УДК 519.6

ОЦЕНКИ РИСКА В БАЙЕСОВСКОЙ МОДЕЛИ РАСПОЗНАВАНИЯ ПОРЯДКОВОЙ ПЕРЕМЕННОЙ ПО КОНЕЧНОМУ МНОЖЕСТВУ СОБЫТИЙ [1]

© Бериков В.Б.

Институт математики им. С.Л.Соболева СО РАН
пр-т Академика Коптюга, 4, г. Новосибирск, 630090, Россия

e-mail: *berikov@math.nsc.ru*

Abstract. In the paper we consider the ordered regression problem with use of logical decision functions. We suggest the Bayesian model of rank variable recognition on a finite set of events, which is applied for finding the optimal complexity of the class. The evaluations of the risk, obtained by the Bayesian model, are given.

Введение

Одним из наиболее перспективных подходов к решению задач распознавания и прогнозирования является подход, основанный на логико-вероятностных моделях (ЛВМ) [1]. Такого рода модели используются, например, в широко известных методах построения решающих деревьев или логических решающих функций (ЛРФ). *Разработка указанных методов особенно важна для внедрения информационно-вычислительных технологий в различных трудноформализуемых областях исследований (генетика, медицина, экономика и т.д.).* В этих областях существуют некоторые особенности, учет которых в наибольшей степени возможен только при использовании ЛВМ. Этими особенностями являются: недостаточность знаний об изучаемых объектах, что затрудняет формулировку их математических моделей; большое число разнотипных (количественных или качественных) факторов при сравнительно малом объеме данных; нелинейность взаимосвязей; наличие пропусков, погрешностей измерения характеристик; требование представления результатов анализа в форме, понятной специалистам прикладной области.

Проблема построения моделей, обладающих минимальным риском ошибочного прогноза, является одной из важнейших при решении задач как методами, основанными на ЛВМ, так и другими методами анализа данных. Известно, что сложность модели (где под сложностью может пониматься размерность Вапника-Червоненкиса, число логических закономерностей или листьев решающего дерева и т.д.) является существенным фактором, влияющим на качество решений. Для наилучшего качества должен достигаться определенный компромисс между сложностью и точностью решений на обучающей выборке. В достаточно большом круге прикладных задач, наряду с обучающей выборкой, могут быть использованы различного рода экспертные знания, не связанные с жестким заданием модели распределения. При выборе оптимальной сложности ЛВМ возникает проблема совместного учета имеющихся эмпирических данных и экспертных знаний. Провести такой учет позволяет, в частности,

[1]При поддержке международного фонда "Научный потенциал" (грант №144), Российского фонда фундаментальных исследований (гранты №07-01-00331-а, 08-07-00136а)

байесовская теория обучения. *В рамках этого направления в работах* [1, 2, 3] *были разработаны байесовские модели распознавания по конечному множеству событий, созданы алгоритмы построения этих моделей.* Предложенный подход обладает тем преимуществом, что он не ориентирован на самый "неблагоприятный" вид распределения и на асимптотический случай.

В работе рассматривается распространение разработанного подхода на задачу распознавания порядковой переменной. Указанная задача (известная также как порядковая регрессия (ordered regression)) в некотором смысле является "промежуточной" между распознаванием образов и регрессионным анализом. Обзор имеющихся методов решения можно найти, например, в работе [4].

Требуется с использованием байесовской модели оценить оптимальную сложность класса ЛРФ, при которой ожидаемый риск минимален. Для этого необходимо в рамках модели найти риск, усредненный по множеству стратегий природы и по всевозможным обучающим выборкам заданного объема. Кроме того, расматривается задача нахождения апостериорной оценки риска и ее применения как критерия качества решающей функции.

1. Байесовская модель распознавания по конечному множеству событий

При распознавании образов требуется предсказать номер образа ω для произвольного объекта a генеральной совокупности Γ, описываемого набором некоторых переменных $X_1, \ldots, X_n$. При этом предсказание осуществляется на основе анализа случайной обучающей выборки, в которой для каждого объекта указаны значения этих переменных вместе с номером соответствующего образа. Переменные могут быть разнотипными, т.е. часть из них может иметь количественную, а часть – качественную природу. Как правило, для решения задачи используется некоторый класс решающих функций Φ, в котором ищется оптимальная по заданному критерию функция. Класс логических решающих функций определяется на множестве разбиений пространства переменных на конечное число подобластей, описываемых конъюнкциями предикатов простого вида. Число подобластей M определяет степень сложности логической функции.

