О комплексировании разделяющих функций для повышения точности классификации данных

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 точность классификации данных в терминах соотношения вероятности ошибки и количества обрабатываемой информации для различных схем комплексирования. Рассматриваются наборы слабых разделяющих функций и способы их комплексирования на множестве данных одной модальности и на ансамбле данных от источников различной модальности. Экспериментально показано уменьшение вероятности ошибки и ее избыточности относительно теоретико-информационной нижней границы с увеличением количества обрабатываемой информации.**

Ключевые слова — классификация, вероятность ошибки, взаимная информация, энтропия, разделяющая функция, ансамбль, комплексирование, избыточность.

1. Введение

В большинстве известных работ по классификации данных критерием качества решающих алгоритмов является точность, определяемая вероятностью ошибки. Этот же критерий применяется к алгоритмам, строящимся на основе схем комплексирования [1]. Учитывая, что вероятность ошибки классификации должна уменьшаться с ростом количества обрабатываемой информации, целесообразно использовать двухфакторный критерий качества, который задается вероятностью ошибки и количеством информации. Применение такого критерия позволяет оценить избыточность вероятности ошибки любого решающего алгоритма относительно теоретико-информационной нижней границы при фиксированном количестве обрабатываемой информации [2]. При этом нижняя граница определяется модификацией функции "скорость-погрешность" (rate distortion function) для кодирования дискретных сообщений с допустимой погрешностью по мере Хемминга [3]. Исследование качества решающих алгоритмов в терминах двухфакторного критерия выполняется с использованием разделяющих функций [4], которые позволяют получить оценки вероятности ошибки и количества обрабатываемой информации в форме энтропии множества решений о классах.

В настоящей работе на множестве объектов одной модальности рассматриваются слабые наборы разделяющих функций, обеспечивающие достаточно низкую точность, и способ комплексирования слабых наборов, который позволяет уменьшить вероятность ошибки за счет увеличения количества обрабатываемой информации. Демонстрируется возможность существенного повышения качества классификации на ансамбле источников путем комплексирования разделяющих функций для объектов различной модальности.

2. Формализация задачи

Рассматривается схема классификации

$$\Omega \to \mathbf{X}^M \to \hat{\Omega}, \qquad (1)$$

в которой $\Omega = \{\omega_i\}_{i=1}^{c}$ и $\hat{\Omega} = \{\omega_j\}_{j=1}^{c}$ – множества классов и их оценок, где $c \geq 2$, а $\mathbf{X}^M = \mathbf{X}_1, ..., \mathbf{X}_M$ – ансамбль множеств различной модальности, в котором каждое множество $\mathbf{X}_m, m = 1, ..., M$ содержит объекты одной модальности. В схеме (1) любой составной объект $\mathbf{x}^M = (\mathbf{x}_1, ..., \mathbf{x}_M) \in \mathbf{X}^M$ представлен набором объектов одного класса, взятых по одному $\mathbf{x}_m \in \mathbf{X}_m, m = 1, ..., M$ от каждого источника.

Пусть на множествах объектов источников заданы наборы разделяющих функций

$$\mathrm{G}(\mathbf{X}_m) = \{g_j(\mathbf{x}_m), \mathbf{x}_m \in \mathbf{X}_m\}_{j=1}^{c}, m = 1, ..., M, \qquad (2)$$

которые определяют "меру правдоподобия" оценок классов по объектам источников. Комплексирование функций (2) позволяет сформировать набор разделяющих функций на ансамбле источников

$$\mathrm{G}(\mathbf{X}^M) = \{g_j(\mathbf{x}^M), \mathbf{x}^M \in \mathbf{X}^M\}_{j=1}^{c}. \qquad (3)$$

Нормирование функций (3) дает оценки апостериорных вероятностей классов

$$\left\{Q_{\hat{\Omega}|\mathbf{X}^M}(\omega_j \mid \mathbf{x}^M) = g_j(\mathbf{x}^M) \Big/ \sum_{i=1}^{c} g_i(\mathbf{x}^M)\right\}_{j=1}^{c}, \qquad (4)$$

которые при $M = 1$ соответствуют оценкам апостериорных вероятностей по объектам источников $\mathbf{x}_m \in \mathbf{X}_m, m = 1, ..., M$.

