COMMON APPROACH TO SOLUTION OF SOME PROBLEMS OF DISCRIMINANT AND CLASTER ANALYSIS

O.V.Babak[1], *A. S. Gasanov*[1], *T.A. Babayev*[2], *J. V. Backlan*[3]

1. International Research and Training Centre of Information Technologies and Systems,
40, prospect Academica Glushkova, 03680, Kiev, Ukraine.
Ph.: (38044) 266-41-87 fax: (38044) 266-15-70,
E-mail: vig@tel.dlab.kiev.ua

2. Baku Scientific and Training Centre
11, I. Idayat-Zade, 370154, Baku, Azerbaijan,
Ph.: (99412) 66-38-81, fax: (99412) 66-39-95
E-mail: tbabayev@azeri.com

3. Education-Scientific complex "Institute for Applied System Analysis",
37, Peremoghy ave, 01056, Kiev, Ukraine.
Ph: (38044) 243-59-20, fax: (38044) 243-59-20
E-mail: kopernik2000@mail.ru

Abstract In this work a common approach is proposed, that simplifies solution of some problems of the discriminant and claster analysis. The approach is based on restoration of an objectively existing relation, hidden in input data. The relation can be synthesized as a multiplicative function in accordance with obtained information on direction of gradient components of linear model of object classification. The revealed relation projects input information onto an axis and makes it possible to get an enough simple decision rule of dividing objects onto classes, or to establish, that in a given class of uniform objects there exist subclasses of objects, similar in a sense.

ВСТУП

Статистичний підхід щодо розпізнавання образів, методами якого є дискримінантний та кластерний аналіз, займає значне місце у різних сферах науки та техніки. Статистичні моделі алгоритмів розпізнавання засновані на використанні апарату математичної статистики. Вони застосовуються загалом у тих випадках, коли відомі або можуть бути просто визначені ймовірні характеристики класів, що, наприклад, відповідають функції розподілу.

З математичної точки зору засоби розв'язання задач у вказаних двох галузях розпізнавання образів мають істотні відмінності. Тому створення єдиного підходу до їх розв'язання навіть щодо окремих випадків, викликає певний практичний та теоретичний інтерес. Така можливість виникає, якщо у вихідній інформації об'єктивно існують приховані залежності одного характеру, виявлення яких дозволяє здобути бажаний результат. Наслідком цього є істотне спрощення розв'язання задачі, що одержується за допомогою одного і того ж алгоритмічного та програмового забезпечення.

Методологія розпізнавання при цьому виглядає приблизно так. Допустимо, наприклад, що є дані, які отримано в результаті фізичного або імітаційного експерименту. Ці дані в деякому дуже обмеженому змісті, що характеризують досліджуваний об'єкт або явище, необхідно спробувати звести разом для того, щоб встановити, які закономірності відбуваються в наявному матеріалі. Для цього висувається деяка робоча гіпотеза, що породжується евристикою, якій придається математичний образ – потім покладений в основу моделі розпізнавання тієї чи іншої задачі дискримінантного або кластерного аналізу.

1. ЕЛЕМЕНТИ ТЕОРІЇ ТА ОСОБЛИВОСТІ АЛГОРИТМІВ

Перш ніж перейти до особливостей побудови єдиного підходу до розв'язування деяких завдань дискримінантного та кластерного аналізу відзначимо:

При дискримінантному аналізі, у разі розпізнавання, наприклад, двох класів об'єктів, за наявними даними або відновлюються, або виявляються дискримінантні залежності y^* від n – числа ознак (змінних) x. Доречно відзначити, що тут відновлення та виявлення – це принципово різні задачі. Якщо відновлення – це вибирання за заданим

набором ознак її структури з певного класу залежностей [1], то при виявленні використовується не весь набір ознак, а лише кращі з них або їх комбінації. Особливо важливим є виявлення комбінацій ознак, тому що вже давно спостережено, що часто "розрізнювання" приховане не у самих ознаках, а в їх сполученнях [2]. Таким чином, здійснюється перехід до нового, лінійного за параметрами й нелінійному за змінними простору, вимірність якого завдяки редуціювання вхідної інформації можна істотно знизити (аж до одновимірного).

Під кластеризацією об'єктів зазвичай вважається виявлення у певній матриці даних звичайного, об'єктивно існуючого порядку, завдяки чому стає можливим виділення кластерів – деяких підмножин множин, що досліджуються.