Байесовская модель распознавания образов по конечному множеству событий вводится путем формулирования некоторых положений, смысл которых состоит в абстрагировании от локальных метрических свойств пространства переменных - перехода от точек пространства к "событиям", где под событием понимается принятие исходными переменными значений из некоторой подобласти разбиения; рассмотрении задачи распознавания по значениям дискретной неупорядоченной переменной; использовании понятия метода обучения как отображения из множества всевозможных выборок во множество решающих функций; привлечении различных способов формализации экспертных знаний о задаче распознавания, не требующих жесткого задания модели распределения вероятностей.

Рассмотрим две дискретные случайные переменные: входную (объясняющую или предикторную) переменную X с множеством неупорядоченных значений $D_X = \{c_1, \ldots, c_j, \ldots, c_M\}$, где c_j – j-е значение (ячейка) и выходную (предсказываемую) переменную Y с множеством неупорядоченных значений $D_Y = \{\omega^{(1)}, \ldots, \omega^{(i)}, \ldots, \omega^{(K)}\}$, где $\omega^{(i)}$ – i-е значение, называемое i-м образом (классом), $K \geq 2$ – число классов. Предполагается, что значения переменной Y упорядочены. Закодируем значения переменной X через номера ячеек, а образы – через соответствующие им номера. Для решения задачи распознавания требуется найти решающую функцию $f : D_X \to D_Y$. Предполагается, что задана ограниченная неотрицательная функция потерь $L_{i,q}$, возникающих в случае принятия решения $Y = i$, когда истинный образ есть q. Будем рассматривать квадратичную функцию потерь такого вида: $L_{i,q} = (i - q)^2$, где $i, q = 1, 2, \ldots, K$.

Пусть на подмножествах множества $D_X \times D_Y$ определена вероятностная мера; $p_j^{(i)}$ - вероятность совместного события "$X = j, Y = i$", при этом выполняется $p_j^{(i)} \geq 0$, $\sum\limits_{i,j} p_j^{(i)} = 1$, $j = 1, 2, \ldots, M$, $i = 1, 2, \ldots, K$. Обозначим $\theta = (\theta_1, \ldots, \theta_j, \ldots, \theta_M)$, где $\theta_j = (p_j^{(1)}, p_j^{(2)}, \ldots, p_j^{(K)})$.

Пусть имеется некоторый класс Φ решающих функций распознавания. Под сложностью класса будем понимать величину M. Каждой решающей функции f из Φ можно сопоставить ожидаемые потери (риск) при распознавании произвольного наблюдения

$$R_f(\theta) = \sum_{j=1}^{M} \sum_{i=1}^{K} p_j^{(i)} (f(j) - i)^2.$$

Если бы вектор θ был известен, можно было бы построить оптимальную решающую функцию ("функцию регрессии") f_r, для которой риск минимален:

$$f_r(j) = l : \sum_{i=1}^{K} p_j^{(i)} (l - i)^2 = \min_{\rho} \sum_{i=1}^{K} p_j^{(i)} (\rho - i)^2,$$

где $\rho = 1, 2, \ldots, K$, $j = 1, 2, \ldots, M$. Легко показать, что значение $f_r(j)$ находится в ближайшей окрестности точки $f^* = \dfrac{1}{p_j} \sum\limits_{i} p_j^{(i)} i$, где $p_j = \sum\limits_{i} p_j^{(i)}$.

