Journal of the Korean Data & Information Science Society
2011, **22**(3), 477–485

한국데이터정보과학회지

호흡곤란 환자에 대한 혈액검사 결과들의 수량화 연구[†]

박철용[1]

[1]계명대학교 통계학과

접수 2011년 4월 12일, 수정 2011년 5월 11일, 게재확정 2011년 5월 16일

요 약

Park 등 (2010)은 호흡곤란을 주호소로 내원한 668명의 환자를 대상으로 11개 혈액검사 결과 중 퇴원구간에 속한 결과의 개수를 가지고 입퇴원 결정을 위한 간편한 통계모형을 제안하였다. 그런데 11개 혈액검사의 결과에 대한 중요성을 고려하지 않아 모형의 성능이 떨어질 수 있다는 문제점이 있었다. 이 연구에서는 수량화 방법에 의해 11개 혈액검사 결과의 중요성을 평가해보고, 이 중요성을 고려한 통계모형을 도출하였다. 그 결과 중요성을 고려한 새로운 모형이 중요성을 고려하지 않은 기존 모형보다 다소 성능이 향상된 것을 발견할 수 있었다.

주요용어: 불균형 집단, 수량화 방법, 커널밀도함수, 퇴원 결정, 호흡곤란.

1. 머리말

호흡곤란 (dyspnea)은 환자의 주관적인 증상으로 응급실에서 볼 수 있는 가장 흔한 주호소 (chief complaint) 중 하나이다. Jevon과 Ewens (2001)에 의하면 응급실에 호흡곤란을 주호소로 내원한 환자는 크게 폐인성 (respiratory) 질환과 심인성 (cardiac) 질환 등으로 구분할 수 있는데, 폐인성 질환은 천식, 만성 폐쇄성 폐질환, 폐렴, 폐결핵 등이 주요 원인이며 심인성 질환은 좌심실 부전, 폐부종, 울혈성 심부전 등이 주요 원인이다. 이러한 호흡곤란의 원인질환은 짧은 시간의 문진으로 진단을 하기 어려우며 임상전문가들은 혈액검사나 흉부 방사선 검사 등을 이용하여 진단을 하고 있다.

Park 등 (2010)에서는 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 혈액검사 결과를 설명변수로 이용하였다. 먼저 11개 혈액검사 각각에 대해 이산화 (discretization)를 시도하였다. 구체적으로 입원 및 퇴원 환자수의 불균형이 뚜렷하게 나타나고 있기 때문에 절대도수가 아닌 상대도수에 근거하여 퇴원환자의 비율이 높은 구간들로 퇴원 구간 (discharge interval)을 설정하였다. 그런 다음에 11개 혈액검사 결과 중 퇴원구간에 속한 개수를 가지고 환자의 퇴원여부를 결정하는 최적 모형을 선택하였다.

Park 등 (2010)의 연구는 불균형 집단 (imbalance class)에 대해 절대도수에 의한 퇴원구간 설정이 정분류율 (correct classification rate)만 크게 만들기 때문에 상대적으로 빈도가 작은 퇴원환자에 대한 민감도 (sensitivity)가 작아지는 단점을 극복하기 위한 하나의 노력이다. 그런데 간편한 모형을 추구하는 과정에서 퇴원구간에 속하는 혈액검사 결과들에 대해 동일한 가중값을 주기 때문에 그 성능이 떨어질 수 있다는 문제점이 있었다.

[†] 본 연구는 지식경제부 지방기술혁신사업 (RTI04-01-01) 지원으로 수행되었음.
[1] (704-701) 대구광역시 달서구 신당동 1000번지, 계명대학교 통계학과, 교수. E-mail: cypark1@kmu.ac.kr

이 연구는 Park 등 (2010)에 의해 설정된 (혈액검사 결과들의) 퇴원구간이 입퇴원 판단에 미치는 영향력에는 차이가 있을 것이라는 점에 착안하여 시작되었다. 구체적으로 Park 등 (2010)과 동일한 퇴원구간을 사용하되 수량화 방법 (quantification method)에 의해 각 퇴원구간의 중요성을 평가하고자 한다. 수량화 방법에 의해 각 퇴원구간의 가중값을 구한 다음, 이 가중값을 고려한 입퇴원 결정을 위한 새로운 통계모형을 제공하고자 한다. 또한 새로운 모형과 기존 모형의 성능을 여러 가지 모형평가 기준에서 비교하고자 한다.

