

СВОЙСТВА КОМБИНИРОВАННОЙ ОЦЕНКИ В ЗАДАЧЕ КЛАССИФИКАЦИИ НАБЛЮДЕНИЙ

С.В. Скрипин

Томский государственный университет

Томский научный центр СО РАН

E-mail: skripin@ef.tsu.ru

Представлены свойства комбинированной непараметрической оценки классификации, использующей оценки параметрической и непараметрической моделей дискриминации. Комбинированная оценка использует дополнительно оценки некоторых функционалов плотности. Статистическим моделированием показано, что комбинированные оценки для выборок конечного объема предпочтительнее, чем каждая из оценок построенных моделей дискриминации.

Ключевые слова:

Непараметрическая классификация, комбинированная оценка, бутстреп-оценка, статистическое моделирование.

Введение

Задачи отнесения объектов наблюдений по их признакам к одному из известных классов при наличии обучающей (эталонной) выборки распространены во многих областях деятельности человека – в науке, технике, промышленности, медицине, геологии и так далее. Они решаются различными методами, в том числе методами математической статистики (задачи дискриминации). В них задача оценки вероятности отнесения к известному классу может решаться с использованием множества различных математических моделей, как из класса параметрических (когда имеется априорная информация или предположения о законах распределения плотности вероятности в классах), так и непараметрических (когда априорной информации о классах нет). Часто встречаются ситуации, когда информация о классах отсутствует, а объем эталонной выборки N конечен (не может быть дополнен) и мал (несколько десятков наблюдений). В этих условиях качество оценок классификации у выбранных моделей может оказаться неприемлемо низким, а их перебор из разных классов и сравнение – трудоемким и малоэффективным.

Другим подходом в построении моделей дискриминации является получение комбинированных оценок, использующих оценки построенных моделей дискриминации и априорную информацию о классах из эталонной выборки. Среди различных разработок указанного подхода интересен метод, использующий априорную информацию о некоторых функционалах плотности [1]. Его преимущества при получении комбинированных оценок регрессии, как в случае конечного объема выборок N , так и в асимптотическом случае, показаны в [2]. Варианты применения указанного метода для задачи классификации наблюдений, в частном случае ограничения числа построенных моделей дискриминации до двух, представлены в [3, 4]. В данной работе рассмотрен случай получения комбинированных оценок вероятности классификации, построенных по оценкам двух моделей – параметрической и непараметрической, с использованием подхода, предложенного в [2–4].

Параметрическая и непараметрическая оценки

Пусть имеется эталонная выборка объема $N=n_1+\dots+n_k$, содержащая наблюдения из k классов. Пусть $(X_1^t, \dots, X_m^t)$ – независимые наблюдения случайного вектора X из класса $t=1, \dots, k$, размерности m с неизвестной плотностью вероятности $f^t(x)$; $x \in R^m$. Ограничимся рассмотрением двух случаев: когда функция плотности известна с точностью до параметра θ (параметрический случай), и когда функция плотности полностью неизвестна. Это приводит к двум моделям дискриминации: параметрической $J(x; \theta)$ и непараметрической $J(x)$.

В условиях малых объемов выборки N и отсутствия информации о классах априорные вероятности классификации в каждый класс примем равными. Тогда параметрическую оценку апостериорной вероятности классификации наблюдения в точке x в класс t (классифицирующей функции в класс t) представим в виде

$$J(x; \hat{\theta}^t) = \hat{p}(t|x) = f^t(x; \hat{\theta}^t) / \sum_{j=1}^k f^j(x; \hat{\theta}^j), \quad t = \overline{1, k}. \quad (1)$$

Здесь $f^t(x; \hat{\theta}^t)$ – параметрическая оценка плотности вероятности классификации в класс t . Если каждый класс описывается m -мерной нормальной генеральной совокупностью, то функция плотности вероятности классификации наблюдения в точке x в класс t может быть представлена параметрической оценкой вида [5]

$$f^t(x; \hat{\theta}^t) = (2\pi)^{-m/2} |\mathbf{V}^t|^{-1/2} \exp(-0.5 d^2(x; \hat{\theta}^t)), \quad t = \overline{1, k}, \quad (2)$$

