응용통계연구 (2011)
24(6), 987–994
DOI: http://dx.doi.org/10.5351/KJAS.2011.24.6.987

ROC 함수 추정

홍종선[1] · Lin, Mei Hua[2] · 홍선우[3]

[1]성균관대학교 통계학과, [2]성균관대학교 응용통계연구소, [3]성균관대학교 응용통계연구소

(2011년 7월 접수, 2011년 10월 채택)

요 약

모집단이 부도와 정상상태로 구분되는 신용평가 관점에서 부도와 정상 상태의 조건부 누적분포함수를 추정하는 방법으로 정규혼합 분포추정과 kernel density estimation을 이용하는 분포추정을 고려한다. 정규혼합 분포의 모수를 EM 알고리즘을 사용해 추정하고, KDE 방법에서는 많이 사용하는 다섯 종류의 커널 함수와 네가지의 띠폭을 이용한다. 그리고 추정한 분포로부터 구한 각각의 ROC 함수를 구한다. 추정한 분포들의 적합도를 비교 분석하고, 이를 바탕으로 구한 ROC 곡선의 성과를 비교 토론한다. 본 연구에서는 KDE 방법으로 추정한 분포함수가 더 적합하고, 추정한 정규혼합 분포를 이용한 ROC 함수가 더 좋은 성과를 나타내는 것을 발견하였다.

주요용어: 띠폭, 분포추정, 성과, ROC 함수, 적합도, 정규혼합, 커널.

1. 서론

ROC 곡선은 성과(performance)를 기반으로 분류자(classifiers)를 시각화하고, 조직화하고, 선정하는 방법이며 (Fawcett, 2003), 진단 시스템의 동작을 시각화하고 분석하는데 사용이 확장되고 (Swets, 1988), 이항적 결정 규칙의 성과를 요약하는데 사용된다 (Lloyd와 Yong, 1999). ROC 곡선은 분류자의 'hit rate'(이익)과 'false alarm rate'(비용) 사이에 교환(tradeoff)을 묘사하기 위해 신호탐지 이론에서 오래전부터 사용되었다 (Egan, 1975; Swets 등, 2000). ROC 곡선의 특성에 관한 설명과 실증연구에서 ROC 분석을 응용하는데 관련된 정보는 Fawcett (2003)과 Provost와 Fawcett (1997, 2001), 홍종선과 최진수 (2009), 홍종선 등 (2010)에서 발견할 수 있다.

본 연구에서는 진단 결과를 의학적 관점이 아닌 신용평가(credit evaluation)적 관점으로 논의하기 위하여 차주(borrower)는 스코어(score) 확률변수 S와 모수공간 $D = \{d, n\}$에 의해서 특성을 나타낸다고 가정하자. 여기서 확률변수 S는 대출기관에서 차주의 신용가치를 예상하기 위해 차주에게 부여한 연속형 값을 갖는 스코어이다. 스코어 변수 S를 통하여 대출기관은 궁극적으로 차주의 신용가치에 관한 정보에 의거하여 차주의 미래상태 D를 예상하는 것이다. D의 원소인 d는 default(부도) 또는 disease(질병)을 나타내며 다른 원소인 n은 non-default(정상) 또는 non-disease(정상)으로 설정한다. 차주의 모집단은 두 개의 부모집단으로 구성한다고 가정한다. 부모집단은 미래시점에 대출상환능력이 없는 부도상태와 대출상환능력이 있는 정상상태로 구분된다. 차주의 모수 D가 d일 때($D = d$) 부도차주의 모집단에 속하고, 차주의 D가 n일 때($D = n$) 정상차주의 모집단에 속한다. 그리고 주어진 모수공간 D에서 스코어 변수의 조건부 누적분포함수를 각각 $F_d(s) = P(S \le s \,|\, D = d)$와 $F_n(s) = P(S \le s \,|\, D = n)$

[1]교신저자: (110-745) 서울시 종로구 명륜동 3-53, 성균관대학교 경제학부 통계학전공, 교수.
E-mail: cshong@skku.ac.kr

으로 나타낼 때, 스코어 S의 분포함수는 다음과 같이 표현한다.

$$F(s) = \gamma F_d(s) + (1-\gamma)F_n(s), \tag{1.1}$$

여기서 γ는 전체부도율(total probability of default)이며, $\gamma = P(D = d)$이다.

