УДК 519.21

Я. МИХАЛЕК, В. ВЕСЕЛЫЙ

СРАВНЕНИЕ ОЦЕНОК ROC-КРИВЫХ МЕТОДАМИ МОДЕЛИРОВАНИЯ[1]

Ключевые слова: *ROC-кривая, непараметрическая оценка, полупараметрическая оценка, моделирование, бинормальная модель.*

ROC-кривые часто используются в различных задачах диагностики, например для идентификации факторов, влияющих на точность диагностики, либо для определения погрешности работы диагностических систем. ROC-кривые также применяются в задачах классификации, например для диагностики раковых заболеваний.

Существует несколько разных подходов для нахождения подходящих оценок ROC-кривых для бинормальной модели. В различных приложениях очень эффективны методы оценки при малых выборках. Ниже, предполагая бинормальность модели, сравним несколько параметрических, полупараметрических и непараметрических оценок ROC-кривых на численных примерах. Параметрическая оценка использует метод обобщеных наименьших квадратов, полупараметрическая — функциональное моделирование. Непараметрическая оценка базируется на выборочной функции распределения (sdf).

ROC-КРИВАЯ И БИНОРМАЛЬНАЯ МОДЕЛЬ

ROC-кривая используется в бинарных классификаторах для разделения объектов на два класса: с положительными исходами и с отрицательными исходами соответственно. По определению ROC-кривая есть график зависимости количества верно классифицированных положительных входов от количества неверно классифицированных отрицательных входов при изменении пороговой переменной X. Предположим, что класс с положительным исходом характеризуем малыми значениями X; X — случайная переменная с функцией распределения $F(x)$, класс с отрицательным исходом характеризуем большими значениями X и функцией распределения $G(x)$. Исходя из сказанного выше, ROC-кривая определяется как $\mathrm{ROC}(t)=1-G(F^{-1}(1-t))$ для $0<t<1$, если обратная функция $F^{-1}(x)$ существует.

Обычно предполагается, что распределения $F(x)$ и $G(x)$ нормальны. Такая модель названа бинормальной моделью. Без потери общности, предположим, что F — это $N(0{,}1)$, а G — $N(\mu,\sigma^2)$, где μ и σ — неизвестные параметры. Тогда ROC-кривую запишем $\mathrm{ROC}(t)=\mathrm{ROC}(t;\mu,\sigma)$. Пусть Φ — стандартная нормальная функция распределения. Тогда ROC-кривую запишем

$$\mathrm{ROC}(t)=1-G(F^{-1}(1-t))=1-\Phi\left(\frac{\Phi^{-1}(1-t)-\mu}{\sigma}\right),\ 0<t<1. \qquad (1)$$

В дальнейшем будем изучать оценки в случае бинормальной модели.

ЭЛЕМЕНТАРНЫЕ ОЦЕНКИ ФУНКЦИИ РАСПРЕДЕЛЕНИЯ

В этом разделе предполагается, что $X_1,\dots,X_m$ обозначают случайную выборку из нормальной популяции $N(\mu,\sigma^2)$ с функцией распределения $F_0(x;\mu_0,\sigma_0)$. Далее будем применять следующие оценки.

Элементарная оценка $\widetilde{F}_0$

$$\widetilde{F}_0(x)=F_0(x;\overline{X},S^2),$$

где $\overline{X}$ и S^2 — несмещенные оценки среднего и дисперсии.

[1] Поддержано исследовательским грантом VZ04-FEM-K01-13-SJA и грантом MSMT CR согласно исследовательскому договору MSM0021622418.

Выборочная оценка $\widetilde{F}_s$

$$\widetilde{F}_s(x)=\frac{1}{m}\sum_{i=1}^{m} I_{[X_i \le x]} ,$$

где I_A — индикатор множества A.

Кусочно-линейная оценка $\widetilde{F}_l$.

