УДК 004.93'1: 519.157

М.И. Шлезингер, А.В. Бондаренко

Как формулировать задачи обучения в распознавании образов

Исследованы задачи распознавания образов в ситуации, когда статистическая модель распознаваемого объекта известна лишь частично. Выполнен критический анализ минимаксного подхода к решению таких задач и подхода, основанного на максимально правдоподобном оценивании модели по обучающей выборке. Сформулирована постановка задачи, покрывающая весь спектр ситуаций для обучающих выборок любого объема, от нулевого до бесконечного. Выполнен формальный анализ задач обучения в этой новой постановке и показано ее решение в некоторых простейших случаях.

Pattern recognition problems are considered for a case when a statistical model of an object is not completely known. A minimax approach to solution of such problems is critically analyzed as well as an approach based on the maximal likelihood model estimation with respect to given training multiset. The suggested formulation of the recognition learning problem embraces a whole spectrum of situations for training sets of an arbitrary size: from zero to infinite ones. Main formal properties of the suggested problem formulation are analyzed and its solution in several simplest cases is shown.

Досліджено задачі розпізнавання образів у ситуації, коли статистична модель розпізнаваного об'єкта відома лише частково. Виконано критичний аналіз мінімаксного підходу до розв'язання таких задач та підходу, заснованого на максимально правдоподібному оцінюванні моделі за навчальною вибіркою. Сформульовано постановку задачі, яка покриває весь спектр ситуацій для навчальних вибірок будь-якого об'єму, від нульового до безкінечного. Виконано формальний аналіз задач навчання у цій новій постановці та показано її розв'язання у деяких найпростіших випадках.

Введение. В работе исследуется задача распознавания образов в обычной для практики ситуации, когда статистическая модель распознаваемого объекта известна не полностью, а известен лишь класс моделей, которому она принадлежит. Известно, например [1], что задача распознавания допускает разумную, так называемую минимаксную формулировку и в такой, не полностью определенной статистической ситуации. Задача в этой формулировке состоит в построении стратегии, обеспечивающей удовлетворительное качество распознавания при любой модели из заданного класса.

В случае, когда недостающие знания о модели объекта частично восполняются так называемой обучающей выборкой, возникает вопрос, как эти дополнительные сведения использовать для улучшения качества распознавания. Совокупность подходов к решению этого вопроса составляет современную теорию обучения в распознавании образов. Большая часть этих подходов состоит в разделении процесса распознавания на два этапа. Сначала находят модель из заданного класса, которая в определенном смысле наилучшим образом согласуется с обучающей выборкой. Затем с этой наилучшей моделью поступают так, будто она совпадает с действительной. Это отождествление правомерно лишь для обучающих выборок достаточно большой длины. В противном случае возникает масса вопросов, известных как проблема коротких выборок.

Указанные подходы охватывают далеко не всю проблему, а лишь два ее крайних частных случая: в отсутствие обучающей выборки (минимаксный подход) и при неограниченно большой ее длине (существующие методы обучения). Дадим такую формулировку задачи обучения в распознавании, которая охватывает весь спектр ситуаций для обучающих выборок любой длины: от нулевой до бесконечной.

Формулировка основных понятий

Будем говорить об объекте, который характеризуется двумя случайными параметрами x и k. Эти параметры принимают значения из двух конечных множеств, X и K соответственно. Параметр x будем называть *наблюдаемым признаком* объекта, или *наблюдением*, а параметр k – его *скрытым*, непосредственно *не наблюдаемым состоянием*. Также будем говорить о

конечном множестве M моделей и об $m \in M$, как о некоторой модели из этого множества. Пусть функция $p : X \times K \times M \to R$ такова, что ее значение $p(x, k; m)$, $x \in X$, $k \in K$, $m \in M$, есть совместная вероятность наблюдения x и состояния k в модели m.

Множество $\{(x,k) | x \in X,\ k \in K\}$ обозначим Λ, а через Λ^n обозначим множество обучающих выборок длины n, $n = 1, \ldots$. Одна отдельная выборка $L^n \in \Lambda^n$ имеет вид

$$L^n = (x_1, k_1; x_2, k_2; \ldots, x_n, k_n).$$

Введем также обозначение L^0 для обучающей выборки нулевой длины и Λ^0 для непустого множества, содержащего единственную выборку L^0.

Примем, что на множестве Λ^n задано распределение вероятностей, зависящее от модели m, так что

$$p(L^n; m) = \prod_{i=1}^{n} p(x_i, k_i; m),\ L^n \in \Lambda^n,\ m \in M.$$

При этом считается, что $p(L^0; m) = 1$.

Функцию $q : X \times K \to R$ будем называть *стратегией распознавания*. Ее значение $q(k/x)$ понимается как условная вероятность принятия решения, что объект находится в состоянии k, при условии, что на вход стратегии подано наблюдение x. На данной стадии изложения введение такой рандомизированной стратегии распознавания может показаться ненужным усложнением в сравнении с детерминированной стратегией вида $X \to K$. Однако в дальнейшем определение стратегии именно в рандомизированном виде существенно упростит доказательство некоторых утверждений.

Множество $R^{X \times K}$ всех возможных стратегий вида $q : X \times K \to R$ обозначим Q. При этом, естественно, предполагается, что Q содержит только те функции q, для которых

$$\sum_{k \in K} q(k/x) = 1,\ x \in X,\ q(k/x) \geq 0,\ x \in X,\ k \in K.$$

Под стратегией обучения, или алгоритмом обучения, будем понимать функцию, которая для каждой обучающей выборки L^n указывает стратегию распознавания из множества Q. Обозначим $q^n : \Lambda^n \to Q$ стратегию обучения по обучающей выборке длины n. Для каждой обучающей выборки $L^n \in \Lambda^n$ она определит стратегию $q^n(L^n) : X \times K \to R$ распознавания. Эта стратегия распознавания, примененная к некоторому конкретному значению x признака, выдает случайное решение k о состоянии объекта с вероятностью $q^n(L^n)(k/x)$. Для каждой стратегии $q^n : \Lambda^n \to Q$ определим рандомизированную стратегию $\tilde{q}^n : L^n \times X \times K \to R$, так что

$$\tilde{q}^n(k/x, L^n) = q^n(L^n)(k/x),$$

и назовем ее *стратегией обучения-распознавания*. Стратегии $\tilde{q}^n$ и q^n – формально различные функции, так как первая имеет формат $\Lambda^n \times X \times K \to R$, а вторая – $\Lambda^n \to Q$. Однако их содержательный смысл один и тот же. Стратегия обучения $q^n : \Lambda^n \to Q$ указывает, как стратегия распознавания зависит от обучающей выборки, а стратегия обучения-распознавания $\tilde{q}^n : L^n \times X \times K \to R$ указывает, как решение о состоянии k объекта по признаку x должно зависеть от обучающей выборки L^n. Поэтому в дальнейшем будем использовать как понятие стратегии обучения q^n, так и равноценное ему понятие стратегии $\tilde{q}^n$ обучения-распознавания.

Пусть $w : K \times K \to R$ – функция потерь со значениями $w(k, k')$, обозначающими потери в случае, когда об объекте, находящемся в состоянии k, принимается решение, что он находится в состоянии k'. Риском стратегии $q : X \times K \to R$ на модели m называется математическое ожидание штрафа, т.е. число

$$R(q, m) = \sum_{\substack{k \in K \\ x \in X}} \sum_{k' \in K} q(k'/x) \cdot p(x, k; m) \cdot w(k, k').$$

Для каждой модели $m \in M$ определим риск $R^B(m) = \min_{q \in Q} R(q, m)$, который назовем байесовским риском для модели m, и стратегию

$q^B(m) = \arg\min_{q \in Q} R(q,m)$, которую назовем байесовской стратегией для модели m. Величина $R\left(q^n\left(L^n\right),m\right)$ – это риск, который на модели m достигается распознающей системой после ее обучения на обучающей выборке L^n. В силу случайного характера выборки L^n риск $R\left(q^n\left(L^n\right),m\right)$ также будет случайным. Его математическое ожидание по всем выборкам $L^n \in \Lambda^n$ обозначим

$$R^n\left(q^n,m\right) = \sum_{L^n \in \Lambda^n} p\left(L^n;m\right) \cdot R\left(q^n\left(L^n\right),m\right) =$$

$$= \sum_{L^n \in \Lambda^n} \prod_{i=1}^{n} p\left(x_i,k_i;m\right) \cdot R\left(q^n\left(L^n\right),m\right). \qquad (1)$$

Задача обучения в распознавании образов в ее неформальном понимании состоит в том, чтобы для заданного числа n и заданных функций $p: X \times K \times M \to R$ и $w: K \times K \to R$ построить удовлетворительную в определенном смысле стратегию q^n обучения или, что то же самое, стратегию $\tilde{q}^n$ обучения-распознавания. Однако качество искомой стратегии определяется не одним числом, а совокупностью чисел $R^n(q^n, m)$, $m \in M$.