이 논문은 다음과 같이 구성되어 있다. 2절에서는 Park 등 (2010)에 소개된 불균형 집단 (imbalance class)에 대한 연속형 변수의 이산화 방법과 최적 모형 선택기준들에 대해 간략히 설명하였다. 3절에서는 2절의 이산화 방법을 실제 호흡곤란 환자에 대해 적용하여 얻은 퇴원구간의 중요성을 수량화 방법으로 평가하는데 중점을 둔다. 이 중요성을 고려한 새로운 입퇴원 결정모형을 제공하게 되며, 이 새로운 모형의 성능을 기존의 중요성을 고려하지 않은 모형과 비교하였다. 4절의 결론 및 논의에서는 이 연구의 결과들을 요약하고 논의하였다.

2. 불균형 집단에 대한 이산화 방법과 모형 선택기준들

이 절에서는 Park 등 (2010)에서 제시된 불균형 집단 (imbalance class)에 대한 연속형 변수의 이산화 방법과 최적 모형 선택기준들을 간략하게 설명하고자 한다.

먼저 불균형 집단에 대한 연속형 변수의 이산화 방법을 설명한다. Park 등 (2010)에서 제안하는 이산화 방법은 다변량분석 (multivariate analysis)의 분류 (classification)에서 흔히 사용되는 우도비 (likelihood ratio)에 근거한 방법으로, 두 모집단 A (admission; 입원), D (discharge; 퇴원)의 확률밀도함수를 각각 $f_A(x), f_D(x)$라고 했을 때 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 퇴원구간 (discharge interval)을 $\{x : f_D(x)/f_A(x) > 1\}$로 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 규칙으로 알려져 있다 (Johnson과 Wichern, 1992). 여기서 $P(F|E)$는 E집단 중에서 F로 분류될 조건부확률을 나타내는 부호이다. 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 최소화시키는 퇴원구간이 된다. 이 절의 나중에 소개되는 모형평가 기준 설명에서 이 방법이 불균형 집단을 극복하는 하나의 방법으로 많이 채택되고 있음을 설명한다.

식 (2.1)은 모집단에 근거한 방법이기 때문에 표본에 근거한 방법의 제시가 필요하다. 확률밀도함수의 추정량으로는 커널밀도함수 (kernel density function)

$$\widehat{f_h}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left[(x - x_i)/h\right] \tag{2.2}$$

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

$$h = [4/(3n)]^{1/5} S \tag{2.3}$$

를 사용한다. 여기서 S는 표본 표준편차이다. 따라서 Park 등 (2010)에서 이산화 방법으로 제안하는 규칙은 다음과 같이 나타낼 수 있다.

$$\widehat{f}_{Dh_1}(x)/\widehat{f}_{Ah_2}(x) > 1\text{이면 } x\text{는 } D\text{집단으로 분류한다 .} \tag{2.4}$$

여기서 $\widehat{f}_{Dh_1}(x)$ $\widehat{f}_{Ah_2}(x)$는 식 (2.2)에 의해 퇴원과 입원 집단에서 각각 구해진 커널밀도함수이며, h_1 h_2는 각각의 집단에서 식 (2.3)에 의해서 구해진 것이다.

