С.И. Гуров

ОЦЕНКИ НАДЁЖНОСТИ АЛГОРИТМОВ КЛАССИФИКАЦИИ. I. ВВЕДЕНИЕ В ПРОБЛЕМУ. ТОЧЕЧНЫЕ ЧАСТОТНЫЕ ОЦЕНКИ[1]

Введение

В науке распознавания образов методы оценки надежности выбранного решающего правила развиты значительно слабее, чем теория построения распознающих алгоритмов. Проблема определения надёжности построенного распознающего алгоритма усугубляется ещё и тем, что при решении практических задач распознавания 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 неприемлемыми 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. Этой проблеме и посвящена настоящая работа.

1 Цель исследования. Основные понятия и определения

Под *пространством образов* $\mathcal{X}$ будем понимать произвольный непустой компакт[2]. Элементы $\mathcal{X}$ называются *образами*. Множество $\mathcal{X}$ полагается

[1]Работа выполнена при поддержке гранта РФФИ № 01-01-00885-а.

[2]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 их надежности.

разбитым на конечное число $s \geqslant 2$ 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 $x_1, x_2, \ldots, x_m, x_{m+1}, \ldots, x_{m+l}$ элементов $\mathcal{X}$. Такие образы с известной классификацией называют *прецедентами*. Мы разбиваем список прецедентов на две подпоследовательности: начальную $x_1, x_2, \ldots, x_m$ и заключительную $x_{m+1}, \ldots, x_{m+l}$, и считаем при этом, что последняя используется для построения алгоритма классификации, а первая – для оценки качества построенного алгоритма. Эти подпоследовательности образуют *обучающую* и *экзаменационную* выборки. Полагаем, что все элементы внутри *каждой* выборки различны. Будем обозначать $L = m + l$.

Здесь следует сделать важное замечание[3]. Далее мы считаем, что указанные выборки не имеют общих элементов. Это требование (при условии выполнения сформулированной ниже гипотезы представительности) гарантирует корректность применения результатов классификации на экзаменационной последовательности к задаче оценки качества обучения. 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 $\{f^*(x_i)\}_{i=1}^m = \bar{f}^*(\bar{x}_m)$ в точках $\bar{x}_m$. Функция f^* называется *истинным классификатором*. Заметим, что $y \in \mathcal{Y}$ является *номинальной переменной*.

Рассматривается задача классификации с непересекающимися классами в детерминированной постановке. *Классификатором* или *решающим правилом* (р.п.) называется любая функция $f : \mathcal{X} \to \mathcal{Y}$ (хотя на класс таких функций на практике накладываются те или иные ограничения). 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 оценённая степень уверенности ν в справедливости

[3]на него указал автору К.В. Воронцов

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 $\{\gamma_i = \gamma(x_i)\}_{i=1}^{L} = \bar{\gamma}_L$ прецедентов $\bar{x}_L$ мы будем включать в понятие прецедентной информации вместе с самими прецедентами и указанными наборами значений классификатора 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[4]. В этих случаях оценивать величину ν приходиться лишь по значениям функций $\{f^*(x_i)\}$, $f(x_i)\}$ и весов $\gamma(x_i)$ прецедентов $x_1, x_2, \ldots, x_m$, входящих в экзаменационную последовательность. 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 предположение о том, что *прецедентная информация отражает свойства*

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

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)[5] является определяющим фактором при оценке надежности построенного решающего правила, на котором основываются все дальнейшие выводы.

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[6]. Для этого предполагают, что $\mathcal{X}$ обладает вероятностной мерой $\mu(\cdot)$, т.е. для любого подмножества X из некоторой σ-алгебры подмножеств пространства образов существует интеграл

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

$P(X)$ называется, как известно, вероятностью или распределением вероятностей на $\mathcal{X}$. Вероятность события A будем обозначать $\mathrm{P}(A)$ или $\mathrm{P}\{A\}$. Для упрощения выкладок предполагают и существование 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.

