Журнал обчислювальної та прикладної математики **2014, № 3(117)** Journal of Computational & Applied Mathematics

УДК 519.71

КВАДРАТИЧНЫЙ РЕЛЯЦИОННЫЙ ДИСКРИМИНАНТНЫЙ АНАЛИЗ МАТРИЧНЫХ ДАННЫХ И УСТРАНЕНИЕ НЕОПРЕДЕЛЕННОСТИ

Д. А. Клюшин, М. В. Присяжная, Е. С. Бондарь

Резюме. Предложен новый метод однозначного квадратичного реляционного дискриминантного анализа матричных данных с помощью эллипсов Петунина. Для перехода из пространства признаков в пространство близости используется p-статистика. Устранение неопределенности осуществляется на основе ранжирования многомерных выборок с помощью статистической глубины. Продемонстрирована практическая ценность предложенного подхода.

1. Введение

В статье предлагается новый метод однозначной классификации многомерных выборок, представляющий собой результат синтеза идей, лежащих в основе двух направлений современной теории дискриминантного анализа — квадратичного и реляционного, или беспризнакового. Предметом исследования являются данные, образующие не вектор, как в классической схеме, а матрицу. Такие ситуации часто встречаются в цитометрических исследованиях, когда у пациентов измеряется ряд показателей множества клеток. В этом случае строки матрицы соответствуют отдельным клеткам, а столбцы — отдельным показателям.

Наиболее распространенными инструментами квадратичного дискриминантного анализа являются эллипсоиды минимального объема, содержащие заданное множество точек (minimal volume ellipsoid — MVE). Разнообразные алгоритмы построения эллипсоидов MVE изложены в работах [1–5]. Реляционный дискриминантный анализ основан на идее перехода из пространства признаков в пространство близости и гипотезе компактности, утверждающей, что объекты одного и того класса больше похожи друг на друга, чем на объекты альтернативного класса. Гипотеза компактности в пространстве близости означает, что сходные объекты находятся на примерно одинаковом расстоянии от альтернативного класса. Впервые эти идеи были развиты (независимо) в работах Ю. Петунина и Д. Клюшина [6], а также R. Duin [7]. В дальнейшем они нашли развитие в работах E. Pekalska, О. Середина, В. Моттля и др. [8–13].

Применение эллипсоидов при классификации многомерных данных часто приводит к неопределенности, вызванной пересечением доверительных

областей. Естественным способом устранения неопределенности такого рода является многомерное ранжирование [14]. По классификации, предложенной V.Barnett методы многомерного ранжирования подразделяются на маргинальные, редуцированные, частичные и условные. Среди этих методов следует выделить методы, основанные на оценке статистической глубины выборок относительно центра распределения и соответствующих методах пилинга (см. [15–20] и др.) В этой статье мы предлагаем новый метод устранения неопределенности, возникающей в квадратичном реляционном дискриминантном анализе, основанный на эллипсах Петунина [3, 5] и методе, предложенном авторами в работе [21].

2. Пространство близости и понижение размерности

Опишем общую многоэтапную схему перехода из многомерного пространства признаков в двумерное пространство близости. Пусть G_1 и G_2 — две генеральные совокупности (классы), из которых извлечено по N обучающих выборок $u_k = (x_1, x_2, ..., x_n)$, $k = 1, ..., N$ и $v_l = (y_1, y_2, ..., y_n)$, $l = 1, ..., N$, где n — количество признаков, а x_i и y_j — векторы размерности m. Следовательно, n — количество столбцов, а m — количество строк исходной матрицы.

Процедура перехода состоит из следующих этапов.

1. Перекрестное вычисление меры близости между выборками из генеральных совокупностей G_1 и G_2 (внутри классов и между классами).
2. Построение плоскостей близости и вычисление эллипсов Петунина, содержащих заданное множество точек в пространстве близости. При этом оси координат соответствуют средним значениям меры близости между выборками по разным показателям.
3. Многомерное ранжирование, позволяющее устранить неопределенность, возникающую из-за точек, попадающих в пересечение эллипсов.

В результате этой процедуры возникают $n\,(n-1)\,/2$ пар доверительных эллипсов. Рассмотрим каждый из перечисленных этапов.

3. Перекрестное вычисление меры близости

Для перехода из пространства признаков в пространство близости применим p-статистику Петунина [22].