Современные исследования проблемы обучения состоят совсем не в конкретизации приведенной формулировки и получении процедур обучения-распознавания как решения конкретно сформулированной задачи, а в фиксировании некоторых общепризнанных процедур обучения и последующем анализе ситуаций, когда они оказываются особенно плохими. Укажем наиболее известные такие процедуры.

Известные подходы к распознаванию при не полностью определенной модели распознаваемого объекта

Минимаксная стратегия

Потребуем, чтобы стратегия распознавания q обеспечивала удовлетворительное качество распознавания для любой модели m из заданного класса M, т.е. чтобы

$$\left\{R(q,m) \le c,\ m \in M\right. . \qquad (2)$$

При малых значениях величины c требования (2) могут оказаться противоречивыми. Поэтому сформулируем задачу отыскания минимального значения c, при котором требования (2) еще выполнимы. Пусть это минимальное значение равно c^*. Тогда стратегию q^*, удовлетворяющую требованиям $\left\{R\left(q^*,m\right) \le c^*,\ m \in M\right.$ назовем минимаксной стратегией распознавания на множестве M моделей. Поскольку задача построения минимаксной стратегии сформулирована для случая отсутствия обучающей выборки, для некоторых классов моделей гарантированный уровень c^* может оказаться слишком плохим. Само по себе это не является недостатком минимаксных стратегий, потому что для некоторых моделей даже байесовский риск может оказаться неприемлемо большим. Существенный изъян минимаксного подхода состоит в том, что минимаксные стратегии распознавания без обучения не допускают разумного обобщения на минимаксные стратегии обучения. А именно, наилучший в минимаксном смысле алгоритм обучения – это алгоритм, который обучающую выборку игнорирует. Далее подробно рассмотрим этот существенный недостаток минимаксного подхода.

Наиболее правдоподобное оценивание модели

Пусть для каждой обучающей выборки $L^n = \left(x_1,k_1;x_2,k_2;\ldots,x_n,k_n\right)$ определена наиболее правдоподобная модель

$$m^*\left(L^n\right) = \arg\max_{m \in M} \prod_{i=1}^{n} p\left(x_i,k_i;m\right),$$

а для каждой модели $m \in M$ определена байесовская стратегия $q^B(m) = \arg\min_{q \in Q} R(q,m)$.

Стратегия обучения $q^n: \Lambda^n \to Q$ определена так, что

$$q^n\left(L^n\right) = q^B\left(m^*\left(L^n\right)\right),\ L^n \in \Lambda^n. \qquad (3)$$

Стратегии такого вида глубоко укоренились в теории и практике распознавания. Именно в силу того, что они хорошо исследованы, об-

щеизвестны их недостатки, обозначаемые кратко, как проблема коротких выборок. Покажем их на простом примере в следующем разделе. Несмотря на эти известные недостатки, стратегия обучения (3) показала свою плодотворность в таких практических задачах, в которых возможны достаточно длинные обучающие выборки. Следует отметить также, что для многих классов моделей как отыскание наиболее правдоподобной модели $m^*(L^n)$, так и построение байесовской стратегии $q^B(m)$ оказывается далеко не тривиальной задачей, представляющей самостоятельный интерес.

Минимизация эмпирического риска

Пусть $Q(M) = \{q^B(m) \mid m \in M\} \subset Q$ – множество байесовских стратегий для моделей из класса M, а $R^э\left(q, L^n\right) = \sum_{i=1}^{n} \sum_{k \in K} q\left(k / x_i\right) \cdot W\left(k_i, k\right)$ – так называемый эмпирический риск стратегии $q : X \times K \to R$ на обучающей выборке $L^n = \left(x_1, k_1; x_2, k_2; \ldots, x_n, k_n\right)$. Стратегия $q^n : \Lambda^n \to Q$ определена как

$$q^n\left(L^n\right) = \arg \min_{q \in Q(M)} R^э\left(q, L^n\right),\ L^n \in \Lambda^n .$$

Наряду с методом наибольшего правдоподобия минимизация эмпирического риска является одним из наиболее популярных методов обучения. Критический анализ этого подхода к обучению в распознавании выполнен в [1].

Иллюстрация введенных понятий и известных подходов

Пусть X – множество вещественных чисел, $K = \{1, 2\}$ а $p(x/k)$, $k = 1, 2$, – распределение плотности условной вероятности, определенное как

$$p\left(x / k\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(x-\mu_k\right)^2},\ k = 1,2,\ \mu_1 = 1,\ \mu_2 = -1 .$$

Пусть p_k, $k = 1, 2$, – априорная вероятность состояния k. Вероятности p_1 и p_2 неизвестны, и пара (p_1, p_2) – неизвестная модель m объекта, так что $p\left(x, k; m\right) = p_k \cdot p\left(x / k\right)$.

Таким образом, в данном примере статистическая модель распознаваемого объекта почти полностью известна. Неизвестны только априорные вероятности пребывания объекта в первом или втором состояниях. Этот специально упрощенный пример предназначен для предельно ясной иллюстрации введенных формальных понятий, недостатков метода наибольшего правдоподобия и сущности предлагаемого далее подхода к обучению в распознавании.

Пусть потери равны

$w\left(k, k'\right) = 1$, если $k \neq k'$, $w\left(k, k'\right) = 0$, если $k = k'$.

Байесовская стратегия $q : X \times K \to R$ для модели $m = (p_1, p_2)$ имеет вид

$q\left(k = 1 / x\right) = 1$, если $p_1 \cdot p\left(x / 1\right) > p_2 \cdot p\left(x / 2\right)$,

$q\left(k = 2 / x\right) = 1$, если $p_1 \cdot p\left(x / 1\right) < p_2 \cdot p\left(x / 2\right)$.

На рис. 1 показана зависимость риска различных стратегий от модели m, в данном случае, от вероятности p_1 первого состояния. Риск $R^B(m)$ байесовской стратегии представлен сплошной линией.

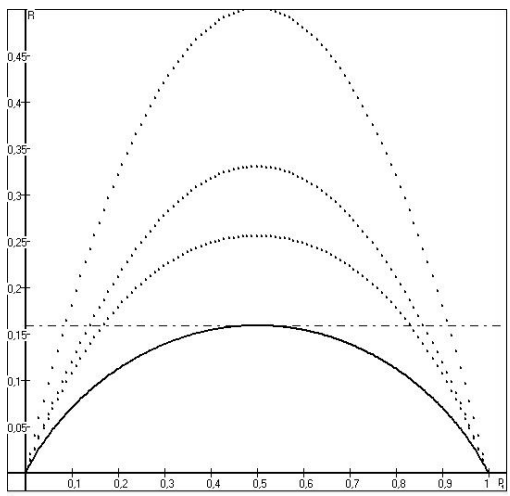

Рис. 1. Зависимость вероятности ошибочного распознавания от априорных вероятностей состояний для различных стратегий

Минимаксная стратегия $q'(x) : X \times K \to R$ имеет вид

$$q'\left(k = 1 / x\right) = 1, \text{ если } x > 0,$$

$$q'\left(k = 2 / x\right) = 1, \text{ если } x < 0 .$$

Риск $R(q', m)$ для этой стратегии показан в виде штрихпунктирной линии.

Пусть L^n – обучающая выборка длины $n > 0$. Наиболее правдоподобная модель $m^*\left(L^n\right)$ или,

что то же самое, наиболее правдоподобные значения p_1^*, p_2^* априорных вероятностей равны $m^*\left(L^n\right)=\left(p_1^*,p_2^*\right)=\left(\frac{n_1}{n},\frac{n_2}{n}\right)$, где n_k, $k=1,2$, – количество элементов *i* в выборке L^n, для которых $k_i=k$. Стратегия $\tilde{q}^n:\Lambda^n\times X\times K\to R$ в этом случае имеет вид

$\tilde{q}^n\left(k=1/L^n,x\right)=1$, если $n_1\cdot p\left(x/1\right)>n_2\cdot p\left(x/2\right)$,

$\tilde{q}^n\left(k=2/L^n,x\right)=1$, если $n_1\cdot p\left(x/1\right)<n_2\cdot p\left(x/2\right)$.

Риск $R^n\left(\tilde{q}^n,m\right)$ этой стратегии зависит от длины *n* выборки и для значений *n*, равных 1, 2 или 3, приведен на рис. 1 в виде пунктирной линии. Очевидно, что при больших значениях длины выборки риск $R^n\left(\tilde{q}^n,m\right)$ становится неотличимым от байесовского риска $R^B(m)$. Однако, как видно из рис. 1, при значениях длины *n*, равных 1, 2 или 3, риск $R^n\left(\tilde{q}^n,m\right)$ заметно отличается от байесовского. Существенно, что риск $R^n\left(\tilde{q}^n,m\right)$ отличается в худшую сторону и от риска минимаксной стратегии, которая обучающую выборку просто игнорирует. Следовательно, метод наибольшего правдоподобия неэффективно использует информацию, содержащуюся в обучающейся выборке, потому что ее использование приводит к результатам, худшим, чем ее игнорирование.

