Оценки вероятности ошибки в байесовской логико-вероятностной модели распознавания образов*

В. Б. Бериков
Институт математики им. С.Л. Соболева СО РАН, Новосибирск, Россия
e-mail: berikov@math.nsc.ru

In this paper, we consider Bayes logical-and-probabilistic model of pattern recognition on a finite set of events, where event is defined as an assignment of some values from a subregion of the partition space described by a logical statement, to the given heterogeneous variables. Using this model, we found an expression for the expected probability of errors for the optimal Bayes decision function, obtain a posteriori estimations of the classifications probability for a given sample. These estimations are used for a construction and studying of the logical decision function.

Введение

Одним из известных подходов к решению задач интеллектуального анализа данных является подход, основанный на классе логических решающих функций (ЛРФ; наиболее часто используемая форма ЛРФ — дерево решений). Применение ЛРФ позволяет решать задачи, характеризуемые разнотипностью переменных, малым объемом данных и наличием в них пропущенных значений; результаты анализа представляются в форме логических закономерностей, легко интерпретируемых специалистом прикладной области.

При использовании методов, основанных на ЛРФ (как, впрочем, и других методов анализа данных), возникает проблема выбора оптимальной сложности класса решающих функций. Известно, что сложность класса (где под сложностью может пониматься размерность Вапника—Червоненкиса, максимальное число логических закономерностей или листьев решающего дерева и т. д.) — существенный фактор, влияющий на качество решений. Для наилучшего качества (т. е. минимального риска неправильного распознавания новых объектов генеральной совокупности) должен достигаться определенный компромисс между сложностью класса 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 закона распределения характеристик объектов. При выборе оптимальной сложности ЛРФ возникает проблема совместного учета имеющихся эмпирических данных и экспертных знаний. Провести такой учет позволяет, в частности, байесовская теория обучения. В рамках этого направления были предложены байесовские логико-вероятностные модели распознавания по конечному множеству событий, разработаны

*Работа выполнена при финансовой поддержке фонда "Научный потенциал" (грант № 144) и Российского фонда фундаментальных исследований (гранты № 08-07-00136а и № 07-01-00331а).

соответствующие алгоритмы построения ЛРФ [1 – 3]. Эти модели обладают той особенностью, что они не ориентированы на самый "неблагоприятный" вид распределения и на асимптотический случай.

В работах [1 – 3] были исследованы некоторые свойства предложенных моделей: в частности, найдено выражение для ожидаемой вероятности ошибки оптимальной байесовской решающей функции для случая, когда отсутствуют априорные предпочтения на множестве событий; получены апостериорные точечные и интервальные оценки риска для исследуемых моделей. Найденные оценки могут применяться для нахождения риска неправильного прогнозирования новых объектов, а также как критерии качества логических решающих функций. При этом интервальные оценки предпочтительнее в случае значительной дисперсии ошибки, что характерно для малого объема выборки. Вывод интервальных оценок основывался на применении вероятностных неравенств чебышевского типа. В предлагаемой работе ставится задача распространения полученных результатов на случай, когда на множестве событий имеются априорные предпочтения; а также повышения качества найденных оценок путем использования более точных вероятностных неравенств.

1. Байесовская логико-вероятностная модель распознавания образов по конечному множеству событий

При распознавании образов требуется предсказать номер класса для произвольного объекта генеральной совокупности, описываемого набором некоторых переменных. При этом предсказание осуществляется на основе анализа 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, в котором ищется оптимальная по заданному критерию функция. Класс логических решающих функций [4] определяется на множестве разбиений пространства переменных на конечное число подобластей, описываемых конъюнкциями предикатов простого вида. Число подобластей определяет степень сложности логической функции.

Байесовская логико-вероятностная модель распознавания образов по конечному множеству событий вводится путем абстрагирования от локальных метрических свойств пространства переменных (перехода от точек пространства к "событиям", где под событием понимается принятие исходными переменными значений из некоторой подобласти разбиения, описываемой соответствующим логическим высказыванием); рассмотрения задачи распознавания по конечному множеству событий (или по значениям дискретной неупорядоченной переменной); сопоставления каждому из возможных вероятностных распределений, принадлежащих заданному классу, некоторого веса, отражающего интуитивную уверенность эксперта в том, что неизвестное истинное распределение совпадает с рассматриваемым. Для такого сопоставления могут привлекаться различные способы формализации экспертных знаний о задаче распознавания, не требующие жесткого задания модели распределения.

Рассматриваемая модель распознавания удобна для теоретического изучения, так как снимаются проблемы, связанные с многомерностью пространства переменных (сложный вид модели, большое число параметров и т. д.) и их разнотипностью. Произвольное распределение в исходном пространстве переменных аппроксимируется некоторым полиномиальным распределением, определенным на подобластях разбиения. Точность аппроксимации можно повышать, увеличивая число подобластей. Логические решающие функции определены на разбиениях пространства переменных, и поэтому при изучении ЛРФ естественно использовать модель распознавания по таким подобластям.