---

[5]"Образам соответствуют компактные множества в пространстве выбранных свойств" [1]. По мнению автора, данная формулировка гипотезы компактности нуждается в существенной корректировке, однако этот вопрос не относится к теме данного исследования. Более развернутую формулировку ГК см. в [17].

[6]Было бы крайне интересно предложить невероятностную формулировку гипотезы представительности. Это позволило бы подойти к рассматриваемой проблеме с совершенно новой стороны. Нельзя ли использовать для этого нечеткие множества или теорию возможностей [27]?

Гипотеза 1. *На пространстве образов $\mathcal{X}$ задано (может быть неизвестное) распределение 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 $\mathrm{P}(x)$ известно, то оценка надежности построенного р.п. не представляет труда (см. ниже формулы (2) и (3)). Далее мы считаем функцию $\mathrm{P}(x)$ неизвестной.

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

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

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

Здесь $Q(K_i, K_j) = c_{ij} \geqslant 0$ – некоторая выбранная функция потерь или штрафа за отнесение образа 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 $\bar{x}_m$ используют функционал *эмпирического риска* $R^e_m(f)$:

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

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 [5], [6]. 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 $\bar{x}_m$.

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

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

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 образов $\mathcal{X}$, 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 $Z = Z(\mathcal{X},\ s,\ L,\ m,\ \bar{x}_L,\ \bar{\gamma}_L,\ \bar{f}^*(\bar{x}_L))$ состоит в выборе р.п. f, минимизирующего тот или иной функционал $R_Z(\cdot)$ (обычно это средний риск) и оценки полученной величины $R_Z(f)$. 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. Различные подходы к построению классификаторов рассматриваются, например, в [1], [32], [33], [26], [34] и в других монографиях и учебных пособиях. Также существует [12], [29] универсальный метод построения *корректных* (точных на прецедентах) алгоритмов классификации. В настоящей работе рассматриваются методы решения подзадачи $Z2$ задачи Z при выбранном классификаторе f (т.е. подзадача Z1 считается уже решённой).

В конце данного раздела уточним, что подразумевается под малой выборкой. Разные авторы по разному определяют это понятие. Выборку считают малой, если её объём не превосходит 200 [15], или 50 [37], или 30 [8], [30], или "нескольких десятков" [36], или 10-20 [18], или 10-15 [30], или "меньше расчетного числа, определенного при помощи специальной номограммы достаточно больших чисел" [22]. Часто вообще не определяют это понятие. Наша точка зрения основана на соображениях, изложенных в [7]. Здесь справедливо замечено, что при работе с выборками небольших объёмов приходится отказываться от классических способов статистической обработки, основанных на группировке наблюдений (гистограммы, критерии типа χ^2 и т.д.) и переходить к методам основанных на использовании каждой отдельной реализации (статистическая функция распределения, порядковые критерии типа критерия Уилкоксона и др.). Итак, выборку считаем *малой*, если при её обработке методами, основанными на группировке наблюдений и аппроксимационными методами, нельзя достичь заданных точности и достоверности[7]. Таким образом понятие малой выборки является условным и зависящим от поставленной задачи.

[7]Ср. определение малой выборки в [36] где за основу взят "факт отсутствия устойчивости информативных свойств и статистических характеристик".

2 Аналитические методы получения оценок надежности алгоритмов классификации

В данном разделе рассматриваются методы определения вероятностей ошибки распознавания, основанные на использовании только прецедентной информации при одном выбранном р.п., т.е. когда классификатор задан и фиксирован.

В основу различных методов определения надежности классификации кладутся те или иные предположения. Однако Гипотеза 1 является общим для всех из них: автору не известны подходы к решению рассматриваемой задачи, базирующиеся на иных предположениях.

В том случае, когда известен тип, к которому принадлежит неизвестное распределение $p(x)$, применяют различные методы параметрического оценивания, описанные, например, в [2], [5], [34]. Заметим, что даже в этом случае наличия достаточно большой информации о свойствах пространства образов, надежные оценки получаются лишь при значительных объёмах обучающей выборки.

