УДК 519.7

А.А. ГАЛКИН

АЛГОРИТМИЧЕСКИЕ АСПЕКТЫ ОПРЕДЕЛЕНИЯ ФУНКЦИЙ ГЛУБИНЫ В ПРОЦЕДУРЕ ВЫБОРА ОПТИМАЛЬНОЙ ГИПОТЕЗЫ ДЛЯ ЗАДАЧ КЛАССИФИКАЦИИ ДАННЫХ

Аннотация. Исследуются проблемы выбора оптимальной гипотезы в задачах классификации на основе класса гипотез, распределенного относительно апостериорной вероятности. Предложен метод, базирующийся на концепции относительного взвешенного среднего значения и функциях глубины, которые выполняются в пространстве функций классификации. Разработаны алгоритмы для аппроксимации относительной глубины данных и относительного взвешенного среднего значения, обеспечивающие полиномиальные приближения к полупространственным аналогам.

Ключевые слова: взвешенное среднее значение, функция относительной глубины, оптимальная гипотеза Байеса.

ВВЕДЕНИЕ

Наряду с открытыми проблемами машинного обучения одной из актуальных задач многомерной классификации является выбор оптимальной гипотезы. Существующие методы решения данной проблемы имеют ограничения относительно воспроизводимости результатов, что обусловлено стохастичностью в прогнозных моделях. Кроме того, при использовании большинства таких методов достаточно распространенным фактором является возможность применения гипотезы только из заданного класса, что вызвано характерными практическими ограничениями. Учитывая недостатки существующих методов, в данной статье предлагается новый метод решения проблемы выбора оптимальной гипотезы для эффективного обобщения на основе апостериорной вероятности. В качестве критерия для выбора гипотезы используется концепция глубины данных, в которой для нахождения наиболее глубокой функции применяются соответствующие методы аппроксимации. С учетом полученных результатов исследованы алгоритмические аспекты определения функций глубины и предложены алгоритмы для равномерной аппроксимации глубины на общем классе функций, а также алгоритмы для аппроксимации взвешенного среднего значения.

ФУНКЦИИ ГЛУБИНЫ В ПРОЦЕДУРЕ ВЫБОРА ОПТИМАЛЬНОЙ ГИПОТЕЗЫ

Суть нового метода заключается в анализе функций глубины, функционирующих в сопряженном пространстве (пространстве функций классификации), вместо их исследования в пространстве характеристик. В данном случае функция глубины, определяющая согласованность функции h со взвешенным мажоритарным голосованием на z, всегда будет иметь высокую степень согласованности с ее функцией прогнозирования в классе H.

Рассмотрим модель $z \in \mathrm{Z}$. Относительная глубина функции h относительно P определяется как

$$E_{\mathrm{P}}(h|\, z) = \underset{c \approx \mathrm{P}}{P}[c(z) = h(z)],$$

где H — класс функций, P — вероятностная мера на H. Заметим, что в общем случае относительная глубина функции h относительно P определяется как

$$E_{\mathrm{P}}(h) = \inf_{z \in \mathrm{Z}} E_{\mathrm{P}}(h|\, z) = \inf_{z \in \mathrm{Z}} \underset{c \approx \mathrm{P}}{P}[c(z) = h(z)].$$

Введем понятие относительного взвешенного среднего значения на основе класса функций H, на котором определена вероятностная мера P. При условии, что $E_{\mathrm{P}}(h) \le E_{\mathrm{P}}(h^*)$, величина h^* является относительным взвешенным средним значением на H относительно P $\forall\, h \in \mathrm{H}$. Отметим, что взвешенное среднее значение всегда существует для каждой вероятностной меры P только при условии, что класс H замкнут. Несмотря на то что инфимумом по всем точкам $z \in \mathrm{Z}$ является глубина $E_{\mathrm{P}}(h)$ и большинство исследуемых моделей имеют достаточно большую глубину, будем допускать существование моделей $z \in \mathrm{Z}$, имеющих небольшую глубину.

Далее предположим, что ε — вероятностная мера на Z, а $\omega \ge 0$. Для ослабления инфимума в определении глубины введем понятие ω-эфферентной глубины h относительно P и ε, определяемое как

$$E_{\mathrm{P}}^{\omega,\varepsilon}(h) = \sup_{Z' \subseteq \mathrm{Z},\, \varepsilon(Z') \le \omega} \inf_{z \in \mathrm{Z} \setminus Z'} E_{\mathrm{P}}(h|\, z),$$

где H — класс функций, P — вероятностная мера на H. Отметим, что функция ω-эфферентной глубины устанавливает высокую степень согласованности в классе H функции h на всем множестве моделей с вероятностной массой ω.

Рассмотрим частный случай, когда класс гипотез состоит из линейных классификаторов. Будем использовать линейные пороговые функции, являющиеся модификацией линейных классификаторов для определения факта существования функции глубины, а также того, что полупространственная глубина — частный случай относительной глубины [1].

Определим единичную сферу как $\mathrm{W} = \{z \in \mathrm{R}^r : ||\, z|| = 1\}$, где $\mathrm{H} = \mathrm{R}^r$ и $\mathrm{Z} = \mathrm{W} \times \mathrm{R}$ такие, что $h \in \mathrm{H}$ выполняется на $z = (z_b, z_\lambda) \in \mathrm{Z}$ с $h(z) = \mathrm{sign}(h \cdot z_b - z_\lambda)$. В данном случае модель $z \in \mathrm{Z}$ можно определить как комбинацию направления z_b и смещения z_λ.

Теорема 1. Пусть H — класс линейных пороговых функций на Z, где $\mathrm{Z} = \mathrm{W} \times \mathrm{R}$. Пусть $v(h)$ — такая функция плотности, что $v(h) = \dfrac{1}{Y}\exp(-\Omega(h))$, при которой P является вероятностной мерой на H, а $\Omega(h)$ — выпуклой функцией. Кроме того, пусть $h^* = \Omega_{h \approx \mathrm{P}}[h]$. Тогда имеет место неравенство $E_{\mathrm{P}}(h^*) \ge 1/e$.

