Ученые записки Таврического национального университета им. В. И. Вернадского

Серия "Математика. Механика. Информатика и Кибернетика"N 1 (2002) 199 – 210

УДК 519.8

С.И. Гуров

ОЦЕНКИ ОШИБОК АЛГОРИТМОВ РАСПОЗНАВАНИЯ

В работе предложены новые подходы к построению точечных оценок алгоритмов классификации, применимых к случаю малого числа прецедентов.

Введение

При разработке систем распознавания для заказчика важно не только получить алгоритм, реализующий требуемое разделение классов, но и знать, как часто данный алгоритм будет ошибаться при классификации вновь предъявляемых объектов. Ясно, что указанная оценка напрямую определяет качество решения поставленной задачи. На практике же дать такую обоснованную оценку часто оказывается затруднительным.

Несмотря на указанную важность, методы оценки надежности выбранного решающего правила развиты значительно слабее, чем теория построения распознающих алгоритмов. Проблема усугубляется ещё и тем, что при решении практических задач распознавания образов часто приходиться довольствоваться малым числом имеющихся в наличии прецедентов. В этом случае типичной является ситуация, когда либо параметры формул оценки ошибок распознавания находятся вне границ применимости метода, либо полученные оценки оказываются сильно заниженными или завышенными и интуитивно неприемлимыми для заказчика, как, например, нулевая точечная оценка ошибки при корректном алгоритме распознавания.

Вышесказанное свидетельствует о необходимости предложить новые подходы к построению оценок алгоритмов распознавания, способных охватить важный случай малого числа прецедентов. В настоящей работе рассмотрены только точечные оценки ошибок алгоритмов классификации. Подходы к получению интервальных оценок намечены в [5]; им будет посвящена отдельная работа.

1. Основные понятия и определения

Под *пространством образов* $\mathcal{X}$ будем понимать произвольный непустой компакт [6]. Элементы $\mathcal{X}$ называются *образами*. Множество $\mathcal{X}$ полагается разбитым на конечное число $s \geq 2$ попарно непересекающихся областей $\{\mathcal{X}_t\}$, $t = \overline{1,s}$, называемых *классами*. Существенным является то, что информация о разбиении $\mathcal{X}$ на классы ограничивается знанием о принадлежности к тому или иному классу конечного числа $x_1, x_2, \ldots, x_m$ элементов $\mathcal{X}$. Такие образы с известной классификацией называют *прецедентами*, а их совокупность — *обучающей выборкой* (или *последовательностью*) $\overline{x}_m$ (длины или объёма m). Обозначив через $\mathcal{Y}$ множество символов классов $\{K_1, \ldots, K_s\}$ можно сказать, что существует функция $f^* : \mathcal{X} \to \mathcal{Y}$, о которой известен лишь набор ее значений $\{f^*(x_i)\}_{i=1}^m = \overline{f}^*(\overline{x}_m)$ в точках $\overline{x}_m$. Функция f^* называется *истинным классификатором*.

[6] Обычно также считают, что $\mathcal{X}$ есть подмножество прямого произведения конечного числа n метрических пространств, соответствующих *признакам*, и называют его *признаковым пространством*. Однако это предположение, существенное при построении классификаторов, не будет использоваться нами при оценке надежности построенных решающих правил.

Классификатором или *решающим правилом* (р.п.) называется любая функция $f : \mathcal{X} \to \mathcal{Y}$ (рассматривается, следовательно, задача распознавания с непересекающимися классами в детерминированной постановке). Классификация образа x состоит в вычислении значения $f(x)$. Мы не будем различать функцию f и реализующий ее алгоритм.

При решении задач распознавания образов требуется построить оптимальный в некотором смысле классификатор $f(x)$, а именно такой, чтобы при предъявлении элементов x из $\mathcal{X}$ в процессе классификации на практике равенство

$$f(x) = f^*(x)$$

(правильная классификация), выполнялось "как можно чаще". Количественно оцененная степень уверенности ν в справедливости данного равенства для произвольного $x \in \mathcal{X}$ называется *надежностью классификации*. Задача оценки надежности р.п. и состоит в определении ν.

На практике часто встречается ситуация, когда для оценки надежности р.п. в распоряжении разработчика имеются лишь наборы значений на прецедентах истинного и построенного классификаторов и, возможно, некоторая дополнительная информация о «важности» самих прецедентов. Набор образов с известной классификацией, использующийся для оценки надежности выбранного р.п. называется *экзаменационной последовательностью (выборкой)*. Важность прецедентов, учитывающая их значимость с точки зрения потерь при ошибочной их классификации и/или отражающая частоту встречаемости аналогичных образов на практике описывается, как правило, в виде неотрицательных весов. Вектор весов $\{\gamma_i = \gamma(x_i)\}_{i=1}^m = \overline{\gamma}_m$ прецедентов $\overline{x}_m$ мы будем включать в понятие прецедентной информации вместе с самими прецедентами и указанными наборами значений классификаторов на них.

