Journal of the Korean Data & Information Science Society
2010, **21**(2), 279–289

한국데이터정보과학회지

호흡곤란 환자의 입퇴원 결정을 위한 간편 통계모형[†]

박철용[1] · 김태윤[2] · 권오진[3] · 박형섭[4]

[123]계명대학교 통계학과
[4]계명대학교 의과대학 내과학교실

접수 2010년 2월 5일, 수정 2010년 3월 17일, 게재확정 2010년 3월 22일

요 약

이 논문에서는 호흡곤란을 주호소로 내원한 668명의 환자를 대상으로 입퇴원 결정을 위한 간편한 통계모형을 제안한다. 이것을 위해 55개 변수 중 임상전문가에 의해 중요하다고 선택된 11개 변수를 설명변수로 이용하였다. 먼저 변환과정으로 11개 연속형 변수 각각에 대해 실제 입원과 퇴원 환자의 커널밀도함수에 의해 퇴원구간을 설정하였다. 다음으로 11개 설명변수 중 퇴원구간에 속한 변수의 개수를 가지고 환자의 퇴원여부를 결정하는 최적 모형을 선택하였다. 입원과 퇴원 환자수의 불균형 때문에 최적 모형의 선택기준으로는 민감도와 특이도의 산술평균과 민감도와 정확률의 조화평균을 이용하였다. 그 결과 11개의 검사결과 중 7개 이상에서 퇴원구간이 나오면 퇴원을 결정하는 것이 최적 모형이 되었다.

주요용어: 민감도, 정확률, 커널밀도함수, 특이도, 퇴원 결정, 호흡곤란.

1. 머리말

호흡곤란 (dyspnea 혹은 shortness of breath)은 환자의 주관적인 증상으로 빈호흡 (tachypnea), 서호흡 (bradypnea), 기좌호흡 (orthopnea), 체인스토크스 (cheyne-stokes) 호흡, 쿠스마울 (Kussmaul) 호흡, 과도호흡 (hyperventilation) 등의 형태로 관찰 가능하고 (Jevon과 Ewens, 2001), 응급실에서 볼 수 있는 가장 흔한 주호소 (chief complaint) 중 하나이다. 응급실에 호흡곤란을 주호소로 내원한 환자는 크게 폐인성 (respiratory) 질환과 심인성 (cardiac) 질환 등으로 구분할 수 있는데, 폐인성 질환은 천식 (asthma), 만성 폐쇄성 폐질환 (chronic obstructive pulmonary disease), 폐렴 (pneumonia), 폐결핵 (tuberculosis) 등이 주요 원인이며 심인성 질환은 좌심실 부전 (left ventricular failure), 폐부종 (pulmonary edema), 울혈성 심부전 (congestive cardiac failure) 등이 주요 원인이다. (Jevon과 Ewens, 2001). 이러한 호흡곤란의 원인질환은 짧은 시간의 문진으로 진단을 하기 어렵기 때문에 임상전문가들은 피검사나 흉부 방사선 검사 등을 이용하여 진단을 하고 있으며 검사된 항목들의 결과로부터 중요특징을 분석하고 감별하는데 많은 시간을 투자하고 있다.

이 논문에서는 호흡곤란을 주호소로 내원한 환자를 대상으로 입원 혹은 퇴원 결정을 위한 간편한 통계모형을 제안하고자 한다. 이것을 위해 임상 데이터베이스에서 얻을 수 있는 55개 변수 중 임상전문가

[†] 본 연구는 지식경제부 지방기술혁신사업 (RTI04-01-01) 지원으로 수행되었음.
[1] 교신저자: (704-701) 대구광역시 달서구 신당동 1000번지, 계명대학교 통계학과, 교수.
E-mail: cypark1@kmu.ac.kr
[2] (704-701) 대구광역시 달서구 신당동 1000번지, 계명대학교 통계학과, 교수.
[3] (704-701) 대구광역시 달서구 신당동 1000번지, 계명대학교 통계학과, 박사과정생.
[4] (704-712) 대구광역시 중구 달성로 216번지, 계명대학교 의과대학 내과학교실, 전임강사.

에 의해 중요하다고 선택된 11개 변수를 설명변수로 이용하였다. 먼저 변환과정으로 11개 연속형 설명변수 각각에 대해 이산화 (discretization)를 시도하였다. 이는 범주형 값이 분류 규칙을 도출하기에 더 용이하다는 장점과 더불어 임상전문가들이 흔히 사용하는 이산화 방법과 비교가 용이하기 때문에 사용되었다. 이산화 방법으로는 반응변수인 입원여부와 관련성을 고려한 지도방법 (supervised method)을 사용하였다. 그런데 11개 설명변수에는 근본적으로 내재된 문제가 있었다. 그래서 구체적으로 이 문제점을 살펴보기 위해서 11개 설명변수의 입원 및 퇴원 환자의 절대도수 분포를 동시에 그린 것이 그림 1.1에 주어져 있다.