Итак, рассмотрим две дискретные случайные переменные: переменную X со множеством неупорядоченных значений $D_X = \{c_1, \ldots, c_j, \ldots, c_M\}$, где c_j — j-е значение ("ячейка"), и переменную Y с множеством неупорядоченных значений $D_Y = \{\omega^{(1)}, \ldots, \omega^{(K)}\}$, называемых образами. Закодируем значения переменной X через номера ячеек, а образы — через соответствующие им номера. Пусть $p_j^{(i)}$ — вероятность совместного события "$X = j, Y = i$", где $j = 1, \ldots, M$, $i = 1, \ldots, K$, $\sum\limits_{i,j} p_j^{(i)} = 1$. Обозначим вектор $(p_j^{(1)}, \ldots, p_j^{(K)})$ через $\boldsymbol{\theta}_j$, а вектор $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M)$ через $\boldsymbol{\theta}$. Предполагается, что задана индикаторная функция потерь $L_{r,l}$, возникающих в случае принятия решения $Y = r$, когда истинный образ есть l: $L_{r,l} = 0$ при $r = l$ и $L_{r,l} = 1$ при $r \neq l$. Пусть имеется некоторый класс Φ решающих функций распознавания, т. е. отображений $D_X \to D_Y$. Величину M будем называть сложностью класса. Каждой решающей функции f из Φ можно сопоставить ожидаемые потери (риск или вероятность ошибки) при распознавании произвольного наблюдения: $P_f(\boldsymbol{\theta}) = \sum\limits_{i,j} L_{f(j),i} p_j^{(i)}$. Функция f выбирается из Φ некоторым алгоритмом на основе анализа обучающей выборки $\mathbf{s}$ наблюдений над X и Y, где $\mathbf{s} = (n_1^{(1)}, \ldots, n_M^{(K)})$, $n_j^{(i)}$ — частота наблюдений i-го образа, соответствующих j-й ячейке; $\sum\limits_{i,j} n_j^{(i)} = N$, N есть объем выборки.

Пусть $\mathbf{S}$ — случайный вектор частот. Рассмотрим семейство полиномиальных моделей распределения вектора частот с множеством параметров $\Lambda = \{\boldsymbol{\theta}\}$. Это семейство (класс распределений) будем также называть множеством стратегий природы. Под сложностью класса распределений понимается величина M. Используем байесовский подход: предположим, что на Λ определена случайная величина $\boldsymbol{\Theta} = (P_1^{(1)}, \ldots, P_M^{(K)})$ с некоторой известной плотностью априорного распределения $p(\boldsymbol{\theta})$ при $\boldsymbol{\theta} \in \Lambda$. Будем полагать, что $\boldsymbol{\Theta}$ подчиняется распределению Дирихле с параметрами $d_1^{(1)}, \ldots, d_M^{(K)}$: $\boldsymbol{\Theta} \sim \mathrm{Dir}(d_1^{(1)}, \ldots, d_M^{(K)})$, т. е.

$$p(\boldsymbol{\theta}) = \frac{1}{Z} \prod_{i,j} (p_j^{(i)})^{d_j^{(i)} - 1},$$

где $d_j^{(i)} > 0$ — заданные вещественные числа, $i = 1, \ldots, K$, $j = 1, \ldots M$,

$$Z = \frac{\prod\limits_{i,j} \Gamma\left(d_j^{(i)}\right)}{\Gamma\left(\sum\limits_{i,j} d_j^{(i)}\right)}$$

— нормализующая константа, $\Gamma(\cdot)$ — гамма-функция. Параметры $d_j^{(i)}$ аналогичны числу попаданий в ячейки наблюдений различных образов и выражают априорные предпочтения между ячейками.

В случае, когда априорные предпочтения отсутствуют, можно полагать, что $d_j^{(i)} \equiv d$. При $d = 1$ получим случай равномерного априорного распределения, который означает, что все стратегии природы имеют равные шансы на осуществление. Если $d \neq 1$, то это означает, что нет априорной информации о предпочтении одних ячеек перед другими, но при этом априорное распределение на множестве стратегий природы — неравномерное. Как показано в [1, 2], априорное распределение такого вида особенно удобно для выражения экспертных знаний, имеющих вид оценки степени "пересечения" между образами. При увеличении параметра d априорное распределение меняется так, что образы в среднем более "пересекаются" (т. е. увеличивается ожидаемая по стратегиям природы вероятность ошибки, соответствующая оптимальной байесовской решающей функции). Это свойство позволяет задать величину d по предполагаемой экспертом степени пересечения между образами.