где $|\mathbf{V}^t|$ – детерминант матрицы ковариаций из класса t , элементы которой получены по эталонной выборке с помощью метода максимального правдоподобия; $d^2(x; \hat{\theta}^t)$ – обобщенный квадрат расстояния от точки x до центра группы наблюдений эталонной выборки из класса t . Если используется объединенная матрица ковариаций, его можно записать

$$d^2(x; \hat{\theta}^t) = (x - \hat{\mathbf{a}}^t)^T (\mathbf{V}^t)^{-1} (x - \hat{\mathbf{a}}^t), \quad t = \overline{1, k}, \quad (3)$$

где $\hat{\mathbf{a}}^t = (\hat{a}_1^t, \dots, \hat{a}_m^t)$ – оценка вектора математических ожиданий компонент случайного вектора X из

класса t ; $\mathbf{V}^t$ — объединенная матрица ковариаций из класса t , полученная по эталонной выборке с помощью метода максимального правдоподобия.

Пренебрегая в (3) элементами перекрестных связей и квадратичными, выражение показателя функции экспоненты в (2) можно приближенно представить в виде линейной функции дискриминации

$$-0,5d^2(\mathbf{x}; \hat{\theta}^t) \approx -0,5(\hat{\mathbf{a}}^t)^T (\mathbf{V}^t)^{-1} \hat{\mathbf{a}}^t + \mathbf{x}(\mathbf{V}^t)^{-1} \hat{\mathbf{a}}^t = \\ = \hat{\theta}_0^t + \hat{\theta}_1^t x_1 + \dots + \hat{\theta}_m^t x_m, \quad t = \overline{1, k}, \quad (4)$$

где $\hat{\theta}_0^t = -0,5(\hat{\mathbf{a}}^t)^T (\mathbf{V}^t)^{-1} \hat{\mathbf{a}}^t$, $(\hat{\theta}_1^t, \dots, \hat{\theta}_m^t) = (\mathbf{V}^t)^{-1} \hat{\mathbf{a}}^t$.

После получения оценок вида (4) в каждом классе наблюдение в точке $\mathbf{x}$ отнесем к тому классу, в котором оценка вида (1) получит максимальное значение.

Непараметрическую оценку апостериорной вероятности классификации наблюдения в точке $\mathbf{x}$ в класс t представим в виде [4]:

$$J_N^t(\mathbf{x}) = \hat{p}(t | \mathbf{x}) = f_{n_t}^t(\mathbf{x}) / \sum_{j=1}^k f_{n_j}^j(\mathbf{x}), \quad t = \overline{1, k}. \quad (5)$$

Здесь $f_{n_t}^t(\mathbf{x})$ — непараметрическая оценка плотности вероятности классификации в класс t вида:

$$f_{n_t}^t(x^{(1)}, \dots, x^{(m)}) = \frac{1}{n_t H_t} \sum_{i=1}^{n_t} \prod_{l=1}^m K[(x^{(l)} - X_{i,t}^{(l)}) / h_t^{(l)}],$$

где n_t — число наблюдений из класса t ; $K(u)$ — заданное ядро (некоторая функция плотности вероятности); $h_t^{(l)}$ — параметры масштаба в классе t , удовлетворяющие следующим требованиям: при объеме выборки $N = n_1 + \dots + n_k$ и числа наблюдений класса n_t , стремящихся к бесконечности

$$H_t = \prod_{l=1}^m h_t^{(l)}, \quad h_t^{(l)} \rightarrow 0; \quad n_t H_t \rightarrow \infty; \quad h_t^{(l)} = h_t^{(l)}(n_t, N).$$

При рассмотрении выражений для средних удельных потерь и оптимальных процедур (Байеса) классификации очевидно, что для сравнения оценок вида (4) и (5) с данными эталонной выборки выполняется замена данных о принадлежности к классу (номинальная шкала) оценками вероятности отнесения к каждому из классов эталонной выборки (числовая шкала). Если данные о принадлежности наблюдений эталонной выборки к классам «абсолютно» достоверны, оценки отнесения наблюдений в свой класс приравняются к единице, а к остальным классам — к нулю. Т. е., для наблюдения эталонной выборки $\mathbf{X}_i$ из класса t оценки будут

$$q_i(t | \mathbf{X}_i) = \begin{cases} 1, & \mathbf{X}_i \in t, \\ 0, & \mathbf{X}_i \notin t, \end{cases} \quad t = \overline{1, k}, \quad i = \overline{1, N}. \quad (6)$$

Выражение (6) нельзя считать наилучшим во всех случаях. Для эталонных выборок с малым числом наблюдений в классах многие значения случайной величины $\mathbf{X}$ из разных классов не реализуются. Вероятности их реализации приравняются к нулю, что ухудшает качество классификации.

