УДК 519.68+519.2

КАК ОЦЕНИТЬ НАДЕЖНОСТЬ АЛГОРИТМА КЛАССИФИКАЦИИ I. ВВЕДЕНИЕ В ПРОБЛЕМУ И ТОЧЕЧНЫЕ ОЦЕНКИ [1]

С.И. Гуров

Россия, г. Москва, ф-т ВМиК МГУ им. М.В. Ломоносова
E-mail: *sgur@cs.msu.su, gurov@ccas.ru.*

Abstract.
In this paper, the new dot statistical estimations of pattern algorithms reliability are presented. These estimations can be used when shot training samples are given.

Введение

Как известно, распознавание образов не сводится к методам разделения наборов подмножеств в признаковом пространстве. Для заказчика важно не только получить алгоритм, реализующий (возможно с некоторыми ошибками) требуемое разделение классов, но и иметь оценку надежности решения поставленной задачи, т.е. 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 неприемлемыми для заказчика, как, например, нулевая точечная оценка ошибки при корректном алгоритме распознавания.

Вышесказанное свидетельствует о необходимости предложить новые подходы к построению оценок надёжности алгоритмов распознавания, способных охватить важный случай малого числа прецедентов. В данной работе дается общее введение в проблематику и рассматриваются точечные оценки ошибок классификации. Интервальным оценкам будет посвящена следующая статья.

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

1. Основные понятия и определения

Под *пространством* образов X будем понимать произвольный непустой компакт. Обычно также считают, что 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.

Элементы X называются *образами*. Множество X полагается разбитым на конечное число $s \geqslant 2$ попарно непересекающихся областей $\{\mathsf{X}_t\}, t = \overline{1,s}$, называемых *классами*.

Существенным является то, что информация о разбиении X на классы ограничивается знанием о принадлежности к тому или иному классу конечного числа $x_1, x_2, \ldots, x_m$ элементов X . Такие образы с известной классификацией называют *прецедентами*, а их совокупность — *обучающей выборкой* (или *последовательностью*) $\bar{x}_m$ (длины или объёма m). Обозначив через Y множество символов классов $\{K_1, \ldots, K_s\}$ можно сказать, что существует функция $f^* : \mathsf{X} \to \mathsf{Y}$, 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*.

Рассматривается задача классификации с непересекающимися классами в детерминированной постановке. *Классификатором* или *решающим правилом* (р.п.) называется любая функция $f : \mathsf{X} \to \mathsf{Y}$ (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 X в процессе классификации на практике равенство $f(x) = f^*(x)$ (правильная классификация), выполнялось «как можно чаще». Количественно оценённая степень уверенности ν в справедливости данного равенства для произвольного $x \in \mathsf{X}$ 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)*. В экзаменационную последовательность могут входить частично или полностью элементы обучающей последовательности, возможно пополненные за счет набора дополнительных контрольных прецедентов. 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 $\bar{x}_m$ 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[2]. 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 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 пространства образов, связанные с определённым распределением появляющихся образов по различным подобластям* X *в процессе классификации на практике.*

Гипотеза представительности, принятая в той или иной форме в рамках конкретной задачи, вместе с гипотезой компактности (ГК)[3] 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[4]. Для этого предполагают, что X обладает вероятностной мерой $\mu(\cdot)$, т.е. для любого подмножества X

[2]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.).

[3]**«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 [16].

[4]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]?

пространства образов существует интеграл

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

$P(X)$ называется, как известно, функцией распределения вероятностей на X. Вероятность события A будем обозначать P(A) или P{A}. Для упрощения выкладок полагают и существование функции плотности вероятности p(x) на 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 X задана (может быть неизвестная) функция распределения вероятностей $P(X)$, $X \subseteq \mathsf{X}$, и любой рассматриваемый набор образов $x_i, x_2, \ldots, x_l$ 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* $R(f)$:

$$R(f) \triangleq \int_X \left(\sum_{f^*(x)\in Y} \sum_{f(x)\in Y} Q(f^*(x), f(x)) \right) p(x)dx, \tag{1}$$

где $Q : \mathsf{Y} \times \mathsf{Y} \to \mathbf{R}_{\geqslant \mathbf{0}}$ ($\mathbf{R}_{\geqslant \mathbf{0}}$ — множество неотрицательных действительных чисел, $\triangleq$ означает «равно по определению»).

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* $R^e_m(f)$:

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

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 [4], [5]. К сожалению оказывается, что в рамках VC гарантировать малость $R(f_{min})$ при малом $R^e_m(f_m in)$, где

$$f_{\min} = \arg\min_f\{R^e_m(f)\}$$

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}

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

где $X_{ij} \triangleq \{x\,|x\ \in \mathsf{X}, f^*(x), K_i, f(x) = K_j\}$. Подобласти $\{X_{ij}\}^{s,s}_{i,j=1}$ — это s^2 областей разбиения пространства образов 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 $c_{ii} = c_r, c_{ii} = c_w, (i \neq j)$ можно полагать 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. определяеться набором вероятностей $\{p_{ij}\}^{s,s}_{i,j=1}$ или величиной ν (вероятность правильной классификации).

Задача классификации $Z = Z(\mathsf{X}, s, m, \overline{x}_m, \overline{\gamma}_m, \overline{f}^*(\overline{x}_m))$ состоит в выборе р.п. f, минимизирующего тот или иной функционал $R_Z(\cdot)$ (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 паралельно и согласовано. На практике же, в силу вышеупомянутых причин, обе подзадачи решают, как правило, приближённо 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], [34], [26], [35] и в других известных монографиях и учебных пособиях. Также существует [10], [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. $Z1$ подзадача 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 [14] или 50 [38] или 30 [30] или «несколько десятков» [37] или 10-20 [17] 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 [6]. Здесь справедливо замечено, что приработе с выборками небольших объёмов приходиться 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 доверительности*[5]. Таким образом, понятие малой выборки является условным и зависящим от поставленной задачи.

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, базирующихся на иных предположениях.

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

[5]Ср. определение малой выборки в [37] где за основу взят «факт отсутствия устойчивости информативных свойств и статистических характеристик».