Заданные априорные вероятности классов $\{P_\Omega(\omega_i)\}_{i=1}^{c}$ и условные по классам вероятности $\{P_{\mathbf{X}^M|\Omega}(\mathbf{x}^M \mid \omega_i)\}_{i=1}^{c}$ дают вероятности

$$P_{\mathbf{X}^M}(\mathbf{x}^M) = \sum_{i=1}^{c} P_\Omega(\omega_i) P_{\mathbf{X}^M|\Omega}(\mathbf{x}^M \mid \omega_i), \ \mathbf{x}^M \in \mathbf{X}^M \qquad (5)$$

составных объектов. Тогда для решающего алгоритма по максимуму разделяющих функций, распределения (4) и (5) позволяют вычислить вероятность ошибки

$$E_{\mathrm{G}}(\mathbf{X}^M; \hat{\Omega}) = 1 - \sum_{\mathbf{x}^M \in \mathbf{X}^M} P_{\mathbf{X}^M}(\mathbf{x}^M) \max_{j=1}^{c} Q_{\hat{\Omega}|\mathbf{X}^M}(\omega_j \mid \mathbf{x}^M) \qquad (6)$$

и энтропию множества решений

$$H_{\mathrm{G}}(\hat{\Omega}) = -\sum_{j=1}^{c} P_{\hat{\Omega}}(\omega_j) \ln P_{\hat{\Omega}}(\omega_j), \qquad (7)$$

где $P_{\hat{\Omega}}(\omega_j)$ – вероятность подмножества $\mathbf{X}_j^M \subset \mathbf{X}^M$ объектов, по которым принимается решение $\omega_j \in \hat{\Omega}$.

Нижняя граница минимума средней взаимной информации $I(\mathbf{X}^M; \hat{\Omega})$ при ограничении средней

вероятности ошибки $E(\mathbf{X}^M;\Omega) \le \varepsilon$ имеет вид монотонно убывающей функции [2]

$$\underline{R}_M(\varepsilon) = I(\mathbf{X}^M;\Omega) - h(\varepsilon - \varepsilon^{(c)}_{M\_\min}) - (\varepsilon - \varepsilon^{(c)}_{M\_\min})\ln(c-1) \quad (8)$$

на отрезке $\varepsilon^{(c)}_{M\_\min} \le \varepsilon \le \varepsilon^{(c)}_{M\_\max}$, где $I(\mathbf{X}^M;\Omega)$ – средняя взаимная информация между множествами $\mathbf{X}^M$ и Ω, $h(z) = -z\ln z - (1-z)\ln(1-z)$, $\underline{R}_M(\varepsilon^{(c)}_{M\_\min}) = I(\mathbf{X}^M;\Omega)$ и $\underline{R}_M(\varepsilon^{(c)}_{M\_\max}) = 0$. Тогда избыточность вероятности ошибки (6) относительно обращения $\underline{R}_M^{-1}(H_{\mathrm{G}})$ границы (8) в значении энтропии (7) определяется величиной

$$r_{\mathrm{G}} = \begin{cases} E_{\mathrm{G}} - \underline{R}_M^{-1}(H_{\mathrm{G}}), & H_{\mathrm{G}} \le I(\mathbf{X}^M;\Omega), \\ E_{\mathrm{G}} - \varepsilon^{(c)}_{M\_\min}, & H_{\mathrm{G}} > I(\mathbf{X}^M;\Omega). \end{cases} \quad (9)$$

Пусть для всех источников заданы слабые наборы разделяющих функций $\left\{\mathrm{G}^{(k)}(\mathbf{X}_m)\right\}_{k=1}^{K}, m = 1,...,M$, вида (2). Задача состоит в нахождении способа комплексирования слабых разделяющих функций в наборы $\mathrm{G}(\mathbf{X}_m), m = 1,...,M$, которые обеспечивают вероятность ошибки $E_{\mathrm{G}}(\mathbf{X}_m;\hat{\Omega}) < \min_{k=1}^{K} E_{\mathrm{G}^{(k)}}(\mathbf{X}_m;\hat{\Omega})$ и избыточность $r_{\mathrm{G}} < \min_{k=1}^{K} r_{\mathrm{G}^{(k)}}$, определенные в (6) и (9) при $M = 1$. Необходимо также предложить способ комплексирования разделяющих функций (2) в функции (3), который приводит к дополнительному снижению вероятности ошибки и избыточности.