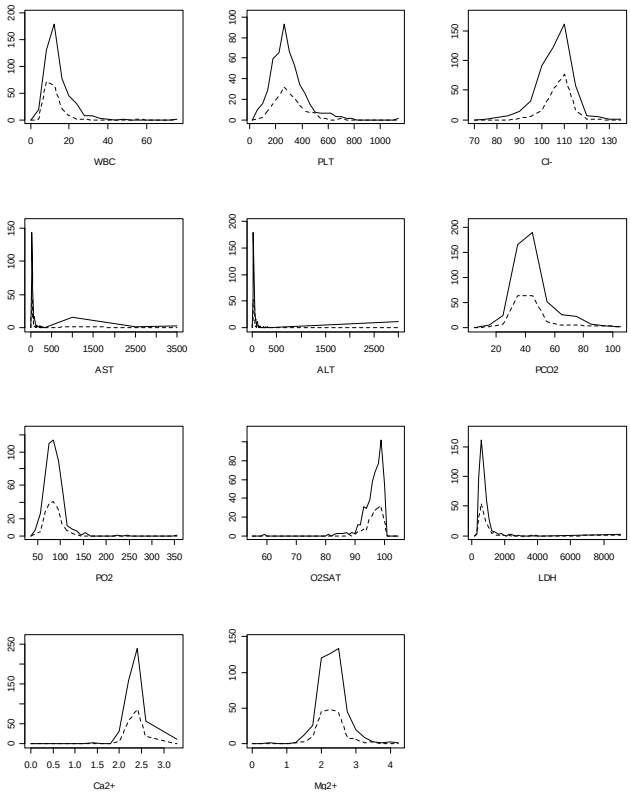

그림 1.1 11개 설명변수의 입원 및 퇴원 환자의 절대도수 분포: 실선-입원 환자, 점선-퇴원환자

11개 설명변수의 구체적인 이름은 3절에서 살펴보기로 하고 여기서는 절대도수 분포에만 초점을 맞추어 살펴보기로 하자. 이 분포들에는 두 가지 두드러진 특징이 있다. 먼저 입원 및 퇴원 환자수의 불균형이 뚜렷하게 나타나고 있으며 다음으로 입원 및 퇴원 분포의 최빈값이 거의 비슷하다는 것이다. 따라서 전체 환자 중에서의 오분류율 (misclassification rate)이 최소가 되는 규칙은 모든 설명변수에서 퇴원구간을 거의 설정하지 않고 대부분 입원구간으로 설정하면 된다 (Johnson과 Wichern, 1992). 또한 두 분포의 최빈값이 비슷하기 때문에 경계값을 정해 상하로 이분하여 이산화하는 것은 결코 좋은 방법이 되지 못한다. 이 두 문제를 동시에 해결하는 방법으로 제안한 것이 입원 및 퇴원에 각각 1/2의 사전확률을 주어 퇴원구간을 설정하는 방식이다. 이 방법은 결국 입원 및 퇴원 환자의 상대도수 분포를 비교하여 퇴원

구간을 설정하게 되며, 이 때 이 퇴원구간은 사전확률이 1/2인 경우에 오분류율을 최소화시키기 때문에 결국 민감도 (sensitivity)와 특이도 (specificity)의 산술평균을 최대화시키게 된다 (구체적인 입증 과정은 2.1 소절에서 설명된다). 구체적으로 이 논문에서 상대도수 분포로 이용한 것은 입원 및 퇴원 환자의 커널밀도함수 (kernel density function)이며, 이것에 의해 퇴원환자의 비율이 높은 구간들로 구성된 퇴원구간 (discharge interval)을 설정하였다.

이렇게 11개 설명변수를 이산화한 다음에 11개 설명변수 중 퇴원구간에 속한 변수의 개수를 가지고 환자의 퇴원여부를 결정하는 최적 모형을 선택하였다. 입원과 퇴원 환자수의 불균형 때문에 최적 모형의 선택기준으로는 전체 환자의 정분류율 (correct classification rate) 대신에 Chawla 등 (2004)과 Weiss (2004)에 의해 해결책으로 소개된 민감도 (sensitivity)와 특이도 (specificity)의 산술평균 및 민감도와 정확률 (precision)의 조화평균을 이용하였다.

이 논문은 다음과 같이 구성되어 있다. 2절에서는 불균형집단 (imbalance class)에 대한 연속형 변수의 이산화 방법과 이 논문에서 사용하고자 하는 최적 모형의 선택기준들에 대해 자세히 설명하였다. 3절에서는 2절의 방법을 실제 호흡곤란 환자에 대해 적용하여 최적 모형을 구하고 이것의 성능을 살펴보았다. 4절의 결론에서는 이 연구의 결과들을 정리하였다.

2. 불균형집단에 대한 이산화 방법과 최적 모형 선택기준들

이 절에서는 먼저 불균형집단 (imbalance class)에 대한 연속형 변수의 이산화 방법을 제시하고, 그 다음에 이 연구에서 사용하고자 하는 최적 모형 선택기준들을 설명하고자 한다. 연속형 설명변수에 대한 이산화를 시도하는 이유는 범주형 값이 분류 규칙을 도출하기에 더 용이하다는 장점과 더불어 임상전문가들이 흔히 사용하는 이산화 방법과 비교가 용이하기 때문이다. 그런데 불균형집단이라는 특성 때문에 그에 합당한 이산화 방법을 제안하게 된다. 불균형집단의 특성 때문에 또한 최적 모형 선택기준으로 정분류율 (correct classification rate)이 아닌 다른 합당한 기준을 제안하게 된다.

