Об эффективности классификации данных в терминах взаимной информации и вероятности ошибки

М.М. Ланге[1], А.М. Ланге[1]

[1]Федеральный исследовательский центр "Информатика и управление" РАН, Вавилова 40, Москва, Россия,119333

Аннотация. Исследуется модель классификации данных на основе зависимости средней взаимной информации, содержащейся в множестве классифицируемых объектов относительно множества возможных решений по этим объектам, от вероятности ошибки классификации. Оптимизация модели заключается в нахождении обменного соотношения между наименьшей средней взаимной информацией и вероятностью ошибки, которое аналогично известной функции "скорость-погрешность" (Rate Distortion Function) для модели кодирования сообщений с допустимой погрешностью, переданных по каналу с искажениями. Строится нижняя граница введенного соотношения, которая дает нижнюю оценку вероятности ошибки классификации на заданном множестве объектов при любом фиксированном значении средней взаимной информации. Предлагаемая граница может быть использована для оценивания избыточности вероятности ошибки решающих алгоритмов, реализуемых на различных наборах разделяющих функций.

1. Введение

В теории кодирования источников с допустимой погрешностью Шенноном введена функция скорость-погрешность [1], которая при заданной погрешности дает нижнюю границу скорости кодирования, либо при заданной скорости определяет нижнюю границу погрешности для всевозможных способов (алгоритмов) кодирования. Функция скорость-погрешность определяется параметрами источника и метрикой погрешности и не зависит от выбранного алгоритма кодирования. Поэтому качество любого алгоритма кодирования может быть оценено избыточностью скорости кода относительно нижней границы при заданной погрешности, либо избыточностью погрешности кода относительно нижней границы при заданной скорости.

Следуя результатам теории кодирования источников, для модели классификации данных целесообразно найти аналогичную функцию в виде зависимости наименьшей средней взаимной информации между множеством классифицируемых объектов и множеством решений о классах этих объектов от заданной допустимой вероятности ошибки. Такая функция дает нижнюю границу вероятности ошибки на заданном множестве объектов при фиксированной средней взаимной информации и, следовательно, позволяет оценить избыточность вероятности ошибки решающего алгоритма, реализуемого на заданном наборе разделяющих функций.

Известны работы, в которых приводятся теоретические оценки точности для классификаторов с различными решающими правилами [2,3]. Однако эти работы не содержат общего подхода к построению границы потенциально достижимой точности, которая не зависит от решающих правил конкретных алгоритмов. Для получения нижней границы

вероятности ошибки классификации на заданном множестве данных предлагается использовать теоретико-информационную модель на основе известной схемы кодирования источника с допустимой погрешностью при наличии канала наблюдения с шумом [4]. В предлагаемой модели метки классов и классифицируемые объекты рассматриваются как входные и выходные данные канала наблюдения, вероятностные характеристики которого определяются условными по классам вероятностями или их аппроксимациями, построенными с использованием метрики на множестве объектов. Средняя погрешность между исходными метками классов и решениями измеряется в метрике Хемминга и эквивалента вероятности ошибки. Для такой модели вводится функция скорость-погрешность в форме зависимости наименьшей средней взаимной информации в классифицируемых объектах относительно принимаемых решений от вероятности ошибки и строится нижняя граница этой функции. Предлагаемая граница базируется на результатах работы [5] и является обобщением нижней границы Шеннона для схемы кодирования дискретных сообщений с допустимой погрешностью при отсутствии шума в канале наблюдения. Построенная граница достигается на байесовском решающем алгоритме с разделяющими функциями вида апостериорных вероятностей классов [6]. Приводятся численные реализации полученного соотношения между средней взаимной информацией и вероятностью ошибки классификации для множеств изображений подписей и лиц.

2. Теоретико-информационная модель классификации и задача исследования

Пусть $\Omega = \{\omega_1, ..., \omega_c\}, c \geq 2$ – множество классов с априорными вероятностями $P(\omega_i), i = 1, ..., c$ и $\mathbf{X}$ – множество объектов с условными по классам вероятностями $P(\mathbf{x} \mid \omega_i), \forall \mathbf{x} \in \mathbf{X}, \ i = 1, ..., c$. Будем считать, что множества Ω и $\mathbf{X}$ являются входными и выходными данными отображения $\Omega \to \mathbf{X}$. Множество объектов $\mathbf{X}$ и множество решений $\hat{\Omega} = \{\omega_j = 1, ..., c\}$ о классах этих объектов являются входом и выходом отображения $\mathbf{X} \to \hat{\Omega}$. Пусть $\mathrm{X}^N = (\mathbf{x}_1, ..., \mathbf{x}_N)$ – блок объектов $\mathbf{x}_n \in \mathbf{X}, \ n = 1, ..., N$, и $\mathbf{X}^N$ – множество всевозможных блоков длины N. Условные по классам распределения вероятностей блоков $\mathrm{X}^N \in \mathbf{X}^N$ являются характеристиками канала наблюдения $\Omega \to \mathbf{X}^N$ и образуют множество распределений

