УДК 519.95

НЕЛИНЕЙНОЕ ОТОБРАЖЕНИЕ НАБОРОВ ПРИЗНАКОВ НА ЧИСЛОВУЮ ОСЬ И ДИСКРИМИНАНТНЫЙ АНАЛИЗ ДАННЫХ

Саидов Д.Ю.
старший научный сотрудник-исследователь Национального университета Узбекистана имени Мирзо Улугбека,
тел.: +(99890) 973-70-66, e-mail: doniyor_2286@mail.ru

Рассматривается алгоритм нелинейного отображения описания объектов на числовую ось с использованием иерархической агломеративной группировки. Из каждой группы исходных признаков формируется один латентный признак. Латентные признаки могут применяться для распознавания как по прецедентам, так и по правилам. Для получения аналитического представления правил используются методы формальной грамматики. Приводится сравнительный анализ использования результатов отображения по точности распознавания с линейным дискриминантом Фишера.

Ключевые слова: латентные признаки, линейный дискриминант Фишера, нелинейное отображение, информативный признак, пороговое значение.

NONLINEAR MAPPING FEATURE SET ON THE NUMERIC AX AND THE DISCRIMINANT ANALYSIS DATA

Saidov D.Y.

An algorithm of nonlinear mapping the description of the objects in the numerical axis using a hierarchical agglomerative grouping is considered. From each group of initial features one latent feature is formed. Latent features may be used to recognize both by precedents and the rules. The methods of formal grammar had used to obtain an analytical representation of the rules. A comparative analysis of using the mapping results on the accuracy of recognition with linear discriminant of Fisher is provided.

Keywords: latent features, linear discriminant of Fisher, nonlinear mapping, informative feature, the threshold value.

ALOMATLAR TO'PLAMINI SON O'QIDA NOCHIZIQLI AKSLANTIRISH VA BERILGANLARNING DISKRIMINANT TAHLILI

Saidov D.Y.

Ierarxik aglomerativ guruhlashdan foydalangan holda obyektlarni son o'qida tasvirlashning nochiziqli akslantirish algoritmi qaraladi. Dastlabki alomatlarning har bir guruhidan bitta latent alomat shakllanadi. Latent alomatlar ham pretsedent bo'yicha, ham qoida bo'yicha anglash uchun qo'llaniladi. Formal grammatika usullaridan qoidalarning analitik tasvirlanishini olish uchun foydalanilgan. Anglash aniqligi bo'yicha akslantirish natijalaridan foydalanish bilan Fisher chiziqli diskriminanti qiyosiy tahlili keltirilgan.

Tayanch iboralar: latent alomatlar, Fisher chiziqli diskriminanti, nochiziqli akslantirish, informativ alomat, ostona qiymat.

1. Введение

Отображение наборов признаков в описании объектов на числовые оси широко используется в анализе данных. По большей части эти отображения представляют линейные проекции. К числу таковых относятся методы главных компонент (PCA) [1], обобщённых оценок [2], линейный дискриминант Фишера [3] и т.д.

Нелинейные отображения можно реализовать через использование обобщенных функций от n исходных значений признаков. Например, для обобщенных функций степени не выше 2 (квадратичных) размерность нового пространства равна $\frac{n(n+1)}{2}$.

Расширенное за счёт этих функций пространство считается линейным. Проблемы, с которыми приходится сталкиваться при этом:

- нет критериев выбора обобщенных функций;
- в расширенном признаковом пространстве использование многих решающих функций становится не эффективным (проклятие размерности Белмана);
- поиск наборов информативных признаков труднореализуем из-за комбинаторной сложности алгоритмов отбора;

- численные результаты теряют практический смысл из-за размеров вычислительной и машинной погрешности.

В работе [4] впервые сделана попытка использования нелинейных преобразований (отображений) для случаев, когда аналитический вид (класс функций) заранее не известен. Определены правила формирования латентных признаков на основе агломеративной иерархической группировки исходных (сырых) признаков. Результаты группировки использовались для аналитического представления таких функций [5] и обоснования обобщающей способности алгоритма ближайшего соседа.

В настоящее время нет единого мнения насчёт использования методов проверки обобщающей способности алгоритмов распознавания [6]. В качестве таковых в порядке убывания значимости имеют место: метод скользящего экзамена (Cross Validation); отступ между объектами классов; характеристики компактности при минимальном покрытии выборки объектами – эталонами [7].