Для учета указанных недостатков используется практика присвоения некоторой доли вероятности $\alpha < 1$ на значения случайной величины $\mathbf{X}$ из других классов. Можно выбрать разные способы такого присвоения [6]. Представим значения переменной — отклика (в номинальной шкале измерений) о принадлежности каждого наблюдения эталонной выборки $\mathbf{X}_i$ из класса t оценками отнесения к своему классу равной $(1-\alpha)$, а к другим классам — $\alpha/(k-1)$. Тогда вместо (6) представим это определение в виде

$$Y_i^j(t | \mathbf{X}_i) = \begin{cases} 1-\alpha > 0, & \mathbf{X}_i \in t, \\ \frac{\alpha}{k-1} > 0, & \mathbf{X}_i \notin t, \end{cases} \\ i = \overline{1, N}, \quad j = \overline{1, k}, \quad t = \overline{1, k}. \quad (7)$$

Для каждого наблюдения $\mathbf{X}_i$ получим k значений оценок отнесения к классам — $\mathbf{Y}_i = (Y_i^1, \dots, Y_i^k)$. В результате замены данных переменной — отклика о принадлежности к классу оценками вида (7) для всех наблюдений эталонной выборки получим матрицу $\mathbf{Y}$, содержащую k столбцов и i строк оценок отнесения каждого наблюдения $\mathbf{X}_i$ к каждому из k классов. Оценки представлены в числовой шкале и допускают любые математические операции. Следовательно, становится возможным применение методов улучшения оценок классификации, представленных в [1–4].

Выбор оптимального значения параметра α для получения наилучших оценок классификации является отдельной задачей и в данной работе не рассматривается. Зафиксируем его значение, соблюдая условие — $(1-\alpha) > \alpha$.

Комбинированная оценка

Комбинированную оценку вероятности классификации наблюдения в точке $\mathbf{x}$ в класс t представим в виде

$$J_N^t(\mathbf{x}; \lambda^t) = J_N^t(\mathbf{x}) - \lambda^t (J_N^t(\mathbf{x}) - J(\mathbf{x}; \hat{\theta}^t)), \quad t = \overline{1, k}, \quad (8)$$

где λ^t — коэффициент в классе t , выбираемый согласно требованиям, предъявляемым к качеству оценки.

Выберем параметр λ^t в выражении (8) для случая конечного объема эталонной выборки N из следующего критерия

$$Q(\lambda^t) = \sum_{i=1}^{n_t} [Y_i^t - J_N^t(X_i^{(1)}, \dots, X_i^{(m)}; \lambda^t)]^2 \rightarrow \min_{\lambda^t}. \quad (9)$$

Из выражения (9) получим оптимальный коэффициент $\lambda^t = \lambda_o^t$:

$$\lambda_o^t = \frac{\sum_{i=1}^N \left((J_N^t(X_i^{(1)}, \dots, X_i^{(m)}) - Y_i^t)(J_N^t(X_i^{(1)}, \dots, X_i^{(m)}) - J(X_i^{(1)}, \dots, X_i^{(m)}; \hat{\theta}^t)) \right)}{\sum_{i=1}^N (J_N^t(X_i^{(1)}, \dots, X_i^{(m)}) - J(X_i^{(1)}, \dots, X_i^{(m)}; \hat{\theta}^t))^2}. \quad (10)$$

Выражение (10) не зависит от выбора точки $\mathbf{x}$.

Рассмотрим другую комбинированную оценку (8) с выбором коэффициента λ' (в асимптотическом случае) на основе критерия

$$Q(\lambda'(\mathbf{x})) = E[J'_N(\mathbf{x}; \lambda'(\mathbf{x})) - J'(\mathbf{x})]^2 \rightarrow \min_{\lambda'(\mathbf{x})}, \quad (11)$$

где E – оператор математического ожидания. Из выражения (11) получим оптимальный коэффициент $\lambda'_o(\mathbf{x}) = \lambda'_o(\mathbf{x})$:

$$\lambda'_o(\mathbf{x}) = \frac{E(J'_N(\mathbf{x}) - J'(\mathbf{x}))(J'_N(\mathbf{x}) - J(\mathbf{x}; \hat{\theta}^t))}{E(J'_N(\mathbf{x}) - J(\mathbf{x}; \hat{\theta}^t))^2}. \quad (12)$$