У праці [2] наведено основні різновиди кластерного аналізу, які зустрічаються найчастіше. Найбільш розповсюджений метод виявлення кластерів пов'язаний із завданням на багатьох спостереженнях належної метрики, наприклад, евклідової та вирахуванням за допомогою її відстаней між усіма точками вказаної множини.

Доречно відмітити, що особливістю процесу виділення кластерів є те, що він може розглядатися й як геометричне завдання про виявлення у певному просторі "щільних" накопичень точок. Вказана особливість обумовлює можливість розвинути зовсім специфічні методи її вирішення. Один з таких методів, що дозволяє не лише спростити рішення, але й візуально "побачити" підсумки кластеризації розглядається у поданій роботі. Ідея запропонованого методу базується на редукції інформації, в результаті чого вихідну, як правило, складну багатомірну задачу перетворюють у більш просту, наприклад, трьохмірну, двомірну або одномірну. При цьому існують два принципово різних напрямки скорочення розмірності простору наявних даних.

В основі першого напрямку лежить стиснення набору вихідних факторів (незалежних змінних) максимум до трьох за рахунок відсіювання решти, якщо вдається встановити, що останні практично не впливають на виділення кластерів. Вказана процедура носить назву цілеспрямованого проектування й, зазвичай, опирається на метод головних компонент [5], технічна реалізація якого, відрізняючись складністю, і є можливою лише тоді, коли фактори корельовані між собою. Другий напрямок, що розвивається нами, базується на пошукові деякої компактної функціональної залежності, що об'єднує усі значні і некорельовані (корельовані) між собою фактори, число яких може бути більше трьох, у вигляді так званої узагальненої змінної (поняття введено у працях О. Г. Івахненко [4]), завдяки чому стає можливим перетворення вихідної багатомірної задачі в зручну одномірну.

Звичайно, відображення багатомірного простору на одномірне теоретично не може зменшити мінімально досяжний рівень помилки виявлення кластерів. Однак, з успіхом можна вважати, що якщо дані, які породжуються певним джерелом, підпорядковуються нормальному закону розподілу, то ці жертви будуть мінімальними [7]. Відрізняючись, як буде вказано нижче, простотою технічної реалізації, другий напрямок є плідним для виділення кластерів, практично некорельованих між собою значущих факторів.

Оскільки при кластерному аналізі об'єкти поділяються на групи за їх подібністю у певному понятті, що можна інтерпретувати й так: дані, що належать до певного класу однорідних об'єктів поділяються на підкласи. Саме таке розуміння розбиття класу подібних об'єктів на підкласи разом з проектуванням вхідної інформації на одновимірний простір створює базу єдиного підходу до розв'язання деяких задач дискримінантного та кластерного аналізу.

Виникає питання: як знайти оптимальні у певному сенсі сполучення ознак, не вдаючись до комбінаторного перебору, що потребує великого обсягу обчислень, не гарантуючи вірності вибору того чи іншого критерію. Відповідь на це питання базується на таких положеннях:

1. Більшість явищ довкілля і відповідно їх моделей нелінійні. При цьому найважливіша інформація про характер структури такої моделі є в лінійній за параметрами й змінними моделі явища, що може бути представлена в виді (1) в якій є інформація про складові градієнта функції відгуку:

$$\hat{y} = \sum_{i=1}^{n} a_i x_i , \qquad (1)$$

де $\hat{y}$ - прогнозоване значення функції відгуку y, a_i - оцінки коефіцієнтів, x- число ознак (змінних).

2. Більшість моделей класичної фізики є мультиплікативні і можуть бути представлені у виді (2):

$$y = k\prod_{i=1}^{n} x_i^{p_i} , \qquad (2)$$

де $k>0$ – коефіцієнт (раціональне число), p_i – показник степені позитивне (негативне) раціональне число.

3. Розв'язання деяких задач дискримінантного та кластерного аналізу зводиться до виявлення прихованих у вихідних даних закономірностей, що об'єктивно існують, у вигляді мультиплікативних моделей типу (2).

В роботі [3] показано: якщо у лінійній моделі (1) коефіцієнти a є оцінки напрямку складових градієнту функції y, то можна синтезувати узагальнену змінну.

$$v = \prod_{i=1}^{n} x_i^{p_i}, \qquad (3)$$

де p_i дорівнюють ±1, залежно від знаку відповідної компоненти градієнта.