В статье затрагиваются вопросы нелинейного отображения исходных (сырых) признаков в описании объектов на числовую ось. Предлагается использовать результаты отображения для распознавания принадлежности к классам как по правилам, так и по прецедентам. Эффект, получаемый при линейном и нелинейном отображениях, демонстрируется в форме сравнительного анализа предлагаемого алгоритма и линейного дискриминанта Фишера. При сравнении используются такие понятия, как порог и отступ между классами, информативность наборов и отдельных признаков.

2. Постановка задачи. Нелинейное отображение на числовую ось

Рассматривается двухклассовая задача распознавания в стандартной постановке. Объекты выборки $E_0 = \{S_1,\ldots,S_m\}$ принадлежат к одному из классов K_1 или K_2 $(E_0 = K_1 \bigcup K_2)$ и описываются с помощью n количественных признаков $X(n) = (x_1,\ldots,x_n)$. На E_0 задано последовательное разбиение набора $X(n)$ по правилу иерархической агломеративной группировки на непересекающиеся подмножества $X_1(k_1),\ldots,X_\tau(k_\tau), \tau \geq 1, k_1+\ldots+k_\tau \leq n$. Каждое подмножество $X_1(k_t), t = 1,\ldots,\tau$ отображается на числовую ось и рассматривается как новый латентный признак в описании объектов. Требуется:

- сформировать новое пространство из латентных признаков в описании объектов;

- сравнить точность алгоритмов распознавания при линейном и нелинейном отображении с использованием линейного дискриминанта Фишера на определяемых значениях порогов и наборах признаков.

В описании алгоритма нелинейного отображения признаков на числовую ось используются символьные обозначения из [8]. Для идентификации признаков как исходных, так и полученных при вычислении обобщённых оценок на p-м шаге $(0 \leq \mathrm{p} < n)$ иерархической агломеративной группировки, будем использовать $\left\{\mathrm{x}_\mathrm{i}^\mathrm{p}\right\}_{\mathrm{i}=1}^{\mathrm{n}-\mathrm{p}}$. Множество номеров количественных признаков будем обозначать как I. При $\mathrm{p}=0$ число групп совпадает с числом признаков в $\mathrm{X(n)}$ и $\mathrm{I}=\{1,\ldots,\mathrm{n}\}$.

В процессе группировки и формирования обобщённых оценок состав элементов и мощность множества I, $|I| \leq n$ будут изменяться.

Упорядоченное множество значений признака $x_j^p, j \in I, p \geq 0$ объектов из E_0 разделим на два интервала $[c_1^{jp}, c_2^{jp}]$, $(c_2^{jp}, c_3^{jp}]$, каждый из которых рассматривается как градация номинального признака. Критерий для определения границы c_2^{jp} основывается на проверке гипотезы (утверждения) о том, что каждый из двух интервалов содержит значения количественного признака объектов только одного класса.

Пусть u_i^1, u_i^2 – количество значений признака x_j^p, $j \in I$ класса $K_i, i = 1,2$ соответственно в интервалах $[c_1^{jp}, c_2^{jp}]$, $(c_2^{jp}, c_3^{jp}]$; $|K_i| > 1$, ϑ – порядковый номер элемента упорядоченной по возрастанию последовательности $r_{j1},\ldots,r_{jv},\ldots,r_{jm}$ значений x_j^p из E_0, определяющий границы интервалов как $c_1^{jp} = r_{j1}$, $c_2^{jp} = r_{jv}$, $c_3^{jp} = r_{jm}$. Критерий

$$\left(\frac{\sum_{d=1}^{2}\sum_{i=1}^{2} u_i^d\left(u_i^d - 1\right)}{\sum_{i=1}^{2}|K_i|\left(|K_i|-1\right)}\right) \times \left(\frac{\sum_{d=1}^{2}\sum_{i=1}^{2} u_i^d\left(|K_{3-i}| - u_{3-i}^d\right)}{2|K_1||K_2|}\right) \to \max_{c_1<c_2<c_3}. \quad (1)$$

позволяет вычислять оптимальное значение границы между интервалами $[c_1^{jp}, c_2^{jp}]$, $(c_2^{jp}, c_3^{jp}]$. Выражение в левых скобках (1) представляет внутриклассовое сходство, в правых – межклассовое различие.