Доказательство. Очевидно, что P является логарифмически вогнутой функцией тогда и только тогда, когда v — логарифмически вогнутая функция. Поскольку в условиях теоремы $v(h)$ — логарифмически вогнутая функция, то P также логарифмически вогнутая функция. Используя теорему о «медианном избирателе», можно утверждать, что $\forall\, z\; E_{\mathrm{P}}(h|\, z) \ge 1/e$, если h — центр тяжести P. Отсюда следует, что центр тяжести P имеет глубину не менее $1/e$, где e — число Эйлера. Итак, центр тяжести находится в H, поскольку $\mathrm{H} = \mathrm{R}^r$.

Теорема доказана.

Отметим, что использование выпуклых оценочных функций $\Omega(h) = \sum_{i=1}^{m} j(h(z_i), x_i) + d(h)$, в которых как функция потерь, так и функция регуляризации выпуклые, имеет широкое распространение в машинном обучении [2]. Поэтому наиболее глубокая точка, которая является взвешенным средним значением, имеет глубину не менее $1/e$.

Покажем, что если класс функций H замкнутый, то взвешенное среднее значение существует, а уровневые множества функций глубины выпуклые.

Проводя исследование в рамках функциональных классов, рассмотрим класс функций H, где $c, h_1, \dots, h_m \in \mathrm{H}$. При условии, если $\forall z$ существует такое $l \in 1, \dots, m$, что $c(z) = h_l(z)$, можно утверждать, что c находится в выпуклой оболочке $h_1, \dots, h_m$.

Лемма 1. Если c находится в выпуклой оболочке $h_1, \dots, h_m$, то

$$E_{\mathrm{P}}(c) \geq \min_l E_{\mathrm{P}}(h_l),$$

где H — класс функций с вероятностной мерой P. Кроме того, если ε — мера на Z, то

$$E_{\mathrm{P}}^{\omega,\varepsilon}(c) \geq \min_l E_{\mathrm{P}}^{\frac{\omega}{m},\varepsilon}(h_l),$$

где $\omega \geq 0$.

Доказательство. Определим условие, при котором функция c находится в выпуклой оболочке $h_1, \dots, h_m$. Поскольку для каждого z существует такое l, что $c(z) = h_l(z)$, имеем

$$E_{\mathrm{P}}(c \mid z) = E_{\mathrm{P}}(h_l \mid z) \geq \min_l E_{\mathrm{P}}(h_l),$$

откуда следует $E_{\mathrm{P}}(c) \geq \min_l E_{\mathrm{P}}(h_l)$.

Предположим, что $\Psi = \{z : E_{\mathrm{P}}(c \mid z) \leq E_{\mathrm{P}}^{\omega,\varepsilon}(c)\}$ и $\Psi_l = \{z \in \Psi : h_l(z) = c(z)\}$ для $\omega > 0$ и $l = 1, \dots, m$ соответственно.

Можно утверждать, что $\bigcup_l \Psi_l = \Psi$, поскольку функция c находится в выпуклой оболочке $h_1, \dots, h_m$. Следовательно, имеет место неравенство

$$\sum_l \varepsilon(\Psi_l) \geq \varepsilon(\Psi) \geq \omega.$$

В результате получаем

$$E_{\mathrm{P}}^{\omega,\varepsilon}(c) \geq E_{\mathrm{P}}^{\frac{\omega}{m},\varepsilon}(h_l) \geq \min_l E_{\mathrm{P}}^{\frac{\omega}{m},\varepsilon}(h_l),$$

поскольку существует такое l, что $\varepsilon(\Psi_l) \geq \omega / m$. Лемма доказана.

Далее исследуем условия существования взвешенного среднего значения, когда класс функций является замкнутым.

Отметим, что класс функций H замкнутый, если для каждой последовательности $h_1, h_2, \ldots \in \mathrm{H}$ существует такое $h^* \in \mathrm{H}$, что для каждого $z \in \mathrm{Z}$, если $\lim_{m\to\infty} h_m(z)$ существует, то $h^*(z) = \lim_{m\to\infty} h_m(z)$.

Теорема 2. Если класс функций H замкнутый, то взвешенное среднее значение H существует относительно произвольной вероятностной меры P.

Доказательство. Пусть $h^* \in \mathrm{H}$ — граница ряда $h_1, h_2, \ldots$ Предположим также, что существует такое h_m, что $E_{\mathrm{P}}(h_m) > E - 1/m$, где $E = \sup_h E_{\mathrm{P}}(h)$. Можно утверждать, что $E_{\mathrm{P}}(h^*) = E$. Однако, поскольку E является супремумом значений глубины, выполняется неравенство $E_{\mathrm{P}}(h^*) \leq E$.

Отметим, что $\forall z \in \mathrm{Z}$ и $\forall M$ существует такое $m > M$, что $h^*(z) = h_m(z)$. Поэтому, если $E_{\mathrm{P}}(h^*) < E$, существует такое z, что $E_{\mathrm{P}}(h^* \mid z) < E$.

Таким образом,

$$E_{\mathrm{P}}(h_{m_k}) \le E_{\mathrm{P}}(h_{m_k} | z) = E_{\mathrm{P}}(h^* | z) < E,$$

поскольку существует такая подпоследовательность $m_k \to \infty$, что $h_{m_k}(z) = h^*(z)$. Заметим, что, так как $\lim\limits_{k\to\infty} E_{\mathrm{P}}(h_{m_k}) = E$, полученный результат является противоречием.

Исходя из того, что $E_{\mathrm{P}}(h^* | z) \ge E \ \forall z$, имеет место конечное неравенство $E_{\mathrm{P}}(h^*) \ge E$.

Теорема доказана.

