ОПТИМИЗАЦИЯ ОЦЕНКИ ВЕРОЯТНОСТИ ОШИБОЧНОЙ КЛАССИФИКАЦИИ В ДИСКРЕТНОМ СЛУЧАЕ[1]

Виктор Неделько

***Abstract**: The goal of the paper is to investigate what training sample estimate of misclassification probability would be the best one for the histogram classifier. Certain quality criterion is suggested. The deviation for some estimates, such as resubstitution error (empirical risk), cross validation error (leave-one-out), bootstrap and for the best estimate obtained via some optimization procedure, is calculated and compared for some examples.*

***Keywords**: pattern recognition, classification, statistical robustness, deciding functions, complexity, capacity, overfitting, overtraining problem.*

***ACM Classification Keywords**: G.3 Probability and statistics, G.1.6. Numerical analysis: Optimization; G.2.m. Discrete mathematics: miscellaneous.*

***Conference**: The paper is selected from XVth International Conference "Knowledge-Dialogue-Solution" KDS 2009, Varna, Bulgaria, June-July 2009*

Введение

Для оценивания качества решающих функций (одна из первых работ [Лбов, 1965]) в задачах распознавания образов (классификации с учителем) на практике обычно используются точечные оценки риска, т.е. вероятности ошибочной классификации. В роли таких оценок, как правило, выступают эмпирический риск (resubstitution error) оценка скользящего экзамена (cross validation) или оценка bootstrap. При этом эмпирический риск является смещенной оценкой риска. Для величины смещения в общем случае существуют лишь приближенные интервальные оценки в рамках подхода Вапника–Червоненкиса [Вапник, Червоненкис, 1974], хотя для частных случаев возможно точное оценивание смещения, например, для дискретного пространства [Неделько, 2003]. Также имеет смысл использование эмпирических интервальных оценок риска [Неделько, 2008].

Наилучшей с практической точки зрения из точечных оценок риска считается bootstrap, чье преимущество продемонстрировано на многочисленных примерах. Естественным образом напрашивается вопрос, насколько эта оценка близка к оптимальной, и в каком смысле можно вообще говорить об оптимальности такого рода оценки [Неделько, 2007].

Стандартной мерой качества точечной оценки является ее эффективность, которая характеризуется средним квадратом отклонения (deviation) от оцениваемой величины. Однако эта величина зависит от вероятностной модели, т.е. распределения, из которого взята выборка, и для разных распределений оптимальными будут разные оценочные функционалы. Получаем ситуацию многокритериального выбора. В этом случае можно рассматривать множества Парето-оптимальных оценок. Но в данной ситуации критерии сравнимы, поскольку являются фактически одним критерием при разных моделях. Это позволяет сравнивать оценки, считая, что один функционал лучше другого, если его выигрыш в лучшей ситуации превосходит проигрыш в худшей.

[1] Работа выполнена при поддержке РФФИ, гранты 07-01-00331-а и 08-01-00944-а.

Если бы на множестве всех распределений была задана некоторая мера [Лбов, Старцева, 1999], адекватно отражающая «важность» этих распределений, или их «встречаемость» в реальных задачах, то можно было бы просто использовать усредненный критерий. Но так как такой меры нет, разумным представляется использование различных вариаций минимаксного подхода.

Задача нахождения оптимального оценочного функционала в общем случае является сложной, поэтому в данной работе исследуется частный случай задачи классификации в дискретном пространстве (histogram classifier), при котором все требуемые статистики могут быть вычислены аналитически [Braga–Neto, Dougherty, 2005].

Постановка задачи

Для введения основных понятий рассмотрим сначала общую постановку задачи построения решающих функций.

Пусть X – пространство значений переменных, используемых для прогноза, а Y – пространство значений прогнозируемых переменных, и пусть C – множество всех вероятностных мер на $D = X \times Y$. Тогда элементом $c \in C$ будет $P_c[D]$. Здесь и далее квадратные скобки используются для указания множества, на σ-алгебре подмножеств которого задана мера.