Поскольку плотность вероятности $J'(\mathbf{x})$ неизвестна, то неизвестна и $\lambda'_o(\mathbf{x})$. Выделив в (12) главную часть относительно N и оценив ее по исходной выборке, получим оценку $\lambda'_o(\mathbf{x})$ в точке $\mathbf{x}$ [2, 3]:

$$\hat{\lambda}'_{o,ig}(\mathbf{x}) = \frac{C'(\mathbf{x})}{V'(\mathbf{x})} \left(1 - \frac{NH_t(\Delta'_1(\mathbf{x}))^2 / V'(\mathbf{x})}{1 + NH_t(\Delta'_1(\mathbf{x}))^2 / V'(\mathbf{x})} \right).$$

Здесь $C(\mathbf{x}) = g'_2(\mathbf{x}) - J'_N(\mathbf{x})g'_1(\mathbf{x})$,

$$V'(\mathbf{x}) = g'_2(\mathbf{x}) - J'_N(\mathbf{x})g'_1(\mathbf{x}) + (J'_N(\mathbf{x}))^2 - J'_N(\mathbf{x})g'_1(\mathbf{x}),$$

$$\Delta'_1(\mathbf{x}) = \Delta'_0(\mathbf{x}) \left(1 - \frac{1}{1 + (NH)^{\frac{2}{3}}(\Delta'_0(\mathbf{x}))^2} \right),$$

$$\Delta'_0(\mathbf{x}) = g'_1(\mathbf{x}) - J(\mathbf{x}; \hat{\theta}^t)g'_0(\mathbf{x}),$$

$$g'_0(\mathbf{x}) = \frac{1}{NH_t} \sum_{i=1}^N \prod_{l=1}^m K[(x^{(l)} - X_i^{(l)}) / h_t^{(l)}],$$

$$g'_1(\mathbf{x}) = \frac{1}{NH_t} \sum_{i=1}^N Y_i \prod_{l=1}^m K[(x^{(l)} - X_i^{(l)}) / h_t^{(l)}],$$

$$g'_2(\mathbf{x}) = \frac{1}{NH_t} \sum_{i=1}^N (Y_i)^2 \prod_{l=1}^m K[(x^{(l)} - X_i^{(l)}) / h_t^{(l)}].$$

Оценку коэффициента $\lambda'_o(\mathbf{x})$ в выражении (12) можно выполнить и бутстреп-методом [7]. Причем, как и в [2], оценку можно представить в двух вариантах.

Первый вариант предусматривает получение для каждой бутстреп-выборки оценок классификации двух моделей – параметрической и непараметрической. Обозначим оценки, получаемые по бутстреп-методу символом звездочка, а по исходной выборке – без звездочки. В соответствии с (12), оценка оптимального коэффициента $\lambda'_o(\mathbf{x})$ по бутстреп-выборкам примет вид

$$\hat{\lambda}'_{o,*}(\mathbf{x}) = \sum_{s=1}^B \frac{(J'_{N,s}(\mathbf{x}) - J'_N(\mathbf{x}))(J'_{N,s}(\mathbf{x}) - J^*(\mathbf{x}; \hat{\theta}^{t,s}))}{\sum_{j=1}^B (J'_{N,s}(\mathbf{x}) - J^*(\mathbf{x}; \hat{\theta}^{t,s}))^2}, \quad (13)$$

где B – размер серии сформированных бутстреп-выборок, соответствующих одной исходной выборке, для оценки коэффициента $\lambda'_o(\mathbf{x})$ в точке $\mathbf{x} \in R^n$.

Второй вариант предусматривает получение для каждой бутстреп-выборки оценок классификации только одной модели – непараметрической. Тогда оценка оптимального коэффициента $\lambda'^{ig}_o(\mathbf{x})$ по бутстреп-выборкам будет иметь вид

$$\hat{\lambda}'_{o,*}(\mathbf{x}) = \sum_{s=1}^B \frac{(J'_{N,s}(\mathbf{x}) - J'_N(\mathbf{x}))(J'_{N,s}(\mathbf{x}) - J(\mathbf{x}; \hat{\theta}^{t,s}))}{\sum_{s=1}^B (J'_{N,s}(\mathbf{x}) - J(\mathbf{x}; \hat{\theta}^{t,s}))^2}. \quad (14)$$

Оценка (14), в отличие от оценки (13), позволяет экономить вычислительные ресурсы и избегать проблем с получением оценок параметрической модели дискриминации в случае вырожденной матрицы ковариаций у сформированных бутстреп-выборок.

