Оценки вероятности переобучения и комбинаторные отступы объектов в задачах классификации

Евгений Соколов
ВМК МГУ, кафедра ММП, студент 4-го курса

Конференция «Ломоносов-2012»
11 апреля 2012 г.

Задача

Основная цель: построение композиций классификаторов над пространствами малой размерности.

Шаги:

1. Отбор подпространств признаков;
2. Построение классификаторов в найденных подпространствах;
3. Объединение классификаторов в композицию.

Определения

Дано:

- генеральная выборка $\mathbb{X}^L = \{x_1, \dots, x_L\}$;
- семейство алгоритмов $\mathcal{A}$;
- индикатор ошибки $I : \mathcal{A} \times \mathbb{X} \to \{0, 1\}$;
- бинарный вектор ошибок алгоритма a: $\vec{a} = (I(a, x_1), \dots, I(a, x_L))$;
- метод обучения $\mu : 2^{\mathbb{X}} \to \mathcal{A}$.

Аксиома: Все разбиения генеральной выборки на обучающую X^ℓ и контрольную X^k равновероятны.
Вероятность переобучения:

$$Q_\varepsilon(\mu, X^L) = \mathsf{P}[\nu(\mu X, X^k) - \nu(\mu X, X^\ell) \geqslant \varepsilon],$$

$\nu(a, X)$ — частота ошибок алгоритма a на выборке X.

Критерий отбора признаков

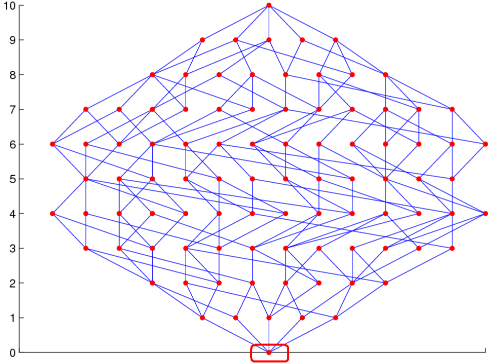
Пусть известна оценка вероятности переобучения $Q_\varepsilon(\mu, X^L) \leqslant \eta(\varepsilon)$. Тогда, если $\varepsilon(\eta)$ — функция, обратная к $\eta(\varepsilon)$, то с вероятностью не менее $(1 - \eta)$ справедлива оценка

$$\nu(\mu X, X^k) \leqslant \nu(\mu X, X^\ell) + \varepsilon(\eta)$$

Предлагается минимизировать величину $\nu(\mu X, X^\ell) + \varepsilon(\eta)$, используя ее как внешний критерий для отбора признаков.

Необходимы точные оценки вероятности переобучения и эффективные способы их вычисления.

Граф расслоения-связности



Граф расслоения-связности — граф Хассе естественного отношения порядка на множестве векторов ошибок алгоритмов:
$a \le a \Leftrightarrow \forall x I(a,x) \le I(b,x)$.
Ребру (a,b) соответствует объект $x_{ab}: \; I(a,x_{ab}) = 0, I(b,x_{ab}) = 1$.

Теорема (Воронцов, Решетняк, Ивахненко, 2010)

Для пессиместичного метода минимизации эмпирического риска μ *и любых* $\mathbb{X}$, $\mathcal{A}$ *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 u *— верхняя связность алгоритма* a *(число ребер, исходящих из* a*),*
q *— неполноценность алгоритма* a *(мощность множества объектов, соответствующих всем ребрам на путях, ведущих в* a*),*
m *— число ошибок алгоритма* a*,*

$$\mathcal{H}_L^{\ell,m}(z) = \sum_{s=0}^{\lfloor z \rfloor} \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}$$

— функция гипергеометрического распределения.

Улучшенная оценка вероятности переобучения

Теорема

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

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

где u — верхняя связность алгоритма a (число ребер, исходящих из a), m — число ошибок алгоритма a,
$|A_{is}|$ — число объектов, на которых ошибается a_s и не ошибается a_i,
$|B_{is}|$ — число объектов, на которых не ошибается a_s и ошибается a_i,

$$\mathcal{H}_L^{\ell,m}(z) = \sum_{s=0}^{\lfloor z \rfloor} \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}$$

— функция гипергеометрического распределения.

Проблема эффективного вычисления оценки

Проблема: для вычисления оценки необходимо обойти весь граф расслоения-связности, что невозможно на практике.

Требуется упростить граф так, чтобы

- можно было быстро осуществить его обход;
- вычисленная по нему оценка была близка к истинной.

Комбинаторный отступ

Определение

Комбинаторным отступом объекта x_0 называется величина

$$d(x_0) = \min_{a_s \in S} \min\{d \mid \exists a_i : \ I(a_s, x_0) \neq I(a_i, x_0), |B_{is}| = d\}$$

Можно эффективно оценить отступы объектов с помощью сэмплирования.

Свойства отступов

Теорема

Вклад алгоритма a в оценку вероятности переобучения экспоненциально убывает с ростом $\max_{x \in N(a)} d(x)$:

$$Q(a) = \mathcal{O}\left(\frac{1}{2^{d(a)}}\right),$$

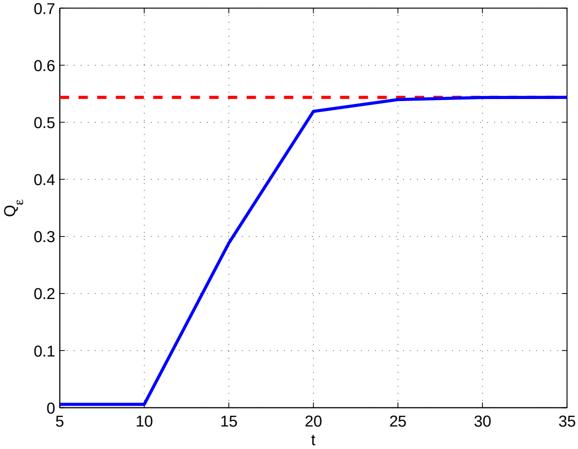
где $d(a) = \max_{x \in N(a)} d(x)$, $N(a) = \{x_{a,a'} \mid a' \in \mathcal{A}\}$.

Теорема

Если из алгоритма a выходит ребро, соответствующее объекту x с отступом $d(x) = t$, то существует такой путь из одного из истоков до этого алгоритма, что все ребра в нем соответствуют объектам с отступами не больше t.

Вывод: если удалить из графа все ребра, соответствующие объектам с $d(x) > t$, то недостижимыми могут стать лишь алгоритмы с незначительными вкладами в оценку.

Эксперимент



$L = 200$, $\ell = 100$, $\varepsilon = 0.1$; $\mathcal{A}$ — семейство линейных классификаторов.
Время работы для $t = 20$ — 20 секунд, полный обход графа занял более часа.

Заключение

- Предложена более точная оценка вероятности переобучения;
- Введено понятие комбинаторного отступа объекта, характеризующее его «важность» для вычисления оценки вероятности переобучения;
- Показано, что можно значительно упростить вычисление оценки вероятности переобучения, оставив в графе только те ребра, которые соответствуют объектам с маленькими отступами.