Экстремум критерия (1) используется в качестве веса $\omega_j^p \left(0 \leq \omega_j^p \leq 1\right)$ признака x_j^p. При $\omega_j^p = 1$ значения признака x_j^p у объектов из классов K_1, K_2 не пересекаются между собой. Значение обобщённой оценки b_{rij}^p объекта $S_r = (a_{r1}^p,\ldots,a_{r,n-p}^p)$, $S_r \in E_0$ по паре $\left(x_i^p, x_j^p\right)$, $1 \leq p < n, i, j \in I, i \neq j$ вычисляется как

$$b_{rij}^p = \mu_{ij}\left(\begin{array}{l} t_i\omega_i^p\left(\dfrac{a_{ri}^p - c_2^{ip}}{c_3^{ip} - c_1^{ip}}\right) + \\ +t_j\omega_j^p\left(\dfrac{a_{rj}^p - c_2^{jp}}{c_3^{jp} - c_1^{jp}}\right) \end{array}\right) + \\ +\left(1-\mu_{ij}\right)t_{ij}\omega_{ij}^p\left(\frac{a_{ri}^p a_{rj}^p - c_2^{ijp}}{c_3^{ijp} - c_1^{ijp}}\right), \quad (2)$$

где $\omega_i^p, \omega_j^p, \omega_{ij}^p$ – веса признаков, определяемые по (1) соответственно по множеству значений признаков x_i^p, x_j^p и их произведения $x_i^p x_j^p$, значения $t_i, t_j, t_{ij} \in \{-1,1\}$, $\mu_{ij} \in [0,1]$ выбираются по экстремуму функционала

$$\varphi(p,i,j) = \frac{\min\limits_{S_r \in K_1} b_{rij}^p - \max\limits_{S_r \in K_2} b_{rij}^p}{\max\limits_{S_r \in E_0} b_{rij}^p - \min\limits_{S_r \in E_0} b_{rij}^p} \to \max. \quad (3)$$

Значение (3) интерпретируется как отступ между объектами классов K_1 и K_2.

Обозначим через $\left\{z_{ij}^p\right\}_{i,j=1}^{n-p}, p \ge 0$ матрицу, значения элемента z_{ij}^p которой определяются как

$$z_{ij}^p = \begin{cases} 0,\ i = j, \\ значению\,(1)\ на \left\{b_{rij}^p\right\}_{r=1}^m,\ i \ne j, \end{cases} \quad (4)$$

через $G_\mu, \mu > 0$ – подмножество номеров признаков из $X(n)$.

Пошаговая реализация алгоритма итеративной группировки следующая.

1 шаг. $p = 0$, $\lambda c = 0$, $\eta = 1$. **Выполнять** $\Gamma_\eta = \{\eta\}$, $Margin_\eta = -2$, $\eta = \eta + 1$, **пока** $\eta \le n$.

2 шаг. Вычислить значения элементов матрицы $\left\{z_{ij}^p\right\}_{i,j \in I}$ по (4).

3 шаг. Выделить

$$\Phi = \left\{ z_{uv}^p \,\middle|\, z_{uv}^p \ge \max\left(w_u^p, w_v^p\right) and \quad u \ne v, u, v \in I \right\}.$$

Если $\Phi = \varnothing$, то перейти к п. 9.

4 шаг. Вычислить $\lambda n = \max\limits_{z_{uv}^p} z_{uv}^p$. Выделить $\Delta = \left\{ (s,t) s,t \in I \,\middle|\, z_{st}^p = \lambda n \ and \ s < t \right\}$. Определить пару $\{i,j\}, i < j$ как

$$\{i,j\} = \begin{cases} \Delta, |\Delta| = 1, \\ \{s,t\}, (s,t) \in \Delta \ and \ \varphi(\mathrm{p,s,t}) > \\ > \max\limits_{(u,v) \in \Delta \setminus (s,t)} \phi(p,u,v). \end{cases}$$

5 шаг. Если $\lambda n > \lambda c$ или $\lambda c = \lambda n$ и $Margin_i < \varphi(p,i,j)$, $Margin_j < \varphi(p,i,j)$, то $\Gamma_i = \Gamma_i \cup \Gamma_j$, $\Gamma_j = \varnothing$ $Margin_i = \varphi(p,i,j)$, перейти к п. 7.

6 шаг. Вывод номеров признаков из Γ_i, $\Gamma_i = \varnothing$, $I = I \setminus \{i\}$, перейти к п. 3.