Поскольку обычно неизвестен даже тип распределения $p(x)$, для восстановления последнего по прецедентной информации могут быть применены непараметрические методы (см, например, [6]). При этом, как правило, используется непараметрическое оценивание [25], основанное на подходе, восходящему к работам Розенблатта [42] и Парзена [41]. Основная идея связана здесь с "размазыванием" информации, полученной от каждого прецедента с помощью специальных функций, называемых *ядрами*. В многомерном случае выбирают ядра колоколообразного вида. Искомое распределение ищется в виде суперпозиции ядерных функций, привязанных к прецедентам. Не отрицая возможности такого подхода, отметим, что он требует задания коэффициента размытости, являющегося параметром ядерных функций. Вопрос о выборе такого параметра открыт. При малых объёмах обучающей выборки предлагается метод генерации новых m прецедентов в некоторой окрестности каждого прецедента соответствии с видом ядра (т.н. метод динамических сгущений). Однако оказывается, что при фиксированном объёме l выборки и росте числа m полученное распределение, вообще говоря, не стремится к истинному.

Перспективным представляется подход [7], основанный на объединении априорной и эмпирической информации об искомом распределении.

Важным является то, что задача (параметрического или непараметрического) восстановления $p(x)$ является, вообще говоря, более сложной [5], чем задача классификации[8]. Восстановление распределения вероятностей по эмпирическим данным является генеральной проблемой математической статистики. Искомая плотность вероятностей $p(x)$ полностью определяет все вероятностные свойства пространства $\mathcal{X}$, а не только используемые в задаче Z в связи с конкретным фиксированным его разбиением. Таким образом, восстановление неизвестной функции распределения в задачах распознавания образов, как правило, не является рациональным шагом. Исключения могут составлять лишь сильно вырожденные случаи[9]. В силу этого указанные подходы могут оказаться эффективными лишь при наличии большого объёма прецедентной информации.

Отметим, что в обоих описанных выше подходах рассмотренные методы применяют, как правило, для нахождения явного вида условных распределений $p(x \mid K_t)$ образов x из классов K_t, $t = \overline{1,s}$. Затем, считая набор вероятностей $\{p(K_t)\}_{t=1}^{s}$ появления образов данного класса известным, по формуле Байеса вычисляют вероятности $\{p(K_t \mid x)\}_{t=1}^{s}$ принадлежности образа x классу K_t. По данному набору распределений вычисляют отношения логарифмов средних рисков при данном р.п., на основе чего принимается решение о классификации данного образа. В некоторых частных случаях данный метод может быть доведен до получения оптимального классификатора в явном виде[10]. Однако и в этом случае вероятности ошибок классификации представляются в виде интегралов от условных вероятностей по определённым подобластям пространства признаков, причем границы этих областей оказываются заданными неявно и имеют, как правило, сложную форму. Ясно, что такие формулы для определения величин $\{p_{ij}\}_{i,j=1}^{s,s}$ непригодны для практического использования.

[8]При этом обе задачи являются некорректно поставленными по Адамару, т.к. допускают, очевидно, неединственность решения.

[9]Например, когда $p(x) = \prod_{i=1}^{n} p_i(x)$ в n-мерном признаковом пространстве. Сюда же, впрочем, относится и случай параметрического оценивания.

[10]В классическом случае нормальных (многомерных) условных распределений образов из каждого класса оптимальный разделитель двух классов есть квадрика, которая при дополнительном равенстве ковариционных матриц распределений становится линейной формой, называющейся *(линейной) дискриминантной функцией Фишера*.На практике же часто функцию Фишера находят и используют не проверяя ни нормальности распределений, ни равенства ковариционных матриц, получая при этом вполне приемлемые результаты.

Наиболее разработанные результаты в области надежности алгоритмов классификации получены в рамках уже упоминавшейся теории VC Вапника-Червоненкиса. В теории VC найдены необходимые и достаточные условия равномерной сходимости частот $\nu_l(A)$ появления событий A в l экспериментах по схеме Бернулли на заданном подмножестве $\mathcal{F}^*$ σ-алгебры событий к их вероятностям $\mathrm{P}\{A\}$, т.е. критерий выполнения соотношения

$$\mathrm{P}\left\{ \sup_{A\in\mathcal{F}^*} \left| \mathrm{P}(A) - \nu_l(A) \right| > \varepsilon \right\} \xrightarrow[l\to\infty]{} 0, \quad 0 < \varepsilon < 1 .$$

Для применения теории VC не требуется восстанавливать плотности распределения вероятностей, что является безусловным её достоинством.