Поскольку обычно неизвестен даже тип распределения $p(x)$, для восстановления последнего по прецедентной информации могут быть применены непараметрические методы (см., например, [5]). При этом,как правило, используется параметрическое оценивание [25], основанное на подходе, восходящему к работам Розенблантта [43] и Парзена [42]. Основная идея 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 эадания 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 окресности 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 [6], основанных 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 [4], чем задача классификации[6]. Восстановление распределения вероятностей по эмпирическим данным является генеральной проблемой математической статистики. Искомая плотность вероятностей $p(x)$ плотностью определяет все вероятностные свойства пространства 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[7]. 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 $p(x|K_t)$ образов x из классов K_t, $t = \overline{1,s}$. Затем, считая набор вероятностей $\{p(K_t)\}_{t=1}^{s}$ появления образа данного класса известным, по формуле Байеса вычисляют вероятности $\{p(K_t|x)\}_{t=1}^{s}$ 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

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

[7]Например, когда $p(x) = \prod_{i=1}^{n} p_i(x)$ в n-мерном признаковом пространстве или в случае параметрического оценивания $p(x)$.

случаях данный метод может быть доведен до получения оптимального классификатора в явном виде[8]. 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.

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 F^* σ-алгебры к их вероятностям $P\{A\}$, т.е. критерий выполнения соотношения

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

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

В применении к задаче надежности р.п. теория VC базируется на следующих двух предположениях.

VC-1. Гипотеза 1.

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

Семейство F, задаёт подмножество 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 комбинированным вычислениям, которые не всегда можно провести» [21]. Имеется также большое число методов классификации с континуальными признаками (например, метод потенциальных

[8]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 используя не проверяя ни нормальности распределений, ни равенства ковариционных матриц, получая при этом вполне приемлемые результаты.

функций в «машинной» реализации [1] или алгебраический подход к построению корректных распознавающих алгоритмов [11], [10], 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 [4], [5] 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 [40]

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

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], [2], [3][9]) развивается байесовский подход к оценке качества р.п. (см. ниже п. 4). В его основе лежит предположение, что искомый параметр (например, ν) распределен в соответствии с некоторым 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[10] , обычно 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

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

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

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 [7], [8] и [9].

3. Постановка задачи

Пусть в результате решения подзадачи $Z1$ задачи распознавания

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

построено р.п. $f(x)$. Предположим пока, что $\gamma_1 = \gamma_2 = \ldots = \gamma_m$ и примем гипотезу представительности в «Гипотеза 1». Случай неравных весов прецедентов будет рассмотрен в п. 5.2.4.

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

1. $\nu = 2$. Здесь X_1 и X_2 суть области правильных и неправильных классификаций.
2. $\nu = s^2$. Здесь $\{X_k\}_{k=1}^{\nu}$ суть переобозначенные области $\{X_{ij}\}_{i,j=1}^{s,s}$ пространства образов, т.е. $X_{ij} = \{x \,|\, x \in \mathsf{X},\ f^*(x) = K_i,\ f(x) = K_j\} = \{X_1, X_2, \ldots, X_\nu\}$ (см. п. 1).
3. $\nu = s^2 + 1$. Здесь к определённым выше областям добавляется область, соответствующая случаю отказа от классификации.

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

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

и при данном ν мы имеем $(\nu - 1)$-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 $p\,(m_1, m_2, \ldots, m_\nu)$ того, что при независимой случайной выборке m элементов из X в соответствии с распределением $P(\cdot)$ соотношение $x \in X_k$ будет выполняться m_k раз, $k = \overline{1,\nu},\ \sum\limits_{k=1}^{\nu} m_k = m$ имеет $(\nu - 1)$-мерное полиномиальное (мультиномиальное)

распределение $M\,(m;\,p_1,\,p_2,\,\ldots,\,p_\nu)$, функция плотности вероятности которого дается формулой

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

Отметим, что первые моменты полиномиального распределения суть

$$\mu_k = mp_k,\quad k = \overline{1,\,\nu-1} \tag{7}$$

а матрица ковариаций —

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

При $\nu = 2$, $p_1 = p$ имеем биномиальное распределение $B_i(m,\,p)$ с функцией плотности вероятности

$$p\,(m_1) = \left(\begin{array}{c} m \\ m_1 \end{array}\right) p^{m_1}(1-p)^{m-m_1};\quad p \in (0,\ 1).$$

Наша задача (статистического оценивания) состоит в том, чтобы построить точечные и интервальные оценки неизвестных, но фиксированных величин $p_1,\,p_2,\,\ldots,\,p_\nu$ по случайным значениям $m_1,\,m_2,\,\ldots,\,m_\nu,\ \sum\limits_{k=1}^{\nu} m_k = m$. Построенные функции оценки должны быть применимы для случая малого числа m прецедентов.

Заметим, наконец, что, поскольку искомые вероятности принадлежат подмножествам (0, 1) конечномерного евклидова пространства, рассматриваемые ниже методы относятся к параметрическим методам математической статистики.

4. Два подхода к построению оценок

Общая задача построение оценок значений неизвестных величин состоит в следующем. Необходимо ответить на вопрос: ***«Какое из событий $\{B_i\}_{i\in I}$, составляющих полную группу B несовместных событий, имеет место в действительности?».***

Область изменения индекса I есть некоторое непустое конечное или континуальное измеримое множество. К сожалению, сами события из B не наблюдаемы. А наблюдаемо некоторое (возможно сложное) событие A, как-то связанное с событиями $\{B_i\}_{i\in I}$. Требуется предложить способ, позволяющий по наблюдённому событию A ответить на поставленный выше вопрос. Здесь ещё раз важно подчеркнуть, что искомое событие хоть и неизвестно, но фиксировано.

В математической статистике имеется два подхода к получению оценок характеристик распределений по случайным наблюдениям: *частотный* и *байесовский*.