Часто заказчику необходимо иметь обоснованную оценку надежности полученного алгоритма классификации в условиях наличия лишь данной прецедентной информации и невозможности ни её пополнения, ни организации проверки в ходе практического проведения процесса классификации [7]. В этих случаях оценивать величину ν приходиться лишь по значениям функций $\{f^*(x_i)\}$, $f(x_i)\}$ и весов $\gamma(x_i)$ прецедентов $x_1, x_2, \ldots, x_m$. Ясно, что такая оценка будет адекватной в той или иной степени, если состав экзаменационной выборки будет отражать характер появления новых предъявляемых для классификации образов при практическом применении алгоритма классификации. Здесь имеется в виду, что образы из одних подобластей $\mathcal{X}$ могут встречаться чаще, чем из других, и состав набора прецедентов должен отражать этот факт.

Указанное предположение о свойствах обучающей и экзаменационной последовательностей назовем *гипотезой представительности* (ГП). Точнее, под ГП мы будем понимать принятие положения о том, что прецедентная информация отражает свойства пространства образов, связанные с определённым распределением появляющихся образов по различным подобластям $\mathcal{X}$ в процессе классификации на практике.

Гипотеза представительности, принятая в той или иной форме в рамках конкретной задачи, вместе с гипотезой компактности (ГК) [8] является определяющим фактором при оценке надежности построенного решающего правила, на котором основываются все дальнейшие выводы.

---

[7] Например, когда получение нового прецедента связано с проведением дорогостоющего исследования или невозможно принципиально (распознавание и прогнозирование экономических, социальных процессов, в медицине, политике, военном деле и т.д.).

[8] «Образам соответствуют компактные множества в пространстве выбранных свойств» [1].

Для практического использования данная весьма общая формулировка гипотезы представительности формализуется в точной математической форме. Такая формализация (одновременно с приведенным выше интуитивным критерием оптимальности классификатора) проводится в вероятностных терминах. Для этого предполагают, что $\mathcal{X}$ обладает вероятностной мерой $\mu(\cdot)$, т.е. для любого подмножества X пространства образов существует интеграл

$$\int_X \mu(dx) = P(X) \geq 0, \qquad P(\mathcal{X}) = 1.$$

$P(X)$ называется, как известно, функцией распределения вероятностей на $\mathcal{X}$. Вероятность события A будем обозначать $\mathcal{P}(A)$ или $\mathcal{P}\{A\}$. Для упрощения выкладок полагают и существование функции плотности вероятности $p(x)$ на $\mathcal{X}$: $p(x) = \mu(dx)/dx$. Далее принимают, что и обучающая выборка, и образы с неизвестной принадлежностью к подмножествам X_t, $t = \overline{1,s}$, которые будут в дальнейшем предъявляться для классификации, получены из пространства образов в результате подобных процедур выбора, что обеспечивает их аналогичные статистические свойства.

Таким образом, при отсутствии информации о весах прецедентов (или, что то же, при равенстве всех весов) гипотеза представительности принимается в следующей форме.

Гипотеза 1. На пространстве образов $\mathcal{X}$ задана (может быть неизвестная) функция распределения вероятностей $P(X)$, $X \subseteq \mathcal{X}$, и любой рассматриваемый набор образов $x_1, x_2, \ldots, x_l$ является, если явно не указано иначе, реализацией независимой выборки l случайных величин из генеральной совокупности с распределением $P(X)$.

Ясно, что Гипотеза 1 является условием репрезентативности выборки в математической статистике.

Если $P(x)$ известно, то оценка надежности построенного р.п. не представляет труда (см. ниже формулы (2) и (3)). Далее мы считаем функцию $P(x)$ неизвестной.

Степень удовлетворенности (точнее, неудовлетворенности) исследователя полученным классификатором $f(x)$ выражается значением функционала *среднего риска* $R(f)$:

$$R(f) = \int_{\mathcal{X}} \left(\sum_{f^*(x)\in\mathcal{Y}} \sum_{f(x)\in\mathcal{Y}} Q(f^*(x), f(x)) \right) p(x)\, dx, \tag{1}$$

где $Q : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ ($\mathbb{R}_{\geq 0}$ – множество неотрицательных действительных чисел).

Здесь $Q(K_i, K_j) = c_{ij} \geq 0$ — некоторая выбранная функция потерь или штрафа за отнесение объекта из класса K_i в класс K_j. Часто можно полагать, что

$$c_{ii} = 0,\ c_{ij} = 1,\ i \neq j,\ i,j = \overline{1,s}.$$

Тогда $R(f)$ есть вероятность ошибочной классификации при применении р.п. f.

Ясно, что прямое использование зависимости (1) для вычисления среднего риска невозможно в силу неизвестности $f^*(x)$ даже при известном распределении $p(x)$. Чтобы обойти данную трудность, при построении классификатора по прецедентам $\overline{x}_m$ используют функционал *эмпирического риска* $R^e_m(f)$:

$$R^e_m(f) = \frac{1}{m} \sum_{i=1}^{m} Q(f^*(x_i), f(x_i)). \tag{2}$$

Однако такая замена функционалов тут же порождает вопрос о связи минимальных значений эмпирического и среднего рисков. Ответ на этот вопрос дает теория VC равномерной сходимости частот к вероятностям в условиях конечности выборок, предложенная

В.Н. Вапником и А.Я. Червоненкисом [3], [4]. К сожалению оказывается, что в рамках VC гарантировать малость $R(f_{min})$ при малом $R_m^e(f_{min})$, где

$$f_{min} = \arg\min_f \{R_m^e(f)\}$$

можно лишь при достаточно больших объёмах m обучающей выборки $\overline{x}_m$.

