СИСТЕМНЫЙ АНАЛИЗ

И.В. СЕРГИЕНКО, А.М. ГУПАЛ, А.А. ВАГИС

УДК 519.68

БАЙЕСОВСКИЙ ПОДХОД, ТЕОРИЯ МИНИМИЗАЦИИ ЭМПИРИЧЕСКОГО РИСКА. СРАВНИТЕЛЬНЫЙ АНАЛИЗ

Ключевые слова: *распознавание, эмпирический риск, байесовская процедура, погрешность процедуры, обучающая выборка, цепь Маркова.*

ВВЕДЕНИЕ

Статистическая теория восстановления зависимостей по эмпирическим данным разработана В.Н. Вапником и А.Я. Червоненкисом в конце 60-х — начале 70-х годов [1–3]. Эта теория получила широкую известность в середине 80-х. В настоящее время она активно развивается и применяется для обоснования различных алгоритмов машинного обучения.

Основным результатом данной теории являются количественные оценки, связывающие обобщающую способность алгоритмов с длиной обучающей выборки и сложностью семейства алгоритмов. Эти оценки необходимы для того, чтобы предсказывать, насколько хорошо будет работать построенный алгоритм.

В настоящей работе проводится сравнительный анализ байесовского подхода и методов минимизации эмпирического риска.

ОСНОВНЫЕ ПОНЯТИЯ И ОПРЕДЕЛЕНИЯ

Пусть имеются множество объектов X, множество ответов Y (состояний объектов) и существует целевая функция $y^*: X \to Y$, значения которой $y_i = y^*(x^i)$ известны только на конечном подмножестве объектов $\{x^1,\dots,x^l\} \subset X$. Пара «объект–ответ» (x^i, y_i) называется прецедентом. Совокупность пар $X^l = (x^i, y_i)_{i=1}^l$ называется обучающей выборкой.

Задача обучения заключается в том, чтобы восстановить функциональную зависимость между объектами и ответами, т.е. построить отображение $a: X \to Y$. Алгоритм a должен обладать обобщающей способностью — приближать целевую функцию $y^*(x^i)$ не только на объектах обучающей выборки, но и на всем множестве X.

ВЕРОЯТНОСТНАЯ ПОСТАНОВКА ЗАДАЧИ

Постановка задачи приводится в упрощенном виде, при этом не учитывается, что элементы множества X — не реальные объекты, а лишь их описания, содержащие доступную часть информации об объектах. Невозможно исчерпывающим образом охарактеризовать человека, геологический район, производственное предприятие, экономику страны и т.п. Поэтому одному и тому же описанию x могут соответствовать различные объекты, а значит, и некоторое множество ответов (состояний объектов). Например, в задачах предсказания пространственной структуры белков каждая аминокислота может находиться в трех состояниях, определяющих вторичную структуру белка [4].

Вместо существования неизвестной целевой функции $y^*(x)$ предполагается, что существует неизвестное вероятностное распределение $P(x, y)$ на множестве $X \times Y$, согласно которому сгенерирована выборка пар $X^l = (x^i, y_i)_{i=1}^l$, x — непрерывный вектор $x = (x_1, x_2, \ldots, x_n)$ размерности n, y принимает два значения — нуль и единица.

Следующий шаг в постановке наиболее важен. Он придает точный смысл тому, как выбираются наблюдения, по которым строится правило для классификации и определяется его качество. Принято считать, что на пространстве векторов X существует неизвестная вероятностная мера $P(x)$. В соответствии с $P(x)$ случайно и независимо возникают ситуации x, которые классифицируются с помощью правила $P(y|x)$, т.е. строится обучающая последовательность $X^l = (x^i, y_i)_{i=1}^l$. Для всякого решающего правила $a(x)$ определяет качество обучения как вероятность различной классификации с помощью правила $a(x)$ и правила $P(y|x)$. Чем меньше эта вероятность, тем выше качество обучения. Формально качество решающего правила можно записать в виде

$$I(a) = \int (a(x) - y)^2 P(x, y) dx dy. \tag{1}$$

Минимизация среднего риска (1) является обобщением классических задач, решаемых на основе метода наименьших квадратов, т.е. когда наблюдению $x = (x_1, x_2, \ldots, x_n)$ соответствует не одно, а несколько состояний объектов (исходов экспериментов). В.Н. Вапник и А.Я. Червоненкис были одними из первых, кто придал задачам распознавания строгую математическую трактовку. В дальнейшем задача минимизации среднего риска (1) привлекла внимание многих ученых.

Некоторые исследователи считают, что задачи распознавания сводятся к минимизации среднего риска (1) в специальном классе решающих правил [1]. Существует и другая точка зрения: требуется найти такие структуры описания объектов, для которых можно построить эффективные (оптимальные) процедуры распознавания.

Для булевых векторов $x = (x_1, x_2, \ldots, x_n)$ число различных функций $a(x) \in \{0,1\}$ равно 2^{2^n}, т.е. проблема распознавания относится к задачам большой размерности. Средний риск (1) в булевом случае для двух состояний записывается в виде

$$I(a) = \sum_{x \in X} \sum_{y=0}^{1} (a(x) - y)^2 P(x, y), \tag{2}$$

где усреднение проводится по всем векторам $x = (x_1, x_2, \ldots, x_n)$; он отличен от нуля, поскольку при суммировании исключается только одна из двух вероятностей $P(x, y)$. Функции распознавания $a(x)$ можно легко построить программно: первая функция $a(x)$ на всех 2^n булевых векторах принимает значение нуль, вторая (в двоичной записи) — единица и т.д., наконец, последняя на всех векторах принимает значение единица; число построенных функций равно 2^{2^n}. Очевидно, реализовать данную процедуру на компьютере можно лишь для небольших величин n.