Для $i=1,\ldots,m-1$ положим $c_i=\dfrac{X_{(i+1)}+X_{(i)}}{2}$ — середины интервалов $X_{(i)}+X_{(i+1)}$, и пусть $c_0=\dfrac{3X_{(1)}-X_{(2)}}{2}$, $c_m=\dfrac{3X_{(m)}-X_{(m-1)}}{2}$, где $X_{(1)}, \ldots, X_{(m)}$ — упорядоченные статистики выборки $X_1, \ldots, X_m$. Положим далее $f(x)=(m(c_{i+1}-c_i))^{-1}$ для $x\in\langle c_i, c_{i+1}\rangle$, $i=0,\ldots,m-1$, и $f(x)=0$ в противном случае. После этого кусочно-линейная оценка функции $F(x)$ определяется как

$$\widetilde{F}_l(x)=\int_{-\infty}^{x} f(t)dt .$$

Наилучшая несмещенная точечная оценка $\widetilde{F}_k$ (предложенная А.Н. Колмогоровым). Запишем ее (см. [1]):

$$\widetilde{F}_k(x)=\begin{cases} 0 \text{ для } Q(x)\le -1, \\ \dfrac{1}{2}-\dfrac{1}{2}\beta_{Q^2(x)}\left(\dfrac{1}{2},\dfrac{m}{2}-1\right) \text{ для } -1<Q(x)\le 0, \\ \dfrac{1}{2}+\dfrac{1}{2}\beta_{Q^2(x)}\left(\dfrac{1}{2},\dfrac{m}{2}-1\right) \text{ для } 0<Q(x)\le 1, \\ 1 \text{ для } Q(x)>1, \end{cases}$$

где $Q(x)=\frac{x-X}{(m-1)S}\sqrt{m}$ и $\beta_a(p,q)=\int_0^a t^{p-1}(1-t)^{q-1}dt$ — неполная бета-функция, $a\in\langle 0,1\rangle$, $p\ge 0$, $q\ge 0$.

ЭЛЕМЕНТАРНАЯ ОЦЕНКА ROC-КРИВЫХ

Рассмотрим две независимые популяции: $X_1, \ldots, X_m$ и $Y_1, \ldots, Y_n$ с функциями распределения $F(x)$ и $G(x)$ соответственно. Простой подход к оценке ROC основан на замене функций $F(x)$ и $G(x)$ их выборочными оценками $\widetilde{F}_s(x)$ и $\widetilde{G}_s(x)$. Будем заниматься оценками кривых $\mathrm{ROC}(t)=\mathrm{ROC}(t;\mu,\sigma)$, которые основаны на оценках $\widetilde{G}_0$, $\widetilde{G}_s$, $\widetilde{G}_l$, $\widetilde{G}_k$ функции распределения $G(x)$ и на оценках квантильных функций $\widetilde{F}_0^{-1}$, $\widetilde{F}_s^{-1}$, $\widetilde{F}_l^{-1}$, $\widetilde{F}_k^{-1}$, которые отвечают оценкам $\widetilde{F}_0$, $\widetilde{F}_s$, $\widetilde{F}_l$, $\widetilde{F}_k$ функции $F(x)$. Элементарные оценки (EE-оценки) ROC-кривых $\mathrm{ROC}(t)=\mathrm{ROC}(t;\mu,\sigma)$ имеют следующий вид.

EE1. Элементарная оценка $\widetilde{\mathrm{ROC}}_0(t)=1-\widetilde{G}_0(\widetilde{F}_0^{-1}(1-t))$.

EE2. Выборочная оценка $\widetilde{\mathrm{ROC}}_s(t)=1-\widetilde{G}_s(\widetilde{F}_s^{-1}(1-t))$.

EE3. Кусочно-постоянная оценка $\widetilde{\mathrm{ROC}}_c(t)=1-\widetilde{G}_c(\widetilde{F}_c^{-1}(1-t))$.

Здесь $\widetilde{F}_c^{-1}$ — кусочно-линейная аппроксимация для кусочно-постоянной квантильной функции $\widetilde{F}_s^{-1}$, полученной из точек $\widetilde{F}_s^{-1}\left(\dfrac{i}{m}\right)$, $i=0,1,\ldots,m$.

