

А.Б. Исаев, А. Алнадфа

ИТЕРАЦИОННО-ВЗВЕШЕННЫЙ АЛГОРИТМ ПОСТРОЕНИЯ ЛИНЕЙНОЙ РАЗДЕЛЯЮЩЕЙ ФУНКЦИИ РЕГРЕССИИ В ЗАДАЧЕ КЛАССИФИКАЦИИ ОБРАЗОВ, ПРИЗНАКИ КОТОРЫХ ОТЯГОЩЕНЫ ПОГРЕШНОСТЯМИ ИЗМЕРЕНИЙ

Математически строго сформулирована и решена задача о наилучшем проведении прямой, разделяющей два класса образов, представляющих собой два конечных, дискретных и счетных множества, выпуклые оболочки которых не пересекаются. Для нахождения состоятельных оценок параметров разделяющей линии регрессии сконструирован итерационный алгоритм, получаемый из функции максимального правдоподобия, в предположении нормального распределения погрешностей измерений координат признаков предъявляемых двумерных образов в условиях конфлюентной ситуации.

Ключевые слова: истинные значения переменных, конфлюентный анализ, погрешности измерений, выпуклая оболочка, функция максимального правдоподобия, образ, признаки образов.

Основой создания алгоритмом идентификации образов в классе линейных решающих правил являются алгоритмы построения гиперплоскости, разделяющей два класса образов w_1 и w_2 .

Известно, что процедура классификации образов может производиться как с учетом погрешностей, так и без учета погрешностей измерений их признаков предъявляемых образов. Однако, если процедура классификации образов не учитывает погрешности измерений их признаков, даже в том случае, если предполагается известный вид разделяющих (решающих) функций (поверхностей), оценки их аналитических параметров, полученные без учета погрешностей измерений их признаков, становятся не точными, т.е. сама оценка разделяющих функций (поверхности) становится не точной, смещенной, приводя к ошибкам идентификации вследствие увеличения площади зон неопределенности из-за появления случайных и систематических погрешностей у разделяющих функций [1].

В данной статье предложен алгоритм построения разделяющей линии в форме линейного уравнения регрессии $\hat{Y} = b_0 + b_1 X + \varepsilon$, состоящий из нескольких возможных вариантов постановки задачи классификации:

Вариант I. Результаты измерений признаков не содержат погрешностей измерений (детерминированный случай).

Вариант II. Результаты измерений признаков содержат погрешности измерений только в одном признаке –

$$Y: y_i^1 = \eta_i^1 + \varepsilon_i; x_i^1 = \xi_i^1; y_i^2 = \eta_i^2 + \varepsilon_i; x_i^2 = \xi_i^2 (i = \overline{1, N});$$

Вариант III. Результаты измерений признаков содержат погрешности измерений только в одном признаке –

$$X: y_i^1 = \eta_i^1; x_i^1 = \xi_i^1 + \delta_i; y_i^2 = \eta_i^2; x_i^2 = \xi_i^2 + \delta_i (i = \overline{1, N});$$

Вариант IV. Конфлюентная ситуация: присутствуют случайные погрешности измерений в результатах измерений координат как признака X , так и признака:

$Y : y_i^1 = \eta_i^1 + \varepsilon_i; x_i^1 = \xi_i^1 + \delta_i; y_i^2 = \eta_i^2 + \varepsilon_i; x_i^2 = \xi_i^2 + \delta_i (i = \overline{1, N});$ – нами рассмотрен вариант IV, как наиболее общий.

Заметим, что во всех четырех вариантах случайные погрешности измерений ε и δ распределены по нормальным законам:

$$\varepsilon \sim N(0; \sigma_{yi}^2 = D\varepsilon_i^2), i = \overline{1, N}$$

$$\delta \sim N(0; \sigma_{xi}^2 = D\delta_i^2), i = \overline{1, N}$$

Предстоит оценить параметры регрессионной линейной зависимости:

$$M(y / x) = \tilde{Y} = b_0 + b_1 X + \varepsilon \quad (1).$$

Вопросы построения оценок параметров для регрессии (1) $\sim \tilde{b}_0, \tilde{b}_1 \sim$ таких, которые будут сохранять, по крайней мере, свои оптимальные свойства асимптотической несмещенности, эффективности или состоятельности при нарушении каких-либо основных предпосылок классического регрессионного анализа [2] на сегодняшний день чрезвычайно актуальны, особенно для задач классификации объектов различной природы.