Байесовский подход основан на использовании известной формулы, носящей его имя, и которая в простейшем случае конечного I имеет вид

$$\mathsf{P}\{B_i \,|\, A\} = \frac{\mathsf{P}\{B_i\} \cdot \mathsf{P}\{A \,|\, B_i\}}{\sum\limits_{i \in I} \mathsf{P}\{B_i\} \cdot \mathsf{P}\{A \,|\, B_i\}}. \tag{8}$$

Напомним, что вероятности в (8) носят названия: $\mathsf{P}\{B_i\}$ — *априорных* (условных), $\mathsf{P}\{B_i \,|\, A\}$ — *апостериорных*, а значения $\mathsf{P}\{A \,|\, B_i\}$ (с точностью до мультипликативной константы) — *правдоподобий*. Формула Байеса позволяет, таким образом, находить апостериорные вероятности, как соответствующие априорные, умноженные на правдоподобия. Последние обычно могут быть установлены исходя из той или иной принятой модели появления событий. Априорные вероятности мы считаем мерой нашего незнания, таким образом, мы придерживаемся т.н. «субъективного» подхода в статистических задачах оценивания (см. [14]).

Пусть апостериорные вероятности вычислены. Теперь необходимо определить, какое событие из системы B имеет место в действительности. В простейшем случае за него может быть принято событие B_i с максимальной апостериорной вероятностью. Такая функция оценки называется, как известно, *оценкой по максимуму апостериорной вероятности*. В общем случае полученные апостериорные вероятности рассматриваются как распределение на множестве $I = \Theta$, задающие на нём некоторые «веса». Θ обозначает множество значений неизвестного параметра θ. Далее с каждым B_i, 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 и 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 $I = \Theta$. 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, 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)*[11]. Принцип Байеса является одним из важнейших моментов в математической статистике. Обсуждение вопросов, связанных с ПБ, можно найти, например, в [14], [15], [18] и др.

Мы видим, что байесовский подход основан на максимизации совместных распределений событий A и $\{B_i\}_{i\in I}$, и для его применения 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 события из группы B равновероятны. Это допущение известно под названием *принципа неопределённости Лапласа*[12]. 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 [14]. Г. Джеффрис [41] 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. [18], [39]).

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 A при условиях реализации B_i, $i \in I$. 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.

Частотный подход не ограничивается, естественно, только принципом МП и методами, на нём основанными. Просто этот подход, в отличие от байесовского, просто не связан ни с какими априорными предположениями о том или ином распределении каких-либо величин. Понятно, что это есть и сильная, и слабая его сторона. В целом, преобладание положительных или отрицательных сторон любого подхода, как

[11] Это определение отличается от приведенного в известной монографии [14].

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

частотного, так и байесовского, зависит от конкретного их применения к конкретной задаче.

5. Точечные оценки

5.1. **Частотный подход.** Рассмотрим сразу многомерный случай $\nu > 2$.

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 *функция правдоподобия* L, которая для нашего случая определяется следующим образом.

Результат определения количества прецедентов в областях $\{X_k\}_{k=1}^{\nu}$ представим в виде (0, 1)-таблицы $T = \{t_{k,\,i}\}_{k,\,i=1}^{\nu,\,m}$, где

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

Ясно, что

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

Тогда функции правдоподобия

$$L\,(T; p_1, p_2, \ldots, p_\nu) = const \cdot p_1^{t_{1,1}+\ldots+t_{1,m}} p_2^{t_{2,1}+\ldots+t_{2,m}} p_i^{t_{1,1}+\ldots+t_{i,m}+\ldots+t_{2,m}} \ldots p_\nu^{t_{\nu,1}+\ldots+t_{\nu,m}} =$$

$$= const \cdot p_1^{m_1} p_2^{m_2} \ldots p_\nu^{m_\nu}$$

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

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

при условии нормировки (5).

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

$$\mathsf{L}(p_1, p_2, \ldots, p_\nu, \lambda) = \ln L(p_1, p_2, \ldots, p_\nu) + \lambda \cdot \{1 - \sum_{k=1}^{\nu} p_k\}$$

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

$$\begin{cases} \frac{m_k}{p_k} - \lambda = 0, \ k = \overline{1,\nu}, \\ \sum\limits_{k=1}^{\nu} p_k = 1, \end{cases}$$

решения которой суть $\lambda = m, \ \widehat{p}_k = m_k/m, \ k = \overline{1,\nu}$.

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

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 — $m_k, k = \overline{1,\nu}$.

Метрические методы основаны на рассмотрении различных мер расхождения между наблюдёнными величинами $m_1, m_2, \ldots, m_\nu$ и их математическими ожиданиями $mp_1, mp_2, \ldots, mp_\nu$. Оценка $(\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_\nu)$ определяется как значения вероятностей, минимизирующие эту меру. Для оценивания используются такие меры, как «X^2», «модифицированный X^2», «расстояние Хеллингера», «дивергенция Калбэка-Лейбера», «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 X^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 $M\{\hat{\bar{p}}\}$ вектора оценок $\{p_k\}_{i=k}^{\nu}$ есть (с учетом (7) и обозначений $\bar{m}(m_1, m_2, \ldots, m_\nu)^T$ и $\bar{p}^*$ — ν-ичный вектор истинных значений вероятностей)

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

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

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

Здесь 1 — ν-ичный вектор $(1, 1, \ldots, 1)^T$ и имеется ввиду адамарово (покомпонентное) произведение векторов. Естественно, здесь и далее только $\nu - 1$ компонент векторов будут независимы.

Известно, что это оценка с минимальной значением дисперсии в неравенстве Крамера-Рао для несмещённых оценок[13]. Поскольку $D\{\hat{\hat{p}}\}$ сходится по вероятности к 0 при возрастании m, то оценка является состоятельной.

Можно показать, что несмещённая оценка для $p_k^*(1-p_k^*),\ k=1$, 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 $\overline{D\{\hat{\hat{p}}\}}$ для дисперсии $D\{\hat{\hat{p}}\}$ будет ν-ичный вектор с компонентами

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

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[14].

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 $\Theta=(0,1)^\nu$. 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 [12], где указано, что процентная относительная частота $\frac{r}{n}100\%$ при $25\leqslant n\leqslant 200$ должна записываться без знаков после запятой (а начиная $c=2000$ — с двумя знаками после запятой).

В дальнейшем мы будем придерживаться данного правила (из него, в частности, следует, что при $n<25$ выборка считается малой[15]).

[13] т.е. *эффективной* в общепринятом смысле.

[14] «Поэтому при оценивании по конечному малому числу набору данных асимптотические характеристики могут ввести в заблуждение» [37].

[15] и тогда только одна цифра является значащей?