7 шаг. $p = p+1$, $I = I \setminus \max(i,j)$, $k = \min(i,j)$, $\lambda c = \lambda n$. Заменить значения признаков в описании объекта $S_r = \left\{a_{ru}^{p-1}\right\}_{u \in I}$, $r = 1,...,m$ на

$$a_{ru}^p = \begin{cases} a_{ru}^{p-1}, u \in I \setminus k, \\ b_{rij}^p, u = k. \end{cases}$$

8 шаг. Для каждой пары (u,v), $u,v \in I$ определить значение

$$z_{uv}^p = \begin{cases} z_{uv}^{p-1}, u \in I \setminus \{k\}, v \in I, \\ значению \quad (1) \quad на \quad \left\{a_{rv}^p\right\}_{r=1}^m, u = k, v \in I. \end{cases}$$

Если $n - p > 1$, то перейти к п. 3.

9 шаг. Конец.

Значения латентных признаков образуют новое пространство для описания объектов в алгоритмах распознавания по прецедентам. Аналитическое представление нелинейных преобразований, предложенное в [5], позволяет использовать результаты описанного алгоритма в качестве правила для распознавания.

3. О выборе порога в линейных решающих функциях

Для точности распознавания выбор порога правила играет определяющую роль. В линейном дискриминанте Фишера [3]

$$d(x) = w_1 x_1 + w_2 x_2 + \ldots + w_n x_n + w_0$$

решение о принадлежности объекта к классам K_1, K_2 принимается на основе правила: $x \in K_1$, если $d(x) > 0$, и $x \in K_2$, если $d(x) < 0$. В качестве значения порога выбирается

$$w_0 = -\sum_{i=1}^n m_i w_i, \quad (5)$$

где m_i – элемент вектора математических ожиданий по признаку $x_i \in X(n)$, вычисляемый на E_0.

Другой способ вычисления порога описывается в [9]. Для сокращения размерности разнотипного признакового пространства авторы предлагают использовать линейное отображение всех номинальных признаков в один количественный. Для линейной проекции количественных признаков в описании объектов на числовую ось используется разбиение на интервалы $[c_1, c_2]$, $(c_2, c_3]$ по критерию (1), а значение порога вычисляется как

$$w_0 = \frac{c_2 + z}{2}, \quad (6)$$

где z - ближайшее к c_2 значение из интервала $(c_2, c_3]$.

Вычислительный эксперимент проводился на медицинских данных [8] с показателями гипертонической болезни. Выборка из 147 объектов

была разделена на два класса: K_1 (здоровые) содержал показатели 111 объектов, K_2 (больные) – 36 объектов. Каждый объект описывается 29-ю признаками. В табл. 1 приведены результаты последовательного объединения по (2) и (3) исходных признаков в группы с учётом вложенности скобок.

Таблица 1

Результаты группировки признаков

Номер группы	Последовательность объединения признаков	Значение критерия (1)	Отступ по (3)
1	$(((((((x_4, x_{20}), x_9), x_{18}), x_8), x_2), x_{10}), x_{12})$	1,0000	0,0105
2	$((((((x_5, x_{14}), x_{16}), x_{19}), x_{23}), x_3), x_7)$	1,0000	0,0086
3	$((((((((((x_{26}, x_{28}), x_{27}), x_{15}), x_1), x_{25}), x_{13}), x_{21}), x_{22}), x_{24}), x_{11}), x_{17})$	0,9672	-0,0042
4	(x_6, x_{29})	0,7331	-0,2380

Согласно правилам формальной грамматики, используемой в [5, 10], аналитическое представление группы № 1 из табл. 1 такое:

$$x_4^1 = -0.0069(АДС - 140) + 0.3949(ДИАСТОЛА - 0.42) - -0.00413(АДС * ДИАСТОЛА - 68.2);$$

$$x_4^2 = 0.8013(-0.0094) + 3.1287(QRS - 0.08);$$

$$x_4^3 = 0.5811(x_4^2) + 0.5518(\text{СИСПОК} - 0.485) + + 0.8920(x_4^2 * \text{СИСПОК});$$

$$x_4^4 = 0.2117(x_4^3 + 0,0388) + 0.2756(QT - 0.36) + + 2.1363(x_4^3 * QT + 0.0124);$$

$$x_2^5 = 0.0057(x_4^4 * \text{РОСТ} + 1.9623);$$

$$x_2^6 = 0.3107(x_2^5 * ПЛП);$$

$$x_2^7 = 0.1804(x_2^6 * \text{КДР}).$$

В [2] показано, что для точности линейного разделения основную роль играют веса по (1). Порядок следования весов признаков по (1) можно проследить в табл. 2.

Согласно табл. 1, состав группы 1 и набор $\{x_4, x_{14}, x_{15}, x_5, x_{10}, x_1, x_{28}, x_{22}\}$ из табл.2 пересекаются только по двум признакам $\{x_4, x_{10}\}$.