2.1. 불균형집단에 대한 연속형 변수의 이산화 방법

1절의 그림 1.1에서 살펴보았듯이 이 연구에서 다루고자 하는 호흡곤란 자료는 입원과 퇴원 환자수의 불균형이 뚜렷한 형태이다. 구체적으로 3절에서 우리가 분석하게 되는 자료에는 입원 환자수가 500명, 퇴원 환자수가 168명으로 약 3:1의 불균형을 이루고 있다. 또한 우리가 사용하게 되는 11개 설명변수의 입원 및 퇴원 환자의 중앙값이 거의 비슷하게 나타나고 있다.

이 논문에서 제안하는 이산화 방법은 다변량분석의 분류규칙에서 흔히 사용되는 우도비 (likelihood ratio)에 근거한 방법이다. 구체적으로 두 모집단 A (admission; 입원), D (discharge; 퇴원)의 확률밀도함수를 각각 $f_A(x), f_D(x)$라고 했을 때 우도비에 근거한 방법은 다음과 같다.

$$f_D(x)/f_A(x) > 1\text{이면 } x\text{는 } D\text{집단으로 분류한다 .} \tag{2.1}$$

따라서 식 (2.1)에 근거하여 $\{x : f_D(x)/f_A(x) > 1\}$를 퇴원구간 (discharge interval)으로 잡을 수 있다. 이 방법은 두 모집단의 사전확률이 $p_A = p_D = 1/2$로 동일할 경우 오분류확률 (misclassification probability) $p_A P(D|A) + p_D P(A|D)$를 최소화하는 규칙이다 (Johnson과 Wichern, 1992). 여기서 일반기호 $P(F|E)$는 E집단 중에서 F로 분류될 확률이다. 따라서 상대적으로 크기가 작은 퇴원집단 D를 목표집단 (target class)으로 잡았을 때

$$P(D|D) = 1 - P(A|D),\ P(A|A) = 1 - P(D|A)$$

는 각각 민감확률 (sensitivity probability)과 특이확률 (specificity probability)이 된다. 그러므로 식 (2.1)에 근거한 규칙은 $p_A = p_D = 1/2$일 때

$$p_A P(D|A) + p_D P(A|D) = 1 - \{P(A|A) + P(D|D)\}/2$$

를 최소화시키며 이는 민감확률과 특이확률의 산술평균을 최대화시키게 된다. 따라서 이 규칙은 $p_A = p_D = 1/2$일 때 민감확률과 특이확률이 비슷하게 되었을 때 오분류확률이 작아지며 정분류확률 $p_A P(A|A) + p_D P(D|D)$이 커지게 된다.식 (2.1)의 규칙을 제안한 이유는 그림 1.1에서 관찰된 불균형집단 때문이다. 두 모집단의 사전확률이 동일하지 않을 경우 일반적으로 오분류확률을 최소화시키는 최적 규칙은 다음과 같다 (Johnson과 Wichern, 1992).

$$[p_D f_D(x)]/[p_A f_A(x)] > 1\text{이면 } x\text{는 } D\text{집단으로 분류한다 .} \tag{2.2}$$

그러나 우리가 분석할 자료는 $p_A \approx 3p_D$이기 때문에 식 (2.2)에 의한 기준은

$$[p_D f_D(x)]/[p_A f_A(x)] \approx f_D(x)/[3f_A(x)]$$

에 기초하게 되어 실상 두 모집단의 절대빈도 분포를 사용하는 것과 동일한 효과를 가지게 된다. 그런데 우리가 분석할 자료에는 거의 모든 구간에서 $3f_A(x)$가 $f_D(x)$보다 커지게 되어 거의 항상 입원을 결정하게 되는 문제점이 발생하게 된다. 이러한 문제점을 해결하기 위해서 정분류확률 $p_A P(A|A) + p_D P(D|D)$이 아닌 민감확률과 특이확률의 산술평균인 $[P(A|A) + P(D|D)]/2$를 최대화시키는 규칙을 사용하게 된 것이다. 2.2 소절에서 다시 살펴보겠지만 이 방법은 불균형집단을 극복하는 하나의 방법으로 많이 채택되고 있다. 식 (2.1)은 모집단에 근거한 방법이기 때문에 표본에 근거한 방법의 제시가 필요하다. 이 연구에서는 확률밀도함수의 추정량으로 흔히 사용되는 커널밀도함수 (kernel density function)

$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left[(x - x_i)/h\right] \tag{2.3}$$

를 이용한다. 여기서 $x_1, x_2, \ldots, x_n$은 표본이며 $K(x)$는 커널함수인데, 이 연구에서는 커널함수로는 표준정규분포의 확률밀도함수를 사용하며, 또한 평활모수 (smoothing parameter)로는 표준정규 커널함수에서 쉽게 사용할 수 있는

$$h = [4/(3n)]^{1/5} S \tag{2.4}$$

를 사용한다. 여기서 S는 표본표준편차이다. 따라서 이 연구에서 이산화 방법으로 제안하는 규칙은 다음과 같다.

$$\widehat{f}_{Dh_1}(x)/\widehat{f}_{Ah_2}(x) > 1\text{이면 } x\text{는 } D\text{집단으로 분류한다 .} \tag{2.5}$$

여기서 $\widehat{f}_{Dh_1}(x)$ $\widehat{f}_{Ah_2}(x)$는 식 (2.3)에 의해 퇴원과 입원 집단에서 각각 구해진 커널밀도함수이며, h_1 h_2는 각각의 집단에서 식 (2.4)에 의해서 구해진 것이다.

