Инвариантный метод настройки параметров в разреженном байесовском обучении

Ветров Д. П., Кропотов Д. А.
vetrovd@yandex.ru, dkropotov@yandex.ru
Москва, ВМиК МГУ, ВЦ РАН

Одним из популярных современных подходов для решения задач классификации является байесовское обучение, в частности, метод релевантных векторов (RVM), в котором используется независимая регуляризация весов объектов, а параметры регуляризации определяются автоматически в процессе обучения [2]. В RVM применяется регуляризация с помощью гауссовского априорного распределения. Однако, известно, что лапласовское априорное распределение может обеспечивать существенно более разреженные решающие правила [1]. Тем не менее, непосредственное применение лапласовского априорного распределения в RVM приводит к интегралам, недоступным для вычисления как аналитически, так и численно. Кроме того, метод RVM оказывается неинвариантным относительно линейных преобразований базисных функций, входящих в решающее правило.

В данной работе предлагается подход, в рамках которого возможно применение любых типов априорных распределений в разреженном байесовском обучении. При этом классификатор становится инвариантным относительно линейных преобразований базисных функций.

Допустим, что имеется набор объектов обучения $\{(\mathbf{x}_i, t_i)\}_{i=1}^n = (\mathcal{X}, \mathcal{T})$, представленных d-мерным вектором признаков $\mathbf{x} \in \mathbb{R}^d$ и меткой класса, принимающей два значения $t \in \{-1, +1\}$. В качестве семейства классификаторов выберем обобщенные линейные модели

$$y(\mathbf{x}, \mathbf{w}) = \sum_{i=1}^{M} w_i \varphi_i(\mathbf{x}) = \langle \mathbf{w}, \varphi(\mathbf{x}) \rangle, \tag{1}$$

где $\mathbf{w}$ — набор весов, определяющих классификатор, $\varphi(\mathbf{x}) = \{\varphi_i(\mathbf{x})\}_{i=1}^N$ — набор базисных функций (обобщенных признаков). Тогда логарифм правдоподобия корректной классификации обучающей выборки может быть записан как

$$L(\mathcal{T}|\mathcal{X}, \mathbf{w}) = -\sum_{i=1}^{n} \log\bigl(1 + \exp(-t_i y(\mathbf{x}_i, \mathbf{w}))\bigr). \tag{2}$$

Множество возможных классификаторов определяется априорной функцией распределения на веса $P(\mathbf{w}|\alpha)$. Оптимальные значения весов находятся как $\mathbf{w}_{MP} = \arg\max_{\mathbf{w}} \exp(L(\mathcal{T}|\mathcal{X}, \mathbf{w}))P(\mathbf{w}|\alpha)$. В RVM в качестве априорного распределения выбирается нормальное распределение

$w_i \sim \mathcal{N}(0, \alpha_i^{-1})$ со своим параметром регуляризации α_i для каждого веса. Для поиска значений α используется максимизация обоснованности:

$$P(\mathcal{T}|\alpha) = \int P(\mathcal{T}|\mathcal{X}, \mathbf{w}) P(\mathbf{w}|\alpha) d\mathbf{w} \to \max_{\alpha}. \tag{3}$$

Интеграл (3) не берется аналитически. В RVM используется аппроксимация подынтегральной функции (регуляризованного правдоподобия) с помощью гауссианы, от которой интеграл может быть найден аналитически. При применении других типов априорных распределений, в частности, лапласовского, аппроксимация регуляризованного правдоподобия с помощью гауссианы является неадекватной.

Основная идея предлагаемого подхода заключается в аппроксимации функции правдоподобия гауссианой, интерпретации собственных векторов матрицы ковариации гауссианы в качестве новых координатных осей в пространстве весов и регуляризации вдоль этих осей с подбором коэффициентов регуляризации путем максимизации обоснованности. После такой аппроксимации значение обоснованности (3) может быть записано как

$$P(\mathcal{T}|\alpha) \approx P(\mathcal{T}|\mathcal{X}, \mathbf{w}_{ML}) \int \exp\left(\frac{1}{2}(\mathbf{w} - \mathbf{w}_{ML})^T H (\mathbf{w} - \mathbf{w}_{ML})\right) P(\mathbf{w}|\alpha) d\mathbf{w},$$

где $\mathbf{w}_{ML}$ и $(-H)^{-1}$ — математическое ожидание и ковариационная матрица аппроксимирующей гауссианы.