В следующем разделе исследуется вопрос о том, можно ли построить минимаксные стратегии обучения-распознавания так, чтобы использование коротких выборок пусть незначительно, но улучшало бы результаты в сравнении с теми, которые получаются в отсутствие обучающей выборки. На этот вопрос будет получен отрицательный ответ.

Минимаксные стратегии обучения-распознавания

Сформулируем задачу отыскания такого минимального значения *c*, при котором еще не противоречивы требования $\left\{R^n\left(\tilde{q}^n,m\right)\le c,\ m\in M,\right.$ к стратегии $\tilde{q}^n$ обучения-распознавания.

Учитывая, что для $n=0$ стратегия $q^0:L^0\times\times X\times K\to R$ есть стратегия вида $X\times K\to R$ и

$$R^0\left(q^0,m\right)=R\left(q,m\right)=$$
$$=\sum_{\substack{k'\in K\\ x\in X}}q\left(k'/x\right)\sum_{k\in K}p\left(x,k;m\right)\cdot w\left(k,k'\right),$$

запишем эту задачу для $n=0$ в виде следующей пары двойственных задач линейного программирования.

Прямая задача:

$$\min c \tag{4}$$

$$\tau(m)\ \left|\ \sum_{\substack{k'\in K\\ x\in X}}q\left(k'/x\right)\sum_{k\in K}p\left(x,k;m\right)\cdot w\left(k,k'\right)\le c, \quad m\in M; \right. \tag{5}$$

$$t(x)\ \left|\ \begin{array}{l}\sum\limits_{k'\in K}q\left(k'/x\right)=1,\ x\in X;\\ q\left(k'/x\right)\ge 0,\ x\in X;\ k'\in K.\end{array}\right. \tag{6}$$

Двойственная задача

$$\max\sum_{x\in X}t\left(x\right) \tag{7}$$

$$q\left(k'/x\right)\ \left|\ t\left(x\right)\le\sum_{k\in K}\left(\sum_{m\in M}\tau\left(m\right)\cdot p\left(x,k;m\right)\right)\times \right. \tag{8}$$
$$\left.\times w\left(k,k'\right),\quad k'\in K,\quad x\in X;\right.$$

$$c\ \left|\ \sum_{m\in M}\tau\left(m\right)=1;\ \tau\left(m\right)\ge 0,\ x\in X.\right. \tag{9}$$

В приведенных формулах слева от каждого ограничения прямой задачи записана соответствующая ей двойственная переменная; слева от каждого ограничения двойственной задачи записана соответствующая ей переменная в прямой задаче.

Функцию вида $\tau:M\to R$ назовем весовой функцией со значениями $\tau\left(m\right)$, которые назовем весами. Множество весовых функций, удовлетворяющих ограничениям $\sum_{m\in M}\tau(m)=1,\ \tau(m)\ge 0$, $m\in M$, обозначим *T*. Определим модель $m\left(\tau\right)$ так, что совместная вероятность *x* и *k* в модели $m(\tau)$определяется выражением

$$p\left(x,k;m\left(\tau\right)\right)=\sum_{m\in M}\tau\left(m\right)\cdot p\left(x,k;m\right). \tag{10}$$

Модели этого вида назовем *смешанными* моделями в отличие от моделей из первоначального класса M, которые будем называть *чистыми*. Ранее введенные стратегии $q^B(m)$, $m \in M$, будем называть чистыми в отличие от стратегий $\arg\min_{q\in Q} R\left(q, m(\tau)\right)$, $\tau \in T$, которые будем называть смешанными.

Лемма 1. Пусть q^* и c^* – решение прямой задачи (4). В таком случае существует такой набор $\tau^*(m)$, $m \in M$, весов моделей, что

$$\max_{m\in M} R\left(q^*, m\right) = R^B\left(m\left(\tau^*\right)\right). \qquad (11)$$

Доказательство. Пусть $t^*(x)$ и $\tau^*(m)$ – решение двойственной задачи (7). В силу известной теоремы двойственности [4]

$$\sum_{x\in X} t^*(x) = c^*. \qquad (12)$$

Выражение в левой части (5) есть риск $R\left(q^*, m\right)$, а c^* – минимальное число, удовлетворяющее условиям $R\left(q^*, m\right) \le c^*$, $m \in M$.

Следовательно,

$$c^* = \max_{m\in M} R\left(q^*, m\right). \qquad (13)$$

Учитывая, что по определению (10)

$$\sum_{m\in M} \tau(m)\cdot p(x,k;m) = p\left(x,k;m(\tau)\right),$$

запишем ограничение (8) в виде

$$t(x) \le \sum_{k\in K} p\left(x,k;m(\tau)\right)\cdot w(k,k'),\ k' \in K.$$

Число $t^*(x)$ есть максимальное число, удовлетворяющее условиям

$$t^*(x) \le \sum_{k\in K} p\left(x,k;m\left(\tau^*\right)\right)\cdot w(k,k'),\ k' \in K.$$

Следовательно,

$$t^*(x) = \min_{k'\in K} \sum_{k\in K} p\left(x,k;m\left(\tau^*\right)\right)\cdot w(k,k'),$$

что в свою очередь обозначает, что

$$\sum_{x\in X} t^*(x) = R^B\left(m\left(\tau^*\right)\right). \qquad (14)$$

Подставляя (13) и (14) в (12), получаем (11). Таким образом, набор весов $\tau^*(m)$, существование которого требуется доказать, – это веса $\tau^*(m)$, являющиеся решением двойственной задачи (7). **Лемма доказана**.

Лемма 2. Для любой стратегии q^n обучения и любой модели $m \in M$ справедливо неравенство $R^n\left(q^n, m\right) \ge R^B(m)$.

Доказательство. По определению байесовской стратегии неравенство $R\left(q^n\left(L^n\right), m\right) \ge R^B(m)$ справедливо для любой выборки $L^n \in \Lambda^n$. Следовательно, справедливо и неравенство

$$\sum_{L^n\in\Lambda^n} p\left(L^n; m\right)\cdot R\left(q^n\left(L^n\right), m\right) \ge R^B(m).$$

Левая часть этого неравенства есть $R^n(q^n, m)$ по определению (1). **Лемма доказана**.

Теорема 1. Если для класса M моделей решение $q^0 : X\times K \to R$ минимаксной задачи распознавания (4) есть чистая стратегия, то для любого $n = 1, 2, \ldots$ и любой стратегии $q^n : L^n \to Q$ выполняется неравенство

$$\max_{m\in M} R^n\left(q^n, m\right) \ge \max_{m\in M} R\left(q^0, m\right).$$

Доказательство. Пусть $m^0 = \arg\max_{m\in M} R\left(q^0, m\right)$. В таком случае $\max_{m\in M} R^n\left(q^n, m\right) \ge R^n\left(q^n, m^0\right) \ge$

$$\ge R^B\left(m^0\right) = \max_{m\in M} R\left(q^0, m\right).$$

Первое неравенство в этой цепочке очевидно. Второе – справедливо в силу леммы 2. В силу предположения о том, что стратегия q^0 – чистая, и в силу леммы 1 справедливо равенство в третьем звене цепочки. **Теорема доказана**.

Поскольку данная теорема представляет достаточно существенный отрицательный результат, рассмотрим несколько примеров классов моделей, где минимаксная стратегия распознавания оказывается чистой.

1. Минимаксная стратегия оказывается чистой, конечно, в том случае, когда множество смешанных моделей совпадает с множеством чистых моделей. Эта ситуация, в частности, имеет место в приведенном примере, когда модель распознаваемого объекта почти полностью известна, а неизвестны только априорные вероятности состояний.

2. Пусть $K \in \{1, 2\}$ – множество состояний объекта, а x – многомерная гауссова случайная величина с единичной ковариационной матрицей. Условное математическое ожидание этого вектора при условии, что объект находится в состоянии k, есть μ_k. Математические ожидания μ_1 и μ_2 неизвестны. Известны, однако, два выпуклых множества M_1 и M_2, которым они принадлежат. Множество чистых моделей в этом случае есть $M_1 \times M_2$ и оно, конечно, не равно множеству смешанных моделей. Тем не менее минимаксная стратегия распознавания – чистая.

3. Для теории и практики распознавания обычна ситуация, когда функции $p_k : X \to R$, $k \in K$, со значениями $p(x/k)$ неизвестны, но известно лишь, что все они входят в заданный класс P функций, один и тот же для всех $k \in K$. В этом случае минимаксная стратегия распознавания также чистая. Эта наилучшая в минимаксном смысле стратегия оказывается совсем плохой. Она гарантирует лишь, что вероятность ошибочного распознавания не превосходит 50% для любой модели. Однако в силу того, что эта стратегия чистая, вступает в силу вывод доказанной теоремы, что никакая стратегия обучения этот показатель не может улучшить.