2. Ожидаемая вероятность ошибки байесовской решающей функции

Если бы вектор $\boldsymbol{\theta}$ был известен, можно было бы построить оптимальную байесовскую решающую функцию f_B, для которой вероятность ошибки минимальна:

$$f_B(j) = l : \sum_{\substack{i=1,\\ i\neq l}}^{K} p_j^{(i)} = \min_{\rho} \sum_{\substack{i=1,\\ i\neq \rho}}^{K} p_j^{(i)},$$

где $\rho = 1, \ldots, K$, $j = 1, \ldots, M$. Далее в этом разделе рассматривается случай задачи распознавания двух образов, когда вероятность ошибки распознавания байесовской решающей функции равна

$$P_{f_B}(\boldsymbol{\theta}) = \sum_j \min\{p_j^{(1)}, p_j^{(2)}\}.$$

Рассмотрим ожидаемую по стратегиям природы вероятность ошибки:

$$\mathbb{E}P_{f_B}(\boldsymbol{\Theta}) = \int_{\Lambda} P_{f_B}(\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

Теорема 1. *При сформулированных выше предположениях выполняется*

$$\mathbb{E}P_{f_B}(\boldsymbol{\Theta}) = W(d_1^{(1)}, \ldots, d_M^{(2)}),$$

где $W(d_1^{(1)}, \ldots, d_M^{(2)}) = \dfrac{1}{D} \sum_j \{d_j^{(1)} I_{0,5}(d_j^{(1)} + 1, d_j^{(1)}) + d_j^{(2)} I_{0,5}(d_j^{(2)} + 1, d_j^{(2)})\}$, $D = \sum\limits_{i,j} d_j^{(i)}$, $I_x(p, q)$ *— бета-функция распределения с параметрами* x, p, q.

Доказательство. Так как $\mathbb{E}P_{f_B}(\boldsymbol{\Theta}) = \int\limits_{\Lambda} \sum\limits_j \min\{p_j^{(1)}, p_j^{(2)}\} p(\boldsymbol{\theta}) d\boldsymbol{\theta}$, то получим

$$\mathbb{E}P_{f_B}(\boldsymbol{\Theta}) = \frac{1}{Z}\sum_j \int\limits_{\Lambda} \min\{p_j^{(1)}, p_j^{(2)}\} \prod_{l,r} (p_r^{(l)})^{d_r^{(l)}-1} d\boldsymbol{\theta} =$$

$$= \frac{1}{Z}\sum_j \int\limits_{\substack{\{p_j^{(1)}, p_j^{(2)}: \\ p_j^{(1)}+p_j^{(2)} \le 1\}}} \min\{p_j^{(1)}, p_j^{(2)}\} (p_j^{(1)})^{d_j^{(1)}-1} (p_j^{(2)})^{d_j^{(2)}-1} \times$$

$$\times \left\{ \int\limits_{\substack{\{p_r^{(l)}:\, r\neq j, \\ \sum\limits_{l,r} p_r^{(l)} = 1 - p_j^{(1)} - p_j^{(2)}\}}} \prod_{\substack{l,r: \\ r\neq j}} (p_r^{(l)})^{d_r^{(l)}-1} \prod_{r\neq j} d\boldsymbol{\theta}_r \right\} dp_j^{(1)} dp_j^{(2)} =$$

$$= \frac{1}{Z}\sum_j \int\limits_{\substack{\{p_j^{(1)}, p_j^{(2)}: \\ p_j^{(1)}+p_j^{(2)} \le 1\}}} \min\{p_j^{(1)}, p_j^{(2)}\} (p_j^{(1)})^{d_j^{(1)}-1} (p_j^{(2)})^{d_j^{(2)}-1} \frac{\prod\limits_{\substack{l,r: \\ r\neq j}} \Gamma(d_r^{(l)})}{\Gamma(\sum\limits_{\substack{l,r: \\ r\neq j}} d_r^{(l)})} \times$$

$$\times (1 - p_j^{(1)} - p_j^{(2)})^{\sum\limits_{\substack{l,r: \\ r\neq j}} d_r^{(l)} - 1} dp_j^{(1)} dp_j^{(2)}.$$

Здесь мы воспользовались формулой, обобщающей формулу Дирихле [5]:

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

где $d_1, \ldots, d_m$ — вещественные неотрицательные числа.

Рассмотрим вспомогательную лемму.

Лемма 1. *Пусть p, q, r — вещественные неотрицательные числа и*

$$\chi(p,q,r) = \int\limits_{\substack{\{x,y:\, x+y\le 1, \\ y<x,\, x\ge 0,\, y\ge 0\}}} x^{p-1} y^{q-1} (1-x-y)^{r-1} dx dy.$$

Тогда

$$\chi(p,q,r) = B(p+q,r) B_{0,5}(q,p),$$

где $B_x(p,q) = \int\limits_0^x t^{p-1}(1-t)^{q-1}dt$ — неполная бета-функция, $B(p,q)$ — бета-функция: $B(p,q) = B_1(p,q) = \dfrac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$.