기존 연구에서 이산화 방법으로 흔히 사용되는 것으로는, 비지도 방법 (unsupervised method)인 동일간격 (equal width) 방법, 동일비율 (equal frequency) 방법이 있으며, 지도 방법 (supervised method)으로 카이제곱 방법 (Kerber, 1992), 엔트로피 방법 (Fayyad와 Irani, 1993) 및 분포기반 기법 (이상훈 등, 2003) 등이 있다. 또한 여러 이산화 알고리즘의 비교 (Na 등, 2005; Kim 등, 2005)가 있었다. 비지도 방법은 두 모집단을 고려하지 않기 때문에 여기서는 적절한 방법이 아니며 카이제곱 방법

과 엔트로피 방법은 추가 매개변수를 필요로 하는 약점이 있다. 마지막으로 분포기반 기법은 이 연구와 기본적으로 비슷한 아이디어를 가지고 출발하였지만 커널밀도함수와 같은 정밀한 확률밀도함수 추정량을 이용하지 않고 있다. 그리고 두 확률밀도함수 추정량의 최빈값이 차이가 크면 이 연구와 다른 이산화 결과를 얻게 된다. 따라서 커널밀도함수를 사용하여 이상훈 등 (2003)의 분포기반 기법을 보다 정밀하게 확장할 수도 있으리라 생각된다.

2.2. 불균형집단에 대한 최적 모형 선택기준들

이 연구에서는 11개 설명변수에 대해 식 (2.5)의 규칙에 의해 11개의 퇴원구간을 구하고, 이 중 퇴원구간에 들어가는 변수의 수에 기반하여 입원 혹은 퇴원을 결정하는 최적의 모형을 찾고자 한다. 구체적으로 퇴원구간에 들어가는 설명변수의 수가 $3, 4, \cdots, 11$ 이상이 되는 모형 중에서 최적의 모형을 찾고자 한다. 그런데 2.1 소절에서도 언급되었듯이 불균형집단이기 때문에 단순히 정분류율 (correct classification rate)에만 근거하여 최적의 모형을 구하게 되면 모든 설명변수의 거의 모든 영역에서 입원으로 분류하는 것이 최적의 모형이 되는 불합리한 결과를 얻게 되기 때문에, 적절한 최적 모형 선택기준들을 사용하고자 한다.

불균형집단에 대한 모형평가에 대한 방법론은 Chawla 등 (2004)과 Weiss (2004)에 자세히 소개되어 있다. 그 방법론을 크게 나누어 보면 적절한 모형평가 측도 방법, 표본추출 (sampling) 방법 및 비용민감 (cost sensitive) 방법 등이 있다. 여기서는 상대적으로 쉽게 사용할 수 있는 적절한 모형평가 측도 방법을 사용하도록 한다.

불균형집단에 대해 모형평가 측도를 소개하기 전에 오분류 행렬 (confusion matrix)에서 흔히 사용하는 모형평가 측도들을 먼저 소개하도록 하자. 오분류 행렬이 다음과 같이 주어졌다고 하자.

실제	분류		합
	D	A	
D	TD	FA	n_D
A	FD	TA	n_A

이 오분류 행렬에서 흔히 사용하는 모형평가 측도는 다음과 같다.

$$\text{민감도 (sensitivity) } r = TD/n_D$$
$$\text{특이도 (specificity) } s = TA/n_A$$
$$\text{정확률 (precision) } p = TD/(TD + FD)$$

정보검색 (information retrieval) 분야에서는 민감도를 재현율 (recall)이라고 부르며 정확률과 함께 많이 사용되고 있다. 여기서 민감도와 특이도는 각각 민감확률 $P(D|D)$과 특이확률 $P(A|A)$의 추정량이라 할 수 있다. 여러 가지 모형평가 측도를 하나의 숫자로서 요약하는 것으로 가장 많이 사용되는 것이 정분류율 (correct classification rate)인데 이것은 민감도와 특이도의 표본크기를 이용한 가중평균

$$(n_D r + n_A s)/(n_D + n_A) = (TD + TA)/(n_D + n_A)$$

이다. 그런데 이것은 정분류확률 $p_A P(A|A) + p_D P(D|D)$의 추정량이기 때문에 이 연구에서 분석하는 호흡곤란 자료와 같은 뚜렷한 불균형집단에서는 모형평가 기준으로 사용하기에는 적절하지 못하다.