Условия эксперимента

С целью сравнения качества оценок вида (8) проведен имитационный эксперимент, а также проверка качества оценок на нескольких выборках реальных статистических данных. Для исключения тривиальных результатов классификации из эксперимента исключены оценки в фиксированных точках $\mathbf{x}_i$ с одинаковым объемом выборок n . Имитационный эксперимент проведен с получением оценок классификации для различных объемов эталонных выборок N . Для упрощения интерпретации результатов эксперимент проведен с одномерной случайной величиной $X \in R$. Сравнение свойств оценок выполнено со случайным выбором точек X_i из двух одномерных случайных величин X^1 и X^2 , представляющих два класса; $k=2$. В имитационном эксперименте выбраны следующие условия моделирования:

- для имитационного моделирования данных выборки генерировались с наблюдениями из двух одномерных случайных величин X^1 и X^2 с нормальным законом распределения;
- случайные величины X^1 и X^2 имели одинаковые значения дисперсии и математические ожидания, равные $\sigma_1 = \sigma_2 = 20,0$, $EX^1 = 60,0$, $EX^2 = 100,0$;
- для формирования области перемешивания наблюдений из разных классов расстояние между математическими ожиданиями случайных величин выбрано равным двум среднеквадратическим отклонениям, $EX^2 - EX^1 = 40,0$;
- отбор в исходные выборки наблюдений из разных классов выполнялся с вероятностью 0,5;
- оценки классификации получены для выборок с объемом наблюдений в диапазоне $N=6-95$ с интервалом $\delta=5$, отсчитываемым от конца диапазона;
- для каждой исходной и бутстреп-выборки соблюдалось условие – в выборке должно присутствовать не менее двух наблюдений из каждого класса;
- числовые результаты эксперимента были получены по серии исходных выборок количеством $K=10000$ (с одинаковым объемом наблюдений N в каждой). Для каждого наблюдения каждой исходной выборки из серии K формировалась серия бутстреп-выборок размером $B=1000$.

При выбранном законе распределения случайных величин X^1 и X^2 в имитационном эксперименте

обеспечивается преимущество (в смысле качества) оценке параметрической модели дискриминации.

Сравнение качества оценок классификации у шести моделей $J_1=J(x;\hat{\theta}^1)$, $J_2=J_N(x)$, $J_3=J_N(x;\hat{\lambda}_o^1)$, $J_4=J_N(x;\hat{\lambda}_o^2)$, $J_5=J_N(x;\hat{\lambda}_o^3)$, $J_6=J_N(x;\hat{\lambda}_o^4)$ проводилось по следующим критериям.

Критерии, вычисленные для каждой исходной выборки.

1. Оценка вероятности безошибочной классификации наблюдений эталонной выборки по всем классам

$$S_r^1 = \frac{1}{N} \sum_{j=1}^k u_j, \quad r = \overline{1,6}, \quad (15)$$

где u_j – количество правильно классифицированных наблюдений эталонной выборки из класса j ; r – номер модели классификации в соответствии с нумерацией, указанной выше.

2. Оценка вероятности классификации наблюдений эталонной выборки в точке X_i в свой класс t в связи с дискретными свойствами выражения (15) для случая равных значений величин S_r^1

$$S_r^2 = \frac{1}{N} \sum_{i=1}^N \hat{p}_i(t | X_i \in t), \quad r = \overline{1,6}, \quad t = \overline{1,k}. \quad (16)$$

Критерии, вычисленные по серии K исходных выборок.

3. Среднее оценок вероятностей безошибочной классификации наблюдений эталонной выборки по всем классам

$$Q_r^1 = \frac{1}{K} \sum_{i=1}^K S_{r,i}^1, \quad r = \overline{1,6}. \quad (17)$$

4. Среднее оценок вероятностей классификации наблюдений эталонной выборки в свой класс

$$Q_r^2 = \frac{1}{K} \sum_{i=1}^K S_{r,i}^2, \quad r = \overline{1,6}. \quad (18)$$

Выражения (15)–(18) нормированы в диапазоне $[0,1]$.

Результаты эксперимента

Результаты имитационного эксперимента по критериям (17) и (18) представлены на графиках, рис. 1, 2. Номера кривых совпадают с номерами моделей дискриминации в соответствии с нумерацией, указанной выше.