$$\mathrm{P} = \left\{ P(\mathrm{X}^N \mid \omega_i) = \prod_{n=1}^{N} P(\mathbf{x}_n \mid \omega_i) : \sum_{\mathrm{X}^N \in \mathbf{X}^N} P(\mathrm{X}^N \mid \omega_i) = 1, \ \ i = 1, ..., c \right\},$$

а условные распределения решений о классах $\omega_j \in \hat{\Omega}$ предъявляемых блоков являются характеристиками тест-канала $\mathbf{X}^N \to \hat{\Omega}$ и образуют множество распределений

$$\mathrm{Q} = \left\{ Q(\omega_j \mid \mathrm{X}^N) : \sum_{j=1}^{c} Q(\omega_j \mid \mathrm{X}^N) = 1, \forall \mathrm{X}^N \in \mathbf{X}^N \right\}.$$

В принятых обозначениях, множества Ω и $\hat{\Omega}$ состоят их одних и тех же элементов, однако вероятности элементов множества $\hat{\Omega}$ отличаются от априорных вероятностей соответствующих элементов множества Ω. Тогда множества $\Omega, \mathbf{X}^N, \hat{\Omega}$ совместно с распределениями P и Q порождают следующую схему классификации

$$\Omega \to \boxed{P(\mathrm{X}^N \mid \omega_i), i = 1, ..., c} \to \mathbf{X}^N \to \boxed{Q(\omega_j \mid \mathrm{X}^N), \ j = 1, ..., c} \to \hat{\Omega} \qquad (1)$$

В схеме (1) распределения из множества P считаются известными, а распределения из множества Q подлежат оптимизации.

Для оптимизации Q введем функционалы средней взаимной информации $I_{\mathrm{Q}}(\mathbf{X}^N; \hat{\Omega}) \geq 0$ и вероятности ошибки $E_{\mathrm{Q}}(\mathbf{X}^N, \hat{\Omega}) \geq 0$, зависящие от Q. Согласно [1], средняя взаимная информация имеет вид

$$I_{\mathrm{Q}}(\mathbf{X}^N;\hat{\Omega}) = \sum_{\mathrm{X}^N\in\mathbf{X}^N} P(\mathrm{X}^N)\sum_{j=1}^{c} Q(\omega_j\,|\,\mathrm{X}^N)\ln\left(Q(\omega_j\,|\,\mathrm{X}^N)/Q(\omega_j)\right)$$
$$= -\sum_{j=1}^{c} Q(\omega_j)\ln\,Q(\omega_j) + \sum_{\mathrm{X}^N\in\mathbf{X}^N} P(\mathrm{X}^N)\sum_{j=1}^{c} Q(\omega_j\,|\,\mathrm{X}^N)\ln Q(\omega_j\,|\,\mathrm{X}^N) \qquad (2)$$
$$= H(\hat{\Omega}) - H(\hat{\Omega}\,|\,\mathbf{X}^N),$$

где $P(\mathrm{X}^N)=\sum_{i=1}^{c} P(\omega_i)P(\mathrm{X}^N\,|\,\omega_i)$ и $Q(\omega_j)=\sum_{\mathrm{X}^N\in\mathbf{X}^N} P(\mathrm{X}^N)Q(\omega_j\,|\,\mathrm{X}^N)$

безусловные вероятности блоков объектов $\mathrm{X}^N\in\mathbf{X}^N$ и решений $\omega_j\in\hat{\Omega}$ по блокам объектов, а

$H(\hat{\Omega})=-\sum_{j=1}^{c} Q(\omega_j)\ln Q(\omega_j)$ и $H(\hat{\Omega}\,|\,\mathbf{X}^N)=-\sum_{\mathrm{X}^N\in\mathbf{X}^N} P(\mathrm{X}^N)\sum_{j=1}^{c} Q(\omega_i\,|\,\mathrm{X}^N)\ln Q(\omega_j\,|\,\mathrm{X}^N)$