3.1. **Мера близости между выборками.** Пусть $x = (x_1, x_2, ..., x_m) \in G_1$ и $y = (y_1, y_2, ..., y_m) \in G_2$ — выборки, полученные путем простого случайного выбора из генеральных совокупностей G_1 и G_2, имеющих функции распределения F_1 и F_2 соответственно. Предположим, что $F_1\,(x) \equiv F_2\,(x)$ и построим вариационный ряд $x_{(1)} \leq ... \leq x_{(m)}$. Обозначим через A_{ij} событие, которое состоит в том, что выборочное значение y_k попадает между

порядковыми статистиками $x_{(i)}$ и $x_{(j)}$: $A_{ij} = \left\{ y_k \in \left(x_{(i)}, x_{(j)} \right) \right\}$. По гипотезе Хилла (см. [20, 22]),

$$P(A_{ij}) = P\left(x_k \in (x_{(i)}, x_{(j)}) \right) = p_{ij} = \frac{j-i}{m+1}.$$

Построив доверительный интервал для неизвестной вероятности события A_{ij} по формулам Вильсона

$$p_{ij}^{(1)} = \frac{h_{ij}m + g^2/2 - g\sqrt{h_{ij}(1-h_{ij})m + g^2/4}}{m+g^2},$$

$$p_{ij}^{(1)} = \frac{h_{ij}m + g^2/2 + g\sqrt{h_{ij}(1-h_{ij})m + g^2/4}}{m+g^2},$$

где h_{ij} – частота события A_{ij} в m испытаниях, получим доверительный интервал $I_{ij} = \left(p_{ij}^{(1)}, p_{ij}^{(2)} \right)$ с уровнем значимости, который определяется параметром g. В частности, при $g = 3$ уровень значимости этого интервала не превышает 0,05.

Обозначим через B событие, которое состоит в том, что вероятность p_{ij} попадает в интервал I_{ij}, а через L – количество интервалов I_{ij}, содержащих вероятности p_{ij}. Тогда $h = \rho(x, y) = \frac{L}{N}$ — мера близости между выборками x и y, или р-статистика. Полагая в формулах Вильсона, приведенных выше, $h_{ij} = h, m = N$ и $g = 3$, получим доверительный интервал $I = \left(p^{(1)}, p^{(2)} \right)$ для вероятности $p(B)$.

3.2. **Плоскость близости.** Вычислим меру близости между выборками объектов из классов G_1 и G_2 по двумпоказателям. Рассмотрим матрицы показателей k-го объекта из группы G_1 и l-го объекта из группы G_2:

$$u_k = \begin{pmatrix} x_{11}^{(k)} & x_{12}^{(k)} & \cdots & x_{1n}^{(k)} \\ x_{21}^{(k)} & x_{22}^{(k)} & \cdots & x_{2n}^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1}^{(k)} & x_{m2}^{(k)} & \cdots & x_{mn}^{(k)} \end{pmatrix},$$

$$v_l = \begin{pmatrix} y_{11}^{(l)} & y_{12}^{(l)} & \cdots & y_{1n}^{(l)} \\ y_{21}^{(l)} & y_{22}^{(l)} & \cdots & y_{2n}^{(l)} \\ \vdots & \vdots & \ddots & \vdots \\ y_{m1}^{(l)} & y_{m2}^{(l)} & \cdots & y_{mn}^{(l)} \end{pmatrix}.$$

Выделим из матриц u_k и v_l столбцы, соответствующие i-му показателю: $X_i^{(k)} = \left(x_{1i}^{(k)}, x_{2i}^{(k)}, ..., x_{mi}^{(k)} \right)$ і $Y_i^{(l)} = \left(y_{1i}^{(l)}, y_{2i}^{(l)}, ..., y_{mi}^{(l)} \right)$. Теперь мы можем вычислить значения меры близости между выборками $X_i^{(k)}$ и $Y_i^{(l)}$ и сформировать вектор значений меры близости между объектами u_k и v_l по

каждому показателю:

$$\mu_{kl}^{(1)} = \rho\left(X_1^{(k)}, Y_1^{(l)}\right),$$
$$\mu_{kl}^{(2)} = \rho\left(X_2^{(k)}, Y_2^{(l)}\right),$$
$$\ldots$$
$$\mu_{kl}^{(n)} = \rho\left(X_N^{(k)}, Y_N^{(l)}\right),$$

а затем найти значения усредненных p-статистик

$$\nu_k^{(1)} = \frac{1}{N}\sum_{t=1}^{N}\mu_{kt}^{(1)},$$
$$\nu_k^{(2)} = \frac{1}{N}\sum_{t=1}^{N}\mu_{kt}^{(2)},$$
$$\ldots$$
$$\nu_k^{(n)} = \frac{1}{N}\sum_{t=1}^{N}\mu_{kt}^{(n)},$$