Weiss (2004)에서 불균형집단에서 적절한 모형평가 측도로서 소개된 것 중 대표적인 것이 AUC (Area Under Curve)이며 또한 정보검색 분야에서 흔히 사용되는 것으로 민감도와 정확률의 조화평균

도 함께 소개되고 있다. 이 조화평균은 van Rijsbergen (1979)에서 F-측도 (F-measure)라는 이름으로 제시되었다. 또한 AUC는 ROC (Receiver Operating Characteristic) 곡선 아래의 면적을 나타내는 것으로 이 연구에서 사용하고 있는 한 단계 이지분리 (binary splitting)에서는 민감도와 특이도의 산술평균이 된다.한 단계 이지분리 (binary splitting)에서 AUC가 민감도와 특이도의 산술평균이 된다는 것은 한 단계 이지분리의 대표적 ROC 곡선이 그림 2.1과 같이 된다는 것에서 쉽게 알 수 있다.

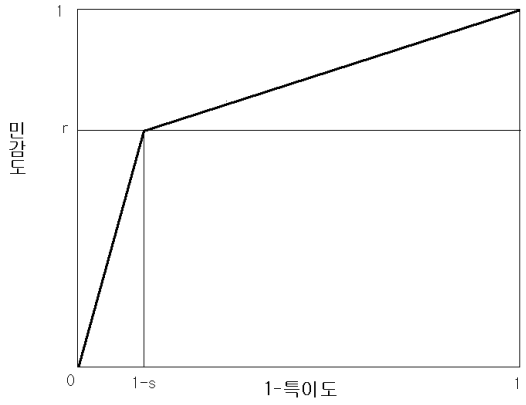

그림 2.1 한 단계 이지분리의 대표적 ROC 곡선

그러므로 이 ROC 곡선의 아래 면적인 AUC는

$$\begin{aligned}&(\text{왼쪽 아래 삼각형 면적}) + (\text{오른쪽 아래 사각형 면적}) + (\text{오른쪽 위 삼각형 면적})\\ &= \frac{1}{2}(1-s)r + sr + \frac{1}{2}s(1-r) = \frac{1}{2}(r+s)\end{aligned}$$

가 된다. 다시 말해 한 단계 이지분리에서는 민감도와 특이도의 산술평균 $AM \equiv (r+s)/2$이 가장 대표적인 모형평가 기준이 되는 것이다. 식 (2.1)의 규칙이 민감확률과 특이확률의 산술평균을 최대화시키기 때문에, 식 (2.5)의 규칙이 민감도와 특이도의 산술평균을 근사적으로 최대화시켜 불균형집단에서 적절한 분류규칙이 되는 것을 알 수 있다.

또한 van Rijsbergen (1979)에서 F-측도로 제시되었고 정보검색 분야에서 모형평가 기준으로 흔히 사용되고 있는 민감도와 정확률의 조화평균 $HM \equiv 2/(1/r + 1/p)$도 사용하도록 한다.

3. 호흡곤란 자료에의 적용

이 연구에서 분석하고자 하는 자료는 A 의료원에 2006년 7월부터 2007년 6월 사이에 호흡곤란을 주호소로 내원한 환자 1129명의 의무기록에서 추출되었다. 구체적으로 대상자의 인적사항을 제외한 등록번호, 성별, 나이, 응급실 내원일자 및 시간, 진료결과, 입원시 진단, 초기 검사항목 등의 자료들을 데이터웨어하우스에서 추출하였다. 이렇게 추출된 자료 중 타병원으로 전원된 환자, 도착 직후 사망 (Death On Arrival; DOA), 심폐소생술 (Cardio-Pulmonary Resuscitation; CPR) 후 혹은 심폐소생술 하지 않기 (Do Not Resuscitate; DNR)로 사망한 환자, 자의 퇴원 혹은 미상의 기타 환자, 의무기록이 불완전한 경우를 제외한 668명의 환자를 이용하였다. 이 중 500명이 입원 환자였으며 나머지 168명이 퇴원 환

자였다. 또한 원래 데이터웨어하우스에서 추출된 55개의 변수 중 임상전문가에 의해 중요하다고 판단된 11개의 변수를 분석에 사용하였다. 전문가에 의해 선택되어 이 연구에서 사용되는 11개의 설명변수는 다음의 표 3.1에 자세히 설명되어 있다.

표 3.1 11개 설명변수의 약자와 설명

영문약자	영어 설명	한글 설명	단위
WBC	White Blood Cell [count]	백혈구 [수]	$\times 10^3/uL$
PLT	Platelet count	혈소판 수	$\times 10^3/uL$
Cl-	Chloride	염소농도	$mmol/L$
AST	Aspartate Transaminase	아스파르테이트아미노전이효소	U/L
ALT	Alanine Transaminase	알라닌 아미노전이효소	U/L
$PCO2$	Pressure of Carbon dioxide	이산화탄소 압	$mmHg$
$PO2$	Pressure of Oxygen	산소 압	$mmHg$
$O2SAT$	Oxygen Saturation	산소포화도	%
LDH	Lactate Dehydrogenase	젖산 탈수소효소	U/L
$Ca2+$	Calcium	칼슘	mEg/L
$Mg2+$	Magnesium	마그네슘	mEg/L

이 11개 설명변수에 대해 입원 및 퇴원 환자의 절대도수 분포를 그린 것이 1절에서도 살펴보았던 그림 1.1이다. 이 그림에서 모든 변수의 거의 모든 구간에서 퇴원 환자의 절대도수가 입원 환자의 절대도수보다 큰 불균형집단의 문제가 두드러지게 나타났기 때문에 절대도수에 의해 이산화를 사용하는 것이 적절하지 않게 되었다.

