голосування. Реалізовано процедуру використання методу крос-перевірки для вибору ступеня розділового многочлена в алгоритмі реалізації Σ-класифікатора. Досліджено автоматичну технологію визначення форм розділових кривих Σ-класифікатора по геометричній структурі даних, що лежать в основі Σ-схеми.

Ключові слова: Σ-класифікатор, задача класифікації, непараметричний розподіл

The article is devoted to developing nonparametric classification algorithm that does not require prior information about distribution or form of dividing curve. The new Σ-classifier is proposed to solve multiclass classification problems with non-elliptical of distribution data on the basis of majority voting method. The procedure of using the cross-validation method is implemented for selecting the degree of the dividing polynomial in the algorithm of Σ-classifier. This procedure obviously outperforms the linear division when the linear division fails to reach the efficiency of the Bayes classifier. Such approach can also be a more objective approach to select the suitable concept of depth in Σ-classifiers. The automatic technology is studied to determine the form of dividing curves of Σ-classifier by the geometric structure of the data that are the basis of the scheme. The advantage of the proposed method is that there is no need to evaluate parameters such as means and scales, which is often required by the most classification methods. Comparative analysis shows that the Σ-classifier frequently is more effective than the maximum depth classifier, and is comparable with the k-nearest neighbor method. The constructed Σ-classifier is absolutely data operated and its classification result can be reproduced on a two-dimensional diagram despite the dimension of the data.

Key Words: Σ-classifier, classification problem, nonparametric distribution

Статтю представив д.ф.-м.н., проф. Анісімов А.В.

Вступ

Непараметричні класифікатори є досить гнучкими щодо адаптації різних структур даних. Враховуючи актуальність даної проблематики, у даній роботі запропоновано та досліджено новий непараметричний класифікатор на основі глибиннозалежної Σ-схеми, що визначає значення глибини точок двох заданих вибірок відносно двох відповідних розподілів, а також трансформує вибірки довільної розмірності в двовимірну діаграму розсіювання. Ідея запропонованого Σ-класифікатора полягає в отриманні кривої, що найбільш ефективно розділяє дві вибірки в Σ-схемі таким чином, що розділення забезпечує мінімальний коефіцієнт помилкової класифікації в заданій Σ-схемі. Виходячи з цього, розглядається Σ-схема, де

$\{Z_1,...,Z_n\}(\equiv \mathrm{Z})$ та $\{X_1,...,X_m\}(\equiv \mathrm{X})$ є двома випадковими вибірками, відповідно з розподілів H та U, що є визначеними на R^r. Отже, Σ-схема визначається як

$$\Sigma(H,U)=\{(E_H(z),E_U(z)),\ z\in \mathrm{Z}\cup \mathrm{X}\}, \quad (1)$$

де $E(\cdot)$ означає $\forall$ дійсне поняття глибини. Зазначимо, що Σ-схема є завжди двовимірним графіком, незалежно від розмірів вибірки та визначається як

$$\Sigma(H_n,U_m)=\{(E_{H_n}(z),E_{U_m}(z)),\ z\in \mathrm{Z}\cup \mathrm{X}\}, \quad (2)$$

якщо H та U є невідомими. Далі буде показано, що Σ-схеми можуть виявляти певні відмінності між двома множинами даних, на основі яких може бути побудований новий класифікатор.

Дана стаття охоплює лише двокласові задачі, хоча запропонований метод класифікації може бути застосований до багатокласових задач з використанням методу мажоритарного голосування [1].

Нехай $\{Z_1,...,Z_n\}(\equiv \mathrm{Z})$ та $\{X_1,...,X_m\}(\equiv \mathrm{X})$ є двома випадковими вибірками, відповідно з розподілів H та U, що є визначеними на R^r. Якщо $H=U$, тоді Σ-схема повинна бути зосереджена вздовж прямої в 45 градусів, як слідує з визначення Σ-схеми в (2). Зазначимо, що Σ-схема матиме певне відхилення від прямої в 45 градусів за умови, якщо розподіли H та U є різними.

