Вероятность ошибки и вычислительная сложность классификации объектов в пространстве многоуровневых представлений

М.М. Ланге
Федеральный исследовательский центр «Информатика и управление» РАН
Москва, Россия
lange_mm@ccas.ru

С.В. Парамонов
Федеральный исследовательский центр «Информатика и управление» РАН
Москва, Россия
psvpobox@gmail.com

***Аннотация*—Исследуется схема классификации объектов, заданных многоуровневыми древовидными представлениями, в терминах зависимости вероятности ошибки от количества обрабатываемой информации. Используя разделяющие функции на последовательных уровнях представления объектов, предлагается стратегия направленного поиска решения. Вводится избыточность вероятности ошибки относительно нижней границы, зависящая от параметра стратегии поиска решения. Экспериментальные оценки, полученные на множествах изображений лиц и подписей, демонстрируют размен между вычислительной сложностью и качеством классификации при различных значениях параметра стратегии поиска решения.**

Ключевые слова— классификация, вероятность ошибки, взаимная информация, разделяющие функции, многоуровневое представление, вычислительная сложность.

1. Введение

В работе [1] получена аналитическая нижняя граница вероятности ошибки классификации на множестве объектов с заданной метрикой, и для любого набора разделяющих функций введена избыточность вероятности ошибки относительно нижней границы. Найденная граница является обобщением границы Шеннона для скорости кодирования символов дискретного источника с допустимой погрешностью в метрике Хемминга [2], когда символы источника передаются по каналу без памяти с искажениями. Приведенные в [1] численные оценки вероятности ошибки и избыточности получены для разделяющих функций экспоненциального типа на множествах древовидных представлений изображений лиц и подписей при переборном поиске решения.

В настоящей работе исследуются характеристики качества и вычислительной сложности классификации в пространстве древовидных представлений информативных объектов, заданных изображениями, с использованием разделяющих функций на всех уровнях представления. Для параметрической стратегии направленного поиска решения приводится оценка вычислительного выигрыша по сравнению с полным перебором. Предлагаемый подход может быть полезен для получения соотношения характеристик качества и быстродействия в схемах поиска приближенного ближайшего соседа на множестве объектов, заданных изображениями с высоким разрешением [3].

Исследуется схема классификации $\Omega \to \mathbf{X} \to \hat{\Omega}$, в которой Ω и $\hat{\Omega}$ – множества меток и их оценок для $c \geq 2$ классов по объектам из множества $\mathbf{X}$. При заданных априорных вероятностях на множестве Ω и условных по классам вероятностях на множестве $\mathbf{X}$, безусловные вероятности объектов образуют распределение $\mathrm{P} = \{P(\mathbf{x}), \mathbf{x} \in \mathbf{X}\}$. Классификация выполняется в пространстве многоуровневых представлений, которое образовано набором $\mathbf{X}^L = \{\mathbf{X}_l\}_{l=1}^L$ представлений множества $\mathbf{X}$ с нарастающим разрешением. В пространстве $\mathbf{X}^L$ каждый объект $\mathbf{x} \in \mathbf{X}$ задан последовательностью представлений $\mathbf{x}^L = (\mathbf{x}_1, ..., \mathbf{x}_L)$, в которой $\mathbf{x}_l \in \mathbf{X}_l, l = 1,,,,L$ – поддерево глубины l в бинарном дереве $\mathbf{x}_L$ глубины L [1].

2. Основные результаты

Используя квадратичную метрику на множествах представлений $\mathbf{X}_l, l = 1,...,L$, введены наборы разделяющих функций

$$\mathrm{G}_l = \{g_{jl}(\mathbf{x}), \mathbf{x} \in \mathbf{X}\}_{j=1}^c, l = 1,...,L, \quad (1)$$

которые являются «мерами правдоподобия» оценок классов по предъявляемым объектам. Безусловное распределение P и условные по объектам распределения

$$\mathrm{Q}_l = \left\{Q_l(j \mid \mathbf{x}) = g_{jl}(\mathbf{x}) \Big/ \sum\nolimits_{i=1}^c g_{il}(\mathbf{x})\right\}_{j=1}^c, l = 1,...,L$$

вероятностей оценок меток классов позволяют ввести среднюю взаимную информацию $I_{\mathrm{G}_l}(\mathbf{X}; \hat{\Omega})$ и среднюю вероятность ошибки $E_{\mathrm{G}_l}(\mathbf{X}; \hat{\Omega})$ для разделяющих функций (1) на уровнях представления $l = 1,...,L$.

Для направленного поиска решения по объекту $\mathbf{x}$ с последовательностью представлений $\mathbf{x}^L$ предложена параметрическая стратегия направленного отбора меток классов с наибольшими значениями разделяющих функций на уровнях $l = 1,...,L$. Согласно такой стратегии, количество меток l-го уровня, среди которых отбираются наиболее правдоподобные по функциям (1)

метки $l+1$-го уровня, определяется экспоненциально убывающей функцией

$$c_l = \left\lfloor c2^{-\alpha(l-1)} \right\rfloor, l = 1,...,L \ , \qquad (2)$$

где $\alpha = (L-1)^{-1}\log(c/c^*) < 1$ и c^* – параметр, принимающий целые значения на отрезке $[1,c]$. На L-м уровне принимается решение j^* по наибольшему значению $g_{j^*L}(\mathbf{x})$ среди c^* меток, отобранных на $(L-1)$-м уровне. В случае $c^* = c$ предлагаемая стратегия эквивалентна полному перебору.