EE4. Кусочно-линейная оценка $\widetilde{\mathrm{ROC}}_l(t)=1-\widetilde{G}_l(\widetilde{F}_l^{-1}(1-t))$.

EE5. Оценка, основанная на оценке Колмогорова $\widetilde{\mathrm{ROC}}_k(t)=1-\widetilde{G}_k(\widetilde{F}_k^{-1}(1-t))$.

Рассмотрим сначала бинормальную модель (1). Определим кривую ODC(u) (Ordinal Dominance Curve), $0 \le u \le 1$, с помощью замены

$$1 - \mathrm{ROC}(t) = u, \qquad 1 - t = \mathrm{ODC}(u).$$

Тогда, используя модель (1), запишем выражение для ODC-кривой:

$$\mathrm{ODC}(u) = F(G^{-1}(u)) = \Phi(\mu + \sigma\,\Phi^{-1}(u)), \; 0 < u < 1. \tag{2}$$

В задачах оценивания использование ODC-кривых проще, чем соответствующее выражение для ROC-кривой, и взвешенный регрессивный метод для оценки неизвестных параметров μ и σ рассматриваем для ODC-модели. ODC-кривая будет задана в k точках $0 < t_1 < t_2 < \ldots < t_k < 1$. Положим

$$\beta_i = \mathrm{ODC}(t_i) = F(G^{-1}(t_i)) = \Phi(\mu + \sigma\,\Phi^{-1}(t_i)). \tag{3}$$

Тогда естественной оценкой параметра β_i будет

$$\hat{\beta}_i = \widetilde{F}_s(\widetilde{G}_s^{-1}(t_i)), \; i = 1, \ldots, k, \tag{4}$$

где $\widetilde{F}_s$ и $\widetilde{G}_s$ — выборочные функции распределения и $\widetilde{G}_s^{-1}$ — квантильная функция, соответствующая функции $\widetilde{G}_s$. Асимптотическое распределение вектора $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_k)'$ можно найти с помощью ковариационной структуры случайного процесса (броуновского моста, см. [2]). Покажем, что для фиксированных $0 < t_1 < t_2 < \ldots < t_k < 1$ и в рамках бинормальной модели имеют место:

$$\sqrt{n}\,(\hat{\beta} - \beta) \sim^A N(0, \lambda\Sigma_1 + \Sigma_2) \tag{5}$$

$$\sqrt{n}\,(\Phi^{-1}(\hat{\beta}) - \Phi^{-1}(\beta)) \sim^A N(0, \Sigma). \tag{6}$$

- $\sim^A$ обозначает асимптотическое распределение при $m \to \infty$, $n \to \infty$ и $\dfrac{m}{n} \to \lambda$,

где λ — фиксированная постоянная;

- $\Sigma = C[\lambda\Sigma_1 + \Sigma_2]C$ и $C^{-1} = \mathrm{diag}\,(\ldots, \varphi(\mu + \sigma\,\Phi^{-1}(t_i)), \ldots)$, φ — плотность $N(0,1)$;
- $\Sigma_1 = (\sigma_{ij}^1)$ и $\sigma_{ij}^1 = \min\{\beta_i, \beta_j\} - \beta_i\beta_j$;
- $\Sigma_2 = A\Sigma_0 A$, где $A = \mathrm{diag}\,(\ldots, \sigma\,\varphi(\mu + \sigma\,\Phi^{-1}(t_i)))/\varphi(\Phi^{-1}(t_i))), \ldots)$ и $\Sigma_0 = (\sigma_{ij}^0)$, $\sigma_{ij}^0 = \min\{t_i, t_j\} - t_i t_j$.

Используя (3) и (4), рассмотрим линейную регрессивную модель

$$\Phi^{-1}(\hat{\beta}_i) = \mu + \sigma\,\Phi^{-1}(t_i) + \varepsilon_i, \; i = 1, \ldots, k. \tag{7}$$

Вектор ошибки $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_k)'$ имеет распределение, определенное выражением (6). Поскольку Σ зависит от неизвестных параметров μ и σ, то для их оценки требуется итерационная процедура. Получены следующие результаты.