соответственно энтропия и условная энтропия на множестве решений $\hat{\Omega}$, причем $H(\hat{\Omega})\ge H(\hat{\Omega}\,|\,\mathbf{X}^N)$. Вероятность ошибки определяется средним значением индикатора $[\omega_i\ne\omega_j]$ в метрике Хемминга:

$$E_{\mathrm{Q}}(\mathbf{X}^N,\hat{\Omega}) = \sum_{i=1}^{c} P(\omega_i)\sum_{\mathrm{X}^N\in\mathbf{X}^N} P(\mathrm{X}^N)\sum_{j=1}^{c} Q(\omega_j\,|\,\mathrm{X}^N)[\omega_i\ne\omega_j]$$
$$= \sum_{\mathrm{X}^N\in\mathbf{X}^N} P(\mathrm{X}^N)\sum_{j=1}^{c} Q(\omega_j\,|\,\mathrm{X}^N)\sum_{i=1}^{c} P(\omega_i\,|\,\mathrm{X}^N)[\omega_i\ne\omega_j], \qquad (3)$$

где $\sum_{i=1}^{c} P(\omega_i\,|\,\mathrm{X}^N)[\omega_i\ne\omega_j]=P(\omega_j\,|\,\mathrm{X}^N)$ – вероятность ошибки по блоку X^N на решении $\omega_j\in\hat{\Omega}$.

Функционалы (2) и (3) позволяют ввести функцию

$$R(\varepsilon)=\min_{N}\ \min_{\mathrm{Q}:\ E_{\mathrm{Q}}(\mathbf{X}^N,\hat{\Omega})\le\varepsilon} I_{\mathrm{Q}}(\mathbf{X}^N;\hat{\Omega}), \qquad (4)$$

где внутренний минимум берется по распределениям Q при $\varepsilon>0$ и $N\ge 1$.

Функция вида (4) определена для множества объектов $\mathbf{X}$ с условными по классам вероятностями $P(\mathbf{x}\,|\,\omega_i),\forall\mathbf{x}\in\mathbf{X};i=1,...,c$, которые либо известны, либо вычисляются по заданной метрике $d(\mathbf{x},\hat{\mathbf{x}})\ge 0$ для любой пары объектов $\mathbf{x}\in\mathbf{X},\hat{\mathbf{x}}\in\mathbf{X}$. В последнем случае используются условные по классам вероятности

$$P(\mathbf{x}\,|\,\omega_i)=\frac{e^{-vd(\mathbf{x},\mathbf{x}_i)}}{\sum_{\mathbf{x}\in\mathbf{X}} e^{-vd(\mathbf{x},\mathbf{x}_i)}},\ i=1,...,c \qquad (5)$$

где $\mathbf{x}_i\in\mathbf{X}_i, i=1,...,c$ – "центральные" представители кластеров $\mathbf{X}_i:\bigcup_{i=1}^{c}\mathbf{X}_i=\mathbf{X}$, а $v>0$ – параметр. В обоих случаях задача состоит в нахождении нижней границы функции $R(\varepsilon)$ для заданного множества объектов $\mathbf{X}$.

3. Нижняя граница функции $R(\varepsilon)$

Используя технику, предложенную в работе [5], вычисление функции (4) эквивалентно нахождению минимума

$$R(\varepsilon)=\min_{\mathrm{Q}_s:\ E_{\mathrm{Q}_s}(\Omega^*,\hat{\Omega})\le\varepsilon-\varepsilon_{\min}} I_{\mathrm{Q}_s}(\Omega^*;\hat{\Omega}), \qquad (6)$$

где $\Omega^*=\{\omega_k,k=1,...,c\}$ – множество классов, получаемых на отображении $\mathbf{X}^N\to\Omega^*$ с наименьшей вероятностью ошибки $\varepsilon_{\min}$, которая реализуется на байесовском решающем правиле для блоков $\mathrm{X}^N\in\mathbf{X}^N$ длины $N=N^*$. Здесь

$$I_{\mathrm{Q}_s}(\Omega^*;\hat{\Omega}) = \sum_{k=1}^{c} P(\omega_k) \sum_{j=1}^{c} Q_s(\omega_j \mid \omega_k) \ln Q_s(\omega_j \mid \omega_k) / Q(\omega_j) \ ,$$

$$E_{\mathrm{Q}_s}(\Omega^*,\hat{\Omega}) = \sum_{k=1}^{c} P(\omega_k) \sum_{j=1}^{c} Q_s(\omega_j \mid \omega_k)[\omega_k \neq \omega_j]$$

средняя взаимная информация и вероятность ошибки между входом и выходом тест-канала $\Omega^* \to \hat{\Omega}$, характеристики которого задаются условными распределениями

$$\mathrm{Q}_s = \left\{ Q_s(\omega_j \mid \omega_k) = \frac{\exp\left(-s[\omega_k \neq \omega_j]\right)}{\sum_{i=1}^{c} \exp\left(-s[\omega_k \neq \omega_i]\right)} : \sum_{j=1}^{c} Q_s(\omega_j \mid \omega_k) = 1,\ k = 1,...,c \right\} \tag{7}$$