Исходя из указанного существенного изъяна минимаксного критерия обучения-распознавания и общеизвестных недостатков максимально правдоподобного оценивания модели, приходим к выводу, что проблема распознавания при не полностью известной статистической модели исследована в весьма незначительной части. Известные и широко применяемые методы охватывают лишь два крайних ее случая: когда имеется обучающая выборка неограниченно большой длины и когда обучающей выборки нет вообще. Желательно сформулировать задачу обучения-распознавания так, чтобы риск распознавания при короткой обучающей выборке оказывался хоть ненамного, но лучше, чем риск распознавания без учета этой выборки. Кроме того, необходимо, чтобы при неограниченном увеличении длины выборки качество распознавания становилось неотличимым от байесовского.

Общий вид стратегий обучения-распознавания

Задача обучения состоит в выборе стратегии $\tilde{q}^n$, оптимальной в определенном смысле, из множества всех возможных стратегий вида $\tilde{q}^n : L^n \times X \times K \to R$. Эта задача допускает множество точных формулировок в зависимости от того, как определено качество стратегии, которое следует оптимизировать. При этом следует учесть, что совокупность $R^n\left(\tilde{q}^n, m\right)$, $m \in M$, рисков позволяет задать на множестве всех возможных стратегий вида $\tilde{q}^n : L^n \times X \times K \to R$ естественное отношение частичной упорядоченности. А именно, стратегия $\tilde{q}_1^n$ не хуже стратегии $\tilde{q}_2^n$ на классе моделей M, если для всех моделей $m \in M$ выполняется неравенство $R^n\left(\tilde{q}_1^n, m\right) \le R^n\left(\tilde{q}_2^n, m\right)$, а стратегия $\tilde{q}_2^n$ во всех отношениях хуже стратегии $\tilde{q}_1^n$, если указанные неравенства выполняются строго. Формулировка задачи обучения должна быть разумной в том смысле, что ее решение не должно оказаться стратегией, которая во всех отношениях хуже некоторой другой стратегии. На основании этих соображений можно априори сузить множество стратегий обучения-распознавания еще до формулировки задачи обучения, требуя лишь, чтобы эта формулировка была разумной в указанном выше смысле.

Для заданной весовой функции τ определим стратегию обучения-распознавания так, что $\tilde{q}^n\left(k' / L^n, x\right) = 1$, если для всех $k'' \ne k'$ выполняется неравенство

$$\sum_{k \in K}\left(\sum_{m \in M} \tau(m) \cdot p\left(L^n; m\right) \cdot p\left(x, k; m\right)\right) \cdot w\left(k, k'\right) <$$

$$< \sum_{k \in K}\left(\sum_{m \in M} \tau(m) \cdot p\left(L^n; m\right) \cdot p\left(x, k; m\right)\right) \cdot w\left(k, k''\right).$$

Стратегию такого вида назовем байесовской для весовой функции τ. Иными словами, при обучающей выборке L^n и наблюдении x байе-

совская стратегия принимает решение в пользу состояния k^*,

$$k^* = \arg\min_{k' \in K} \sum_{k \in K} \left(\sum_{m \in M} \tau(m) \cdot p\left(L^n; m\right) \cdot p\left(x, k; m\right) \right) \times \times w\left(k, k'\right).$$

При фиксированном классе M моделей байесовская стратегия может быть той или иной в зависимости от весовой функции τ. Докажем, что решение любой разумно поставленной задачи обучения – это байесовская стратегия для некоторой весовой функции $\tau : M \to R$.

Теорема 2. Для любой стратегии $\tilde{q}^{*n} : L^n \times X \times K \to R$ обучения-распознавания существует байесовская стратегия $\tilde{q}^n : L^n \times X \times K \to R$, для которой неравенство $R^n\left(\tilde{q}^n, m\right) \le R^n\left(\tilde{q}^{*n}, m\right)$ выполняется для любой модели $m \in M$.

Доказательство. Представим искомую стратегию $\tilde{q}^n$ как решение следующей задачи линейного программирования:

$$\min c$$

$$\tau(m) \left| \begin{array}{l} c - \sum\limits_{L^n \in \Lambda^n} p\left(L^n; m\right) \sum\limits_{x \in X} \sum\limits_{k' \in K} \tilde{q}^n\left(k' / L^n, x\right) \times \\ \times \sum\limits_{k \in K} p\left(x, k; m\right) \cdot w\left(k, k'\right) \ge -R^n\left(\tilde{q}^{*n}, m\right), \quad (15) \\ m \in M; \end{array} \right.$$

$$t\left(L^n, x\right) \left| \begin{array}{l} \sum\limits_{k' \in K} \tilde{q}^n\left(k' / L^n, x\right) = 1, x \in X, L^n \in \Lambda^n, \quad (16) \\ \tilde{q}^n\left(k' / L^n, x\right) \ge 0, k' \in K, L^n \in \Lambda^n, x \in X, \end{array} \right.$$

которой соответствует двойственная задача

$$\max \left[\sum_{x \in X} \sum_{L^n \in \Lambda^n} t\left(L^n, x\right) - \sum_{m \in M} \tau(m) \cdot R^n\left(\tilde{q}^{*n}, m\right) \right]$$

$$\tilde{q}^n\left(k' / L^n, x\right) \left| \begin{array}{l} t\left(L^n, x\right) \le \sum\limits_{k \in K} \left(\sum\limits_{m \in M} \tau(m) \times \right. \\ \left. \times p\left(L^n; m\right) \cdot p\left(x, k; m\right) \right) \times \\ \times w\left(k, k'\right), \; k' \in K, L^n \in \Lambda^n, x \in X; \quad (17) \end{array} \right.$$

$$c \left| \begin{array}{l} \sum\limits_{m \in M} \tau(m) = 1; \; \tau(m) \ge 0, m \in M. \end{array} \right.$$

Как и ранее, слева от уравнений-ограничений задач записаны соответствующие им двойственные переменные.

Система линейных ограничений прямой задачи не противоречива, так как ей удовлетворяют значение $c = 0$ и значения $\tilde{q}^n\left(k' / L^n, x\right) = \tilde{q}^{*n}\left(k' / L^n, x\right)$, $k' \in K$, $L^n \in \Lambda^n$, $x \in X$. Нетрудно также увидеть, что величина c ограничена снизу на множестве решений этой системы. Отсюда, в силу известных теорем двойственности, следует, что ограничения в задаче (17) также непротиворечивы, а целевая функция в этой задаче ограничена сверху.

Пусть $\tilde{q}_0^n : L^n \times K \times X \to R$, c_0 – решение прямой задачи, а $t_0\left(L^n, x\right)$, $L^n \in \Lambda^n$, $x \in X$, и $\tau_0(m)$, $m \in M$ – решение двойственной. Пусть для некоторой тройки $k_0 \in K$, $L_0^n \in \Lambda^n$, $x_0 \in X$ неравенство

$$\sum_{k \in K} \left(\sum_{m \in M} \tau_0(m) \cdot p\left(L_0^n; m\right) \cdot p\left(x_0, k; m\right) \right) \cdot w\left(k, k_0\right) < < \sum_{k \in K} \left(\sum_{m \in M} \tau_0(m) \cdot p\left(L_0^n; m\right) \cdot p\left(x_0, k; m\right) \right) \cdot w\left(k, k'\right)$$

выполняется для всех $k' \ne k_0$. Это значит, что неравенство (17) выполняется строго для данных L_0^n, x_0 и всех $k' \ne k_0$. Отсюда, в силу известной теоремы двойственности, равенство $q_0^n\left(k' / L_0^n, x_0\right) = 0$ выполняется для всех $k' \ne k_0$. В свою очередь, в силу (16) отсюда следует, что $q_0^n\left(k_0 / L_0^n, x_0\right) = 1$.

Таким образом, стратегия q_0^n является байесовской для весовой функции τ_0.

Поскольку значение $c = 0$ достигается на множестве допустимых решений системы ограничений прямой задачи, то $c_0 \le 0$. Отсюда, в силу неравенств (15), следует, что неравенство

$$\sum_{L^n \in \Lambda^n} p\left(L^n; m\right) \sum_{x \in X} \sum_{k' \in K} \tilde{q}_0^n\left(k' / L^n, x\right) \sum_{k \in K} p\left(x, k; m\right) \times \times w\left(k, k'\right) \le R^n\left(q^{*n}, m\right)$$

выполняется для всех $m \in M$. Выражение в левой части этого неравенства есть $R^n\left(\tilde{q}_0^n, m\right)$ и $R^n\left(\tilde{q}_0^n, m\right) \le R^n\left(\tilde{q}^{*n}, m\right),\ m \in M$. **Теорема доказана**.

Таким образом, решение любой разумно поставленной задачи обучения существенно отличается от общепринятых стратегий, указанных ранее. В решении разумно поставленной задачи отсутствует такой этап, как выбор одной-единственной модели на основании обучающей выборки, пусть даже наиболее правдоподобной модели или модели с минимальным эмпирическим риском. В распознавании после обучения учитываются все модели из заданного класса, но с различными весами в зависимости от того, насколько та или иная модель согласуется с обучающей выборкой.

