И.Л. Толмачев, М.В. Хачумов

Бинарная классификация на основе варьирования размерности пространства признаков и выбора эффективной метрики[1]

Аннотация. Рассматривается метод решения задачи бинарной классификации, основанный на снижении размерности *n*-мерного пространства признаков до двумерного, построении разделяющей гиперплоскости и ее обратном отображении в *n* -мерное пространство. Обратное преобразование обеспечивает удобство решения задачи классификации непосредственно в системе исходных признаков. Предлагаемый способ построения разделяющей гиперплоскости по заданной учебной выборке основан на последовательном применении известных алгоритмов и их адаптации к условиям задачи. Дано сравнение качества распознавания объектов при использовании расстояний Евклида, Махаланобиса и Евклида-Махаланобиса.

Ключевые слова: бинарная классификация, разделяющая гиперплоскость, комитет большинства, пространство признаков, МГК, МНК, ZET, метрики и расстояния.

Введение

Есть два основных подхода к варьированию пространства признаков, направленных на решение задачи бинарной классификации. Одним из них является метод опорных векторов, который осуществляет перевод исходных векторов значений признаков, описывающих заданные объекты, в пространство более высокой размерности и поиск разделяющей гиперплоскости с максимальным «зазором» в этом пространстве [1].

Другой подход, напротив, направлен на уменьшение размерности признакового пространства с целью снижения сложности задачи и, как правило, опирается на метод главных компонент (МГК) [2]. Задача первоначального снижения размерности пространства признаков, нахождения комитета большинства на плоскости и его обратного отображения в n-мерное пространство рассматривалась в работе [3]. Данный подход лежит несколько в стороне от двух основных, представляет определенный интерес с теоретической стороны, но исследован недостаточно полно, что снижает его практическую значимость.

Рациональная сторона варьирования размерности пространства признаков заключена в относительной простоте решения задачи на плоскости, причем прямое преобразование является решающим для распознавания, а обратное − преследует своей единственной целью удобство работы с исходными признаками, имеющими физический смысл, в отличие от преобразованных координат, для которых такой смысл теряется.

В настоящей работе предлагается метод решения задачи бинарной классификации, основанный на известных алгоритмах МГК, метода наименьших квадратов (МНК) и восстановления пропущенных элементов таблиц ZET, применение которых обеспечивает необходимую однозначность преобразований. Комитет большинства необходим, когда одна гиперплоскость не обеспечивает разделение входных векторов на два непересекающихся класса. В настоящем исследовании его построение не является

[1] Работа выполнена при частичной поддержке РФФИ (проект № 09-07-00006) и Программы фундаментальных исследований ОНИТ РАН «Информационные технологии и методы анализа сложных систем» (проект 2.2).

принципиальным, поскольку предлагаемая технология бинарной классификации с варьированием пространства признаков без потери общности достаточно подробно поясняется на примере одной разделяющей гиперплоскости. Дополнительно в работе исследована эффективность применения различных метрик для распознавания входных векторов.

1. Постановка задачи

Пусть задано m объектов $\{\omega_1,...,\omega_m\}$, каждый из которых представлен вектором $W_i=(x_{i1},...,x_{in})$, $(i=1,...,m)$ значений признаков $X=(x_1,...,x_n)$. Объекты отнесены экспертами к классам Ω_1,Ω_2 следующим образом: $\Omega_1=\{\omega_1,...,\omega_{m_1}\}$, $\Omega_2=\{\omega_{m_1+1},...,\omega_m\}$. Исходная таблица с прецедентами представлена в Табл.1.

