Минимизация вероятности переобучения для композиций линейных классификаторов низкой размерности

Евгений Соколов
ВМК МГУ

Конференция «Интеллектуализация обработки информации»
ИОИ-9
16-22 сентября 2012 г.

Задача классификации

- Объекты описаны d действительными признаками: $x_i \in \mathbb{R}^d$;
- Выборка объектов: $\mathbb{X} = (x_i)_{i=1}^L \subset \mathbb{R}^d$;
- Каждому объекту x_i однозначно соответствует ответ: $y_i \in Y = \{-1, +1\}$;
- Задача: восстановить зависимость $a : \mathbb{R}^d \to Y$ по обучающей выборке $X \subset \mathbb{X}$.

В данной работе рассматриваются композиции вида

$$a(x) = \operatorname{sign} \sum_{i=1}^{p} \operatorname{th}\langle w_i, x\rangle,$$

где

- $\langle x, y\rangle = \sum_{i=1}^d x_i y_i$ — скалярное произведение;
- $w_i \in \mathbb{R}^d$ — вектор весов;
- $\operatorname{th} x = \frac{2}{1+\exp(-x)} - 1$.

Задача построения композиции классификаторов

$$a(x) = \operatorname{sign} \sum_{i=1}^{p} \operatorname{th}\langle w_i, x \rangle$$

Требования:

- Небольшое число базовых классификаторов;
- Интерпретируемость базовых классификаторов;
- Высокая обобщающая способность.

Предлагается строить базовые классификаторы с учетом оценок обобщающей способности.

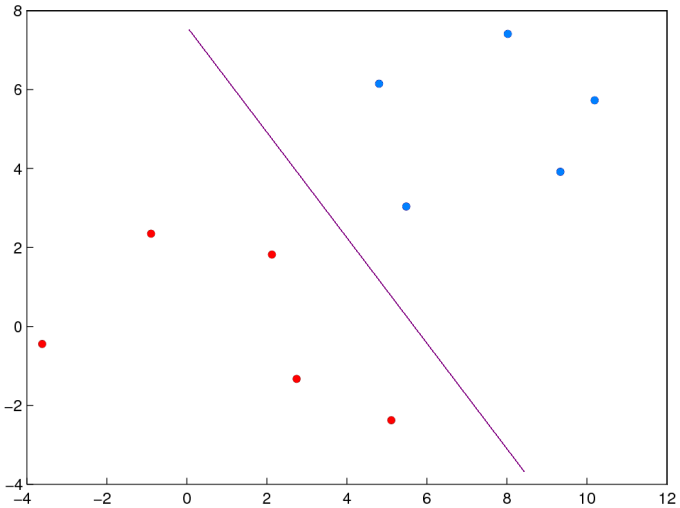
Определения

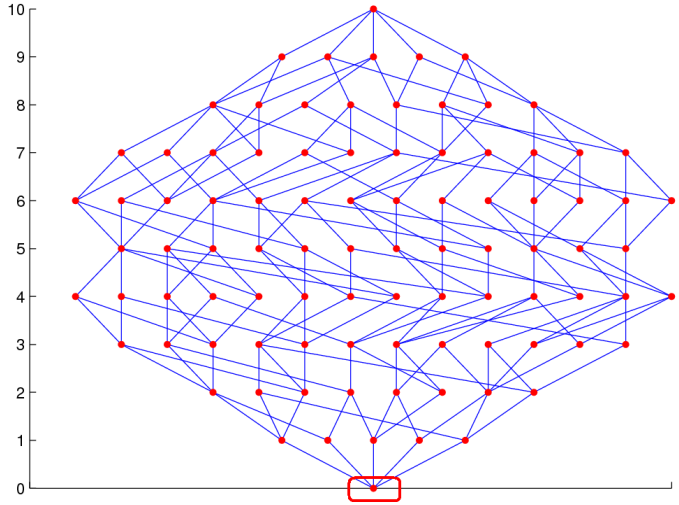
- $\mathbb{X} = \{x_1, \ldots, x_L\}$ — конечное генеральное множество объектов;
- $A = \{a_1, \ldots, a_D\}$ — конечное множество алгоритмов;
- $I(a, x) = [\text{алгоритм } a \text{ ошибается на объекте } x]$ — индикатор ошибки;
- $n(a, X)$ — *число ошибок* алгоритма a на выборке X.
- $\nu(a, X) = n(a, X)/|X|$ — *частота ошибок* алгоритма a на выборке X.
- *Метод обучения* $\mu : 2^{\mathbb{X}} \to A$ по произвольной выборке $X \subset \mathbb{X}$ выбирает алгоритм $a \in A$.
- *Вероятность переобучения:*

$$Q_\varepsilon = \mathsf{P}[\nu(\mu X, \bar{X}) - \nu(\mu X, X) \geqslant \varepsilon] = \frac{1}{C_L^\ell} \sum_{X \sqcup \bar{X}} [\nu(\mu X, \bar{X}) - \nu(\mu X, X) \geqslant \varepsilon]$$