Одна из возможных формулировок задачи обучения-распознавания и ее формальные свойства

Риск $R^n(q^n, m)$ зависит не только от стратегии q^n, но и от модели m. Для некоторых моделей m даже байесовский риск $R^B(m)$ может оказаться неприемлемо большим. Поэтому характеристикой собственно стратегии q^n естественно считать не столько величину $R^n(q^n, m)$, сколько ее отклонение $R^n(q^n, m) - R^B(m)$ от байесовского риска $R^B(m)$, который характеризует собственно модель m, вне ее связи со стратегией q^n. Именно исходя из разности $R^n(q^n, m)$ — $R^B(m)$ исследуются отдельные процедуры обучения в работах [2, 3] и других. Качеством же стратегии q^n безотносительно модели m примем величину $\max\limits_m\left[R^n\left(q^n, m\right) - R^B\left(m\right)\right]$ и сформулируем задачу обучения, как поиск стратегии q^{*n} с наилучшим качеством, т.е.

$$q^{*n} = \arg\min_{q^n} \max_m \left[R^n\left(q^n, m\right) - R^B\left(m\right)\right]. \qquad (18)$$

Теорема 3. Искомая стратегия $\tilde{q}^{*n}$ есть байесовская стратегия для смешанной модели с весовой функцией

$$\tau^* = \arg\max_{\tau \in T}\left[\min_{q^n} R^n\left(q^n, m(\tau)\right) - \sum_{\tau \in T} \tau(m) R^B\left(m\right)\right].$$

Доказательство. Запишем искомую стратегию как решение следующей задачи линейного программирования.

Прямая задача

$$\min c$$

$$\begin{array}{l|l}
\tau(m) & \sum\limits_{L^n \in \Lambda^n} \sum\limits_{x \in X} \sum\limits_{k' \in K} \tilde{q}^n\left(k' / L^n, x\right) \times \\
& \times p\left(L^n; m\right) \sum\limits_{k \in K} p\left(x, k; m\right) \times \\
& \times w\left(k, k'\right) - R^B\left(m\right) \le c,\ m \in M; \qquad (19) \\
t(L^n, x) & \sum\limits_{k' \in K} \tilde{q}^n\left(k' / L^n, x\right) = 1; L^n \in \Lambda^n, x \in X; \\
& \tilde{q}^n\left(k' / L^n, x\right) \ge 0, k' > K, L^n \in \Lambda^n, x \in X.
\end{array}$$

Переменными в этой задаче являются числа $\tilde{q}^n\left(k' / L^n, x\right),\ k' \in K,\ L^n \in \Lambda^n,\ x \in X$, и число c. Задаче соответствует двойственная задача с двойственными переменными $\tau(m),\ m \in M$, и $t\left(L^n, x\right),\ L^n \in \Lambda^n,\ x \in X$.

Двойственная задача

$$\max\left[\sum_{x \in X} \sum_{L^n \in \Lambda^n} t\left(L^n, x\right) - \sum_{m \in M} \tau(m) \cdot R^B\left(m\right)\right]$$

$$\begin{array}{l|l}
\tilde{q}^n\left(k' / L^n, x\right) & t\left(L^n, x\right) \le \sum\limits_{k \in K}\left(\sum\limits_{m \in M} \tau(m) \times\right. \\
& \left. \times p\left(L^n; m\right) \times p\left(x, k; m\right)\right) \times w\left(k, k'\right), \\
& k' \in K, L^n \in \Lambda^n, m \in M; \\
c & \sum\limits_{m \in M} \tau(m) = 1; \tau(m) \ge 0, m \in M.
\end{array}$$

Слева от ограничений прямой (или двойственной) задачи записаны соответствующие им переменные двойственной (или прямой) задачи.

Пусть c^* и q^{*n} – решение прямой задачи, а $t^*\left(L^n, x\right),\ L^n \in \Lambda^n,\ x \in X$ и $\tau^*\left(m\right),\ m \in M$, – решение двойственной. Пусть $L_0^n,\ x_0,\ k_0$ – такие значения, что для любого $k' \ne k_0$ выполняется строгое неравенство

$$\sum_{k\in K}\left(\sum_{m\in M}\tau^*(m)\cdot p\left(L_0^n;m\right)\cdot p\left(x_0,k;m\right)\right)\cdot w(k,k_0)<$$

$$<\sum_{k\in K}\left(\sum_{m\in M}\tau^*(m)\cdot p\left(L_0^n;m\right)\cdot p\left(x_0,k;m\right)\right)\cdot w(k,k').$$

Отсюда немедленно следует, что и неравенство

$$t^*\left(L^{*n},x_0\right)<$$

$$<\sum_{k\in K}\left(\sum_{m\in M}\tau^*(m)\cdot p\left(L_0^n;m\right)\cdot p\left(x_0,k;m\right)\right)\cdot w(k,k')$$

также выполняется строго для всех $k'\neq k_0$. Отсюда, в силу известной теоремы двойственности о дополняющей жесткости, следует, что $\tilde{q}^{*n}\left(k_0/L_0^n,x_0\right)=1$, т.е., что искомая стратегия есть байесовская стратегия для весовой функции τ^*, являющейся решением двойственной задачи. Докажем теперь, что решение τ^* двойственной задачи совпадает с весовой функцией, обеспечивающей максимальное значение числа $\min\limits_{q^n}R^n\left(q^n,m(\tau)\right)-\sum\limits_{m\in M}\tau(m)\cdot R^B(m)$, как это формулируется в теореме.

Риск $R^n\left(q^n,m(\tau)\right)$ есть по определению

$$R^n\left(q^n,m(\tau)\right)=\sum_{k'\in K}\sum_{x\in X}\sum_{L^n\in\Lambda^n}\tilde{q}^n\left(k'/L^n,x\right)\times$$

$$\times\sum_{k\in K}\left(\sum_{m\in M}\tau(m)\cdot p\left(L^n;m\right)\cdot p\left(x,k;m\right)\right)\cdot w(k,k').$$

Поскольку все числа $\tilde{q}^n\left(k'/L^n,x\right)$ не отрицательные, причем для любых $L^n\in\Lambda^n$, $x\in X$ должно выполняться $\sum\limits_{k'\in K}\tilde{q}^n\left(k'/L^n,x\right)=1$, то

$$\min_{q^n}R^n\left(q^n,m(\tau)\right)=$$

$$=\sum_{x\in X}\sum_{L^n\in\Lambda^n}\min_{k'\in K}\sum_{k\in K}\left(\sum_{m\in M}\tau(m)\cdot p\left(L^n;m\right)\cdot p\left(x,k;m\right)\right)\times$$

$$\times w(k,k').$$

Поскольку в двойственной задаче величины $t\left(L^n,x\right)$ должны принимать как можно большие значения, то их можно однозначно определить через весовые коэффициенты $\tau(m)$, так что

$$t\left(L^n,x\right)=$$

$$=\min_{k'}\sum_{k\in K}\left(\sum_{m\in M}\tau(m)\,p\left(L^n;m\right)\cdot p\left(x,k;m\right)\right)\cdot w(k,k')$$

и

$$\sum_{x\in X}\sum_{L^n\in\Lambda^n}t\left(L^n,x\right)-\sum_{m\in M}\tau(m)\cdot R^B(m)=$$

$$=\min_{q^n}R^n\left(q^n,m(\tau)\right)-\sum_{m\in M}\tau(m)\,R^B(m).$$

Теорема доказана.

Поиск оптимальной стратегии обучения-распознавания сводится, таким образом, к отысканию весовой функции $\tau:M\to R$, обеспечивающей максимальное значение числа

$$F(\tau)=\min_{q^n}R^n\left(q^n,m(\tau)\right)-\sum_{m\in M}\tau(m)\,R^B(m). \quad (20)$$

Эта задача – задача выпуклой оптимизации, как утверждает следующая теорема.

Теорема 4. Для любого класса M моделей функция $F(\tau)$ является вогнутой функцией.

Доказательство. Докажем, что для любой весовой функции $\tau_0\in T$ существует линейная функция $L_{\tau_0}:T\to R$, такая, что неравенство

$$L_{\tau_0}(\tau-\tau_0)\geq F(\tau)-F(\tau_0) \quad (21)$$

выполняется для всех $\tau\in T$.