На рис. 1-2 показано побудову Σ-схеми за допомогою глибини Махаланобіса для двох двовимірних нормальних вибірок розміру 300 з різницею у розташуванні даних. Зазначимо, що вибірка 1 має стандартний двовимірний нормальний розподіл даних, а вибірка 2 – із середнім зміщенням до $(2,0)'$. На рис. 1 квадрати червоного кольору означають елементи даних з Z, а символи "×" - елементи даних з X, що полегшує ідентифікацію опорних точок однієї вибірки в порівнянні з іншою. Тобто, Σ-схема на рис. 1 демонструє, що елементи даних з двох різних вибірок відображаються симетрично навколо прямої в 45 градусів. Крім того, пряма в 45 градусів є найбільш ефективним способом для розділення двох вибірок в Σ-схемі. У даному випадку правило класифікації розділової прямої в 45 градусів призначитиме z до H, якщо $E_{H_n}(z)>E_{U_m}(z)$ та z до U в іншому випадку.

Отже, найбільш ефективна розділова пряма між двома вибірками в Σ-схемі повинна забезпечувати продуктивність найбільш ефективної розділової прямої між двома вибірками у вихідному просторі R^2.

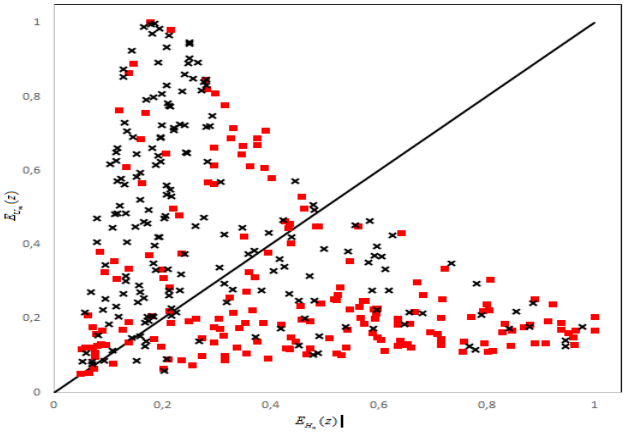

Рис. 1. Експеримент 1-1

Вибірки в їх вихідному просторі R^2 можна побачити на рис.2, де квадрати червоного кольору означають елементи даних з Z, а "×" - елементи даних з X. Обидві криві, що є дуже близькими, отримані таким чином: товста крива - шляхом відображення в Σ-схемі прямої в 45 градусів у простір R^2, а тонка крива - генерується з правила Байеса [2].

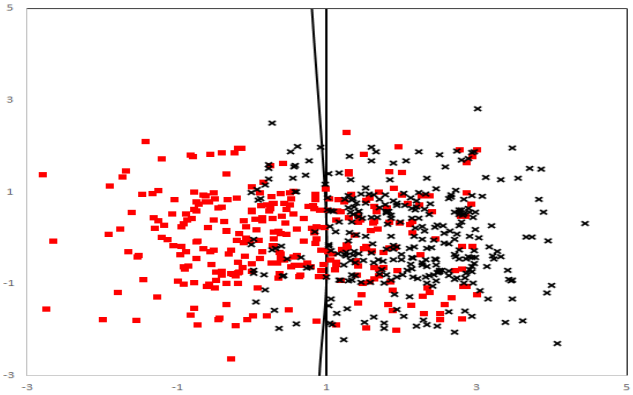

Рис. 2. Експеримент 1-2

Для дослідження відмінностей у масштабі ми множимо всі елементи даних в Z на 3 та розглядаємо нову вибірку Z, що має двовимірний нормальний розподіл із середнім значенням $(0,0)'$ та коваріаційною матрицею $7\Lambda_2$, де Λ_2 є двовимірною одиничною матрицею.

На рис. 3 представлено Σ-схему X та нову вибірку Z, де обидві двовимірні нормальні вибірки з різницею в розташуванні даних та масштабі з розділенням прямої в Σ-схемі більше не відображаються симетрично. У даному випадку елементи даних з Z розташовані в напрямку осі x, а елементи даних з X - до вертикальної прямої $x=1$. Для цих двох вибірок ми застосовуємо класифікатор максимальної глибини, що еквівалентний побудові прямої в 45 градусів та присвоєнню елементів даних до U та

H, коли вони розташовані вище та нижче прямої, відповідно. Очевидно, що дане правило класифікації призначатиме більшість елементів даних з X в H та забезпечить високий коефіцієнт помилкової класифікації [3]. Таким чином, на основі результатів Σ-схеми можна зробити висновок, чому класифікатор максимальної глибини має низьку продуктивність, коли розподіли мають різні дисперсійні структури [4,5].