5.2. **Байесовский подход.** Байесовские точечные оценки $\hat{\bar{p}}_W$ получаются как решения задачи минимизации функционала среднего риска записываемой как

$$\int\limits_{S_{\nu-1}(\bar{p})} W(\bar{p},\bar{q}) f(\bar{p}|m_1,m_2,\ldots,m_\nu)d\bar{p} = R(\bar{q}),\ \bar{p}_W = \arg\min_{\bar{q}\in S_{k-1}(\bar{x})} R_(\bar{q})$$

Здесь и далее

- $S_{\nu-1}(\overline{x}) = \{(x_1,x_2,\ldots,x_\nu); x_k \geqslant 0, k = \overline{1,\nu}; \sum\limits_{k=1}^{\nu} p_k = 1\}$ – $(\nu-1)$-мерный симплекс в пространстве $\mathbf{R}^\nu$
- $\bar{p},\bar{q},\hat{\bar{p}}_W$ — векторы из $S_{\nu-1}(\overline{x})$, причем последний — вектор оценок вероятностей при данной функции потерь W;
- $W(\bar{p},\bar{q}) : S_{\nu-1}(\overline{x}) \times S_{\nu-1}(\overline{x}) \to \mathbf{R}_{\geqslant 0}$ —функция потерь для выбранных значений $\bar{q}$, когда $\bar{p}$ суть истинные значения искомых вероятностей;
- $f(\bar{p}\,|\,m_1,m_2,\ldots,m_\nu)$ — апостериорная плотность вероятности вектора $\bar{p}$ при наблюдённых значениях $m_1,m_2,\ldots,m_\nu$ попадания прецедентов в соответствующие области пространства образов.

Решение данной задачи в значительной мере определяется видом функции потерь.

«Простая» функция потерь (п. 4) приводит к методу максимизации апостериорной вероятностей, которая, при использовании принципа неопределённости Лапласа, даёт, как мы видели, полученную ранее в рамках частотного подхода МП-оценку.

Практически используют либо квадратичную

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

либо нормированную квадратичную функцию потерь

$$W(\bar{p},\bar{q}) = c(\bar{p}) = \|\bar{p}-\bar{q}\|^2 \Big/ \prod_{k=1}^{\nu} p_k\ ,$$

где $c(\bar{p})$ — весовая функция вектора вероятностей $c(\bar{p})$; обычно полагают $c(\bar{p}) = 1$.

Отметим, что в общем случае получить байесовскую функцию оценки для произвольной функции потерь, как правило, нелегко. Однако общепринято, что наиболее адекватные результаты получаются при использовании именно квадратичной функции потерь. Тот же результат — математическое ожидание апостериорной плотности вероятности искомого параметра — получается для широкого класса апостериорных распределений и при использовании любой другой выпуклой симметричной функции потерь [36][16].

[16]Единственное существенное возражение против применения квадратичной функции потерь состоит в том, что она «подчеркивает хвосты» распределений, приписывая слишком большой вес редким, вообще говоря, значениям параметра. Однако для задачи оценки вероятностей это возражение снимается, поскольку область изменения параметра в этом случае конечна.

5.2.1. *Одномерный случай.* Рассмотрим для простоты сначала случай $\nu = 2$, 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. Построим байесовские точечные функции оценки $\hat{p}$ неизвестной вероятности $p^* = 1 - \nu$ ошибочной классификации при различном задании функции потерь.

Формула Байеса в нашем случае имеет вид

$$f(p|m_w m_r) = \frac{f(p)f(m_w m_r|p)}{\int\limits_0^1 f(p)f(m_w m_r|p)dp} \tag{9}$$

Здесь $f(m_w m_r|p) = p^{m_\nu}(1-p)^{m_r}$ — правдоподобие. В качестве априорного распределения мы будем использовать бетта-распределение ($\boldsymbol{B}$) $Be(a,b)$ с параметрами $a > 0$, $b > 0$, плотность которого равна

$$f(p) = f(p|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1},\ p \in (0,1)$$

$\boldsymbol{B}$-распределение очень удобно для наших целей, поскольку в этом случае вычисления апостериорного распределения наиболее просто. С другой стороны, формы кривых плотностей $Be(a,b)$ при различных $a > 0$, $b > 0$, как известно, весьма разнообразны. Заметим здесь, что математическое ожидание и дисперсия $\boldsymbol{B}$-распределения равны

$$\mu_\beta = \frac{a}{a+b},\ \sigma_\beta^2 = \frac{ab}{(a+b)^2(a+b+1)}.$$

С учётом сделанного выбора плотность вероятности апостериорного распределения будет

$$f(p|m_w, m_r) = \frac{\Gamma(d_1+d_2+m)}{\Gamma(m_w+d_1)\Gamma(m_r+d_2)} p^{m_w+d_1-1}(1-p)^{m_r+d_2-1},\ p \in (0,1), \tag{10}$$

т.е. $be(m_w + d_1, m_r + d_2)$.

Укажем, что для вычисления знаменателя (9) и подобных выражений используют формулу Лиувилля [13]:

$$\int\limits_{S_{\nu-1}(\overline{x})} \prod_{i=1}^{n} x_i^{m_i} dx_1 \ldots dx_n = \frac{m_1! \ldots m_n!}{\left(\sum\limits_{i=1}^{n} m_i + n - 1\right)!},$$

где $m_1, m_2, \ldots, m_n$ -натуральные числа.

При $\nu = 2$ и учетом $p_1 + p_2 = 1, p = p$ сформулированная в начале п. 5.2 задача минимизации принимает вид

$$\int_0^1 W(p,q) f(p|m_w, m_r) dp = R(q) \to \min, \quad q \in S_{\nu-1}(x).$$

Как указывалось выше, при квадратичной

$$W_1(p,q) = (p-q)^2$$

функции потерь байесовская оценка совпадает с математическим ожиданием апостериорного распределения. Математическое ожидание μ апостериорного распределения (10) есть

$$\mu = \frac{m_w + a}{m + a + b}.$$

Полученная оценка может рассматриваться как модификация МП-оценки с учётом априорной информации относительно p^* или как модификация априорной оценки $\mathrm{a}/(\mathrm{a}+b)$ с учётом наблюдённых величин m_w, и m_r.