Используя ту или иную теорию для решения частной задачи мы вынуждены принимать, соответствующим образом адаптируя, предположения, на которых эта теория базируется. Для нашей задачи оценки надежности р.п. эти предположения теории VC суть:

VC-1. Гипотеза 1.

VC-2. Классификатор $f(x)$ выбирается из фиксированного заранее семейства р.п. F.

Семейство F, задаёт подмножество $\mathcal{F}^*$ (обычно является параметрическим и записывается в виде $F(\tau)$, где τ – вектор параметров). Для получения оценок в теории VC требуется также вычислять меру разнообразия правил, составляющих класс F – его *ёмкость*. В случае конечности семейства F роль ёмкости играет его мощность.

Представляется ясным, что если использование ГП в форме "Гипотеза 1" не может вызвать серьёзных возражений, то принятие условия VC-2 при решении задачи распознавания Z далеко не всегда является оправданным. Это условие имеет место, например, в случае конечного признакового пространства, где семейство р.п. F всегда явно определено и конечно. Однако и в этих случаях, когда класс F зафиксирован перед решением задачи, часто не удаётся вычислить его ёмкость, поскольку нахождение её "сводится к громоздким комбинаторным вычислениям, которые не всегда можно провести" [21]. Имеется также большое число методов классификации с континуальными признаками (например, метод потенциальных функций в "машинной" реализации [1] или алгебраический подход к построению корректных распознающих алгоритмов [13],

[12]), когда классификатор конструируется непосредственно в процессе решения задачи и семейство F заранее не фиксируется. Более того, всегда можно сначала определить оптимальный в смысле минимума (2) классификатор f_{min}, а затем заново формально решить задачу Z, полагая $F = \{f_{min}\}$ и $|F| = 1$. Это ставит под вопрос применимость наиболее интересных результатов теории VC к нашей задаче. Кроме того, оценки полученные авторами теории [5], [6] в подавляющем числе случаев, к сожалению, оказываются непригодными для прямого использования на практике: значения надежности р.п. при имеющихся объёмах l выборок получаются крайне низкими и для получения оценок требуемой точности и достоверности необходимы величины l в десятки и сотни раз превышающие длину выборок, с которыми обычно приходится иметь дело. Между тем, опыт успешного решения самых разных задач распознавания свидетельствует о том, что эти оценки требуемых длин l сильно завышены (а для коэффициента доверия η, соответственно, занижены). Одной из причин этого является неявное предположение, что предъявляемое для оценки р.п. выбрано случайно из множества F. Как следствие, для оценки вероятности отклонения частоты $\nu_l(A)$ события A от его вероятности $P(A)$ используются оценка Хёфдинга [39]

$$\mathrm{P}\{\,|P(A) - \nu_l(A)| > \varepsilon\,\} < 2e^{-2\varepsilon^2 l}$$

или несколько более грубая оценка Бернштейна, которые не могут быть радикально усилены. Обобщая сказанное необходимо признать, что даже при принятии условия (VC-2) вопрос обоснования качества алгоритма распознавания для небольших значений l остаётся, открытым, а именно этот случай и представляют наибольший прикладной интерес.

В последнее время ([26], [21], [3], [4][11]) развивается байесовский подход к оценке качества р.п. В его основе лежит предположение, что искомый параметр (например, ν) распределен в соответствии с некотором априорном распределением, которое характеризует степень нашего знания о его значении. По данному распределению, используя формулу Байеса, определяется апостериорное распределение как функция от наблюдаемых величин. При этом происходит усреднение параметра по всевозможным распределениям в соответствии с выбранной функцией потерь[12], обычно

[11]Две последние работы наиболее близки к нашей. Заметим, что вначале формулы для точечной оценки вероятности ошибки получены здесь без привлечения условия VC-1.