Табл. 1

Объекты	Векторы признаков	Признаки и их значения			Класс
		x_1	⋯	x_n	
ω_1	W_1	x_{11}	⋯	x_{1n}	Ω_1
...	...	...	...	...	...
ω_{m_1}	W_{m_1}	$x_{m_1 1}$	⋯	$x_{m_1 n}$	Ω_1
ω_{m_1+1}	W_{m_1+1}	$x_{(m_1+1)1}$	⋯	$x_{(m_1+1)n}$	Ω_2
...	...	...	...	...	...
ω_m	W_m	x_{m1}	⋯	x_{mn}	Ω_2

Табл.1 содержит данные об объектах обучающей выборки. В соответствии с алгебраической теорией распознавания [4] и предложениями работы [3] можно составить систему из m неравенств вида:

$$\sum_{j=1}^{n} x_{ij}x_j > 0,\ \ \forall\omega_i\in\Omega_1,\ \ i=1,...,m_1;$$
$$\sum_{j=1}^{n} x_{ij}x_j < 0,\ \ \forall\omega_i\in\Omega_2,\ \ i=(m_1+1),...,m. \qquad (1)$$

Если система неравенств (1) совместна, то имеется гиперплоскость, разделяющая множество объектов на два класса в соответствии с обучающей выборкой, представленной в Табл. 1. Уравнение разделяющей гиперплоскости может быть записано в виде

$$f(x_1,...,x_n)=\sum_{i=1}^{n}\alpha_i x_i+\alpha_{n+1}=0, \qquad (2)$$

где $\alpha_1,...,\alpha_{n+1}$ – коэффициенты в уравнении (2). При подстановке в (2) вектора признаков $W(\omega)$ тестируемого объекта ω по знаку $f(x_1,...,x_n)$ относим объект к соответствующему классу или имеем неопределенность на основании неравенств (1).

Заметим, в работе [3] принято α_{n+1}=0, что ограничивает возможности получения решения. Если система несовместна, то задача решается проведением p гиперплоскостей с построением решающей функции по методу комитета большинства. Под комитетом понимают наборы векторов $x^1=(x_1^1,...,x_n^1)$, $x^2=(x_1^2,...,x_n^2),...,x^p=(x_1^p,...,x_n^p)$, такие, что каждому из неравенств (1) удовлетворяет большее число таких векторов. Тогда принадлежность объекта ω с вектором признаков $W(\omega)$ классу $\Omega(\omega)$ будет определяться значением решающей функции $F(\omega)$:

$$F(\omega)=\sum_{i=1}^{p} sign(W(\omega),x^i), \qquad (3)$$

причем:

$$\Omega(\omega)=\begin{cases}\Omega_1, & \text{если}\ \ F(\omega)>0,\\ \Omega_2, & \text{если}\ \ F(\omega)<0,\\ \Delta, & \text{если}\ \ F(\omega)=0,\ \ (\textit{неопределенность}).\end{cases} \qquad (4)$$

Таким образом, задача бинарной классификации заключается в построении разделяющей гиперплоскости $f(x_1,...,x_n)$ или решающей функции $F(\omega)$ по данным Табл.1. Будем решать ее, используя переходы в новое пространство признаков $Y=(y_1,...,y_n)$ с последующим возвратом в исходное $X=(x_1,...,x_n)$ на основе принципов варьирования размерности и детерминированности действий. Дополнительно исследуем эффективность некоторых метрик для задач (1)-(4).

2. Метод решения

Предлагаемый метод решения поставленной задачи включает несколько этапов.

1. Предварительное преобразование пространства признаков и условий задачи в новую

систему координат на основе МГК [6]. Преобразование предполагает выполнение следующих действий:

а) Построение матрицы ковариаций C размерности $n \times n$ по данным Табл.1

$$C = \begin{pmatrix} c_{11} & c_{12} & \cdot & \cdot & \cdot & c_{1n} \\ c_{21} & c_{22} & \cdot & \cdot & \cdot & c_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ c_{n1} & c_{n2} & \cdot & \cdot & \cdot & c_{nn} \end{pmatrix}, \quad (5)$$

где:

$c_{ij} = \frac{1}{m-1} \sum_{k=1}^{m} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$; $\quad i, j = 1,...,n$;

$\bar{x}_i, \bar{x}_j$ – средние значения признаков;

b) Определение собственных чисел $(\lambda_1,...,\lambda_n)$ матрицы (5), путем решения определителя матрицы A:

$$|A| = |C - \lambda \cdot E| = 0, \quad (6)$$

где E - единичная матрица размерности $n \times n$.

2. Преобразование условий исходной задачи путем пересчета входных векторов $W_i = (x_{i1},...,x_{in})$, в вектора $V_i = (y_{i1},...,y_{in})$, $(i = 1,...,m)$ новой системы признаков $Y = (y_1,...,y_n)$. Составляется Табл. 2, содержащая множество векторов $V = \{V_1,...,V_m\}$ пространства R^n.