Перейдем к новым переменным в пространстве весов, определяемых собственными векторами матрицы H: $\mathbf{u} = Q\mathbf{w}$, где $H = Q^T \Lambda Q$, $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_M)$, $\{\lambda_i\}_{i=1}^M$ — собственные значения H. Логарифм правдоподобия (2) — вогнутая функция, поэтому гессиан $H \leqslant 0$ и все его собственные значения $\{\lambda_i\}_{i=1}^M$ неположительны. Обозначим $h_i = -\lambda_i \geqslant 0$. Независимая регуляризация относительно новых переменных $\mathbf{u}$ означает, что априорная функция распределения может быть записана как

$$P(\mathbf{u}|\alpha) = \prod_{i=1}^{M} P(u_i|\alpha_i).$$

Основная цель подобной регуляризации — это представление обоснованности (3) в виде произведения одномерных интегралов

$$P(\mathcal{T}|\alpha) \approx P(\mathcal{T}|\mathcal{X}, \mathbf{u}_{ML}) \prod_{i=1}^{M} \underbrace{\int \exp\left(-\frac{h_i}{2}(u_i - u_{ML,i})^2\right) P(u_i|\alpha_i) du_i}_{f_i(h_i, u_{ML,i}, \alpha_i)}. \tag{4}$$

Каждый из одномерных интегралов может быть взят аналитически либо численно в зависимости от используемого типа регуляризатора. Поиск оптимальных значений α можно осуществлять независимо для каждого α_i, решая одномерную задачу оптимизации. Такая процедура обучения получила название метода релевантных собственных векторов (REVM).

В случае использовании гауссовского априорного распределения $u_i \sim \mathcal{N}(0, \alpha_i^{-1})$ значения одномерных интегралов $f_i(h_i, u_{ML,i}, \alpha_i)$ в выражении (4) могут быть вычислены аналитически:

$$f_i(h_i, u_{ML,i}, \alpha_i) = \sqrt{\frac{\alpha_i}{h_i + \alpha_i}} \exp\left(-\frac{h_i \alpha_i u_{ML,i}^2}{2(h_i + \alpha_i)}\right). \tag{5}$$

В зависимости от значений h_i и $u_{ML,i}$ интеграл (5), как функция от α_i, имеет один максимум либо монотонно возрастает на интервале $[0, +\infty)$. Приравнивая производную (5) по α_i к нулю, получаем оптимальное значение α_i:

$$\alpha_i^{\text{опт}} = \begin{cases} \frac{h_i}{h_i u_{ML,i}^2 - 1}, & \text{если } h_i u_{ML,i}^2 > 1; \\ +\infty, & \text{иначе.} \end{cases}$$

Теорема 1. *Рассмотрим наборы базисных функций $\varphi_1(\mathbf{x})$ и $\varphi_2(\mathbf{x}) = A\varphi_1(\mathbf{x})$, где A — невырожденная матрица размера $M \times M$. Обозначим решающее правило вида* (1)*, полученное методом классикации по набору базисных функций φ по выборке $(\mathcal{X}, \mathcal{T})$ через $y(\mathbf{x}; \varphi, \mathcal{X}, \mathcal{T})$. Тогда*

$$y_{RVM}(\mathbf{x}; \varphi_1, \mathcal{X}, \mathcal{T}) \not\equiv y_{RVM}(\mathbf{x}; \varphi_2, \mathcal{X}, \mathcal{T});$$
$$y_{REVM}(\mathbf{x}; \varphi_1, \mathcal{X}, \mathcal{T}) \equiv y_{REVM}(\mathbf{x}; \varphi_2, \mathcal{X}, \mathcal{T}).$$

Результаты экспериментов показывают, что по сравнению с RVM, REVM работает на порядок быстрее и приводит к более разреженным решающим правилам (в терминах переменных $\mathbf{u}$).

Работа выполнена при поддержке РФФИ, проекты № 06-01-08045, № 05-07-90333, № 06-01-00492, № 07-01-00211.

Литература

[1] *Williams P. M.* Bayesian regularization and pruning using a Laplace prior // Neural Computation. — 1995. — V. 7, № 1. — Pp. 117–143.

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