УДК 621.391

Ю.М. Коршунов

ПРИМЕНЕНИЕ МЕТОДА ДВУХАЛЬТЕРНАТИВНОГО ДИСКРИМИНАНТНОГО АНАЛИЗА ДЛЯ ОЦЕНКИ КАЧЕСТВА ВЫПУСКАЕМОЙ ПРОДУКЦИИ

Дается описание одного из многих известных алгоритмов дискриминантного анализа, реализованного в пакете Matlab на кафедре АИТУ, позволяющего путем измерения небольшого числа признаков, характеризующих качество выпускаемого изделия, обнаруживать с высокой достоверностью непригодные к эксплуатации изделия

Ключевые слова: Кластерный анализ, дискриминантный анализ, класс, объект, признаки, обучающая выборка.

Быстрое развитие производства, которое ожидается в связи с внедрением инновационных технологий, потребует повышенного внимания к оценке качества выпускаемой продукции. Современные средства вычислительной техники позволяют привлечь для решения подобных задач мощный аппарат многомерного статистического анализа, особенно те его разделы, которые относятся к решению задач классификации, как, например, ***кластерный анализ*** (“классификация без учителя”) и ***дискриминантный анализ*** (“классификация с учителем”). Для решения поставленной задачи более приемлемым является дискриминантный анализ, на котором мы остановимся более подробно.

Дискриминантный анализ ставит задачу отнести исследуемый объект, характеризуемый m признаками, к одному из k заранее заданных классов. Роль учителя выполняют заданные для каждого класса статистические выборки размером $n_i \times m$, где n_i число объектов в выборке i-го класса. Задача дискриминантного анализа сводится к тому, чтобы на основе имеющихся выборок сформировать решающее правило, позволяющее отнести любой новый объект к наиболее близкому к нему по своим свойствам классу.

Однако в такой общей постановке решение задачи методом дискриминантного анализа имеет весьма сложное математическое описание [1] и ее трудно реализовать на практике. Мы ограничимся рассмотрением более простого варианта, содержащего всего два класса объектов, названного ***двухальтернативным дискриминантным анализом.*** Под словом ***объект*** будем понимать выпускаемое промышленностью изделие.

Хотя выпуску нового вида изделий предшествует длительный период разработки конструкции и технологии производства, Однако, полной гарантии того, что все выпускаемые при массовом производстве изделия будут полностью удовлетворять всем требованиям, быть не может. Поэтому выделяется m признаков, которые могут быть сравнительно легко измерены на каждом выпущенном изделии и по значениям которых можно хотя бы приближенно судить о качестве изделия. Точная оценка потребовала бы более глубокого анализа, на проведение которого при массовом производстве времени нет. Результаты грубой оценки позволяют каждый объект отнести или к классу 1 -объект исправен и может быть пущен в эксплуатацию, или к классу 2-объект непригоден к эксплуатации (брак). Но это решение нельзя признать окончательным, так как среди объектов, признанных пригодными к эксплуатации, могут оказаться бракованные. Однако обнаружить это можно только, запустив объект в эксплуатацию и проверив на практике качество его работы.

Таким образом, вырисовывается следующая процедура проверки качества объектов. После того, как запущен процесс производства, каждый выпущенный объект проверяется по имеющимся признакам, и те объекты, которые признаны годными, поступают в продажу. За ними устанавливается жесткий контроль и объекты, на которые поступили жалобы, заносятся в класс бракованных. Через некоторое время класс 2 бракованных объектов будет содержать достаточно большое число объектов и эксперимент можно прекратить. Объекты, на которые жалобы не поступали, заносятся в класс 1. Два полученных класса будем рассматривать как ***обучающую выборку*** для реализации процедуры дискриминантного анализа. Приведем математическое описание рассмотренной процедуры.

Обозначим X_j статистическую выборку объектов класса j с числом объектов n_j, $j=1,2$, а через x_{ji}, $i=1,\cdots,n_j$ i - й объект в выборке X_j. Будем также считать, что обе выборки подчиняются нормальному закону распределения

$$w(X_j)=N(M_j,Q_j),$$

где M_j – вектор математических ожиданий, а Q_j – ковариационная матрица j-й выборки.