Доказательство. Для доказательства применяется способ, описанный в [5, с. 213]. Воспользуемся следующей заменой переменных:

$$x = u(1-t), \quad y = ut.$$

Тогда

$$\chi(p,q,r) = \int\limits_0^1 du \int\limits_0^{0.5} u^{p-1}(1-t)^{p-1}u^{q-1}t^{q-1}(1-u)^{r-1}u dt =$$

$$= \int\limits_0^1 u^{p+q-1}(1-u)^{r-1}du \int\limits_0^{0.5} t^{q-1}(1-t)^{p-1}dt = B(p+q,r)B_{0.5}(q,p),$$

что и требовалось доказать. □

Вернемся к вычислению ожидаемой вероятности ошибки. Имеем

$$\mathbb{E}P_{f_B}(\boldsymbol{\Theta}) = \sum_j \frac{\Gamma(D)}{\Gamma(d_j^{(1)})\Gamma(d_j^{(2)})\Gamma(D-d_j^{(1)}-d_j^{(2)})} \times$$

$$\times \left\{ \int\limits_{\substack{\{p_j^{(1)},p_j^{(2)}:\, p_j^{(1)}<p_j^{(2)} \\ p_j^{(1)}+p_j^{(2)}\le 1\}}} (p_j^{(1)})^{d_j^{(1)}}(p_1^{(2)})^{d_j^{(2)}-1}(1-p_1^{(1)}-p_1^{(2)})^{D-d_j^{(1)}-d_j^{(2)}-1}dp_j^{(1)}dp_j^{(2)} + \right.$$

$$\left. + \int\limits_{\substack{\{p_j^{(1)},p_j^{(2)}:\, p_j^{(2)}<p_j^{(1)} \\ p_j^{(1)}+p_j^{(2)}\le 1\}}} (p_j^{(1)})^{d_j^{(1)}-1}(p_1^{(2)})^{d_j^{(2)}}(1-p_1^{(1)}-p_1^{(2)})^{D-d_j^{(1)}-d_j^{(2)}-1}dp_j^{(1)}dp_j^{(2)} \right\}.$$

Воспользовавшись доказанной леммой, получим

$$\mathbb{E}P_{f_B}(\boldsymbol{\Theta}) = \sum_j \frac{\Gamma(D)}{\Gamma(d_j^{(1)})\Gamma(d_j^{(2)})\Gamma(D-d_j^{(1)}-d_j^{(2)})} \times$$

$$\times \Big\{ B(d_j^{(1)}+d_j^{(2)}+1, D-d_j^{(1)}-d_j^{(2)})B_{0,5}(d_j^{(1)}+1, d_j^{(2)}) +$$

$$+ B(d_j^{(1)}+d_j^{(2)}+1, D-d_j^{(1)}-d_j^{(2)})B_{0,5}(d_j^{(2)}+1, d_j^{(1)}) \Big\} =$$

$$= \sum_j \frac{\Gamma(D)}{\Gamma(d_j^{(1)})\Gamma(d_j^{(2)})\Gamma(D-d_j^{(1)}-d_j^{(2)})} B(d_j^{(1)}+d_j^{(2)}+1, D-d_j^{(1)}-d_j^{(2)}) \times$$

$$\times \{B_{0,5}(d_j^{(1)}+1, d_j^{(2)}) + B_{0,5}(d_j^{(2)}+1, d_j^{(1)})\} =$$

$$= \sum_j \frac{\Gamma(D)}{\Gamma(d_j^{(1)})\Gamma(d_j^{(2)})\Gamma(D-d_j^{(1)}-d_j^{(2)})} \frac{\Gamma(d_j^{(1)}+d_j^{(2)}+1)\Gamma(D-d_j^{(1)}-d_j^{(2)})}{\Gamma(D+1)} \times$$

$$\times\{B_{0,5}(d_j^{(1)}+1,d_j^{(2)})+B_{0,5}(d_j^{(2)}+1,d_j^{(1)})\}=$$

$$=\frac{1}{D}\sum_j\frac{\Gamma(d_j^{(1)}+d_j^{(2)}+1)}{\Gamma(d_j^{(1)})\Gamma(d_j^{(2)})}\{B_{0,5}(d_j^{(1)}+1,d_j^{(2)})+B_{0,5}(d_j^{(2)}+1,d_j^{(1)})\}=$$

$$=\frac{1}{D}\sum_j\{d_j^{(1)}I_{0,5}(d_j^{(1)}+1,d_j^{(1)})+d_j^{(2)}I_{0,5}(d_j^{(2)}+1,d_j^{(2)})\},$$