ROC 곡선은 각 절단점(cut-off value, threshold)의 스코어에서 얻는 비율들로 구성되어 있으며, 실제 부도를 부도로 정확히 예측하는 비율 TPR(true positive rate)과 실제정상을 부도로 잘못 예측하는 비율 FPR(false positive rate)을 각각 X축과 Y축 좌표에 대응시킨 그래프로 다음과 같이 표현된다 (상세한 정보는 Tasche (2006) 참조).

$$(F_n(s), F_d(s)), \quad s \in (-\infty, \infty),$$
$$(u, \mathrm{ROC}(u)), \quad u \in (0, 1),$$

여기서 $\mathrm{ROC}(u) = F_d(F_n^{-1}(u))$이다.

Pepe (1998, 2003)는 정규분포를 가정하여 ROC(u)를 이봉정규(binormal)형태로 표현하였다. 실제 자료는 꼬리가 두껍고 대칭적이 아닌 성격을 지니고 있어서 정규분포에 적합하는 경우는 쉽게 찾아보기 어렵다. 기존의 연구에서는 $F_n(\cdot)$과 $F_d(\cdot)$를 추정하는 방법으로 비모수적 방법과 회귀모형을 이용한 모수적 방법 (Pepe, 1998, 2003) 그리고 준모수적 방법 (McCullagh와 Nelder, 1983) 등이 있다.

본 연구에서는 자료에 적합한 분포함수를 추정하기 위하여 자료가 정규혼합(normal mixture) 분포를 따른다는 가정 하에서 혼합된 정규분포 추정과 Kernel Density Estimation(이하 KDE)을 이용하여 분포를 추정하는 방법을 연구한다. 정규혼합 분포의 모수는 EM 알고리즘을 사용해 추정하고, KDE 방법에서는 많이 사용하는 다섯 종류의 커널 함수와 네가지의 띠폭(bandwidth) 중에서 최적을 선정하여 분포를 추정한다. 그리고 추정한 분포로부터 구한 각각의 ROC 함수를 구하고, ROC 곡선의 성과를 비교한다.

본 연구의 2절에서는 자료에 가장 적합한 분포함수를 추정하기 위하여 KDE 방법을 간략히 설명하고, EM 알고리즘을 이용하여 정규혼합분포의 모수 및 분포추정을 정리한다. 3절에서는 두 종류의 실증예제를 통해 추정한 분포가 자료에 적합한지를 살펴보고, ROC 곡선을 구하여 성과를 비교 분석하기 위하여 AUC(area under ROC curve)를 구하여 토론한다. 마지막으로 4절에서는 결론을 유도한다.

2. 분포추정

2.1. Kernel Density Estimation

주어진 확률표본(random sample) $S_1, \ldots, S_n$의 확률밀도함수를 $f(s)$라 하자. Rossenblatt (1956)에 의해 제안된 커널밀도함수의 추정(KDE)은 다음과 같이 표현된다.

$$\hat{f}(s) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{s - S_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n} k_h(s - S_i),$$

여기서 함수 $k(\cdot)$을 커널(kernel)이라 하며 $k_h(y) = k(y/h)/h$으로 다음 조건을 만족한다.

$$\int_{-\infty}^{\infty} k(s)ds = 1, \quad \int_{-\infty}^{\infty} sk(s)ds = 0, \quad \int_{-\infty}^{\infty} s^2k(s)ds = \sigma_k^2 > 0.$$

$\{S_{d1}, S_{d2}, \ldots, S_{dn}\}$과 $\{S_{n1}, S_{n2}, \ldots, S_{nm}\}$을 각각 $f_d(\cdot)$와 $f_n(\cdot)$로부터의 표본크기 n과 m의 확률표본이라 하면, Zou 등 (1997)과 Lloyd (1998)이 제안한 ROC 함수의 커널 추정량은 식 (2.1)과 같이 정의

된다.

$$\widetilde{\mathrm{ROC}}(\cdot) = \tilde{F}_d\left(\tilde{F}_n^{\,-1}(\cdot)\right), \tag{2.1}$$