Обычная оценка методом наименьших квадратов для μ и σ для линейной регрессивной модели выражается формулой

$$\begin{pmatrix} \hat{\mu}_0 \\ \hat{\sigma}_0 \end{pmatrix} = (M'M)^{-1} M'\,\Phi^{-1}(\hat{\beta}),$$

где матрица M имеет вид

$$M' = \begin{pmatrix} 1 & \ldots & 1 \\ \Phi^{-1}(t_1) & \ldots & \Phi^{-1}(t_k) \end{pmatrix}$$

и вектор $\Phi^{-1}(\hat{\beta}) = (\Phi^{-1}(\hat{\beta}_1), \ldots, \Phi^{-1}(\hat{\beta}_k))'$.

Подставляя $\hat{\mu}_0$ и $\hat{\sigma}_0$ для μ и σ в выражение (6) и обозначая соответствующее Σ как $\hat{\Sigma}$, получаем, что одношаговая обобщенная оценка методом наименьших квадратов определяется формулой

$$\begin{pmatrix} \hat{\mu} \\ \hat{\sigma} \end{pmatrix} = (M'\,\hat{\Sigma}^{-1} M)^{-1} M'\,\hat{\Sigma}^{-1} \Phi^{-1}(\hat{\beta}).$$

Процедура может повторяться итеративно, что приводит к взвешенной регрессивной оценке ODC, часто достаточно одного шага. Заметим, что асимптотика распределения $(\hat{\mu},\ \hat{\sigma})$ имеет вид $\sqrt{n}\begin{pmatrix}\hat{\mu}-\mu\\ \hat{\sigma}-\sigma\end{pmatrix} \sim^A \ N(0,(M'\,\Sigma^{-1}M)^{-1})$.

Преобразование Φ^{-1} в левой части регрессивной модели (7) может привести к неустойчивой взвешенной регрессивной оценке, если размеры выборок m и n уменьшаются. Для улучшения оценки в [2] предложена адаптивная процедура выбора точек $t_1,\ldots,t_k$, согласно которой величины t_i сконцентрированы в окрестности быстрого роста оценки ROC-кривой. Процедура имеет следующий вид:

1) фиксируем положительное целое q;

2) принимаем за $t_1 = \min\{j/n\colon \widetilde{F}_s(\widetilde{G}_s^{-1}(j/n)) \geq q/m,\ j=1,\ldots,n\}$;

3) находим $t_{i+1} = \min\{j/n\colon \widetilde{F}_s(\widetilde{G}_s^{-1}(j/n)) - \widetilde{F}_s(\widetilde{G}_s^{-1}(t_i)) \geq q/m,\ j=1,\ldots,n\}$ для $i=1,\ldots,k(q)$, где $k(q)$ — наибольшее целое такое, что $t_{k(q)} < 1$.

В дальнейшем будем использовать описанную улучшенную адаптивную взвешенную регрессивную оценку. Выбор параметра q существенен, так как связан с количеством точек дискретизации (узлов сети) k. Для малых q величина k больше. Автоматизация выбора q является сложной задачей. Далее выбор q проводится на основе эксперементально выведенной формулы $q \approx 1{,}5n^{\frac{1}{3}} \log_{10}(n)$. Соответсвующую оценку назовем взвешенной регрессивной оценкой (WRE) и обозначим $\widetilde{\mathrm{ROC}}_{wr}(t)$.

ФУНКЦИОНАЛЬНАЯ МОДЕЛЬ ДЛЯ ОЦЕНКИ ROC-КРИВОЙ, ОСНОВАННАЯ НА АТОМАРНОМ РАЗЛОЖЕНИИ

Сначала рассмотрим теоретический подход к функциональному моделированию специальными атомарными формулами (CAD).