что и требовалось доказать. □

3. Апостериорные оценки риска

Пусть задана выборка $\mathbf{s}=(n_1^{(1)},\ldots,n_j^{(i)},\ldots,n_M^{(K)})$. Если придерживаться байесовского подхода, то функцию риска можно рассматривать как случайную функцию $P_f(\widetilde{\boldsymbol{\Theta}})$, зависящую от случайного вектора $\widetilde{\boldsymbol{\Theta}}=\boldsymbol{\Theta}|\mathbf{s}$ с апостериорной плотностью $p(\boldsymbol{\theta}|\mathbf{s})$. По свойству распределения Дирихле, $\widetilde{\boldsymbol{\Theta}}\sim\mathrm{Dir}(d_1^{(1)}+n_1^{(1)},\ldots,d_j^{(i)}+n_j^{(i)},\ldots,d_M^{(K)}+n_M^{(K)})$.

Заметим, что из теоремы 1 следует, что ожидаемая апостериорная вероятность ошибки байесовской решающей функции (при $K=2$):

$$\mathbb{E}P_{f_B}(\widetilde{\boldsymbol{\Theta}})=W(d_1^{(1)}+n_1^{(1)},\ldots,d_j^{(i)}+n_j^{(i)},\ldots d_M^{(2)}+n_M^{(2)}).$$

Следующая теорема описывает вероятностные свойства функции риска для случая фиксированной решающей функции.

Теорема 2. *Пусть величина* $\boldsymbol{\Theta}$ *подчиняется распределению Дирихле с параметрами* $d_1^{(1)},\ldots,d_M^{(K)}$*; из множества* Φ *выбрана произвольная решающая функция* f*; по заданной выборке* $\mathbf{s}\in\mathbf{S}$ *определено число* n_{er} *неправильно распознанных объектов. Обозначим* $D_{er}=D-\sum\limits_j d_j^{(f(j))}$, $a=n_{er}+D_{er}$, $c=N+D$. *Тогда производящая функция моментов величины* $P_f(\widetilde{\boldsymbol{\Theta}})$ *равна* $\psi(t)=H(a,c;t)$, *где* $H(a,c;t)=\sum\limits_{q=0}^{\infty}\dfrac{a_{(q)}}{c_{(q)}}\dfrac{t^q}{q!}$ *— гипергеометрическая функция Куммера, через* $a_{(q)}$ *обозначено произведение* $a(a+1)\cdots(a+q-1)$.

Доказательство. Для случайной величины $U=P_f(\widetilde{\boldsymbol{\Theta}})=1-\sum\limits_j\widetilde{P}_j^{(f(j))}$ производящая функция моментов есть

$$\psi(t)=\mathbb{E}e^{tU}=e^t\frac{\Gamma(c)}{\prod\limits_{j,l}\Gamma(d_j^{(l)}+n_j^{(l)})}\int\limits_{\Lambda}\exp\left(-t\sum_j p_j^{(f(j))}\right)\prod_{l,j}(p_j^{(l)})^{n_j^{(l)}+d_j^{(l)}-1}d\boldsymbol{\theta}=$$

$$=e^t\frac{\Gamma(c)}{\prod\limits_{j,l}\Gamma(d_j^{(l)}+n_j^{(l)})}\int\limits_{\substack{\{p_1^{(f(1))},\ldots,p_M^{(f(M))}:\\ \sum\limits_j p_j^{(f(j))}\le 1\}}}\exp\left(-t\sum_j p_j^{(f(j))}\right)\prod_j(p_j^{(f(j))})^{n_j^{(f(j))}+d_j^{(f(j))}-1}\times$$

$$\times \left\{ \int\limits_{\substack{\{p_j^{(l)}: \\ p_j^{(l)}=1-\sum_{l,j} p_j^{(f(j))}, \\ (l,j)\neq(f(j),j)\}}} \prod_{\substack{l,j \\ f(j)\neq l}} (p_j^{(l)})^{n_j^{(l)}+d_j^{(l)}-1} \prod_{\substack{l,j \\ f(j)\neq l}} dp_j^{(l)} \right\} dp_1^{(f(1))} \dots dp_M^{(f(M))}.$$

Обозначим интеграл в фигурных скобках через I. Заметим, что подынтегральная функция в I зависит от $MK - M$ переменных. Из (1) получим

$$I = \frac{\left(1 - \sum_j p_j^{(f(j))}\right)^{\sum_{l,j:f(j)\neq l}(n_j^{(l)}+d_j^{(l)})-1} \prod_{l,j:f(j)\neq l} \Gamma(d_j^{(l)} + n_j^{(l)})}{\Gamma\left(\sum_{l,j:f(j)\neq l}(n_j^{(l)} + d_j^{(l)})\right)} =$$

$$= \frac{\left(1 - \sum_j p_j^{(f(j))}\right)^{a-1} \prod_{l,j:f(j)\neq l} \Gamma(d_j^{(l)} + n_j^{(l)})}{\Gamma(a)}.$$