따라서 2.1 소절에서 제안된 이산화 방법인 분류규칙 (2.5)를 사용하였다. 먼저 각 변수의 입원 및 퇴원 환자의 커널밀도함수 (kernel desntiy function)의 평활모수 (smoothing parameter) 값으로 선택된 것을 정리한 것이 다음 표에 주어져 있다.

변수	입원 환자	퇴원 환자
WBC	14.42	0.94
PLT	38.55	38.17
Cl-	2.27	1.91
AST	22.00	3.00
ALT	50.31	9.74
$PCO2$	4.05	4.00
$PO2$	8.00	6.00
$O2SAT$	1.06	0.88
LDH	160.89	174.84
$Ca2+$	0.054	0.058
$Mg2+$	0.119	0.125

이 평활모수 값은 식 (2.4)에 기초하되 분포 모양의 왜곡이 심할 경우 약간 조정된 것이다. 이 평활모수를 사용하여 (2.3)의 커널밀도함수를 그린 것이 그림 3.1에 주어져 있다.

이 커널밀도함수를 살펴보면 두 가지 두드러진 특징이 있다. 먼저 절대도수 분포에서도 관찰할 수 있었듯이 두 분포의 최빈값이 거의 일치한다는 것이다. 또한 LDH 변수를 제외하고는 퇴원 환자의 분포가 입원 환자의 분포보다 최빈값에 밀집되어 있다는 것을 알 수 있다. 이 커널밀도함수에 근거하여 (2.5)의 분류규칙에 의한 퇴원구간 (discharge interval) $\left\{x : \widehat{f}_{Dh_1}(x)/\widehat{f}_{Ah_2}(x) > 1\right\}$을 계산하였더니 표 3.2와 같이 되었다.

이 표에는 A 의료원에서 임상전문가들이 정상이라고 간주하는 구간인 정상간주구간을 동시에 나타내고 있다. LDH 변수를 제외하고는 퇴원구간과 정상간주구간이 많이 겹쳐 있음을 알 수 있다. 그런데

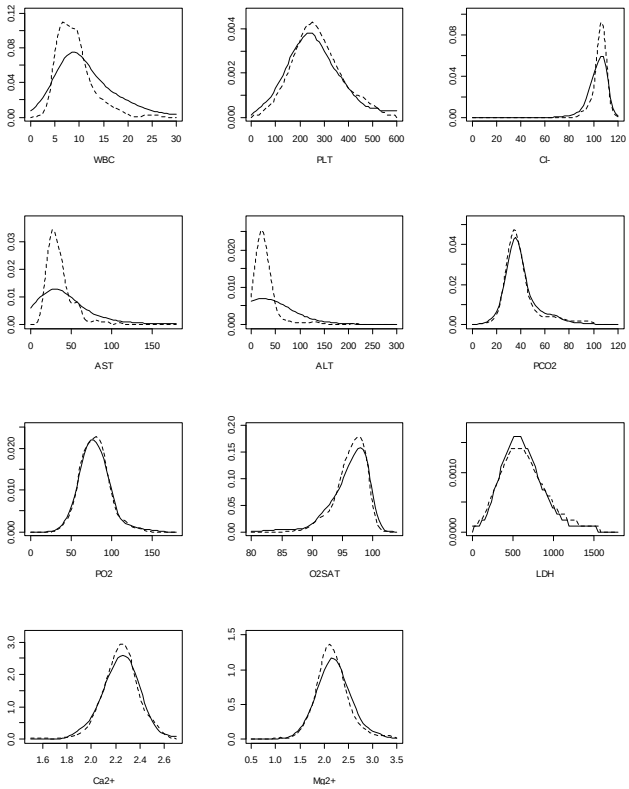

그림 3.1 11개 설명변수의 입원 및 퇴원 환자의 커널밀도함수:실선 - 입원 환자, 점선 - 퇴원 환자

표 3.2 11개 설명변수의 퇴원구간과 정상간주구간

변수	퇴원구간	정상간주구간	변수	퇴원구간	정상간주구간
WBC	4.5 10.5	5.2 12.4	*PO2*	57 96	83 108
PLT	200 520	130 400	*O2SAT*	94.4 98.8	95 99
Cl-	104 110	95 108	*LDH*	0 300, 810 이상	211 423
AST	12 48	13 36	*Ca2+*	2.1 2.32	1.2 3.2
ALT	0 45	5 44	*Mg2+*	1.8 2.3	1.5 2.7
PCO2	26 40	35 48(남) 32 45(여)			

LDH 변수의 경우 퇴원구간이 오히려 최빈값의 외곽에 위치하고 있어, 최빈값 근처에 위치하고 있는 정상간주구간과 많은 차이를 보이고 있다. 이 부분에 대해서는 좀 더 규명이 필요하리라 생각된다.

