Концентрация меры в комбинаторных оценках обощающей способности

Н. К. Животовский

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра "Интеллектуальные Системы"

Научный руководитель д.ф.-м.н., с.н.с. ВЦ РАН К. В. Воронцов

20 июня 2013 г.

Содержание

Определения

- Задана конечная генеральная выборка объектов: $\mathbb{X} = \{x_1, \dots, x_L\}$.
- Задано множество $A(\mathbb{X})$ алгоритмов (классификаторов), сопоставляющих объекту его класс.
- Бинарная функция потерь $I \colon A \times \mathbb{X} \to \{0, 1\}$, где $(I(a, x) = 1) \leftrightarrow (a$ допускает ошибку на объекте $x)$.
- Метод обучения $\mu : X \to A$, $X \subseteq \mathbb{X}$.
- Разбиение генеральной выборки на обучение и контроль $\mathbb{X} = X \sqcup \bar{X}$.
- Отношения порядка $\forall a, b \in A, (a \leq b) \leftrightarrow (I(a, x) \leq I(b, x), \forall x \in \mathbb{X})$, причем $(a < b) \leftrightarrow (a \neq b, a \leq b)$, $(a \prec b) \leftrightarrow (a < b, \rho(a, b) = 1)$
- Пессимистичная минимизация эмпирического риска $\mu X = \arg \max\limits_{a \in A(X)} n(a, \bar{X})$, где $A(X) = \arg\min\limits_{a \in A} n(a, X)$.

Вероятностное предположение

Предположение

Все разбиения $\mathbb{X} = X \cup \bar{X}$, $|X| = \ell, |\bar{X}| = k$ *равновероятны.*

Цель

Для метода обучения μ *и обучающей выборки* X *оценить частоту ошибок на контрольной выборке* $\nu(\mu X, \bar{X})$.

Существующие оценки вероятности переобучения

$$Q_t(\mu, \mathbb{X}) = \mathsf{P}\left[\nu(\mu X, \bar{X}) - \nu(\mu X, X) \geqslant t\right]$$

существенно используют информацию о семействе алгоритмов на всей генеральной выборке.

Возникающие задачи

Цель

Получение верхних оценок обобщающей способности, зависящих только от наблюдаемой обучающей выборки.

Определение

Функция

$$ {}_2F_1(a, b, c, z) = 1 + \sum_{k=1}^{\infty} \left[\prod_{l=0}^{k-1} \frac{(a+l)(b+l)}{(1+l)(c+l)} \right] z^k $$

с параметрами a, b, c, *определённая в круге* $|z| < 1$, *называется гипергеометрической.*

Оценка ожидаемой переобученности

Теорема

Пусть $\ell = k = \frac{L}{2}$ *и* $\max\limits_{a \in A} m(a) \leq \frac{L}{2}$, *тогда*

$$\mathcal{EOF}_{\max}(\mathbb{X}) = \mathsf{E} \max_{a \in A} (\nu(a, \bar{X}) - \nu(a, X)) \leq$$

$$\min_{\lambda > 0} \frac{1}{\lambda} \left(\ln(\cosh(\lambda)) + \frac{1}{\ell} \ln \left(\sum_{s=0}^{l} \Delta_s \varphi(s, \ell, \lambda) \right) \right) \leq$$

$$\sqrt{\frac{2 \ln(|A(\mathbb{X})|)}{\ell}},$$

где Δ_s — *число алгоритмов в* s*-ом слое семейства алгоритмов,*

$$\varphi(s, \ell, \lambda) = {}_2F_1 \left(\tfrac{1-s}{2}, -\tfrac{s}{2}, \tfrac{1}{2} - \ell, (\tanh(\lambda))^2 \right)$$

Энтропия семейства алгоритмов

Пусть π – некоторая вероятностная мера на $A(\mathbb{X})$.
Алгоритм a_0 – корректный на $\mathbb{X}$.

$B(a, \varepsilon)$ – замкнутый шар с центром в алгоритме a и радиусом ε по метрике $\sqrt{\rho(a, b)}$.

