РОЗРОБКА МЕТОДУ КЛАСИФІКАЦІЇ ДЛЯ СИСТЕМ, ЩО НАВЧАЮТЬСЯ З ВЧИТЕЛЕМ

Єськов Сергій Сергійович[1]

У роботі пропонується уявлення завдання класифікації у вигляді задачі визначення належності точки замкнутій геометричній області. Розроблено метод класифікації для систем що навчаються з учителем – метод крайніх точок.

1 Постанова задачі

Розглянемо випадок, коли класифікація об'єкту проводиться за двома параметрами x1 і x2, або, іншими словами, в деякому просторі ознак X = {x1, x2}. Візуально таке простір може бути представлено у вигляді декартової площини, осями якої виступають елементи множини X.

Окрема точка може розглядатися як відображення деякого об'єкту в даному просторі ознак. Якщо вибраний об'єкт достовірно відноситься до деякого класу об'єктів, то точка що характеризує його в просторі ознак вважається еталонною. Домовимося, що одна й та ж точка може бути еталонної тільки для одного класу.

У загальному випадку на заданому просторі ознак може бути визначено безліч класів. Процес формування множини класів назвемо кластеризацією [1-3] заданого простору ознак. Таким чином, кожному класу повинен відповідати певний кластер (геометрична область) простору ознак.

На рис. 1 показано графічне представлення простору ознак X у вигляді геометричної площини.

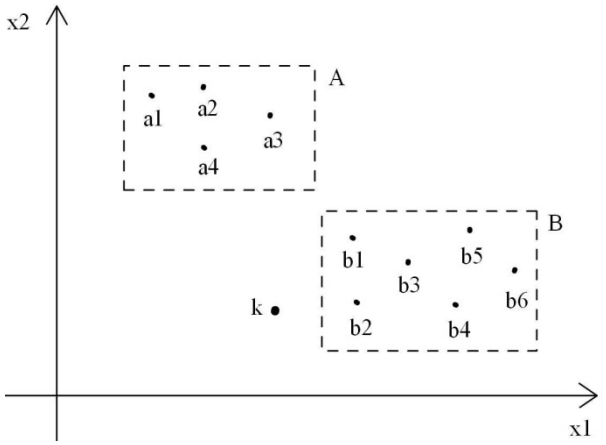

Рисунок 1. Уявлення простору ознаку вигляді геометричної площини

На рис. 1: a1–a4 – еталонні точки класу A, b1–b6 – еталонні точки класу B, k – точка, відповідна до об'єкта що класифікуються. Пунктиром показаний приклад формування кластерів, відповідно до класів A і B.

Формально задача кластеризації в системах що навчаються із вчителем (або у іншій літературі – задача навчаються за прецедентами) полягає в тому, щоб за допомогою заданих еталонних точок побудувати вирішальне правило (decision function), яке наближало б цільову залежність, причому не тільки на еталонних об'єктах, а і на всьому просторі ознак.

[1] аспірант кафедри систем штучного інтелекту Донецького національного технічного університету

Другим етапом вирішення задачі класифікації в системах що навчаються із вчителем є безпосереднє віднесення вибраної зовнішнім фактором точки з простору ознак до одного з класів за допомогою отриманого на попередньому етапі вирішального правила.

2 Запропонований метод

Розглянемо вибірку точок як «скелет», а клас, що необхідно побудувати як «тіло». Як і для будь-якого фізичного тіла, для математичного тіла можна визначити форму, а також щільність. Щільність для тіла буде дискретною – кількість точок, що припадає на деяку площу.

Звичайно, якщо судити строго, то будь-яка точка простору ознак, яка не належить до вхідної вибірки має нульову щільність, а точки що належать до вхідної вибірки матимуть абсолютну щільність, але такий підхід до розглядання проблеми не є правильним з точки зору необхідності класифікувати невідомі точки простору ознак. Тому, для розуміння суті методу, домовимось що ми маємо право самостійно визначати щільність будь-якої точки простору ознак.

Припустимо, що ми вирішили завдання «чи належить точка тілу» за допомогою функції Access (параметри: номер класу, об'єкт). Тепер, якщо є n класів, то для того щоб визначити до якого класу належить новий об'єкт, достатньо перевірити чи належить вхідний об'єкт кожному з тіл (рис. 2).