Принципы CAD. Пусть (X,ρ) — линейное функциональное пространство, $H \subset X$ — его сепарабельное подпространство и $\rho\colon X \times X \to R$ — метрика на X. Для функции $f \in X$ определим ρ-приближение как элемент $\hat{f} \in H$, минимизирующий (максимизирующий) $\rho(f,\hat{f})$. Пусть $T\colon P \to H$ — линейный сюрьективный оператор из пространства параметров $P \subseteq \ell^2(J) := \left\{\{\xi_j\}_{j\in J} \mid \sum_{j\in J} |\xi_j|^2 < \infty, \xi_j \in C\right\}$ высокой размерности $\dim P \leq \mathrm{card}(J) \leq \aleph_0$ на H. Возьмем канонический ортонормальный базис $E := \{\varepsilon_j\}_{j\in J}$ на $\ell^2(J)$, $\varepsilon_j = \{\delta_{jk}\}_{k\in J}$, в силу определения T имеем разложение

$$\hat{f} = T\xi = \mathrm{T}\left(\sum_{j\in J} \xi_j\,\varepsilon_j\right) = \sum_{j\in J} \xi_j T\varepsilon_j =: \sum_{j\in J} \xi_j \phi_j\,, \tag{8}$$

где $\phi_j := T\varepsilon_j$ называют атомами, а само разложение $\hat{f} = \sum_{j\in J} \xi_j \phi_j$ — атомарным в терминах словаря $\Phi := \{\phi_j\}_{j\in J}$. Параметр ξ — параметрическое представление $\hat{f}$ в терминах словаря $\Phi := \{\phi_j\}_{j\in J}$. Если J бесконечно, то суммирование в (8) следует понимать как безусловно сходящийся ряд. В случае конечных J с большим количеством элементов, как правило, существует не одно параметрическое представление ξ для $\hat{f}$, удовлетворяющее условию $\mu\left(\hat{f},\ \sum_{j\in J} \xi_j \phi_j\right) < \varepsilon$, где μ — некоторая метрика. В этом случае как разложение, так и параметризацию называют избыточными.

Выбирая желаемую точность $\varepsilon > 0$, найдем конечное наименьшее подмножество $F^* \subset J$ такое, что $\mu\left(\hat{f}, \sum_{j \in F^*} \xi_j^* \phi_j\right) < \varepsilon$. Выражение $\sum_{j \in F^*} \xi_j^* \phi_j$ назовем разреженным ε-субоптимальным атомарным разложением $\hat{f}$, а $\xi^* = \{\xi_j^*\}_{j \in F^*}$ — разреженной ε-субоптимальной параметризацией. Отметим, что в стандартных постановках задач обычно используют гильбертовы пространства [3, 4].

Множество алгоритмов было предложено разными авторами (см. [3]) для поиска разреженных представлений по избыточным. Используем универсальную многошаговую итеративную процедуру (реализованную как функцию MATLAB), устойчивую по отношению к ошибкам аппроксимаций в некорректных плохо обусловленных обратных задачах. Процедура базируется на алгоритме BPA (Basis Pursuit Algorithm), предложенном в [3] для конечномерного случая и распространенного на случай функционального пространства в [4]. Алгоритм BPA имеет следующий вид:

1) найти $\xi_\varepsilon = \arg\min \|\xi\|_1$ при условии $\mu\left(\hat{f}, \sum_{j \in J} \xi_j \phi_j\right) < \varepsilon$;

2) выбрать $\delta > 0$ как можно большим при условии $\mu\left(\hat{f}, \sum_{j \in F^*} \xi_j^* \phi_j\right) < \varepsilon$, где $F^* = \{j \in J \,\big|\, |\xi_{\varepsilon j}| \geq \delta\}$ и ξ^* минимизирует $\mu\left(\hat{f}, \sum_{j \in F^*} \xi_j^* \phi_j\right)$.

В работе [5] описанный метод используется для нового подхода к сглаживанию ядер, который имеет преимущества перед стандартными методами.