Пусть q_0^n – стратегия, обеспечивающая минимальный риск $R^n\left(q^n,m(\tau_0)\right)$.

$$R^n\left(q_0^n,m(\tau_0)\right)=\min_{q^n}R^n\left(q^n,m(\tau_0)\right). \quad (22)$$

Определим линейную функцию $L_{\tau_0}(\tau)$ как

$$L_{\tau_0}(\tau)=\sum_{m\in M}\tau(m)\left[R^n\left(q_0^n,m\right)-R^B(m)\right]. \quad (23)$$

Справедлива следующая цепочка:

$$L_{\tau_0}(\tau-\tau_0)=\sum_{m\in M}\tau(m)\left[R^n\left(q_0^n,m\right)-R^B(m)\right]-$$

$$-\sum_{m\in M}\tau_0(m)\left[R^n\left(q_0^n,m\right)-R^B(m)\right]\geq$$

$$\geq\min_{q^n}\sum_{m\in M}\tau(m)\left[R^n\left(q^n,m\right)-R^B(m)\right]-$$

$$-\sum_{m\in M}\tau_0(m)\left[R^n\left(q_0^n,m\right)-R^B(m)\right]=$$

$$= \min_{q^n} \sum_{m \in M} \tau(m) \left[R^n\left(q^n, m\right) - R^B\left(m\right) \right] -$$

$$- \min_{q^n} \sum_{m \in M} \tau_0(m) \left[R^n\left(q^n, m\right) - R^B\left(m\right) \right] =$$

$$= F(\tau) - F(\tau_0).$$

Равенство в первом звене следует из (23). Неравенство во втором звене очевидно. Равенство в третьем звене следует из (22). Наконец, последнее равенство следует из определения (20). Таким образом, линейная функция (23) есть искомая линейная функция, доказывающая вогнутость функции $F(\tau)$. **Теорема доказана**.

В приведенном доказательстве получен вспомогательный результат: обобщенный градиент g_{τ_0} функции $F(\tau)$в точке τ_0 есть $|M|$-мерный вектор с компонентами $R^n\left(q_0^n, m\right) - R^B\left(m\right)$. Это указывает пути отыскания весовой функции, а следовательно, и стратегии обучения методом обобщенных градиентов. В экспериментах, приводимых далее, использован другой, теоретически не обоснованный алгоритм.

Алгоритм работает по шагам $i = 1, 2, \ldots$, и на каждом шаге строит стратегию обучения q_i^n и весовую функцию $\tau_i = \left(\tau_i\left(m\right) \middle| m \in M\right)$.

Исходными для алгоритма являются конечное множество M моделей и число $\varepsilon > 0$ – требуемая точность решения задачи.

1. Для каждой модели m определить целое число $n(m) = 1$.

2. Для каждой модели $m \in M$ определить ее вес $\tau\left(m\right) = \dfrac{n\left(m\right)}{\sum\limits_{m \in M} n\left(m\right)}$.

3. Построить байесову стратегию обучения-распознавания q^n для весов $\tau(m)$.

4. Для каждой модели $m \in M$ определить число $R^n(q^n, m)$ – качество текущей стратегии на модели m.

5. Найти гарантированное качество текущей стратегии $R^n\left(q^n\right) = \max\limits_m \left[R^n\left(q^n, m\right) - R^B\left(m\right) \right]$ и наихудшую модель

$$m^* = \arg\max_m \left[R^n\left(q^n, m\right) - R^B\left(m\right) \right].$$

6. Выполнить $n\left(m^*\right)$++.

7. Найти среднее качество

$$\bar{R}^n\left(q^n\right) = \sum_{m \in M} \tau(m) \left[R^n\left(q^n, m\right) - R^B\left(m\right) \right].$$

8. Найти наилучший во всей предыстории алгоритма гарантированный уровень $R^{\min} = \min\left(R^{\min}, R^n\left(q^n\right)\right)$ и наилучшую во всей предыстории стратегию q_0^n.

9. Найти наихудшее во всей предыстории среднее качество $R^{\max} = \max\left(R^{\max}, \bar{R}^n\left(q^n\right)\right)$.

10. Если $R^{\min} - R^{\max} \le \varepsilon$, **останов**.

11. Перейти на п. 3.

В настоящее время теоретически не доказано, хотя экспериментально и подтверждается, что приведенный алгоритм непременно выходит на останов при любой требуемой точности $\varepsilon > 0$. Однако, если он вышел на останов, то его результатом является стратегия q_0^n, решающая задачу обучения с заданной точностью ε в том смысле, что

$$\max_{m \in M} \left[R^n\left(q_0^n, m\right) - R^B\left(m\right) \right] -$$

$$- \max_{m \in M} \left[R^n\left(q^{*n}, m\right) - R^B\left(m\right) \right] \le \varepsilon,$$

где q^{*n} – решение задачи обучения (18) или, что то же самое, задачи (19).

Иллюстрация метода обучения-распознавания

Покажем, как предложенный метод обучения-распознавания реализуется для простейшего класса моделей, описанного выше. В этом классе наблюдение x есть одномерная вещественная величина, $K = \{1, 2\}$, условные распределения $p(x/k)$ полностью известны и равны

$$p\left(x / k\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(x - \mu_k\right)^2}, \quad k = 1, 2, \ \mu_1 = 1, \ \mu_2 = -1.$$

Неизвестными являются только априорные вероятности p_1 и p_2 состояний $k = 1$ и $k = 2$. Одна конкретная модель m, таким образом, есть пара (p_1, p_2). Поскольку в рассматриваемом примере условные распределения $p(x/k)$ полностью известны, нет необходимости рас-

сматривать обучающие последовательности вида $L^n = (x_1, k_1; x_2, k_2; \ldots; x_n, k_n)$, а достаточно рассматривать обучающие последовательности вида $L^n = (k_1, k_2, \ldots, k_n) \in \Lambda^n = K^n$.

Для специального вида потерь $w(k, k') = |k - k'|$, исходя из теоремы 3, стратегия обучения-распознавания должна иметь вид

$$q\left(k' = 1 \middle/ L^n, x\right) = 1,$$

$$\text{если } \sum_{m \in M} \tau(m) \left(\prod_{i=1}^{n} p_{k_i} \right) \cdot p_1 \cdot p\left(x \middle/ 1\right) >$$

$$> \sum_{m \in M} \tau(m) \left(\prod_{i=1}^{n} p_{k_i} \right) \cdot p_2 \cdot p\left(x \middle/ 2\right);$$

$$q\left(k' = 2 \middle/ L^n, x\right) = 1,$$

$$\text{если } \sum_{m \in M} \tau(m) \left(\prod_{i=1}^{n} p_{k_i} \right) \cdot p_1 \cdot p(x/1) <$$

$$p_1 \cdot p\left(x \middle/ 1\right) < \sum_{m \in M} \tau(m) \left(\prod_{i=1}^{n} p_{k_i} \right) \cdot p_2 \cdot p(x/2).$$

В этих выражениях для краткости не указано, что величины p_1 и p_2 и, вообще, p_{k_i}, $i = 1, 2, \ldots, n$, зависят от модели m. Указанную стратегию можно представить в более наглядном виде

$$k^* = 1, \text{ если } \frac{p(x/1)}{p(x/2)} > \theta(\tau, n_1, n_2);$$

$$k^* = 2, \text{ если } \frac{p(x/1)}{p(x/2)} < \theta(\tau, n_1, n_2),$$

где k^* – решение о состоянии k по наблюдению x и обучающей последовательности. Числа n_1 и n_2 обозначают, сколько раз в обучающей последовательности L^n встретилась ситуация, когда $k_i = 1$ и $k_i = 2$. Решение о состоянии объекта принимается на основании сравнения отношения правдоподобия $p(x/1)/p(x/2)$ с порогом, зависящим от обучающей выборки. Однако значение этого порога определяется как

$$\theta(\tau, n_1, n_2) = \frac{\sum_{m \in M} \tau(m)\, p_1^{n_1} \cdot p_2^{n_2+1}}{\sum_{m \in M} \tau(m)\, p_1^{n_1+1} \cdot p_2^{n_2}}, \qquad (24)$$

а совсем не так просто и грубо,

$$\theta(\tau, n_1, n_2) = \frac{n_1}{n_2}, \qquad (25)$$

как в методе наиболее правдоподобных моделей. По-видимому, стратегии, основанные на сравнении с порогами (24) и (25), становятся неотличимыми при неограниченном росте обучающей последовательности. Однако при малых значениях длины, как показали эксперименты, это различие существенно.

Описанный в предыдущей главе метод обучения-распознавания включает в себя и отыскание значений весов $\tau(m)$, $m \in M$. Для реализации этого метода следует располагать конструктивным методом вычисления риска $R^n(q^n, m)$, который определен как

$$R^n(q^n, m) = \sum_{L^n \in \Lambda^n} \sum_{x \in X} \sum_{k' \in K} \tilde{q}^n(k' / L^n, n)\, p(L^n; n) \times$$
$$\times \sum_{k \in K} p(x, k; m)\, w(k, k').$$

В общем случае наибольшую трудность здесь представляет суммирование по всем обучающим выборкам $L^n \in \Lambda^n$. В рассматриваемом примере этот риск равен

$$R^n(q^n, m) = \sum_{L^n \in \Lambda^n} \left(\prod_{i=1}^{n} p_{k_i} \right) \sum_{k=1}^{n} p_k \cdot \Phi_k\left(\theta(\tau, n_1, n_2)\right), \quad (26)$$

где

$$\Phi_1(\theta) = \int_{-\infty}^{\theta} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-1)^2}, \quad \Phi_2(\theta) = \int_{\theta}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x+1)^2}.$$

Представим множество Λ^n в виде

$$\Lambda^n = \bigcup_{\substack{n_1, n_2 \\ n_1 + n_2 = n}} \Lambda(n_1, n_2),$$