В непрерывном случае мощность множества решающих правил составляет величину

$$2^{\aleph^n} = 2^{\aleph} > \aleph,$$

где $\aleph$ — мощность континуума, т.е. мощность множества решающих правил не только бесконечна, но и превосходит мощность континуума [5]. Вначале рассматривается случай конечного множества параметрических функциональных зависимостей $F(x,\alpha)$ (класс решающих правил). Все функции $F(x,\alpha)$ — характеристические, т.е. принимают два значения — нуль и единица.

В работах В.Н. Вапника изучается задача минимизации среднего риска

$$I(\alpha) = P(\alpha) = \int (y - F(x,\alpha))^2 P(x, y)\, dx\, dy \tag{3}$$

по эмпирическим данным

$$x^1, y_1, \ldots, x^l, y_l. \tag{4}$$

Функционал (3) для каждого решающего правила определяет вероятность ошибочной классификации. Вместо среднего риска (3) минимизируется эмпирический риск, согласно которому за точку минимума (3) принимается точка минимума эмпирического функционала

$$I_э(\alpha) = \nu(\alpha) = \frac{1}{l}\sum_{i=1}^{l}(y_i - F(x^i, \alpha))^2, \tag{5}$$

построенного по случайной независимой выборке (4). Если минимум функционала (5) достигается на функции $F(x,\alpha_э)$, необходимо установить, в каких случаях найденная функция $F(x,\alpha_э)$ близка к функции $F(x,\alpha_0)$, которая минимизирует (3) в классе функций $F(x,\alpha)$.

Данная проблема связана с проблемой существования равномерной сходимости частот к математическим ожиданиям: близость найденного решения к наилучшему следует из достаточно сильного условия, когда для любого ε выполняется равенство

$$\lim_{l\to\infty} P\{\sup_{\alpha} |P(\alpha) - \nu(\alpha)| > \varepsilon\} = 0. \tag{6}$$

Условие равномерной сходимости (6) в реальных задачах распознавания выполнить невозможно, поскольку объем выборки фиксирован и его нельзя по своему усмотрению увеличивать до бесконечности, так как это связано с выполнением, как правило, дорогостоящих экспериментов. Для булевых задач число различных векторов составляет конечную величину 2^n, объем обучающей выборки l — лишь небольшую долю от экспоненты 2^n. Поэтому операция предельного перехода в (6) невыполнима для дискретных задач.

Кроме того, в теории вероятности пытаются избегать определения вероятности массового события в виде предела частоты $\frac{m}{n}$, когда число испытаний $n \to \infty$. В качестве численного значения вероятности выбирается число, вблизи которого колеблется частота события при большом, но конечном числе наблюдений [6].

Показано, что для конечного множества решающих правил $F(x,\alpha_1),\ldots,F(x,\alpha_N)$ равномерная сходимость частот появления событий к их вероятностям вытекает из условия $l \to \infty$. Каждому решающему правилу $F(x,\alpha_i)$ можно поставить в соответствие событие A_i, состоящее из тех пар x, y, на которых $(y - F(x,\alpha_i))^2 = 1$, т.е. определено конечное число N событий $A_1,\ldots,A_N$.

Из неравенства Бернштейна [7] вытекает оценка

$$P\{|P(\alpha_i) - \nu(\alpha_i)| > \varepsilon\} < \exp\{-\varepsilon^2 l\}, \tag{7}$$

поскольку дисперсия отдельного члена в эмпирическом риске (5) не превосходит 1/4. Из (7) следует

$$P\{\sup_i |P(\alpha_i) - \nu(\alpha_i)| > \varepsilon\} < N\exp\{-\varepsilon^2 l\}. \tag{8}$$

Для конечного множества решающих правил из (8) вытекают следующие оценки.

Теорема 1. Пусть множество решающих правил состоит из N элементов и для решающих правил $F(x,\alpha_i)$ частоты ошибок на обучающей последовательности длины l равны $\nu(\alpha_i)$. Тогда с вероятностью $1-\eta$ можно утверждать, что одновременно для всех решающих правил выполняются неравенства

$$\nu(\alpha_i) - \sqrt{\frac{\ln N - \ln\eta}{l}} < P(\alpha_i) < \nu(\alpha_i) + \sqrt{\frac{\ln N - \ln\eta}{l}}. \tag{9}$$

Поскольку неравенства справедливы всех N правил, теорема 1 устанавливает доверительный интервал для качества решающего правила $F(x,\alpha_э)$, которое минимизирует среди N правил эмпирический риск. Он равен

$$\nu(\alpha_э) - \sqrt{\frac{\ln N - \ln\eta}{l}} < P(\alpha_э) < \nu(\alpha_э) + \sqrt{\frac{\ln N - \ln\eta}{l}}. \tag{10}$$

Оценки (9), (10) имеют асимптотический характер, т.е. выполняются для больших выборок. Кроме того, в них присутствует вероятностный параметр η. Необходимо задать этот параметр, подставить в формулы и вычислить оценку погрешности. Для достижения высокой вероятности параметр η следует положить близким к нулю, вследствие этого оценки завышены.