여기서 $\tilde{F}_d(s) = \sum_{i=1}^{n} K((s - S_{di})/h_d)/n$와 $\tilde{F}_n(s) = \sum_{j=1}^{m} K((s - S_{nj})/h_n)/m$는 $F_d(\cdot)$와 $F_n(\cdot)$의 커널 추정량이다. 그리고 $K(s) = \int_{-\infty}^{s} k(u)du$은 커널 누적분포함수이며, h_d과 h_n는 띠폭으로 Lloyd와 Yong (1999) 그리고 Hall과 Hyndman (2003)이 제시한 네가지의 띠폭선택(Normal, Lloyd, Plug-in, Mix)을 이용하여 최적의 띠폭선택방법을 사용한다. 커널밀도함수 $k(\cdot)$는 함수의 모양에 따라서 다섯 가지의 Gaussian, Rectangular, Triangular, Epanechnikov, Biweight으로 분류한다 (Silverman, 1986).

2.2. 정규혼합 방법

홍종선과 이원용 (2011)의 연구에서는 정규혼합 분포를 이용한 ROC 곡선이 자료에 가장 잘 적합함을 보였다. 현실적인 상황에서 분포함수를 모르고 평균함수에 영향을 주는 공변량도 고려하지 않는다고 가정한다. 이런 상황에서 자료에 가장 적합한 분포함수를 추정하는 편리한 방법은 정규혼합(normal mixture) 분포를 이용한다.

스코어에 대한 조건부 분포함수 $F_d(\cdot)$와 $F_n(\cdot)$를 각각 p개와 q개의 정규분포함수의 선형결합(linear combination)으로 구성되었다고 가정하고 다음과 같이 표기한다.

$$\hat{F}_d(s) = \sum_{i=1}^{p} \alpha_i \Phi\left(s; \mu_{d_i}, \sigma_{d_i}^2\right), \quad \hat{F}_n(s) = \sum_{j=1}^{q} \beta_j \Phi\left(s; \mu_{n_j}, \sigma_{n_j}^2\right),$$

여기서 $\sum_{i=1}^{p} \alpha_i = 1$, $\sum_{j=1}^{q} \beta_j = 1$ 그리고 $\Phi(s; \mu, \sigma^2)$는 평균과 분산이 각각 μ, σ^2인 정규분포함수를 나타낸다.

EM 알고리즘은 다양한 모형에 적용될 수 있으며 혼합모형에서의 모수 추정에서도 유용하게 사용되고 있다 (Mclachlan과 Krishnan, 1997). Everitt (1984)는 두 정규분포가 혼합되었을 때, 모수를 추정하는 다양한 알고리즘 중에서 EM 알고리즘의 우수성을 토론하고, Aitkin과 Wilson (1980)은 다양한 평균과 분산의 경우의 혼합모형(mixture model)에서 EM 알고리즘을 연구하였다. EM 알고리즘은 여러 분야의 혼합분포에 대한 모수추정에서도 많이 사용한다.

본 연구에서 두 정규분포가 혼합된 즉 $p = q = 2$인 간단한 경우를 고려하고, EM 알고리즘을 이용하여 정규혼합 분포의 모수를 추정한다. 추정된 분포함수를 바탕으로 ROC 함수는 식 (2.2)처럼 표현된다.

$$\widehat{\mathrm{ROC}}(\cdot) = \hat{F}_d\left(\hat{F}_n^{-1}(\cdot)\right). \tag{2.2}$$

3. 실증예제

3.1. DP21 자료

첫 번째 자료는 1,848명의 사람들 중 489명의 사람은 청각장애가 있고 1,359명의 정상인 자료(이하 DP21)를 분석한다 (Pepe, 2003). 이 자료는 식 (1.1)과 같은 혼합분포를 따르며 $\gamma = 489/1848$이다. 이 자료에 적합한 정규혼합 분포를 2.2절에서 언급한 가정 하에서 EM 알고리즘으로 추정한 모수는 다음과 같다.

$$F_d(s) = 0.99\,\Phi\left(s, -11.38, \sqrt{80.96}\right) + 0.01\,\Phi\left(s, 10.56, \sqrt{3.83}\right),$$
$$F_n(s) = 0.23\,\Phi\left(s, -0.90, \sqrt{78.40}\right) + 0.77\,\Phi\left(s, 8.36, \sqrt{35.11}\right).$$