СОГЛАСОВАННОСТЬ ГИПОТЕЗЫ ВЗВЕШЕННОГО СРЕДНЕГО ЗНАЧЕНИЯ.

С учетом того, что оценка максимума апостериорной вероятности определена только в вершине распределения, она может быть недостоверной. Далее проанализируем случай, когда на каждой модели оценка максимума апостериорной вероятности не согласовывается с оптимальной гипотезой Байеса, в то время как гипотеза взвешенного среднего значения согласовывается с ней на каждом шагу [3]. Кроме того, оптимальная гипотеза Байеса также является взвешенным средним значением, поскольку она член класса гипотез. Отсюда следует, что оценка максимума апостериорной вероятности не может быть прогнозируемой.

Предположим, что множество из M дискретных элементов, проиндексированных целыми числами $1, \dots, M$, является выборочным пространством Z, т.е. $\mathrm{Z} = \{1, \dots, M\}$. Класс функций H состоит из $M+2$ функции, где функции h_i определяются следующим образом: $\forall i \in \{1, \dots, M\}$ $h_i(z) = 1$, если $z \equiv i$, и $h_i(z) = 0$ в противном случае. Также класс функций H содержит постоянные функции h_0 и h_{M+1}, принимающие значения 0 и 1 соответственно для каждого входа.

Рассмотрим апостериорную вероятность

$$\mathrm{P}\{h_{M+1}\} = \frac{1-\mu/2}{Y},\ \mathrm{P}\{h_0\} = \frac{\mu/2}{Y},\ \mathrm{P}\{h_z\} = \frac{1-\mu}{Y}\ \ \forall i \ne z,\ \mathrm{P}\{h_i\} = \frac{\mu}{Y},$$

где $0 < \mu < 1/2$ и $M > \dfrac{2}{\mu} - 1$.

Таким образом, оценкой максимума апостериорной вероятности является h_{M+1}. Однако согласно P вероятность того, что меткой i ($\forall i$) является $x = 1$, равна $\dfrac{2 - \frac{3}{2}\mu}{Y}$, а в случае $x = 0$ равна $\dfrac{\left(M - \frac{1}{2}\right)\mu}{Y}$. В результате можно утверждать, что оценка Байеса — функция h_0, поскольку в соответствии с P метка $x = 0$ имеет более высокую вероятность, чем метка $x = 1$.

Отметим, что оптимальная гипотеза Байеса также является взвешенным средним значением, поскольку находится в классе H. В данном случае на целом выборочном пространстве оценка максимума апостериорной вероятности не согласовывается со взвешенным средним значением с оптимальной гипотезой Байеса.

Минимизация апостериорной вероятности — более эффективная оценка, чем максимизация апостериорной вероятности. Это обусловлено тем, что оптимальная гипотеза Байеса h_0 имеет минимальную плотность в распределении P. Кроме того, апостериорную вероятность можно получить в результате выполне-

ния следующей процедуры. Предположим такую априорную вероятность на H, что $\Pi(h_i) = 1/M + 2$, а также такой зашумленный аналог, что для $i = 1, \ldots, M$ результат h_i меняется с вероятностью $0 < \mu < 1/2$, тогда как для h_0 и h_{M+1} результат меняется с вероятностью $\mu/2$.

В ходе исследования было установлено, что пороговая точка гипотезы взвешенного среднего значения ограничена по нижней границе функцией от ее глубины.

Далее проведем сравнение пороговой точки гипотезы взвешенного среднего значения и пороговой точки оценки максимума апостериорной вероятности. Полагаем, что пороговая точка оценки максимума апостериорной вероятности равна нулю для непрерывных классов.

Предположим, что P — такая мера Лебега, что v является функцией плотности P и ограничена некоторым бесконечным N. Заметим, что такое предположение сделано в целях эффективного определения оценки максимума апостериорной вероятности. Кроме того, рассмотрим P′ с функцией плотности $v'(h) = N + 1$, если $h = h_0$, и $v'(h) = v(h)$ в противном случае, где $h_0 \in$ H. Отметим, что оценка максимума апостериорной вероятности для P′ равна h_0, в то время как общее расстояние вариации между P и P′ равно нулю [4]. В результате для определения h_0 в качестве оценки максимума апостериорной вероятности можно ввести нулевую меру P $\forall\, h_0$. Отсюда следует вывод, что пороговая точка оценки максимума апостериорной вероятности равна нулю.

АЛГОРИТМЫ ОПРЕДЕЛЕНИЯ ФУНКЦИЙ ГЛУБИНЫ

Предлагается эффективный алгоритм оценки глубины, который принимает в качестве входа две выборки: $W = \{z_1, \ldots, z_g\}$ и $Q = \{h_1, \ldots, h_m\}$, первая из которых является выборкой точек из генеральной совокупности Z, а вторая — выборкой функций из класса функций H. На первом этапе алгоритм оценивает глубину заданной функции h на элементах $z_1, \ldots, z_g$, затем использует минимальное значение в качестве оценки общей глубины. Отметим, что глубина элемента z_i оценивается путем подсчета компонент функций $h_1, \ldots, h_m$. Итак, приведем алгоритм оценки глубины, входом которого является выборка $W = \{z_1, \ldots, z_g\}$, где $z_i \in$ Z, выборка $Q = \{h_1, \ldots, h_m\}$, где $h_l \in$ H, и функция h, а выходом — функция $\overline{E}_Q^W(h)$, которая является приближением для глубины h:

1) для $i = 1, \ldots, g$ вычислить $\overline{E}_Q(h \mid z_i) = \frac{1}{m}\sum_l 1_{h_l(z_i)=h(z_i)}$; 2) возвратить $\overline{E}_Q^W(h) = \min_i \overline{E}(h \mid z_i)$. Отметим, что эмпирической глубиной h является значение $\overline{E}_Q^W(h)$, которое возвращено алгоритмом оценки глубины. Кроме того, данный алгоритм позволяет получить эффективные оценки истинной глубины [5].