Очевидно, что конечное множество решающих правил является ограниченным множеством для непрерывных задач распознавания. Поэтому В.Н. Вапник делает попытку обобщить эти результаты на случай бесконечного числа решающих правил.

Пусть задано множество S решающих правил $F(x,\alpha)$ и дана выборка $x^1, \ldots, x^l$. Ее можно разделить на два класса 2^l способами. (С помощью правила $F(x,\alpha)$ множество $x^1, \ldots, x^l$ делится на два подмножества — подмножество, на котором $F(x,\alpha)=1$, и подмножество, на котором $F(x,\alpha)=0$.)

Число таких способов разделения зависит как от класса решающих правил $F(x,\alpha)$, так и от состава выборки. Это число обозначается $\Delta^S(x^1, \ldots, x^l)$ и называется индексом системы S относительно выборки $x^1, \ldots, x^l$.

Рассматривается система событий

$$S(\alpha)=\{x, y: (y-F(x,\alpha))^2=1\},$$

образованных множеством решающих правил $F(x,\alpha)$. Функция

$$m^S(l)=\max_{x^1,\ldots,x^l}\Delta^S(x^1, \ldots, x^l) \tag{11}$$

называется функцией роста системы событий S, где максимум берется по всем возможным выборкам длины l. Функция роста вычисляет максимальное число способов разделения l точек на два класса с помощью решающих правил.

Определение (11) — ключевой момент теории. Отмечается, что максимум всегда достигается, поскольку индекс $\Delta^S(x^1, \ldots, x^l)$ принимает конечное число значений. Однако это может быть не так, если область определения компонент вектора x — вещественная прямая $(-\infty,\infty)$. Согласно определению (11) нужно организовать бесконечное (континуальное) множество выборок длины l и к каждой выборке применить бесконечное число решающих правил, разделяющих выборку на два класса. Поэтому в (11) должен быть указан эффективный алгоритм вычисления $m^S(l)$. В противном случае определение функции роста воспринимается как акт веры, а в математике такие определения считаются некорректными [8].

В тексте работы [2] приводится фраза, которая вводит читателя в заблуждение тем, что функцию $m^S(l)$ можно легко определить: «для функции роста справедлива замечательная теорема, которая позволяет легко ее оценить».

Теорема 2. Функция роста либо тождественно равна 2^l, либо при $l>h$ мажорируется функцией $m^S(l)<1{,}5\dfrac{l^h}{h!}$, где $h+1$ — минимальный объем выборки, при котором нарушается условие $m^S(l)=2^l$. Иначе говоря,

$$m^S(l)=\begin{cases}\text{либо} \equiv 2^l,\\ \text{либо} < 1{,}5\dfrac{l^h}{h!}\quad (l>h).\end{cases}$$

Для того чтобы оценить функцию роста, необходимо показать, что либо для любого l найдутся точки $x^1, \ldots, x^l$ такие, что с помощью решающих правил $F(x,\alpha)$ их можно разбить на два класса всеми 2^l возможными способами, либо существует число h такое, что h точек можно, но никакие $h+1$ точек нельзя разбить на два класса всеми возможными способами.

Даже если теорема 2 верна, то неясно, как определить переход, при котором нарушается условие $m^S(l)=2^l$, поскольку мощность множества S решающих правил может превосходить мощность континуума.

Класс характеристических функций имеет емкость h, если справедливо неравенство $m^S(l) < 1{,}5\dfrac{l^h}{h!}$ $(l > h)$. В случае выполнения равенства $m^S(l) = 2^l$ считается, что емкость класса характеристических функций $F(x,\alpha)$ бесконечна.

В качестве примера приводится оценка функции роста для суммы линейных по параметру решающих правил (этот случай играет важную роль в дальнейшей теории В.Н. Вапника):

$$F(x,\alpha) = \theta\Big(\sum_{i=1}^{n} \alpha_i \varphi_i(x)\Big); \qquad \theta(z) = \begin{cases} 1, & \text{если } z \ge 0, \\ 0, & \text{если } z < 0. \end{cases} \tag{12}$$

Множество параметрических функций (12) образует континуальный класс функций, так как параметры α_i — вещественные числа. Как отмечалось, функция роста эффективно не вычисляется в точке. В работе [2] автор прибегает к следующему приему. Приведем ход его рассуждений: «Нетрудно найти функцию роста для класса событий, заданных линейными решающими правилами (12). Для этого определяется максимальное число точек h в пространстве размерности n, которые можно с помощью гиперплоскости разбить на два класса всеми 2^h способами. Известно, что это число равно n. Поэтому для класса линейных решающих правил функция роста оценивается формулой $m^S(l) < 1{,}5\dfrac{l^n}{n!}$ $(l > n)$».

Данные рассуждения неубедительны, поскольку функция $\sum_{i=1}^{n} \alpha_i \varphi_i(x)$ не является гиперплоскостью и, кроме того, как легко заметить, при $n = 2$ три точки можно разбить на два класса 2^3 способами. Это свидетельствует о математических просчетах.