Проблема оценки надежности р.п. была бы снята, если бы удалось определить или хотя бы оценить вероятности

$$p_{ij} = P(X_{ij}) = \int_{X_{ij}} p(x)\,dx, \quad i,j = \overline{1,s}\,, \tag{3}$$

где $X_{ij} = \{x \mid x \in \mathcal{X},\ f^*(x) = K_i,\ f(x) = K_j\}$. Подобласти $\{X_{ij}\}_{i,j=1}^{s,s}$ — это s^2 областей разбиения пространства объектов $\mathcal{X}$, соответствующих ситуациям, когда x принадлежит классу K_i, а решающее правило относит его к классу K_j. При $i \neq j$ p_{ij} суть вероятности ошибок классификации соответствующего рода.

Теперь можно явно вычислить средний риск

$$R(f) = \sum_{i=1}^{s}\sum_{j=1}^{s} c_{ij}\,p_{ij}\,. \tag{4}$$

В предположениях $c_{ii} = c_r$, $c_{ij} = c_w$, $(i \neq j)$ можно полагать $\mathcal{X}$ разбитым на две подобласти — правильных X_r и неправильных X_w классификаций и обозначить $\nu = P(X_r)$. Тогда

$$R(f) = c_r\nu + c_w(1-\nu)\,,$$

а при $c_r = 0$, $c_w = 1$ имеем $R(f) = 1 - \nu$.

Итак, надежность классификации р.п. определяется набором вероятностей $\{p_{ij}\}_{i,j=1}^{s,s}$ или величиной ν (вероятность правильной классификации).

Задача классификации $Z = Z(\mathcal{X},\ s,\ m,\ \overline{x}_m,\ \overline{\gamma}_m,\ \overline{f}^*(\overline{x}_m))$ состоит в выборе р.п. f, минимизирующего тот или иной функционал $R^0(\cdot)$ (обычно это средний риск) и оценки полученной величины $R^0(f)$. Указанные подзадачи будем обозначать Z1 и Z2. Когда позволяет имеющаяся информация (удается восстановить плотности соответствующих распределений), эти подзадачи решаются параллельно и согласовано. На практике же, в силу вышеупомянутых причин, обе подзадачи решают, как правило, приближенно и раздельно (хотя, возможно, и используют результаты Z2 для корректировки или выбора решающих правил Z1).

Заметим, что предложить для решения Z1 решающее правило, основанное на тех или иных идеях, вообще говоря, несложно. Также существует [6], [10] универсальный метод построения *корректных* (точных на прецедентах) алгоритмов классификации. В настоящей работе сначала рассматриваются методы решения подзадачи Z2 задачи Z при выбранном классификаторе f (т.е. подзадача Z1 считается уже решённой).

2. Постановка задачи

Пусть в результате решения подзадачи Z1 задачи распознавания

$$Z = Z(\mathcal{X},\ s,\ m,\ \overline{x}_m,\ \overline{\gamma}_m,\ \overline{f}^*(\overline{x}_m))$$

построено р.п. $f(x)$. Предположим пока, что $\gamma_1 = \gamma_2 = \cdots = \gamma_m$ и примем гипотезу представительности в форме «Гипотеза 1». Случай неравных весов прецедентов будет рассмотрен в п. 6.

Далее мы считаем, что пространство образов $\mathcal{X}$ разбито на $v \geq 2$ подобластей $\{X_k\}_{k=1}^{v}$ и обозначаем через m_k количество прецедентов, попавших в область X_k, $k = \overline{1,v}$, $\sum_{k=1}^{v} = m$.

В задачах классификации встречаются только следующие случаи значений v (напомним, что $s \geq 2$).

(1) $v = 2$. Здесь X_1 и X_2 суть области правильных и неправильных классификаций.

(2) $v = s^2$. Здесь $\{X_k\}_{k=1}^{v}$ суть переобозначенные области $\left\{X_{ij}\right\}_{i,j=1}^{s,s}$ пространства образов, т.е. $X_{ij} = \left\{x | x \in \mathcal{X},\, f^*(x) = K_i,\, f(x) = K_j\right\} = \{X_1, X_2, \ldots X_v\}$ (см. п. 1).

(3) $v = s^2 + 1$. Здесь к определённым выше областям добавляется область соответствующая случаю отказа от классификации.

(4) $v = s^2 + s$, когда мы хотим специфицировать класс прецедента, на котором произошёл отказ.

Обозначим $p_k = P(X_k) \geq 0$, $k = \overline{1, v}$. Мы будем определять оценки значений данных вероятностей. Ясно, что справедливо условие нормировки

$$\sum_{k=1}^{v} p_k = 1 \tag{5}$$

и при данном v мы имеем $(v-1)$-мерную задачу.