с параметром $s > 0$. Соотношения (6) и (7) сводят вычисление нижней границы функции $R(\varepsilon)$ к известной нижней границе Шеннона [1] в схеме кодирования $\Omega^* \to \hat{\Omega}$ при условии, что средняя погрешность по метрике Хемминга $[\omega_k \neq \omega_j]$ не превосходит величины $\varepsilon - \varepsilon_{\min} \geq 0$. Полученная граница сформулирована в следующей теореме.

Теорема. Нижняя граница функции $R(\varepsilon)$ имеет вид

$$R_L(\varepsilon) = I(\mathbf{X};\Omega) - h(\varepsilon - \varepsilon_{\min}) - (\varepsilon - \varepsilon_{\min}) \ln(c-1),\ \varepsilon_{\min} \leq \varepsilon \leq \varepsilon_{\max} \ ,$$

где $h(z) = -z \ln z - (1-z) \ln(1-z)$, $R_L(\varepsilon_{\min}) = I(\mathbf{X};\Omega)$, $R_L(\varepsilon_{\max}) = 0$ и $I(\mathbf{X};\Omega)$ – средняя взаимная информация между множествами $\mathbf{X}$ и Ω.

Справедливо соотношение $I(\mathbf{X};\Omega) = H(\Omega) - H(\Omega \mid \mathbf{X})$, где $H(\Omega) = -\sum_{i=1}^{c} P(\omega_i) \ln P(\omega_i)$ – энтропия множества Ω и $H(\Omega \mid \mathbf{X}) = -\sum_{\mathbf{x} \in \mathbf{X}} P(\mathbf{x}) \sum_{i=1}^{c} P(\omega_i \mid \mathbf{X}) \ln P(\omega_i \mid \mathbf{X})$ – условная энтропия множества Ω при заданном множестве $\mathbf{X}$, причем $H(\Omega \mid \mathbf{X}) \leq H(\Omega)$. Необходимо отметить, что граница $R_L(\varepsilon)$ является обобщением нижней границы Шеннона в схеме кодирования независимых дискретных сообщений с допустимой погрешностью, измеряемой в метрике Хемминга [1]. В схеме Шеннона условные вероятности $P(\mathbf{x} \mid \omega_i) = [\mathbf{x} = \mathbf{x}_i], i = 1,...,c$ порождают апостериорные вероятности $P(\omega_i \mid \mathbf{X})$, которые принимают значения 1 и 0, что обеспечивает $H(\Omega \mid \mathbf{X}) = 0$ и $I(\mathbf{X};\Omega) = H(\Omega)$. В этом случае $\varepsilon_{\min} = 0$ и $R_L(\varepsilon_{\min}) = H(\Omega)$. В общем случае $\varepsilon_{\min} \geq 0$, а наибольшая вероятность ошибки равна $\varepsilon_{\max} = (c-1) \min\limits_{i=1}^{c} P(\omega_i)$ и для равновероятных классов $\varepsilon_{\max} = (c-1)/c$.

Наименьшая вероятность ошибки $\varepsilon_{\min} \geq 0$ определяется условной энтропией $H(\Omega \mid \mathbf{X}) \geq 0$. Для малых значений $\varepsilon_{\min}$ получена асимптотическое значение

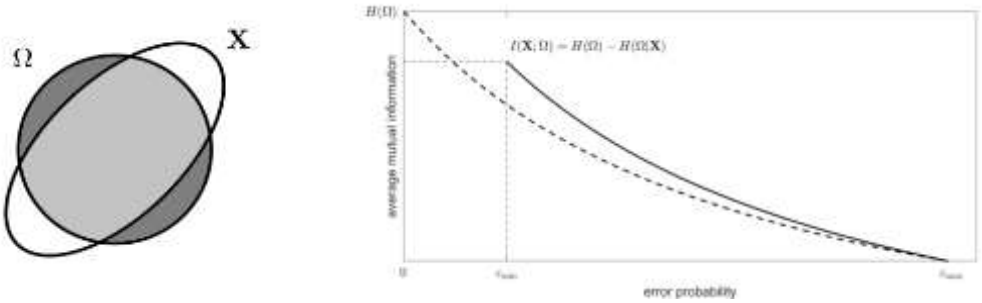
$$\begin{aligned} \varepsilon_{\min} = {} & \left(\varepsilon_{\max}(1-\varepsilon_{\max})\right) \sqrt{\ln^2\left((c-1)(1-\varepsilon_{\max})/\varepsilon_{\max}\right) + 2H(\Omega \mid \mathbf{X}) \left(\varepsilon_{\max}(1-\varepsilon_{\max})\right)^{-1}} \\ & - \left(\varepsilon_{\max}(1-\varepsilon_{\max})\right) \ln\left((c-1)(1-\varepsilon_{\max})/\varepsilon_{\max}\right), \end{aligned} \tag{8}$$