표 **3.1.** DP21 자료의 K-S 통계량

방법	K-S 통계량
KDE	0.0185
정규혼합	0.0357

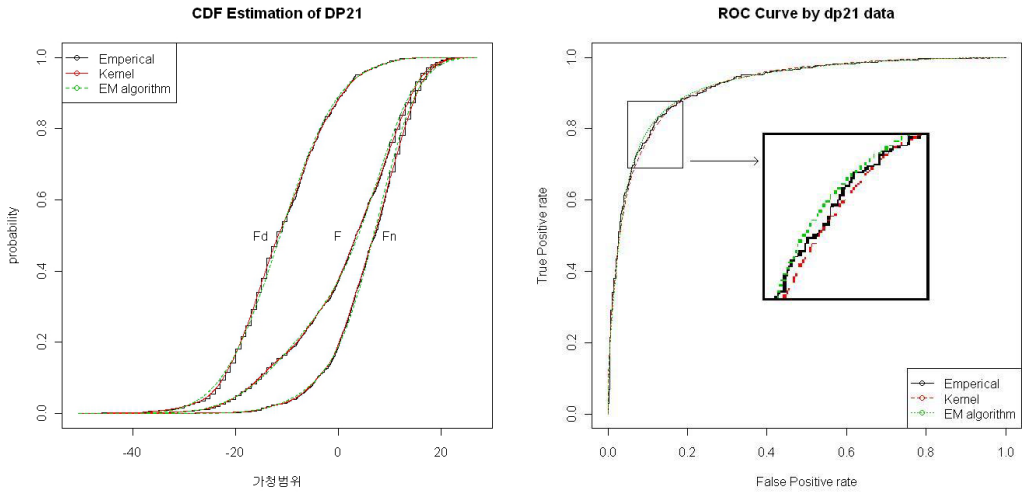

그림 **3.1.** DP21 자료의 CDF와 ROC 곡선

또한 DP21 자료의 분포를 커널밀도함수를 추정하기 위하여 다섯 가지의 커널함수(Gaussian, Rectangular, Triangular, Epanechnikov, Biweight)와 네 가지의 띠폭(Normal, Lloyd, Plug-in, Mix)을 사용하였다.

경험적 누적분포함수에 대하여 EM 알고리즘을 사용하여 추정한 정규혼합 분포함수와 KDE 방법을 이용하여 추정한 함수와의 적합도를 알아보기 위하여 콜모고로프-스미르노프(Kolmogorov-Smirnov; K-S) 검정을 실시한 결과를 표 3.1에 정리하였다. 여러 종류의 커널함수와 띠폭 중에서 K-S 통계량이 제일 작은 값을 갖는 Gaussian 커널함수와 Mix 띠폭을 이용한다.

표 3.1의 K-S 통계량은 임계값(0.0447)보다 작아서 두 방법에 의해 추정한 분포함수가 경험적 누적분포함수에 적합하다고 판단할 수 있으며, KDE 방법을 사용했을 때의 K-S 통계량값은 정규혼합 방법에 의한 K-S 통계량값보다 작기 때문에 경험적 누적분포함수에 더욱 적합한 분포함수는 KDE 방법으로 추정한 경우라고 파악할 수 있다. DP21 자료 전체 그리고 장애와 정상의 경험적 누적분포함수와 정규혼합 방법과 KDE 방법으로 추정된 각각의 분포함수를 그림 3.1의 왼쪽에 나타내었다. 경험적 누적분포함수는 계단식 실선으로 나타나고 정규혼합 분포함수는 점선으로, 커널 분포함수는 실선으로 표현하였다. 정규혼합 분포함수와 커널 분포함수 모두가 중복되기 때문에 경험적 분포함수에 매우 적합하다고 판단할 수 있다.

DP21 자료의 경험적 ROC 함수 ROC(·)와 KDE 방법을 이용하여 추정한 식 (2.2)의 ROC 함수 $\widehat{\mathrm{ROC}}(\cdot)$, 정규혼합 방법을 이용하여 추정한 식 (2.2)의 ROC 함수 $\widehat{\mathrm{ROC}}(\cdot)$를 구하고 그림 3.1의 오른쪽 그림에 구현하였다. 그림 3.1의 오른쪽을 통해 경험적 ROC 곡선을 기준으로 두 종류의 방법으로 추정한 ROC 곡선을 살펴보면, 경험적 ROC 곡선에 매우 유사하게 중복되며 큰 차이가 없음을 파악할 수 있다.