기존의 이산화 방법으로는 동일간격 방법, 동일비율 방법, 카이제곱 방법 (Kerber, 1992), 엔트로피 방법 (Fayyad와 Irani, 1993) 및 분포기반 기법 (이상훈 등, 2003) 이 있으며, 여러 이산화 알고리즘을 비교한 Na 등 (2005)과 Kim 등 (2005)의 연구가 있었다. 그러나 기존의 이산화 방법들에는 약간의 문제점이 존재하고 있었다. 동일간격과 동일비율 방법은 두 모집단을 고려하지 않기 때문에 우리 방법과 비교 대상이 되지 않으며, 카이제곱 방법과 엔트로피 방법은 추가 매개변수를 필요로 하는 약점이 있다. 또한 분포기반 기법은 이 연구와 기본적으로 비슷한 아이디어를 가지고 출발하였지만, 커널밀도함수와 같은 정밀한 확률밀도함수 추정량을 이용하지 않고 있으며 두 확률밀도함수 추정량의 최빈값이 차이가 크면 이 연구와 다른 이산화 결과를 얻게 된다.

다음으로 최적의 모형 선택을 위한 모형평가 방법을 간략히 소개하도록 하겠다. 불균형 집단에 대한 모형평가 방법론은 Chawla 등 (2004)과 Weiss (2004)에 자세히 소개되어 있는데 그 방법론들 중 상대적으로 쉽게 사용할 수 있는 적절한 모형평가 측도 방법을 사용한다.

그 방법론의 설명을 위해 다음의 오분류 행렬 (confusion matrix)을 고려해 보도록 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.

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

여기서 민감도와 특이도는 각각 민감확률 $P(D|D)$과 특이확률 $P(A|A)$의 추정량이라 할 수 있다. 또한 정보검색 (information retrieval) 분야에서는 민감도를 재현율 (recall)이라고도 부르며 정확률과 함께 많이 사용되고 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 불균형 집단에서는 모형평가 기준으로 사용하기에는 적절하지 못하다.

Weiss (2004)가 불균형 집단에서 적절한 모형평가 측도로서 소개된 것 중 대표적인 것이 AUC (Area Under Curve)이며, 정보검색 분야에서 흔히 사용되는 것으로 민감도와 정확률의 조화평균도 함께 소개되고 있다. 그런데 퇴원구간 설정에 사용되고 있는 이지분리 (binary split)에서는 AUC가 민감도와 특이도의 산술평균이 된다 (Park 등, 2010). 따라서 이 연구에서는 불균형 집단의 모형평가 측도로서 민감도와 특이도의 산술평균 $AM \equiv (r+s)/2$과 민감도와 정확률의 조화평균 $HM \equiv 2/(1/r + 1/p)$을 사용한다.

3. 호흡곤란 자료에의 적용

이 연구에서 사용된 자료는 A 의료원에 2006년 7월부터 2007년 6월 사이에 호흡곤란을 주호소로 내원한 환자 1129명의 의무기록에서 추출되었다. 이렇게 추출된 자료 중 타병원으로 전원된 환자, 도착 직후 사망, 심폐소생술 후 혹은 심폐소생술 금지 (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 환자였다. 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 설명되어 있다.

표 3.1 분석에 사용되는 11개 설명변수의 약자와 간략 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

입원 환자수 500명이 퇴원 환자수 168명 보다 3배 정도 많기 때문에 절대도수가 아닌 상대도수에 근거한 이산화 방법인 분류규칙 (2.4)를 적용한 결과 표 3.2와 같은 퇴원구간을 얻을 수 있었다.

3.1. 수량화 방법에 의한 중요성 평가

이 소절에서는 표 3.2에서 구한 퇴원구간들이 입퇴원 결정에 미치는 중요성을 평가한다. 이것을 위하여 반응변수가 질적변수이며 설명변수도 모두 질적변수인 경우 적용되는 수량화 방법 II를 사용한다. 수량화 방법 II는 정준상관분석 (canonical correlation analysis) 혹은 정준판별분석 (canonical discriminant analysis)을 통해 질적 반응변수 집단의 선형결합과 가장 높은 상관을 보이는 질적 설명변수들 집단의 선형결합을 찾아내는 과정이다.