Предположим, что C — некоторое множество подмножеств в Z, а ε — вероятностная мера, определенная на генеральной совокупности Z. Можно определить μ-сеть как такое конечное подмножество $G \subseteq$ Z, что $\forall\, d \in C$, если $\varepsilon(d) \geq \mu$, то $G \cap d \neq \varnothing$.

Далее приведем теорему о равномерной сходимости глубины, что является важным фактором при нахождении взвешенного среднего значения. Из теоремы следует, что эмпирическая глубина — эффективная оценка истинной глубины, если h_l отбираются из вероятностной меры P, а z_i — из распределения по Z. Кроме того, данная оценка равномерно эффективна на всех функциях $h \in$ H.

Теорема 3. Пусть ε — вероятностная мера на Z, $\mu, \omega > 0$, а P — вероятностная мера на H. Пусть функция h_ω ($\forall\, h \in \mathrm{H}$) такова, что $h_\omega(z) = 1$, если $E_{\mathrm{P}}(h|z) \le$ $\le E_{\mathrm{P}}^{\omega,\varepsilon}(h)$, и $h_\omega(z) = -1$ в противном случае. Определим $f(r,t) = \sum_{i=0}^{r} \binom{t}{i}$, если $r < t$, и $f(r,t) = 2^t$ в противном случае. Также предположим, что при условии, что H_ω имеет конечную размерность Вапника–Червоненкиса $r < \infty$, $\mathrm{H}_\omega = \{h_\omega\}_{h\in\mathrm{H}}$. Если W и Q выбраны случайным образом из ε^g и P^m соответственно, где $g \ge 8/\omega$, то с вероятностью

$$1 - g\exp(-2m\mu^2) - f(r,2g)2^{1-\omega g/2}$$

имеет место неравенство

$$\forall\, h \in \mathrm{H}, \quad E_{\mathrm{P}}(h) - \mu \le E_{\mathrm{P}}^{0,\varepsilon}(h) - \mu \le \overline{E}_Q^W(h) \le E_{\mathrm{P}}^{\omega,\varepsilon}(h) + \mu,$$

где $\overline{E}_Q^W(h)$ — эмпирическая глубина, которая вычисляется по алгоритму измерения глубины.

Доказательство. Как было показано, с вероятностью, большей или равной $1 - f(r,2g)2^{1-\omega g/2}$, случайная выборка $W = \{z_i\}_{i=1}^g$ является ω-сетью для $\{h_\omega\}_{h\in\mathrm{H}}$. Используем h_ω для обозначения как функции, так и подмножества Z, включающего все $z \in \mathrm{Z}$, для которых $E_{\mathrm{P}}(h|z) \le E_{\mathrm{P}}^{\omega,\varepsilon}(h)$.

Можно утверждать, что $\forall\, h \in \mathrm{H}$ имеет место выражение $\exists\, i \in [1,\dots,g]$ $s.t.\, z_i \in h_\omega$, так как $\forall\, h \in \mathrm{H}$ выполняется неравенство $\varepsilon(h_\omega) \ge \omega$. Поскольку $E_{\mathrm{P}}(h|z_i) \le E_{\mathrm{P}}^{\omega,\varepsilon}(h)$, что следует из $z_i \in h_\omega$, получаем $\forall\, h \in \mathrm{H}$

$$E_{\mathrm{P}}(h) \le \min_i E(h|z_i) \le E_{\mathrm{P}}^{\omega,\varepsilon}(h)$$

с вероятностью $1 - f(r,2g)2^{1-\omega g/2}$ при случайном выборе $z_1,\dots,z_g$.

Далее, используя неравенство Хёфдинга для фиксированных z_i, имеем

$$\underset{h_1,\dots,h_m}{P}\left[\left|\frac{1}{m}|\,h_l : h_l(z_i) = 1| - \varepsilon\{h : h(z_i) = 1\}\right| > \mu\right] \le 2\exp(-2m\mu^2),$$

где $h_1,\dots,h_m$ — независимая одинаково распределенная выборка из P.

Итак, $\forall\, i$ получаем

$$\left|\frac{1}{m}|\,h_l : h_l(z_i) = 1| - \varepsilon\{h \in \mathrm{H} :\, h(z_i) = 1\}\right| \le \mu$$

с вероятностью $1 - g\exp(-2m\mu^2)$.

Кроме того, $\forall\, i$ имеет место неравенство

$$\left|\frac{1}{m}|\,h_l : h_l(z_i) = -1| - \varepsilon\{h \in \mathrm{H} : h(z_i) = -1\}\right| \le \mu.$$

Учитывая полученные результаты, можно утверждать, что $\forall\, h \in \mathrm{H}$

$$\overline{E}_Q^W(h) \le E_{\mathrm{P}}^{\omega,\varepsilon}(h) + \mu$$

с вероятностью не менее $1-g\exp(-2m\mu^2)-f(r,2g)2^{1-\mu g/2}$ при случайном выборе $z_1,\dots,z_g$ и $h_1,\dots,h_m$.

Заметим, что с вероятностью 1 в выборке не будет существовать такого i, что $E_{\mathrm{P}}(h|z_i)< E_{\mathrm{P}}^{0,\varepsilon}(h)$. Отсюда следует, что $\forall h\in \mathrm{H}\ E_{\mathrm{P}}^{0,\varepsilon}(h)-\mu\le \overline{E}_Q^W(h)$, а также выполняется неравенство $E_{\mathrm{P}}(h)\le E_{\mathrm{P}}^{0,\varepsilon}(h)$.

Теорема доказана.

Исходя из теоремы 3, можно сделать вывод, что расчетная глубина равномерно сходится к истинной глубине.