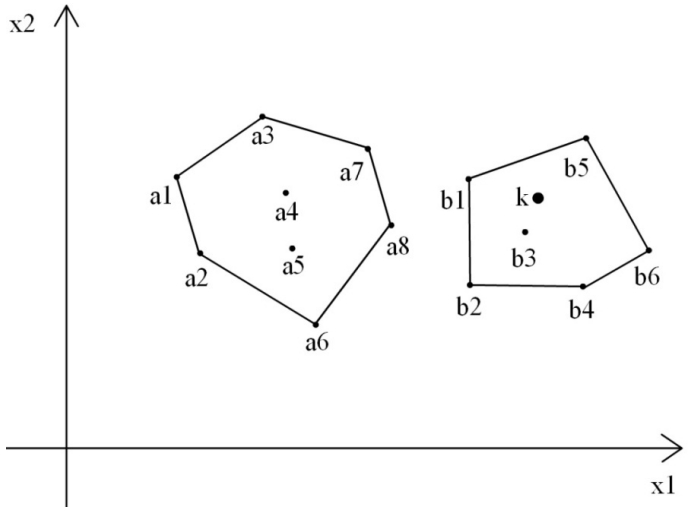

Рисунок 2. Приклад визначення належності об'єкта до тіла

З рисунку видно, що точка, що класифікується, належить до тіла, сформованого на основі точок класу B, тобто за допомогою запропонованого алгоритму визначено, що точка k належить до класу B. Через те, що порівняння досить проводити, використовуючи тільки крайні опуклі точки вибірки, назвемо цей метод – метод крайніх точок.

Тепер розглянемо випадок, коли сформовані тіла будуть перетинатися. Назвемо область, що одночасно належить до декількох тіл областю конфлікту. Очевидно, що при попаданні об'єкта що класифікується в область конфлікту перевагу про приналежність слід віддавати класу, щільність тіла що сформовано на основі його еталонних точок вище в цій точці.

Можна довго міркувати про найкращий, «самий оптимальний» крок дискретизації для визначення щільності, але об'єктивно він буде змінюватися від сфери застосування методу. Також можна побудувати функцію щільності, яка показує щільність «тіла» на кожній ділянці простору, але це далеко не тривіальна задача [2] і її розгляд планується проводити в ході подальших досліджень.

На сьогодні, для досягнення універсальності у рамках методу крайніх точок, пропонується вчинити наступним чином. Знайти середню щільність кожного з тіл, що входять в область конфлікту і віднести об'єкт і-му класу з імовірністю пропорційно його щільності щодо тіл інших класів, що входять в цю область.

Для двох тіл які перетинаються, досить вчинити наступним чином. Порахувати (за вже отриманою функцією Access) скільки точок першого класу входить у друге тіло (k12), і скільки точок другого класу входить в перше тіло (k21). І віднести об'єкт до першого класу з імовірністю p1 рівної відношенні k12 до суми k12 і k21. Віднесення до другого класу буде мати ймовірність (використовуючи формулу повної ймовірності) одиниця відняти p1 (рис. 3).

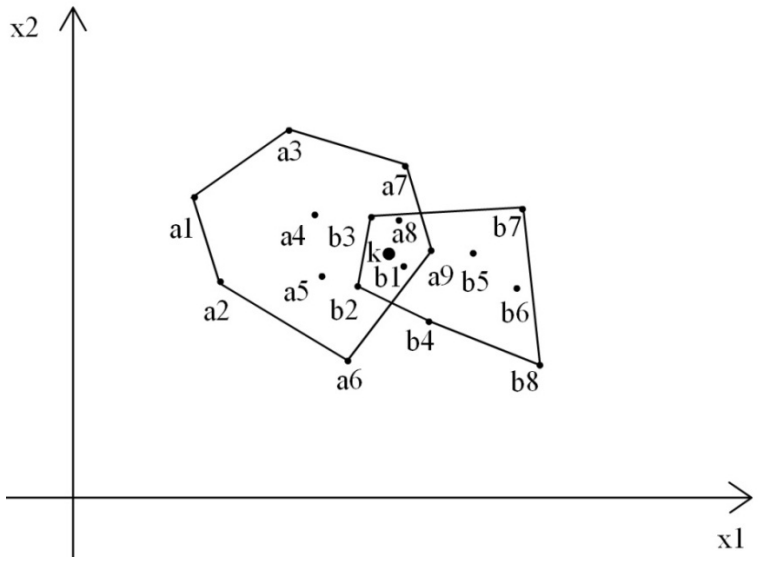

Рисунок 3. Вирішення конфліктних ситуацій, засноване на визначенні середньої щільності тіл в області конфлікту

У розглянутому прикладі (рис. 3) об'єкт повинен бути віднесений до класу А з імовірністю дві п'ятих. До класу В новий об'єкт k буде віднесений з ймовірністю три п'ятих. Для більшої кількості вхідних тіл, можна робити по запропонованій аналогії, з тією різницею, що вже n чисел будуть формувати повну ймовірність.