которое в случае равновероятных классов преобразуется к виду

$$\varepsilon_{\min} = \left((c-1)/c\right) \sqrt{2H(\Omega \mid \mathbf{X})/(c-1)} \ . \tag{9}$$

Оценки (8) и (9) демонстрируют уменьшение вероятности ошибки $\varepsilon_{\min}$ с уменьшением условной энтропии $H(\Omega \mid \mathbf{X})$ и, следовательно, с увеличением средней взаимной информации $I(\mathbf{X};\Omega)$ при фиксированной энтропии $H(\Omega)$. При этом $H(\Omega \mid \mathbf{X}) = 0$ обеспечивает $\varepsilon_{\min} = 0$.

Рисунок 1 (а) иллюстрирует среднюю взаимную информацию $I(\mathbf{X};\Omega)$ как теоретико-информационную меру пересечения $\Omega \cap \mathbf{X}$ (серая заливка) и условную энтропию $H(\Omega \mid \mathbf{X})$ как теоретико-информационную меру разности $\Omega \setminus \mathbf{X}$ (черная заливка). На рисунке 1 (б) показан характер границы $R_L(\varepsilon)$ (сплошная кривая) и нижней границы Шеннона (пунктирная кривая).



а) б)

Рисунок 1. Интерпретация $I(\mathbf{X};\Omega)$ и $H(\Omega\,|\,\mathbf{X})$ а) и характер границы $R_L(\varepsilon)$ б).

4. Условные по классам распределения на множестве объектов с заданной метрикой

В модели с заданной метрикой $d(\mathbf{x},\hat{\mathbf{x}}) \geq 0$ условные по классам распределения определяются условными вероятностями вида (5) с "центрами" $\mathbf{x}_i \in \mathbf{X}, i = 1,...,c$ и параметром $v > 0$. В качестве центральных представителей кластеров $\mathbf{X}_i, i = 1,...,c$ выбираются объекты

$$\mathbf{x}_i = \arg\min_{\hat{\mathbf{x}} \in \mathbf{X}_i} \sum_{\mathbf{x} \in \mathbf{X}_i} d(\mathbf{x},\hat{\mathbf{x}}),\ i = 1,...,c\ , \tag{10}$$

которые обеспечивают наименьшие внутриклассовые рассеяния.

Полагая, что значения $d(\mathbf{x},\hat{\mathbf{x}})$ вычисляются в метрике L1, для нахождения параметра v воспользуемся экспоненциальными плотностями распределения $p(\theta_i) = ve^{-v\theta_i}$, $i = 1,...,c$ случайных величин $\theta_i = d(\mathbf{x},\mathbf{x}_i)$. Используя указанные плотности и требуя максимума вероятности

$$\Pr\left[|\,\theta_i - \mu\,| \leq \alpha\mu\right] = \int_{\mu(1-\alpha)}^{\mu(1+\alpha)} p(\theta_i)d\theta_i = e^{-v\mu(1-\alpha)} - e^{-v\mu(1+\alpha)} \to \max_v$$

значений θ_i на отрезке $\left[\mu(1-\alpha), \mu(1+\alpha)\right]$ с параметрами $\mu > 0$ и $0 < \alpha < 1$, получим

$$v = \frac{1}{2\alpha\mu}\ln\frac{1+\alpha}{1-\alpha} \leq \frac{1}{\mu(1-\alpha)}\ . \tag{11}$$

Параметры μ и α определяются статистиками расстояний $d(\mathbf{x},\mathbf{x}_i)$ в кластерах $\mathbf{X}_i, i = 1,...,c$ и должны давать наибольшее значение оценки (11), которая обеспечивает наибольшую среднюю взаимную информацию $I(\mathbf{X},\Omega)$ и, соответственно, наименьшую вероятность ошибки $\varepsilon_{\min}$. В качестве статистик выбраны средние значения μ_i и среднеквадратические отклонения σ_i, $i = 1,...,c$ внутриклассовых расстояний, которые дают $\mu = \min_{i=1}^{c} \mu_i$ и $\alpha = \max_{i=1}^{c} \max\left\{\frac{\sigma_i}{\mu_i + \sigma_i}, \frac{\mu_i}{\mu_i + \sigma_i}\right\}$.