Теорема 4. Пусть h_ω такое, что $h_\omega(z)=1$, если $E_{\mathrm{P}}(h|z)< E$, и $h_\omega(z)=-1$ в противном случае $\forall h\in \mathrm{H}$. Также предположим, что $E=\sup_{h\in \mathrm{H}}E_{\mathrm{P}}(h)$ и $\mathrm{H}_\omega=\{h_\omega\}_{h:E_{\mathrm{P}}^{\omega,\varepsilon}(h)<E}$. Тогда размерность Вапника–Червоненкиса H_ω ограничена сверху размерностью Вапника–Червоненкиса H.

Доказательство. Для каждой последовательности $x\in\{\pm1\}^n\ \exists h^x$ такое, что h_ω^x генерирует метки x на $z_1,\dots,z_n$ при условии, что $z_1,\dots,z_n$ модифицированы H_ω. Необходимо показать, что выборка модифицирована H, поскольку функции h^x и $h^{x'}$ генерируют различные метки на $z_1,\dots,z_n\ \ \forall x\ne x'$.

Предположим, что $x\ne x'$, $x_i=1$ и $x'_i=-1$. Отсюда следует, что z_i такое, что

$$E_{\mathrm{P}}(h^x|z_i)< E\le E_{\mathrm{P}}(h^{x'}|z_i).$$

При условии $h^x(z_i)\ne h^{x'}(z_i)$ имеем $E_{\mathrm{P}}(h^x|z_i)\ne E_{\mathrm{P}}(h^{x'}|z_i)$. Заметим, что данный результат вытекает из определения глубины на точке z_i.

Итак, выборка $z_1,\dots,z_n$ является модифицированной H, исходя из ее модификации H_ω. Поэтому можно сделать вывод, что размерность Вапника–Червоненкиса H_ω ограничена размерностью Вапника–Червоненкиса H.

Теорема доказана.

АЛГОРИТМ ПРИБЛИЖЕНИЯ ВЗВЕШЕННОГО СРЕДНЕГО ЗНАЧЕНИЯ

Проанализировав методы определения глубины, установили, что расчетная глубина принимает точные значения равномерно для всех функций $h\in \mathrm{H}$ с достаточно большой вероятностью при условии, что выборки W и Q большие. Кроме того, имеем $h=\arg\max_{h\in \mathrm{H}}E_{\mathrm{P}}(h)$, откуда следует, что относительное взвешенное среднее значение является функцией h, которая максимизирует глубину.

На основе полученных результатов строим алгоритм для приближения относительного взвешенного среднего значения, а также алгоритм для нахождения функции h, максимизирующей эмпирическую глубину, т.е. $h=\arg\max_{h\in \mathrm{H}}\overline{E}_Q^W(h)$.

Итак, предположим, что $W=\{z_i\}_{i=1}^g$. Отметим, что функция с большой эмпирической глубиной будет согласовываться с большинством характеристик данных точек. Однако иногда имеют место случаи, когда такой функции не существует и возникает необходимость нахождения гипотезы, которая не согласовывается с большинством характеристик на соответствующих моделях [6]. В таких случаях эмпирическая глубина будет принимать большее значение, если большинство характеристик на таких элементах данных доминирует с небольшим отрывом. Для решения данной проблемы используем выборку функций $Q=\{h_l\}_{l=1}^m$ для вычисления большинства характеристик каждого z_i и часть s_i функций, которые не согласовываются с большинством характеристик.

Предложенный подход базируется на нахождении функции, которая согласовывается с большинством характеристик всех элементов в W. При условии, если такой функции не существует, удаляем элемент данных, для которого s_i принимает наибольшее значение, и находим функцию, которая согласовывается с большинством характеристик других элементов данных. Процедура выполняется до тех пор, пока не будет найдена согласованная функция. Данная функция является максимизирующей оценкой $\overline{E}_Q^W(h)$, поэтому в алгоритме приближения взвешенного среднего значения такая процедура ускоряется с помощью двоичного поиска [7].

Отметим, что метод двоичного поиска уменьшает сложность алгоритма до $O(mg + g\log(g) + g^q \log(g))$, в то время как линейный поиск требует $O(mg + g\log(g) + g^{q+1})$ операций. Указанная сложность обеспечивается при условии, что алгоритм согласованности требует $O(g^q)$ операций для некоторого q при работе на выборке размера g.

Далее представлен алгоритм приближения взвешенного среднего значения, преимуществом которого является использование модели согласованности вместо модели минимизации эмпирической ошибки. Отметим, что условием выполнения алгоритма приближения взвешенного среднего значения считается лишь доступ к модели, которая способна находить согласованную гипотезу, если таковая существует. Указанная особенность алгоритма позволяет уменьшить сложность минимизации и аппроксимации эмпирической ошибки.

Алгоритм приближения взвешенного среднего значения можно представить следующим образом. На вход алгоритма подается выборка $W = \{z_1, \dots, z_g\} \in \mathrm{Z}^g$ и выборка $Q = \{h_1, \dots, h_m\} \in \mathrm{H}^m$, где $h_l \in \mathrm{H}$. Также имеет место алгоритм Σ, возвращающий согласованную функцию, если таковая существует. На выходе получаем функцию $h \in \mathrm{H}$, приближающую относительное взвешенное среднее значение с его оценкой глубины $\overline{E}_Q^W(h)$.

Алгоритм состоит из следующих шагов: 1) $\forall i = 1, \dots, g$ вычисляем $\pi_i^+ = \frac{1}{m} |\{l : h_l(z_i) = 1\}|$ и $s_i = \min\{\pi_i^+, 1 - \pi_i^+\}$; 2) классифицируем $z_1, \dots, z_g$ таким образом, что $s_1 \geq s_2 \geq \dots \geq s_n$; 3) $\forall i = 1, \dots, g$ пусть $x_i = 1$, если $\pi_i^+ \geq 0.5$, и пусть $x_i = -1$ в противном случае; 4) используем двоичный поиск для нахождения i^*, являющегося наименьшим из i, для которого алгоритм Σ может найти согласованную функцию h на выборке $W^i = \{(z_t, x_t)\}_{t=i}^g$; 5) если $i^* \equiv 1$, возвращаем h и глубину $\overline{E} = 1 - s_1$, иначе возвращаем h и глубину $\overline{E} = s_{i^*} - 1$.