На рис. 1 представлены результаты сравнения качества оценок классификации по критерию Q_r^1 . По горизонтальной оси указан объем выборки N , по вертикальной оси – величина среднего оценок вероятностей безошибочной классификации (величины Q_r^1).

На рис. 2 представлены результаты сравнения качества оценок классификации по критерию Q_r^2 . По горизонтальной оси указан объем выборки N , по вертикальной оси – величина среднего оценок вероятностей классификации в свой класс (величины Q_r^2).

Кривые на графиках рис. 1, 2 статистически значимо различаются на уровне значимости мень-

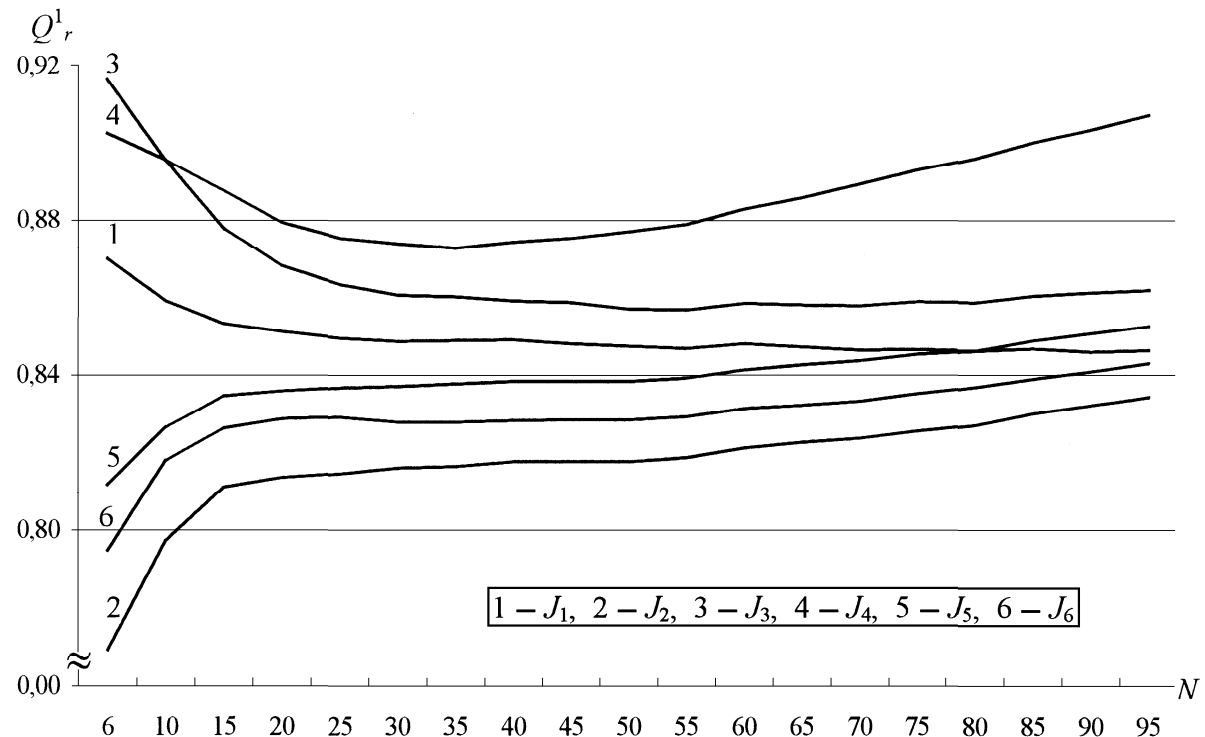


Рис. 1. Среднее оценок вероятностей безошибочной классификации Q_r^1

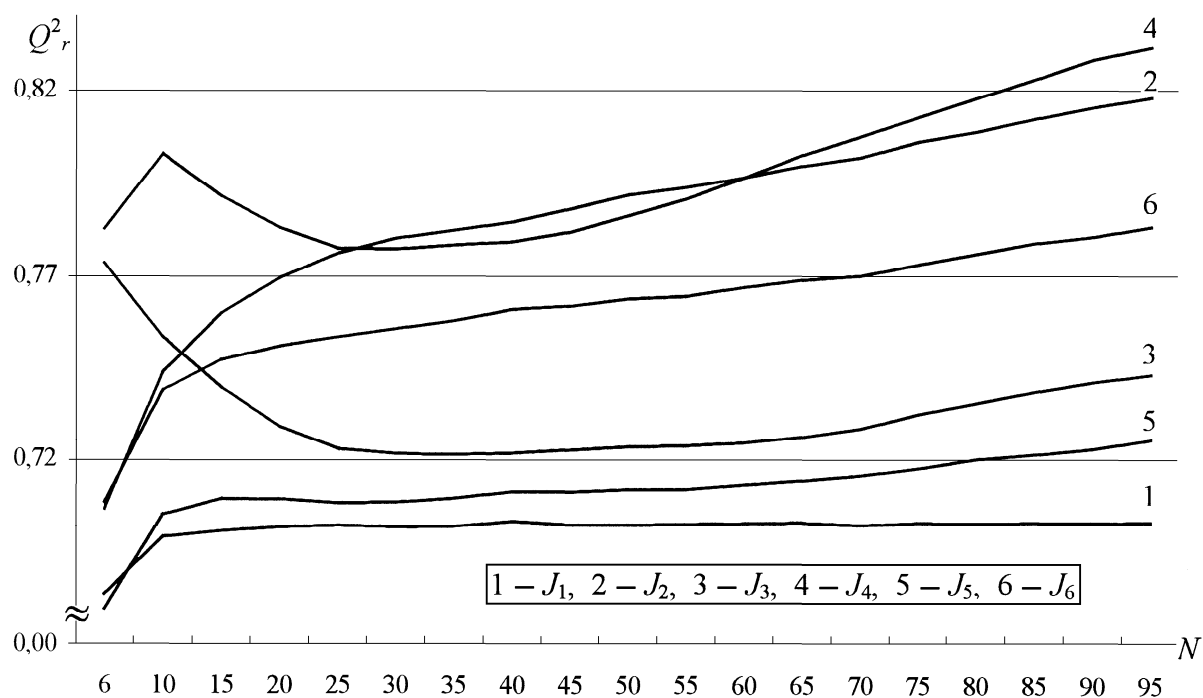


Рис. 2. Среднее оценок вероятностей классификации в свой класс Q_r^2 .

ше 0,05, за исключением областей пересечения кривых.

По результатам моделирования при указанных условиях имитационного эксперимента можно сделать следующие выводы.

1. Оценки модели дискриминации:

- параметрической J_1 при $N < 80$ дают меньшие, чем у непараметрической J_2 и комбинированных моделей J_5 и J_6 , ошибки классификации (рис. 1, кривая 1). Значения среднего оценок вероятностей классификации в свой класс (рис. 2, кривая 1) самые низкие среди рассматриваемых моделей. Следовательно, качество оценок параметрической модели дискриминации J_1 в имитационном эксперименте высокое;
- непараметрической J_2 дают наибольшие ошибки классификации среди рассматриваемых моделей (рис. 1, кривая 2). Значения среднего оценок вероятностей классификации в свой класс одни из самых высоких и занимают второе место среди рассматриваемых моделей (рис. 2, кривая 2). Это свидетельствует о невысоком качестве оценок непараметрической модели J_2 в имитационном эксперименте;