В прикладных задачах вектор θ обычно неизвестен. Решающая функция выбирается из Φ на основе случайной выборки наблюдений над X и Y (обучающей выборки) с помощью некоторого *заданного метода* μ. Будем полагать, что результаты наблюдений являются независимыми, одинаково распределенными случайными величинами. Пусть N– объем выборки, $n_j^{(i)}$- число наблюдений i-го образа, соответствующих j-й ячейке; $\sum\limits_{i,j} n_j^{(i)} = N$. Обозначим вектор частот через $s = (s_1, \ldots, s_j, \ldots, s_M)$, где $s_j = (n_j^{(1)}, n_j^{(2)}, \ldots, n_j^{(K)})$. Оценкой эмпирического риска (или просто эмпирическим риском) для решающей функции f назовем величину

$$\widehat{R}_f(s) = \frac{1}{N}\sum_{i,j} n_j^{(i)}(f(j) - i)^2.$$

Метод μ можно рассматривать как функцию, задающую отображение конечного множества всевозможных векторов $\mathbf{S}{=}\{s\}$ во множество решающих функций Φ: $f = \mu(s)$. Метод μ^* минимизации эмпирического риска (аналог метода наменьших квадратов) состоит в следующем: поставить в соответствие j-й ячейке такой номер образа l, что

$$\sum_i n_j^{(i)}(l - i)^2 = \min_\rho \sum_i n_j^{(i)}(\rho - i)^2,$$

где $\rho = 1, 2, \ldots, K$, $j = 1, 2, \ldots, M$. Обозначим через $\widehat{f_r}$ решающую функцию ("эмпирическую регрессию"), полученную с помощью этого метода. Нетрудно убедиться, что $\widehat{f_r}(j)$ равно округленному до ближайшего целого среднему значению переменной Y по j-й ячейке: $\widehat{f}^* = \frac{1}{n_j}\sum\limits_i n_j^{(i)} i$, где $n_j = \sum\limits_i n_j^{(i)}$. Условимся, для определенности, что в случае, когда округление не может быть проведено однозначно, выбирать наибольший номер класса. Если некоторая ячейка оказалась "пустой", т.е. не содержащей ни одного наблюдения, договоримся приписывать такой ячейке значение Y, равное K.

Пусть S – случайный вектор частот. Этот вектор подчиняется полиномиальному распределению $\mathbf{P}(S = s|\theta) = p(s|\theta)$, где

$$p(s|\theta) = \frac{N!}{\prod\limits_{i,j} n_j^{(i)}!}\prod_{i,j}\left(p_j^{(i)}\right)^{n_j^{(i)}}.$$

Рассмотрим семейство полиномиальных моделей распределения вектора частот с множеством параметров $\Lambda = \{\theta\}$. Будем использовать байесовский подход: предположим, что на множестве Λ определена случайная величина $\Theta = (P_1^{(1)}, \ldots, P_M^{(K)})$ с некоторым известным априорным распределением $p(\theta)$ при $\theta \in \Lambda$. В этом случае риск является функцией $R_f(\Theta)$, зависящей от случайного вектора параметров модели. Значение этой функции можно рассматривать как оценку неизвестного риска, вычисленную в предположении о том, что истинный вектор параметров равен θ.

Будем полагать, что величина Θ подчиняется распределению Дирихле $\Theta \sim Dir(d_1^{(1)}, d_1^{(2)}, \ldots, d_M^{(K)})$: $p(\theta) = \frac{1}{Z}\prod\limits_{i,j}(p_j^{(i)})^{d_j^{(i)}-1}$, где $d_j^{(i)} > 0$ – некоторые заданные вещественные числа, выражающие априорные знания о распределении Θ, $i = 1, 2, \ldots, K$, $j = 1, 2, \ldots M$, Z – нормализующая константа: $Z = \frac{\prod\limits_{i,j}\Gamma(d_j^{(i)})}{\Gamma(D)}$, где $\Gamma(\cdot)$ – гамма-функция, $D{=}\sum\limits_{i,j} d_j^{(i)}$. При $d_j^{(i)} \equiv 1$ получим случай равномерного априорного распределения, использование которого оправдано при отсутствии априорной информации о классе распределений Λ.

Выбор априорного распределения Дирихле объясняется тем, что оно удобно для выражения априорных знаний о распределении на множестве стратегий природы: величины $d_j^{(i)}$ аналогичны числу попаданий в ячейки наблюдений различных образов. Это распределение сопряжено с рассматриваемым полиномиальным распределением вектора частот. В работе [2] было показано, что при уменьшении параметра Дирихле понижается "степень пересечения" между образами, понимаемая как ожидаемое значение вероятности ошибки оптимальной байесовской решающей функции; было найдено выражение для указанной зависимости, согласно которому можно определять значение параметра по экспертной оценке степени пересечения.