표 **3.2.** DP21 자료의 AUC 통계량

모형	경험적 방법	KDE	정규혼합 방법
AUC 통계량	0.9217	0.9196	0.9213
(경험적 방법과 차이)		(0.0021)	(0.0004)

표 **3.3.** 외감기업 자료의 K-S 통계량

방법	K-S 통계량
KDE	0.0042
정규혼합	0.0106

경험적 ROC 곡선과 비교하여 얼마나 근사한지를 판단하기 위하여 각각의 ROC 곡선의 AUC를 구하여 표 3.2에 요약하였다. AUC에 대하여는 많은 문헌이 있으나 특히 Joseph (2005)에 자세히 설명되어 있다. 표 3.2의 결과를 살펴보면, 정규혼합 방법을 이용한 ROC 곡선의 AUC가 KDE 방법에 의한 AUC보다 경험적 분포함수에 의한 AUC에 더욱 근사하기 때문에 정규혼합 방법을 이용하여 추정한 정규혼합 분포가 더 적합하다고 판단할 수 있다.

3.2. 외감기업 자료

두 번째 자료는 1994년부터 2005년까지 한국기업 중에서 외부감사를 받는 기업 중 총자산 규모가 4500억원 이상인 기업에 대한 자료(이하 외감기업 자료)이며, 총표본수는 4,134 ($n = 238$, $m = 3,896$)이며 $\gamma = 238/4134$이다. 이 자료에 적합한 정규혼합분포를 EM 알고리즘으로 추정한 결과는 다음과 같다.

$$F_d(s) = 0.18\ \Phi\left(s, 10.03, \sqrt{11.03}\right) + 0.82\ \Phi\left(s, 25.85, \sqrt{116.17}\right),$$
$$F_n(s) = 0.25\ \Phi\left(s, 33.64, \sqrt{101.15}\right) + 0.75\ \Phi\left(s, 57.88, \sqrt{153.85}\right).$$

외감기업 자료에 분포를 KDE 방법을 사용하기 위하여 3.1절과 같이 다섯 가지의 커널함수 중 Gaussian 커널함수를 사용하였으며, 네 종류의 띠폭 중에서 Mix를 이용한다. 외감기업 자료의 경험적 누적분포함수에 대하여 EM 알고리즘으로 추정한 정규혼합 분포함수와 KDE 방법을 이용하여 추정한 커널분포함수와의 적합성을 파악하기 위하여 K-S 검정 통계량을 표 3.3에 정리하였다.

표 3.3을 살펴보면 두 방법으로 추정한 분포함수에 대한 K-S 통계량값이 임계값(0.0299)보다 모두 작기 때문에 경험적 누적분포함수에 적합하고, KDE 방법을 사용했을 때의 K-S 통계량값은 정규혼합 방법에 의한 K-S 통계량값보다 작기 때문에 경험적 누적분포함수에 더욱 적합한 분포함수는 KDE 방법으로 추정한 분포임을 파악할 수 있다.

그림 3.2의 왼쪽에서는 외감기업 자료의 부도와 정상 그리고 전체자료 각각에 대응하는 경험적 누적분포함수를 계단식 선으로 표현하고, 정규혼합 방법과 KDE 방법으로 추정한 각각의 분포함수를 각각 점선과 실선으로 구현하였다. 정규혼합 분포함수와 커널 분포함수가 중복되기 때문에 경험적 분포함수에 매우 적합하다고 판단한다.

외감기업 자료의 경험적 ROC 함수와 KDE 방법을 이용하여 추정한 ROC 함수, 추정한 정규혼합 분포로부터의 ROC 함수를 구하고 그림 3.2의 오른쪽에 표현하였다. 정규혼합 방법을 이용한 ROC 곡선이 KDE 방법을 이용한 것보다 경험적 방법에 의한 ROC 곡선에 적합함을 파악할 수 있다.