Поскольку случайная величина x распределена в соответствии с $P(\cdot)$, то p_k есть вероятность выполнения соотношения $x \in X_k$. Тогда вероятность $p(m_1, m_2, \ldots, m_v)$ того, что при независимой случайной выборке m элементов из $\mathcal{X}$ в соответствии с распределением $P(\cdot)$ соотношение $x \in X_k$ будет выполняться m_k раз, $k = \overline{1, v}$, $\sum_{i=k}^{v} m_k = m$ имеет $(v-1)$-мерное полиномиальное (мультиномиальное) распределение $M(m; p_1, p_2, \ldots, p_v)$, функция плотности вероятности которого дается формулой

$$p(m_1, \ldots, m_v) = \frac{m!}{m_1!\, m_2! \ldots m_v!}\, p_1^{m_1} p_2^{m_2} \ldots p_v^{m_v};\ p_k \in (0,1),\ k = \overline{1, v}. \tag{6}$$

Отметим, что первые моменты полиномиального распределения суть

$$\mu_k = mp_k,\ k = \overline{1, v-1}$$

а матрица ковариаций —

$$C = (\mu_{ij})_{i,j=1}^{v-1,\,v-1},\ \mu_{ii} = mp_i(1-p_i),\ \mu_{ij} = -mp_ip_j,\ i \neq j. \tag{7}$$

При $v = 2$, $p_1 = p$ имеем биномиальное распределение $Bi(m, p)$ с функцией плотности вероятности

$$p(m_1) = \binom{m}{m_1} p^{m_1}(1-p)^{m-m_1};\ p \in (0,1)$$

для которой

$$\mu = mp,\ \sigma^2 = mp(1-p).$$

Наша задача (параметрического статистического оценивания) состоит в том, чтобы построить точечные оценки неизвестных, но фиксированных величин $p_1, p_2, \ldots, p_v$ по случайным значениям $m_1, m_2, \ldots, m_v$, $\sum_{k=1}^{v} m_k = m$. Построенные оценки должны быть применимы для случая малого числа m прецедентов.

3. Частотный подход

В рамках частотного подхода используются следующие методы получения точечных оценок неизвестных параметров:

- метод максимального правдоподобия;
- метод моментов;
- метрические методы.

Метод максимального правдоподобия (ММП) основан на максимизации *функции правдоподобия* L аргументов $p_1, p_2, \ldots, p_v$. Функции правдоподобия для нашего случая определяется следующим образом. Результат определения количества прецедентов в областях $\{X_k\}_{k=1}^{v}$ представим в виде 0,1-таблицы $T = \{t_{k,i}\}_{k,i=1}^{v,\,m}$, где

$$t_{k,i} = \begin{cases} 1, & \text{если } i\text{-й прецедент принадлежит области } X_k, \\ 0, & \text{иначе.} \end{cases}$$

Ясно, что

$$\sum_{k=1}^{v} t_{k,i} = 1, \quad \sum_{i=1}^{m} t_{k,i} = m_k, \quad \sum_{k=1}^{v} m_k = m\,.$$

Тогда функция правдоподобия есть

$$L(T;\, p_1, p_2, \ldots, p_v) = const{\cdot}p_1^{t_{1,1}+\ldots+t_{i,m}} p_2^{t_{2,1}+\ldots+t_{2,m}} \ldots p_v^{t_{v,1}+\ldots+t_{v,m}} = const{\cdot}p_1^{m_1} p_2^{m_2} \ldots p_v^{m_v}.$$

Теперь, поскольку максимумы L и $\ln L$ совпадают, наша задача состоит в максимзации функции

$$\ln L(p_1, p_2, \ldots, p_v) = const + \sum_{k=1}^{v} m_k \ln p_k$$

при условии нормировки (5).

Данная задача на условный экстремум легко решается методом множителей Лагранжа. Составляя функцию Лагранжа

$$\mathcal{L}(p_1, p_2, \ldots, p_v, \lambda) = \ln L(p_1, p_2, \ldots, p_v) + \lambda \cdot \{1 - \sum_{k=1}^{v} p_k\}$$

и приравнивая $\partial \ln \mathcal{L}/\partial p_i$ и $\partial \ln \mathcal{L}/\partial \lambda$ нулю, получаем СЛАУ порядка $v+1$

$$\begin{cases} \frac{m_k}{p_k} - \lambda = 0\,, \quad k = \overline{1, v}\,, \\ \sum_{k=1}^{v} p_k = 1\,, \end{cases}$$

решения которой суть $\lambda = m$, $\widehat{p_k} = m_k/m$, $k = \overline{1, v}$.

Таким образом, МП-оценками $\widehat{p_k}$ вероятностей p_k будут относительные частоты m_k/m числа прецедентов m_k в областях X_k, $k = \overline{1, v}$.

Нетрудно видеть, что **метод моментов** даёт такие же оценки, поскольку моменты первого порядка μ_k полиномиального распределения равны mp_k, а соответствующие выборочные — m_k, $k = \overline{1, v}$.

Метрические методы основаны на рассмотрении различных мер расхождения меду наблюдёнными величинами $m_1, m_2, \ldots, m_v$ и их математическими ожиданиями $mp_1, mp_2, \ldots, mp_v$. Оценка $\widehat{p}_1, \widehat{p}_2, \ldots, \widehat{p}_v$ определяется как значения вероятностей, минимизурующих эту меру. Для оценивания используются такие меры, как «χ^2», «модифицированный χ^2», «расстояние Хеллингера», «дивергенция Калбэка-Лейбера», «мера расхождения Холдейна» и др. [9]. Изучение их показывает, что к нашей задаче оказывается применим (по крайней мере в своём исходном виде) лишь метод «модифицированный χ^2», который даёт всё ту же оценку в виде относительных частот.