В дальнейшем, при формулировке теоретических утверждений будем считать, что *выполняются все вышеозначенные предположения.*

2. Ожидаемый риск для метода обучения

Найдем математическое ожидание функции риска, где усреднение проводится по всевозможным значениям параметров из множества Λ и всевозможным обучающим выборкам заданного объема N. Заметим, что в методе обучения при принятии решения может учитываться не только эмпирическая ошибка. Например, если есть экспертные знания о предпочтении какого-либо образа, решение может приниматься в его пользу несмотря на увеличение ошибки.

Теорема 1. *Пусть $f = \mu(s)$ – решающая функция, полученная по выборке s с помощью некоторого детерминированного метода μ, такого, что решение для j-й ячейки принимается по набору частот s_j, $j = 1, \ldots, M$. Тогда математическое ожидание $\mathbf{E}R_{\mu(S)}(\Theta)$ функции риска равно*

$$R_\mu = \frac{\Gamma(D)N!}{\Gamma(D+N+1)} \sum_{j=1}^{M} \frac{1}{\Gamma(D-d_j)\prod\limits_{l=1}^{K}\Gamma(d_j^{(l)})} \times$$

$$\times \sum_{s_j} \frac{\Gamma(\overline{n}_j + D - d_j)}{\prod\limits_l n_j^{(l)}!\,\overline{n}_j!} \prod_l \Gamma(d_j^{(l)} + n_j^{(l)}) \sum_{q=1}^{K} (f(j)-q)^2 (d_j^{(q)} + n_j^{(q)}),$$

где $d_j = \sum\limits_l d_j^{(l)}$, $\overline{n}_j = N - \sum\limits_l n_j^{(l)}$, оператор $\sum\limits_{s_j}$ означает, что суммирование ведется по всем наборам $(n_j^{(1)}, n_j^{(2)}, \ldots, n_j^{(K)})$ таким, что $\sum\limits_l n_j^{(l)} \le N$.

Доказательство. Из определения функции риска следует, что

$$\mathbf{E}R_{\mu(S)}(\Theta) = \int\limits_\Lambda \sum_s p(s|\theta)p(\theta) \sum_{j,q} (f(j)-q)^2 p_j^{(q)} d\theta =$$

$$= \frac{1}{Z} \sum_{j,q} \int\limits_\Lambda \prod_{l,m} (p_m^{(l)})^{d_m^{(l)}-1} p_j^{(q)} \sum_{s_j} (f(j)-q)^2 \sum_{\substack{s:\\ s_j\ \text{фиксир.}}} p(s|\theta) d\theta =$$

$$= \frac{1}{Z}\sum_{j,q}\int\limits_{\Lambda}\prod_{l,m}(p_m^{(l)})^{d_m^{(l)}-1}p_j^{(q)}\sum_{s_j}(f(j)-q)^2 p(s_j|\theta_j)d\theta$$

по свойству полиномиального распределения. Здесь $p(s_j|\theta_j)$ – функция вероятности распределения величины S_j: $p(s_j|\theta_j) = \frac{N!}{\prod\limits_l n_j^{(l)}!\overline{n}_j!}\prod\limits_l (p_j^{(l)})^{n_j^{(l)}}(\overline{p}_j)^{\overline{n}_j}$, $\overline{p}_j = 1 - \sum\limits_l p_j^{(l)}$. Далее, $\mathbf{E}R_{\mu(S)}(\Theta) =$

$$= \frac{1}{Z}\sum_{j,q}\sum_{s_j}\frac{N!}{\prod\limits_l n_j^{(l)}!\overline{n}_j!}\int\limits_{\substack{p_j^{(1)},\ldots p_j^{(K)}:\\ \sum_i p_j^{(i)}\le 1}}\prod_l (p_j^{(l)})^{d_j^{(l)}-1}p_j^{(q)}(f(j)-q)^2\prod_l (p_j^{(l)})^{n_j^{(l)}}(\overline{p}_j)^{\overline{n}_j}\times$$

$$\times\left\{\int\limits_{\substack{\{p_m^{(l)}\}:\\ \sum\limits_{\substack{l,m:\\ m\ne j}} p_m^{(l)}=1-\overline{p}_j}}\prod_{\substack{l,m:\\ m\ne j}}(p_m^{(l)})^{d_m^{(l)}-1}d\theta_1\ldots d\theta_{j-1}d\theta_{j+1}\ldots d\theta_M\right\}d\theta_j.$$