Определение

Энтропия семейства $A(\mathbb{X})$

$$\frac{1}{3}\ln\left(\frac{2}{\min\limits_{a\in A\cup a_0}\pi\left(B(a,\frac{d}{2})\right)}\right)+\frac{4}{3}\sum_{k=2}^{\infty}2^{-k}\ln\left(\frac{2}{\min\limits_{a\in A\cup a_0}\pi\left(B(a,\frac{d}{2^k})\right)}\right),$$

где d *– диаметр семейства* $A(\mathbb{X})\cup a_0$.

Энтропийная оценка переобученности

Теорема

Пусть $\ell = k = \frac{L}{2}$ *и* $\max\limits_{a \in A} m(a) \leq \frac{\ell}{2}$, *тогда*

$$\mathcal{EOF}_{\max}(\mathbb{X}) \leq 3\sqrt{\frac{2\mathbf{Q}(A(\mathbb{X}))}{\ell}}$$

Показано, что данная оценка не хуже, чем предыдущая (при оптимальном выборе меры π).
Показано, что в случае, когда ёмкость семейства алгоритмов конечна, энтропия зависит лишь от ёмкости.

Комбинаторные характеристики семейства алгоритмов

Для алгоритма a введем

- Порождающее множество
  $X_a = \bigl\{x \in \mathbb{X} \bigm| \exists b \in A\colon a \prec b,\; I(a,x) < I(b,x)\bigr\}.$
- Запрещающее множество
  $X'_a = \bigl\{x \in \mathbb{X} \bigm| \exists b \in A\colon b < a,\; I(b,x) < I(a,x)\bigr\}.$
- Верхняя связность $u(a) = |X_a|$.
- Неполноценность $q(a) = |X'_a|$.
- Нижняя связность
  $d(a) = \bigl|\bigl\{x \in \mathbb{X} \bigm| \exists b \in A\colon b \prec a,\; I(b,x) < I(a,x)\bigr\}\bigr|.$
- Графом расслоения–связности множества алгоритмов A будем называть направленный граф $\langle A, E\rangle$ с множеством рёбер $E = \bigl\{(a,b)\colon a \prec b\bigr\}$.

Оценка, учитывающая метод обучения

Теорема

Пусть метод обучения μ *– ПМЭР,* $\ell = k = \frac{L}{2}$ *и* $\max\limits_{a \in A} m(a) \le \frac{L}{2}$

$$\mathcal{EOF}_\mu(\mathbb{X}) = \mathsf{E}(\nu(\mu X, \bar{X}) - \nu(\mu X, X)) \le$$

$$\min_{\lambda > 0} \frac{1}{\lambda}\left(\ln(\cosh(\lambda)) + \frac{1}{\ell}\ln\left(\frac{\sum\limits_{a \in A}\varphi(\ell, m, q, u, \lambda)}{C_{2\ell}^{\ell}}\right)\right).$$

$$\varphi(\ell, m, q, u, \lambda) =$$

$$\sum_{j=0}^{m-q}\sum_{i=0}^{j}(-1)^i C_j^i C_{2\ell-j-u-q}^{\ell-q+i-j} C_{m-q}^{j}(1+\tanh(\lambda))^q \tanh^j(\lambda).$$

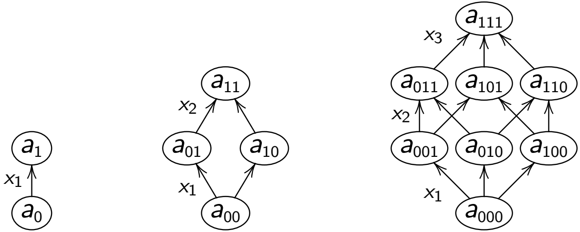
Факторы завышенности

Замечание

Предыдущая оценка не превосходит $\sqrt{\frac{2\ln(|A(\mathbb{X})|)}{\ell}}$

Завышенность предыдущей оценки появляется из-за использование неравенства

$$[\mu X = a] \leq [X_a \subset X][X'_a \subset \bar{X}].$$



Критерий в терминах графа расслоения–связности

Теорема

Пусть $\ell \geq 2 \max\limits_{a \in A} |X_a|$ *и* $k \geq 2 \max\limits_{a \in A} |X'_a|$, G *— граф расслоения–связности множества* A*. Условие*