Решающей функцией назовем соответствие $f : X \to Y$ и введем для нее функцию потерь: $L : Y^2 \to [0, \infty)$.

Под риском будем понимать средние потери:

$$R(c, f) = \int L(y, f(x)) dP_c[D].$$

Пусть $V = \left\{ \left(x^i, y^i\right) \in D \;\middle|\; i = \overline{1, N} \right\}$ – случайная независимая выборка из распределения $P_c[D]$, $V \in D^N$. Эмпирический риск определим как средние потери на выборке:

$$\tilde{R}(v, f) = \frac{1}{N} \sum_{i=1}^{N} L\left(y^i, f\left(x^i\right)\right).$$

Пусть $Q : D^N \to \Phi$ – алгоритм построения решающих функций, а $f_{Q,V} \in \Phi$ – функция, построенная по выборке V алгоритмом Q.

Оценкой скользящего экзамена называется величина

$$\breve{R}(V, Q) = \frac{1}{N} \sum_{i=1}^{N} L\left(y^i, f_{Q,V_i'}\left(x^i\right)\right),$$

где $V_i' = V \setminus \left\{ \left(x^i, y^i\right) \right\}$ – выборка, получаемая из V удалением i-го наблюдения.

Также мы будем использовать оценку bootstrap

$$\hat{R}(V, Q) = \frac{1}{\mathrm{E}|J_0|} \mathrm{E} \sum_{i \in J_0} L\left(y^i, f_{Q,\hat{V}}\left(x^i\right)\right),$$

где $\hat{V}$ – выборка, получаемая из V путем N-кратного случайного (равновероятного) выбора ее значений с повторениями, J_0 – множество индексов объектов из V, ни разу не выбранных в $\hat{V}$, математическое ожидание подразумевает усреднение по выборкам $\hat{V}$. Легко показать, что $\mathrm{E}|J_0| = N\left(1 - \frac{1}{N}\right)^N \approx N e^{-1}$.

Ввиду того, что оценка bootstrap является смещенной, чаще используют ее в комбинации с эмпирическим риском

$$\ddot{R}(V,Q) = e^{-1} \cdot \tilde{R}(V,Q) + \left(1 - e^{-1}\right) \cdot \hat{R}(V,Q).$$

В общем случае оценочный функционал — это некоторая функция выборки (при фиксированном методе построения решающих функций).

Качество эмпирического функционала $\overline{R}\left(V, f_{Q,v}\right)$ как оценки риска естественно характеризовать средним квадратом уклонения, т.е.

$$\Delta = \mathrm{E}\left(\overline{R}\left(V, f_{Q,V}\right) - R\left(c, f_{Q,V}\right)\right)^2.$$

Существенная проблема заключается в том, что выражения зависят от c – распределения, которое неизвестно. Решением может быть взятие супремума по всем распределениям и ориентирование таким образом на «наихудшее» распределение.

Классификация в дискретном пространстве

Будем рассматривать задачу классификации двух образов.

Пусть X дискретно, то есть $X = \{1, \ldots, k\}$, и решающая функция минимизирует эмпирический риск независимо в каждой точке x.

Тогда вероятностная мера $c \in C$ задается набором вероятностей

$$c = \left\{ \varsigma_j^{\omega} = \mathrm{P}\left(x = j, y = \omega\right) \middle| \; j = \overline{1,k},\ \omega = \overline{1,2} \right\}.$$

При этом $Y = \{1, 2\}$, функцией потерь будет: $L\left(y, y'\right) = \begin{cases} 0, & y = y' \\ 1, & y \neq y' \end{cases}$, а риском – вероятность ошибочной классификации.

Обозначим $\alpha_j = P\left(x = i\right) = \varsigma_j^1 + \varsigma_j^2$, $p_j = P\left(y = 1 / x = i\right)$, $q_j = 1 - p_j$, $c_j = \left(\alpha_j, p_j\right)$.