Табл. 2

Объекты	Векторы признаков	Признаки и их значения			Класс
		y_1	···	y_n	
ω_1	V_1	y_{11}	···	y_{1n}	Ω_1
...	...	...	...	...	...
ω_{m_1}	V_{m_1}	$y_{m_1 1}$	···	$y_{m_1 n}$	Ω_1
ω_{m_1+1}	V_{m_1+1}	$y_{(m_1+1)1}$	···	$y_{(m_1+1)n}$	Ω_2
...	...	...	...	...	...
ω_m	V_m	y_{m1}	···	y_{mn}	Ω_2

Пересчет выполняется следующим образом: $V_i^T = A_* \cdot W_i^T$, $(i = 1,...,m)$ или в развернутом виде:

$$\begin{pmatrix} y_{i1} \\ y_{i2} \\ \cdot \\ \cdot \\ \cdot \\ y_{in} \end{pmatrix} = \begin{pmatrix} c_{11} - \lambda_1 & c_{12} & \cdot & \cdot & \cdot & c_{1n} \\ c_{21} & c_{22} - \lambda_2 & \cdot & \cdot & \cdot & c_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ c_{n1} & c_{n2} & \cdot & \cdot & \cdot & c_{nn} - \lambda_n \end{pmatrix} \begin{pmatrix} x_{i1} \\ x_{i2} \\ \cdot \\ \cdot \\ \cdot \\ x_{in} \end{pmatrix}. \quad (7)$$

3. Выбор из Табл.2, заполненной на основе (7), двух признаков (x_i, x_j), соответствующих двум наибольшим собственным числам (λ_i, λ_j) и построение неравенств задачи бинарной классификации в пространстве R^2.

4. Построение разделяющей гиперплоскости или, при необходимости, комитета большинства $B = (b^1,...,b^p)$, где $b^k = (b_1^k, b_2^k)$, $k = 1,...,p$, p – количество членов комитета, и решающей функции $F^{/}(\omega)$, классифицирующей объект ω, $\omega \in \{\omega_1,...,\omega_m\}$ в пространстве R^2:

$$F^{/}(\omega) = \sum_{i=1}^{p} sign(V(\omega), b^i) \quad (8)$$

Представляет теоретический интерес проблема выбора минимально необходимого числа p в функции (8), но она выходит за рамки настоящей работы.

5. Восстановление разделяющей гиперплоскости или комитета большинства в пространстве R^n:

а) Включим в Табл. 2 дополнительно двухкомпонентный вектор (β_i , β_j), определяющий коэффициенты разделяющей гиперплоскости или гиперплоскости из комитета, увеличив тем самым количество строк до $m+1$, и восстановим его недостающие элементы в R^n. Для этого предлагается модифицированный метод восстановления табличных данных ZET [7]. Используем в качестве «компетентных» векторов расширенной таблицы пару заполненных векторов-столбцов с номерами (i, j).

Пусть включенная в Табл. 2 строка V имеет пустое поле в столбце k и заполненные поля в столбцах i, j $(k \neq i, j)$. Находим значения модулей коэффициентов корреляции векторов-столбцов с номерами i, j со столбцом k,

обозначаемых как $\gamma(B_k, B_i)$, $\gamma(B_k, B_j)$. Коэффициент корреляции векторов B_1 и B_2 вычисляют по формуле $\gamma(B_1, B_2) = \frac{(B_1, B_2)}{\|B_1\| \cdot \|B_2\|}$. Используя пару компетентных векторов-столбцов с номерами i, j и данный k- й столбец, содержащий пробел, осуществляем прогнозы пропущенного значения β_k^i, β_k^j на основе уравнений линейной регрессии и МНК. Полученные прогнозы используем для вычисления (восстановления) значения элемента β_k, например, следующим образом:

$$\beta_k = \frac{\beta_k^i \cdot \gamma(B_k, B_i) + \beta_k^j \cdot \gamma(B_k, B_j)}{\gamma(B_k, B_i) + \gamma(B_k, B_j)}. \qquad (9)$$

Процедуру восстановления (9) повторим для всех векторов, имеющих незаполненные элементы в дополнительной строке.

b) Восстановленную строку V в системе координат Y отобразим в строку W в систему координат X, решая обратную задачу:

$$W^T = A_*^{-1} \cdot V^T \qquad (10)$$

6. Экспериментальная проверка качества решения задачи бинарной классификации.