Далее была получена оценка скорости равномерной сходимости частот к вероятности по классу событий $S(\alpha)$. Показано, что имеет место неравенство

$$P\left\{\sup_{\alpha} |P(\alpha) - \nu(\alpha)| > \varepsilon\right\} < 6m^S(2l)\exp\left\{-\frac{\varepsilon^2 l}{4}\right\}.$$

Если емкость класса решающих правил конечна, $m^S(l) < 1{,}5\dfrac{l^h}{h!}$, то приводилась следующая теорема [2].

Теорема 3. Пусть $F(x,a)$ — класс решающих правил ограниченной емкости h и $\nu(\alpha)$ — частота ошибок, вычисленная по обучающей последовательности для правила $F(x,a)$. Тогда с вероятностью $1-\eta$ для всех правил $F(x,a)$ вероятность ошибочной классификации заключена в пределах

$$\nu(\alpha) - \sqrt{\frac{h\left(\ln\frac{2l}{h}+1\right) - \frac{\ln\eta}{9}}{l}} < P(\alpha) < \nu(\alpha) + 2\sqrt{\frac{h\left(\ln\frac{2l}{h}+1\right) - \frac{\ln\eta}{9}}{l}}.$$

При выводе свойств функции роста $m^S(l)$ и теоремы 3 требуется неоднократное вычисление этой функции. Поскольку функцию $m^S(l)$ эффективно вычислить невозможно, полученные результаты для бесконечного числа решающих правил нельзя считать обоснованными.

ЗАМЕЧАНИЯ ОТНОСИТЕЛЬНО ТЕОРИИ МИНИМИЗАЦИИ ЭМПИРИЧЕСКОГО РИСКА

Если алгоритм a доставляет минимум эмпирическому функционалу на заданной обучающей выборке X^l, то это еще не означает, что он будет хорошо приближать целевую зависимость на произвольной контрольной выборке. Специфика методов минимизации эмпирического риска состоит в том, что выборка X^L разбивается на обучающую выборку X^l длины l и контрольную X^k длины k,

$L = l + k$. На этапе обучения на основе выборки X^l строится алгоритм $a = \mu(X^l)$, затем его работа проверяется на контрольной выборке X^k. Естественно, что при таком сокращении выборки эффективность алгоритма уменьшается, поскольку в приведенных ранее оценках фигурирует длина выборки.

Теория минимизации эмпирического риска построена для непрерывных объектов $x = (x_1, x_2, \ldots, x_n)$ размерности n. Верхние оценки погрешности метода минимизации эмпирического риска построены для конечного числа параметрических функций. Поэтому функция $a(x)$, которая минимизирует средний риск (1), может не принадлежать классу параметрических функций. Поскольку минимум среднего риска неизвестен, построенные оценки (9), (10) не определяют величину отклонения минимума эмпирического риска от минимума (1).

Оценки (9), (10) в теории минимизации эмпирического риска получены на основе эмпирических данных $x^1, y_1, \ldots, x^l, y_l$, которые являются некоторой случайной последовательностью описаний объектов и их состояний, поэтому они носят вероятностный характер. В эти оценки входит длина выборки, конечное число решающих правил N и вероятностный параметр η.

В обучающей выборке количества объектов различных классов известны, более того, на практике они часто определяются заранее. Если в выборке отсутствует один из классов объектов, то оценки (9), (10) дают пользователю неверное представление о работе метода. Представим, что медицинская экспертная система строится только на классе «больных» либо «здоровых» пациентов (или размеры этих классов значительно отличаются). Очевидно, что эффективных процедур распознавания в таком случае построить нельзя.

Если обучающая выборка содержит только один класс объектов, то можно построить пример, когда эмпирический риск окажется нулевым, в то время как средний риск (2) будет максимальным. Пусть $n = 1$ и для булевого случая возьмем следующие вероятностные распределения:

$$P(x = 0, y = 0) = 0{,}1, \quad P(x = 0, y = 1) = 0{,}3,$$
$$P(x = 1, y = 0) = 0{,}2, \quad P(x = 1, y = 1) = 0{,}4.$$

Тогда функций распознавания будет четыре:

$$a_1(x = 0) = 1, \quad a_1(x = 1) = 1; \qquad a_2(x = 0) = 1, \quad a_2(x = 1) = 0;$$
$$a_3(x = 0) = 0, \quad a_3(x = 1) = 1; \qquad a_4(x = 0) = 0, \quad a_4(x = 1) = 0.$$

Функция a_1 минимизирует риск (2), он равен 0,3. Если обучающая выборка содержит объекты только класса 0 и в выборке присутствуют объекты $x = 0$ и $x = 1$, то функция a_4 дает нулевой эмпирический риск, однако у этой функции наблюдается максимальный риск (2), равный 0,7.

Наличие такого рода контрпримеров показывает, что подобные «плохие» задачи в совокупности могут составлять как раз тот диапазон вероятности η, при котором оценки (9), (10) не выполняются. В этом заключается недостаток вероятностных оценок. Другими словами, в оценки методов распознавания должны входить размеры классов, а не общая длина выборки; в этом случае контрпримеры исключаются.