Граф расслоения-связности

Множество вершин графа — все алгоритмы $a \in A$.
Множество ребер графа E — все пары вершин (a, a') такие, что $n(a, \mathbb{X}) + 1 = n(a', \mathbb{X})$ и $I(a, x_i) \leqslant I(a', x_i)$, $\forall x_i \in \mathbb{X}$.





Опр. $A_m = \{a \in A \mid n(a, \mathbb{X}) = m\}$ — m-й слой множества A.
Опр. $u(a) = \sharp\{a' \in A \mid (a, a') \in E\}$ — *связность* алгоритма a.
Опр. $q(a) = \sharp\{x \in \mathbb{X} \mid \exists b \in A : (b \leqslant a)\ \&\ (I(b, x) < I(a, x))\}$ — *неполноценность* алгоритма a.

Опр. Минимизация эмпирического риска (МЭР):

$$\mu X \in \operatorname*{Argmin}_{a \in A} n(a, X).$$

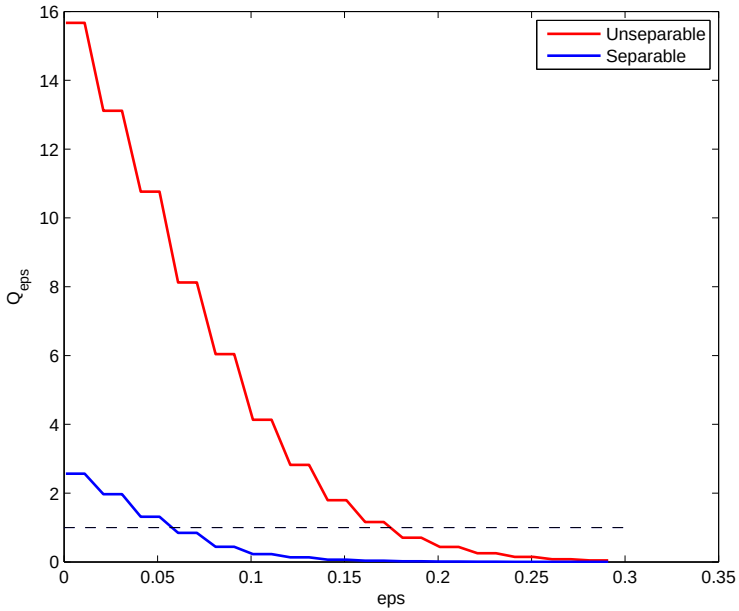
Теорема (Воронцов, Решетняк, Ивахненко, 2010)

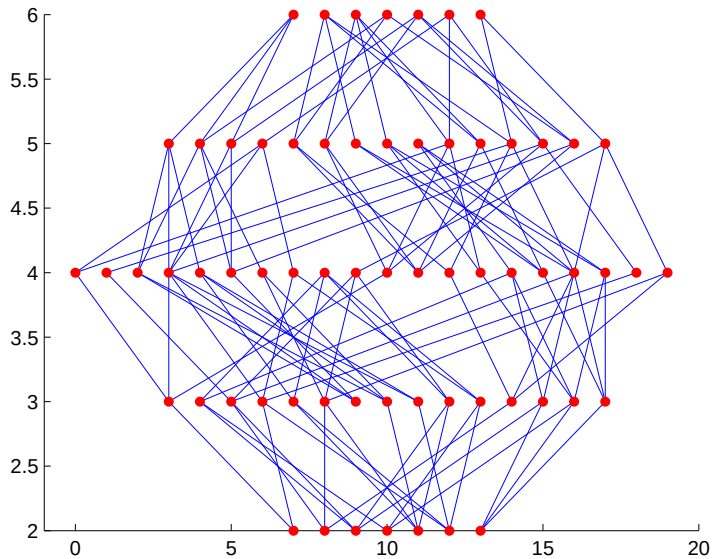
Для любого метода минимизации эмпирического риска μ *и любых* $\mathbb{X}$, A *и* $\varepsilon \in (0,1)$

$$Q_\varepsilon \leqslant \sum_{i=1}^{D} \frac{C_{L-u-q}^{\ell-u}}{C_L^\ell} \mathcal{H}_{L-u-q}^{\ell-u,\ m-q}\left(\frac{\ell}{L}(m-\varepsilon k)\right),$$