Предложенный подход, разумеется, не является единственно возможным, но он позволяет проводить целенаправленные действия (5-10), обеспечивая отсутствие неопределенности в выборе параметров преобразования, наблюдаемой в работе [3].

3. Проведение экспериментов

Рассмотрим пример из работы [3]. Пусть заданы следующие вектора в системе признаков $X = (x_1, ..., x_6)$ пространства R^6.

Класс Ω_1:

$W_1 = (1,\ 2,\ 0,\ 1,\ -1,\ 0)$;

$W_2 = (-1,\ 0,\ -1,\ -2,\ -2,\ 1)$;

$W_3 = (1,\ -1,\ 2,\ -1,\ -1,\ 0)$.

Класс Ω_2:

$W_4 = (0,\ 2,\ 1,\ -1,\ 0,\ 1)$;

$W_5 = (2,\ -1,\ 0,\ -1,\ -1,\ 0)$.

Построим систему неравенств:

$$\begin{aligned}
&1x_1 + 2x_2 + 0x_3 + 1x_4 - 1x_5 + 0x_6 > 0;\\
&-1x_1 + 0x_2 - 1x_3 - 2x_4 - 2x_5 + 1x_6 > 0;\\
&1x_1 - 1x_2 + 2x_3 - 1x_4 - 1x_5 + 0x_6 > 0;\\
&0x_1 + 2x_2 + 1x_3 - 1x_4 + 0x_5 + 1x_6 < 0;\\
&2x_1 - 1x_2 + 0x_3 - 1x_4 - 1x_5 + 0x_6 < 0.
\end{aligned} \qquad (11)$$

Знаки неравенств могут быть без потери общности заменены на обратные, так как принципиальным для бинарной задачи является только наличие разных знаков для объектов, относящихся к классам Ω_1 и Ω_2. В работе [3] система неравенств (11) была отнесена авторами к несовместной. На самом деле можно показать, что существует решение в виде гиперплоскости $f(x_1, ..., x_6) = \sum_{i=1}^{6} \alpha_i x_i + \alpha_7 = 0$, разделяющей элементы обучающей выборки на два класса. Решение α =(1, 0.2, –0.2, –0.4, 1, 0.2, 0) было получено путем настройки нейрона с функцией активации типа единичный скачок по методу Видроу-Хоффа без смещения (α_7 = 0), что соответствовало условиям задачи [3]. Проверка показывает, что для исходных векторов имеет место: $f(W(\omega_1)) = 0$, $f(W(\omega_2)) = -1.8$, $f(W(\omega_3)) = -0.8$, $f(W(\omega_4)) = 0.8$, $f(W(\omega_5)) = 3.2$. Т.е лишь одна точка пространства, соответствующая вектору W_1, оказывается непосредственно на разделяющей гиперплоскости и может быть отнесена как к одному, так и другому классу. Более тонкая настройка нейрона способна снять неопределенность за счет введения смещения, например, α_7 = -0.1. В связи с этим задача построения комитета большинства здесь теряет смысл и далее не рассматривается.

Рассмотрим выполнение установленных этапов преобразования.

1. Занесем данные задачи в Табл. 3.