Функції (1) і (3) у певному змісті подібні. Подібність їх є у тому, що із зменшенням (зростання) будь-якої змінної відповідно змінюватимуться (зростатимуть або зменшуватимуться) значення y та v. Саме це, зберігаючи фізичні особливості явища при переході від його лінійної моделі до нелінійної, забезпечує вибір сполучень ознак, не удаючись до комбінаторного перебору. Цей результат закладено в основу єдиного підходу до розв'язання деяких задач дискримінантного та кластерного аналізу. Схема алгоритмів їх рішення включає такі етапи:

1. За даними навчальної вибірки, де для позначення належності об'єктів до двох класів або одного, використовується, відповідно, значення ± 1 або +1, відновляється лінійна функція (1) у результаті чого знаходяться оцінки коефіцієнтів a, що є оцінками напрямку складових її градієнту.

2. Знаючи напрямок складових градієнту функції відгуку, синтезується узагальнена змінна v (3).

3. Значення v_j, $j = \overline{1,l}$, де l – довжина навчальної вибірки, якщо це необхідно – масштабуються і потім розміщуються на одній координатній осі.

4. Характер розміщення v_j аналізується в залежності від роду задачі.

Примітка. За обчислення v можуть бути використані, як нормовані, так і ненормовані, значення незалежних змінних. При цій ситуації, значення, що включають їхнє нульове не розглядаються, оскільки в даному випадку класифікація стає неможливою.

Далі розглядатимуться особливості алгоритмів розв'язання таких задач.

2. ЗАДАЧА ДИСКРИМІНАНТНОГО АНАЛІЗУ

Нехай задано навчаючу вибірку довжиною l

$$\{x_{ij}, y_j^*\}, \quad i = \overline{1,n}, \quad j = \overline{1,l}, \qquad (4)$$

причому $\forall x_{ij} \neq 0$, де y_j^* приймає значення ±1 в залежності від належності класу A або B елементів вибірки за якою відбудовується лінійна дискримінантна функція y^*. При цьому стають відомими оцінки напрямку складових її градієнту і можливий синтез v [3]. Оскільки значення v_j, $j = \overline{1,l}$, розташовані на одній координатній осі, нескладно встановити межі належності їх до класів A і B та одержати у даному простому випадку розв'язальне правило.

При побудові його необхідно врахувати важливе поняття дикримінантного аналізу – "відмова від розпізнавання" [2]. Сутність його полягає в тому, що в сумнівних випадках (існує область, що включає об'єкти як класу A, так і класу B), або при поганій згоді (існує область в якій об'єкти розташовані далеко від середніх значень класів) класифікація неможлива. Отже, при необхідності в таких випадках вона повинна здійснюватися іншими засобами, зокрема, наприклад, виходячи з евристичних розумінь. Природно, використовуючи відмову від класифікації, можна різко підвищити її якість при зменшенні числа правильних класифікаційних рішень. Таким чином, з урахуванням сумнівних випадків (область з межами v_a і v_b, $v_a < v_b$, де a, b деякі числа з вибірки $\overline{1,l}$) і при поганій згоді (область з межами $-\infty$, v_c і v_d, $+\infty$ де $v_c < v_a$ і $v_b < v_d$ і c, d теж деякі числа з вибірки $\overline{1,l}$) вирішальне правило матиме вигляд:

$$v_c < v < v_a, \quad v \in \text{класу } A,$$
$$v_b < v < v_d, \quad v \in \text{класу } B.$$

При цьому оцінка вирішальних правил повинна бути виконана на основі оцінювання класифікаційної помилки. Найпростіша можливість полягає у застосуванні його до об'єктів навчання і підрахунку числа невірних рішень. Інша можливість більш суворої оцінки пов'язана з поділом множини об'єктів на навчальне і контрольне. При цьому навчальна множина використовується при побудові вирішального правила, а контрольна - для оцінки його якості. Найбільш сильний результат може бути отриманий із застосуванням методу контролю, що стежить і який передбачає вилучення з навчальної множини лише одного об'єкта, який потім використовується для оцінки якості вирішального правила [1].