которые характеризуют сходство между объектом u_k и объектами группы G_2 по i-му показателю. Эту схему можно применить и для сравнения объекта u_k с остальными объектами группы G_1:

$$\bar{\nu}_k^{(1)} = \frac{1}{N-1}\sum_{s=1,s\neq k}^{N}\mu_{ks}^{(1)},$$
$$\bar{\nu}_k^{(2)} = \frac{1}{N-1}\sum_{s=1,s\neq k}^{N}\mu_{ks}^{(2)},$$
$$\ldots$$
$$\bar{\nu}_k^{(n)} = \frac{1}{N-1}\sum_{s=1,s\neq k}^{N}\mu_{ks}^{(n)},$$

где индекс s пробегает множество объектов в классе G_1.

Объединим усредненные меры близости в пары, образующие координаты в плоскости близости (i,j) для i-го и j-го показателей $\left(\nu_t^{(i)}, \nu_t^{(j)}\right)$, $\left(\bar{\nu}_s^{(i)}, \bar{\nu}_s^{(j)}\right)$ $(i,j = 1,2,...,m; t,s = 1,2,...,n)$. В итоге в каждой плоскости (i,j) получаем два множества точек, состоящих из точек, характеризующих среднюю внутриклассовую близость между объектами, и точек, характеризующих среднюю междуклассовую близость между объектами.

3.3. **Доверительные эллипсы Петунина.** Алгоритм построения эллипса Петунина, содержащего множество точек $M_n = \{(x_1, y_1), ..., (x_n, y_n)\}$, состоит из следующих этапов [5]. Построим выпуклую оболочку точек $M_n = \{(x_1, y_1), ..., (x_n, y_n)\}$. Найдем вершины выпуклой оболочки (x_k, y_k)

и (x_l, y_l), лежащие на диаметре выпуклой оболочки, т.е. вершины, наиболее удаленные друг от друга. Соединим точки (x_k, y_k) и (x_l, y_l) отрезком L. Найдем вершины выпуклой оболочки (x_r, y_r) и (x_q, y_q), наиболее удаленные от L. Соединим точки (x_r, y_r) и (x_q, y_q) отрезками L_1 и L_2, параллельными к отрезку L. Проведем через точки (x_k, y_k) и (x_l, y_l) отрезки L_3 и L_4, перпендикулярные к отрезку L. Пересечения отрезков L_1, L_2, L_3 и L_4 образуют прямоугольник П, стороны которого имеют длины a и b.

Допустим, что $a \leq b$. Переведем левый нижний угол прямоугольника в начало новой системы координат с осями Ox' и Oy' с помощью поворота и параллельного переноса. Точки (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) перейдут в точки (x'_1, y'_1), (x'_2, y'_2), ..., (x'_n, y'_n). Отобразим точки (x'_1, y'_1), (x'_2, y'_2), ..., (x'_n, y'_n) в точки $(\alpha x'_1, y'_1)$, $(\alpha x'_2, y'_2)$, ..., $(\alpha x'_n, y'_n)$, где $\alpha = \frac{a}{b}$. В результате получим совокупность точек, лежащих в квадрате S.

Вычислим центр (x'_0, y'_0) квадрата S и найдем расстояния $r_1, r_2, ..., r_n$ от него до каждой точки $(\alpha x'_1, y'_1)$, $(\alpha x'_2, y'_2)$, ..., $(\alpha x'_n, y'_n)$. Наибольшее число $R = \max(r_1, r_2, ..., r_n)$ определяет круг с центром в точке (x'_0, y'_0) и радиусом R. В итоге все точки $(\alpha x'_1, y'_1)$, $(\alpha x'_2, y'_2)$, ..., $(\alpha x'_n, y'_n)$ оказываются внутри круга с радиусом R. Растягивая этот круг вдоль оси Ox' с коэффициентом $\beta = \frac{1}{\alpha}$ и выполняя обратные преобразования поворота и переноса, получим эллипс Петунина. В результате происходит их естественное ранжирование точек, каждая из которых лежит на отдельном эллипсе [21].