При отсутствии какой-либо информации о значениях вероятности $p(\gamma_i = 1,\ i = \overline{1,m})$ по принципу неопределённости Лапласа полагаем, что априорная вероятность имеет равномерное на $(0,1)$ распределение. Равномерное распределение — это **B**-распределение с параметрами $d_1 = d_2 = 1$. Тогда получаем апостериорную плотность в виде

$$f(p|m_r,\, m_w) = \frac{\Gamma(m+2)}{\Gamma(m_w+1)\Gamma(m_r+1)} p^{m_w}(1-p)^{m_r},$$

т.е. плотность **B**-распределения $Be(m_r + l,\, m_w + 1)$ у которого $\mu = (m_w+1)(m+2)$.

Таким образом получена точечная функция оценка $\widehat{p}w_1 = \widehat{p}w$ вероятности ошибки распознавания $1 - \nu$:

$$\hat{p}w = \frac{m_w + 1}{m + 2}. \tag{11}$$

Найдем теперь функцию оценки $\widehat{p}w_2 = \widehat{p}w$ при нормированной функции потерь W_2. Имеем:

$$R(q) = \int_0^1 \frac{(p-q)^2 (m+1)!}{p(1-p) m_r! m_w!} p^{m_w}(1-p)^{m_r} dp =$$

$$= \int_0^1 (p-q)^2 \frac{(m-1)!1m(m+1)}{(m_r-1)!1m_r(m_w-1)!m_w} p^{m_w-1}(1-p)^{m_r-1} dp =$$

$$= \frac{m(m+1)}{m_r m_w} \int_0^1 (p-q)^2 \frac{(m-1)!}{(m_r-1)!(m_w-1)!} p^{m_w-1}(1-p)^{m_r-1} dp =$$

$$= \frac{m(m+1)}{m_r m_w} \int_0^1 (p-q)^2 f(p|\, m_w - 1,\, m_r - 1)dp$$

Минимум значения интеграла в последнем выражении будет достигаться при $q = \widehat{p}w_2 = m_w/m$, и мы снова получаем оценку максимального правдоподобия.

Возвратимся к оценке (11). Ясно, что она является смещенной: если $\widehat{p}$ — МП-оценка, то

$$\widehat{p}w = \frac{m}{m+2}\widehat{p} + \frac{1}{m+2},$$

и с учетом свойств $\widehat{\mathrm{p}}$, приведённых в предыдущем разделе 5.1,

$$M\{\widehat{p}w\} = M\left\{\frac{m}{m+2}\widehat{p} + \frac{1}{m+2}\right\} = \frac{mp^* + 1}{m+2} \neq p^*.$$

Также ясно, что оценка $\widehat{p}w$ не смещена асимптотически.

Дисперсия $\mathbf{D}\{\widehat{p}w\}$ полученной оценки равна

$$\mathbf{D}\{\widehat{p}w\} = \mathbf{D}\left\{\frac{m}{m+2}\widehat{p} + \frac{1}{m+2}\right\} = \left(\frac{m}{m+2}^2\right)\mathbf{D}\{\widehat{p}\} = \frac{mp^*(1-p^*)}{(m+2)^2},$$

и оценка, очевидно, состоятельна.

Легко видеть, что несмещённая оценка $\overline{\{D\widehat{p}w\}}$ дисперсии полученной оценки 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. Естественно, тот же результат получится, если сразу воспользоваться формулой для нижней границы смещённой оценки [33] [17]. Ясно, что выигрыш в дисперсии оценки будет особенно существенным при малых выборках. Следует, однако, иметь в виду, что для смещённой оценки дисперсия служит мерой близости не к оцениваемому параметру, а к математическому ожиданию оценки. Поэтому важное значение приобретает вопрос об «истинном» виде распределения вероятности p.

[17] Дисперсии смещённых D_d и несмещённых D оценок параметра p связаны формулой

$$D_d = (1 + b'_m(p))^2 D,$$

где $b_m(\mathrm{p})$ — смещение. В нашем случае

$$\widehat{p} = p + \frac{1-2p}{m+2}, \quad b_m(p) = \frac{1-2p}{m+2}, \quad b'_m(p) = -2\frac{p}{m+2}, \quad (1+b'_m(p))^2 = \left(\frac{m}{m+2}\right)^2.$$

5.2.2. *Обсуждение полученных оценок. Другие точечные 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 несмещённость или несмещённость по моде. В нашем случае мы столкнулись с ситуацией, когда смещённая оценка имеет дисперсию меньше, чем несмещённая. Мы считаем это достаточным основанием для того, чтобы отказаться от рассмотрения лишь класса несмещённых оценок. Кроме того, обоснованность использования байесовских оценок подтверждается и проведённым стохастическим моделированием (см. ниже).

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 $x_{(\beta)1/2}$ 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. Легко показать, что она будет являться байесовской с функцией штрафа $W_3(p,\, q) = |p - q|$.

В [18] для малых p^* предлагается в качестве априорного распределения брать $Be(1,\, b)$ с большим b. Тогда байесовской функцией оценки будет

$$\hat{p} = \frac{m_w + 1}{m + b + 1}.$$

Для нашей задачи можно попытаться использовать т.н. W - *минимаксную* оценку, при которой максимальные потери для некоторой выбранной функции потерь W минимальны по $p^* \in (0, 1)$. Понятие W -минимаксности вводится независимо от задания какого-либо априорного распределения и поэтому, вообще говоря, может рассматриваться в рамках частотного подхода. Иногда оказывается возможным подобрать априорное распределение, при котором полученная минимаксная оценка оказывается также равной и соответствующей байесовской. Такое априорное распределение называют *наименее благоприятным.*

Если выбрать функцию потерь квадратичной ($W = W_1$), то минимаксная оценка параметра p биномиального распределения будет иметь вид

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

Представляется, однако, что использование полученной функции 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.

Если же выбрать нормированную квадратичную функцию потерь ($W = W_1$), то W_2-минимаксными оценками искомых вероятностей будут являться относительные частоты. При этом наименее благоприятном распределением оказывается равномерное. Неприемлемость же точечных оценок в виде относительных частот для случая малых выборок обсуждалась выше.