где $\mathcal{H}_L^{\ell,m}(z) = \sum_{s=0}^{\lfloor z \rfloor} \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}$ *— функция гипергеометрического распределения,* $u \equiv u(a)$, $q \equiv q(a)$, $m \equiv n(a, \mathbb{X})$.

Причина завышенности оценки





Улучшенная оценка вероятности переобучения

Определим для произвольных двух алгоритмов a_i и a_j из A множество A_{ij} объектов, на которых a_i не допускает ошибку, а a_j допускает:

$$A_{ij} = \{x \in \mathbb{X} \mid I(a_i, x) = 0, I(a_j, x) = 1\}$$

Теорема

Для пессимистичного метода минимизации эмпирического риска μ и любых $\mathbb{X}$, A и $\varepsilon \in (0, 1)$

$$Q_\varepsilon \leqslant \sum_{i=1}^{D} \min_{s \in S} \left\{ \sum_{t=0}^{T_{is}} \frac{C_{|A_{si}|}^{t} C_{L-u-|A_{si}|}^{\ell-u-t}}{C_L^\ell} \mathcal{H}_{L-u-|A_{si}|}^{\ell-u-t,\ m-|A_{si}|} \left(\frac{\ell}{L}(m - \varepsilon k) - t \right) \right\},$$

где $\mathcal{H}_L^{\ell,m}(z) = \sum_{s=0}^{\lfloor z \rfloor} \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}$ — функция гипергеометрического распределения, $u \equiv u(a)$, $m \equiv n(a, \mathbb{X})$, $T_{is} = \min(|A_{is}|, |A_{si}|)$, S — произвольное подмножество множества истоков графа расслоения-связности.

Опр.

$b(a) = \min_{s \in S} \left\{ \sum_{t=0}^{T_{is}} \frac{C^t_{|A_{si}|} C^{\ell-u-t}_{L-u-|A_{si}|}}{C^\ell_L} \mathcal{H}^{\ell-u-t,\ m-|A_{si}|}_{L-u-|A_{si}|} \left(\frac{\ell}{L}(m - \varepsilon k) - t \right) \right\}$ — вклад алгоритма a в оценку вероятности переобучения.

Задача: приближенно вычислить $B_\varepsilon = \sum_{a \in A} b(a)$.

$b(a)$ — функция, заданная на вершинах графа.
Значит, для приближенного вычисления суммы $\sum b(a)$ подходят методы, основанные на случайном блуждании.

Существенный вклад в оценку вносят только алгоритмы из нижних слоев графа расслоения-связности, поэтому перейдем к графу $G_t = (V_t, E_t)$, образованному его нижними t слоями.

Для генерации выборки алгоритмов предлагается использовать метод *Frontier Sampling*:

Вход: Граф $G = (V, E)$; число итераций N;
набор стартовых вершин $P = (v^1, \ldots, v^s)$;
Выход: Выборка вершин графа $v_1, v_2, \ldots, v_N$;

1: **для** $i = 1, \ldots, N$
2: выбрать $v \in P$ с вероятностью $\frac{\deg(v)}{\sum_{u \in P} \deg(u)}$;
3: **с вероятностью** $\frac{1}{2}$
4: выбрать вершину v' из равномерного
5: распределения на $\{v' \in V \mid (v, v') \in E\}$;
6: $v_i := v'$;
7: **иначе**
8: $v' := v$; $v_i := v$;
9: Заменить в P вершину v на v';

Ribeiro B., Towsley D. *Estimating and sampling graphs with multidimensional random walks* // 10th Conf. on Internet Measurement, 2010. — Pp. 390–403.

Теорема

Пусть $a_1, \ldots, a_n$ — случайная независимая выборка алгоритмов из первых t слоев семейства A, причем вероятность выбрать алгоритм $a \in A$ в результате случайного блуждания равна $p(a)$. Тогда следующая оценка является несмещенной и состоятельной для B_ε:

$$\hat{B}_\varepsilon = \frac{1}{n} \sum_{i=1}^{n} \frac{b(a_i)}{p(a_i)}$$

У нас: $p(a) = \frac{\deg(a)}{2|E_t|}$.