Для выборки V объема N пусть n_j – число точек выборки, для которых $x = j$, и m_j – число точек, для которых $x = j$ и $y = 1$. Таким образом, выборка в дискретном случае задается совокупностью пар $\nu_j = \left(m_j, n_j\right)$, т.е. $V = \left\{ \nu_j \middle| \; j = \overline{1,k} \right\}$. Описывая выборку, мы будем иногда для краткости говорить, что в «ячейке» j находится m_j точек первого и $n_j - m_j$ точек второго класса.

Будем рассматривать алгоритм Q, который минимизирует эмпирический риск независимо в каждой точке пространства X, т.е. $f_{Q,V}\left(j\right) = 2$, при $n_j - m_j > m_j$, $f_{Q,V}\left(j\right) = 1$, при $n_j - m_j < m_j$, и $f_{Q,V}\left(j\right)$ принимает равновероятно значения 1 и 2, при $n_j - m_j = m_j$.

На выборках имеет место полиномиальное распределение $P\left(V\right)$, суммируя по которому, можно вычислять в том числе моменты различных функций выборки. Однако осуществлять перебор всех выборок – трудоемкая в вычислительном плане процедура, поэтому непосредственное суммирование по выборкам осуществимо только для небольших N и k.

При этом для аддитивных функций выборки вычисление моментов может быть произведено с полиномиальной трудоемкостью.

Вычисление моментов для аддитивных функций

Пусть $f(V)=\sum_{j=1}^{k}\varphi(\nu_j,c_j)=\sum_{j=1}^{k}\varphi(m_j,n_j,\alpha_j,p_j)$ – аддитивная функция выборки и распределения.

Математическое ожидание $\mathrm{E}\,f(V)=\sum_{j=1}^{k}\mathrm{E}\varphi(\nu_j,c_j)$ также аддитивно.

Обозначим $B(m,n,p)=C_n^m p^m(1-p)^{n-m}$ – биномиальное распределение.

Введем функцию $\mu_\varphi(c)\equiv\mu_\varphi(\alpha,p)=\mathrm{E}\varphi(\nu,c)$. Легко получить, что

$$\mu_\varphi(\alpha,p)=\sum_{n=0}^{N}B(n,N,\alpha)\sum_{m=0}^{n}B(m,n,p)\varphi(m,n,\alpha,p)=\sum_{n=0}^{N}B(n,N,\alpha)\pi_\varphi(n,\alpha,p),$$

где $\pi_\varphi(n,\alpha,p)=\sum_{m=0}^{n}B(m,n,p)\varphi(m,n,\alpha,p)$.

Окончательно, математическое ожидание есть $\mathrm{E}\,f(V)=\sum_{j=1}^{k}\mu_\varphi(c_j)$.

Для вычисления дисперсии имеем $\mathrm{D}\,f(V)=\mathrm{E}\,f^2(V)-(\mathrm{E}\,f(V))^2$.

$$\mathrm{E}\,f^2(V)=\sum_{j=1}^{k}\mathrm{E}\varphi^2(\nu_j,c_j)+\sum_{i\neq j}\mathrm{E}\,\varphi(\nu_i,c_i)\varphi(\nu_j,c_j).$$

Введем функции

$$\sigma_\varphi(c)\equiv\sigma_\varphi(\alpha,p)=\mathrm{E}\varphi^2(\nu,c),\ \ \omega_\varphi(c_1,c_2)\equiv\omega_\varphi(\alpha_1,p_1,\alpha_2,p_2)=\mathrm{E}\varphi(\nu_1,c_1)\varphi(\nu_2,c_2).$$

Имеем

$$\sigma_\varphi(\alpha,p)=\sum_{n=0}^{N}B(n,N,\alpha)\pi_\varphi^2(n,\alpha,p),\ \text{где}\ \ \pi_\varphi^2(n,\alpha,p)=\sum_{m=0}^{n}B(m,n,p)\varphi^2(m,n,\alpha,p).$$

$$\omega_\varphi(\alpha_1,p_1,\alpha_2,p_2)=\sum_{n=0}^{N}B(n,N,\alpha_1+\alpha_2)\sum_{n_1+n_2=n}B(n_1,n,\alpha_1')\,\pi_\varphi(n_1,\alpha_1,p_1)\pi_\varphi(n_2,\alpha_2,p_2),$$

где $\alpha_1'=\frac{\alpha_1}{\alpha_1+\alpha_2}$.