[12]Не путать с функцией $Q(K_i, K_j)$ в (1)!

выбираемой квадратичной. В силу этого интервальные оценки параметров здесь получаются лучше, чем при применении теории VC, где оценки рассчитаны исходя из предположения о наихудшем случае. Чтобы обойти трудности, связанные с условием VC-2, рассматривается задача Z с логическими р.п., для которых мощность F конечна.

В работах [20], [23], [24], предприняты попытки улучшения оценок теории VC, используя полученное значение эмпирического риска как новое событие, а также некоторые правдоподобные априорные гипотезы. Заметим, что здесь также рассматриваются логические р.п.

Наше исследование в целом лежит в русле байесовского подхода. Впервые полученные результаты опубликованы в [9], [10] и [11].

3 Постановка задачи

Пусть в результате решения подзадачи $Z1$ задачи распознавания

$$Z = Z\left(\mathcal{X},\, s,\, L,\, m,\, \bar{x}_L,\, \bar{\gamma}_L,\, \bar{f}^*(\bar{x}_L)\right)$$

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 следующей статье.

Далее мы считаем, что пространство образов $\mathcal{X}$ разбито на $v \geqslant 2$ подобластей $\{X_k\}_{k=1}^{v}$ и обозначаем через m_k количество прецедентов, попавших в область X_k, $k = \overline{1, v}$; $\sum_{k=1}^{v} m_k = m$. В задачах классификации встречаются только следующие случаи значений v (напомним, что $s \geqslant 2$).

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.

$$X_{ij} = \left\{ x \mid x \in \mathcal{X},\, f^*(x) = K_i,\, f(x) = K_j \right\} = \\ = \{X_1,\, X_2,\, \ldots X_v\}$$

   (см. п. 1).

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

Обозначим $p_k = P(X_k) \geqslant 0$, $k = \overline{1, v}$. 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 вероятности $M(m; p_1, p_2, \ldots, p_v)$, плотность которого дается формулой

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

Отметим, что первые моменты полиномиального распределения суть

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

а матрица ковариаций –

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

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

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

Наша задача (статистического оценивания) будет состоять в том, чтобы построить точечные и интервальные 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 функции оценки должны быть применимы для случая малого числа m прецедентов.

В конце данного раздела остановимся кратко на двух основных группах методов математической статистики. Речь идёт о *частотном* и *байесовском* подходах.

Согласно последнему, по формуле Байеса, определяются апостериорные вероятности событий как априорные вероятности, умноженные на правдоподобия. Далее по полученным апостериорным вероятностям необходимо определить, какое событие имеет место в действительности. В простейшем случае за него может быть принято событие с максимальной апостериорной вероятностью. Такая функция оценки называется *оценкой по максимуму апостериорной вероятности.* В общем случае полученные апостериорные вероятности рассматриваются как распределение на некотором множестве, задающие на нём некоторые "веса". Далее с каждым событием, выбранным в качестве истинного значения, связывается величина, определяющая риск, связанный с данным выбором или соответствующие потери. Выбор события, считающегося реализующимся в действительности, производится исходя из минимума потерь. Таким образом байесовское решение есть решение *минимизирующее среднее значение риска.*

Могут быть предложены различные виды указанной функции потерь. В частности, оценка по максимуму апостериорной вероятности есть оценка с т.н. "простой" функцией потерь, которая приписывает нулевые потери точке, которая апостериори наиболее вероятна и единичные потери остальным точкам множества событий. В подавляющем же большинстве случаев при применении байесовского подхода используют квадратичную функцию потерь, у которой потери пропорциональны квадрату расстояния между даваемой оценкой и истинным значением параметра. Преимущество квадратичной функции потерь состоит в том, что она "подавляет" большие ошибки. Поэтому в тех задачах, где большие ошибки в оценивании параметра крайне нежелательны (к ним относится и наша задача оценки качества алгоритма классификации при малом числе прецедентов), следует использовать квадратичную функцию потерь. Легко показать [26], [35], что при квадратичной функции потерь оптимальная байесовская оценка будет совпадать с математическим ожиданием полученного распределения апостериорных вероятностей.