경험적 ROC 곡선과 비교하기 위하여 각각의 AUC를 구하여 비교하고자 표 3.4에 정리하였다. 정규혼

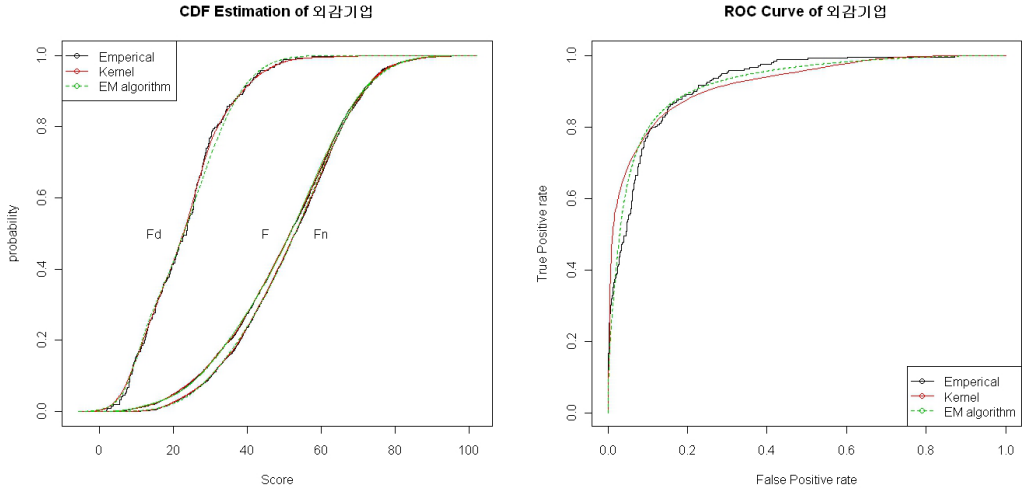

그림 **3.2.** 외감기업 자료의 CDF와 ROC 곡선

표 **3.4.** 외감기업 자료의 AUC 통계량

모형	경험적 방법	KDE	정규혼합 방법
AUC 통계량	0.9227	0.9209	0.9224
(경험적 방법과 차이)		(0.0019)	(0.0003)

합 방법을 이용한 AUC가 KDE 방법에 의한 AUC보다 경험적 분포함수에 의한 AUC에 더욱 근사하기 때문에 3.1절에서와 같이 정규혼합 분포가 더 적합하다고 판단한다.

4. 결론

본 연구에서는 일반적으로 분포함수가 주어지지 않은 경우를 고려하였으며, 두 종류의 실증예제 자료에 대하여 KDE 방법으로 추정한 분포함수와 정규혼합 방법을 이용하여 추정한 분포함수를 비교 분석하였다. 추정한 분포함수가 적합한지를 판단하기 위하여 K-S 검정을 실시하였는데 그 결과 KDE 방법과 정규혼합 방법을 이용한 방법으로 추정한 분포함수 모두 적합하나 KDE 방법으로 추정한 분포가 실제 분포에 더 적합함을 탐색하였다.

그리고 두 종류의 방법으로 추정한 분포함수를 바탕으로 ROC 함수를 구하고 경험적 분포함수로 구한 ROC 곡선에 적합한지를 AUC 통계량을 구하여 비교하였다. 정규혼합 분포함수로 구한 ROC 곡선의 AUC가 KDE 방법으로 추정한 AUC보다 더 작으며 이 결과를 바탕으로 정규혼합 방법으로 추정한 정규혼합분포가 경험적인 ROC 곡선에 잘 적합함을 발견하였다.

그러므로 분포함수 추정은 KDE 방법과 정규혼합 방법 모두 좋은 방법이며 특히 KDE 방법으로 추정한 분포가 더 적합하고, ROC 함수 추정은 KDE 방법과 정규혼합 방법 모두 좋은 방법이나 정규혼합 방법으로 추정한 ROC 함수가 더 좋은 성과를 나타낸다고 결론을 유도할 수 있다.

참고문헌

홍종선, 이원용 (2011). 정규혼합분포를 이용한 ROC 곡선연구, <응용통계연구>, **24**, 269–278.

홍종선, 주재선, 최진수 (2010). 혼합분포에서의 최적분류점, <응용통계연구>, **23**, 13–28.