Для выяснения вопроса: *«Какая из возможных точечных оценок наиболее адекватна реальным практическим ситуациям?»* был проведен численный эксперимент. Для разных значений $p \in [0, 1]$ появления условного события A генерировались выборки объёмам $n = 1, 2, \ldots, 20$ и фиксировалось количество r наблюдённых событий. Затем вычислялось наиболее вероятное (среднее) значение p, для которой при данном n наблюдается r появлений события A, т.е. определялась стохастическая оценка $\check{p}$ вероятности p(A) появления события A. Она сравнивалась с

$$\text{МП}_{\hat{p}_{ML}} = \frac{r}{n}$$

и байесовской

$$\hat{p}_B = \frac{r+1}{n+2}$$

оценками по формуле

$$\check{p} = \lambda \cdot \hat{p}_{ML} + (1 - \lambda) \cdot \hat{p}_B$$

(для четных n и $r = \frac{n}{2}$ указанные оценки совпадают и значение λ не определено).

В результате оказалось, что полученные стохастические оценки, как правило, очень близки к соответствующим байесовским ($\lambda \approx 0$). Наибольшие относительные отклонения значений λ наблюдались в области

$$r = \frac{n}{2} \pm \frac{1}{2}$$

для нечётных n, где рассматриваемые оценки мало различаются и величина λ плохо обусловлена. В интересующей нас области малых n и r значения стохастической и байесовской оценок совпадали с большой точностью (для приблизительно 10000 наблюдений значений r при данном n величина λ составляла порядка нескольких процентов). Таким образом, целесообразность использования байесовские оценок, особенно в случае малых выборок, можно считать подтвержденным стохастическим моделированием[18].

[18]Программа стохастического моделирования написана А. Лапшиным в среде Delphi 5.0 для ПК. Для генерации случайной величины r имеющей биномиальное распределение использовался метод

5.2.3. *Многомерный случай.* Пусть теперь $\nu > 2$. В многомерном случае формула Байеса имеет вид

$$f(\bar{p}|m_1,\, m_2,\, \ldots m_\nu) = \frac{f(\bar{p})f(m_1,\, m_2,\, \ldots\, m_\nu|\bar{p})}{\int\limits_{S_{\nu-1}(\bar{p})} f(\bar{p})f(m_1,\, m_2,\, \ldots\, m_\nu|\bar{p})d\bar{p}}. \tag{12}$$

Здесь

$$f(m_1, m_2, \ldots m_\nu|\bar{p}) = \prod_{k=1}^{\nu} p_k^{m_k}$$

является функцией правдоподобия и, естественно, выполняется условие нормировки (5). Как отмечалось в п. 3, искомые вероятности $\bar{p} = \{p_k\}_{k=1}^{\nu}$ подчиняются полиномиальному распределению (6).

В качестве априорного распределения мы будем использовать $(\nu-1)$-мерное распределение Дирихле $Di = (d_1, d_2, \ldots, d_{\nu-1}, d_\nu)$ с параметрами $d_1, d_2, \ldots, d_{\nu-1}, d_\nu$, имеющее плотность

$$f(\bar{p}|d_1, d_2, \ldots, d_{\nu-1}, d_\nu) = \frac{\Gamma(d_1 + d_2 + \ldots + d_\nu)}{\Gamma(d_1)\Gamma(d_2)\ldots\Gamma(d_\nu)} \prod_{k=1}^{\nu} p_k^{d_k-1} \tag{13}$$

в любой точке симплекса $S_{\nu-1}(\bar{x})$ и равную нулю в других точках $\mathbf{R}^\nu$. Здесь все $(d_1, d_2, \ldots, d_\nu)$ — вещественные положительные числа. При $\nu = 2$ $Di(d_1; d_2)$ сводится к $\mathrm{Be}(a, b)$. С помощью формулы Лиувилля легко установить, что математическое ожидание $(\nu-1)$-мерного распределения Дирихле есть

$$\mu_{Di} = \frac{d_k}{d_1 + \ldots + d_\nu}, \qquad k = \overline{1, \nu}.$$

Из (12) и (13) следует, что плотность вероятности апостериорного распределения есть

$$f(p\,|\,d_1,\, \ldots,\, d_\nu) = \frac{\Gamma(d_1 + \ldots + d_\nu + m)}{\Gamma(d_1 + m_1)\, \ldots\, (d_\nu + m_\nu)} \prod_{k=1}^{\nu} p_k^{d_k+m_k-1},$$

т.е. будет являться плотностью $(\nu-1)$-мерного распределения Дирихле

$$Di\,(m_1 + d_1,\, \ldots,\, m_{\nu-1} + d_{\nu-1},\, m_\nu + d_\nu).$$

Для квадратичной функции потерь $W\,(\bar{p},\, \bar{q}) = ||\,\bar{p} - \bar{q}\,||^2$ байесовскими оценками $\hat{p}_i$ вероятностей p_i^* будут являться компоненты вектора μ_k апостериорного среднего $\bar{\mu} = (\mu_1,\, \ldots,\, \mu_\nu)^T$, равные

$$\hat{p}_k = \mu_k \frac{m_k + d_k}{m + \sum\limits_{j=1}^{\nu} d_j}, \quad k = \overline{1,\, \nu}.$$

«браковки». В программе моделировалось 10000 экспериментов, соответствующих каждому p при данном n. Время счета при этом не превосходило трех минут (процессор Pentium-III).

В условиях отсутствия информации о весах прецедентов принимаем в качестве распределения $\bar{p}$ равномерное. Равномерное распределение есть распределение Дирихле $Di(1, \ldots, 1; 1)$. Получаем отсюда, что апостериорная плотность вероятностей имеет вид

$$f(\bar{p}\,|\,m_1, m_2, \ldots, m_\nu) = \frac{\Gamma(m+\nu)}{\Gamma(m_1+\nu)\ldots\Gamma(m_\nu+1)} \prod_{k=1}^{\nu} p_k^{m_k} = \frac{(m+\nu-1)}{m_1!\ldots m_\nu!}\, p_1^{m_1}\, p_2^{m_2} \ldots p_\nu^{m_\nu},$$

где $\bar{p} \in S_{\nu-1}(\bar{x})$, т.е. является плотностью $(\nu-1)$-мерного распределения Дирихле

$$Di\,(m_1+1; \ldots, m_{\nu-1}+1; m_\nu+1),$$

а байесовскими оценками $\hat{p}_k$ вероятностей p_k^* будут являться величин

$$\hat{p}_k = \mu_k = \frac{m_k+1}{m+\nu}, \quad k = \overline{1,\nu}. \tag{14}$$

Если формально положить $m = 0$ (отсутствие прецедентов) получаем

$$\hat{p}_1 = \hat{p}_2 = \ldots = \hat{p}_{\nu=1/\nu}$$

— принцип неопределенности Лапласа, использованный нами при выводе (14).