3 Оцінка ефективності запропонованого методу

Щоб показати ефективність запропонованого методу крайніх точок, проведемо його порівняльний аналіз з відомим методом потенційних функцій на одних й тих же наборах вхідних даних за допомогою оцінки функціоналу якості. Проведемо порівняння на трьох наборах вхідних даних.

Отже, для простоти візьмемо дві ознаки і дви класи: А і В. Для наглядності позначимо еталонні точки класу А хрестиками, а класу В – точками. Перший набір

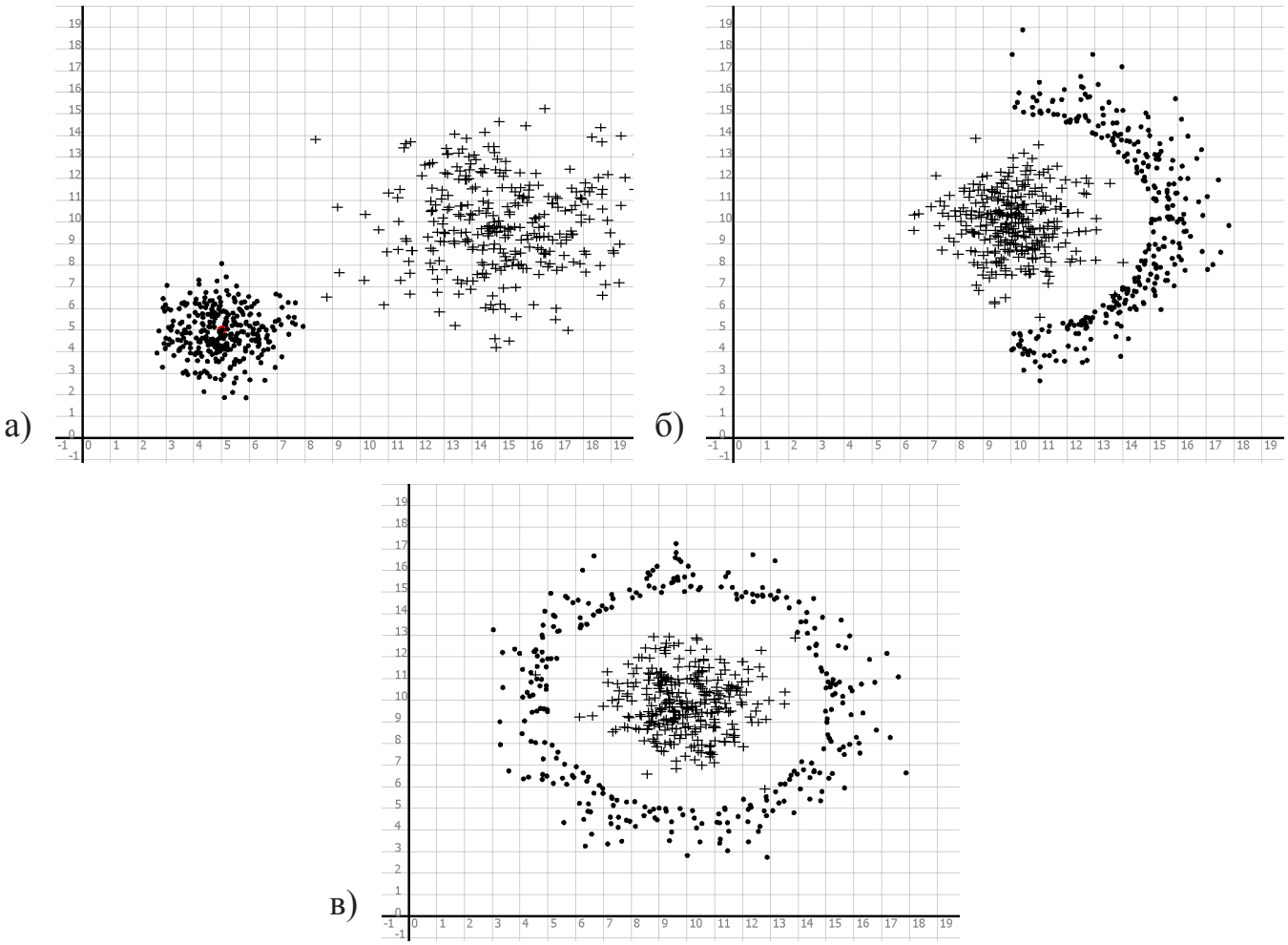

Рисунок 4. Набори еталонних вхідних даних: а) набір 1; б) набір 2; в) набір 3.