При выполнении условий варианта IV (конфлюентная ситуация вариант IV) мы попадаем в область «робастного» оценивания, т.е. возникают проблемы построения таких оценок для параметров $\tilde{b}_0, \tilde{b}_1$ регрессии (1), которые будут мало чувствительны к «засорению» выборки экспериментальных данных $\{x_i, y_i\}$ погрешностями результатов измерений типа «выбросов» (или «промахов») или, наконец, к наличию в выборке «чужих» распределений, представленных соответствующими наблюдениями. В результате, распределение выборки может быть представлено в виде смеси основного Гауссова распределения $N(x)$ с другим (засоряющим) распределением $h(x) : P_\varepsilon(x) = (1 - \varepsilon)N(x) + \varepsilon h(x)$, где $\varepsilon \ll 1$, $N(x)$ – плотность нормального распределения, – плотность не гауссовского распределения, $P_\varepsilon(x)$ – результирующее распределение оценок выборки. На этот счет сконструировано большое количество оценок – оценок Андриуса, Хьюбера, L – оценки, M – оценки, и т.д. [1].

В нашей работе рассматриваются и предлагаются оценки (обусловленные нарушением постулата об отсутствии погрешностей измерений признака по оси OX), робастные к присутствию погрешностей в результатах измерения независимой переменной X , т.е. $x_i = \xi_i + \delta_i (i = \overline{1, N}), \delta \in N(0, \sigma_{xi}^2)$. В этом случае, многие оценки теряют свои оптимальные свойства, становятся смещенными, несостоятельными и малоэффективными [1, 2].

Эта экспериментальная ситуация («конфлюентная» ситуация) существенно усложняет задачу построения уравнения регрессии, находящегося по всей области определения в наибольшей близости (например, в смысле евклидовой нормы 2) к неизвестной, так называемой истинной зависимости $\eta(\xi)$ (или же $\xi(\eta)$).

Напомним [1, 2], что традиционные оценки метода наименьших квадратов в конфлюентной ситуации становятся «грубыми», смещенными оценками, и для частичного восстановления их утраченных оптимальных свойств применяются оценки, реализующие принцип максимального правдоподобия на основе модифицированной функции максимального правдоподобия, содержащей ин-

формацию о погрешностях измерений всех факторов – т.е. «независимой» и «зависимой» переменной в той или иной форме.

Алгоритм получения оценок и сами оценки, использованы для построения «разделяющей» линии регрессии $\hat{Y} = b_0 + b_1 X + \varepsilon$ – регрессии, построенной по каждой из выпуклых оболочек $H(w_1)$ или $H(w_2)$, причем таким точкам из разных оболочек (например, $(x_k^1, y_k^1) \in H(w_1)$, $(x_{k+1}^1, y_{k+1}^1) \in H(w_1)$ и $(x_e^2, y_e^2) \in H(w_2)$, $(x_{e+1}^2, y_{e+1}^2) \in H(w_2)$, которые являются «ближайшими соседями» по отношению друг к другу), см. рисунок.

Нами реализован алгоритм на основе модифицированной функции максимального правдоподобия, в предположении нормальности распределения погрешностей координат наблюдений над образами из классов w_1 и w_2 . Перейдем к формализованной записи постановки задачи, предварительно описав содержательную сущность задачи.

Пусть необходимо построить гиперплоскость (линию), разделяющую два множества образов, w_1 и w_2 , которые, являясь дискретными, счетными и конечными множествами, *a priori* не пересекаются $w_1 \cap w_2 = \emptyset$.

Образуем их выпуклые оболочки – $H(w_1)$ и $H(w_2)$, наименьшие выпуклые множества, содержащие все первоначальные точки первоначального множества точек.

Определение. Два точечных (дискретных), счетных ограниченных множества S_1 и S_2 назовем непересекающимися, если не пересекаются их выпуклые оболочки: $S_1 \cap S_2 = \emptyset \Leftrightarrow H(w_1) \cap H(w_2) = \emptyset$.