$discharge$를 입원이면 0, 퇴원이면 1의 값을 취하는 반응변수로 놓고, WBC, PLT, $Cl-$, AST, ALT, $PCO2$, $PO2$, $O2SAT$, LDH, $Ca2$ 및 $Mg2$를 각각 해당 변수가 퇴원구간이면 1, 그 외에는 0의

표 3.2 11개 설명변수의 퇴원구간

변수	퇴원구간
WBC	4.5 ∼ 10.5
PLT	200 ∼ 520
Cl-	104 ∼ 110
AST	12 ∼ 48
ALT	0 ∼ 45
$PCO2$	26 ∼ 40
$PO2$	57 ∼ 96
$O2SAT$	94.4 ∼ 98.8
LDH	0 ∼ 300, 800 ∼
$Ca2+$	2.1 ∼ 2.32
$Mg2+$	1.8 ∼ 2.3

값을 취하는 퇴원구간 여부를 나타내는 변수로 놓기로 하자. 그리고 정준판별분석에 의해 $discharge$와 가장 상관계수가 높은 질적 설명변수들의 최적 선형결합 can을 구하였더니 다음과 같았다.

$$can = 1.011\ WBC + 0.555\ PLT + 1.038\ Cl_{+}0.248\ AST + 0.602\ ALT + 0.276\ PCO2$$
$$+ 0.124\ PO2 + 0.344\ O2SAT + 0.466\ LDH + 0.491\ Ca2 + +0.366\ Mg2$$

이렇게 구한 최적 선형결합 can이 입원과 퇴원 환자들을 잘 구분하는지 알아보기 위해서 두 집단의 상자그림 (box plot)을 그렸더니 그림 3.1과 같았다.

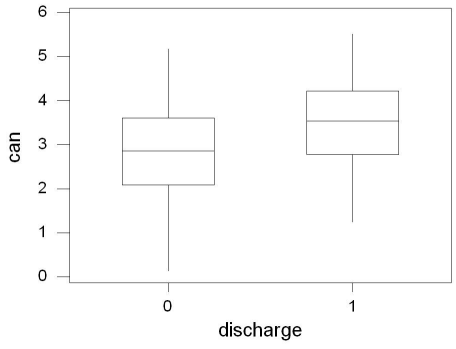

그림 3.1 입원 및 퇴원 환자의 최적 선형결합 can의 상자그림

이 상자그림을 통해 퇴원 환자의 can 값이 입원 환자의 can 값보다 상대적으로 크게 나타나는 경향이 있음을 알 수 있다. 실제로 can의 입원 환자 평균 (표준편차) 값은 2.82 (1.010)이고 퇴원 환자 평균 (표준편차) 값은 3.49 (0.968)가 되어 이표본 t-검정 (two-sample t-test)에 의해 유의적인 차이가 있는 것을 확인할 수 있었다. 따라서 can을 통해 입원과 퇴원 집단을 예측하는 데 도움이 될 수 있을 것이라는 것을 알 수 있었다.

각 퇴원구간이 최적 선형결합 can에 미치는 영향력으로 그 중요성을 판단할 수 있다. 중요성을 판단

하는 대표적인 방법으로는 긱 설명변수 집단 계수값들의 범위 (range)와 반응변수와 각 설명변수 사이의 편상관 (partial correlation)이 있다 (허명회, 1997). 각 퇴원구간의 범위와 편상관 값을 구하여 정리한 것이 표 3.3에 주어져 있다.

표 3.3 퇴원구간 변수들의 범위와 편상관에 의한 중요성 평가

변수	범위	편상관
WBC	1.011	0.140
PLT	0.555	0.071
Cl-	1.038	0.135
AST	0.248	0.023
ALT	0.602	0.050
$PCO2$	0.276	0.038
$PO2$	0.124	0.014
$O2SAT$	0.344	0.046
LDH	0.466	0.052
$Ca2+$	0.491	0.068
$Mg2+$	0.366	0.050

표 3.3에 의하면 범위 기준으로 중요성이 가장 높은 네 개의 변수는 $Cl-$, WBC, ALT, PLT이며, 편상관 기준으로 중요성이 가장 높은 세 개의 변수는 WBC, $Cl-$, PLT, $Ca2$로서 두 방법에 의한 중요성 평가에 약간의 차이가 있다. 특히 LDH 퇴원구간의 경우 다른 변수와 달리 정상구간이 아닌 꼬리 부분에서 퇴원이 결정되는데도 불구하고 중요성이 그리 떨어지지 않는 것은 놀라운 일이다.