Функциональные аппроксимации ROC-кривой. Положим $X = L^2[0,1]$, $\phi_j(t) = \mathrm{ROC}(t; \mu_j, \sigma_j) - t$, $[\mu_j, \sigma_j] \in M \times S$ для $j \in J$, где M и S — подходящие (т.е. однородные) разбиения доверительных интервалов $I_{1-\alpha}(\mu)$ и $I_{1-\alpha}(\sigma)$ соответственно ($\alpha = 0{,}05$). Выбирая сетки размера 30–50 точек, получаем словарь $\Phi = \{\phi_j(t)\}_{j \in J}$, состоящий из 900–2500 атомов, поскольку $\mathrm{card}(J) = \mathrm{card}(M) \times$ $\times \mathrm{card}(S)$. Ниже использовано $30 \times 30 = 900$ атомов для моделирования. Вычитание t обеспечивает равенство нулю в граничных точках как для атомов, так и для элементов их линейных комбинаций. Аналогично положим $f(t) := \widetilde{\mathrm{ROC}}(t) - t$, где $\widetilde{\mathrm{ROC}}(t)$ — кусочно-линейная оценка ROC. Выбор метрики ρ зависит от конкретной постановки задачи: например, $\rho(f, \hat{f}) = \|f - \hat{f}\| \to \min$, где $\|\cdot\|$ — норма в X, либо метрика Леви, т.е. метрика, полученная на основе нормы $L^\infty[0, \sqrt{2}]$ после поворота на 45° по часовой стрелке. Также возможны вариации на основе нормы L^p. Такие метрики позволяют избавиться от асимметричности в границах интервала.

В следующем разделе используется норма пространства $L_2[0,1]$. Полученная ROC-оценка называется функциональной оценкой (FE) и обозначается $\widetilde{\mathrm{ROC}}_F(t)$. После того как найдено разреженное разложение $\hat{f}^* := T\xi^*$, положим $\mathrm{ROC}^*(t) = t + \hat{f}^* =$ $= t + \sum_{j \in F^*} \xi_j^* (\mathrm{ROC}(t; \mu_j, \sigma_j) - t)$. Эту оценку назовем разреженной функциональной оценкой (SFE) и обозначим $\widetilde{\mathrm{ROC}}_{SF}(t)$.

МОДЕЛИРОВАНИЕ

Описанные выше оценки сравниваются на примере трех теоретических бинормальных моделей, в которых $(\mu, \sigma) = (2{,}1)$, $(1{,}1)$ и $(0{,}1)$. Размеры выборки $m = n = 10$, 5, 30, 50, 100, 500, число численных экспериментов для каждой модели 100, расстояние между точными значениями $\mathrm{ROC}(t;\mu,\sigma)$ и оценкой $\widetilde{\mathrm{ROC}}(t)$ измерялось с помощью метрики Леви:

$$\rho_\infty^{(L)}:=\|\mathrm{ROC}_L-\widetilde{\mathrm{ROC}}_L\|_\infty:=\sup_{u\in[0,\sqrt{2}]}|\mathrm{ROC}_L(u)-\widetilde{\mathrm{ROC}}_L(u)|,$$

где ROC_L и $\widetilde{\mathrm{ROC}}_L$ — функции, полученные из функций ROC и $\widetilde{\mathrm{ROC}}$ соответственно после поворота системы координат $\pi/4$ на по часовой стрелке. Супремум считается численно на неоднородной сетке $0<t_0<t_1<\ldots<t_N=1$, полученной трансформацией $t_i=x_i^{2,5}$ из однородной сетки $x_i=0{,}1+i(0{,}9/N)$, $i=0,1,\ldots,N$, что позволяет лучше адаптировать сеть к показателям графика ROC. Для метрики Леви эта сеть заменяется на другую для каждой ROC_L и $\widetilde{\mathrm{ROC}}_L$. Поэтому произвели повторную выборку для $\widetilde{\mathrm{ROC}}_L$ в точках новой сети для ROC_L, взяв кусочно-линейную аппроксимацию.

После моделирования среднее и стандартное отклонения для метрики Леви подсчитаны для каждой модели. Моделирование проведено для EE1–EE5 элементарных оценок $\mathrm{ROC}(t)$, для WRE взвешенной регрессивной оценки, FE и SFE функциональной оценки. Последние две были приближены с помощью кусочно-линейной оценки, что несколько улучшило результат по сравнению с выборочными ROC-оценками. Результаты представлены на рис. 1–6 соответственно для $m=n=10$, 15, 30, 50, 100, 500.