Исследование оценок методов минимизации эмпирического риска проводилось для непрерывного случая и двух классов объектов. Легко заметить, что эмпирический риск (если его рассматривать как случайную величину) имеет биномиальное распределение с параметром l и вероятностью $P(a(x) \neq y)$, так как единица в эмпирическом риске появляется с вероятностью $P(a(x) \neq y)$. Поэтому для конечного числа решающих правил оценки (9) очевидным образом вытекают из обобщенного неравенства Чебышова (неравенство Бернштейна), и в таком случае в оценку входит длина выборки l и число решающих правил N.

Методы минимизации эмпирического риска нерациональны по своему способу построения. В этих методах оптимальные параметры α^* решающих правил $F(x, \alpha)$ определяются путем минимизации эмпирического риска по некоторой случайной обучающей выборке $x^1, y_1, \ldots, x^l, y_l$. Очевидно, что для последовательности новых объектов, которые не присутствуют в обучающей выборке, найденные параметры

уже могут не быть оптимальными и их нужно вычислять заново. Другими словами, находить точный минимум эмпирического риска (5) по отдельным обучающим последовательностям не имеет смысла. Поэтому возникает проблема разбиения выборки на обучающую и контрольную. Естественно, что методы, в которых присутствуют трудоемкие процедуры настройки оптимальных параметров по отдельным обучающим выборкам, неприменимы в живой природе, а также при распознавании объектов в быстроменяющейся обстановке.

БАЙЕСОВСКИЕ ПРОЦЕДУРЫ РАСПОЗНАВАНИЯ

В работах [10, 11] построена теория статистического оценивания дискретных процедур распознавания. Выбран прямой подход к построению методов распознавания, а именно, найти такие структуры описания объектов, для которых можно построить эффективные и даже оптимальные процедуры распознавания. Данный подход имеет очевидное преимущество перед другими подходами: вместо использования достаточно сложных процедур минимизации эмпирического риска или других функционалов качества (в специальном классе параметрических функций) применяется простое байесовское правило классификации, которое используется для решения задач большой размерности.

Оптимальным правилом классификации является байесовское правило, которое максимизирует апостериорную вероятность распределения классов при заданном наборе признаков объекта. Однако проблема заключается в том, что распределение вероятностей неизвестно, известны лишь эмпирические данные относительно значений признаков объектов и их состояний, которые составляют обучающую выборку. Построенное правило классификации должно хорошо работать за пределами обучающей выборки, т.е. обладать обобщающей способностью; желательно, чтобы оценка погрешности процедуры достаточно быстро уменьшалась при увеличении размеров классов объектов, составляющих обучающую выборку.

Рассматривается задача минимизации среднего риска

$$I(a) = \sum_{x \in X} \sum_{y=0}^{1} (a(x) - y)^2 P(x, y). \tag{13}$$

Показано, что в булевом случае байесовская процедура для объектов с независимыми признаками является субоптимальной, построена верхняя и нижняя оценки погрешности процедуры в зависимости от размеров обучающей выборки, которые совпадают с точностью до абсолютной константы.

Наблюдаемый объект описывается вектором $x_1, x_2, \ldots, x_n$, f, где $x_1, x_2, \ldots, x_n$ — признаки (измерения) объекта, f — состояние объекта. Пусть проведено m наблюдений над объектами и в каждом случае зафиксировано состояние объекта. Имеем обучающую выборку $V = (V_0, V_1, V_2)$ следующего вида.

На множестве X описаний объектов $(x_1, x_2, \ldots, x_n, f)$ задано некоторое распределение вероятностей P, которое заранее неизвестно. Первая часть V_0 — булева матрица размерности $m_0 \times n$, где m_0 – число строк. Каждая строка представляет собой вектор $x = (x_1, x_2, \ldots, x_n, f)$, который выбран в соответствии с распределением P при условии $f = 0$. Вторая часть V_1 — булева матрица размерности $m_1 \times n$, где m_1 — число строк. Каждая строка матрицы — вектор x, который выбран на основе распределения P при условии $f = 1$. Последняя часть V_2 — булев вектор размерности m_2. Каждая компонента этого вектора — наблюдаемое значение состояния f, которое выбирается в соответствии с распределением P. Поскольку наблюдения проводятся так же, как и в теории минимизации эмпирического риска, можно считать, что $m_2 = m = m_0 + m_1$.

Индуктивный шаг. Требуется построить такую процедуру индуктивного вывода, которая по измерениям $x_1, x_2, \ldots, x_n$ любого следующего объекта и обучающей выборке $V = (V_0, V_1, V_2)$ определяет состояние f объекта.

Поступивший на вход обучающей системы вектор $x = (x_1, x_2, \ldots, x_n)$ может присутствовать в выборке V_0 или выборке V_1 либо в обеих выборках вместе. Однако указать, к какому классу принадлежит $x = (x_1, x_2, \ldots, x_n)$, нельзя, так как не известны вероятности $p(x, 0)$ и $p(x, 1)$. Этот вопрос решает байесовская процедура распознавания.

Погрешность процедуры. Сложность класса задач. Полагаем, что процесс определения состояния f объекта по известным значениям вектора входа $x=(x_1,x_2,\dots,x_n)$ проводится с помощью функции $a(x)$, т.е. $f=a(x)$. Поскольку число функций $a(x)$ конечно, среди них существует наилучшая функция $a^*(x)$, такая, что $P(x,a^*(x))=\max(P(x,0),P(x,1))$. Легко заметить, что функция $a^*(x)$ минимизирует средний риск (13).