Таким образом,

$$\psi(t) = e^t \frac{\Gamma(c)}{\Gamma(a)} \frac{\prod_{l,j:f(j)\neq l} \Gamma(d_j^{(l)} + n_j^{(l)})}{\prod_{l,j} \Gamma(d_j^{(l)} + n_j^{(l)})} \int\limits_{\{p_j^{(f(j))}:\sum p_j^{(f(j))}\leq 1\}} \exp(-t \sum_j p_j^{(f(j))}) \times$$

$$\times \prod_j \left(p_j^{(f(j))}\right)^{n_j^{(f(j))}+d_j^{(f(j))}-1} \left(1 - \sum_j p_j^{(f(j))}\right)^{a-1} dp_1^{(f(1))} \dots dp_M^{(f(M))}.$$

Пусть n_r обозначает число правильно классифицированных объектов: $n_r = N - n_{er} = \sum_j n_j^{(f(j))}$. Обозначим $D_r = \sum_j d_j^{(f(j))}$. Воспользуемся также следующей интегральной формулой [5]:

$$\int\limits_{\substack{\{x_1,\dots,x_m: \\ x_1+\dots+x_m\leq 1, \\ x_i\geq 0\}}} x_1^{d_1-1} \dots x_m^{d_m-1} \phi(x_1 + \dots + x_m) dx_1 \dots dx_m =$$

$$= \frac{\Gamma(d_1) \dots \Gamma(d_m)}{\Gamma(d_1 + \dots + d_m)} \int\limits_0^1 \phi(u) u^{d_1+\dots+d_m-1} du;$$

имеем

$$\psi(t) = e^t \frac{\Gamma(c)}{\Gamma(a)} \frac{1}{\Gamma(n_r + D_r)} \int\limits_0^1 e^{-tu} u^{n_r+D_r-1} (1-u)^{a-1} du.$$

Используя другую известную интегральную формулу [6, формула (13.2.1), с. 505], получим

$$\psi(t) = e^t H(n_r + D_r, c; -t) = H(a, c; t)$$

по свойству гипергеометрической функции [7]. Теорема 2 доказана. □

Из нее можно получить следствие, вытекающее из свойства гипергеометрической функции: производная l-го порядка

$$\frac{d^l}{dt^l} H(a, c; t) = \frac{a_{(l)}}{c_{(l)}} H(a + l, c + l; t).$$

Следствие. *l-й абсолютный момент величины $P_f(\widetilde{\mathbf{\Theta}})$ равен*

$$\mathbb{E}(P_f(\widetilde{\mathbf{\Theta}}))^l = \frac{(n_{er} + D_{er})_{(l)}}{(N + D)_{(l)}}.$$

Очевидно также, что ожидаемая вероятность ошибочного распознавания будет равна

$$\mathbb{E}P_f(\widetilde{\mathbf{\Theta}}) = \frac{n_{er} + D_{er}}{N + D}.$$

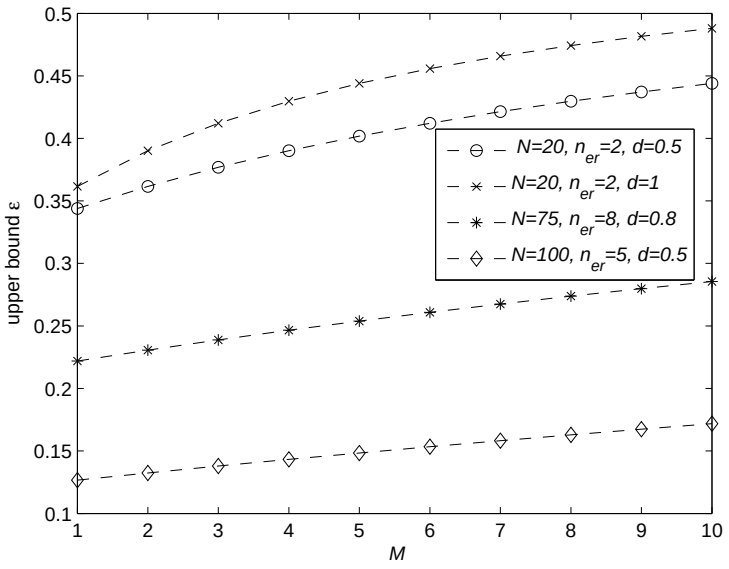
Заметим, что, как следует из теории оценивания, апостериорное математическое ожидание $\mathbb{E}P_f(\widetilde{\mathbf{\Theta}})$ — это оптимальная байесовская оценка вероятности ошибки при квадратичной функции потерь.