вхідних даних має вигляд, що представлено на рис. 4 (а). Будемо називати його набір 1. Другий набір вхідних даних для порівняльної демонстрації роботи метода крайніх точок показано на рис.4 (б). Цей набір будемо називати набір 2. Третій набір вхідних даних, який було сгенеровано за допомогою псевдовипадкового алгоритму, реалізованому у сучасному комп'ютері, показано на рис. 4 (в). Будемо називати його набір 3.

Порівняльне оцінювання

Для того, щоб мати можливість і візуально оцінити роботу реалізованих алгоритмів, напишемо програми, що матимуть візуальне відображення вирішальних правил, що будуються алгоритмами на основі методів потенційних функцій і крайніх точок.

На рис. 5 наведено вирішальні правила (позначено кривими), що були отримані для вхідних даних при навчанні на наборі 1 за допомогою методів що розглядаються: за допомогою методу потенційних функцій (рис. 5, а) та за допомогою методу крайніх точок (рис. 5, б).

На рис. 6 наведено вирішальні правила (позначено кривими), що були отримані для вхідних даних при навчанні на наборі 2 за допомогою методів що розглядаються: за допомогою методу потенційних функцій (рис. 6, а) та за допомогою методу крайніх точок (рис. 6, б). Слід зазначити, що дані навмисно підібрані таким чином, що щільність класу B (точки) в області конфлікту дорівнює нулю.

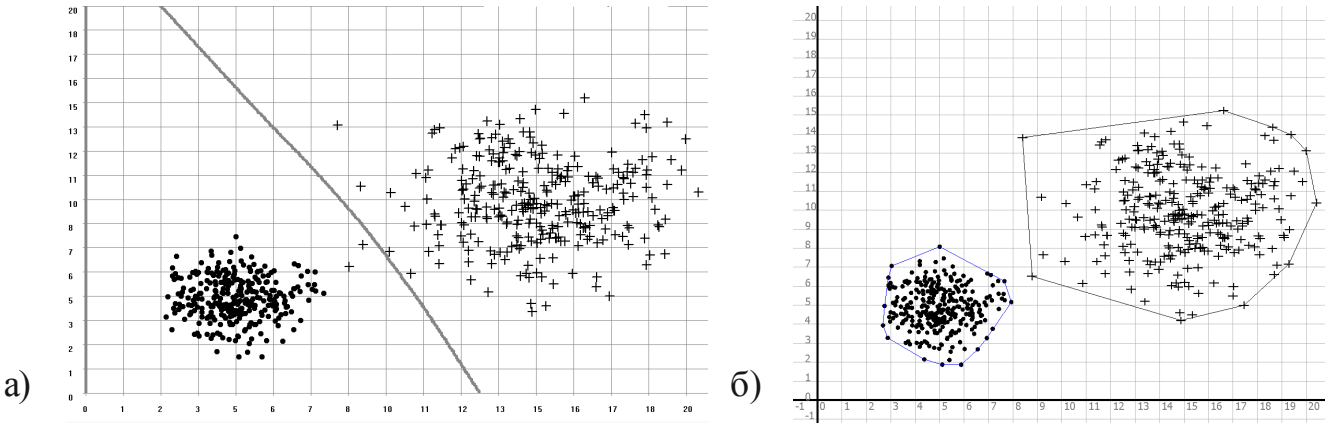

Рисунок 5. Вирішальні правила для набору даних 1: а) отримано за допомогою методу потенційних функцій; б) за допомогою методу потенційних функцій.