Погрешность функции $a(x)$ — усредненная величина

$$v(a,P)=\sum_x P(x,a^*(x))-P(x,a(x)), \tag{14}$$

т.е. $v(a^*)\equiv 0$. Погрешность (14) в отличие от методов минимизации эмпирического риска указывает на отклонение погрешности функции $a(x)$ от минимума среднего риска, поэтому для практических расчетов с ней удобнее работать.

Индуктивная процедура распознавания Q строит по данной выборке $V=(V_0,V_1,V_2)$ и вектору x функцию $a(x)=Q(V,x)$.

Погрешность процедуры Q на распределении P — усредненная величина

$$v(Q,P)=\sum_{V\in W} v(a,P)P_1(V). \tag{15}$$

Усреднение в (15) проводится по всем обучающим выборкам V, имеющим заданные размеры классов, из некоторого конечного множества W, $P_1(V)$ — вероятность получения выборки V. Для независимых признаков вероятность $P_1(V)$ полностью определяется распределением P. Усреднение (15) проводится для того, чтобы получить детерминированные оценки погрешности байесовской процедуры распознавания. Операция усреднения, по сути, выполняет функцию контроля: формула (15) оценивает качество работы процедуры на всевозможных новых объектах, не входящих в состав обучающей выборки. При этом у всех обучающих выборок размеры классов одинаковы.

Класс задач $C\equiv C(m_0,m_1,m_2,n)$ — совокупность всевозможных распределений вероятностей P, детерминированные числа m_0,m_1,m_2,n — вход задачи, они определяют размеры выборки.

Погрешностью процедуры распознавания Q на классе C называется число $v(Q,C)=\sup\limits_{P\in C} v(Q,P)$. Сложность класса C определяется величиной

$$\mu(C)=\inf_Q v(Q,C)=\inf_Q \sup_{p\in C} v(Q,P).$$

Нужно построить такую процедуру распознавания Q, для которой число $v(Q,C)$ мало отличается от числа $\mu(C)$.

Пусть $d=(d_1,d_2,\dots,d_n)$ — булев вектор. Считаем, что распределения P из класса C при каждом d удовлетворяют условию $P(x_1=d_1,x_2=d_2,\dots,x_n=d_n\mid f=i)=\prod\limits_{j=1}^{n} P(x_j=d_j\mid f=i)$, $i=0,1$, что означает независимость признаков x_j для каждого класса объектов.

Рассматриваются случайные величины $\xi(d,i)$, которые зависят от d и i как от параметров:

$$\xi(d,i)=\prod_{j=1}^{n}(k(d_j,i)/m_i)k_i/m_2,\ i=0,1. \tag{16}$$

Здесь $k(d_j,i)$ — количество значений, равных d_j, j-го признака в j-м столбце матрицы V_i; k_i — количество значений целевого признака, равных i, в векторе V_2. Тогда функция распознавания определяется формулой

$$a(d)=\begin{cases}0, & \text{если } \xi(d,0)\ge\xi(d,1),\\ 1, & \text{если } \xi(d,0)<\xi(d,1).\end{cases} \tag{17}$$

Процедуру распознавания, определяемую соотношениями (16), (17), обозначим Q_B. Заметим, что величины $\zeta(d,i)/(\zeta(d,0)+\zeta(d,1))$ представляют собой приближенные значения вероятностей $P(f=i\mid x_1=d_1,x_2=d_2,\dots,x_n=d_n)$, вычислен-

ных по теореме Байеса. При подсчете оценок этих вероятностей существенным образом используется информация относительно размеров классов в обучающей выборке, поэтому данные значения входят в приводимые далее оценки погрешности байесовской процедуры. Для байесовской процедуры проводить разбиение на обучающую и контрольную выборки не нужно, поскольку при подсчете погрешности (15) учитывается работа процедуры на всем множестве обучающих выборок. Байесовская процедура (16), (17) строится программным образом, найти ее аналитический вид невозможно. Методы минимизации эмпирического риска работают с известными функциями.

В работе [10] получена оценка сверху погрешности байесовской процедуры на классе C.

Теорема 4. Существует абсолютная константа $\alpha < \infty$ такая, что справедливо неравенство

$$v(Q_B, C) \le a\sqrt{\frac{n}{\min(m_0, m_1)} + \frac{1}{m_2}}. \tag{18}$$

При условии $\min(m_0, m_1) \ge 2n$ оценка сверху погрешности байесовской процедуры задается квадратным корнем, в противном случае (для малых выборок) она не превосходит единицы. Из доказательства теоремы вытекает, что абсолютная константа $a = 4\sqrt{2}$. При выводе этой теоремы не нужно требовать, как в методах минимизации эмпирического риска, равномерную сходимость частот к их вероятностям, поскольку в силу усреднения (15) математическое ожидание и дисперсии частот в формуле (16) совпадают с вероятностями и дисперсиями бернуллиевских случайных величин.

В [12] доказана теорема о том, что если в обучающей выборке отсутствует один из классов, т.е. $\min(m_0, m_1) = 0$, то любая процедура, в том числе и байесовская, работает плохо и ее погрешность строго положительна. Результат этой теоремы показывает, что в оценки погрешности процедур распознавания должны входить размеры классов, а не длина выборки.