홍종선, 최진수 (2009). ROC와 CAP 곡선에서의 최적분류점, <응용통계연구>, **22**, 911–921.

Aitkin, M. and Wilson, T. G. (1980). Mixture models, outliers, and the EM algorithm, *Technometrics*, **22**, 325–331.

Egan, J. P. (1975). *Signal Detection Theory and ROC Analysis*, Series in Cognitition and Perception, Academic Press, New York.

Everitt, B. S. (1984). Maximum likelihood estimation of the parameters in a mixture of two univariate normal, *Journal of the Royal Statistical Society*, **33**, 205–215.

Fawcett, T. (2003). ROC graphs: Notes and practical considerations for data mining researchers, *Technical Report HPL-2003-4*, HP Laboratories, 1–28.

Hall, P. G. and Hyndman, R. J. (2003). Improved methods for bandwidth selection when estimating ROC curves, *Statistics and Probability Letters*, **64**, 181–189.

Joseph, M. P. (2005). *A PD Validation Framework for Basel II Internal Ratings-Based Systems*, Credit Scoring and Credit Control *IV*.

Lloyd, C. J. (1998). The use of smoothed ROC curves to summarise and compare diagnostic systems, *Journal of the American Statistical Association*, **93**, 1356–1364.

Lloyd, C. J. and Yong, Z. (1999). Kernel estimators of the ROC curve are better than empirical, *Statistics and Probability Letters*, **44**, 221–228.

McCullagh, P. and Nelder, J. A. (1983). Quasi-likelihood functions, *Annals of Statistics*, **11**, 59–67.

McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*, John Wiley & Sons, New York.

Pepe, M. S. (1998). Three approaches to regression analysis of receiver operating characteristic curves for continuous test results, *Biometrics*, **54**, 124–135.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*, University Press, Oxford, New York.

Provost, F. and Fawcett, T. (1997). Analysis and visualization of classifier performance comparison under imprecise class and cost distributions, In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, 43–48.

Provost, F. and Fawcett, T. (2001). Robust classification for imprecise environments, *Machine Learning*, **42**, 203–231.

Rossenblatt, M. (1956). Remarks on some nonparametric estimates of a density function, *Annals of Mathematical Statistics*, **27**, 832–837.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems, *American Association for the Advancement of Science*, **240**, 1285–1293.

Swets, J. A., Dawes, R. M. and Monahan, J. (2000). Better decisions through science, *Scientific Americal*, **283**, 82–87.

Tasche, D. (2006). Validation of internal rating systems and PD Estimates, On-line bibliography available from: http://arXiv:physics/0606071.

Zou, K. H., Hall, W. J. and Shapiro, D. E. (1997). Smooth non-parametric receiver operating characteristic(ROC) curves for continuous diagnostic tests, *Statistics in Medicine*, **16**, 2143–2156.

ROC Function Estimation

Chong Sun Hong[1] · Mei Hua Lin[2] · Sun Woo Hong[3]

[1]Department of Statistics, Sungkyunkwan University
[2]Research Institute of Applied Statistics, Sungkyunkwan University
[3]Research Institute of Applied Statistics, Sungkyunkwan University

(Received July 2011; accepted October 2011)

Abstract

From the point view of credit evaluation whose population is divided into the default and non-default state, two methods are considered to estimate conditional distribution functions: one is to estimate under the assumption that the data is followed the mixture normal distribution and the other is to use the kernel density estimation. The parameters of normal mixture are estimated using the EM algorithm. For the kernel density estimation, five kinds of well known kernel functions and four kinds of the bandwidths are explored. In addition, the corresponding ROC functions are obtained based on the estimated distribution functions. The goodness-of-fit of the estimated distribution functions are discussed and the performance of the ROC functions are compared. In this work, it is found that the kernel distribution functions shows better fit, and the ROC function obtained under the assumption of normal mixture shows better performance.

Keywords: Bandwidth, density estimation, goodness-of-fit, normal mixture, kernel, performance, ROC function.

[1]Corresponding author: Professor, Department of Statistics, Sungkyunkwan University, 3-53, Myungryun-Dong, Jongro-Gu, Seoul 110-745, Korea. E-mail: cshong@skku.ac.kr