В качестве одной из характеристик линейного разделения так же, как и в [2], предлагается использовать отступ между классами, определяемый как

$$\min_{x \in K_1} d(x) - \max_{x \in K_2} d(x). \qquad (7)$$

Таблица 2

Порядок следования весов признаков по (1)

№	Признаки	Значение признаков по критерию 1
1	x_4	0,9031
2	x_{14}	0,8890
3	x_{15}	0,8096
4	x_5	0,7515
5	x_{10}	0,6268
6	x_1	0,6066
7	x_{28}	0,5336
8	x_{22}	0,5251
-	-	-
-	-	-
22	x_{19}	0,2756
23	x_{20}	0,2685
24	x_7	0,2674
25	x_6	0,2529
26	x_{13}	0,2529
27	x_{26}	0,2526
28	x_2	0,2526
29	x_9	0,2503

Анализ влияния выбора порогов по (5) и (6) на указанных выше наборах из табл. 1 и 2 приводится в табл. 3.

Таблица 3

Точность в % при значении порогов по (5) и (6)

№ п/п	Набор признаков	Порог		Отступ по (7)
		по (5)	по (6)	
1	$x_2, x_4, x_8, x_9, x_{10}, x_{12}, x_{18}, x_{20}$	91,84%	98,64%	-0,03521
2	$x_1, x_4, x_5, x_{10}, x_{14}, x_{15}, x_{22}, x_{28}$	91,84%	98,64%	-0,02938

Согласно табл. 3, точность распознавания при использовании порогов по (5) и (6) совпадает на двух наборах, взятых из табл. 1 и 2, а различие происходит по значению отступа. В пересечение этих наборов входят признаки $\{x_4, x_{10}\}$ с высоким уровнем информативности (см. табл. 2). О роли набора $\{x_4, x_{10}\}$ и восьми наименее информативных признаков из табл. 2 можно судить по точности распознавания на них из табл. 4.

Таблица 4

Точность распознавания с учетом пересечения и информативности

№ п/п	Набор признаков	Порог		Отступ по (7)
		по (5)	по (6)	
1	x_4, x_{10}	91,84%	97,96%	-0,03332
2	$x_2, x_6, x_7, x_9, x_{13}, x_{19}, x_{20}, x_{26}$	60,54%	72,79%	-0,01600

Значения 91,84% (97,96%) на $\{x_4, x_{10}\}$ из табл. 4 указывают на его (подмножества) определяющую роль при вычислении точности распознавания на наборах из табл. 3. Также подтверждается истинность гипотезы о влиянии значений информативности каждого признака из табл. 2 на точность распознавания.

Литература

[1] *Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика. – М.: Финансы и статистика, 1989. – 608 с.

[2] *Игнатьев Н.А.* Вычисление обобщённых показателей и интеллектуальный анализ данных // Автоматика и телемеханика. – 2011. – № 5. – С. 183-190.

[3] *Дуда Р., Харт П.* Распознавание образов и анализ сцен. – М.: Мир, 1976. – 512 с.

[4] *Игнатьев Н.А.* Вычисление обобщенных оценок и иерархическая группировка признаков // Вестник Томского государственного университета. – Томск, 2015. - С. 31-38.

[5] *Саидов Д.Ю.* Нелинейные преобразования признакового пространства и их аналитические представления // Международный молодежный научный форум «ЛОМОНОСОВ-2015». – 2015.

[6] *Воронцов К.В.* Обзор современных методов по проблеме качества обучения алгоритмов // Таврический вестник информатики и математики. – 2004. – № 1. – С. 5-25.

[7] *Загоруйко Н.Г., Кутненко О.А., Зырянов А.О., Леванов Д.А*. Обучение распознаванию образов без переобучения // Машинное обучение и анализ данных. – 2014 . – Т. 17. – С. 891-901.

[8] *Игнатьев Н.А*. Кластерный анализ данных и выбор объектов-эталонов в задачах распознавания с учителем // Вычислительные технологии. – 2015. – Т. 20, № 6. – С. 34-43.

[9] *Игнатьев Н.А., Нуржонов Ш.Ю.* Выбор параметров регуляризации для повышения обобщающей способности дискриминантных функций // Ўзбекистон Республикаси Курол Кучлари академиясининг хабарлари. – 2014. – № 1(14). – С. 81-87.

[10] *Саидов Д.Ю.* Аналитическое представление распознающих операторов для вычисления обобщённых оценок // ЎзМУ ХАБАРЛАРИ. – 2015. – № 2/1. – С. 121-125.