Окончательно, второй момент есть $\mathrm{E}\,f^2(V)=\sum_{j=1}^{k}\sigma_\varphi(c_j)+\sum_{i\neq j}\omega_\varphi(c_i,c_j)$.

Пусть $g(V)=\sum_{j=1}^{k}\psi(\nu_j,c_j)$ – также аддитивная функция выборки и распределения.

Смешанный момент

$$\mathrm{E}\,f(V)g(V)=\sum_{j=1}^{k}\mathrm{E}\varphi(\nu_j,c_j)\psi(\nu_j,c_j)+\sum_{i\neq j}\mathrm{E}\,\varphi(\nu_i,c_i)\psi(\nu_j,c_j)$$

вычисляется аналогично рассмотренным.

Оптимизация оценки риска

Пусть $f(V)$ – некоторая аддитивная оценка риска, а $g(V)$ – фактическое значение риска (вероятности ошибочной классификации), который в рассматриваемом дискретном случае также является аддитивной функцией.

Функция $f(V)$ полностью определяется функцией $\varphi(\nu, c)$, которая на самом деле не может зависеть от c, поскольку при построении оценки риска распределение неизвестно. Кроме того, данная функция дискретна и определяется счетным набором значений. Обозначим $\varphi(\nu) \equiv \varphi(m,n) = x_{mn}$.

Требуется подобрать x_{mn} так, чтобы минимизировать погрешность оценивания риска, т.е. величину

$$\Delta_{fg} = \mathrm{E}(f-g)^2 = \mathrm{E}\, f^2 - 2\mathrm{E}\, fg + \mathrm{E}\, g^2.$$

Пусть $\alpha_j = \alpha = \frac{1}{k}$ и $p_j = p$.

Вычислим частные производные

$$\frac{\partial \Delta_{fg}}{\partial x_{mn}} = 2k\, B(m,n,p)\Big((x_{mn} - \psi(m,n,\alpha,p))B(n,N,\alpha) + (k-1)c_{\varphi\psi}(n,N-n,\alpha,p)\Big), \text{ где}$$

$$c_{\varphi\psi}(n,N-n,\alpha,p) = \sum_{i=0}^{N-n} B(i+n,N,2\alpha)\, B(n,i+n,0{,}5)\big(\pi_{\varphi}(i,\alpha,p) - \pi_{\psi}(i,\alpha,p)\big).$$

Вторая производная

$$\frac{\partial^2 \Delta_{fg}}{\partial x_{mn}^2} = 2k\, B(m,n,p)\big(B(n,N,\alpha) + (k-1)B(2n,N,2\alpha)\, B(n,2n,0{,}5)\, B(m,n,p)\big).$$

Пусть $\delta^+(x_{mn}) = \max\limits_p \frac{\partial \Delta_{fg}}{\partial x_{mn}}$, $\delta^-(x_{mn}) = \min\limits_p \frac{\partial \Delta_{fg}}{\partial x_{mn}}$, а $p_{\max}$ и $p_{\min}$ – значения параметра p, при которых соответственно достигаются указанные максимум и минимум, и $\delta_2(x_{mn}) = \frac{\partial^2 \Delta_{fg}}{\partial x_{mn}^2}(p_{\max}) + \frac{\partial^2 \Delta_{fg}}{\partial x_{mn}^2}(p_{\min})$.

Наилучшей оценкой риска x_{mn}^* будем считать значение, при котором $\delta^+(x_{mn}^*) = -\delta^-(x_{mn}^*)$. При изменении оценки в окрестности точки x_{mn}^* максимальное по всем распределениям улучшение точности оценки будет равно максимальному ее ухудшению. Это значение представляется в определенном смысле оптимальным выбором, т.к. при других вариантах мы можем взять близкое значение, при котором максимальное уменьшение погрешности Δ_{fg} будет больше ее максимального увеличения. Оценку x_{mn}^* будем называть *сбалансировано-оптимальной*.