Указанные положения, применяемые для получения оценок и составляют *принцип Байеса (ПБ)*[13]. Принцип Байеса является одним

[13]Это определение отличается от приведенного в известной монографии [15].

из важнейших моментов в математической статистике. Обсуждение вопросов, связанных с ПБ можно найти, например, в [15], [16], [19] и др.

Мы видим, что для вычисления байесовских оценок необходимо знать распределение априорных вероятностей. Однако очень часто *априорные вероятности неизвестны, и их приходится определять, исходя* из дополнительной информации, специфичной для данной задачи. В случае же, когда такая информация отсутствует, вынужденно считают, что **все возможные события равновероятны**. Это допущение известно под названием *принципа неопределённости Лапласа*[14]. Хотя данный принцип является одним из наиболее спорных моментов в статистической теории, на практике в рамках байесовский подхода он применяется очень часто. Заметим, что в современных формулировках этого принципа допускается и не равновероятный характер априорного распределения [15]. Г. Джеффрис [40] развил указанный подход. Он предложил *неинформативное* априорное распределение для неизвестного параметра θ, с плотностью, пропорциональной $\sqrt{|I(\theta)|}$, где $|I(\theta)|$ есть определитель т.н. *информационной матрицы* (см. [19], [38]).

Естественно, и принцип неопределённости Лапласа, и сам принцип Байеса могут быть оспорены. В то же время ясно: если данные принципы отвергаются, они должны быть заменены чем-либо другим.

В частотном подходе предлагается считать, что в действительности имеет место событие, имеющее максимальное правдоподобие. Данное допущение называется *принципом максимального правдоподобия (МП)*. Таким образом, принцип МП основан на максимизации не апостериорной, а лишь условной вероятности наблюдаемого события. Ясно, что и против принципа МП могут быть высказаны возражения. С другой стороны, в случае принятия принципа неопределённости Лапласа и оценки по максимуму апостериорной вероятности (при строгой положительности апостериорных вероятностей, чего всегда можно добиться), результаты обоих подходов, очевидно, совпадут и методы на основе МП могут считаться частными случаями байесовского подхода [15].

В целом, преобладание положительных или отрицательных сторон любого подхода, как частотного, так и байесовского, зависит от конкретного их применения к конкретной задаче.

[14] а также *постулата Байеса* или *принципа равновероятности.*

4 Точечные оценки

4.1 Частотный подход

В рамках частотного подхода используются следующие методы получения точечных оценок неизвестных параметров [31]:

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

Метод максимального правдоподобия прямо основан на принципе МП. По этому методу максимизируется *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

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

Тогда функция правдоподобия есть

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

Мы видим, что функция правдоподобия зависит только от величин $m_1, m_2, \ldots, m_v$ и $p_1, p_2, \ldots, p_v$ и не зависит от T.

Теперь, поскольку максимумы L и $\log L$ совпадают, наша задача состоит в максимизации функции

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

при условии нормировки (5).

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

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

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

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

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

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

Метод моментов. Нетрудно видеть, что метод моментов, основанный на приравнивании выборочных моментов теоретическим, 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. Данные методы основаны на рассмотрении различных мер расхождения между наблюдёнными величинами $m_1, m_2, \ldots, m_v$ и их математическими ожиданиями $mp_1, mp_2, \ldots, mp_v$. Оценка $(\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_v)$ определяется как значения вероятностей, минимизирующие эту меру. Для оценивания используются такие меры, как "χ^2", "модифицированный χ^2", "расстояние Хеллингера", "дивергенция Кульбаха-Лейблера", "мера расхождения Холдейна" и др. [31], [28]. 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 функцию оценки в виде относительных частот.

Из сказанного выше ясно, что в основе метода максимального правдоподобия не лежит никаких строго обоснованных соображений,

а широкое использование МП-оценок и вера в их хорошие качества основаны, отчасти, на асимптотической оптимальности, как правило, их свойств. Речь идет об известных свойствах несмещенности, состоятельности и эффективности МП-оценок.