Будем рассматривать вариант IV (см. выше) постановки задачи классификации: конфлюентная ситуация. Присутствуют погрешности измерений координат признаков Y и X :

$$x_i^j = \xi_i^j + \delta_i \quad (i = \overline{1, N}; j = 1, 2) \quad y_i^j = \eta_i^j + \delta_i \quad (i = \overline{1, N}; j = 1, 2)$$

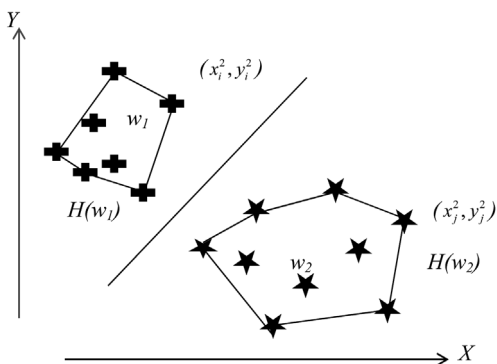
$$\delta_i \sim N(0, \delta_{xi}^2), \quad \varepsilon_i \sim N(0, \delta_{yi}^2);$$

Требуется провести разделяющую линию, которая будет являться классификатором в этом варианте, и на основании которого может быть принято достоверно (в определенном смысле) решение принадлежности предъявляемых к распознаванию объектов к одному из классов: $w_1 \vee w_2$.

Построенная разделяющая линия должна быть оптимальной в определенном смысле. В случае варианта IV нами предлагается алгоритм построения этой линии, основанный на получении оценок ее параметров $\hat{b}_0, \hat{b}_1$ (для неизвестных «истинных» параметров β_0, β_1) с помощью модифицированной функции максимального правдоподобия, в которую введена информация о существующих погрешностях в измерениях координат признаков X и Y :

$$\sigma_{xi}^2 \neq 0 \quad \text{и} \quad \sigma_{yi}^2 \neq 0 \quad (i = \overline{1, N})$$

Построенная в условиях конфлюентной ситуации разделяющая линия регрессии должна иметь по крайней



мере такие оптимальные статические свойства, как состоятельность и асимптотическую несмещенность (или величину смещения меньшую, чем у смещения оценок традиционного метода наименьших квадратов).

На основании этой линии должно быть принято решение о принадлежности образов классам w_1 , w_2 , и ошибка принятого решения должна быть минимизирована.

При построении несмещенной и состоятельной оценки линейной связи $\eta = M(y / x) = \widetilde{b}_0 + \widetilde{b}_1 x + \varepsilon_x$ с помощью обычного метода наименьших квадратов минимизируется взвешенная сумма квадратов отклонений в направлении той оси координат, которую измеряли с погрешностями $S_\eta = \sum_{i=1}^N \omega_{y_i} (y_i - b_0 - b_1 x_i)^2 \equiv \sum_{i=1}^N \varepsilon_{y_i}^2$ или $S_\xi = \sum_{i=1}^N \omega_{x_i} (x_i - b_0^* - b_1^* y_i)^2 \equiv \sum_{i=1}^N \varepsilon_{x_i}^2$, где ε_{x_i} , ε_{y_i} – невязки по осям OX и OY , $(b_0, b_1), (b_0^*, b_1^*)$ – соответственно две пары оценок параметров линий регрессии Y на X и X на Y .

В конфлюентной ситуации при одновременном существовании погрешностей измерений вдоль двух осей координат логично допустить, что одновременно должны минимизироваться суммы квадратов невязок вдоль двух осей: $\varepsilon_\Sigma^2 = \sum_{i=1}^N \varepsilon_{x_i}^2 + \sum_{i=1}^N \varepsilon_{y_i}^2$, что соответствует минимизации некоторой обобщенной квадратичной формы

$$S_\Sigma = \sum_{i=1}^N \{ \omega(x_i)(x_i - \xi_i)^2 + \omega(y_i)(y_i - \eta_i)^2 \} \quad (2)$$

Логично также допустить, что наилучшие (наилучшая) прямые пройдут через «центр тяжести» экспериментальных данных $\{x_i, y_i\}$:

$$\left\{ \frac{\sum_{i=1}^N \widetilde{\omega}_i x_i}{\sum_{i=1}^N \widetilde{\omega}_i}, \frac{\sum_{i=1}^N \widetilde{\omega}_i y_i}{\sum_{i=1}^N \widetilde{\omega}_i} \right\},$$

где $\widetilde{\omega}_i$ – некоторая весовая функция.