3. Основные результаты

Для множества $\mathbf{X}_m$ объектов $m-$ го источника с разделяющими функциями, принимающими значения на отрезке [0,1], предложен способ комплексирования функций (2) по схеме

$$g_j(\mathbf{x}_m) = \begin{cases} \max_{k=1}^{K} g_j^{(k)}(\mathbf{x}_m), & \exists g_j^{(k)}(\mathbf{x}_m) \ge \Delta \\ \min_{k=1}^{K} g_j^{(k)}(\mathbf{x}_m), & \forall g_j^{(k)}(\mathbf{x}_m) < \Delta \end{cases} \quad (10)$$

где Δ – порог, выбираемый из условия оптимизации. Разделяющие функции (3) на ансамбле источников $\mathbf{X}^M = \mathbf{X}_1,...,\mathbf{X}_M$ заданы произведениями

$$g_j(\mathbf{x}^M) = \prod_{m=1}^{M} g_j(\mathbf{x}_m). \quad (11)$$

Экспериментальные оценки характеристик качества схем комплексирования (10) и (11) получены на множествах изображений лиц [5] и подписей [6] в пространстве их древовидных представлений с заданным расстоянием в квадратичной метрике. Источники содержали по 1000 объектов от $c = 25$ персон, по 40 объектов в равновероятных классах. В экспериментах использовались экспоненциальные разделяющие функции

$$\left\{g_j^{(k)}(\mathbf{x}_m) = \exp\left(-v_{jm}^{(k)} d^2(\mathbf{x}_m, \mathbf{x}_{jm}^{(k)})\right)\right\}_{j=1}^{c}, k = 1,2, \quad (12)$$

где $v_{jm}^{(k)} > 0$ – свободный параметр, $\mathbf{x}_{jm}^{(k)}$ – представитель класса, $d(\mathbf{x}_m, \mathbf{x}_{jm}^{(k)}) \ge 0$ – расстояние между $\mathbf{x}_m$ и $\mathbf{x}_{jm}^{(k)}$. Номера $m = 1$ и $m = 2$ соответствуют источникам лиц и подписей. В качестве представителей $\mathbf{x}_{jm}^{(k)}$ выбраны ближайшие $(k = 1)$ и вторые ближайшие $(k = 2)$ объекты к представителям классов, относительно которых сумма квадратов внутриклассовых расстояний минимальна. Параметры $v_{jm}^{(k)}$ вычислялись с использованием средних значений и дисперсий внутриклассовых расстояний [2]. В Таблице 1 даны оценки характеристик эффективности классификации, полученные для источников и ансамбля в режиме скользящего контроля, при пороге $\Delta = 0,05$. Показано уменьшение E_{G} и r_{G} с ростом H_{G} для источников за счет комплексирования наборов $\mathrm{G}^{(1)}$ и $\mathrm{G}^{(2)}$ и дополнительное уменьшение E_{G} и r_{G} для ансамбля.

Таблица 1. Оценки эффективности классификации

	$\mathrm{G}^{(1)}$	$\mathrm{G}^{(2)}$	G
Лица $\mathbf{X}_1$			
H_{G}	3,159	3,139	3,162
E_{G}	0,168	0,188	0,155
r_{G}	0,076	0,096	0,063
Подписи $\mathbf{X}_2$			
H_{G}	3,217	3,216	3,219
E_{G}	0,139	0,185	0,106
r_{G}	0,087	0,133	0,054
Ансамбль $\mathbf{X}_1\mathbf{X}_2$			
H_{G}	3,215	3,211	3,217
E_{G}	0,021	0,030	0,018
r_{G}	0,010	0,018	0,008

4. Заключение

Для повышения точности классификации данных предложены способы комплексирования разделяющих функций на множестве объектов одной модальности и на ансамбле множеств различной модальности. Продемонстрировано уменьшение вероятности ошибки принимаемых решений и ее избыточности относительно нижней границы с ростом количества обрабатываемой информации. Предложенный подход позволяет повысить точность классификации за счет увеличения числа наборов разделяющих функций по отдельным источникам и числа источников в ансамбле.

Литература

[1] Kuncheva, L. I. Limits on the majority vote accuracy in classifier fusion / L. I. Kuncheva, C. J. Whitaker, C. A. Shipp , R.P.W. Duin // Pattern Analysis and Applications. – 2003.– Vol. 6. – P. 22–31. DOI: 10.1007/s10044-002-0173-7

[2] 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 of ITNT-2021. – 2021. – P. 1–6. DOI:10.1109/ITNT52450.2021.9649088

[3] Gallager, R.G. Information Theory and Reliable Communication / R.G. Gallager, – New York: Wiley & Sons, 1968. – 608 p.

[4] Duda, R.O. Pattern Classification, 2nd ed. / R.O Duda, P.E. Hart, D.G. Stork – New York: Wiley & Sons, 2001.– 688 p.

[5] Distance matrices for face dataset [Electronic resource]. — Access mode: http://sourceforge.net/projects/distance-matrices-face (01.03.2023).

[6] Distance matrices for signature dataset [Electronic resource]. — Access mode: http://sourceforge.net/projects/distance-matrices-signature (01.03.2023).