Табл. 3

Исходные векторы	x_1	x_2	x_3	x_4	x_5	x_6	Класс
W_1	1	2	0	1	-1	0	1
W_2	-1	0	-1	-2	-2	1	1
W_3	1	-1	2	-1	-1	0	1
W_4	0	2	1	-1	0	1	2
W_5	2	-1	0	-1	-1	0	2

2. Построим матрицу ковариаций C:

$$C = \begin{pmatrix} 1.3 & -0.55 & 0.45 & 0.60 & 0.25 & -0.55 \\ -0.55 & 2.30 & -0.20 & 0.90 & 0.50 & 0.30 \\ 0.45 & -0.20 & 1.30 & 0.15 & 0.50 & -0.20 \\ 0.60 & 0.90 & 0.15 & 1.20 & 0.25 & -0.35 \\ 0.25 & 0.50 & 0.50 & 0.25 & 0.50 & 0.00 \\ -0.55 & 0.30 & -0.20 & -0.35 & 0.00 & 0.30 \end{pmatrix} \quad (12)$$

3. Найдем собственные числа матрицы ковариаций (12), решая определитель:

$$\begin{vmatrix} c_{11}-\lambda & c_{12} & c_{13} & c_{14} & c_{15} & c_{16} \\ c_{21} & c_{22}-\lambda & c_{23} & c_{24} & c_{25} & c_{26} \\ c_{31} & c_{32} & c_{33}-\lambda & c_{34} & c_{35} & c_{36} \\ c_{41} & c_{42} & c_{43} & c_{44}-\lambda & c_{45} & c_{46} \\ c_{51} & c_{52} & c_{53} & c_{54} & c_{55}-\lambda & c_{56} \\ c_{51} & c_{52} & c_{53} & c_{54} & c_{55} & c_{56}-\lambda \end{vmatrix} = 0 \quad (13)$$

Из решения (13) получим вектор собственных чисел:

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
0	2.96	1.18	2.45	0	0.29

Результирующая таблица векторов в системе признаков $Y = (y_1,...,y_n)$:

Табл. 4

Векторы в новой системе.	y_1	y_2	y_3	y_4	y_5	y_6	Класс
V_1	-0.03	0.27	-0.01	-0.43	-0.02	-0.85	1
V_2	-0.16	0.87	-0.06	0.39	0.18	0.08	1
V_3	-0.32	0.07	0.75	-0.38	0.34	0.2	1
V_4	0.64	-0.06	0.47	0.45	0.24	-0.29	2
V_5	-0.45	-0.36	-0.22	0.31	0.66	-0.26	2

Как видно из Табл. 4, использование даже одного признака y_2 обеспечивает правильное разделение векторов на два класса. Но для дальнейшего хода исследований нам важно использовать наибольшие собственные числа $\lambda_2 = 2.96$, $\lambda_4 = 2.45$ и соответствующие им собственные векторы в R^2.

4. Выпишем условие задачи в виде неравенств для плоскости с координатами (y_2, y_4):

$$\begin{aligned} &0.27\,y_2 - 0.43\,y_4 > 0, \\ &0.87\,y_2 + 0.39\,y_4 > 0, \\ &0.07\,y_2 - 0.38\,y_4 > 0, \\ &-0.06\,y_2 + 0.45\,y_4 < 0, \\ &-0.36\,y_2 + 0.31\,y_4 < 0. \end{aligned} \quad (14)$$

Покажем, что система неравенств (14) является совместной.

Рассмотрим уравнение разделяющей линии $\beta_2 y_2 + \beta_4 y_4 = 0$. Коэффициенты β_2, β_4 можно найти различными способами, например, с помощью нейронной сети или, поскольку задача решается на плоскости, то подбором поворота вектора вокруг начала координат.

Если принять $y_2 = 1$, то получим систему неравенств:

$$y_4 < \frac{0.27}{0.43},\ y_4 > -\frac{0.87}{0.39},\ y_4 < \frac{0.07}{0.38},\ y_4 < \frac{0.06}{0.45},\ y_4 < \frac{0.36}{0.31},$$

откуда следует $-2.23 < y_4 < 0$. Таким образом, имеет место бесконечное число решений, например, системе (14) удовлетворяет вектор: $1 \cdot y_2 - 1 \cdot y_4 = 0$ и соответственно пара коэффициентов ($\beta_2 = 1$, $\beta_4 = -1$) разделяющей линии позволяет решить задачу классификации векторов на плоскости с ординатами (y_2, y_4).

5. Отобразим разделяющую линию пространства R^2 в гиперплоскость пространства R^6.

