Устойчивость обучения метода релевантных векторов

Васильев О. М., Ветров Д. П., Кропотов Д. А.
ovasiliev@inbox.ru, vetrovd@yandex.ru, dkropotov@yandex.ru
Москва, ВМиК МГУ, ВЦ РАН

Рассматривается проблема оценивания качества обучения метода релевантных векторов в классической постановке задачи классификации с двумя классами. Получена оценка отклонения эмпирического риска от риска на скользящем контроле для метода релевантных векторов.

Определения и инструментарий

В задаче классификации восстанавливаемая величина y принимает значения из множества ответов $Y = \{-1, 1\}$, а независимая величина $\mathbf{x}$ принимает значения из множества объектов $X = \mathbb{R}^n$. Задана прецедентная информация (выборка) $S = \{\mathbf{z}_1 = (\mathbf{x}_1, y_1), \ldots, \mathbf{z}_m = (\mathbf{x}_m, y_m)\}$, $(\mathbf{z}_1, \ldots, \mathbf{z}_m) \in Z^m = (X \times Y)^m$. Рассмотрим также выборки $S^i = S \setminus \{\mathbf{z}_i\}$, $i = 1, \ldots, m$, получаемые удалением из S одного наблюдения. *Алгоритмическим оператором* называется отображение из X в $\mathbb{R}$, выбранное из некоторого параметризованного семейства. В случаях, если необходимо подчеркнуть роль параметров $\mathbf{u}$ некоторого алгоритмического оператора g, будем использовать альтернативную запись $g = [\mathbf{u}]$. Алгоритмы классификации, рассматриваемые в работе, имеют вид $\mathrm{sign}(g)$, где g — некоторый алгоритмический оператор.

Методом обучения μ называется отображение, сопоставляющее произвольной выборке S' алгоритм классификации $\mu_{S'}$. Далее рассматривается единственный метод обучения, поэтому введём специальные обозначения для результатов его обучения (алгоритмов и соответствующих алгоритмических операторов) на выборках S и S^i, $i = 1, \ldots, m$. Будем обозначать $\mu_S = \mathrm{sign}(f) = \mathrm{sign}([\mathbf{w}])$ и $\mu_S^i = \mathrm{sign}(f^i) = \mathrm{sign}([\mathbf{w}^i])$ результат обучения метода μ на выборке S и S^i соответственно. Для задачи классификации используем ценовую функцию $c\colon \mathbb{R}^2 \to \mathbb{R}$ вида

$$c(y, y') = \begin{cases} 1, & yy' \leqslant 0; \\ 1 - yy', & 0 \leqslant yy' \leqslant 1; \\ 0, & yy' \geqslant 1. \end{cases} \tag{1}$$

Эмпирическим риском алгоритма μ_S будем называть величину

$$R_{\mathrm{em}}(f) = \frac{1}{m} \sum_{i=1}^{m} c\big(f(\mathbf{x}_i), y_i\big),$$

Риском на скользящем контроле будем называть величину

$$R_{\text{lo}}(f) = \frac{1}{m}\sum_{i=1}^{m} c\big(f^i(\mathbf{x}_i), y_i\big).$$

Определение 1. *Метод обучения для задачи классификации будем называть β-устойчивым в обучении, если для всех $S \in Z^m$, всех $(\mathbf{x}, y) \in S$ и всех $i = 1, \ldots, m$ выполняется условие $\big|[\mathbf{w}](\mathbf{x}) - [\mathbf{w}^i](\mathbf{x})\big| \leqslant \beta$.*

Метод релевантных векторов RVM [1] в своем наиболее распространенном варианте строит алгоритмы классификации в форме

$$y = \operatorname{sign}(g(\mathbf{x})) = \operatorname{sign}([\mathbf{u}](\mathbf{x})) = \operatorname{sign}\left(\sum_{i=1}^{m} u_i K(\mathbf{x}, \mathbf{x}_i)\right),$$

где $\mathbf{u} \in \mathbb{R}^m$, $K(\cdot,\cdot)$ — некоторая функция, называемая *ядром*.