Теорема [20]. Если $x_1, x_2, ..., x_n$ — независимые и одинаково распределенные случайные выборки из генеральной совокупности G, E_n — эллипс Петунина, содержащий точки $x_1, x_2, ..., x_n$, и $x_{n+1} \in G$, то

$$P(x_{n+1} \in E_n) = \frac{n}{n+1}.$$

4. Решающее правило

Гипотеза компактности в пространстве близости подразумевает симметричность отношения близости между объектами внутри и между классами. По этой причине схема принятия решений должна учитывать внутриклассовые и межклассовые расстояния между объектами классов G_1 и G_2. Следовательно, необходимо построить четыре вида эллипсов:

1. E_{11} — эллипсы, содержащие точки, представляющие собой значения меры близости между объектами внутри класса G_1.
2. E_{21} — эллипсы, содержащие точки, представляющие собой значения меры близости между объектами класса G_2 и класса G_1.
3. E_{22} — эллипсы, содержащие точки, представляющие собой значения меры близости между объектами внутри класса G_2.
4. E_{12} — эллипсы, содержащие точки, представляющие собой значения меры близости между объектами класса G_1 и G_2.

Классификация выполняется на основе совокупности пар полуплоскостей, содержащих эллипсы E_{11} и E_{21} на первой полуплоскости и E_{22} и E_{12} на второй полуплоскости. Количество полуплоскостей зависит от количества пар показателей, котоорые можно образовать. Поскольку из n

показателей можно образовать $n(n-1)/2$ пар, умножив это число на 2, получаем общее количество пар эллипсов, равное $n(n-1)$.

Если точка попадает в эллипс E_{11}, значит, средняя мера близости между объектом, которому она соответствует, и объектами класса G_1, характерна для объектов класса G_1. Следовательно, объект следует отнести к классу G_1. Аналогично, если точка попадает в эллипс E_{22}, значит, средняя мера близости между объектом, которому она соответствует, и объектами класса G_2, характерна для объектов класса G_2. Следовательно, объект следует отнести к классу G_2.

Если точка попадает в эллипс E_{21}, значит, средняя мера близости между объектом, которому она соответствует, и объектами класса G_1, характерна для объектов класса G_2. Следовательно, объект следует отнести к классу G_2. Аналогично, если точка попадает в эллипс E_{12}, значит, средняя мера близости между объектом, которому она соответствует, и объектами класса G_2, характерна для объектов класса G_1. Следовательно, объект следует отнести к классу G_1.

Если точка попадает к пересечение эллипсов, то сравниваются ранги точки в каждом из эллипсов. Это не требует дополнительных вычислений, поскольку ранги автоматически определяются в ходе построения эллипса Петунина. Точка относится к тому эллипсу, в котором он имеет меньший ранг, т. е. лежит "глубже".

Окончательный результат для тестируемого объекта принимается простым голосованием.

5. Результаты

Для проверки практической ценности предложеного метода были проведены эксперименты на основе данных о состоянии буккального эпителия у 25 больных раком молочной железы и 25 больных фиброаденоматозом. Эти данные были предоставлены Институтом экспериментальной патологии, онкологии и радиобиологии им. Р. Е. Кавецкого НАН Украины (доктор медицинских наук Н. В. Бородай). Содержание этих данных и результаты классификации без учета ранжирования приведены в монографии [23]. Чувствительность распознавания рака с помощью квадратичного метода без ранжирования составляла 72%, а специфичность была очень низкой (12%). Это привело к необходимости усложнить процедуру принятия решения, включив в нее дополнительные критерии. После устранения неопределенности чувствительность и специфичность квадратичного критерия (без дополнительных критериев) составили 80% и 44% соответственно. Несмотря на то, что специфичность остается недостаточно высокой, проведенные эксперименты позволяют утверждать, что процедура ранжирования на 10% повышает чувствительность и почти в 4 раза повышает специфичность распознавания рака молочной железы.

Для повышения чувствительности и специфичности критерия решающее правило можно усилить, например, за счет бустинга или применения

генетического алгоритма построения эллипса, уменьшающего площадь доверительной области без уменьшения уровня статистической значимости.

6. Выводы

Предложенный метод устранения неопределенности в квадратичном реляционном дискриминантном анализе на основе ранжирования многомерных выборок позволяет значительно снизить неопределенность, вызванную пересечением эллипсоидов, поскольку каждая точка выборки имеет уникальный ранг. Асимптотический уровень значимости статистического критерия, на котором основано решающее правило, не превышает 0,05. Вычислительный эксперимент продемонстрировал практическую ценность разработанной методики для повышения точности диагностики рака молочной железы.