이 연구에서는 11개 설명변수 중 퇴원구간에 들어가는 변수의 숫자에 근거하여 최적의 모형을 선택하고자 한다. 구체적으로 11개 설명변수 중 퇴원구간에 들어가는 변수의 숫자가 3개 이상인 모형부터 11개 이상인 모형까지 고려하여 2.2 소절에서 설명되었던 $AM = (r+s)/2$ 및 $HM = 2/(1/r + 1/p)$을 이용하여 최적 모형을 선정하고자 한다. 이 8개 모형에 대한 여러 모형평가 측도들을 정리한 것이 표

3.3이다.

표 3.3 퇴원구간에 속하는 설명변수의 수에 따른 모형평가 측도들

설명변수 수	민감도(r)	특이도(s)	정확율(p)	정분류율	AM	HM	AM+HM
11개 이상	0.012	0.996	0.500	0.749	0.504	0.023	0.527
10개 이상	0.083	0.968	0.333	0.746	0.526	0.133	0.658
9개 이상	0.315	0.882	0.491	0.739	0.599	0.384	0.982
8개 이상	0.518	0.702	0.400	0.655	0.610	0.451	1.061
7개 이상	0.750	0.472	0.344	0.542	0.611	0.472	1.083
6개 이상	0.887	0.296	0.314	0.445	0.592	0.464	1.055
5개 이상	0.970	0.154	0.290	0.359	0.562	0.447	1.009
4개 이상	1.000	0.068	0.270	0.301	0.534	0.425	0.959
3개 이상	1.000	0.018	0.258	0.264	0.509	0.410	0.919

이 결과에 의하면 먼저 정분류율은 설명변수의 수가 커지면 무조건 커지는 경향이 있어 모형평가 기준으로 적합하지 않다는 것을 알 수 있다. 이는 소수의 퇴원 환자에게는 거의 맞지 않더라고 다수의 입원 환자에 거의 다 맞추면 정분류율이 높아지기 때문에 당연한 현상이다. 그런데 AM 및 HM 기준에 의하면 7개 이상의 설명변수가 퇴원구간에 속하면 퇴원을 결정하는 것이 최적의 모형으로 선택되었다. 약간의 차이를 두고 8개 이상을 기준으로 퇴원을 결정하는 모형이 차선의 모형으로 선택되었다.

참고로 임상전문가에 의해 정상으로 간주되는 정상간주구간을 퇴원구간 대신에 사용하여 분석하였을 때의 모형평가 측도들을 계산하여 보았더니 다음의 표 3.4와 같이 되었다.

표 3.4 정상간주구간에 속하는 설명변수의 수에 따른 모형평가 측도들

설명변수 수	민감도(r)	특이도(s)	정확율(p)	정분류율	AM	HM	AM+HM
11개 이상	0.006	0.996	0.250	0.747	0.501	0.012	0.513
10개 이상	0.036	0.962	0.375	0.729	0.499	0.066	0.565
9개 이상	0.250	0.828	0.333	0.683	0.539	0.286	0.825
8개 이상	0.565	0.610	0.326	0.599	0.588	0.413	1.001
7개 이상	0.792	0.360	0.296	0.468	0.576	0.431	1.007
6개 이상	0.935	0.170	0.272	0.362	0.553	0.421	0.974
5개 이상	0.976	0.062	0.260	0.292	0.519	0.411	0.930
4개 이상	0.994	0.016	0.254	0.262	0.505	0.405	0.910
3개 이상	0.994	0.004	0.251	0.253	0.499	0.401	0.900

이 표에 의하면 역시 7개 이상이 최적 모형, 8개 이상이 거의 비슷한 성능을 가지는 차선의 모형으로 선택되지만, 퇴원구간에 기반한 최적 모형보다 성능이 다소 떨어지고 있음을 알 수 있다.

4. 결론

이 논문에서는 호흡곤란을 주호소로 내원한 환자를 대상으로 입원 혹은 퇴원 결정을 위한 간편한 통계모형을 제안하였다. 이것을 위해 임상 데이터베이스에서 얻을 수 있는 55개 변수 중 임상전문가에 의해 중요하다고 선택된 11개 변수를 설명변수로 이용하였다. 먼저 변환과정으로 11개 연속형 설명변수 각각에 대해 이산화를 시도하였다. 이산화 방법으로는 반응변수인 입원여부와 관련성을 고려한 지도방법을 사용하였다. 그런데 11개 설명변수에는 근본적으로 내재된 문제가 있었다. 먼저 입원 및 퇴원 환자수의 불균형이 뚜렷하게 나타나고 있다. 따라서 전체 환자 중에서의 오분류율 (misclassification rate)이 최소가 되는 규칙은 모든 설명변수에서 퇴원구간을 거의 설정하지 않고 대부분 입원구간으로 설정하게 되는 문제가 발생하는 것이다. 다음으로 입원 및 퇴원 분포의 최빈값이 거의 비슷하다는 문제인

데 이 경우 경계값을 정해 상하로 이분하여 이산화하는 것은 결코 좋은 방법이 되지 못하게 된다. 이 두 문제를 동시에 해결하는 방법으로 제안한 것이 입원 및 퇴원에 각각 1/2의 사전확률을 주어 퇴원구간을 설정하는 방식이다. 이 방법은 결국 입원 및 퇴원 환자의 상대도수 분포를 비교하여 퇴원구간을 설정하게 된다. 구체적으로 이 논문에서 상대도수 분포로 이용한 것은 입원 및 퇴원 환자의 커널밀도함수이며, 이것에 의해 퇴원환자의 비율이 높은 구간들로 구성된 퇴원구간을 설정하였다.