где $\Lambda(n_1, n_2)$ – множество всех тех выборок $L^n = (k_1, k_2, \ldots, k_i, \ldots, k_n)$ длины $n_1 + n_2$, в которых событие $k_i = 1$ произошло n_1 раз, а событие $k_i = 2 - n_2$ раз. С учетом этого обозначения выражение для риска $R^n(q^n, m)$ приобретает вид

$$R^n\left(q^n,m\right)=\sum_{L^n\in\Lambda^n}p_1^{n_1}\cdot p_2^{n_2}\sum_{k=1}^{2}p_k\cdot\Phi_k\left(\theta\left(\tau,n_1,n_2\right)\right)=$$

$$=\sum_{\substack{n_1,n_2\\ n_1+n_2=n}}\ \sum_{L^n\in\Lambda\left(n_1,n_2\right)}p_1^{n_1}\cdot p_2^{n_2}\sum_{k=1}^{2}p_k\cdot\Phi_k\left(\theta\left(\tau,n_1,n_2\right)\right)=$$

$$=\sum_{\substack{n_1,n_2\\ n_1+n_2=n}}p_1^{n_1}\cdot p_2^{n_2}\sum_{k=1}^{2}p_k\cdot\Phi_k\left(\theta\left(\tau,n_1,n_2\right)\right)\sum_{L^n\in\Lambda\left(n_1,n_2\right)}1=$$

$$=\sum_{\substack{n_1,n_2\\ n_1+n_2=n}}C_n^{n_1}\cdot p_1^{n_1+1}\cdot p_2^{n_2}\cdot\Phi_1\left(\theta\left(\tau,n_1,n_2\right)\right)+$$

$$+\sum_{\substack{n_1,n_2\\ n_1+n_2=n}}C_n^{n_1}\cdot p_1^{n_1}\cdot p_2^{n_2+1}\cdot\Phi_2\left(\theta\left(\tau,n_1,n_2\right)\right),$$

и его вычисление перестает быть проблематичным.

На рис. 2–7 показана зависимость риска различных стратегий от модели m, в данном случае, от вероятности p_1 первого состояния. Байесовский риск $R^B(m)$ показан сплошной линией, риск стратегии по методу наибольшего правдоподобия показан пунктирной линией, а риск стратегии по предлагаемому методу – жирной пунктирной линией. Рис. 2–7 соответствуют различным значениями длины n обучающих выборок: 1, 2, 3, 10, 20 и 50.

Расчет весов $\tau(m)$, $m \in M$, выполнялся в соответствии с алгоритмом, описанным в предыдущем разделе.

Заключение. Результаты данной статьи изложены на определенном уровне общности, который мог быть и большим, и меньшим. Обозначим некоторые возможные обобщения приведенных результатов, которые уместно рассмотреть в данной статье.

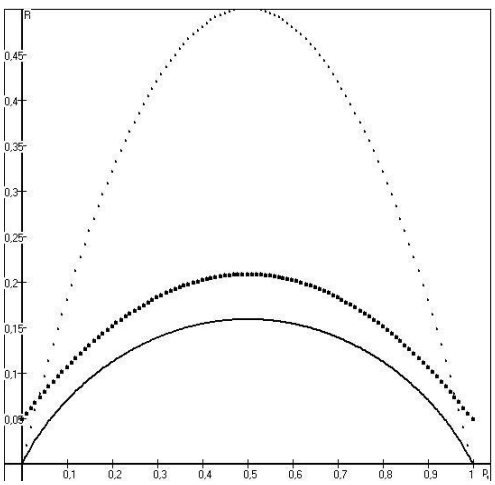

Рис. 2. Вероятности ошибочного распознавания для различных стратегий при длине обучающей выборки, равной 1

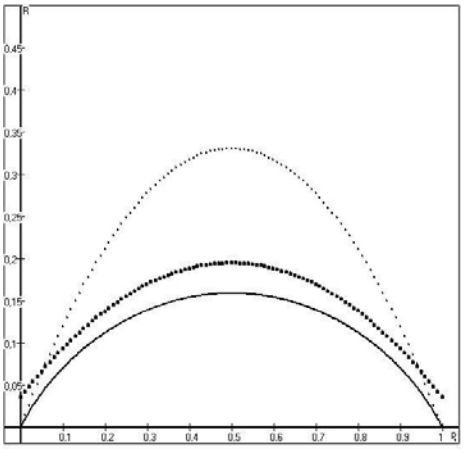

Рис. 3. Вероятности ошибочного распознавания для различных стратегий при длине обучающей выборки, равной 2

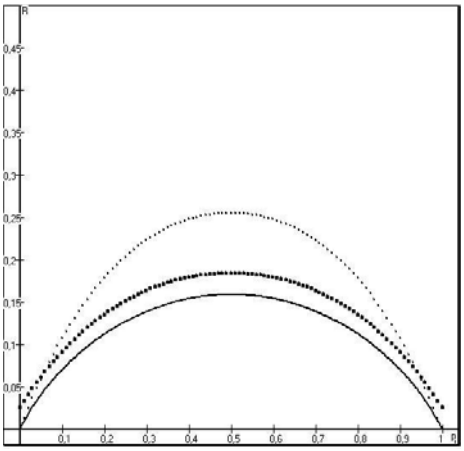

Рис. 4. Вероятности ошибочного распознавания для различных стратегий при длине обучающей выборки, равной 3

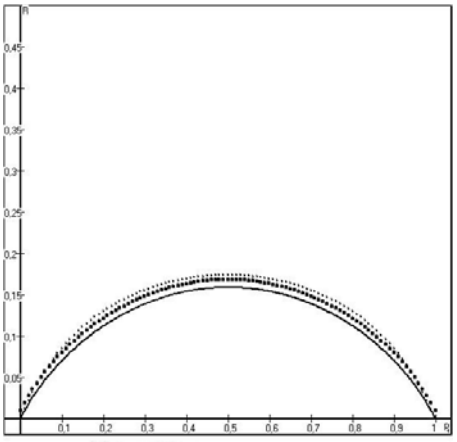

Рис. 5. Вероятности ошибочного распознавания для различных стратегий при длине обучающей выборки, равной 10

Обучение и самообучение

Обучающая информация, на основании которой строится собственно распознавание, совсем не обязательно должна быть выборкой вида $(x_1, k_1; x_2, k_2; \ldots; x_n, k_n) \in (X \times K)^n$. Это может быть результат L любого эксперимента над распознаваемым объектом. Важно только, чтобы этот результат мог считаться случайным и зависящим от модели, так, чтобы имела смысл вероятность $p(L; m)$. Результаты статьи без изменения переносятся и на этот более общий случай. В частности, обучающая информация может быть выборкой $L = (x_1, x_2, \ldots, x_n) \in X^n$, состоящей только из наблюдаемых при распознавании признаков и не включающей в себя информацию о скрытых состояниях. Построение распознающих систем при обучающей информации такого вида получило название *самообучение*. Основным методом построения таких систем стало, как и при обучении, максимально правдоподобное оценивание модели на основании выборки $L = (x_1, x_2, \ldots, x_n)$, для которого широко применяются алгоритмы [5], ставшие известными как *EM*-алгоритмы [6].

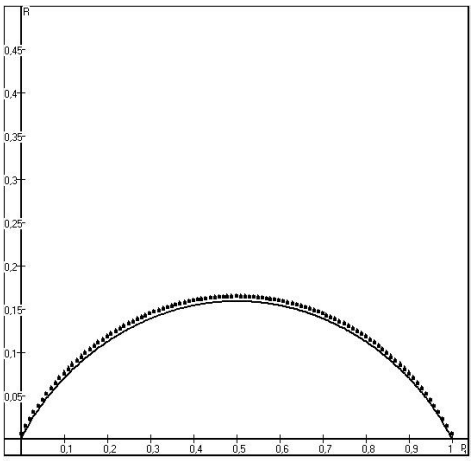

Рис. 6. Вероятности ошибочного распознавания для различных стратегий при длине обучающей выборки, равной 20

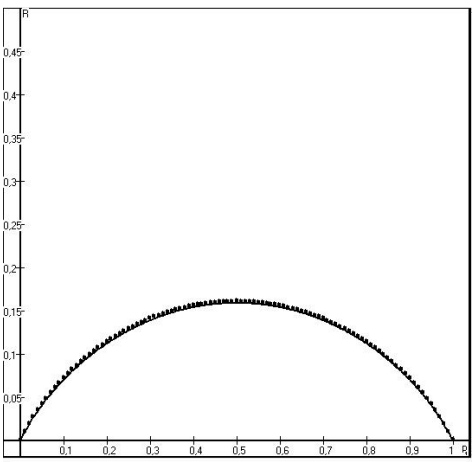

Рис. 7. Вероятности ошибочного распознавания для различных стратегий при длине обучающей выборки, равной 50

Описанный подход вносит коррективы и в эти известные методы распознавания в режиме самообучения. Как и в случае обучения, любые разумные требования к распознаванию в режиме самообучения приводят к алгоритмам, в которых отсутствует такой этап, как поиск наиболее правдоподобной модели на основании выборки $L = (x_1, x_2, \ldots, x_n)$. Эти коррективы могут быть существенными, особенно при малых значениях длины n выборки L.