Легко показать, что применение нормированной многомерная функция потерь

$$W(\bar{p}, \bar{q}) = ||\bar{p} - \bar{q}||^2 \Big/ \prod_{k=1}^{\nu} p_k$$

приводит к оценкам $\hat{p}_k = m_k/m,\ k = \overline{1,\ \nu}$, совпадающим в этом случае МП-оценками.

Аналогично одномерному случаю, используя свойство воспроизводимости[19] по m полиномиального распределения $M(m; \cdot)$ и свойства распределения Дирихле получим, что компоненты вектора дисперсий оценок (14) суть

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

а их несмещенные оценки —

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

[19]Параметрическая с параметром θ функция распределения $p(u, \theta)$ случайной величины u называется *воспроизводящей по* θ, если для независимых случайных величин u_1 и u_2, которые имеют функции распределения $P(u_1, \theta_1)$ и $P(u_2, \theta_2)$ соответственно, величина $u_1 + u_2$ распределена по $P(u_1 + u_2, \theta_1 + \theta_2)$ (см. [33]). Если в (12) $f(p)$ и $f(m_1, m_2, \ldots, m_\nu|\,p)$ принадлежат к одному типу воспроизводящих плотностей, то и плотность $f(p\,|\,m_1, m_2, \ldots, m_\nu)$ будет относится к тому же типу распределений.

5.2.4. *Случай неравных весов прецедентов.* Перейдем теперь к рассмотрению случая, когда прецедентная информация включает в себя вектор весов $\{\gamma_i = \gamma(x_i)\}_{i=m}^{m} = \bar{\gamma}_m$ (где не все компоненты равны) прецедентов $\bar{x}_m$. Значение γ_i показывает частоту встречаемости прецедента x_i. Часто заказчик, готовя исходные данные для решения задачи распознавания и желая дать более полное и компактное описание пространства образов, намеренно или вынужденно[20] 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» $\gamma_i \geqslant 1$ данного прецедента x_i можно трактовать как задание «дополнительных прецедентов» вблизи 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.

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 образов γ_i через количества квазипрецедентов описывают вероятности появления образов в окрестностях x_i с тем же значением истинного классификатора $f^*(x_i)$. Точнее, мы считаем, что веса γ_i образов x_i линейно и аддитивно связаны с вероятностями появления в процессе классификации на практике новых образов в окрестностях x_i с тем же значением истинного классификатора $f^*(x_i)$, $i = 1, 2, \ldots, m$. Конкретно, мы дополняем Гипотезу 1 нижеследующей Гипотезой 2.

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* X, *однако веса прецедентов* $\{\gamma_1, \gamma_2, \ldots, \gamma_m\}$ *отражают априорную информацию о распределении* $P(X)$.

[20] например, из-за отсутствия соответствующих данных.

Поскольку мы трактуем веса как информацию о количестве квазипрецедентов в окрестности x_i, естественно считать, что $\gamma_i \geqslant 1,\ i = \overline{1,\, m}$, (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.

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

Введём обозначение

$$\sum_{i:\, x_i \in X_k} \gamma_i = \mu_k \tag{15}$$

Понятно, что $\mu_k' = \mu_k - m_k \geqslant 0, \quad k = \overline{1,\, \nu}$, поскольку $m_k = \sum\limits_{i:\, x_i \in X_k} 1$.

В качестве априорного распределения вероятностей на $\{X_k\}_{k=1}^{\nu}$ примем распределение Дирихле $Di\,(\mu_1' + 1,\, \mu_2' + 1,\, \ldots,\, \mu_{\nu-1}' + 1; \mu_\nu' + 1)$.

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

Обозначим

$$M = \sum_{k=1}^{\nu} \mu_k. \tag{16}$$

Используя формулу Байеса (12) и выше приведённые зависимости, получим апостериорное распределение вектора вероятностей

$$\bar{p} = \{p_1,\, p_2,\, \ldots,\, p_\nu\},\ p_k \in (0,\, 1),\ k = \overline{1,\, \nu}:$$

$$f(\bar{p}\,|\, m_1,\, m_2,\, \ldots,\, m_\nu) = \frac{\Gamma\left(m + \nu + \sum\limits_{k=1}^{\nu} \mu_k'\right)}{\prod\limits_{k=1}^{\nu} \Gamma(m_k + \mu_k' + 1)} \prod_{k=1}^{\nu} p_k^{m_k + \mu_k'} =$$

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

которое является плотностью $(\nu - 1)$-мерного распределения Дирихле

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

Байесовской оценкой искомых вероятностей при квадратичной функции потерь будет вектор апостериорного среднего с компонентами

$$\hat{p}_k = \frac{\mu_k + 1}{M + \nu}, \quad k = \overline{1,\, \nu} \tag{17}$$

где μ_k и M вычисляются по (15) и (16) соответственно. Эти значения и предлагается использовать в качестве точечных оценок вероятностей событий $x \in X_k$ в общем случае задачи Z. Легко проверить, что при $\gamma_i = const,\ i = \overline{1,\nu}$, формула (17) превращается в (14).

Ясно также, что в рамках частотного подхода формула (17) примет вид

$$\hat{p}_k \frac{\mu_k}{M},\ k = \overline{1,\nu}.$$

Автор глубоко признателен академику РАИ Ю.И. Журавлёву за понимание и поддержку. Автор также благодарен проф. В.Е. Бенингу за ценные консультации и к.ф.-м.н. К.В Воронцову за предоставленные материалы по теории Вапника-Червоненкиса.