Для формирования ***решающего правила*** введем в рассмотрение уравнение гиперплоскости в m – мерном пространстве признаков

$$h(x)=U^T x+u_0, \qquad (1)$$

где $x=(x_{j1},\cdots,x_{jm})$ – вектор значений признаков, $U=(u_1,\cdots,u_m)$ – вектор числовых коэффициентов, u_0 – скаляр.

В выражении (1) $h(x)$ представляет собой скалярную случайную величину с нормальным законом распределения, имеющим для выборок 1 и 2 вид

$$w(h_j)=N(a_j,\sigma_j^2),\ \ j=1,2,$$

где a_j,σ_j^2 – математические ожидания и дисперсии плотности вероятностей $w(h_j)$.

Задача состоит в том, чтобы подобрать значения параметров разделительной гиперплоскости U и u_0, при которых получится разбиение всего множества объектов $X_1 \bigcup X_2$ на два класса, наиболее близких к заданному обучающей выборкой. При этом должно быть

$$h(x)=U^T x+u_0=0, \qquad (2)$$

если x лежит на разделительной гиперплоскости;

$$\begin{aligned} &h(x)<0, \textit{если } x=x_1 \subset X_1,\\ &h(x)>0, \textit{если } x=x_2 \subset X_2. \end{aligned} \qquad (3)$$

Однако нет никакой гарантии, что удастся подобрать параметры U и u_0 такими, что все объекты $x_j \subset X_j$ $j=1,2$ в выборках X_1и X_2 будут лежать по разные стороны от разделительной гиперплоскости. При этом возможны ошибки вида:

$h_1 \mid x_2$ – ошибка первого рода или ***пропуск цели*** (негодный объект x_2 признан годным);

$h_2 \mid x_1$ – ошибка второго рода или ***ложная тревога*** (годный объект x_1- признан бракованным).

Поскольку ошибки такого вида неизбежны, то ставится задача подобрать параметры U и u_0 разделительной гиперплоскости такими, при которых вероятности ошибок первого и второго рода, обозначаемые как $\alpha=p(h_1 \mid x_2)$ и $\beta=p(h_2 \mid x_1)$, были бы минимальны. Трудность здесь состоит в том, что эти ошибки определенным образом связаны между собой, так что с уменьшением ошибки первого рода возрастает ошибка второго рода и наоборот. Поэтому задачу минимизации этих ошибок приходится решать с применением настроечного параметра s, $0<s<1$, определяющего приемлемую вероятность ошибки первого рода α, и применением критерия Неймана-Пирсона, минимизирующего вероятность ошибки второго рода β при заданной вероятности α. При этом приходится опробовать разные значения s и выбрать то значение, при котором вероятности ошибок α и β будут наиболее приемлемыми. Не вдаваясь в подробности, приведем основные соотношения, используемые при решении задачи.

Если параметры U и u_0 уже найдены, то параметры плотности вероятностей $w(h_j)$ найдутся как

$$\begin{aligned} a_j &= E\big(h(x) \mid x_j\big)=E\big(U^T x+u_0 \mid x_j\big)=\\ &=U^T M_j+u_0,\ \ j=1,2. \end{aligned} \qquad (4)$$

$$\sigma_j^2=E[(h(x)-a_j)^2 \mid x_j]=U^T Q_j uU, j=1,2. \qquad (5)$$

Здесь E символ определения математического ожидания. При этом

$$\alpha=\int_{-\infty}^{0} p(h_1 \mid x_2)dh \qquad (6)$$

$$\beta=1-\int_{-\infty}^{0} p(h_2 \mid x_1)dh \qquad (7)$$

При заданном s параметры U и u_0 находятся по соотношениям, полученным в результате решения оптимизационной задачи:

$$U=\big(s \cdot Q_1+(1-s)\cdot Q_2\big)^{-1}\big(M_2-M_1\big). \qquad (8)$$

$$u_0 = -\frac{s\sigma_1^2 U^T M_2 + (1-s)\sigma_2^2 M_1}{s\sigma_1^2 + (1-s)\sigma_2^2} \qquad (9)$$

По найденным параметрам разделительной гиперплоскости следует проверить значения $h(x)$ для всех объектов, входящих в выборку первого $(x = x_1)$ и второго $(x = x_2)$ классов и подсчитать числа $n12$ и $n21$ ошибочно классифицированных объектов, для которых $h(x_1) > 0$ и $h(x_2) < 0$. Отношения $n12/n_1$ и $n21/n_2$ при достаточно большом числе элементов в выборках n_1 и n_2 будут характеризовать вероятности ошибок первого и второго рода.