Для решения уравнения и нахождения x_{mn}^* использован аналог метода касательных, где начальным приближением взят эмпирический риск $x_{mn}^0 = \min(m, n-m)/N$, а последующие приближения вычислялись через предыдущие по формуле

$$x_{mn}^{i+1} = x_{mn}^i - \tau \frac{\delta^+(x_{mn}^i) + \delta^-(x_{mn}^i)}{\delta_2(x_{mn}^i)},$$

где $\tau \approx 0{,}1$ – параметр, введенный для обеспечения устойчивости (сходимости) метода. Заметим, что это не вполне метод касательных, поскольку $\delta_2(x_{mn})$ – не есть производная функции $\delta^+(x_{mn}) + \delta^-(x_{mn})$, но может выступать в роли эвристической оценки последней.

Экспериментальное сравнение оценок

Было проведено численное сравнение точности перечисленных оценок риска при различных значениях параметров задачи: объема выборки N и числа значений k.

Эмпирический риск и оценка скользящего экзамена являются аддитивными функциями и соответствующие им оценки выражаются соответственно

$$\tilde{x}_{mn} = \tfrac{1}{N}\min(m, n-m),$$

$$\breve{x}_{mn} = \tfrac{1}{N}\left(\min(m, n-m) + \max(m, n-m)\cdot\left(I(m = n-m) + \tfrac{1}{2}I(|n-2m| = 1)\right)\right),$$

где $I(\cdot)$ – индикаторная функция (равна 1, если условие истинно, и 0 – иначе).

Оценка bootstrap вычисляется следующим образом

$$\hat{x}_{mn} = \frac{\left(1-\frac{1}{N}\right)^{-N}}{N}\sum_{n'=0}^{N}\sum_{m'=0}^{n'}\sum_{n_0=0}^{n}\sum_{m_0=0}^{n_0}\hat{r}(m', n'-m', m_0, n_0-m_0)\,p^N_{m,n-m}(m', n'-m', m_0, n_0-m_0),$$

где $\hat{r}(i, j, i_0, j_0) = i_0\cdot I(j \ge i) + j_0\cdot I(i \ge j) + \tfrac{1}{2}\left(j_0\cdot I(j = i+1) + i_0\cdot I(i = j+1)\right)$,

а $p^N_{i,j}(i', j', i_0, j_0)$ – вероятность того, что в «ячейке», содержащей i объектов первого и j объектов второго класса, при генерировании bootstrap выборки окажется i' и j' точек первого и соответственно второго класса, и при этом i_0 и соответственно j_0 из исходных объектов не будут выбраны ни разу (по ним будет проводиться контроль). Данная вероятность может быть вычислена рекуррентно:

$$p^0_{i,j}(i', j', i_0, j_0) = I(i' = j' = i_0 = j_0 = 0),$$

$$p^{N+1}_{i,j}(i', j', i_0, j_0) = p^N_{i,j}(i'-1, j', i_0, j_0)\tfrac{i-i_0}{N} + p^N_{i,j}(i'-1, j', i_0-1, j_0)\tfrac{i_0}{N} +$$

$$+ p^N_{i,j}(i', j'-1, i_0, j_0)\tfrac{j-j_0}{N} + p^N_{i,j}(i', j'-1, i_0, j_0-1)\tfrac{j_0}{N} + p^N_{i,j}(i', j', i_0, j_0)\tfrac{N-i-j}{N}.$$

Комбинированная bootstrap оценка есть $\ddot{x}_{mn} = e^{-1}\cdot\tilde{x}_{mn} + \left(1-e^{-1}\right)\cdot\hat{x}_{mn}$.

Приведем численные результаты для $N = 50$, $k = 10$.