이렇게 11개 설명변수를 이산화한 다음에 11개 설명변수 중 퇴원구간에 속한 변수의 개수를 가지고 환자의 퇴원여부를 결정하는 최적 모형을 선택하였다. 입원과 퇴원 환자수의 불균형 때문에 최적 모형의 선택기준으로는 전체 환자의 정분류율 대신에 Chawla 등 (2004)과 Weiss (2004)에 의해 해결책으로 소개된 민감도와 특이도의 산술평균 및 민감도와 정확률의 조화평균을 이용하였다. 그 결과 퇴원구간에 속하는 설명변수의 수가 7개 이상이면 퇴원을 결정하는 모형이 최적의 모형으로 선택되었으며, 약간의 차이를 두고 8개 이상을 기준으로 퇴원을 결정하는 모형이 차선의 모형으로 선택되었다.

참고문헌

이상훈, 박정은, 오경환 (2003). 데이터 분포를 고려한 연속 값 속성의 이산화. <한국퍼지 및 지능시스템 학회 논문지>, **13**, 291-396.

Chawla, N. V., Japkowwicz, N. and Nolcz, A. (2004). Editorial: Special issuse on learning from imbalanced data sets. *SIGKDD Explorations*, **6**, 1-6.

Fayyad, U. M. and Irani, K. B. (1993). Multi-interval discretization of continuous attributes as preprocessing for classification learning. *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, 1022-1027.

Jevon, P. and Ewens, B. (2001). Assessment of a breathless patient. *Nursing Standard*, **15**, 48-53.

Johnson, R. A. and Wichern, D. W. (1992). *Applied multivariate statistical analysis*, 3rd Ed., Prentice Hall, New Jersey.

Kerber, R. (1992). ChiMerge: Discretization of numeric attribute. *Proceedings of the 10th National Conference on Artificial Intelligence (AAAI-92)*, 123-127.

Kim, J. S., Jang, Y. M. and Na, J. H. (2005) Comparison of multiway discretization algorithms for data mining. *Journal of the Korean Data & Information Science Society,* **16**, 801-813.

Na, J. H., Kim, J. M. and Cho, W. S. (2005). Comparison of binary discretization algorithms for data mining. *Journal of the Korean Data & Information Science Society*, **16**, 769-780.

van Rijsbergen, C. J. (1979). *Information retrieval*, Butterworths, London.

Weiss, G. M. (2004). Mining with rarity: A unifying framework. *SIGKDD Explorations*, **6**, 7-19.

Journal of the Korean Data & Information Science Society
2010, **21**(2), 279–289

한국데이터정보과학회지

A simple statistical model for determining the admission or discharge of dyspnea patients[†]

Cheolyong Park[1], Tae Yoon Kim[2], O Jin Kwon[3], Hyoung-Seob Park[4]

[123]Department of Statistics, Keimyung University
[4]Department of Internal Medicine, School of Medicine, Keimyung University

Received 5 February 2010, revised 17 March 2010, accepted 22 March 2010

Abstract

In this study, we propose a simple statistical model for determining the admission or discharge of 668 patients with a chief complaint of dyspnea. For this, we use 11 explanatory variables which are chosen to be important by clinical experts among 55 variables. As a modification process, we determine the discharge interval of each variable by the kernel density functions of the admitted and discharged patients. We then choose the optimal model for determining the discharge of patients based on the number of explanatory variables belonging to the corresponding discharge intervals. Since the numbers of the admitted and discharged patients are not balanced, we use, as the criteria for selecting the optimal model, the arithmetic mean of sensitivity and specificity and the harmonic mean of sensitivity and precision. The selected optimal model predicts the discharge if 7 or more explanatory variables belong to the corresponding discharge intervals.

Keywords: Admission or discharge, dyspnea patients, kernel density function, precision, sensitivity, specificity.

[†] This work was supported by the grant No. RTI04-01-01 from the Regional Technology Innovation Program of the Ministry of Knowledge Economy (MKE).
[1] Corresponding author: Professor, Department of Statistics, Keimyung University, Daegu 704-701, Korea. E-mail: cypark1@kmu.ac.kr
[2] Professor, Department of Statistics, Keimyung University, Daegu 704-701, Korea.
[3] Ph. D. student, Department of Statistics, Keimyung University, Daegu 704-701, Korea.
[4] Full-time lecturer, Department of Internal Medicine, School of Medicine, Keimyung University, Daegu 704-712, Korea.