Здесь возможны и более существенные выводы, которые покажем на примере. Пусть для заданного класса моделей и обучающей информации произвольного вида построена стратегия $\tilde{q}(k/L, x)$ обучения-распознавания. Коль скоро эта стратегия построена, то она предназначена не для одноразового применения при распознавании какого-то одного объекта, а для распознавания многих объектов. При этом предполагается, что все эти объекты характеризуются той же моделью m, при которой была получена обучающая информация L. Пусть теперь требуется принять решение о состояниях k_1 и k_2 двух объектов по признакам x_1 и x_2 этих объектов. В этом случае можно поступить двумя способами.

Первый способ состоит в применении полученной стратегии $\tilde{q}(k/L, x)$ к каждому из двух предъявленных объектов и с вероятностью $\tilde{q}(k_1/L, x_1)$ принять решение, что первый объект находится в состоянии k_1, и с вероятностью $\tilde{q}(k_2/L, x_2)$ принять решение, что второй объект находится в состоянии k_2.

Второй способ состоит в рассмотрении пары предъявленных объектов как одного объекта с

состоянием из множества $K \times K = \{(k_1, k_2) | k_1 \in K, k_2 \in K\}$, с признаком (x_1, x_2), который принимает значения из множества $X \times X$, и классом M моделей m, таких, что

$$p(x_1, x_2, k_1, k_2; m) = p(x_1, k_1; m) \cdot p(x_2, k_2; m). \quad (27)$$

Для такого класса моделей строится стратегия $q'(k_1, k_2 / L, x_1, x_2)$, в соответствии с которой решения о состояниях (k_1, k_2) пары объектов принимаются на основании пары (x_1, x_2) признаков. Эти решения совсем не обязательно окажутся независимыми одно от другого. Из предположения (27) не следует, что

$$q'(k_1, k_2 / L, x_1, x_2) = q(k_1 / L, x_1) \cdot q(k_2 / L, x_2).$$

Даже в случае, когда два объекта, предъявленные для распознавания, взаимно независимы, решения о их состояниях не являются независимыми. При этом качество стратегии q', конечно, не хуже, чем качество стратегии q, а может быть и значительно лучшим.

Таким образом обнаруживается еще одна предпосылка, принятая в современных подходах к обучению, как сама собою разумеющаяся, но не вытекающая из каких-либо разумных соображений. Считается, что на основании обучающей выборки происходит выбор той или иной стратегии собственно распознавания, которая затем в неизменном виде применяется для всех объектов, предъявленных для распознавания. Это – ненужное и ниоткуда не следующее ограничение.

Необходимость в обучении возникает вообще только при априорной неопределенности модели объекта. Обучающая информация позволяет лишь уменьшить эту неопределенность, а не исключить ее. Каждое из наблюдений, предъявленных для распознавания уже после обработки обучающей информации, статистически зависит от модели и поэтому дополнительно уменьшает ее неопределенность. Если для распознавания предъявлена последовательность из l объектов, т.е. последовательность $x_1, x_2, \ldots, x_l$ признаков этих объектов, то только первый объект должен распознаваться в соответствии со стратегией $q(k_1 / L, x_1)$. При распознавании второго объекта признак x_1 становится уже частью обучающей информации и распознавание должно выполняться в соответствии с другой стратегией вида $q((k_2/(L, x_1), x_2)$. И вообще, i-й объект должен распознаваться в соответствии со стратегией вида $q(k_i/(L, x_1, \ldots, x_{i-1}), x_i)$.

Обучение, т.е. изменение стратегии распознавания, совсем не должно заканчиваться обработкой обучающей информации L. Обработка этой информации является лишь нулевым этапом обучения, которое затем продолжается в течение всего периода эксплуатации распознающей системы.

Эмпирический байесовский подход

Очевидно, что при неполном знании статистической модели распознавание многих объектов одновременно может оказаться заметно лучшим, чем независимое одно от другого распознавание каждого объекта в отдельности. Этот вывод справедлив независимо от вида обучающей последовательности, в том числе и в ее отсутствие. Эту идею в частном виде высказал Г. Роббинс при формулировке предложенного им эмпирического байесовского подхода [7, 8]. Ученый сформулировал его для случая, когда модель распознаваемого объекта почти полностью известна, а неизвестны только априорные вероятности скрытых состояний. Для этого примера была указана реализация предлагаемого подхода при неограниченном росте количества объектов, предъявленных для распознавания. Асимптотический характер подхода выражен в самом его названии «асимптотически субминимаксные решения в составных задачах теории статистических решений». Вопрос о том, каким должно быть решение не в асимптотическом случае, а при распознавании конечных совокупностей объектов, был оставлен для будущих исследований и до сих пор оставался без ответа. Исследования задачи обучения указывают путь к реализации эмпирического байесовского подхода и в случае конечной совокупности объектов, предъявленных для распознавания.

Проблема коротких выборок

В изложенном подходе совсем не обязательно считать, что распознавание объекта состоит

в указании непременно состояния объекта. Можно допустить и специальный ответ распознающей системы, обозначаемый символом $\# \notin K$ и понимаемый как отказ от распознавания или ответ НЕ ЗНАЮ. В этом случае функция потерь имеет вид $w: K \times \left(K \cup \{\#\}\right) \to R$, а не $w: K \times K \to R$, как раньше, а стратегия обучения-распознавания принимает вид $\tilde{q}^n : L^n \times X \times \left(K \cup \{\#\}\right) \to R$, а не $\tilde{q}^n : L^n \times X \times K \to R$.

Изложенные идеи построения стратегии обучения-распознавания без каких-либо затруднений можно обоснованно изложить и при таком расширенном определении основных понятий.

Введение в рассмотрение понятия «отказ от распознавания» приводит к новой точке зрения на проблему коротких выборок. Исследования по проблеме таких выборок имеют целью ответить на вопрос, при какой минимальной длине обучающей выборки последующее распознавание остается еще удовлетворительным в том или ином смысле. Все выборки с длиной, ниже этого предельного значения, объявляются короткими и непригодными для последующего распознавания. Фактически это обозначает отказ от распознавания независимо от того, какой именно оказалась конкретная короткая выборка и каким оказалось текущее наблюдение над объектом, предъявленным для распознавания.

Формулировка задачи обучения как построение оптимальной стратегии обучения-распознавания, допускающей и отказ от распознавания, позволяет принимать решение о хороших или плохих выборках не просто на основании значений их длины, а более точно, на основании того, какой именно оказалась та или иная длинная или короткая выборка. А именно, среди выборок, пусть даже очень коротких, есть такие, на основании которых возможно вполне удовлетворительное последующее распознавание, равно как среди длинных могут оказаться выборки, неудачные в этом отношении.

Понимание того, что результатом обучения есть стратегия обучения-распознавания, позволяет идти еще дальше. А именно, никакая выборка, пусть даже очень короткая или очень длинная, сама по себе не является ни хорошей, ни плохой. Любая выборка позволяет качественно распознавать одни наблюдения и не годится для распознавания других. Хорошей или плохой является не выборка сама по себе, а пара «выборка-наблюдение», и стратегия обучения-распознавания через решение «отказ от распознавания» реализует в том числе и эту классификацию.

Таким образом, классификация выборок на короткие, т.е. плохие, и длинные, т.е. хорошие, есть лишь сильно загрубленная постановка вопроса о построении оптимальной стратегии обучения-распознавания с возможностью отказа от распознавания. Построение оптимальной стратегии обучения-распознавания поглощает проблему коротких выборок и, поглотив ее, решает ее более тонко, более дифференцированно.

1. *Шлезингер М.*, *Главач В.* Десять лекций по статистическому и структурному распознаванию. – К.: Наук. думка, 2004. – 545 с.
2. *Гупал А.М.*, *Пашко С.В.*, *Сергиенко И.В.* Эффективность байесовской процедуры распознавания // Кибернетика и системный анализ. – 1995. – № 4. – С. 76–89.
3. *Гупал А.М.*, *Сергиенко И.В.* Оптимальные процедуры распознавания и их применение // Там же. – 2007. – № 6. – С. 41–54.
4. *Зуховицкий С.И.*, *Авдеева Л.И.* Линейное и выпуклое программирование. – М.: Наука, 1967. – 460 с.
5. *Шлезингер М.И.* Взаимосвязь обучения и самообучения в распознавании образов // Кибернетика. – 1968. – № 2. – С. 81–88.
6. *Demster A.*, *Laird N.*, *Rubin D.* Maximum likelihood from incomplete data via the EM algorithm // J. of the royal Statistic Society. – 1977. – **B39**. – P. 1–38.
7. *Роббинс Г.* Асимптотически субминимаксные решения в составных задачах теории статистических решений // Математика. – 1964. – № 8:2. – С. 141–159.
8. *Роббинс Г.* Эмпирический байесовский подход к статистике // Там же. – С. 133–140.