На основе полученных результатов может быть найдена приближенная оценка верхней границы риска. Пусть необходимо для заданной величины $\eta \in (0; 1)$ оценить верхнюю границу риска ε, для которой выполняется: $\mathbb{P}(P_f(\widetilde{\mathbf{\Theta}}) \geq \varepsilon) \leq 1 - \eta$.

Как известно, справедливо неравенство Чернова

$$\mathbb{P}(U \geq \varepsilon) \leq \mathbb{E}e^{tU} e^{-t\varepsilon},$$

где U — неотрицательная случайная величина, ε, t — произвольные неотрицательные вещественные числа.



Зависимость верхней границы от числа событий M

Взяв в качестве U величину $P_f(\widetilde{\mathbf{\Theta}})$ и воспользовавшись теоремой 1, получим

$$\mathbb{P}(P_f(\widetilde{\mathbf{\Theta}}) \geq \varepsilon) \leq H(a, c; t)e^{-t\varepsilon}.$$

Для заданных t и $\eta > 0$ найдем соответствующее им ε из условия

$$H(a, c; t)e^{-t\varepsilon} = 1 - \eta,$$

откуда

$$\varepsilon = \frac{1}{t}(\ln H(a, c; t) - \ln(1 - \eta)).$$

Для поиска минимального значения ε, как функции от t, можно воспользоваться методами оптимизации (в данной работе применялись соответствующие процедуры пакета Matlab). На рисунке даны примеры полученных графиков зависимости верхней границы от различных значений объема выборки, числа ошибок и параметров Дирихле. Здесь $K = 2$, $\eta = 0.95$, все $d_j^{(i)}$ совпадают и равны величине d, которая принимает значения 0.5, 0.8, 1.0.

4. Применение оценок при построении логических решающих функций

Практическое значение полученных апостериорных оценок риска состоит, в частности, в том, что эти оценки могут использоваться на этапе обучения как критерий оптимальности решающей функции. Рассмотрим задачу нахождения оптимальной логической решающей функции, имеющей форму дерева решений. Будем рассматривать множество листьев исходного дерева (либо некоторого его поддерева с тем же корнем) как конечное множество событий. Байесовская модель позволяет получить оценку качества распознавания, которая может рассматриваться как критерий оптимальности.

Разобьем выборку на две части примерно равного объема. Пусть в результате работы некоторого алгоритма построения дерева решений (например, с помощью метода последовательно ветвления ЛРП [4] или других методов [1, 8]) по первой части обучающей выборки построено дерево решений. Параметры алгоритма должны быть подобраны так, чтобы число листьев было бы достаточно велико (скажем, примерно равно числу наблюдений). Теперь, с использованием второй части обучающей выборки, определим частоты попадания наблюдений каждого класса в вершины этого дерева.

Рассмотрим произвольное редуцированное поддерево T исходного дерева (с той же самой корневой вершиной, что и у исходного дерева). Набор наблюдаемых частот, соответствующий листьям, обозначим через $\mathbf{s}$. Заметим, что структура поддерева не зависит от $\mathbf{s}$, поскольку эти наблюдения не участвовали в его формировании. Определим апостериорную оценку риска для T. Необходимо найти поддерево с минимальным значением оценки.

Используется следующий приближенный алгоритм поиска оптимального варианта редуцирования. Для каждой внутренней вершины дерева определяется значение критерия для дерева, которое получилось бы, если бы данная вершина стала конечной. Вершина с наилучшим значением критерия, если это значение меньше, чем у исходного дерева, объявляется листом. Далее описанная процедура повторяется, пока не останется ни одной вершины, позволяющей уменьшить значение критерия.

Описанный алгоритм применялся, в частности, для решения задачи прогнозирования редких ("нежелательных" или "экстремальных") событий. Особенность задачи состоит в том, что количество соответствующих прецедентов в эмпирической информации мало по отношению к общему объему выборки. Это обусловливает необходимость разработки специальных методов, позволяющих как можно точнее оценивать риск.

Был проведен следующий вычислительный эксперимент. Случайным образом генерировались две булевы и одна числовая последовательности, не зависимые друг от друга. Длина последовательностей была задана равной 1000. Полагалось, что при определенном заданном сочетании предыдущих значений этих последовательностей с вероятностью 0.25 возникает нежелательное событие, а во всех остальных случаях это событие не возникает. Таким образом, формировалась булева последовательность, обозначающая наличие или отсутствие нежелательного события (доля нежелательных событий составляла около 0.08). Полученные последовательности подавались на вход алгоритма, на выходе которого формировалось дерево решений. Для построения исходного дерева решений использовался R-метод [1].