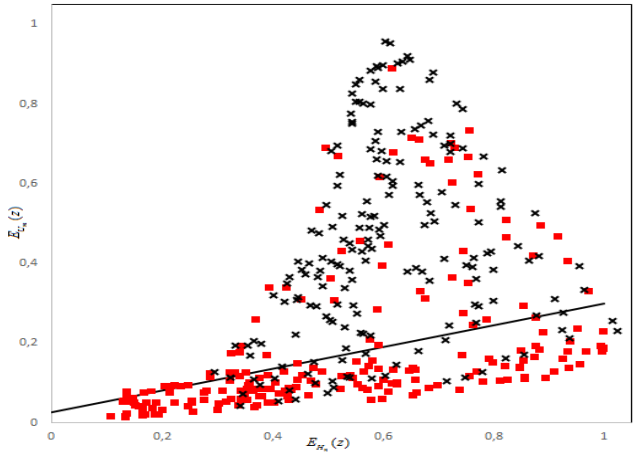

Рис. 3. Експеримент 2-1

З рис. 3 зрозуміло, що існує пряма, яка може ефективно розділити дві вибірки, незважаючи на те, що дві такі вибірки в Σ-схемі більше не мають розкиду в симетричній моделі.

Подальше дослідження показує, що коефіцієнт помилкової класифікації може знизитися, якщо розділова пряма на рис. 3 замінюється відповідним многочленом, як показано на рис. 5.

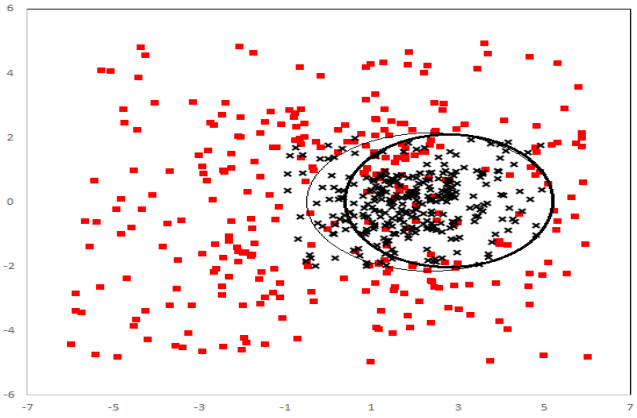

Рис. 4. Експеримент 2-2

Зауважимо, що жирні суцільні кола на рис. 4 та рис. 6, де зображені дві двовимірні нормальні вибірки з різницею в розташуванні та масштабі з поліноміальним розділом даних в Σ-схемі, є розділяючими кривими в початковому просторі, що відповідають прямій на рис. 3 та многочлену на рис. 5, відповідно.

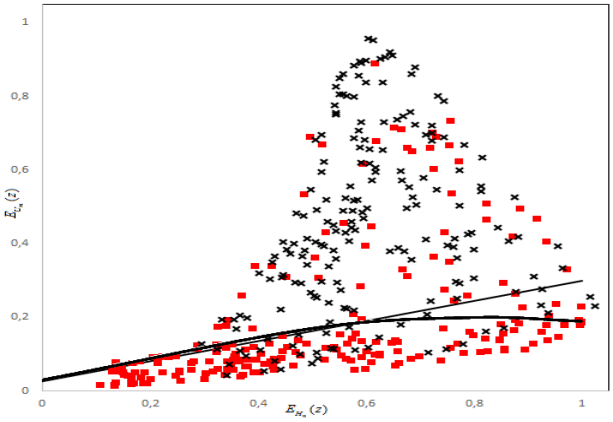

Рис. 5. Експеримент 3-1

Крім того, на рис. 4 найбільше коло є розділовою кривою, отриманою з правила Байеса. Елементи даних з рис. 1-6 припускають, що крива або пряма, що найкраще розділяє дві вибірки в Σ-схемі буде найбільш ефективною розділовою кривою (прямою) між двома вибірками у вихідному просторі.