Для этого сформируем расширенную таблицу:

Табл. 5

Векторы в новой системе	y_1	y_2	y_3	y_4	y_5	y_6	Класс
V_1	-0.03	0.27	-0.01	-0.43	-0.02	-0.85	1
V_2	-0.16	0.87	-0.06	0.39	0.18	0.08	1
V_3	-0.32	0.07	0.75	-0.38	0.34	0.2	1
V	β_1	$\beta_2 = 1$	β_3	$\beta_4 = -1$	β_5	β_6	Δ
V_4	0.64	-0.06	0.47	0.45	0.24	-0.29	2
V_5	-0.45	-0.36	-0.22	0.31	0.66	-0.26	2

Используя описанный способ заполнения позиций Табл. 5 в соответствии с формулой (9), получим Табл.6.

Табл. 6

Векторы в новой системе	y_1	y_2	y_3	y_4	y_5	y_6	Класс
V_1	-0.03	0.27	-0.01	-0.43	-0.02	-0.85	1
V_2	-0.16	0.87	-0.06	0.39	0.18	0.08	1
V_3	-0.32	0.07	0.75	-0.38	0.34	0.2	1
V	-0.400	*1*	0.380	*-1*	0.005	-0.316	Δ
V_4	0.64	-0.06	0.47	0.45	0.24	-0.29	2
V_5	-0.45	-0.36	-0.22	0.31	0.66	-0.26	2

Восстановленная гиперплоскость имеет следующий вид:

$$-0.400\,y_1+1\,y_2+0.380\,y_3-1\,y_4+0.005\,y_5-0.316\,y_6=0$$

Проверка показывает, что гиперплоскость является разделяющей для векторов Табл. 4.

6. Для построения гиперплоскости в системе признаков X необходимо выполнить преобразование (10).

Найдем обратную матрицу:

$$A_*^{-1}=(C-L)=$$

$$=\begin{pmatrix} 6.1561 & 10.3571 & -16.9683 & 8.8350 & -0.8843 & -2.2716 \\ 10.3571 & 15.8928 & -27.3984 & 13.0113 & -0.1786 & 0.2837 \\ -16.9683 & -27.3984 & 44.4382 & -22.1962 & 2.5425 & 0.5916 \\ 8.8350 & 13.0113 & -22.1962 & 9.9664 & -0.2157 & 0.4853 \\ -0.8843 & -0.1786 & 2.5425 & -0.2157 & 0.1861 & 0.0178 \\ -2.2716 & 0.2837 & 0.5916 & 0.4853 & 0.0178 & -4.6314 \end{pmatrix}, \quad (15)$$

где:

$$L=\begin{pmatrix} \lambda_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda_4 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda_5 & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda_6 \end{pmatrix}=\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2.96 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1.18 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2.45 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.29 \end{pmatrix} \quad (16)$$

Восстановим вектор коэффициентов разделяющей функции $\alpha_1,...,\alpha_6$ в системе исходных признаков $X=(x_1,...,x_6)$ в соответствии с (10) и Табл.6:

Таким образом, имеем восстановленную гиперплоскость:

$$-6.682\,x_1-11.776\,x_2+18.318\,x_3-9.088\,x_4+$$
$$+1.353\,x_5+2.398\,x_6=0.$$

Проверка показывает, что данная гиперплоскость не является разделяющей для векторов Табл. 3.

Из этого факта следует, что предложенный метод варьирования размерности пространства признаков и восстановления коэффициентов, к сожалению, не гарантирует получения разделяющей гиперплоскости в системе исходных признаков ровно так же, как его не гарантирует и подход, предложенный в работе [3], основанный на подборе коэффициентов. Введение в уравнение гиперплоскости отрицательного смещения:

$$-6.682\,x_1-11.776\,x_2+18.318\,x_3-9.088\,x_4+$$
$$+1.353\,x_5+2.398\,x_6-6.4=0$$

позволяет классифицировать правильно четыре вектора, при этом неверно классифицируется вектор W_1.

4. Дополнительные исследования

Рассмотрим эффективность применение метрик Евклида, Махаланобиса и Евклида – Махаланобиса [8] для решения задачи бинарной классификации. Расстояние Махаланобиса вычисляется как

$$R_M^2(x,Y)=\left(x-\overline{y}\right)^T C_Y^{-1}\left(x-\overline{y}\right),$$

где $\overline{y}$ – среднее выборочное класса Y, а C_Y^{-1} – матрица, обратная корреляционной матрице C_Y для класса Y, x – рассматриваемый образец, представленный вектором признаков.