С учетом структуры бинарных представляющих деревьев и функции (2), вычислительная сложность поиска решения по объекту в терминах количества обрабатываемых вершин в представляющем дереве удовлетворяет оценке

$$C_\alpha = \sum_{l=1}^{L} c_l 2^l \le 2c\frac{2^{(1-\alpha)L}-1}{2^{(1-\alpha)}-1}. \qquad (3)$$

Оценка (3) обеспечивает вычислительный выигрыш

$$\frac{C_{\alpha=0}}{C_{\alpha>0}} \ge \frac{2^L-1}{2^L}\left(2-\left(c/c^*\right)^{1/(L-1)}\right)c/c^* \qquad (4)$$

направленного поиска решения по сравнению с полным перебором. При достаточно больших значениях $L >> 1+\log(c/c^*)$ правая часть в (4) составляет величину порядка c/c^*.

Семейство наборов разделяющих функций $\mathrm{G}^L = \{\mathrm{G}_l\}_{l=1}^L$ с весами $w_l = \left(\log c_l - \log c_{l+1}\right)/\log c$, $l = 1,...,L$, где $c_1 = c$ и $c_{L+1} = 1$, порождает усредненные по уровням характеристики

$I_{\mathrm{G}^L} = \sum_{l=1}^{L} w_l I_{\mathrm{G}_l}(\mathbf{X};\hat{\Omega})$ и $E_{\mathrm{G}^L} = \sum_{l=1}^{L} w_l E_{\mathrm{G}_l}(\mathbf{X};\hat{\Omega})$,

которые зависят от параметра c^*. Указанным характеристикам соответствует избыточность

$$r_{\mathrm{G}^L} = E_{\mathrm{G}^L} - E_{\min}(I_{\mathrm{G}^L})$$

средней вероятности ошибки E_{G^L} относительно значения нижней границы $E_{\min}$ при среднем количестве обрабатываемой информации I_{G^L}.

Используя на множествах представлений изображений лиц [4] и подписей [5] слабые одномодовые разделяющие функции экспоненциального типа и композиции таких функций для ансамбля указанных источников, получены численные оценки характеристик $I_{\mathrm{G}^L}, E_{\mathrm{G}^L}$ и $E_{\min}(I_{\mathrm{G}^L})$ при различных значениях c^*. Данные источников содержали по 1000 объектов от $c = 25$ персон (классов), по 40 объектов в каждом классе. Для представлений объектов использованы бинарные деревья глубины $L = 8$. Параметры разделяющих функций вычислены в режиме скользящего контроля по схеме «leave-one-out».

Графики указанных характеристик, вычисленных для ансамбля лиц и подписей в диапазоне значений $10 \le c^* \le 25$, даны на рисунке 1. Полученные зависимости демонстрируют слабое увеличение вероятности ошибки и избыточности с уменьшением параметра c^* и, следовательно, с ростом вычислительного выигрыша, определенного в (4).

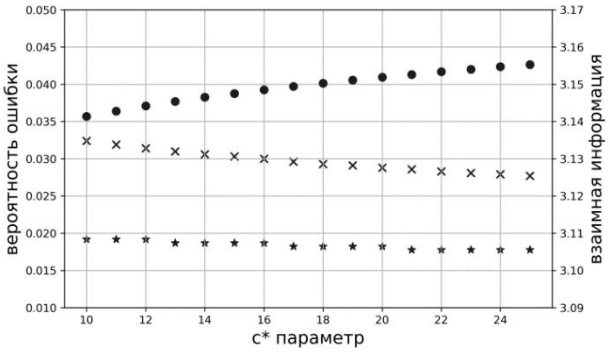

Рис.1. Численные оценки средней вероятности ошибки (×), средней взаимной информации (●) и нижней границы вероятности ошибки (*)

Приближение к нижней границе может быть достигнуто за счет использования многомодовых разделяющих функций для отдельных источников, а также путем увеличения количества источников в ансамбле.

3. Заключение

В рамках теоретико-информационной модели исследованы характеристики качества и вычислительной сложности классификации на множестве древовидных представлений объектов с многоуровневым разрешением. Показана возможность уменьшения времени принятия решения по предъявляемому объекту за счет некоторой потери качества классификации. Универсальность предложенного подхода позволяет применить его для анализа эффективности многоклассовых SVM классификаторов [6] в пространстве многоуровневых представлений с разделяющими функциями сигмоидального типа.

Литература

[1] Lange, M.M. On a Lower Bound to Classification Error Probability in an Ensemble of Data Sources / M.M. Lange, S.V. Paramonov // IEEE Proceedings. – 2021. – P. 1-6. DOI: 10.1109/ ITNT52450.2021.9649088

[2] Gallager, R.G. Information Theory and Reliable Communication – Wiley & Sons, 1968. – 588 c.

[3] Andoni, A. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions / A. Andoni, P. Indyk // Communications of the ACM. – 2008. – Vol. 51(1). – P .117-122. DOI: 10.1145/1327452.1327494.

[4] Distance matrices for face dataset [Electronic resource]. – Access mode: http://sourceforge.net/projects/distance-matrices-face (2020, June).

[5] Distance matrices for signature dataset [Electronic resource]. – Access mode: http://sourceforge.net/projects/distance-matrices-signature (2020, June).

[6] Sueno, H.T. Multi-class document classification using support vector machine (SVM) based on improved naïve Bayes vectorization technique / H.T. Sueno,.B.D. Gerardo, R.P. Medina // International Journal of Advanced Trends in Computer Science and Engineering. – 2020. – Vol. 9(3). – P. 3937-3944. DOI: 10.30534/ijatcse/ 2020/216932020.