Легко показывается, что математическое ожидание $\mathbf{M}\{\widehat{\overline{p}}\}$ вектора оценок $\{p_k\}_{i=k}^{v}$ есть (с учетом (7) и обозначений $\overline{m} = (m_1, m_2 \ldots m_v)^T$ и $\overline{p}^*$ — v-ичный вектор истинных значений вероятностей)

$$\mathbf{M}\{\widehat{\overline{p}}\} = \mathbf{M}\{\overline{m}/m\} = \frac{1}{m}\mathbf{M}\{\overline{m}\} = \frac{m\,\overline{p}^*}{m} = \overline{p}^*\,,$$

и, таким образом, полученная оценка является *несмещённой*. Её дисперсия $\mathbf{D}\{\widehat{\overline{p}}\}$ равна

$$\mathbf{D}\{\widehat{\overline{p}}\} = \mathbf{D}\{\overline{m}/m\} = \frac{1}{m^2}\mathbf{D}\{\overline{m}\} = \frac{m\overline{p}^*(\mathbf{1}-\overline{p}^*)}{m^2} = \frac{\overline{p}^*(\mathbf{1}-\overline{p}^*)}{m}.$$

Здесь $\mathbf{1}$ — v-ичный вестор $(1, 1, \ldots, 1)^T$ и имеется ввиду Адамарово (покомпонентное) произведение векторов. Естественно, здесь и далее только $v-1$ компонент векторов будут независимы.

Известно, что это оценка с минимальной значением дисперсии в неравенстве Крамера–Рао. Таким образом полученная оценка имеет минимальную дисперсию в классе несмещённых (*эффективной* в общепринятом смысле). Поскольку $\mathbf{D}\{\widehat{\overline{p}}\}$ сходится по вероятности к 0 при возрастании m, то оценка является состоятельной.

Можно показать [7], что несмещенная оценка для $p_k^*(1-p_k^*)$, $k=\overline{1,v}$, есть

$$\frac{m}{m-1}\frac{m_k}{m}\left(1-\frac{m_k}{m}\right) = \frac{m_k(m-m_k)}{m(m-1)},$$

поэтому несмещённой оценкой $\overline{\mathbf{D}\{\widehat{\overline{p}}\}}$ для дисперсии $\mathbf{D}\{\widehat{\overline{p}}\}$ будет v-ичный вектор с компонентами

$$\frac{m_k(m-m_k)}{m^2(m-1)}, \quad k=\overline{1,v}.$$

Для наших целей относительные частоты могут быть приняты в качестве точечных оценок искомых вероятностей лишь в случаях больших m. Это связано с тем, что в условиях малой выборки не выполняется основное условие предельных теорем теории вероятностей — существование большого числа случайных событий

С другой стороны, точеные оценки в виде относительных частот в задачах распознавания образов часто становятся неприемлимыми с точки зрения опыта и интуиции. Например, корректное решающее правило мы вынуждены оценивать как 100% безошибочное, что даже при больших объёмах прецедентной информации противоречит здравому смыслу.

Отметим, что в последнем случае полученная оценка должна быть отвергнута и по формальным соображениям: значение $p_k=0$ не принадлежит области изменения параметра $\Theta=(0,1)^v$. Хотя в большинстве статистических моделей оказывается приемлемым рассматривать вместо области Θ ее замыкание $\overline{\Theta}$, но в нашем случае включать в рассмотрение невозможные или достоверные события вида $x\in X_k$ нет никаких оснований.

4. Байесовский подход

Байесовские точечные оценки $\widehat{\overline{p}}_W$ получаются как решения задачи минимизации функционала среднего риска записываемой как

$$\int\limits_{p_1+p_2+\ldots+p_v=1,\, p_1,p_2,\ldots,p_v\geq 0} W(\overline{p},\overline{q})f(\overline{p}\,|\,m_1,m_2,\ldots m_v)\,d\overline{p} = R(\overline{q}),$$

$$\widehat{\overline{p}}_W = \arg\min_{\substack{q_1+q_2+\ldots+q_v=1\\ q_1,q_2,\ldots,q_v\geq 0}} R(\overline{q}).$$

Здесь

- $\overline{p}=(p_1,p_2,\ldots,p_v)$, $\overline{q}=(q_1,q_2,\ldots,q_v)$, $\widehat{\overline{p}}_W$ — векторы из $\mathbb{R}^v_{\geq 0}$ для которых справедливо условие нормировки (5); причем последний — вектор оценок вероятностей при данной функции потерь W;
- $W(\overline{p},\overline{q}):(0,1)^v\times(0,1)^v\to\mathbb{R}_{\geq 0}$ — функция потерь для выбранных значений $\overline{q}$, когда $\overline{p}$ суть истинные значения искомых вероятностей;

– $f(\overline{p}\,|\,m_1, m_2, \ldots m_v)$ — апостериорная плотность вероятности вектора $\overline{p}$ при наблюдённых значениях $m_1, m_2, \ldots m_v$ попадания прецедентов в соответствующие области пространства образов.

Практически используют либо квадратичную

$$W(\overline{p}, \overline{q}) = \| \overline{p} - \overline{q} \|^2,$$

либо "простую" функцию потерь которая приписывает нулевые потери точке, которая апостериори наиболее вероятна и единичные потери остальным точкам области изменения параметра. Последняя приводит к методу максимизации апостериорной вероятностей, что при использовании принципа неопределённости Лапласа даёт, МП-оценку. Общепринято, что наиболее адекватные результаты получаются при использовании именно квадратичной функции потерь. Тот же результат — математическое ожидание апостериорной плотности вероятности искомого параметра (апостериорное среднее) — получается и при использовании любой другой выпуклой симметричной функции потерь [11].