Эта метрика обладает известным недостатком: она не применима, если выборочная дисперсия хотя бы одного из параметров равна нулю. Поэтому для решения задачи классификации уместнее применять обобщенную метрику

$$\begin{pmatrix} 6.1561 & 10.3571 & -16.9683 & 8.8350 & -0.8843 & -2.2716 \\ 10.3571 & 15.8928 & -27.3984 & 13.0113 & -0.1786 & 0.2837 \\ -16.9683 & -27.3984 & 44.4382 & -22.1962 & 2.5425 & 0.5916 \\ 8.8350 & 13.0113 & -22.1962 & 9.9664 & -0.2157 & 0.4853 \\ -0.8843 & -0.1786 & 2.5425 & -0.2157 & 0.1861 & 0.0178 \\ -2.2716 & 0.2837 & 0.5916 & 0.4853 & 0.0178 & -4.6314 \end{pmatrix}\begin{pmatrix} -0.400 \\ 1.000 \\ 0.380 \\ -1.000 \\ 0.005 \\ -0.316 \end{pmatrix}=\begin{pmatrix} -6.682 \\ -11.776 \\ 18.318 \\ -9.088 \\ 1.353 \\ 2.398 \end{pmatrix}$$

Евклида-Махаланобиса [8], которая определяет расстояние между двумя классами X_1 и X_2 в форме

$$R^2_{E-M}(X_1, X_2) = (\bar{x}_1 - \bar{x}_2)^T A^{-1}(\bar{x}_1 - \bar{x}_2),$$

где $\bar{x}_1$ и $\bar{x}_2$ – средние выборочные классов, $A = (C_1 + C_2 + E)$, C_1 и C_2 – ковариационные матрицы для классов X_1 и X_2 соответственно.

Для определения расстояния между вектором x и классом Y используется формула

$$R^2_{E-M}(x, Y) = (x - \bar{y})^T A^{-1}(x - \bar{y}),$$

где $A = C_Y + E$.

Найдем матрицы ковариаций (17)-(18), но теперь раздельно для исходных векторов первого (три первых вектора) и второго классов (два последних вектора).

Матрица ковариаций для класса Ω_1:

$$C_1 = \begin{pmatrix} 1.33 & 0.33 & 1.33 & 1.33 & 0.67 & -0.67 \\ 0.33 & 2.33 & -1.17 & 1.83 & 0.17 & -0.17 \\ 1.33 & -1.17 & 2.33 & 0.33 & 0.67 & -0.67 \\ 1.33 & 1.83 & 0.33 & 2.33 & 0.67 & -0.67 \\ 0.67 & 0.17 & 0.67 & 0.67 & 0.33 & -0.33 \\ -0.67 & -0.17 & -0.67 & -0.67 & -0.33 & 0.33 \end{pmatrix} \quad (17)$$

Матрица ковариаций для класса Ω_2:

$$C_2 = \begin{pmatrix} 2.00 & -3.00 & -1.00 & 0.00 & -1.00 & -1.00 \\ -3.00 & 4.50 & 1.50 & 0.00 & 1.50 & 1.50 \\ 1.00 & 1.50 & 0.50 & 0.00 & 0.50 & 0.50 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ -1.00 & 1.50 & 0.50 & 0.00 & 0.50 & 0.50 \\ -1.00 & 1.50 & 0.50 & 0.00 & 0.50 & 0.50 \end{pmatrix} \quad (18)$$

Результаты вычисления расстояний для векторов $W_1, \ldots, W_5$ с применением различных метрик приведены в Табл. 7.

Как видно по данным Табл.7, результат классификации зависит от используемой метрики и не является однозначным. В данном случае только расстояние Евклида-Махаланобиса позволяет правильно решить задачу классификации, что обосновывает целесообразность его выбора. Уточненные выводы можно будет сделать после проведения экспериментов с реальными данными в конкретных предметных областях.