Приняв гипотезу о нормальном распределении результатов измерений (x_i, y_i) и их некоррелированности, получим, что обращение в максимум функции максимального правдоподобия исследуемой выборки $(x_i, y_i)_{i=1}^N$

$$L_\Sigma = \left[\prod_{i=1}^N 2\pi \sigma_{x_i}^2 \sigma_{y_i}^2 \right]^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \left[\sum_{i=1}^N \left(\frac{x_i - \xi_i}{\sigma_{x_i}} \right)^2 + \sum_{i=1}^N \left(\frac{y_i - b_0 - b_1 \xi_i}{\sigma_{y_i}} \right)^2 \right] \right\} \quad (3)$$

по искомым оценкам параметров $\widetilde{b}_0, \widetilde{b}_1$ эквивалентно обращению в минимум обобщенной квадратичной формы S_Σ (см. (2)), стоящей в показателе степени экспоненты (3), т.е. точки экстремума S_Σ и L_Σ совпадают:

$$\frac{\partial L_\Sigma}{\partial b_j} = \frac{\partial S_\Sigma}{\partial b_j} = 0 \quad (j = 1, 2)$$

В (3), в соответствии с традициями, нами положено $\omega(x_i) = 1 / \sigma_{x_i}^2$; $\omega(y_i) = 1 / \sigma_{y_i}^2$, и b_0, b_1, ξ_1 найдутся из системы уравнений

$$\frac{\partial S_{\Sigma}}{\partial \xi_1} = 0; \frac{\partial S_{\Sigma}}{\partial b_0} = 0; \frac{\partial S_{\Sigma}}{\partial b_1} = 0; \Rightarrow$$

$$\Rightarrow \begin{cases} \widetilde{\xi}_i = \frac{[\sigma_{yi}^2 x_i + \widetilde{b}_1 \sigma_{xi}^2 (y_i - \widetilde{b}_0)]}{\sigma_{yi}^2 + \widetilde{b}_1^2 \sigma_{xi}^2}; \\ \sum_{i=1}^N \frac{y_i - b_0 - \widetilde{b}_1 \widetilde{\xi}_i}{\sigma_{yi}^2} = 0; \\ \sum_{i=1}^N \frac{(y_i - b_0 - \widetilde{b}_1 \widetilde{\xi}_i) \widetilde{\xi}_i}{\sigma_{yi}^2} = 0; \end{cases} \quad (4)$$

Напомним, что $S_{\Sigma} = \sum_{i=1}^N \omega_{xi} (x_i - \xi_i)^2 + \omega_{yi} (y_i - \eta_i)^2$

Данная система кладется в основу итерационно-взвешенного алгоритма вычисления оценок $\widetilde{b}_0, \widetilde{b}_1, \widetilde{\xi}_1$ параметров b_0, b_1, ξ_1 ($i = \overline{1, N}$) оптимальной разделяющей линии регрессии $M(y / x_i, i = \overline{1, N})$ нашей задачи.

Система (4) содержит $N+2$ нелинейных уравнения с $N+2$ неизвестными $(\xi_1, i = \overline{1, N}, \widetilde{b}_0, \widetilde{b}_1)$.

Опишем алгоритм вычисления оценок параметров системы. Решаем систему (4) методом последовательного приближения. Поскольку у нас $N+2$ уравнения с $N+2$ неизвестными, то возникает проблема выбора одного из известных параметров в качестве ведущего приближения. Выберем таковым $\widetilde{\xi}$ и зададим его начальное приближение $\widetilde{\xi}_i^0 = x_i$.

Алгоритм решения системы (4) из $N+2$ нелинейных уравнений состоит в следующем:

1. Задается начальное приближение $\widetilde{\xi}_i^0 = x_i; (i = \overline{1, N})$. Далее подставляем заданное $\widetilde{\xi}_i^0$ в оставшиеся два уравнения системы.

2. Решаем полученную новую систему относительно $\widetilde{b}_0^0, \widetilde{b}_1^0$, в результате чего получим первое приближение $\widetilde{b}_0^{(1)}, \widetilde{b}_1^{(1)}$.

3. Подставляем $\widetilde{b}_0^1, \widetilde{b}_1^1$ в начальное приближение для $\widetilde{\xi}_i^0$ для отыскания следующего приближения $\widetilde{\xi}_i^{(1)}$.