5. Численные реализации функций $R_L(\varepsilon)$ для множеств изображений подписей и лиц

В экспериментах использованы множества подписей и лиц, заданные полутоновыми изображениями размера 256x256 с 8-ми битовым кодированием элементов. Каждое изображение содержит один информативный объект (лицо или подпись). Множества лиц и подписей содержат по 1000 объектов от 25 персон ($c = 25$), по 40 реализаций от каждой персоны. Априорное распределение классов считается равномерным. Информативные объекты заданы древовидными представлениями в виде структурированных наборов эллиптических примитивов [7]. Различие любой пары объектов определяется введенной в [8] метрикой расстояний на множестве древовидных представлений. Эксперимент включает вычисление

функций $R_L(\varepsilon)$ для множеств подписей и лиц с использованием матриц расстояний, заданных ресурсами [9] и [10].

Примеры древовидных представлений лица и подписи даны на рисунке 2. Представления имеют информативные уровни $l = 1,...,8$, которые содержат по 2^l эллиптических примитивов. Параметры примитива нулевого уровня используются для нормировки параметров примитивов последующих уровней. Нормированные примитивы всех уровней задаются в собственных координатных осях примитива нулевого уровня и имеют номера соответствующих им вершин бинарного дерева. Построение примитивов в собственных осях нулевого уровня и нормировка параметров примитивов обеспечивают инвариантность представлений к сдвигу, повороту, масштабу и уровню яркости объектов. Метрика на множестве объектов, представленных деревьями эллиптических примитивов, определяется взвешенной суммой различий нормированных параметров примитивов, имеющих одинаковые номера в сравниваемых деревьях. Выбранная метрика совместно с представителями классов (10) и оценками параметров (11) полностью определяют условные по классам распределения вероятностей объектов. Совместное распределение классов и объектов позволяют вычислить среднюю взаимную информацию $I(\mathbf{X};\Omega)$ и наименьшую вероятность ошибки $\varepsilon_{\min}$, которые являются основными характеристиками нижней границы $R_L(\varepsilon)$, сформулированной в теореме.

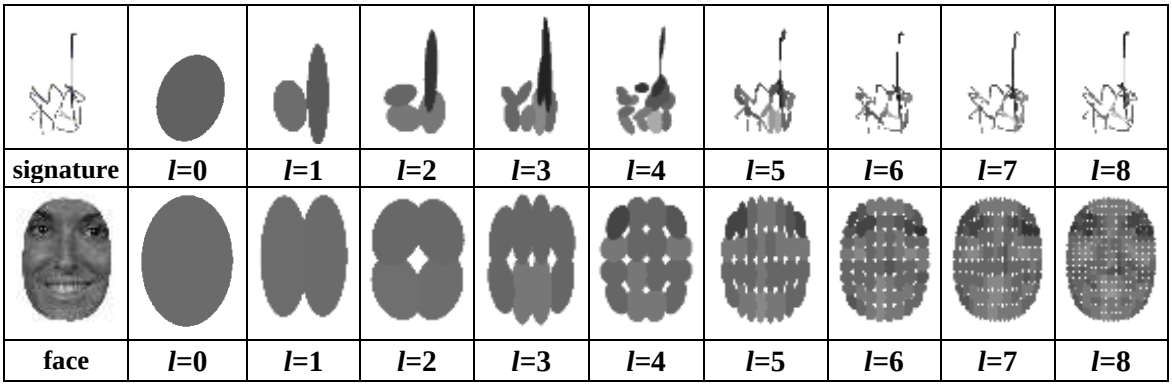


Рисунок 2. Примеры древовидных представлений подписи и лица.

На рисунке 3 приведены численные реализации нижних границ $R_L(\varepsilon)$, полученных на множествах изображений подписей и лиц в пространстве их древовидных представлений с заданной метрикой. Для сравнения дана кривая нижней границы Шеннона, которая обеспечивает $\varepsilon_{\min} = 0$. При $N \geq 1$ и значениях средней взаимной информации $I(\mathbf{X}^N;\hat{\Omega}) \leq I(\mathbf{X}^N;\Omega)$ вычисленные границы демонстрируют нижние оценки вероятности ошибки алгоритмов, принимающих решения по N объектам из множества $\mathbf{X}$. Необходимо отметь, что поскольку ансамбль $\mathbf{X}^N = \mathbf{X}_1...\mathbf{X}_N$, представляющий всевозможные блоки из N объектов, образован тождественными множествами $\mathbf{X}_n = \mathbf{X}, n = 1,..., N$, справедливо равенство $I(\mathbf{X}^N;\Omega) = I(\mathbf{X};\Omega)$.