Литература

1. Silverman B. W., Titterington D. M. Minimum covering ellipses // SIAM J. Sci. Stat. Comput. — 1980. — V. 1. — P. 401–409.
2. Welzl E. Smallest enclosing disks (balls and ellipsoids) // In: Proceedings of New Results and New Trends in Computer Science. LNCS (Springer, Berlin). — 1991. — V. 555. — P. 359–370.
3. Ляшко С. И., Рублев Б. В. Минимальные эллипсоиды и максимальные симплексы в трехмерном евклидовом пространстве // Кибернетика и системный анализ. — 2003. — №6. — С. 65–70.
4. Kumar P., Yildirim E. A. Minimum volume enclosing ellipsoids and core sets // J. Optim. Theory Appl. — 2005. — V. 126 — P. 1–21.
5. Петунин Ю. И., Рублев Б. В. Распознавание образов с помощью квадратичных дискриминантных функций // В сб: Вычислительная и прикладная математика. — 1996. — Вып. 80. — С. 89–104.
6. Andrushkiw R. I. Petunin Yu. I., Klyushin D. A. Nonlinear algorithms of pattern recognition for computer-aided diagnosis of breast cancer // Nonlinear analysis, Theory, Methods & Applications. — Oxford, Elsevier. — 1997. — V. 30, №8 — P. 5431–5336.
7. Duin R. P.W., Ridder D., and Tax D. M. J. Experiments with object based discriminant functions; a featureless approach to pattern recognition // Pattern Recognition Letters. — 1997. — V. 18, №11–13. — P. 1159–1166.
8. Duin R. P. W., Pekalska E., and Ridder D. Relational discriminant analysis // Pattern Recognition Letters. — 1999. — V. 20, №11–13. — P. 1175–1181.
9. Pekalska E., Duin R. P. W. Automatic pattern recognition by similarity representations // Electronics Letters. — 2001. — V. 37, №3. — P. 159–160.
10. Pekalska E., Duin R. P. W. Dissimilarity representations allow for building good classifiers // Pattern Recognition Letters. — 2002. — V. 23, №8. — P. 943–956.
11. Pekalska E., Duin R. P. W. The Dissimilarity Representation for Pattern Recognition, Foundations and Applications. — Singapore: World Scientific, 2005. — p. 636.
12. Середин О.С. Методы и алгоритмы беспризнакового распознавания образов // Автореферат дисс. на соиск. ученой степени канд. наук: 05.13.17 — М., 2001. — 20 с.

13. Mottl V., Dvoenko S., Seredin O., Kulikowski C., Muchnik I. Featureless pattern recognition in an imaginary Hilbert space and its application to protein fold classification // Machine Learning and Data Mining in Pattern Recognition. Lecture Notes in Computer Science. — 2001. — V. 2123.— P. 322–336.
14. Barnett V. The ordering of multivariate data // Journal of the Royal Statistical Society. Series A (General).— V. 139. — 1976. — №3. — P. 318–355.
15. Tukey J. W. Mathematics and the picturing of data // Proceedings of the International Congress of Mathematician, Montreal, Canada. — 1975. — P. 523–531.
16. Titterington D. M. Estimation of correlation coefficients by ellipsoidal trimming // Appl. Statist. — 1978. — V. 27. — P. 227–234.
17. Oja H. Descriptive statistics for multivariate distributions // Statistics and Probability Letters. — 1983. — V. 1. — P. 327–332.
18. Liu R. J. On a notion of data depth based on random simplices //Annals of Statictics. — 1990. — V. 18. — P. 405–414.
19. Zuo Y., Serfling R. General notions of statistical depth function // Annals of Statistics. — 2000. — V. 28. — P. 461–482.
20. Ляшко С. И., Клюшин Д. А., Алексеенко В. В. Многомерное ранжирование и эллиптический пилинг // Кибернетика и системный анализ. — 2013. — №4. — С. 29–36.
21. Клюшин Д. А., Присяжна М. В. Еліптична функція статистичної глибини даних // Вісник Київського національного університету імені Тараса Шевченка, Серія фіз.-мат наук. — 2014. — №1. — С. 179–184.
22. Клюшин Д. А., Петунин Ю. И. Непараметрический критерий эквивалентности генеральных совокупностей, основанный на мере близости между выборками // Український математичний журнал. — 2003. — Т. 5, №2. — С. 147—163.
23. Клюшин Д. А., Петунин Ю. И. Доказательная медицина. Применение статистических методов. — М.: Вильямс, 2008. — 320 с.

Факультет кибернетики, Киевский национальный университет имени Тараса Шевченко, ул. Владимирская, 64, Киев, 01601, Украина.

Поступила 7.09.14