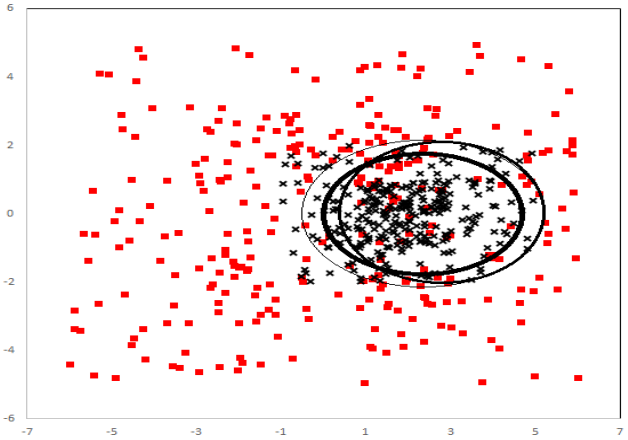

Рис. 6. Експеримент 3-2

Наступна теорема для унімодальних та еліптичних розподілів підтверджує дане явище.

Теорема 1. Нехай $h_1(\cdot)$ та $h_2(\cdot)$ є еліптичними функціями щільності розподілів H та U, відповідно. Також припустимо, що дані функції є такої форми:

$$h_i(z) = v_i \left|\Xi_i\right|^{-1/2} f_i((z-\varepsilon_i)'\Xi_i^{-1}(z-\varepsilon_i)), \qquad (3)$$

де $i = 1,2$, а $f_i(\cdot)$ є строго спадними функціями. Тоді, правило Байеса є еквівалентним такій схемі:

а) якщо $E_U(z) > d(E_H(z))$, то $z \in U$,

б) якщо $E_U(z) < d(E_H(z))$, то $z \in H$,

що має місце зп умови, що E_H та E_U є строго зростаючими функціями від h_1 та h_2, відповідно, а $d(\cdot)$ є деякою дійсною зростаючою функцією.

Доведення. Очевидно, що $h_1(z)=u_1(E_H(z))$ та $h_2(z)=u_2(E_U(z))$, оскільки E_H та E_U є строго спадними функціями від $h_1(\cdot)$ та $h_2(\cdot)$, відповідно. Тому, правило Байеса може бути виражене таким чином:

$$p_2h_2(z)/\,p_1h_1(z)>1 \;\Leftrightarrow\; E_U(z)>u_2^{-1}\left[\left(\frac{p_1}{p_2}\right)u_1(E_H(z))\right]$$

та

$$p_2h_2(z)/\,p_1h_1(z)<1 \;\Leftrightarrow\; E_U(z)<u_2^{-1}\left[\left(\frac{p_1}{p_2}\right)u_1(E_H(z))\right],$$

де $u_i(\cdot)$ є деякими строго зростаючими функціями, що залежать від функцій глибини. Таким чином ми отримуємо результат, беручи $d(\cdot)=u_2^{-1}\left[\left(\frac{p_1}{p_2}\right)u_1(\cdot)\right]$. Теорему доведено.

Зазначимо, що в рамках еліптичних розподілів теорема 1 означає, що найбільш ефективна розділова крива між двома вибірками в просторі R^r еквівалентна кривій, що найбільш ефективно розділяє дві вибірки в Σ-схемі. Крім того, оскільки глибина даних характеризує розподіли по центральності, найбільш ефективна розділова функція в Σ-схемі також призведе до побудови досить ефективного класифікатора для нееліптичних розподілів. Однак, дана розділова функція в Σ-схемі може не бути зростаючою для нееліптичних розподів. Для вирішення даної проблеми, необхідно отримати найбільш ефективну функцію з Σ-схеми з осями для переставлених місцями E_U та E_H, а саме $\Sigma(U,H)$. Така розділова функція забезпечить побудову класифікатоа, що відрізняється від того, що отриманий з $\Sigma(H,U)$. В результаті, в якості рекомендованого класифікатора ми можемо вибрати той, що показує нижчий коефіцієнт помилкової класифікації. Зауважимо також, що аналогічну процедуру можна було б повторити зі зміненими E_H та E_U.