Наиболее сложно получить оценку снизу погрешности процедур распознавания на классе C [10, 12].

Теорема 5. Существует абсолютная константа $a_1 > 0$ такая, что справедливо следующее: каковы бы ни были целые числа m_0, m_1, m_2, удовлетворяющие неравенствам $m_1 \ge m_0 \ge 0$, $m_2 \ge 0$, натуральное число n и процедура распознавания Q, существует такое распределение вероятностей P из класса C, что выполняется неравенство

$$v(Q, P) \ge a_1\sqrt{\frac{n}{\min(m_0, m_1)} + \frac{1}{m_2}}. \tag{19}$$

Из теоремы 5 следует, что погрешность $v(Q_B, C)$ отличается от сложности $\mu(C)$ класса задач C не более чем в константу раз. В этом смысле байесовская процедура распознавания Q_B является субоптимальной. Таким образом, байесовская процедура распознавания реализует сложность класса задач C.

Аналогичные теоремы получены для дискретного случая [11, 13].

В отличие от вероятностных оценок (9), (10), полученных в теории минимизации эмпирического риска, оценки погрешности байесовской процедуры распознавания детерминированы, их удобно использовать для практических расчетов.

Таким образом, проблема минимизации среднего риска (13) полностью решена для дискретных объектов с независимыми признаками.

ЭФФЕКТИВНОСТЬ ПРОЦЕДУРЫ РАСПОЗНАВАНИЯ, ПОСТРОЕННОЙ С ПОМОЩЬЮ ОТДЕЛЯЮЩЕЙ ГИПЕРПЛОСКОСТИ

Для булевого случая можно построить отделяющую гиперплоскость на основе байесовской процедуры распознавания. Рассмотрим процедуру отделяющей гиперплоскоси при условии двух классов объектов, при этом все признаки объектов принимают значения из множества $\{0, 1\}$.

Байесовская процедура Q_B на классе $C \equiv C(m_0, m_1, m_2, n)$ определяется по формулам (16), (17). Вектор $d = (d_1, \ldots, d_n)$ относится к классу объектов 0, если выполняется неравенство

$$q_0 \prod_{j=1}^{n} p_{j0d_j} \ge q_1 \prod_{j=1}^{n} p_{j1d_j}, \tag{20}$$

где $q_i = \dfrac{k_i}{m_2}$, $p_{jid_j} = \dfrac{k(d_j, i)}{m_i}$, $i = 0,1$; $j = 1,2,\ldots,n$. Если (20) не выполняется, то вектор d относится к классу объектов 1.

Процедура отделяющей гиперплоскости (обозначим ее R) состоит в следующем. Если выполняется неравенство

$$\sum_{j=1}^{n} \alpha_j d_j + \alpha_0 \geq 0,$$

то вектор d относится к классу объектов 0, в противном случае — к классу объектов 1; здесь $\alpha_0, \alpha_1, \ldots, \alpha_n$ — действительные числа.

Обозначим

$$J_i = \{j:\ 0 < p_{ji1} < 1\},\ i = 0,1;\ I = \{i:\ 0 < q_i < 1\},$$

$$t = \max\left\{ \max_{i\in\{0,1\}} \max_{j\in J_i} \max_{s\in\{0,1\}} |\ln p_{jis}|,\ \max_{i\in I} |\ln q_i| \right\}$$

(здесь $\max\limits_{k\in\varnothing} r_k = 0$), $t_0 = (n+1)t + 1$, $t_1 = (n+1)t_0 + 1$.

Запишем неравенство процедуры отделяющей гиперплоскости следующим образом:

$$\begin{aligned} &\tau_0(q_0) + \sum_{j=1}^{n} [\tau_{0j}(p_{j01}) d_j + \tau_{0j}(p_{j00})(1-d_j)] \geq \\ &\geq \tau_1(q_1) + \sum_{j=1}^{n} [\tau_{1j}(p_{j11}) d_j + \tau_{1j}(p_{j10})(1-d_j)]. \end{aligned} \tag{21}$$

Здесь $\tau_i(z) = \tau_{ij}(z) = \begin{cases} \ln z, & z > 0, \\ -t_i, & z = 0, \end{cases}$ $i = 0,1,\ j = 1,2,\ldots,n.$

Нетрудно доказать, что неравенства (20) и (21) эквивалентны, т.е. для любого булевого вектора d оба неравенства или одновременно выполняются, или одновременно не выполняются. Докажем этот факт для случая, когда выполняется неравенство $0 < q_i < 1$, $0 < p_{ji1} < 1$, $i = 0,1, j = 1,2,\ldots,n.$ Неравенство (21) принимает вид

$$\begin{aligned} &\ln q_0 + \sum_{j=1}^{n} [(\ln p_{j01}) d_j + (\ln p_{j00})(1-d_j)] \geq \\ &\geq \ln q_1 + \sum_{j=1}^{n} [(\ln p_{j11}) d_j + (\ln p_{j10})(1-d_j)], \end{aligned}$$

или $\ln [q_0 \prod\limits_{j=1}^{n} p_{j01}^{d_j} p_{j00}^{1-d_j}] \geq \ln [q_1 \prod\limits_{j=1}^{n} p_{j11}^{d_j} p_{j10}^{1-d_j}]$.