Отметим, что процедура определения гипотезы, аппроксимирующей гипотезу с минимальной эмпирической ошибкой, является NP-сложной задачей [8]. Однако согласованную гипотезу можно найти за полиномиальное время на основе линейного программирования с использованием линейного классификатора.

АНАЛИЗ АЛГОРИТМА ПРИБЛИЖЕНИЯ ВЗВЕШЕННОГО СРЕДНЕГО ЗНАЧЕНИЯ

Лемма 2. Гипотеза h и глубина $\overline{E}$ всегда будут возвращаться алгоритмом приближения взвешенного среднего значения.

Доказательство. Необходимо показать, что $\exists i$ такое, что алгоритм Σ будет возвращать функцию согласованности h относительно W^i, т.е. установить, что двоичный поиск будет всегда находить $i^* \leq g$.

Можно утверждать, что, поскольку $W^g = \{(z_g, x_g)\}$, выборка будет содержать только один элемент данных z_g с такой меткой x_g, что по крайней мере половина функций в Q таковы, что $h_l(z_g) = x_g$. Отсюда следует, что существует функция h, согласованная с данной выборкой. Лемма доказана.

Далее приводим теорему, доказывающую корректность глубины, вычисленной с помощью алгоритма приближения взвешенного среднего значения.

Теорема 5. Пусть гипотеза h и глубина $\overline{E}$ возвращаются алгоритмом приближения взвешенного среднего значения. Тогда $\overline{E} = \overline{E}_Q^W(h)$.

Доказательство. Предположим, что $X(c) = \{i: c(z_i) = x_i\}$ является множеством образов, на котором функция согласовывается с меткой. Расчетная глубина функции определяется следующим образом:

$$\overline{E}_Q^W(c) = \min\left(\min_{i \in X(c)} (1 - s_i), \min_{i \notin X(c)} s_i \right).$$

Можно предположить, что $i_{\in} = \min\{i: i \in X(c)\}$ и $i_{\notin} = \max\{i: i \notin X(c)\}$. В результате определение расчетной глубины представим в виде

$$\overline{E}_Q^W(c) = \min\left((1 - s_{i_{\in}}), s_{i_{\notin}}\right),$$

учитывая упорядоченность s_i. Заметим, что $s_{i_{\notin}} = 1$, если $X(c)$ включает все i, и $s_{i_{\notin}} = 0$, если $X(c)$ пусто.

Далее предположим, что гипотеза h и вычислительная глубина $\overline{E}$ возвращаются алгоритмом приближения взвешенного среднего значения. В случае, когда $i^* = 1$, и при условии, что i^* является индексом, который возвращается двоичным поиском, величины $X(h) = [1, \ldots, g]$ и $\overline{E}_Q^W(h) = 1 - s_1$ принимают точные значения, возвращенные алгоритмом приближения взвешенного среднего значения. Иначе, если $i^* > 1$, то $i^* - 1 \notin X(h)$ при условии, что $[i^*, \ldots, g] \subseteq X(h)$.

В результате, поскольку $\forall i'$ имеет место неравенство $1 - s_{i'} \geq 0{,}5$ (с учетом $s_{i^*-1} \leq 0.5$), точным значением, возвращенным алгоритмом приближения взвешенного среднего значения, является $\overline{E}_Q^W(h) = s_{i^*-1}$.

Теорема доказана.

Как было показано, эмпирическая глубина функции — это функция множества точек, на которых она согласовывается с большинством образов. Для доказательства того факта, что максимальная оценка эмпирической глубины возвращается алгоритмом приближения взвешенного среднего значения, имеет место следующая теорема.

Теорема 6. Пусть h — функция, возвращенная алгоритмом приближения взвешенного среднего значения. Тогда

$$h = \arg\max_{h \in \mathrm{H}} \overline{E}_Q^W(h).$$

Доказательство. При условии, если $i^* = 1$, эмпирическая глубина h максимальна. Отсюда следует, что $i^* > 1$ и $\overline{E}_Q^W(h) = s_{i^*-1}$, где i^* — значение, возвращенное двоичным поиском, а h — функция, возвращенная моделью согласованности. В случае, если $\exists\, i > i^*$ такое, что $c(z_i) \neq x_i$, выполняется неравенство

$$\overline{E}_Q^W(c) \le s_{i-1} \le s_{i^*-1} \le \overline{E}_Q^W(h),$$

где $c \in \mathrm{H}$. Учитывая тот факт, что этап двоичного поиска в алгоритме приближения взвешенного среднего значения может сгенерировать $i^* - 1$ или большее множество, необходимо, чтобы $c(z_{i^*-1}) \ne x_{i^*-1} \ \ \forall i \ge i^*$, где $c(z_i) = x_i$. В результате имеем $\overline{E}_Q^W(c) = s_{i^*-1} := \overline{E}_Q^W(h)$.

Теорема доказана.

Следующая теорема является основным результатом оценки приближения взвешенного среднего значения.