Таким чином, далі запропоновано підхід з використанням лише $\Sigma(H,U)$-схеми, хоча результати класифікації можуть бути значно покращені, якщо розглянути обидві Σ-схеми. Оскільки будь-яка гладка функція може бути ефективно апроксимована за допомогою многочлена відповідного ступеня, ми можемо обмежити себе до многочленів для знаходження функції $d(\cdot)$, що розділяє дві вибірки в Σ-схемі з мінімальним коефіцієнтом помилкової класифікації [6].

Слід зазначити, що відповідно до правила класифікації, розділовий многочлен повинен проходити через початок координат в Σ-схемі. Щоб переконатися в цьому, розглянемо елементи даних, що відповідають $(0,0)$ в Σ-схемі. Членство у вибірках цих елементів даних є невідомим, оскільки вони мають нульові значення глибини відносно обох вибірок. Тому, правило класифікації фіксуватиме ці елементи даних на розділову криву, що визначає їх приналежність до будь-якої вибірки. Це також слідує з доведення теореми 1, звідки слідує, що $d(\cdot)$ задовольняє $d(0)=0$.

Отже, далі розглянемо многочлени виду $d_{\mathrm{s}}(z)=\sum_{i=1}^{\Delta_0}s_iz^i$ зі ступенем Δ_0, що є наперед визначеним відомим цілим числом та вектором коефіцієнтів многочлена $\mathrm{s}=(s_1,...,s_{\Delta_0})\in\mathrm{R}^{\Delta_0}$. Це дозволить знайти многочлен, що розділяє дві вибірки в Σ-схемі з мінімальним коефіцієнтом помилкової класифікації.

Для знаходження оптимального s, що мінімізує загальний коефіцієнт помилкової класифікації для наперед заданого Δ_0, розглянемо такий алгоритм класифікації:

а) якщо $E_{U_m}(z)>d_{\mathrm{s}}(E_{H_n}(z))$, то $z\in U$,

б) якщо $E_{U_m}(z)<d_{\mathrm{s}}(E_{H_n}(z))$, то $z\in H$.

Далі позначимо через s_0 оптимальне значення s при умові, якщо воно існує та будемо вважати Σ-класифікатор у якості досліджуваного. Отже, відповідно до $E_{U_m}(z)=d_{\mathrm{s}}(E_{H_n}(z))$ в Σ-схемі, ми можемо побудувати многочлен, призначити елементи даних, що розташовані над кривою до U, призначити елементи даних, що розташовані під кривою до H, а потім розрахувати емпіричний коефіцієнт помилкової класифікації для $\forall$ заданого $\mathrm{s}\in\mathrm{R}^{\Delta_0}$, а саме:

$$\bar{\Psi}_M(\mathrm{s})=\frac{p_1}{n}\sum_{i=1}^{n}\Lambda_{\{E_{U_m}(Z_i)>d_{\mathrm{s}}(E_{H_n}(Z_i))\}}+ \\ +\frac{p_2}{m}\sum_{i=1}^{m}\Lambda_{\{E_{U_m}(X_i)<d_{\mathrm{s}}(E_{H_n}(X_i))\}}, \tag{4}$$

де $\Lambda_{\{S\}}$ є характеристичною функцією, яка приймає значення 1, якщо S є істинним та 0 в іншому випадку, а p_i - апріорними ймовірностями двох класів, $M=(n,m)$.

Таким чином, запропонований підхід полягає в оцінці оптимального s_0 по $\bar{\mathrm{s}}_M$, що

мінімізує емпіричний коефіцієнт помилкової класифікації $\bar{\Psi}_M(\mathrm{s})$ в (4). Отже, якщо $\bar{\mathrm{s}}_M = \arg\min\{\bar{\Psi}_M(\mathrm{s})\}$, Σ-класифікатор матиме таку структуру:

а) якщо $E_{U_m}(z) > d_{\bar{\mathrm{s}}_M}(E_{H_n}(z))$, то $z \in U$,

та (5)

б) якщо $E_{U_m}(z) < d_{\bar{\mathrm{s}}_M}(E_{H_n}(z))$, то $z \in H$.

Застосовуючи даний алгоритм до множини даних на рис. 2 та виконавши дії по реалізації Σ-класифікатора, ми отримуємо розділовий многочлен другого ступеня, що мінімізує загальний емпіричний коефіцієнт помилкової класифікації.