6. Избыточность вероятности ошибки решающего алгоритма

Построенные границы позволяют оценить избыточность вероятности ошибки алгоритма, который принимает решения по блокам из $N \geq 1$ объектов, используя разделяющие функции

$$\mathrm{G}_s = \left\{ g_j^s(\mathrm{X}^N), \mathrm{X}^N \in \mathbf{X}^N;\ j = 1,...,c \right\}$$

с параметром $0 \leq s \leq 1$. Значения разделяющих функций определяют сходство блока X^N с классами $\omega_j \in \hat{\Omega},\ j = 1,...,c$ и дают метку класса $j^* = \arg\max_{j=1}^{c} g_j^s(\mathrm{X}^N)$ для X^N.

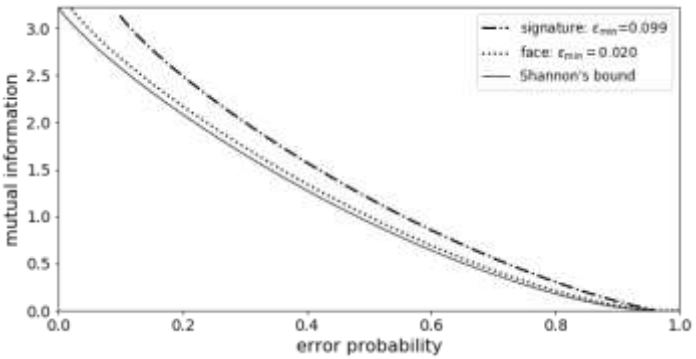


Рисунок 3. Графики нижних границ $R_L(\varepsilon)$ для множеств подписей и лиц.

Разделяющие функции G_s порождают множество распределений

$$\mathrm{Q}_s = \left\{ Q_s(\omega_j \mid \mathrm{X}^N) = \frac{g_j^s(\mathrm{X}^N)}{\sum_{i=1}^{c} g_i^s(\mathrm{X}^N)}, j = 1,...,c; \forall \mathrm{X}^N \in \mathbf{X}^N \right\},$$

которые равномерны, когда $s = 0$, и проявляют растущую неравномерность, когда $s \to 1$. На разделяющих функциях G_s реализуется средняя взаимная информация

$$I_{\mathrm{Q}_s}(\mathbf{X}^N;\hat{\Omega}) = \sum_{\mathrm{X}^N \in \mathbf{X}^N} P(\mathrm{X}^N) \sum_{j=1}^{c} Q_s(\omega_j \mid \mathrm{X}^N) \ln\left(Q_s(\omega_j \mid \mathrm{X}^N) / Q(\omega_j) \right)$$

и вероятность ошибки

$$E_{\mathrm{Q}_s}(\mathbf{X}^N,\hat{\Omega}) = 1 - \sum_{\mathrm{X}^N \in \mathbf{X}^N} P(\mathrm{X}^N) \max_{j=1}^{c} Q_s(\omega_j \mid \mathrm{X}^N) \ .$$

Наибольшая вероятность ошибки $E_{\mathrm{Q}_{s=0}}(\mathbf{X}^N,\hat{\Omega}) = (c-1)/c$ реализуется в точке $s = 0$, которой соответствует $I_{\mathrm{Q}_{s=0}}(\mathbf{X}^N;\hat{\Omega}) = 0$, а наименьшая вероятность ошибки $E_{\mathrm{Q}_{s=1}}(\mathbf{X}^N,\hat{\Omega}) = \hat{\varepsilon}_{\min}$ достигается в точке $s = 1$, при наибольшем значении средней взаимной информации $I_{\mathrm{Q}_{s=1}}(\mathbf{X}^N;\hat{\Omega}) \le I(\mathbf{X};\Omega)$. Для заданного набора разделяющих функций G_s с параметром $0 \le s \le 1$, величины $\hat{R} = I_{\mathrm{Q}_s}(\mathbf{X}^N;\hat{\Omega})$ и $\hat{\varepsilon} = E_{\mathrm{Q}_s}(\mathbf{X}^N,\hat{\Omega})$ образуют обменное соотношение $\hat{R}(\hat{\varepsilon})$.