На рис. 1, 3, 5 на вертикальной оси подсчитаны средние значения метрики Леви, полученные для данного числа симуляций и заданного распределения. На рис. 2, 4, 6 на вертикальной оси отображены стандартные отклонения метрики Леви, на горизонтальной для рис. 1–6 — размеры выборки.

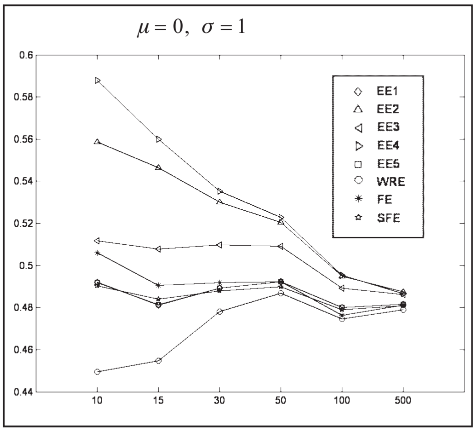

Рис. 1

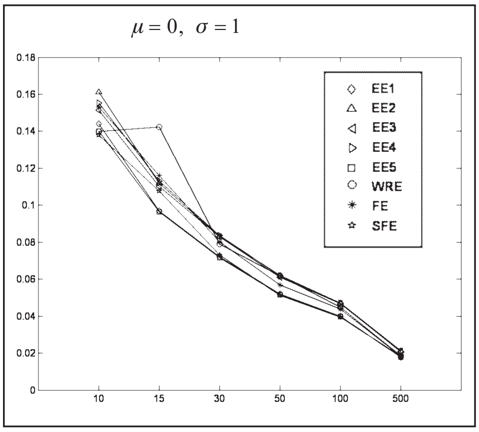

Рис. 2

Качество простой оценки близко к оценке Колмогорова и может считаться хорошей оценкой с точностью до метрики, если средние значения популяций различны. Стандартные отклонения оценок близки между собой. Хорошие свойства параметрическицх оценок — следствие бинормальности модели. Наоборот, наибольшее значение среднего метрики Леви наблюдается при непараметрических ROC-оценках EE2–EE4, которые не извлекают пользы из предположения бинормальности. Качество кусочно-линейной оценки несколько лучше, чем качево выборочной ROC-оценки. Эту оценку можно улучшить за счет функционального моделирования: гладкие оценки можно получить для малых выборок без предположения нормальности.

Функциональные оценки похожи или ненамного хуже взвешенных регрессионных оценок, которые существенно зависят от предположения нормальности. Недостатком взвешенных регрессионных оценок является значительное стандартное отклонение и сложность выбора параметра q. Автоматический выбор q не дает результата в 30 % случаев для малых $m=n=15$ и $\mu=2$, $\sigma=1$. Число сбоев уменьшается при уменьшении среднего и увеличении выборки. Функциональная модель гораздо устойчивей, поскольку сходимость алгоритма отсутствовала лишь в шести случаях из 300.