Заметим, что $p_{ji1}^{d_j} p_{ji0}^{1-d_j} = p_{jid_j}$, поэтому последнее неравенство эквивалентно (20). В общем случае эквивалентность неравенств (20) и (21) вытекает из определения чисел t_0, t_1.

Таким образом, байесовская процедура Q_B эквивалентна процедуре отделяющей гиперплоскости R. Отсюда следует субоптимальность процедуры R на классе C, а также то, что на классе C процедуры R и Q_B имеют одинаковую погрешность

$$v(R,C) = v(Z_B, C) \leq a \sqrt{\frac{n}{\min(m_0, m_1)} + \frac{1}{m_2}},$$

где $a < \infty$ — абсолютная константа.

Теорема 6. Байесовская процедура Q_B в булевом случае эквивалентна процедуре распознавания, построенной с помощью отделяющей гиперплоскости R.

Построение отделяющих гиперплоскостей в [1] сводится к решению задачи квадратичного программирования на основе методов обобщенного портрета. В [3] эти методы составляют основу Support Vector Method.

АНАЛИЗ БАЙЕСОВСКОГО ПОДХОДА И МЕТОДОВ МИНИМИЗАЦИИ ЭМПИРИЧЕСКОГО РИСКА

Оценки погрешности методов минимизации эмпирического риска построены для непрерывного случая. Для конечного числа решающих правил (функций распознавания) получение оценок основано на равномерной сходимости частот к вероятностям и сводится к требованию неограниченного увеличения объема выборки. Эмпирический риск строится на некоторой случайной обучающей выборке, поэтому оценки имеют асимптотический вероятностный характер. Получить детерминированные оценки в этой теории невозможно из-за специфики построения эмпирического риска. В процессе минимизации эмпирического риска находится оптимальная функция распознавания. Ее работа затем проверяется на контрольной выборке, процесс построения которой не формализован.

В работах [1–3] предпринята попытка развития теории минимизации эмпирического риска на случай бесконечного числа решающих правил. В ее основе — построение так называемой функции роста, которая вычисляется путем перебора континуального множества непрерывных векторов, составляющих обучающую выборку, и применения к ним бесконечного числа решающих правил. Известно, что подобные определения в математике считаются неконструктивными. Таким образом, получаемые оценки погрешности методов минимизации эмпирического риска в бесконечном случае нельзя считать обоснованными. В целом положения теории методов минимизации эмпирического риска неприменимы для распознавания дискретных объектов, поэтому данная теория не носит общего законченного характера.

В основе байесовского подхода — определение таких структур описания объектов, для которых возможно построение эффективных (оптимальних) процедур распознавания. В настоящее время эти вопросы решены для дискретних объектов с независимыми признаками и объектов, которые описываются цепями Маркова [14]. С помощью операции усреднения по множеству всех возможных обучающих выборок получены детерминированные оценки погрешности байесовской процедуры распознавания в зависимости от размеров классов, количества признаков и их значений. Эти оценки существенно отличаются от оценок методов минимизации эмпирического риска. Операция усреднения по множеству обучающих выборок выполняет функцию контроля.

СПИСОК ЛИТЕРАТУРЫ

1. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов, статистические проблемы обучения. — М.: Наука, 1974. — 416 с.
2. Вапник В.Н. Восстановление зависимостей по эмпирическим данным. — М.: Наука, 1979. — 448 с.
3. Vapnik V. Statistical learning theory. — New York: Wiley, 1998. — 740 p.
4. Сергиенко И.В., Белецкий Б.А., Васильев С.В., Гупал А.М. Предсказание распознавания вторичной структуры белков на основе байесовских процедур на цепях Маркова // Кибернетика и системный анализ. — 2007. — № 2. — С. 59–64.
5. Хаусдорф Ф. Теория множеств. — М.; Л.: Гл. ред. техн.-теорет. лит., 1937. — 304 с.
6. Вероятность и математическая статистика / Гл. ред. Ю.В. Прохоров. — М.: БРЭ, 1999. — 912 с.
7. Бернштейн С.Н. Теория вероятностей. — М.; Л.: Госиздат, 1927. — 364 с.
8. Клайн М. Математика. Утрата определенности. — М.: Мир, 1984. — 548 с.
9. Boucheron S., Bousquet O., Lugosi G. Theory of classification: A survey of some recent advances // ESAIM: Probability and Statistics. — 2005. — N 9. — P. 323–375.
10. Гупал А.М., Пашко С.В., Сергиенко И.В. Эффективность байесовской процедуры распознавания // Кибернетика и системный анализ. — 1995. — № 4. — С. 76–89.
11. Сергиенко И.В., Гупал А.М., Пашко С.В. О сложности задач распознавания образов // Там же. — 1996. — № 4. — С. 70–88.
12. Сергиенко И.В., Гупал А.М. Принципы построения процедур индуктивного вывода // Там же. — 2006. — № 4. — С. 51–63.
13. Белецкий Б.А., Вагис А.А., Васильев С.В., Гупал Н.А. Сложность байесовской процедуры индуктивного вывода. Дискретный случай // Проблемы управления и информатики. — 2006. — № 6. — С. 55–70.
14. Сергиенко И.В., Гупал Н.А. Оптимальные процедуры распознавания и их применение // Кибернетика и системный анализ. — 2007. — № 6. — С. 41–54.

Поступила 22.02.2008