Данная оценка является хорошим приближением лишь при очень больших n, значительно превосходящих число вершин в графе.

Для повышения точности оценивания предлагается преобразовать B_ε:

$$B_\varepsilon = \sum_{i=1}^{D} b(a_i) = \sum_{m=0}^{L} |A_m| \left(\frac{1}{|A_m|} \sum_{a \in A_m} b(a) \right) = \sum_{m=0}^{L} |A_m| B_m$$

и оценивать все B_m по отдельности по выборке $a_1, \ldots, a_n$:

$$\hat{B}_m = \frac{\sum_{i=1}^{n} \dfrac{[m(a_i) = m]}{p(a_i)} \dfrac{b(a_i)}{|V_t|}}{\sum_{i=1}^{n} \dfrac{[m(a_i) = m]}{p(a_i)}}. \qquad (1)$$

Лемма

Оценка (1) является асимптотически несмещенной: $\mathbb{E}\hat{B}_m \xrightarrow[n \to \infty]{} B_m$.

Основная задача: построение непереобученных композиций линейных классификаторов вида

$$a(x) = \operatorname{sign} \sum_{i=1}^{p} \operatorname{th}\langle w_i, x\rangle.$$

Каждый базовый классификатор строится в подпространстве малой размерности.

Пусть известна некоторая оценка вероятности переобучения

$$\mathsf{P}[\nu(\mu X, \bar{X}) - \nu(\mu X, X) \geqslant \varepsilon] \leqslant \eta(\varepsilon).$$

Обратив ее, можно оценить частоту ошибок на контроле: с вероятностью не менее $1 - \eta$

$$\nu(\mu X, \bar{X}) \leqslant \nu(\mu X, X) + \varepsilon(\eta).$$

Величину в правой части неравенства будем использовать в качестве критерия при отборе признаков.

Построение композиции методом ComBoost

Вход: Выборка X; параметры T, ℓ_0, ℓ_1;
Выход: Базовые линейные классификаторы $b_1, \ldots, b_T$

1: Инициализировать веса и отсутпы:
$w_i = 1$, $M_i = 0$ для всех $i = 1, \ldots, \ell$;
2: **для** $t = 1, \ldots, T$, пока не выполнен критерий останова
3: Обучить базовый алгоритм (SVM по отобранным признакам):
$b_t := \operatorname{argmin}_b Q(b, X, W)$;
4: Обновить значения отступов:
$M_i = M_i + y_i b_t(x_i)$ для всех $i = 1, \ldots, L$;
5: упорядочить выборку X по возрастанию отступов M_i;
6: Отобрать объекты для обучения следующего базового алгоритма:
$w_i = [\ell_0 \leqslant i \leqslant \ell_1]$;

Маценов А. А. *Комитетный бустинг: минимизация числа базовых алгоритмов при простом голосовании* // Всеросс. конф. ММРО-13. — М.: МАКС Пресс, 2007. — С. 180–183.

	Wine Quality	Statlog	Waveform	Faults
Оценка, основанная на эмпирическом риске $\nu(\mu, \mathbb{X})$	64,70	84,92	84,78	73,39
Двухслойная нейронная сеть, настроенная методом backpropagation	72,06	85,41	86,79	74,51
Оценка (2), в которой обращается комбинаторная оценка [Воронцов, 2010], вычисленная с помощью случайных блужданий	69,48	86,26	85,77	77,81
Оценка скользящего контроля, вычисленная по 100 случайным разбиениям	71,06	85,26	86,38	75,76
Оценка (2), в которой обращается предложенная комбинаторная оценка, вычисленная с помощью случайных блужданий	74,68	86,75	86,91	74,03

$$Q_c = \nu(a_0, X) + \varepsilon\left(\tfrac{1}{2}\right) \qquad (2)$$

Зависимость ошибки от длины композиции

Основные результаты

- Предложена новая комбинаторная оценка вероятности переобучения, учитывающая попарные взаимодействия алгоритмов;
- Предложен метод эффективного вычисления данной оценки, основанный на случайных блужданиях;
- Предложенный метод применен к задаче построения композиций классификаторов. Получаемые композиции обладают высокой обобщающей способностью при весьма скромном количестве базовых классификаторов (3-6).