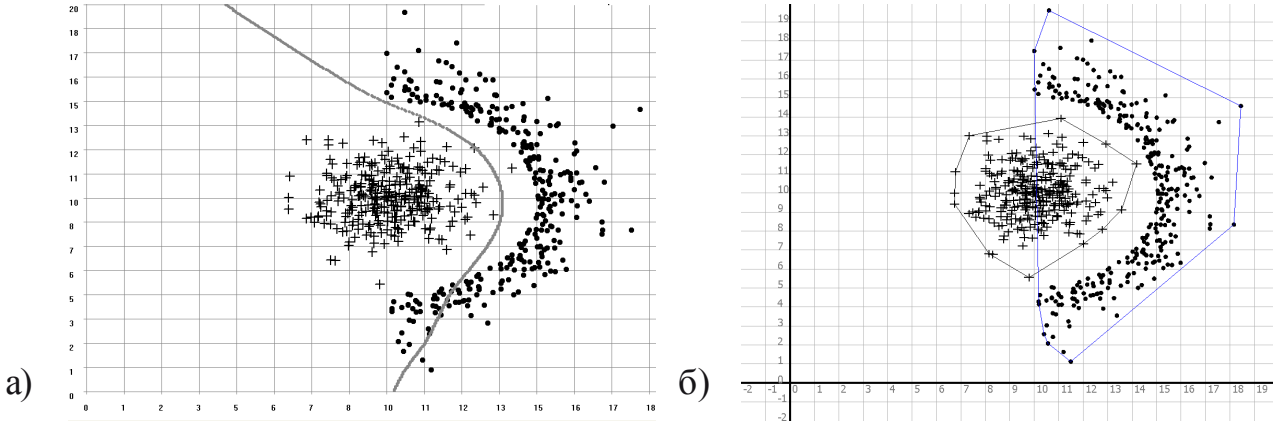

Рисунок 6. Вирішальні правила для набору даних 2: а) отримано за допомогою методу потенційних функцій; б) за допомогою методу потенційних функцій.

На рис. 7 наведено вирішальні правила (позначено кривими), що були отримані для вхідних даних при навчанні на наборі 3 за допомогою методів що розглядаються: за допомогою методу потенційних функцій (рис. 7, а) та за допомогою методу крайніх точок (рис. 7, б). Слід зазначити, що дані навмисно підібрані таким чином, що щільність класу B (точки) в області конфлікту дорівнює нулю.

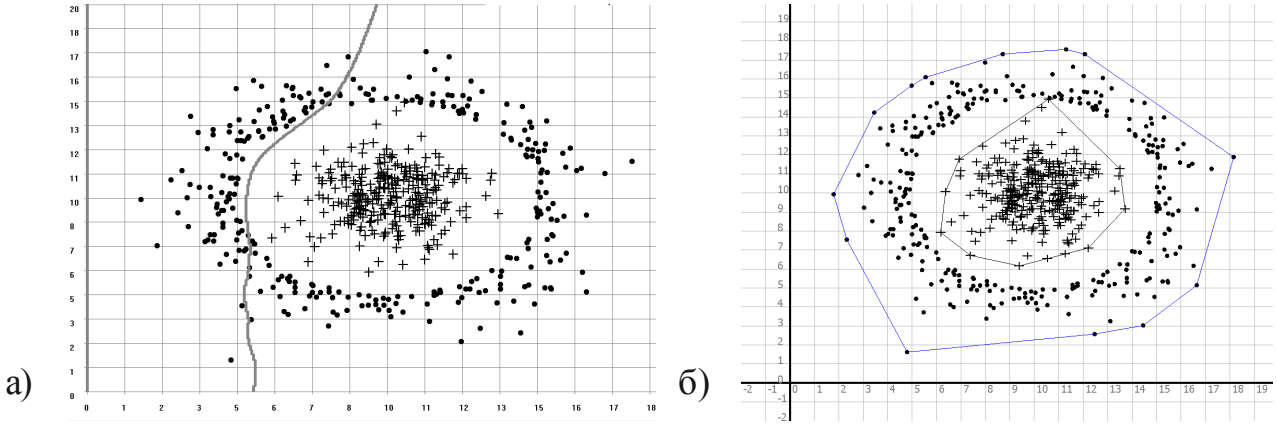

Рисунок 7. Вирішальні правила для набору даних 3: а) отримано за допомогою методу потенційних функцій; б) за допомогою методу потенційних функцій.

Таблиця 1. Результати оцінювання функціоналу якості методів

№ набору	Клас	Метод крайніх точок	Метод потенційних функцій
Набір 1	A	0	2/600
Набір 1	B	0	0
Набір 2	A	0	1/600
Набір 2	B	0	31/600
Набір 3	A	0	0
Набір 3	B	0	236/600

Визначимо функціонали якості методів що розглядаються – методу крайніх точок і методу потенційних функцій на наведених вище наборах еталонних точок. Порівняльні результати оцінювання функціоналу якості методів потенційних функцій і крайніх точок наведено у табл. 1.

З табл. 1 видно, що для деяких випадків метод крайніх точок більш ефективний з точки зору критерію функціоналу якості. Найбільший приріст ефективності методу крайніх точок, порівняно з методом потенційних функцій, відбувається у третьому розглянутому випадку класифікації (набір 3).