В нашем случае формула Байеса имеет вид

$$f(\overline{p}\,|\,m_1, m_1 \ldots m_v) = \frac{f(\overline{p})\, f(m_1, m_1 \ldots m_v\,|\,\overline{p})}{\int\limits_{\substack{p_1+p_2+\ldots+p_v=1 \\ p_1, p_2, \ldots, p_v \geq 0}} f(\overline{p})\, f(m_1, m_1 \ldots m_v\,|\,\overline{p})\, d\overline{p}}. \tag{8}$$

Здесь

$$f(m_1, m_1 \ldots m_v\,|\,\overline{p}) = \prod_{k=1}^{v} p_k^{m_k}$$

является функцией правдоподобия и, естественно, выполняется условие нормировки (5).

Как отмечалось в п. 2, искомые вероятности $\overline{p} = \{p_k\}_{k=1}^{v}$ подчиняются полиномиальному распределению (6). В условиях отсутствия информации о весах прецедентов принимаем в качестве распределения $\overline{p}$ равномерное. Равномерное распределение есть распределение Дирихле $Di\,(1, 1, \ldots, 1; 1)$. Далее, используя (8) и (6) получаем, что апостериорная плотность вероятностей имеет вид

$$f(\overline{p}\,|\,m_1, m_2 \ldots m_v) = \frac{\Gamma(m+v)}{\Gamma(m_1+1)\Gamma(m_2+1)\ldots\Gamma(m_v+1)} \prod_{k=1}^{v} p_k^{m_k} =$$

$$= \frac{(m+v-1)!}{m_1!\, m_2! \ldots m_v!}\, p_1^{m_1} p_2^{m_2} \ldots p_v^{m_v},$$

$\overline{p} \in (0,1)^v$, т.е. является плотностью $(v-1)$-мерного распределения Дирихле

$$Di\,(m_1+1,\ m_2+1,\ \ldots,\ m_{v-1}+1;\ m_v+1).$$

Для квадратичной функции потерь байесовскими оценками $\widehat{p_i}$ вероятностей $\overline{p_i}$ будут являться компоненты вектора μ_k апостериорного среднего $\overline{\mu} = (\mu_1, \mu_2, \ldots \mu_v)^T$, равные [8]

$$\widehat{p_k} = \mu_k = \frac{m_k + 1}{m + v}, \quad k = \overline{1, v}. \tag{9}$$

Как уже указывалось, тот же результат получается и при любой симметричной выпуклой функции потерь.

Используя свойство воспроизводимости по m полиномиального распределения $M(m;\,\cdot\,)$ и свойства распределения Дирихле получим, что компоненты вектора дисперсий оценок (9) суть

$$\mathbf{D}\{\widehat{p}_k\} = \frac{p_k^*(1-p_k^*)\,m}{(m+v)^2},$$

а их несмещенные оценки —

$$\overline{\mathbf{D}\{\widehat{p}_k\}} = \frac{m_k(m-m_k)}{(m-1)(m+v)^2},\ k=\overline{1,v}.$$

Рассмотрим важный одномерный подслучай $v=2$, который соответствует разбиению пространства образов на две подобласти: правильных и неправильных классификаций. Пусть полученное р.п. из имеющихся m прецедентов m_r распознает правильно, а на остальных $m_w = m - m_r$ – ошибается. В соответствии с 9 точечная оценка $\widehat{p}_{W_1} = \widehat{p}_W$ вероятности ошибки распознавания $1-\nu$ есть

$$\widehat{p}_W = \frac{m_w+1}{m+2}. \tag{10}$$

Ясно, что полученные оценки являются смещёнными (но несмещеными асимптотически) и оценка состоятельными.

Легко видеть, что несмещенная оценка $\overline{\mathbf{D}\{\widehat{p}_W\}}$ дисперсии полученной оценки (10) равна

$$\overline{\mathbf{D}\{\widehat{p}_W\}} = \frac{m_w(m-m_w)}{(m+2)^2(m-1)}.$$

Имеем $\mathbf{D}\{\widehat{p}_W\} < \mathbf{D}\{\widehat{p}\}$ и дисперсия оценки $\mathbf{D}\{\widehat{p}_W\}$ в $(m+2)^2/m^2$ раз меньше минимальной граничной по неравенству Крамера-Рао.

Указанное обстоятельство объясняется тем, что полученная байесовская оценка есть оценка смещённая и понизить дисперсию оценки удалось именно за счет выхода класса несмещённых. Ясно, что выигрыш в дисперсии оценки будет особенно существенным при малых выборках.

5. Обсуждение полученных оценок. Оценки по медиане и минимаксные оценки

С общей точки зрения нет никаких оснований, кроме удобства математических свойств (а также традиции практиков), выделять равенство истинному значению именно математического ожидания оценки в качестве критерия несмещённости. Вместо математического ожидания могут также быть выбраны медиана распределения или его мода (т.н. медианная несмещёность или несмещёность по моде. В нашем случае мы столкнулись с ситуацией, когда смещённая оценка имеет дисперсию меньше, чем любая несмещённая, а значит и большую эффективность (оценку с меньшей дисперсией мы считаем более эффективной). Мы считаем это достаточным основанием для того, чтобы отказаться от рассмотрения лишь класса несмещённых оценок.