- комбинированной J_3 дают малые ошибки классификации (рис. 1, кривая 3) и проигрывают только оценкам комбинированной модели J_4 (рис. 1, кривая 4). При этом значения среднего оценок вероятностей классификации в свой класс (рис. 2, кривая 3) выигрывают только у моделей параметрической J_1 и комбинированной J_5 (рис. 2, кривые 1 и 5).

Оценки комбинированной модели J_3 в эксперименте по качеству занимают второе место после оценок комбинированной модели J_4 ;

- комбинированной J_4 дают наименьшие ошибки классификации (рис. 1, кривая 4) среди рассматриваемых моделей. Значения среднего оценок вероятности классификации в свой класс (рис. 2, кривая 4) наибольшие среди рассматриваемых моделей и близки к значениям непараметрической модели J_2 . Следовательно, оценки модели J_4 являются наилучшими в имитационном эксперименте;
- комбинированной J_5 дают меньшие, чем у непараметрической и комбинированной J_6 , ошибки классификации (рис. 1, кривая 5), но проигрывают оценкам комбинированных моделей J_3 и J_4 и параметрической модели J_1 при $N < 80$ (рис. 1, кривые 1, 3 и 4). Значения среднего оценок вероятностей классификации в свой класс (рис. 2, кривая 5) выше оценок только параметрической модели. Следовательно, что применение бутстреп-метода дает невысокое качество оценок классификации при $N < 80$;
- комбинированной J_6 выигрывают только у оценок непараметрической модели J_2 (рис. 1, кривая 6) и немного проигрывают модели J_5 . При этом значения среднего оценок вероятностей классификации в свой класс (рис. 2, кривая 6) проигрывают только оценкам моделей непараметрической J_2 и комбинированной J_4 . Вывод аналогичен выводу для модели J_5 .