실제 임상전문가에게는 이 11개 혈액검사 결과가 모두 중요하게 고려되기 때문에 그 중요성을 가중값으로 선형결합한 can을 입퇴원 결정의 기준 변수로 잡았다. 입퇴원 결정을 위한 can의 최적 경계값 (optimal threshold)을 구하기 위해서 2절에서 설명되었던 $AM = (r+s)/2$ 및 $HM = 2/(1/r+1/p)$를 모형평가 기준으로 사용하였다. can의 입원 환자의 평균이 2.82이고 퇴원 환자의 평균이 3.49이기 때문에 2.8에서 3.5까지 0.1씩 경계값을 증가 시켜가면서 여러 모형평가 측도들을 정리한 것이 표 3.4에 주어져 있다. (참고로 제일 마지막 줄에 각 퇴원구간 변수에 동일한 가중값을 주었을 때 최적 모형으로 판단되었던 '퇴원구간 7개 이상'의 결과를 포함시켰다.)

표 3.4 최적 선형결합 can의 경계값에 따른 모형평가 측도들

경계값	민감도(r)	특이도(s)	정확율(p)	정분류율	AM	HM	AM+HM
3.5	0.530	0.726	0.394	0.731	0.628	0.452	1.080
3.4	0.554	0.702	0.384	0.718	0.628	0.454	1.081
3.3	0.583	0.676	0.377	0.706	0.630	0.458	1.088
3.2	0.613	0.642	0.365	0.686	0.628	0.458	1.085
3.1	0.631	0.596	0.344	0.654	0.613	0.445	1.059
3.0	0.685	0.546	0.336	0.628	0.615	0.451	1.066
2.9	0.702	0.520	0.330	0.612	0.611	0.449	1.060
2.8	0.750	0.476	0.325	0.589	0.613	0.453	1.066
퇴원구간 7개 이상	0.750	0.472	0.344	0.542	0.611	0.472	1.083

모형 평가 결과 경계값 3.3이 최적의 모형으로 나타났고 경계값 3.2와 3.1도 거의 비슷한 성능을 보였다. '퇴원구간 7개 이상'의 결과와 비교하였을 때 '경계값 3.3' 방법이 AM에서 다소 크게 나타나고 HM에서 다소 작게 나타났으나, 두 개의 합 $AM + HM$에서는 근소하게 크게 나타났다.

'퇴원구간 7개 이상'이 HM에서 약간의 우위를 보이는 이유는 민감도가 아주 크기 때문이다. 그러나 이것으로 인해 특이도가 0.5에도 미치지 못하며, 특히 정분류율이 0.55에 미치지 못하는 문제점이 있었

다. 그에 반해 '경계값 3.3' 방법은 민감도를 제외한 특이도, 정확률, 정분류율 값이 모두 커졌으며 특히 특이도와 정분류율이 대폭 커졌다. 만약 민감도와 특이도가 비슷한 값을 가지며 모두 0.6을 상회하기 원한다면 '경계값 3.2' 방법을 이용할 수도 있을 것이다.

4. 결론 및 논의

이 연구는 Park 등 (2010)에 의해 설정된 (혈액검사 결과들의) 퇴원구간 (discharge interval)이 입퇴원 결정에 미치는 영향력에는 차이가 있을 것이라는 점에 착안하여 시작되었다. 실제로 Park 등 (2010)은 호흡곤란을 주호소로 내원한 668명의 환자를 대상으로 11개 혈액검사 결과 중 퇴원구간에 속한 개수를 가지고 입퇴원 결정을 위한 간편한 통계모형을 제안하였다. 그런데 11개 혈액검사의 결과에 대한 중요성을 고려하지 않아 모형의 성능이 떨어질 수 있다는 문제점이 있었다.