Заключение

Выполненные исследования показали, что построение разделяющей гиперплоскости в задаче бинарной классификации путем перехода к системе с существенно меньшим числом признаков и последующим восстановлением размерности возможно путем применения ряда эффективных методов, включая МГК, МНК и ZET. Обратное преобразование служит для удобства работы пользователя с признаками, имеющими физический смысл, но не всегда приводит к получению необходимого решения. Показано, что введение смещения в уравнение гиперплоскости, как правило, улучшает результаты бинарной классификации. Большими возможностями обладают нейронные сети, которые в отличие от рассмотренных подходов восстановления решений обеспечивают универсальность процедуры построения гиперплоскости. Проведенные эксперименты по классификации векторов продемонстрировали определенное преимущество обобщенной мет-

Табл. 7

Исходные векторы	Расстояние Евклида		Расстояние Махаланобиса		Расстояние Евклида-Махаланобиса	
	до класса Ω_1	до класса Ω_2	до класса Ω_1	до класса Ω_2	до класса Ω_1	до класса Ω_2
W_1	6.3333	7	36	3.3750	1.1000	6.5000
W_2	6.3333	10	0	0.3750	1.1000	10.000
W_3	5.3333	5	28	3.6250	1.0667	4.1111
W_4	5.6667	4	243269472	7.6250	5.1417	0.4444
W_5	5	4	276824000	7.6250	3.9333	0.4444

рики Евклида-Махаланобиса, обладающей рядом положительных свойств [9]. Выполненные исследования и представленные выводы по качеству предложенного метода бинарной классификации на основе варьирования размерности пространства признаков не носят окончательного характера и требуют проведения расширенных экспериментов в конкретных приложениях.

Литература

1. Метод опорных векторов. – http://ru.wikipedia.org/wiki/SVM
2. Дубров А.М. Обработка статистических данных методом главных компонент. М.: Финансы и статистика, 1978. -135 с.
3. Саутин С.Н., Пунин А.Е., Савкович-Стеванович Е. Методы искусственного интеллекта в химии и химической технологии. – Л.: Издательство ЛТИ, 1989. – 96 с.
4. Журавлев Ю.И. Об алгебраических методах в задачах распознавания и классификации // Распознавание. Классификация. Прогноз. Математические методы и их применение. Вып. 1. – М.: Наука. 1989, с. 9-16.
5. Журавлев Ю.И., Гуревич И.Б. Распознавание образов и распознавание изображений // Распознавание, классификация, прогноз. Математические методы и их применение. Вып. 2. – М.: Наука, 1989, с. 5-72.
6. Фраленко В.П., Хачумов М.В. Классификация на основе аппарата нейронных сетей с применением метода главных компонент и комитета большинства.- В сб. статей Третьей Всероссийской научной конференции "Нечеткие системы и мягкие вычисления" НСМВ-2009 (Волгоград, 21-24 сентября 2009 г.). – Волгоград: Волгоградский государственный технический университет, т.2, 2009, с. 70-79.
7. Загоруйко Н.Н. Прикладные методы анализа данных и знаний. – Новосибирск: Изд-во Института математики, 1999. – 270 с.
8. Амелькин С.А., Хачумов В.М. Обобщенное расстояние Евклида-Махаланобиса и его применение в задачах распознавания образов. – Доклады 12-ой Всероссийской конференции «Математические методы распознавания образов». – М.: МАКС Пресс, 2005, с. 7-9.
9. Амелькин С.А., Захаров А.В., Хачумов В.М. Обобщенное расстояние Евклида -Махаланобиса и его свойства. – Информационные технологии и вычислительные системы, № 4, 2006, с.40-44.

Толмачев Игорь Леонидович. Заведующий кафедрой Российского университета дружбы народов (РУДН). Окончил в 1964 году Московский государственный университет. Кандидат физико-математических наук, профессор. Автор 88 печатных работ (из них 1 монография). Область научных интересов: информатика и прикладная математика. E-mail: tolmachevil@mail.ru

Хачумов Михаил Вячеславович. Аспирант РУДН. Окончил в 2009 году магистратуру Российского университета дружбы народов. Автор 6 печатных работ. Область научных интересов: искусственный интеллект, машинная графика, кластеризация. E-mail: khmike@inbox.ru