Алгоритмический оператор $f = [\mathbf{w}]$ (или $f^i = [\mathbf{w}^i]$), получаемый этим методом обучения по выборке S (или S^i), доставляет минимум по параметру $g = [\mathbf{u}]$, соответственно, функционалам

$$\frac{1}{m}\sum_{k=1}^{m} \log\big(1 + e^{-y_k g(\mathbf{x}_k)}\big) + u^{\mathrm{T}}\Lambda u, \quad \frac{1}{m}\sum_{k\neq i} \log\big(1 + e^{-y_k g(\mathbf{x}_k)}\big) + u^{\mathrm{T}}\Lambda u,$$

где $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$, $\lambda_i \geqslant 0$, $i = 1, \ldots, m$ — автоматически вычисляемые коэффициенты регуляризации.

Определение 2 (Дивергенция Брегмана [2]). *Пусть $F\colon \mathbb{R}^m \to \mathbb{R}$ — выпуклая функция. Тогда, если $\nabla F(\mathbf{g})$ — градиент F в точке $\mathbf{g}$, то $F(\mathbf{g}') \geqslant F(\mathbf{g}) + \langle \mathbf{g}' - \mathbf{g}, \nabla F(\mathbf{g})\rangle$. Дивергенцией точек $\mathbf{g}$ и $\mathbf{g}'$ называется величина $d_F(\mathbf{g}', \mathbf{g}) \triangleq F(\mathbf{g}) - F(\mathbf{g}') - \langle \mathbf{g} - \mathbf{g}', \nabla F(\mathbf{g}')\rangle \geqslant 0$.*

Лемма 1 (О дивергенциях [2]). *Пусть $N\big([u]\big) = u^T\Lambda u$. Тогда*

$$d_N(f, f^i) + d_N(f^i, f) \leqslant \frac{1}{m}\big|f(\mathbf{x}_i) - f^i(\mathbf{x}_i)\big|.$$

Применим лемму 1 для RVM.

Лемма 2. *Для RVM справедлива оценка суммы дивергенций:*

$$d_N(f, f^i) + d_N(f^i, f) = 2\big\|\Lambda^{\frac{1}{2}}(\mathbf{w} - \mathbf{w}^i)\big\|^2.$$

Следовательно, по лемме о дивергенциях,

$$\big\|\Lambda^{\frac{1}{2}}(\mathbf{w} - \mathbf{w}^i)\big\|^2 \leqslant \frac{1}{2m}\big|f(\mathbf{x}_i) - f^i(\mathbf{x}_i)\big|.$$

Обозначим $m \times m$-матрицу $\big(K(\mathbf{x}_i, \mathbf{x}_j)\big)_{i,j=1}^{m}$ через $\hat{K}$.

Лемма 3. *Справедлива следующая оценка отклонения алгоритмических операторов RVM:*

$$\left|f(\mathbf{x}_j) - f^i(\mathbf{x}_j)\right| \leqslant \left\|\hat{K}^{\mathrm{T}}\Lambda^{-\frac{1}{2}}\right\| \cdot \left\|\Lambda^{\frac{1}{2}}(\mathbf{w} - \mathbf{w}^i)\right\|.$$

Теорема 4. *Метод релевантных векторов является β-устойчивым в обучении с показателем $\beta = \frac{1}{2m}\left\|\hat{K}^{\mathrm{T}}\Lambda^{-\frac{1}{2}}\right\|^2$. При этом $\left|R_{\mathrm{lo}}(f) - R_{\mathrm{em}}(f)\right| \leqslant \beta$.*

Работа поддержана РФФИ, проекты № 07-01-00211, № 05-01-00332.

Литература

[1] *Tipping M. E.* Sparse Bayesian Learning and the Relevance Vector Machines // Journal of Machine Learning Research. — 2001. — Vol. 1, № 5. — P. 211–244.

[2] *Bousquet O., Elisseeff A.* Stability and Generalization // Journal of Machine Learning Research. — 2002. — Vol. 2, № 3. — P. 499–526.