Теорема 7. Имеют место такие свойства алгоритма приближения взвешенного среднего значения:

1) предположим, что $\mu, \omega > 0$; можно утверждать, что с вероятностью не менее

$$1 - g\exp(-2m\mu^2) - f(r, 2g)2^{1-\omega g/2}$$

функция h, возвращенная алгоритмом приближения взвешенного среднего значения, такова, что

$$E_{\mathrm{P}}^{\omega,\varepsilon}(h) \ge \sup_{c \in H} E_{\mathrm{P}}^{0,\varepsilon}(c) - 2\mu \ge \sup_{c \in H} E_{\mathrm{P}}(c) - 2\mu,$$

где r — минимум между размерностью Вапника–Червоненкиса H и размерностью Вапника–Червоненкиса класса H_ω, выборка W взята из ε^g, так что $g \ge 8/\omega$, а выборка Q — из P^m;

2) если h — функция, возвращенная алгоритмом приближения взвешенного среднего значения, то $h = \arg\max_{h \in \mathrm{H}} \overline{E}_Q^W(h)$;

3) если h и $\overline{E}$ — выходы алгоритма приближения взвешенного среднего значения, то $\overline{E} = \overline{E}_Q^W(h)$;

4) алгоритм всегда будет останавливаться и возвращать функцию $h \in \mathrm{H}$ и эмпирическую глубину $\overline{E}$.

Доказательство. Для доказательства п. 1 предположим, что $E = \sup_h E_{\mathrm{P}}^{0,\varepsilon}(h)$, где h — максимальная оценка $\overline{E}_Q^W(h)$. Как было показано, с вероятностью

$$1 - g\exp(-2m\mu^2) - f(r, 2g)2^{1-\omega g/2}$$

имеем

$$\overline{E}_Q^W(h) \ge \max_c \overline{E}_Q^W(c) \ge \sup_c \overline{E}_{\mathrm{P}}^{0,\varepsilon}(c) - \mu = E - \mu$$

при условии, если r, как минимум, меньше размерности Вапника–Червоненкиса H и размерности Вапника–Червоненкиса H_ω. Кроме того, очевидно, что $\overline{E}_Q^W(h) \le E_{\mathrm{P}}^{\omega,\varepsilon}(h) + \mu$, если $\overline{E}_{\mathrm{P}}^{\omega,\varepsilon}(h) < E$. Отсюда следует, что справедливым является неравенство $E_{\mathrm{P}}^{\omega,\varepsilon}(h) \ge E$ или $E_{\mathrm{P}}^{\omega,\varepsilon}(h) \ge \overline{E}_Q^W(h) - \mu \ge E - 2\mu$.

Доказательство пп. 2–4 следует из результатов теоремы 6, теоремы 5 и леммы 2 соответственно.

Теорема доказана.

АЛГОРИТМЫ ПОЛУПРОСТРАНСТВЕННОЙ ГЛУБИНЫ И ПОЛУПРОСТРАНСТВЕННОГО ВЗВЕШЕННОГО СРЕДНЕГО ЗНАЧЕНИЯ

Заключительный этап исследования состоит в определении того, как разработанные алгоритмы можно применить к задачам вычисления функции полупространственной глубины и нахождения полупространственного взвешенного среднего значения. Поскольку полупространственная глубина может рассматриваться как частный случай относительной глубины, алгоритм оценки глубины и алгоритм приближения взвешенного среднего значения могут использоваться для вычисления полупространственной глубины и аппроксимации полупространственного взвешенного среднего значения соответственно.

Предположим, что имеет место выборка $h_1, \dots, h_m$ элементов данных в R^r, а также выборка направлений и смещений интереса [9], т.е. выборка таких z_i, что $z_i \in \mathrm{W} \times \mathrm{R}$, где W — единичная сфера. В данном случае элементы представлены в виде комбинации r-мерного единичного вектора z_i^b и смещения z_i^λ.

Алгоритм оценки полупространственной глубины позволяет использовать данные выборки для оценки полупространственной глубины элемента $h \in \mathrm{R}^r$. Алгоритм приближения полупространственного взвешенного среднего значения показывает, как применять данные выборки для приближения полупространственного взвешенного среднего значения. Заметим, что размерность Вапника–Червоненкиса для данных задач равна r.

С учетом того, что задача вычисления полупространственной глубины требует нахождения инфимума по всем возможным направлениям, алгоритм приближения находит минимум на множестве всех возможных направлений в выборке W. В данном случае выбор z_i^b проводится равномерно из единичной сферы при генерации выборки. Заметим, что выбор z_i^λ в предложенных алгоритмах следует проводить случайным образом. Однако, когда z_i^b фиксирован, нахождение минимальной глубины по всем возможным направлениям z_i^λ возможно для случая линейных функций. Данную процедуру можно выполнить путем подсчета числа h_l таким образом, что $h_l \cdot z_i^b > h \cdot z_i^b$ (или $h_l \cdot z_i^b < h \cdot z_i^b$), после чего необходимо выбрать минимальное значение.

Итак, на вход алгоритма оценки полупространственной глубины подаются: выборка $W = \{z_1, \dots, z_g\}$ такая, что $z_i \in \mathrm{W}$, выборка $Q = \{h_1, \dots, h_m\}$ такая, что $h_l \in \mathrm{R}^r$, и элемент данных $h \in \mathrm{R}^r$. На выходе алгоритма получаем $\overline{E}_Q^W(h)$, что является приближением для глубины h. Алгоритм оценки полупространственной глубины состоит из следующих шагов: 1) для $i = 1, \dots, g$ вычислить $\overline{E}_Q(h \mid z_i) = \frac{1}{m} \min\left(|h_l : h_l \cdot z_i > h \cdot z_i|, |h_l : h_l \cdot z_i < h \cdot z_i|\right)$; 2) возвратить $\overline{E}_Q^W(h) = \min_i \overline{E}(h \mid z_i)$.

Алгоритм приближения полупространственного взвешенного среднего значения позволяет определить множество случайных направлений $z_1, \dots, z_g$. При этом взвешенное среднее значение h должно занимать центральное место в каждом направлении. Кроме того, проекция h должна иметь большую одномерную глубину, т.е. быть близкой к взвешенному среднему значению проекции при проектировании $h_1, \dots, h_m$ и h на z_i. Поэтому первый шаг заключается в нахождении h с максимально возможной глубиной в каждом направлении. В случае, когда такого h не существует, необходимо уменьшить требование глубины в каждом направлении и повторить процедуру для нахождения соответствующего h.