Во-первых, полученная оценка обладает свойством асимптотической несмещённости, а само смещёние невелико.

Во-вторых, представляется ясным, что для случая малых выборок, именно эффективность является основным критерием качества оценки. Наличие у оценок последнего нерассмотренного основного свойства — состоятельности — имеет ценность всё же в основном при теоретических исследованиях.

Заметим, что, неформально рассуждая, принятие МП-оценки (по моде) будет приводить к ошибкам, вообще говоря, редким, но, возможно, значительным, а байесовская оценка (по математическому ожиданию) повлечет, как правило, ошибки частые, но небольшие. Представляется, что данные оценки в силу указанных свойств являются в своём роде граничными, и исходя из специфики конкретных задач Z в качестве точечной оценки искомой вероятности p^* можно выбрать любое значение между модой и математическим ожиданием полученного B-распределения. Можно показать, что, например, его медиана $p_{(\beta)1/2}$

всегда расположена в указанном диапазоне и за оценку вероятности принять именно медиану. Такая оценка будет обладать свойством равновероятной недооценки и переоценки p^*, что может оказаться удобным для некоторых приложений.

Для нашей задачи можно попытаться использовать т.н. *W-минимаксную* оценку $\widetilde{p}$, средние потери которой при некоторой выбранной функции потерь W минимальны по $p^* \in (0,1)$. Если оказывается возможным подобрать априорное распределение, при котором полученная минимаксная оценка оказывается также равной и соответствующей байесовской, то такое априорное распределение называют *наименее благоприятным*.

Если выбрать функцию потерь квадратичной, то минимаксная оценка параметра p биномиального распределения будет иметь вид [2],

$$\widetilde{p} = \frac{\sqrt{m}}{1+\sqrt{m}}\frac{m_1}{m} + \frac{1}{1+\sqrt{m}}\frac{1}{2}.$$

Представляется, однако, что использование полученной оценки в нашем случае недостаточно оправдано с точки зрения «физики» задачи. Действительно, для вышеуказанной оценки наименее благоприятным распределением оказывается B-распределение $Be(\sqrt{m}/2,\ \sqrt{m}/2)$. Неясно, как параметры этого распределения могут быть обоснованы в рамках задачи Z.

6. Байесовские оценки при неравных весах прецедентов

Перейдем теперь к рассмотрению случая, когда прецедентная информация включает в себя вектор весов $\{\gamma_i = \gamma(x_i)\}_{i=m}^{m} = \overline{\gamma}_m$ (где не все компоненты равны) прецедентов $\overline{x}_m$.

Значение γ_i показывает "важность" или частоту встречаемости прецедента x_i. Часто заказчик, готовя исходные данные для решения задачи распознавания и желая дать как можно более полное и компактное описание пространства образов, намеренно или вынужденно [9] предоставляет разработчику список прецедентов более-менее равномерно распределённых по пространству образов, указывая большую или меньшую "типичность" данного прецедента с помощью приписывания ему соответствующего веса. Этот приём может существенно понизить объём предоставляемой прецедентной информации без потери её репрезентативности.

Заметим, что "важность" или "типичность" $\gamma_i \geq 1$ данного прецедента x_i можно трактовать как задание «дополнительных прецедентов» вблизи x_i с аналогичными признаками, и так, что дополнительные прецеденты всегда классифицируются также, как и x_i. Указанные «дополнительные прецеденты» назовем *квазипрецедентами*. Для точного соответствия с информацией, заложенной в весах, их число не обязано быть целым. Действительно, в этом случае та или иная классификация x_i приведет к соответствующему увеличению оценки вероятности p_i, что повысит её вклад в величину среднего риска (4) и отразит, таким образом, значимость данного прецедента. Заметим, что возможность такого представления информации о весах вытекает из гипотезы компактности.

Ясно, однако, что в рассматриваемом случае при остающейся верной гипотезе представительности, её форма в виде «Гипотеза 1» уже становится недостаточной. Поэтому для обоснования определения надежности выбранного р.п. данную гипотезу нужно дополнить предположениями относительно имеющегося вида прецедентной информации.

Наше основное предположение состоит в том, что веса объектов γ_i через количества квазипрецедентов описывают вероятности появления объектов в окрестностях x_i с тем же значением истинного классификатора $f^*(x_i)$. Таким образом в случае наличия в прецедентной информации вектора весов прецедентов для формализации ГП мы дополняем Гипотезу 1 нижеследующей Гипотезой 2.

[9] например, из-за отсутствия соответствующих данных

Гипотеза 2. При неравных весах $\gamma_i \neq const$, $i = \overline{1,m}$, набор прецедентов $\{x_i\}_{i=1}^m$ не является реализацией независимой выборки m случайных величин из генеральной совокупности с распределением $P(X)$ на $\mathcal{X}$, однако веса прецедентов $\{\gamma_1, \gamma_2, \ldots, \gamma_m\}$ отражают априорную информацию о распределении $P(x)$.