Алгоритм реалізації Σ-класифікатора

Крок 1. Для проведення мінімізації $\bar{\Psi}_M(\mathrm{s})$ зазначимо, що Σ-класифікатор вимагає знаходження ступеня многочлена Δ_0, що мінімізує емпіричний коефіцієнт помилкової класифікації $\bar{\Psi}_M(\mathrm{s})$ в (4). У загальному випадку пошук такого многочлена необхідно проводити по всіх многочленах Δ_0 ступеня, що проходять через початок координат. Однак, в лінійному випадку, тобто коли $\Delta_0 = 1$, розглядаються лише такі криві, що проходять через початок координат та на щонайменше одній з $n+m$ точок вибірки, оскільки всі прямі, що проходять між двома сусідніми прямими без опорних точок будуть показувати однаковий коефіцієнт помилкової класифікації. Таким чином, необхідно розглянути щонайбільше $n+m$ прямих, де найбільш ефективною розділовою прямою є та, що забезпечує мінімальний коефіцієнт помилкової класифікації. Зазначимо, що якщо є множина таких прямих, необхідно вибрати пряму з найменшим нахилом.

Те ж саме має місце для випадку, коли $\Delta_0 > 1$. Тобто, необхідно розглянути всі многочлени, що проходять через початок координат та Δ_0 з $n+m$ точок вибірки. Остаточним розділовим многочленом є такий, що забезпечує мінімальний коефіцієнт помилкової класифікації.

Зауважимо, що коли Δ_0 або $n+m$ є великими, обчислення оптимального многочлена може бути досить складним. Це є наслідком того, що цільова функція $\bar{\Psi}_M(\mathrm{s})$ є сумою багатьох характеристичних функцій, які не є всюди диференційованими. Тому, знаходження мінімуму $\bar{\Psi}_M(\mathrm{s})$ вимагає значних обчислювальних витрат. Для знаходження більш ефективного алгоритму для досліджуваної задачі мінімізації, ми використовуємо логістичну функцію $1/(1+e^{-\kappa z})$ для апроксимації характеристичної функції $\Lambda_{\{z>0\}}$ в $\bar{\Psi}_M(\mathrm{s})$. В результаті, мінімум можна знайти за допомогою відповідних чисельних методів на основі похідних. Зауважимо, що коли κ є великим, чисельний метод оптимізації для результуючої цільової функції може бути досить нестабільним. Даний вивід має місце незважаючи на те, що велике κ забезпечує більш ефективне наближення $\Lambda_{\{z>0\}}$. Крім того слід зазначити, що вибір κ також відіграє важливу роль. На підставі чисельних досліджень було встановлено, що результати оптимізації є стабільними, якщо $\kappa \in [100, 400]$, а функція глибини нормалізована з верхньою межею 1. У всіх практичних експериментах було вибрано $\kappa = 200$.

Оскільки функція може мати багато локальних мінімумів, початкове значення для s може вплинути на процедуру оптимізації при використанні чисельних методів для знаходження мінімуму вищенаведеного наближення до $\bar{\Psi}_M(\mathrm{s})$. Тому, ми пропонуємо новий алгоритм вибору початкового значення.

Найбільш ефективною оцінкою для оптимального s_0 є вектор коефіцієнтів s для многочлена, що мінімізує $\bar{\Psi}_M(\mathrm{s})$ серед усіх многочленів, які проходять через початок координат та Δ_0 точок вибірки в Σ-схемі. Тому, замість проходження через всі многочлени, ми випадковим чином вибираємо досить велике число многочленів з цієї множини та вибираємо такий многочлен, що мінімізує $\bar{\Psi}_M(\mathrm{s})$ з цієї підмножини многочленів, а потім використовуємо його вектор коефіцієнтів s, як початкове значення для s в алгоритмі чисельної оптимізації.

Крок 2. Для вибору Δ_0 ми припускаємо, що степінь многочлена Δ_0 є відомою. Однак, вибір оптимального Δ_0 є однією з найбільш важливих практичних задач. Як і у випадку поліноміальної регресії, існує відповідне відношення між прогнозованим зсувом та прогнозованою дисперсією у виборі Δ_0. Дане прогнозування означає прогнозування членства для майбутніх елементів даних на основі виконання Σ-класифікатора. Невелике Δ_0 призведе до невеликої прогнозованої дисперсії, але великого прогнозуваного зсуву, а велике Δ_0 - до невеликого прогнозуваного зсуву, але великої

прогнозованої дисперсії. Для знаходження рішення даної проблеми необхідно використовувати метод крос-перевірки для вибору Δ_0.