Работа алгоритма должна быть проверена при различных значениях настроечного параметра s. Для этого в диапазоне [0,1] выделяется Ls равноотстоящих значений $is = 1,\cdots,Ls$, для каждого из которых значение s находится по соотношению

$$s = (2 \cdot is - 1)/2 \cdot Ls . \qquad (10)$$

За решающее правило принимаются те значения s, U, u_0, при которых общее число ошибочных решений окажется наименьшим.

Теперь можно ввести в программу параметры любого нового объекта и проверить, не будет ли он отнесен к классу 2, т.е. окажется бракованным.

Рассмотренный алгоритм был реализован на языке ***Matlab***. Основная трудность, заключающаяся в получении ***обучающей выборки***, была преодолена тем, что на кафедру обратилась одна организация с готовыми выборками объектов первого и второго классов и с просьбой помочь им в разработке решающего правила для проведения дискриминантного анализа. Эта просьба явилась толчком для разработки алгоритма и его последующей реализации. При реализации алгоритма были использованы некоторые программы, описанные в работах [2,3]. Приведем результаты работы программы.

Фрагмент выборки 1-го класса. $n1 = 60$

i	x1			
18	0.70	1.66	1.04	0.09
19	0.53	2.06	2.32	-0.03
20	0.60	1.49	3.10	-0.02
21	0.64	3.46	5.19	0.06
22	0.61	0.64	0.15	0.18

Фрагмент выборки 2-го класса $n2 = 60$

i	x2			
18	0.11	0.62	0.44	0.20
19	0.33	0.48	1.35	-0.31
20	0.28	1.68	0.49	0.10
21	0.45	1.88	0.33	-0.09
22	0.30	1.30	0.72	0.21

$M_1 =$ 0.7390 2.0442 2.0092 0.1125

$$Q_1 = \begin{matrix} 0.0160 & 0.0354 & -0.0856 & -0.0022 \\ 0.0354 & 0.5476 & 0.3211 & -0.0135 \\ -0.0856 & 0.3211 & 2.4121 & -0.0207 \\ -0.0022 & -0.0135 & -0.0207 & 0.0323 \end{matrix}$$

$M_2 =$ 0.3497 0.9537 0.6217 -0.0558

$$Q_2 = \begin{matrix} 0.0299 & 0.0062 & 0.0022 & -0.0248 \\ 0.0062 & 0.6303 & -0.0985 & 0.0997 \\ 0.0022 & -0.0985 & 0.2731 & -0.0062 \\ -0.0248 & 0.0997 & -0.0062 & 0.1447 \end{matrix}$$

Введите Ls: 18

is	alf	bet
8	0.0910	0.0148
9	0.0670	0.0208
10	0.0468	0.0291
11	0.0307	0.0407
12	0.0185	0.0570
13	0.0101	0.0799

Выбор: $is = 10$

$U =$ -23.7610 -0.2210 -1.7753 -5.7351

$u_0 =$ 16.8956

Выделение ошибочно классифицированных объектов

$n12 = 3$, объекты: 16 22 52.

$n21 = 2$, объекты: 15 35.

Ввод новой выборки

i	x1			
1	0.5986	0.2956	0.5345	0.0279
2	0.8464	0.8096	0.3288	0.6086
3	0.9546	0.5747	0.3773	0.7443
4	0.5547	0.6768	0.6499	0.6861
5	0.4428	0.5636	0.7060	0.5391
6	0.4675	0.5712	0.2353	0.7401

Выделение бракованных объектов

$n12 = 3$, объекты: 1 5 6 .

Библиографический список

1. *Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика: Классификация и снижение размерности. -М.: Финансы и статистика, 1989. 607 с.

2. *Коршунов Ю.М.* Получение многомерной статистической выборки с заданными корреляционными свойствами //Вестник РГРТУ. Вып. 23. 2008. С. 21-24.

3. Математическое моделирование экономических процессов; методическое пособие к

лабораторным работам. Сост. Ю.М. *Коршунов.,* *В.Н. Федоров.* РГРТА, Рязань. 2002. 23 с.