Список литературы

1. **Айзерман М.А., Браверман Э.М., Розоноэр Л.И.** Метод потенциальных функций в теории обучения машин. – М.: Наука, 1970. Математические методы в теории надежности. М.: Наука, – 1965. – 399 с.
2. **Бериков В.Б.** Об устойчивости алгоритмов распознавания в дискретной постановке // Искусственный интеллект. НАН Украины. Ин-т проблем искуст. интелекта. Донецк, 2000, №2. С. 5-8.
3. **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.
4. **Вапник В.Н., Червоненкнс А.Я.** Теория распознавания образов. Стохастические проблемы обучения. – М.: Наука, 1974.
5. **Вапник В.Н.** Воссгановление зависимостей по эмпирическим данным. – М.. Наука, 1979.
6. **Гасканов Д.В., Шаповалов В.И.** Малая выборка. – М.: Статистика, 1978.
7. **Гуров С.И.** Оценки вероятности ошибок классификации при малом числе прецедентов II Интеллектуализация обработки информации. Международная научная конференция ИОИ'2000. Тезисы докладов (Алушта, 10-14 июня 2000 г.). Симферополь: КрНЦ НАНУ, ТНУ, 2000. С. 26.
8. **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 zranch of Moscow State University, Crimean Scientific Centre, Crimean Academy of Sciences, Crimean Mathematical Foundation, 2002 (в печати).
9. **Гуров С.И.** Точечные оценки ошибок распознавания // «Математические мегоды распознавания образов» (ММРО-10). Доклады X Всероссийской конференции. – М.: Российская академия наук, Вычислительный центр, 2001, С. 34-37.
10. **Журавлев Ю.И.** Корректные алгебры над множеством некорректных (эвристических) алгоритмов. I, П, Ш. // Кибернетика, I: №4, 1977, С. 5-17; II: №6, 1977, С. 21-27; Ш: №2, 1978, С. 35-43.
11. **Журавлев Ю.И.** Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. Сб. статей. – М.: Наука, Вып.33, 1978. С. 5-69.
12. **Закс Л.** Статистическое оценивание /Пер. с нем. под ред. Ю.П. Адлера, В.Г. Горского. – М.: Статистика, 1976. – 560 с.

13. **Интегралы** и ряды. Элементарные функции /Прудников А.И., Брычков Ю.А., Маричев О.И. – М.: Наука, 1981. – 800 с.
14. **Кендал М., Стюарт А.** Теория распределений / Пер. с англ. – М.: Наука, 1966. – 588 с.
15. **Кендал М., Стюарт А.** Статистические выводы и связи / Пер. с англ. – М.: Наука, 1973. – 900 с.
16. **Кольцов П.П.** Математические модели теории распознавания образов // Компьютер и задачи выбора. М.: Наука, 1989. С. 89-119. – 200 с.
17. **Кремер Н.Ш.** Теория вероятностей и математическая статистика. – М.: ЮНИТИ-ДАНА, 2000. – 543 с.
18. **Неман Э.** Теория точечного оценивания / Пер. с англ. – М.. Наука, Гл. ред. физ.-мат. лит., 1991. – 448 с.
19. **Леман Э.** Проверка статистических гипотез. – М.: Наука, 1979. – 408 с.
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 решающие функции и вопросы статистической устойчивости решений. – Новосибирск: Изд-во Ин-та математики, 1999. – 212 с.
22. **Методы** статистического анализа и обработка малого числа наблюдений при контроле качества и надежности приборов и машин. – Л., 1974. – 92 с.
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: непараметрический подход / А.В. Лапко, С.В. Ченцов, С.И. Крохов, Л.А. Фельдман. – Новосибирск: Наука. Сибирская издательская фирма РАН, 1996. – 296 с.
26. **Патрик Э.** Основы теории распознавания образов / Пер. с англ. Под. ред. Б.Р. Левина. – М.: Сов. радио, 1980. – 408 с.
27. **Пытев Ю.П.** Возможность. Элементы теории и применения. – М.: Эдиториал УРСС, 2000. – 192 с.
28. **Рао С.Р.** Линейные статистические методы и их применение / Пер. с англ. – М: Накуа, 1968. – 270-272 (несмс. оценки) 548 с.
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 И.В., Дуник-Барковский И.В.** Курс теории вероятностей и математической стстистики для технических приложений. – М.: Наука, 1965. – 512 с.
31. **Справочник** по теории вероятностей и математической стстистике / В.С. Королюк, Н.И. Портенко, А.В. Скороход, А.Ф. Турбин. – М.: Наука, 1985. – 640 с.
32. **Ту Дж., Гонсалес Р.** Принципы распознавания образов – М.: Мир, 1978. – 416 с.
33. **Уилкс С.** Математическая статистика / Пер. с англ. – М.: Наука, 1967. – 632 с.
34. **Фомин В.Н.** Математическая теория обучаемых опознающих систем. – Л: Изд-во Ленингр. ун-та, 1976. – 236 с.

35. **Фу К.** Последовательные методы в распознавании образов и обучении машин / Пер. с англ. – М.: Наука, 1971.– 256 с.
36. **Фукунага К.** Введение в статистическую теорию распознавания образов / Пер. с англ. – М.: Наука, Гл. ред. физ.-мат. лит., 1979. – 368 с.
37. **Фурсов В.А.** Идентификация моделей систем формирования изображений по малому числу наблюдений. – Самара: Самар. гос. аэрокосм, ун-т., 1998. – 218 с.
38. **Шор Я.Б.**Статистические выводы анализа и контроля надежности и качества. – М.: Сов. радио, 1962. – 552 с.
39. **Box G.E., Tiao G.C.** Bayesian Inference in Statistical Analysis. – Mass.: Addisson-Wesley, Reading, 1973.
40. **lloeffding W.** Probability inequalities for sums of founded random variables // J. Amer. Statist. Assoc., 1963, Vol.58. Pp. 13-30.
41. **Jeffreys H.** The Theory of Probability. – Oxford: Oxford University Press, 1961.
42. **Parzen E.** On estimation of a probability density function and mode // Annals of Math. Statist., 1962, v. 33, №3.
43. **Rozenblatt M.** Remarks of some non-parametric estimates of a density function // Annals of Math. Statist., 1956, v. 27, №3.