Крок 3. При виборі відповідної функції глибини варто відзначити, що різні функції глибини відповідають різним характеристикам вихідного розподілу. Тому, якщо різні функції глибини використовуються для побудови Σ-схеми, Σ-класифікатор може демонструвати різну поведінку. При наявності апріорної інформації щодо розподілу, можна використовувати метод крос-перевірки для вибору функції глибину, що забезпечує найбільш низький коефіцієнт помилкової класифікації.

Отже, як показано на рис. 1-6, розділова крива може бути будь-якої форми, залежно від структури даних, коли многочлен відображається у вихідний простір. Даний результат має місце незалежно від того, що ми завжди зосереджені на розділовому многочлені в Σ-схемі.

Висновки. Підводячи підсумки викладеного матеріалу, слід зазначити очевидні переваги запропонованого методу: найбільш ефективна розділова крива в Σ-схемі автоматично визначається по ймовірнісній геометрії даних, тому побудований Σ-класифікатор є повністю непараметричним; оскільки глибинна трансформація володіє ефектом нормалізації на даних, Σ-класифікатор, на відміну від більшості методів класифікації не вимагає проведення оцінки таких параметрів, як середнє значення та масштаб; результат класифікації може бути інформативно представлений на двовимірній Σ-схемі, що є значно простішою задачею, ніж відстеження результату класифікації у вихідному просторі вибірки великої розмірності.

Список використаних джерел

1. *Zuo Y.J.* Projection-based depth functions and associated medians / Y.J. Zuo // The Annals of Statistics. – 2003. – 31. – P. 1463-1484.
2. *Cuesta-Albertos J.A.* The random Tukey depth / J.A. Cuesta-Albertos, A. Nieto-Reyes // Computational Statistics & Data Analysis. – 2008. – 52. – P. 4980-4987.
3. *Анисимов А.В.* Исследование асимптотических свойств непараметрических классификаторов на основе функций глубины / А.В. Анисимов, А.А. Галкин // Проблемы управления и информатики. – 2015. – №4. – С. 147-155.
4. *Li J.* New nonparametric tests of multivariate locations and scales using data depth / J. Li, R.Y. Zuo // Statistical Science. – 2004. – 19. – P. 687-694.
5. *Lange T.* Fast nonparametric classification based on data depth / T. Lange, K. Mosler, P. Mozharovskyi // *Statist. Papers*. – 2014. – 55. – P. 53–67.
6. *Галкін О.А.* Застосування функцій згладжування для апроксимації оцінок непараметричних класифікаторів / О.А. Галкін // Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту (ISDMCI 2015): Матеріали міжнародної наукової конференції. – Херсон, 2015. – С. 266-267.

References

1. ZUO, Y.J. (2003) Projection-based depth functions and associated medians. *The Annals of Statistics*. 31. p. 1463-1484.
2. CUESTA-ALBERTOS, J.A. and NIETO-REYES, A. (2008) The random Tukey depth. *Computational Statistics & Data Analysis*. 52. p. 4980-4987.
3. ANISIMOV, A. and GALKIN, A. (2015) The research of the asymptotic properties of nonparametric classifiers based on depth functions. *Journal of Automation and Information sciences*. 4. p. 147-155.
4. VARDI, Y. & ZHANG, C.H. (2000) The multivariate on L_1-median and associated data depth. *Proceedings of the National Academy of Sciences (USA)*. 97. p.1423-1426.
5. LANGE, T., MOSLER K., MOZHAROVSKYI P. (2014) Fast nonparametric classification based on data depth. *Statist. Papers*. 55. p. 53-67.
6. GALKIN, O. (2015) Application of smoothing functions to approximate estimates of nonparametric classifiers. *Intellectual Systems for Decision Making and Problems of Computational Intelligence*: Conference Proceedings. – KNTU, 2015. p. 266-267.

Надійшла до редколегії 27.08.15