이 연구에서는 Park 등 (2010)과 동일한 퇴원구간을 사용하되 수량화 방법 (quantification method)에 의해 각 퇴원구간의 중요성을 평가하였다. 수량화 방법에 의해 각 퇴원구간의 최적 가중값을 구한 다음, 이 가중값을 이용한 선형결합으로 입퇴원 결정을 위한 기준변수로 삼았다. 이 기준변수의 최적 경계값 선택과 기존 모형과의 성능 비교를 위해 불균형 집단 (imbalance class)에서 흔히 사용하는 기준인 민감도 (sensitivity)와 특이도 (specificity)의 산술평균 AM과 민감도와 정확률 (precision)의 조화평균 HM를 사용하였다.

그 결과 이 최적 선형결합의 전체 평균인 2.99보다 약간 큰 3.3을 경계값 (threshold)으로 잡는 것이 최적의 모형으로 선정되었다. 기존의 동일 가중값을 주는 모형과 비교하였을 때 AM에서 다소 크게 나타났고 HM에서 다소 적게 나타났으나, 두 개의 합 $AM + HM$에서는 근소하게 크게 나타났다. 그러나 동일 가중값 방법의 특이도가 0.5에도 미치지 못하며, 특히 정분류율도 0.55에 미치지 못하는 문제점을 극복한 것으로 나타났다.

이 연구의 결과를 일반 호흡곤란 환자의 입퇴원 결정 모형으로 사용하기에는 한계가 있다. 우선 각 혈액검사 결과의 퇴원구간 설정이 하나의 병원 자료에 근거하기 때문에 일반성이 결여되어 있기 때문이다. 실제로 A 의료원의 임상전문가들이 각 혈액검사 결과의 정상으로 간주하는 구간과 이 연구에서 사용된 퇴원구간이 다소간 차이를 보이고 있다. 또한 이 연구에는 호흡곤란 환자의 상태를 판단하는데 아주 중요한 변수인 흉부 방사선 검사 자료가 누락되어 있기 때문이다. 흉부 방사선 검사 자료를 얻을 수 있다면 훨씬 예측력이 높은 통계모형을 도출할 수 있으리라 기대된다.

참고문헌

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

허명회 (1992). <수량화 방법론의 이해>, 자유아카데미, 서울.

Chawla, N. V., Japkowwicz, N. and Nolcz, A. (2004). Editorial: Special issue 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.

Park, C., Kim, T. Y., Kwon, O. J. and Park, H. S. (2010). A simple statistical model for determining the admission or discharge of dyspnea patients. *Journal of the Korean Data & Information Science Society*, **21**, 279-289.

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

Journal of the Korean Data & Information Science Society
2011, **22**(3), 477–485

한국데이터정보과학회지

A quantification study of blood test results for dyspnea patients[†]

Cheolyong Park[1]

[1]Department of Statistics, Keimyung University

Received 12 April 2011, revised 11 May 2011, accepted 16 May 2011

Abstract

Park *et. al* (2010) proposed a statistical model for determining the admission or discharge of 668 patients with a chief complaint of dyspnea by the number of 11 blood tests belonging to the corresponding discharge intervals. Since this method does not take into consideration the importance of each blood test result, its performance might not be optimally good. In this study, we employ a quantification method to evaluate the importance of those blood test results, and then provide a new statistical model that takes the importance into consideration. The results show that the performance of this new model is a little better than that of the model by Park *et. al* (2010).

Keywords: Admission or discharge, dyspnea patients, imbalance class, kernel density function, quantification method.

[†] 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] Professor, Department of Statistics, Keimyung University, Daegu 704-701, Korea.
E-mail: cypark1@kmu.ac.kr