В таблице 1 приведены значения оценки x^*_{mn}, в таблице 2 — оценки $\ddot{x}_{mn}$. Видим, что при $n = 5$, что является наиболее вероятным числом выборочных точек в ячейке, оценки очень близки. При других значениях n различие более существенно, любопытным представляется отрицательные значения x^*_{mn} вклада в оценку вероятности ошибки для ячеек с большим числом точек и нулевым числом ошибок на обучении.

Таблица 1. Некоторые значения x^*_{mn}.

n	m					
	0	1	2	3	4	5
0	2,21					
1	0,96	0,96				
2	0,65	2,67	0,65			
3	0,41	1,89	1,89	0,41		
4	0,21	1,59	3,35	1,59	0,21	
5	0,03	1,31	2,79	2,79	1,31	0,03
6	–0,16	1,06	2,57	4,02	2,57	1,06
7	–0,36	0,83	2,33	3,61	3,61	2,33
8	–0,55	0,61	2,08	3,45	4,68	3,45

Таблица 2. Некоторые значения $\ddot{x}_{mn}$.

n	m					
	0	1	2	3	4	5
0	0,00					
1	0,32	0,32				
2	0,23	1,41	0,23			
3	0,12	1,59	1,59	0,12		
4	0,054	1,53	2,54	1,53	0,05	
5	0,022	1,39	2,75	2,75	1,39	0,02
6	0,0087	1,26	2,73	3,65	2,73	1,26
7	0,0032	1,16	2,60	3,87	3,87	2,60
8	0,0011	1,09	2,44	3,88	4,74	3,88

Значения всех оценок при $n = 5$ приведены на рис. 1. Цифрами обозначены: 1 – эмпирический риск, 2 – скользящий экзамен, 3 – комбинированная bootstrap оценка, 4 – оптимизированная оценка x^*_{mn}.

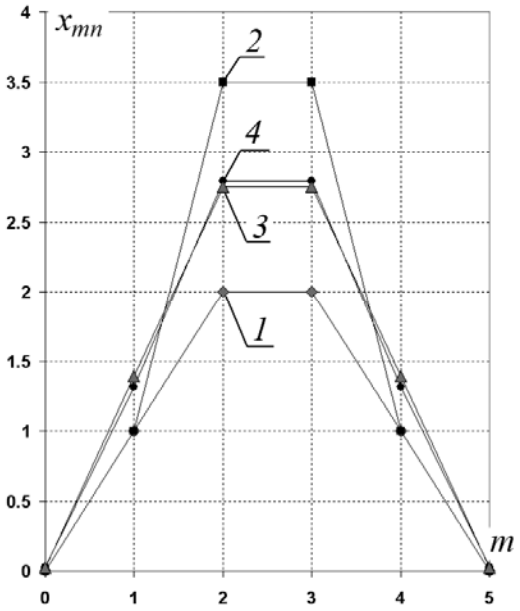

Рис. 1. Различные функции оценки риска.

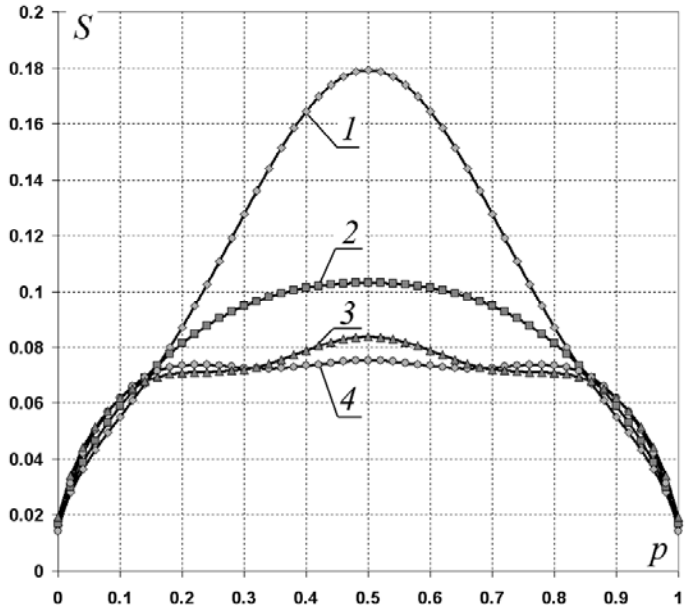

Рис. 2. Среднеквадратичная погрешность оценок.