Алгоритм приближения полупространственного взвешенного среднего значения принимает на входе выборку $W = \{z_1, \ldots, z_g\}$ такую, что $z_i \in \mathrm{W}$, и выборку $Q = \{h_1, \ldots, h_m\}$ такую, что $h_l \in \mathrm{R}^r$. Кроме того, имеем линейный программный алгоритм Σ, который находит элемент данных (если таковой существует), что согласовывается с ограничениями на заданном множестве линейных ограничений [10]. На выходе алгоритма получаем элемент данных $h \in \mathrm{R}^r$ и его оценку глубины $\overline{E}_Q^W(h)$.

Алгоритм приближения полупространственного взвешенного среднего значения состоит из следующих шагов: 1) для $i = 1, \ldots, g$ и $l = 1, \ldots, m$ вычисляем $h_l \cdot z_i$; 2) пусть $w_i^1, \ldots, w_i^m$ являются классифицированными значениями $h_l \cdot z_i$; 3) используем двоичный поиск для нахождения наименьшего $t = 0, \ldots, m/2$, для которого Σ может найти такое h, что $\forall\, i\ w_i^{\left\lfloor \frac{m}{2} \right\rfloor - t} \le h \cdot z_i < w_i^{\left\lceil \frac{m}{2} \right\rceil + t}$; 4) возвратить такое h, которое получено с помощью алгоритма Σ для наименьшего t на шаге 3.

ЗАКЛЮЧЕНИЕ

В данной работе предложен и исследован новый комплексный метод решения проблемы выбора оптимальной гипотезы на основе апостериорного доверия в классе гипотез для задач классификации данных. В рамках исследования введено понятие относительной глубины, от которого напрямую зависит обобщение классификатора. Существенное значение в решении проблемы выбора оптимальной гипотезы имела концепция относительного взвешенного среднего значения для наиболее глубокого классификатора. В процессе исследования проанализированы пороговые свойства взвешенного среднего значения, в результате чего установлена их зависимость от глубины данных. На основе полученных результатов построены эффективные алгоритмы для равномерного измерения относительной глубины и нахождения относительного взвешенного среднего значения. Предложенные алгоритмы позволяют аппроксимировать полупространственную глубину и полупространственное взвешенное среднее значение за полиномиальное время.

СПИСОК ЛИТЕРАТУРЫ

1. Jornsten R., Vardi Y., Zhang C.H. A robust clustering method and visualization tool based on data depth // Statistical data Analysis. — 2002. — P. 354–365.
2. Chacon J.I., Duong T., Wand M.P. Asymptotics for general multivariate kernel density derivative estimators // Statistics. — 2011. — **21.** — P. 810–837.
3. Seeger M. PAC-Bayesian generalisation error bounds for gaussian process classification // The Journal of Machine Learning Research. — 2003. — **3.** — P. 237–268.
4. Herbrich R., Graepel T., Campbell C. Bayes point machines // The Journal of Machine Learning Research. — 2001. — **1.** — P. 247–271.
5. Davies P.L., Gather U. Breakdown and groups // The Annals of Statistics. — 2005. — **33**, N 3. — P. 981–1027.
6. Fine S., Gilad-Bachrach R., and Shamir E. Query by committee, linear separation and random walks // Theoretical Computer Science. — 2002. — **284**, N 1. — P. 27–49.
7. Ben-David S., Eiron N., Long P.M. On the difficulty of approximately maximizing agreements // Journal of Computer and System Sciences. — 2003. — **66**, N 3. — P. 499–511.
8. Rousseeum P.J., Struyf A. Characterizing angular symmetry and regression symmetry // Journal of Statistical Planning and Inference. — 2004. — **122.** — P. 163–170.

9. Oja H., Paindaveine D. Optimal signed-rank tests based on hyperplanes // Journal of Statistical Planning and Inference. — 2005. — **135.** — P. 307–321.
10. Holmes C.C., Adams N.M. A probabilistic nearest neighbor method for statistical pattern recognition // Journal of the Royal Statistical Society. — 2002. — **64.** — P. 295–306.

Надійшла до редакції 09.03.2016

О.А. Галкін

АЛГОРИТМІЧНІ АСПЕКТИ ВИЗНАЧЕННЯ ФУНКЦІЙ ГЛИБИНИ У ПРОЦЕДУРІ ВИБОРУ ОПТИМАЛЬНОЇ ГІПОТЕЗИ ДЛЯ ЗАДАЧ КЛАСИФІКАЦІЇ ДАНИХ

Анотація. Досліджуються проблеми вибору оптимальної гіпотези в задачах класифікації на основі класу гіпотез, розподіленого відносно апостеріорної ймовірності. Запропоновано метод, який базується на концепції відносного зваженого середнього значення та функціях глибини, що виконуються у просторі функцій класифікації. Розроблено алгоритми для апроксимації відносної глибини даних та відносного зваженого середнього значення, що забезпечують поліноміальні наближення до напівпросторових аналогів.

Ключові слова: зважене середнє значення, функція відносної глибини, оптимальна гіпотеза Баєса.

O.A. Galkin

ALGORITHMIC ASPECTS OF DETERMINING THE DEPTH FUNCTIONS IN SELECTING THE OPTIMAL HYPOTHESIS FOR DATA CLASSIFICATION PROBLEMS

Abstract. The paper analyzes optimal hypothesis selection in classification problems based on the hypothesis class distributed with respect to the posterior probability. A method is proposed that is based on the concept of a relative weighted average value and depth functions operating in the space of classification functions. Algorithms are constructed to approximate the relative depth of the data and relative weighted average value providing polynomial approximation to the half-space analogs.

Keywords: weighted average value, relative depth function, optimal Bayesian hypothesis

Галкин Александр Анатольевич,
кандидат физ.-мат. наук, ассистент кафедры Киевского национального университета имени Тараса Шевченко, e-mail: galkin.o.a@gmail.com.