2. Качество оценок параметрической J_1 , непараметрической J_2 , комбинированных моделей J_3 , J_5 и J_6 при $N > 70$ сближается. При этом реализация бутстреп-метода в представленном эксперименте требует больших затрат вычислительных ресурсов и времени, растущих с увеличением N .
3. Уменьшение ошибок классификации у комбинированных моделей J_3 и J_5 выполняется за счет перераспределения оценок вероятностей классификации в свой класс в сторону более «сдержанных» по сравнению с излишне «оптимистичными» у непараметрической модели J_2 , что демонстрирует рис. 2.
4. Невысокое качество оценок у непараметрической модели J_2 не ухудшает свойств комбинированной оценки (8) достигать наилучших результатов.

На реальных данных оценки комбинированных моделей дискриминации J_3 и J_4 получали результат по критерию (15) лучше оценок классификации параметрической и непараметрической моделей.

Заключение

В работе показана принципиальная возможность применения комбинированной оценки, используемой при получении оценок регрессии [2], для решения задачи классификации наблюдений при наличии эталонной выборки.

В имитационном эксперименте подтверждены свойства комбинированной оценки, представленные в [2], позволяющие получать оценки класси-

фикации лучше, чем оценки построенных моделей дискриминации (параметрической и непараметрической) в условиях малых объемов эталонных выборок.

Результаты имитационного эксперимента демонстрируют преимущества оценок комбинированных моделей J_3 и J_4 по сравнению с оценками построенных моделей при $N < 80$. При $N > 100$ оценки всех представленных комбинированных моделей демонстрируют преимущества по сравнению с оценками построенных моделей.

В имитационном эксперименте оценки комбинированных моделей J_5 и J_6 , полученные с применением бутстреп-метода, проигрывают оценкам комбинированных моделей J_3 и J_4 во всем диапазоне изменения величины N . Затраты времени для реализации бутстреп-метода на несколько порядков превышают таковые для других моделей. Следовательно, применение оценок комбинированных моделей J_3 и J_4 позволяет отказаться от использования бутстреп-метода получения комбинированных оценок в указанной задаче классификации.

Результаты эксперимента показывают, что представленные комбинированные оценки классификации предпочтительнее, чем каждая из оценок параметрической и непараметрической моделей. Их преимущества проявляются как при малых объемах эталонных выборок, так и при их интерполяции за пределы эксперимента ($N > 100$).

Результаты моделирования, представленные на рисунках, получены с помощью кластера Межрегионального вычислительного центра ТГУ СКИФ Cyberia (skif.tsu.ru). Автор выражает благодарность сотрудникам Центра за оказанную помощь.

СПИСОК ЛИТЕРАТУРЫ

1. Дмитриев Ю.Г. Непараметрическое условное оценивание функционалов плотности распределения // Математическое моделирование и теория вероятностей / Под ред. И.А. Александрова, А.М. Бубенчикова, В.Н. Берцуна, Ю.К. Устинова. – Томск: Изд-во «Пеленг», 1998. – С. 169–177.
2. Скрипин С.В. Свойства комбинированной оценки регрессии при конечных объемах выборок // Известия Томского политехнического университета. – 2008. – Т. 313. – № 5. – С. 10–14.
3. Скрипин С.В. Комбинированные непараметрические оценки в задаче многомерной дискриминации // Вестник Томского государственного университета. – 2006. – Прилож. № 17. – С. 303–306.
4. Скрипин С.В. Комбинированные оценки в задаче дискриминации // Информационные технологии и математическое моделирование (ИТММ-2007): Матер. VI Всеросс. научно-практ. конф. – Анжиро-Судженск. – Томск: Изд-во Том. ун-та, 2007. – Ч. 2. – С. 53–56.
5. Андерсон Т.В. Введение в многомерный статистический анализ. – М.: Физматгиз, 1963. – 500 с.
6. Орлов А.И. Задачи оптимизации и нечеткие переменные. – М.: Наука, 1980. – 302 с.
7. Эфрон Б. Нетрадиционные методы многомерного статистического анализа. – М: Финансы и статистика, 1988. – 263 с.

Поступила 10.03.2009 г.