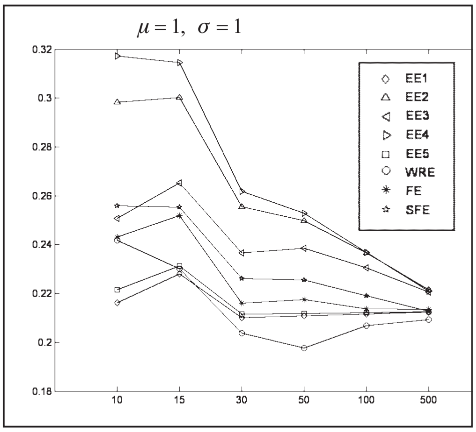

Рис. 3

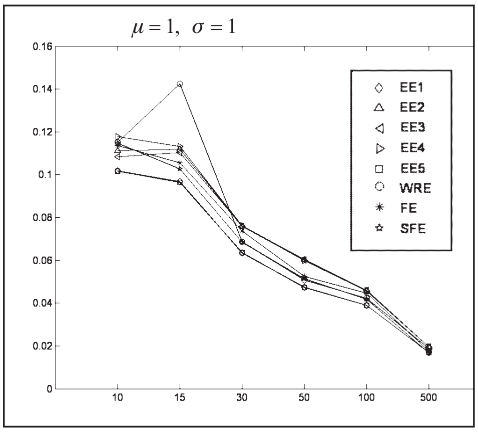

Рис. 4

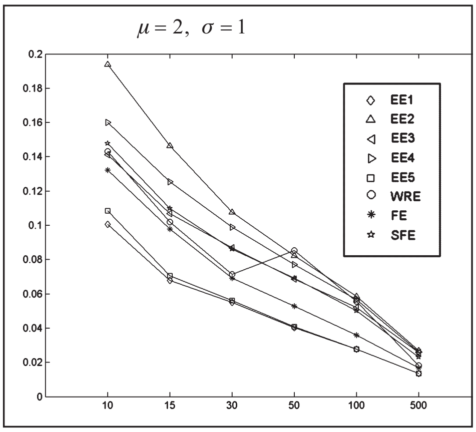

Рис. 5

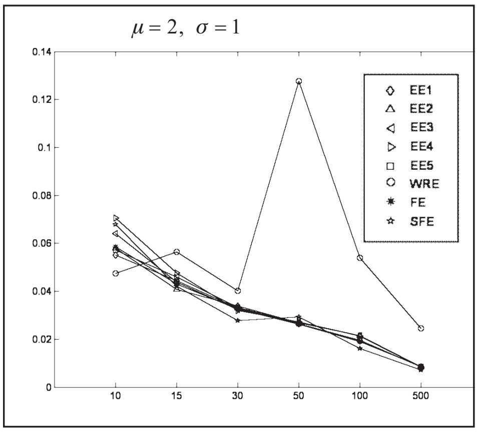

Рис. 6

Из проведенного анализа путем моделирования можно сделать вывод: для больших выборок при условии, что абсолютные значения отклонений средних для популяций велики (что приводит к достаточно быстрому росту ROC-кривой, см. пример с параметрами $\mu = 2$, $\sigma = 1$), элементарная оценка предпочтительнее, чем взвешенная регрессионная оценка WRE. В противном случае для небольших выборок и небольших абсолютных значений отклонений средних для популяций (когда ROC-кривая близка к диагонали единичного квадрата) WRE более предпочтительна. Выборочные стандартные отклонения оценок FE и SFE весьма малы для всех перечисленных выше случаев.

СПИСОК ЛИТЕРАТУРЫ

1 M i c h a l e k J. and V e s e l y V. The ROC and ODC curve estimators in binomial model based on the best unbiased estimator of CDF // XXIII Intern. Colloq. on the Acquisition Process Management. — University of Defence Brno, 2005. — P. 34.

2 H s i e h F. and T u r n b u l l B.W. Nonparametric and semiparametric estimation of receiver operating characteristic curve // The Annals of Statistics. — 1196. — **24**, N 1. — P. 25–40.

3 C h e n S.S., D o n o h o D.L. and S a u n d e r s M.A. Atomic decomposition by basis pursuit. SIAM J. Sci. Cornput., **20(l):33-61**, 1998. REp. in SIAM review. — 2001. — **43**, N 1. — P. 129–159.

4 V e s e l y V. Hilbert-space techniques for spectral representation in terms of overcomplete bases // Proceedings of the summer school DATASTAT’2001, Cihak near Zamberk. Folia Fac. Sci. Nat. Univ. Masaryk. Brunensis, Mathematica. Dept. of Appl. Math., Masaryk Univ of Brno. Czech Rep., 2002. — **11**. — P. 259–273

5. Z e l i n k a J., V e s e l y V. and H o r o v a I. Comparative study of two kernel smoothing techniques // Proceedings of the summer school DATASTAT’2003, Svratka, Folia Fac. Sci. Nat. Univ. Masaryk. Brunensis, Mathematica. Dept. of Appl. Math. Masaryk Univ. of Brno, Czech Rep., 2004. — **15**. — P. 419–436.

Поступила 22.04.2009