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

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

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

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

Здесь $\mathbf{1}$ – v-ичный вектор $(1, 1, \ldots, 1)^T$ 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 (см., например, [16]). Таким образом полученная оценка имеет минимальную дисперсию в классе несмещённых (т.е. эффективной в общепринятом смысле).

Поскольку $\mathbf{D}\{\widehat{\bar{p}}\}$ сходится по вероятности к 0 при возрастании m, то оценка является состоятельной.

Можно показать [16], что несмещённая 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 функцией оценки $\overline{\mathbf{D}\{\widehat{\bar{p}}\}}$ для дисперсии $\mathbf{D}\{\widehat{\bar{p}}\}$ 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 и "поэтому при оценивании по конечному малому числу набору данных асимптотические характеристики могут ввести в заблуждение" [36].

С другой стороны, точеные оценки в виде относительных частот в задачах распознавания образов часто становятся неприемлемыми 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.

Очевидно также, что оценки по малому числу прецедентов по своей сути не могут обладать большой точностью. Данное обстоятельство, например, отражено в [14], где указано, что процентная относительная частота $\frac{r}{n}100\%$ при $25 \leqslant n \leqslant 200$ должна записываться без знаков после запятой (а начиная с $n = 2000$ – с двумя знаками после запятой). Если придерживаться данного правила, то при $n < 25$ выборка должна считаться малой (и тогда только одна цифра является значащей?).

Точечные оценки для одномерного случая элементарно получаются из полученных выше для многомерного:

$$\hat{p}_w = m_w/m,\ \hat{p}_r = \nu = m_r/m\,.$$

Литература

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

2. Андерсон Т. Введение в многомерный статистический анализ /Пер. с англ. – М.: Физматгиз, 1963.

3. Бериков В.Б. Об устойчивости агоритмов распознавания в дискретной постановке //Искусственный интеллект. Научно-теоретический

журнал. НАН Украины. Ин-т проблем искуст. интелекта. Донецк, 2000, № 2. С. 5-8.

4. Бериков В.Б. Байесовский подход к определению качества распознавания //"Математические методы распознавания образов" (ММРО-10). Доклады X Всероссийской конференции. – М.: Российская академия наук, Вычислительный центр, 2001, С. 6-9.

5. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов. Стохастические проблемы обучения. – М.: Наука, 1974.

6. Вапник В.Н. Восстановление зависимостей по эмпирическим данным. – М.: Наука, 1979.

7. Гасканов Д.В., Шаповалов В.И. Малая выборка. – М.: Статистика, 1978.

8. Гмурман В.Е. Теория вероятностей и математическая статистика. – М.: Высшая школа, 1977.

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

10. Гуров С.И. Оценки ошибок алгоритмов распознавания // Spectral end Evaluation Problems: Proceedings of the Eleventh Crimean Autumn Mathematical School-Symposium. Vol. 12. / Simferopol: National Turida V. Vernadsky University, Black Sea Branch of Moscow State University, Crimean Scientific Centre, Crimean Academy of Sciences, Crimean Mathematical Foundation, 2002. – С. 185-196.

11. Гуров С.И. Методы определения ошибок классифицирующих алгоритмов // Искусственный интеллект (Донецк) №2, 2002. с. 88-98.

12. Журавлев Ю.И. Корректные алгебры над множеством некорректных (эвристических) алгоритмов. I, II, III. // Кибернетика, I: № 4, 1977, С. 5-17; II: № 6, 1977, С. 21-27; III: № 2, 1978, С. 35-43.

13. Журавлев Ю.И. Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. Сб. статей. – М.: Наука, Вып. 33, 1978. С. 5-69.

14. Закс Л. Статистическое оценивание /Пер. с нем. под ред. Ю.П. Адлера, В.Г. Горского. – М.: Статистика, 1976.

15. Кендал М., Стюарт А. Теория распределений /Пер. с англ. – М.: Наука, 1966.

16. Кендал М., Стюарт А. Статистические выводы и связи /Пер. с англ. – М.: Наука, 1973.