Воспользуемся формулой, обобщающей формулу Дирихле [5]:

$$\int\limits_{\substack{\{x_1,\ldots,x_{m-1}:\\ x_i\ge 0\\ \sum\limits_{i=1}^{m-1}x_i\le h\}}}\prod_{i=1}^{m-1}x_i^{d_i-1}(h-\sum_i x_i)^{d_m-1}dx_1\ldots dx_{m-1} = \frac{\prod\limits_{i=1}^{m}\Gamma(d_i)}{\Gamma(\sum\limits_{i=1}^{m}d_i)}h^{\sum\limits_{i=1}^{m}d_i-1},$$

где $d_1,\ldots,d_m$ – вещественные неотрицательные числа. Отсюда

$$\mathbf{E}R_{\mu(S)}(\Theta) = \frac{1}{Z}\sum_{j,q}\sum_{s_j}\frac{N!}{\prod\limits_l n_j^{(l)}!\overline{n}_j!}\int\limits_{\substack{p_j^{(1)},\ldots,p_j^{(K)}:\\ \sum_i p_j^{(i)}\le 1}}\prod_l (p_j^{(l)})^{n_j^{(l)}+d_j^{(l)}-1}p_j^{(q)}(f(j)-q)^2\times$$

$$\times(\overline{p}_j)^{\overline{n}_j}\frac{\prod\limits_{\substack{l,m:\\ m\ne j}}\Gamma(d_m^{(l)})}{\Gamma(D-d_j)}(\overline{p}_j)^{D-d_j-1}d\theta_j = \frac{1}{Z}\sum_{j,q}\sum_{s_j}\frac{N!}{\prod\limits_l n_j^{(l)}!\overline{n}_j!}(f(j)-q)^2\frac{\prod\limits_{\substack{l,m:\\ m\ne j}}\Gamma(d_m^{(l)})}{\Gamma(D-d_j)}\times$$

$$\times\frac{\prod\limits_{l:\,l\ne q}\Gamma(d_j^{(l)}+n_j^{(l)})\Gamma(d_j^{(q)}+n_j^{(q)}+1)\Gamma(\overline{n}_j+D-d_j)}{\Gamma(N+D+1)}$$

Воспользуемся свойством гамма-функции: $\Gamma(x+1) = \Gamma(x)x$, получим

$$\mathbf{E}R_{\mu(S)}(\Theta) = \frac{\Gamma(D)}{\prod\limits_{l,m}\Gamma(d_m^{(l)})}\sum_j\sum_{s_j}\frac{N!}{\prod\limits_l n_j^{(l)}!\overline{n}_j!}\frac{\prod\limits_{\substack{l,m:\\ m\neq j}}\Gamma(d_m^{(l)})}{\Gamma(N+D+1)\Gamma(D-d_j)}\times$$
$$\times\prod_l\Gamma(d_j^{(l)}+n_j^{(l)})\Gamma(\overline{n}_j+D-d_j)\sum_q(d_j^{(q)}+n_j^{(q)})(f(j)-q)^2.$$

Отсюда следует справедливость теоремы 1.