Поскольку мы трактуем веса как информацию о количестве квазипрецедентов в окрестности x_i, естественно считать, что $\gamma_i \geq 1$, $i = \overline{1,m}$, (для чего, при необходимости, поделим все веса на $\min \gamma_i$). Точнее, количество дополнительных квазипрецедентов будет описываться величинами $\gamma_i - 1$, т.к. в окрестности x_i уже есть один прецедент — сам x_i. Обозначим $\gamma_i' = \gamma_i - 1$, $i = \overline{1,m}$.

Естественно считать, что априорный вес μ_k' области X_k аддитивен и пропорционален весам попавших в него квазипрецедентов, т.е.

$$\mu_k' = \sum_{i:\, x_i \in X_k} \gamma_i', \quad k = \overline{1,v}.$$

Введём обозначение

$$\sum_{i:\, x_i \in X_k} \gamma_i = \mu_k.$$

Понятно, что

$$\mu_k' = \mu_k - m_k \geq 0, \quad k = \overline{1,v}, \tag{11}$$

поскольку $m_k = \sum\limits_{i:\, x_i \in X_k} 1$.

В качестве априорного распределения вероятностей на $\{X_k\}_{k=1}^v$ примем распределение Дирихле

$$Di\,(\mu_1' + 1,\ \mu_2' + 1,\ \ldots,\ \mu_{v-1}' + 1;\ \mu_v' + 1).$$

Представляется, что такая трактовка весов прецедентов достаточно адекватно отражает рассматриваемую ситуацию.

Обозначим

$$M = \sum_{k=1}^{v} \mu_k.$$

Используя формулу Байеса (8) и (11) получим апостериорное распределение вектора вероятностей $\overline{p} = \{p_1, p_2, \ldots, p_v\}$, $p_k \in (0,1)$, $k = \overline{1,v}$:

$$f(\overline{p}\,|\,m_1, m_2 \ldots m_v) = \frac{\Gamma(m + v + \sum\limits_{k=1}^{v} \mu_k')}{\prod\limits_{k=1}^{v} \Gamma(m_k + \mu_k' + 1)} \prod_{k=1}^{v} p_k^{m_k + \mu_k'} =$$

$$= \frac{\Gamma(M+v)}{\prod\limits_{k=1}^{v} \Gamma(\mu_k + 1)} \prod_{k=1}^{v} p_k^{\mu_k} = \frac{(M+v-1)!}{\mu_1!\,\mu_2! \ldots \mu_v!}\, p_1^{\mu_1} p_2^{\mu_2} \ldots p_v^{\mu_v},$$

которое является плотностью $(v-1)$-мерного распределения Дирихле

$$Di\,(\mu_1 + 1,\ \mu_2 + 1,\ \ldots,\ \mu_{v-1} + 1;\ \mu_v + 1).$$

Байесовской оценкой искомых вероятностей при функции потерь из указанного выше семейства будет вектор апостериорного среднего с компонентами

$$\widehat{p}_k = \frac{\mu_k + 1}{M + v}, \quad k = \overline{1,v}. \tag{12}$$

Эти значения и предлагается использовать в качестве точечных оценок вероятностей событий $x \in X_k$ в общем случае задачи Z.

Автор глубоко признателен академику РАН Ю.И.Журавлёву за понимание и поддержку.

Работа выполнена при поддержке гранта РФФИ N 01-01-00885-а.

Список литературы

[1] **Айзерман М.А., Браверман Э.М., Розоноэр Л.И.** Метод потенциальных функций в теории обучения машин. — М.: Наука, 1970.

[2] **Боровков А.А.** Математическая статистика. — М.: Наука, 1984.

[3] **Вапник В.Н., Червоненкис А.Я.** Теория распознавания образов. Стохастические проблемы обучения. — М.: Наука, 1974.

[4] **Вапник В.Н.** Восстановление зависимостей по эмпирическим данным. — М.: Наука, 1979.

[5] **Гуров С.И.** Оценки вероятности ошибок классификации при малом числе прецедентов. //Интеллектуализация обработки информации. Международная научная конференция ИОИ'2000. Тез. Докл. (Алушта, 10-14 июня 2000 г.). Симферополь: Крымский научный центр НАН Украины, Таврический национальный университет, 2000. С. 26.

[6] **Журавлев Ю.И.** Корректные алгебры над множеством некорректных (эвристических) алгоритмов. I, II, III. // Кибернетика, I: N 4, 1977, С. 5-17; II: N 6, 1977, С. 21-27; III: N 2, 1978, С. 35-43.

[7] **Кендал М., Стюарт А.** Статистические выводы и связи. /Пер. с англ. — М.: Наука, 1973.

[8] **Патрик Э.** Основы теории распознавания образов /Пер. с англ. Под. ред. Б.Р.Левина. — Сов. радио, 1980.

[9] **Рао С.Р.** Линейные ститистические методы и их применение. /Пер. с англ. — М.: Накуа, 1968.

[10] **Рудаков К.В.** Об алгебраической теории универсальных и локальных ограничений для задач классификации. // Распознавание, классификация, прогноз. Математические методы и их применение. Вып. 1. – М.: Наука, 1989. — С. 176-200.

[11] **Фукунага К.** Введение в статистическую теорию распознавания образов. /Пер. с англ. — М.: Наука, Гл. ред. физ.-мат. лит., 1979.

Поступила в редакцию 29.09.2001