Для оценки качества алгоритма генерировалась контрольная выборка, состоящая из 100 нежелательных событий вместе с их описанием. Показателем качества алгоритма служил процент правильно распознанных событий. Описанная процедура была многократно повторена. Оказалось, что в среднем алгоритм в 78 % случаев правильно прогнозировал возникновение нежелательного события. Если же задавалась вероятность возникновения нежелательного события, при заданном сочетании предыдущих значений последовательностей равная 0.75, то средний процент правильного распознавания повышался до 99. В следующем эксперименте дополнительно к предыдущему варианту предполагалось, что среди значений обучающих последовательностей имеется 5 % пропусков (неизмеренных значений). Местонахождение пропусков выбиралось случайно. Результаты показали, что в среднем в 97 % случаев алгоритм правильно определял возникновение нежелательного события.

Данный метод построения дерева решений сравнивался с аналогичным методом, использующим в качестве критерия обычную оценку эмпирического риска (REP-метод [8]). Показатель качества метода, использующего байесовскую оценку, получился на 7 % лучше, чем аналогичный показатель для REP-метода.

Приведем следующий пример прикладной задачи [9]. Имеются ряды данных о стоке реки Обь, об осадках и о температуре в районе Барнаула и Колпашево за последние 80 лет с месячным интервалом. Требуется составить прогноз маловодья в марте по данным ноября. Прогноз необходим, в частности, для заблаговременного предупреждения соответствующих служб водоснабжения населения и предприятий. Возникновение события "маловодье" устанавливается специалистами (как правило, по отклонению величины стока от среднемесячного значения по зимнему периоду на величину, большую среднеквадратического отклонения). Так как за последние 25–30 лет маловодий такого же масштаба, как в первые годы наблюдений, не происходило (видимо, из-за глобальных климатических изменений, а также из-за постройки Новосибирской ГЭС), то данные по первым и последним годам анализировались отдельно. Таким образом, было проанализировано четыре ряда. В итоге средняя частота ошибок 1-го рода на скользящем экзамене составила от 0 до 0.14 (т. е. в последнем случае было неправильно предсказано одно маловодье из семи), а средняя частота ошибок 2-го рода на скользящем экзамене — от 0.13 до 0.26. Получены закономерности, характерные для возникновения маловодий. Например, для Барнаула до постройки ГЭС наблюдалась следующая закономерность:

если расход воды в ноябре не превышает значения 500 м^3/с и средняя температура воздуха в ноябре меньше -5 °C, то в марте будет маловодье.

Заключение

В работе исследованы свойства байесовской логико-вероятностной модели распознавания образов по конечному множеству событий: получено выражение для ожидаемой вероятности ошибки оптимальной байесовской решающей функции; найдены апостериорные точечные и интервальные оценки риска неправильного распознавания произвольной фиксированной решающей функции в случае заданной выборки. Описан подход к построению и исследованию класса логических решающих функций, основанный на применении полученных оценок. Предложенный подход обладает рядом преимуществ: позволяет рассматривать разнотипные переменные, не требует жесткого задания модели распределения, учитывает экспертные знания о классе распределений, не ориентирован на самый "неблагоприятный" вид распределения и на асимптотический случай.

Список литературы

[1] Лбов Г.С., Бериков В.Б. Устойчивость решающих функций в задачах распознавания образов и анализа разнотипной информации. Новосибирск: Ин-т математики СО РАН, 2005.

[2] Berikov V.B. Bayes estimates for recognition quality on a finite set of events // Pattern Recognition and Image Analysis. 2006. Vol. 16, N 3. P. 329–343.

[3] Бериков В.Б., Лбов Г.С. Выбор оптимальной сложности класса логических решающих функций в задачах распознавания образов // Докл. АН. 2007. Т. 417, №1. С. 26–29.

[4] Лбов Г.С. Методы обработки разнотипных экспериментальных данных. Новосибирск: Наука. Сиб. отд-ние. 1981.

[5] Фихтенгольц Г.М. Курс дифференциального и интегрального исчисления. М.: Физматлит, 1960. Т. 3.

[6] Abramowitz M., Stegun I.A. Handbook of Mathematical Functions. Washington, D.C.: NBS, 1972.

[7] Бейтман Г., Эрдейи А. Высшие трансцендентные функции. М.: Наука, 1973. Т. 1.

[8] Esposito F., Malerba D., Semerato G. A comparative analysis of methods for pruning decision trees // IEEE Trans. Pattern Anal. Mach. Intell. 1997. Vol. 19, N 5. P. 476–491.

[9] Лбов Г.С., Бериков В.Б., Герасимов М.К. Прогнозирование экстремальных гидрологических ситуаций на основе анализа многомерных временных рядов // Тр. Междунар. науч. конф. "Экстремальные гидрологические события: теория, моделирование, прогнозирование". Москва, 3–6 ноября 2003. С. 26–30.

Поступила в редакцию 12 марта 2008 г.