17. Кольцов П.П. Математические модели теории распознавания образов //Компьютер и задачи выбора /Автор предисл. Ю.И.Журавлёв. М.: Наука, 1989. С. 89-119.

18. Кремер Н.Ш. Теория вероятностей и математическая статистика. – М.: ЮНИТИ-ДАНА, 2000.

19. Леман Э. Теория точечного оценивания /Пер. с англ. – М.: Наука, Гл. ред. физ.-мат. лит., 1991.

20. Лбов Г.С., Старцева Н.Г. Сложность распределений в задачах классификации //Доклады РАН, 1994, том 338, № 5. С.

21. Лбов Г.С., Старцева Н.Г. Логические решаюшие функции и вопросы статистической устойчивости решений. – Новосибирск: Изд-во Ин-та математики, 1999.

22. Методы статистического анализа и обработка малого числа наблюдений при контроле качества и надежности приборов и машин. – Л., 1974.

23. Неделько В.М. Оценивание доверительного интервала вероятности ошибки решающей функции распознавания по эмпирическому риску // "Математические методы распознавания образов" (ММРО-9). Доклады 9-й Всероссийской конференции. – М.: Российская академия наук, Вычислительный центр, 1999. С. 88-90.

24. Неделько В.М. Критерий оценки качества решающей функции по эмпирическому риску в задаче классификации // Искусственный

интеллект. Научно-теоретический журнал. НАН Украины. Ин-т проблем искуст. интелекта. Донецк, 2000, № 2. С. 172-178.

25. Обучающиеся системы обработки информации и принятия решений: непараметрическимй подход / А.В. Лапко, С.В. Ченцов, С.И. Крохов, Л.А. Фельдман. – Новосибирск: Наука. Сибирская издательская фирма РАН, 1996.

26. Патрик Э. Основы теории распознавания образов /Пер. с англ. Под. ред. Б.Р. Левина. – М.: Сов. радио, 1980.

27. Пытев Ю.П. Возможность. Элементы теории и применения. – М.: Эдиториал УРСС, 2000.

28. Рао С.Р. Линейные статистические методы и их применение /Пер. с англ. – М.: Накуа, 1968.

29. 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.

30. Смирнов Н.В., Дунин-Барковский И.В. Курс теории вероятностей и математической стстистики для технических приложений. – М.: Наука, 1965.

31. Справочник по теории вероятностей и математической стстистике /В.С. Королюк, Н.И. Портенко, А.В. Скороход, А.Ф. Турбин. – М.: Наука, 1985.

32. Ту Дж., Гонсалес Р. Принципы распознавания образов /Пер. с англ. – М.: Мир, 1978.

33. Фомин В.Н. Математическая теория обучаемых опознающих систем. – Л.: Изд-во Ленингр. ун-та, 1976.

34. Фу К. Последовательные методы в распознавании образов и обучении машин /Пер. с англ. – М.: Наука, 1971.

35. Фукунага К. Введение в статистическую теорию распознавания образов /Пер. с англ. – М.: Наука, Гл. ред. физ.-мат. лит., 1979.

36. Фурсов В.А. Идентификация моделей систем формирования изображений по малому числу наблюдений. – Самара: Самар. гос. аэрокосм. ун-т., 1998.

37. Шор Я.Б. Статистические выводы анализа и контроля надежности и качества. – М.: Сов. радио, 1962.

38. Box G.E., Tiao G.C. Bayesian Inference in Statistical Analysis. – Mass.: Addisson-Wesley, Reading, 1973.

39. Hoeffding W. Probability inequalities for sums of founded random variables // J. Amer. Statist. Assoc., 1963, Vol. 58. Pp. 13-30.

40. Jeffreys H. The Theory of Probability. – Oxford: Oxford University Press, 1961.

41. Parzen E. On estimation of a probability density function and mode // Annals of Math. Statist., 1962, v. 33, № 3.

42. Rozenblatt M. Remarks of some non-parametric estimates of a density function // Annals of Math. Statist., 1956, v. 27, № 3.