4. Полученное $\widetilde{\xi}_i^{(1)}$ вновь подставляем в другие два уравнения и находим $\widetilde{b}_0^{(2)}, \widetilde{b}_1^{(2)}$.

5. Продолжаем процедуру до тех пор, пока решение $\widetilde{b}_0^{(k)}, \widetilde{b}_1^{(k)}$ не будет сходиться с истинным решением в рамках заранее заданной точности.

Можно показать, что итерационный процесс, сконструированный на основе уравнений (4), сходится, ибо он эквивалентен итерационной процедуре получения $\widetilde{b}_0, \widetilde{b}_1$ – взвешенного метода наименьших квадратов – сходящейся, как известно [3], от любого начального приближения $(w_i(\xi_i^0), b_0^0, b_1^0)$ при достаточно гладкой поверхности обобщенной квадратичной формы $S_{\Sigma}(b_0, b_1, \xi)$. [2, 3]

Нами сформулировано утверждение: итерационно-взвешенный алгоритм вычисления оценок параметров линейной разделяющей функции регрессии $y = \widetilde{b}_0 + \widetilde{b}_1 x + \varepsilon_x$ в условиях конъюнктивной ситуации и нормального распределения погрешностей измерений признаков образов по осям координат OX и

ОУ сходится к точке глобального минимума квадратичной формы $S_{\Sigma}(\tilde{b}_0, \tilde{b}_1, \tilde{\xi})$ или внутри области определения формы S_{Σ} , или на границе области.

Был проведен вычислительный эксперимент, моделирующий при помощи выборки реальную конфлюентную ситуацию, показавший справедливость данного утверждения и факт меньшего по величине смещения оценок параметров разделяющей линии регрессии (в среднем от двух до пяти процентов), чем у соответствующих оценок традиционного метода наименьших квадратов. Найдена их состоятельность уже при небольших объемах суммарной выборки ($N \leq 50$) для двух классов образов.

СПИСОК ЛИТЕРАТУРЫ

1. Демиденко Е.З. Линейная и нелинейная регрессии. – М.: Финансы и статистика, 1981.
2. Грешилов А.А., Стакун А.А. Статистические методы принятия решений с элементами конфлюентного анализа. – М.: Радио и связь, 1998.
3. Шпапиро Л., Стокман Дж. Компьютерное зрение. – М.: Бином, 2006. **ГИАБ**

КОРОТКО ОБ АВТОРАХ

Алнадфа Антуан – аспирант, e-mail: antwan.tiger@gmail.com, РУДН,
Исаев Андрей Борисович – профессор, e-mail: a.borisovich@mail.ru,
РУДН, Финансовый университет при правительстве РФ.

UDC 519.233.5

ITERATION-WEIGHTED ALGORITHM OF SEPARATION LINE FOR LINEAR REGRESSION FUNCTION IN THE PROBLEM ON CLASSIFICATION OF OBJECTS WITH CHARACTERISTICS BURDENED WITH MEASUREMENT ERRORS

Alnadfa A., Graduate Student, e-mail: antwan.tiger@gmail.com,
Peoples' Friendship University of Russia, Moscow, Russia,
Isaev A.B., Professor, e-mail: a.borisovich@mail.ru,
Peoples' Friendship University of Russia, Moscow, Russia
Financial University under the Government of Russian Federation, Moscow, Russia.

The author gives mathematically rigorous formulation and solution to the problem on constructing a separation line between two classes of codomains that represent two finite, discrete and countable sets convex hulls of which do not intersect.

For finding consistent estimates of parameters of the dividing line of regression, the iteration algorithm is constructed based on the function of the highest likelihood, on the assumption of normally distributed errors of measuring coordinates of characteristics of the two-dimensional codomains under confluence conditions.

Key words: true variables, confluent analysis, measurement errors, convex hull, highest likelihood function, codomain, characteristics of codomains.

REFERENCES

1. Demidenko E.Z. *Lineinaya i nelineinaya regressii* (Linear and nonlinear regression), Moscow, Finansy i statistika, 1981.
2. Greshilov A.A., Stakun A.A. *Statisticheskie metody prinyatiya reshenii s elementami konflyuentnogo analiza* (Statistical decision-making methods with elements of confluent analysis), Moscow, Radio i svyaz', 1998.
3. Shpapiro L., Stokman Dzh. *Komp'yuternoe zrenie* (Computer vision), Moscow, Binom, 2006.