Функция $\hat{R}(\hat{\varepsilon})$ позволяет определить избыточность $\hat{\varepsilon} - \varepsilon$ вероятности ошибки решающего алгоритма относительно нижней границы при условии $\hat{R}(\hat{\varepsilon}) = R_L(\varepsilon)$. Следует отметить, что в случае принятия решений по $N \ge 1$ объектам, наименьшая избыточность реализуется на байесовском алгоритме, разделяющие функции которого дают в точке $s = 1$ апостериорное распределение $\mathrm{Q}_{s=1}$ [6]. При оптимальном значении N^* байесовский алгоритм достигает наименьшей вероятности ошибки $\varepsilon_{\min}$, которая соответствует наибольшей средней взаимной информации $I(\mathbf{X};\Omega)$ и, следовательно, обеспечивает нулевую избыточность.

7. Заключение

Исследована теоретико-информационная модель классификации данных, для которой введено обменное соотношение между средней взаимной информацией множества классифицируемых объектов относительно множества решений по классам и наименьшей вероятностью ошибки. Соотношение взаимная информация-вероятность ошибки является аналогом известной в теории информации функции скорость-погрешность (Rate Distortion Function) для модели кодирования источника с заданной точностью при наличии канала наблюдения с шумом. Построена нижняя граница функции взаимная информация-вероятность ошибки, которая является обобщением нижней границы Шеннона. Полученная нижняя граница зависит от

априорного распределения классов и условных по классам распределений на заданном множестве объектов и не зависит от решающего алгоритма. Независимость найденной границы от решающего алгоритма позволяет использовать ее для оценивания избыточности вероятности ошибки решающих алгоритмов по заданным наборам разделяющих функций. Планируется обобщение полученного обменного соотношения для ансамбля источников, которые с увеличением числа источников обеспечивают уменьшение потенциально возможной вероятности ошибки при фиксированной средней взаимной информации.

8. Благодарности
Работа выполнена при финансовой поддержке РФФИ, проекты: 18-07-01231 и 18-07-01385.

9. Литература
[1] Gallager, R.G. Information Theory and Reliable Communication – New York: Wiley & Sons, Inc., 1968. – 588 с.
[2] 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.
[3] Lam, L. Application of majority voting to pattern recognition: An analysis of its behavior and performance / L. Lam, C.Y. Suen // IEEE Transactions on Systems, Man, and Cybernetics. – 1997. – Vol. 27(5). – P. 553-568. DOI: 10.1109/3468.618255.
[4] Dobrushin, R.L. Information transmission with additional noise / R.L. Dobrushin, B.S. Tsybakov // IRE Trans. Information Theory. – 1962. – Vol. 8(5). – P. 293-304. DOI: 10.1109/TIT.1962.1057738.
[5] Ланге, М.М. О теоретико-информационной модели классификации данных / М.М. Ланге, А.М. Ланге // Машинное обучение и анализ данных. – 2018. – Т. 4, № 3. – С. 165-179. DOI: 10.21469/22233792.4.3.03.
[6] Duda, R.O. Pattern Classification / R.O. Duda, P.E. Hart, D.G. Stork – New York: Wiley & Sons, Inc., 2001.
[7] Lange, M.M. Recognition of objects given by collections of multichannel images / M.M. Lange, D.Y. Stepanov // Pattern Recognition and Image Analysis. – 2014. – Vol. 24(2). – P. 431-442. DOI: 10.1134/S1054661814030122.
[8] Lange, M.M. On fusion schemes for multiclass object classification with reject in a given ensemble of sources / S.N. Ganebnykh // Journal of Physics: Conference Series. – 2018. – Vol. 1096. – P. 012048. DOI: 10.1088/1742-6596/1096/1/012048.
[9] Матрицы расстояний изображений подписей [Электронный ресурс]. – Режим доступа: http://sourceforge.net/projects/distance-matrices-signature (23.12.2019).
[10] Матрицы расстояний изображений лиц [Электронный ресурс]. – Режим доступа: http://sourceforge.net/projects/distance-matrices-face (23.12.2019).

On data classification efficiency based on a trade-off relation between mutual information and error probability

M.M. Lange[1], A.M. Lange[1]

[1]Federal Research Center "Computer Science and Control" of RAS, Vavilov street 40, Moscow, Russia,119333

Abstract. A data classification model based on an average mutual information between a set of the objects and a set of the object class decisions depending on an error probability is developed. The model optimization consists in minimizing the average mutual information by the conditional distribution for the object class decisions subject to a given admissible error probability. It is equivalent to calculating the rate-distortion function in a scheme of coding the class labels with a given distortion when the set of the class labels and the set of the objects are the input and the output of an observation channel with the known class-conditional probability distributions. Given set of the objects and known observation channel, a lower bound to the rate-distortion function is calculated. This bound is independent on a decision algorithm and yields a potentially achievable error probability subject to a fixed value of the average mutual information. The obtained bound can be used for estimating an error probability redundancy of any decision algorithm by the given discriminant functions.