На рис. 2 при различных значениях параметра p приведены графики среднеквадратичной погрешности $S = \sqrt{\Delta_{fg}}$ для всех оценок, нумерация такая же, как на рис. 1.

Из рассмотренных оценок ни одна не доминирует другую, т.е. для каждой пары оценок существуют p, при которых лучше как одна, так и другая. Однако в количественном отношении различие качества при разных p не равноценно. Так эмпирический риск имеет небольшое преимущество при малых p, но существенно проигрывает другим оценкам при p в окрестности 0,5. Сбалансировано-оптимальная оценка x^*_{mn} выглядит действительно наилучшей, при этом комбинированная оценка bootstrap очень близка к ней.

Заключение

В работе рассмотрена задача построения оценки вероятности ошибочной классификации в дискретном пространстве переменных, которая была бы в каком-то смысле наилучшей при различных предположениях о распределениях. Предложен метод решения данной задачи, основанный на построении сбалансировано-оптимальной оценки.

Как показывают численные эксперименты, такая оценка оказывается близкой к оценке, получаемой методом bootstrap. Это позволяет сделать предположение о том, что метод bootstrap в некотором смысле близок к наилучшему способу оценивания вероятности ошибочной классификации. Для проверки данного предположения требуются дополнительные исследования, в частности, нужно построить оценку, оптимизированную по всем распределениям в дискретном пространстве, а не только по заданному их подклассу. Также открытым является вопрос о распространении выводов, полученных при анализе задачи классификации в дискретном пространстве, на непрерывный случай.

Благодарности

Работа выполнена при поддержке РФФИ, гранты 07-01-00331-а и 08-01-00944-а.

Литература

[Лбов, 1965] Лбов Г.С. Выбор эффективной системы зависимых признаков. // Выч. системы, вып. 19, Новосибирск, 1965, с. 21–34.

[Вапник, Червоненкис, 1974] Вапник В. Н., Червоненкис А. Я. Теория распознавания образов. М.: Наука, 1974. 415 с.

[Лбов, Старцева, 1999] Г.С. Лбов, Н.Г. Старцева. Логические решающие функции и вопросы статистической устойчивости решений. Институт математики СО РАН, Новосибирск, 1999, 211 с.

[Неделько, 2003] V. M. Nedelko. Estimating a Quality of Decision Function by Empirical Risk // LNAI 2734. Machine Learning and Data Mining in Pattern Recognition. Third International Conference, MLDM 2003, Leipzig. Proceedings. Springer-Verlag. pp. 182–187.

[Неделько, 2007] Неделько В.М. Об эффективности функционалов эмпирического риска и скользящего экзамена как оценок вероятности ошибочной классификации // Proc. of int. conference, KDS'2007. Sofia. 2007. Vol. 1, P. 111–117.

[Неделько, 2008] V. M. Nedel'ko. Empirical bounds for misclassification probability // 9-th Int. Conf. "Pattern Recognition and Image Analysis: New Information Technologies" (PRIA–9–2008): Conference Proceedings. Vol. 2. – Nizhni Novgorod, 2008. P. 84–87.

[Braga–Neto, Dougherty, 2005] Braga–Neto. U. and Dougherty E.R. Exact performance of error estimators for discrete classifiers. // Pattern Recognition, Elsevier Ltd. 2005. V. 38, N 11. P. 1799-1814.

Информация об авторе

Виктор Михайлович Неделько *– с.н.с. лаборатории Анализа данных Института математики СО РАН, 630090, пр-т Коптюга, 4, Новосибирск, Россия, e-mail: nedelko@math.nsc.ru*