3. Нахождение оптимальной сложности класса логических решающих функций

Покажем, как в рамках байесовской модели можно находить оптимальную сложность класса ЛРФ. Образуем некоторое исходное разбиение пространства переменных $X_1,\ldots,X_n$ на подобласти $E_1,E_2,\ldots,E_{M_0}$. Зададим соответствующую байесовскую модель распознавания по множеству из M_0 событий. Сформируем также множество разбиений, получаемых путем объединения некоторых подобластей исходного разбиения. Пусть произошло объединение подобластей $E_1,\ldots,E_{j_1}$ в новую подобласть $\widetilde{E}_1$, подобластей $E_{j_1+1},\ldots,E_{j_1+j_2}$ в подобласть $\widetilde{E}_2$, …, подобластей $E_{j_1+\cdots+j_{M-1}+1},\ldots,E_{j_1+\cdots+j_M}$ в подобласть $\widetilde{E}_M$. Обозначим $\widetilde{p}_t^{(i)} = \mathbf{P}(Y=i,\widetilde{E}_t)$, где $t=1,2,\ldots,M$, $i=1,2,\ldots,K$. Тогда $\widetilde{p}_1^{(i)} = p_1^{(i)}+\cdots+p_{j_1}^{(i)}$, $\widetilde{p}_2^{(i)} = p_{j_1+1}^{(i)}+\cdots+p_{j_1+j_2}^{(i)}$, …, $\widetilde{p}_M^{(i)} = p_{j_1+\cdots+j_{M-1}+1}^{(i)}+\cdots+p_{j_1+\cdots+j_M}^{(i)}$. По свойству распределения Дирихле [6], вектор $\widetilde{\Theta} = (\widetilde{P}_1^{(1)},\ldots,\widetilde{P}_t^{(i)},\ldots,\widetilde{P}_M^{(K)}) \sim Dir(d_1^{(1)}+\cdots+d_{j_1}^{(1)},\ldots,d_{j_1+\cdots+j_{t-1}+1}^{(i)}+\cdots+d_{j_1+\cdots+j_t}^{(i)},\ldots,d_{j_1+\cdots+j_{M-1}+1}^{(K)}+\cdots+d_{j_1+\cdots+j_M}^{(K)})$.

Теорема 1 позволяет вычислить ожидаемый риск для каждого из вариантов разбиения (число разбиений определяет сложность класса ЛРФ и, одновременно, сложность соответствующей ЛВМ). Будем перебирать различные варианты и найдем такой из них, для которого ожидаемый риск минимален.

Рассмотрим следующий пример. Пусть все подобласти сгруппированы в пары; каждый последующий вариант разбиения образуется из предыдущего путем слияния пар. Сложность исходной модели $M_0=2^u$, где u - целое число. Все параметры Дирихле исходной модели совпадают (что говорит об отсутствии априорных предпочтений между подобластями) и равны некоторой величине d_0. Таким образом, имеем последовательность разбиений сложности $M=M_0,\ldots,4,2,1$, а также набор байесовских моделей с параметрами $d=d_0,2d_0,4d_0,\ldots,M_0d_0$. Пусть метод μ^* минимизирует эмпирическую ошибку. На рис. 1 приведен пример полученного графика зависимости ожидаемого риска от сложности M=1; 2; 4; 8 для объема обучающей выборки N=11. Параметр d_0=0.1, $K=6$.

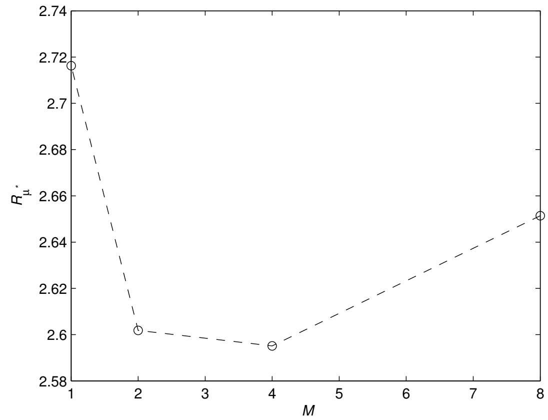

Рис. 1. Зависимость ожидаемого риска R_{μ^*} от сложности класса M.

4. Апостериорный риск

В данном параграфе исследуется оценка функции риска для решающей функции, полученной по заданной выборке. Согласно байесовскому подходу, функция риска $R_f(\Theta)$ при фиксированной выборке является случайной величиной, зависящей от случайного вектора Θ. Рассмотрим апостериорную плотность распределения $p(\theta|S = s) = p(\theta|s)$, где $s \in \mathbf{S}$. По свойству распределения Дирихле [6], $\Theta|s \sim Dir(d_1^{(1)} + n_1^{(1)}, \ldots, d_j^{(i)} + n_j^{(i)}, \ldots, d_M^{(K)} + n_M^{(K)})$.