3. ЗАДАЧА КЛАСТЕРНОГО АНАЛІЗУ

Нехай задано навчаючу вибірку

$$\{x_{ij}\}, \quad i = \overline{1,n}, \quad j = \overline{1,l},$$

при чому $\forall x_{ij} \neq 0$, що репрезентує клас однорідних об'єктів. Аналогічно до вибірки дискримінантного аналізу її можна записати як (4) з

тією різницею, що усі v прийматимуть значення +1 (тільки один клас). Далі, як і раніше, відбудовується лінійна функція (1) і синтезується v. У зв'язку з тим, що v_j розміщені на одній осі, легко побачити чи є на ній "густі" скупчення точок (підкласи подібних об'єктів або кластери). Необхідно відмітити, що проблема остаточного виявлення кластерів зовсім не проста, оскільки оцінка міри подібності вельми суб'єктивна.

Пояснюється це тим, що в кластерному аналізі не існує однозначного кількісного критерію подібного помилці класифікації в дискримінантному аналізі. Такий критерій не можна сформувати, оскільки в різноманітних прикладних задачах різноманітними можуть бути і цілі аналізу, спрямовані на одержання нової інформації. У аналізованому випадку усе ж доцільно ввести ряд понять, що дозволяють якоюсь мірою порівняти і сприяти виділенню кластерів. До таких понять ставляться:

по-перше, довжина кластера L визначається наступним способом:

$$v_{j\min} = \min\{v_j; \overline{1,k}\},$$
$$v_{j\max} = \max\{v_j; \overline{1,k}\},$$
$$L = (v_{j\max} - v_{j\min}),$$

де k - число кластерів;
по-друге, щільність кластера:

$$d_j = \frac{m_j}{L_j}, \quad j \in \overline{1,k},$$

де m - число елементів j-го кластера.

З огляду на те, що кластер є утворенням, де внутрішні структурні зв'язки в декілька разів більші порівняно з зовнішніми зв'язками одного кластера з іншим, для виділення кластерів також корисно ввести поняття міри близькості елементів кожного кластера у вигляді

$$\forall(v_{i+1} - v_i) \le \varepsilon,$$

де i - елемент кластера, ε - деяке чисельне значення, що може бути задане відповідно до характеру тієї чи іншої задачі кластеризації.

ВИСНОВКИ

У відповідності з викладеним можна зробити наступні висновки.

1. Запропоновано єдиний підхід до вирішення деяких задач дискримінантного та кластерного аналізу, особливостями яких є наявність у вхідних даних прихованих об'єктивно існуючих залежностей, які можна відобразити у вигляді мультиплікативних функцій [3].

2. В основі єдиного підходу до вирішення деяких задач дискрімінантного і кластерного аналізу лежить у першому випадку розбиття багатьох різного роду об'єктів на класи, а у другому випадку розбиття багатьох подібних об'єктів (класу об'єктів) на підкласи. Ця обставина дозволяє розглядати вищевказані задачі з однієї і тієї ж математичної точки зору, що спирається на побудову лінійних моделей, які утримують інформацію про спрямування складових градієнта лінійних функцій.

3. Наявність інформації про спрямування складових градієнта лінійної моделі при рішенні деяких задач дискримінантного і кластерного аналізу дозволяє здійснити синтез узагальненої змінної у вигляді мультиплікаційної функції, зберігаючи фізичні особливості явища.

4. Значення мультиплікаційної функції у задачах дискримінантного і кластерного аналізу можуть розташовуватися на одній координатній осі, що суттєво спрощує процедуру побудови алгоритмів класифікації та кластеризації і робить її доступною для широкого кола користувачів.

ЛІТЕРАТУРА

1. Алгоритмы и программы восстановления зависимостей (под ред. В. Н. Вапника). - М.: Наука. - 1984. - 816 с.
2. Распознавание образов: состояние и перспективы. Пер. с англ. (К. Верхаген, Р. Дейн и др.). - М.: Радио и связь. - 1985. - 104 с.
3. Бабак О. В. Об одном подходе к решению задачи восстановления зависимости в классе кусочно-линейных функций // Проблемы управления и информатики. - 1995. - N6. - С. 134-141.
4. Ивахненко А. Г., Мюллер Й. А. Самоорганизация прогнозирующих моделей. - К.: Техника, 1985: Берлин: ФЕБ Ферлаг Техник, 1984. - 233 с.
5. Прикладная статистика: Классификация и снижение размерности: Справ. Изд./С.А.Айвазян и др., под ред. С. А. Айвазяна. - М.: Финансы и статистика, - 1989. – 426 с.
6. Распознование, классификация, прогноз/ Отв. ред. Ю. И. Журавлев. – М.: Наука, 1989. – 280 с.
7. Дуда Р., Харт П. Распознавание образов и сцен. – М.: Мир, 1976. – 511 с.8