$$[\mu X = a] = [X_a \subset X][X'_a \subset \bar{X}].$$

выполнено тогда и только тогда, когда

1. *граф* G *имеет только один сток и один исток,*
2. *если* a *— произвольная вершина* G, $x_1, \ldots, x_n$ *— объекты* $\mathbb{X}$*, соответствующие выходящим из нее ребрам, то* G *содержит направленный* n*-мерный куб с нижней вершиной* a*, построенный с помощью ребер, соответствующих* $x_1, \ldots, x_n$.

Неравенство концентрации в равномерном случае.

Теорема

При $\ell = k = \frac{L}{2}$ для $t > 0$ имеет место неравенство

$$\mathsf{P}\left(\max_{a \in A(\mathbb{X})} (\nu(a, \bar{X}) - \nu(a, X)) \geq \mathcal{EOF}_{\max}(\mathbb{X}) + \frac{8\sqrt{\pi}\sigma}{\ell} + t\right)$$
$$\leq 2\exp\left(-\frac{t^2\ell^2}{16\sigma^2}\right), \text{ где } \sigma = \sqrt{\max_{a \in A(\mathbb{X})} m(a)} \leq \sqrt{2\ell}$$

Замечание

- $\mathcal{EOF}_{\max}(\mathbb{X})$ *оценивается сверху величинами, не зависящими от ненаблюдаемой выборки.*
- *Локализация (уменьшение* $\max_{a \in A(\mathbb{X})} m(a)$*) улучшает оценку.*

Неравенство концентрации с методом обучения.

- Для разбиения $\mathbb{X} = X \sqcup \bar{X}$ и метода обучения μ функционал переобученности $\delta(\mu, X \sqcup \bar{X}) = \nu(\mu X, \bar{X}) - \nu(\mu X, X)$.
- Для разбиения $X \sqcup \bar{X}$ разбиение $X' \sqcup \bar{X}'$ получается перестановкой некоторого элемента из X с элементом $\bar{X}$.
- Для чётных ℓ естественным образом вводится $\mathcal{EOF}_\mu(X)$

Определение

ПМЭР μ вместе с семейством $A(\mathbb{X})$ называется (α, β)-устойчивым, если для всех пар $\{X \sqcup \bar{X}, X' \sqcup \bar{X}'\}$

$$\left|\delta(\mu, X \sqcup \bar{X}) - \delta(\mu, X' \sqcup \bar{X}')\right| \leq \frac{2\beta}{\ell},$$

$$\mathcal{EOF}_\mu(\mathbb{X}) - \mathcal{EOF}_\mu(X) \leq \alpha.$$

Оценка ненаблюдаемой частоты ошибок.

Теорема

Пусть $\ell = k = \frac{L}{2}$, μ — (α, β)*-устойчивый ПМЭР,* $t > 0$

$$\mathsf{P}\left(\nu(\mu X, \bar{X}) - \nu(\mu X, X) \geq \mathcal{EOF}_\mu(X) + \alpha + t\right) \leq \exp\left(\frac{-(\ell+1)t^2}{4\beta^2}\right).$$

Следствие

С вероятностью не меньшей $1 - \delta$

$$\nu(\mu X, \bar{X}) < \nu(\mu X, X) + \mathcal{EOF}_\mu(X) + \alpha + 2\beta\sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{\ell+1}}.$$

Замечание

Оценка вычислима только по наблюдаемой выборке.

Выводы

- Получены оценки ожидаемой переобученности с нелинейным вкладом алгоритмов.
- Исследованы степени вкладов алгоритмов в эти оценки.
- Получен критерий точности комбинаторных оценок обобщающей способности.
- Получены необходимые для комбинаторного подхода неравенства концентрации меры.
- Получена оценка ненаблюдаемой частоты ошибок в равномерном по семейству алгоритмов случае.
- Получена оценка ненаблюдаемой частоты ошибок для устойчивой минимизации эмпирического риска.