Теорема 2. *Пусть f – произвольная решающая функция, выбранная из множества решающих функций Φ, и задана выборка $s \in \boldsymbol{S}$. Тогда апостериорное математическое ожидание функции риска у решающей функции f равно*

$$\mathbf{E}(R_f(\Theta)|s) = \frac{1}{N+D}\sum_{j,q}(f(j)-q)^2(n_j^{(q)} + d_j^{(q)}).$$

Доказательство. По определению функции риска,

$$\mathbf{E}(R_f(\Theta)|s) = \mathbf{E}_{\Theta|s}\sum_{j,q}(f(j)-q)^2 P_j^{(q)} = \sum_{j,q}(f(j)-q)^2(\mathbf{E}_{\Theta|s}P_j^{(q)}) =$$

$$= \sum_{j,q}(f(j)-q)^2\int\limits_{\Lambda} p(\theta|s)p_j^{(q)}d\theta = \sum_{j,q}\frac{\Gamma(N+D)}{\prod\limits_{m,l}\Gamma(d_m^{(l)}+n_m^{(l)})}(f(j)-q)^2\times$$

$$\times\int\limits_{\Lambda}\prod_{m,l}(p_m^{(l)})^{d_m^{(l)}+n_m^{(l)}-1}p_j^{(q)}d\theta =$$

$$= \sum_{j,q}\frac{\Gamma(N+D)}{\prod\limits_{m,l}\Gamma(d_m^{(l)}+n_m^{(l)})}(f(j)-q)^2\frac{\prod\limits_{\substack{m,l:\\(m,l)\neq(j,q)}}\Gamma(d_m^{(l)}+n_m^{(l)})\Gamma(d_j^{(q)}+n_j^{(q)}+1)}{\Gamma(N+D+1)} =$$

$$= \frac{1}{N+D} \sum_{j,q} (f(j) - q)^2 (n_j^{(q)} + d_j^{(q)}),$$

что и требовалось доказать.

Назовем найденную выше величину $\mathbf{E}(R_f(\Theta)|s)$ байесовской оценкой риска, соответствующей решающей функции f. Полученная оценка может использоваться на этапе обучения как критерий оптимальности решающей функции.

Для поиска оптимальной ЛРФ может применяться следующий способ. Выборка разбивается на две части; по первой части строится дерево регрессии, возможно переобученное, т.е. имеющее излишнюю сложность в результате чрезмерной "настройки" на объекты обучающей выборки. По полученному дереву формируется множество событий, связанных с разбиениями пространства переменных в соответствии со структурой дерева. Затем вторая часть выборки используется для оценки качества распознавания с применением байесовской модели распознавания по конечному множеству событий. При этом рассматриваются различные варианты усечения дерева; ищется такой вариант, для которого ожидаемый апостериорный риск минимален.

Заключение

В работе впервые предлагается использовать байесовскую модель распознавания по конечному множеству событий в задаче порядковой регрессии. В рамках байесовской модели найдено выражение для риска, усредненного по множеству стратегий природы и по всевозможным обучающим выборкам заданного объема. Предложен способ определения оптимальной сложности класса логических решающих функций, при которой ожидаемый риск минимален. Найдена апостериорная оценка риска, предложен способ ее применения при построении логической решающей функции.

В перспективе предполагается *исследовать и другие виды априорных распределений в байесовской модели, а также применить полученные результаты для решения таких задач анализа данных, как группировка объектов и адаптивное построение логических решающих функций.*

Список литературы

1. *Лбов Г.С., Бериков В.Б.* Устойчивость решающих функций в задачах распознавания образов и анализа разнотипной информации. Новосибирск: Изд-во Ин-та математики, 2005.
2. *Berikov V.B.* Bayes estimates for recognition quality on a finite set of events // Pattern Recognition and Image Analysis. 2006. V. 16, N 3. P. 329–343.
3. *Бериков В.Б., Лбов Г.С.* Выбор оптимальной сложности класса логических решающих функций в задачах распознавания образов // Доклады Академии наук. 2007. Том 417, №1. С. 26-29.
4. *Kramer S., Widmer G., Pfahringer B., DeGroeve M.* Prediction of ordinal classes using regression trees // Fundamenta Informaticae. 2000. V. 34. P.1-15.
5. *Фихтенгольц Г.М.* Курс дифференциального и интегрального исчисления. Т. 3. М.: Физматлит, 1960.
6. *Уилкс С.* Математическая статистика. М.: Наука, 1967.

Статья поступила в редакцию 30.04.2008