대한의료정보학회지 제15권 제4호
J Kor Soc Med Informatics 2009;15(4):475-481

원저

데이터의 공간적인 분포를 이용한 가변 임계값 기반 특징선택

손창식[1], 신아미[2], 이영동[1], 박희준[2], 박형섭[3], 김윤년[3]

계명대학교 생체정보기술개발사업단[1], 계명대학교 의과대학 의료정보학교실[2], 내과학교실[3]

Variable Threshold based Feature Selection using Spatial Distribution of Data

Chang-Sik Son[1], A-Mi Shin[2], Young-Dong Lee[1], Hee-Joon Park[2], Hyoung-Seob Park[3], Yoon-Nyun Kim[3]

Biomedical Informatics Technology Center, Keimyung Univ.[1],
Dept. of Medical Informatics, School of Medicine, Keimyung Univ.[2],
Dept. of Internal Medicine, School of Medicine, Keimyung Univ.[3]

Abstract

Objective: In processing high dimensional clinical data, choosing the optimal subset of features is important, not only for reduce the computational complexity but also to improve the value of the model constructed from the given data. This study proposes an efficient feature selection method with a variable threshold. **Methods:** In the proposed method, the spatial distribution of labeled data, which has non-redundant attribute values in the overlapping regions, was used to evaluate the degree of intra-class separation, and the weighted average of the redundant attribute values were used to select the cut-off value of each feature. **Results:** The effectiveness of the proposed method was demonstrated by comparing the experimental results for the dyspnea patients' dataset with 11 features selected from 55 features by clinical experts with those obtained using seven other classification methods. **Conclusion:** The proposed method can work well for clinical data mining and pattern classification applications. ***(Journal of Korean Society of Medical Informatics 15-4, 475-481, 2009)***

Key words: Feature Selection, Variable Threshold, Pattern Classification, Dyspnea Patients

Received for review: August 11, 2009; **Accepted for publication:** November 28, 2009

Corresponding Author: Hee-Joon Park, Department of Medical Informatics, School of Medicine, Keimyung University, 194, Dongsan-dong, Jung-gu, Daegu 700-712, Korea
Tel: +82-53-428-7952, **Fax:** +82-53-428-7953, **E-mail:** hjpark@dsmc.or.kr

* This work was supported by the grant No. RTI04-01-01 from the Regional Technology Innovation Program of the Ministry of Knowledge Economy (MKE)
DOI:10.4258/jksmi.2009.15.4.475

Ⅰ. 서론

호흡곤란은 환자의 주관적인 증상으로 빈호흡, 기좌호흡, 체인스톡(cheyne-stokes) 호흡, kussmaul 호흡의 형태로 관찰 가능하고, 응급실에서 볼 수 있는 가장 흔한 주호소(chief complaint) 중 하나이다. 응급실에 호흡곤란을 주호소로 내원한 환자는 크게 심인성 질환과 폐인성 질환으로 구분할 수 있는데, 심인성 질환은 좌심실 부전, 폐부종, 울혈성 심주전 등이 주요 원인이며, 폐인성 질환은 만성 폐쇄성 폐질환, 폐렴, 폐암 등이 주요 원인이다[1]. 이러한 호흡곤란의 원인 질환은 짧은 시간의 문진으로 진단을 하기가 어렵기 때문에 임상전문가들은 피검사나 흉부 방사선 검사 등을 이용하여 진단을 하고 있으며, 검사된 항목들의 결과로부터 중요특징을 분석하고 감별하는데 많은 시간을 투자하고 있다.

일반적으로 특징선택방법은 주어진 데이터로부터 관련된 특징들의 하위집합을 선택하거나, 새로운 특징들을 결합하는 방법으로 고차원의 문제를 저차원으로 변화하여 처리한다. 게다가 특징 수의 증가에 따른 계산 복잡도(computational complexity)나 차원의 저주(curse of dimensionality)를 효과적으로 해결할 수 있다는 장점 때문에 패턴분류나 의사결정등과 같은 문제에 전처리과정으로 사용되고 있으며[2-4], 특히 신경망(neural network)과 유전자알고리즘(genetic algorithm)과 같은 기계학습 알고리즘을 이용한 방법들이 주목할만한 성능을 제공하고 있다. 하지만 신경망[5-7], 유전자알고리즘에 기반한 방법[8-11]들은 관련 특징들을 선택하는데 있어서 여러 학습 매개변수(즉 신경망의 경우, 학습률, 모멘텀, 연결가중치 가지치기 수준, 유전자알고리즘의 경우 목적함수에서 사용된 매개변수 등)들을 조정해야 하고, 선택된 특징들 간의 관계에 대한 해석(interpretability)이 어렵거나 불가능하다는 제약점을 가진다. 그러므로 이들 기계학습 알고리즘은 다변수 혹은 고차원으로 이루어진 임상 데이터에서 중요특징을 선택하기 위한 도구로는 적합하지 않으며, 선택된 특징들과 cut-off 값이 임상기준과 비교하였을 때의 신뢰성 여부가 추가적으로 평가되어야 한다.

본 연구에서는 이러한 점들을 고려하기 위해서 속성값들의 공간적인 분포를 이용한 가변 임계값 기반 특징선택방법을 제안한다. 실험에서는 제안한 방법의 타당성을 검증하기 위해서 응급실에 호흡곤란으로 내원한 환자들의 검사된 항목으로 구성된 데이터-셋을 이용하여 기존의 7가지 분류방법들과의 교차검정 결과를 비교하였고, 가변적인 임계값의 변화에 따라 선택된 중요특징들의 cut-off 값들을 제시하였다.

Ⅱ. 재료 및 방법

1. 연구대상 및 자료수집

본 연구에서는 대구광역시에 소재한 동산의료원에 2006년 7월에서 2007년 6월 사이에 호흡곤란을 주호소로 응급실에 내원한 환자 1,129명의 의무기록을 대상으로 하였다. 대상자의 인적 사항을 제외한 등록번호, 성별, 나이, 응급실 내원일자 및 시간, 진료결과, 입원 시 진단, 초기 검사 항목 등의 자료를 데이터웨어하우스에서 추출하였다. 초기 검사 항목으로는 전혈구 검사(common blood cell & differential count, CBC & diff. count), 프로트롬빈 시간(prothrombin time, PT), 활성화 부분 트롬보플라스틴 시간(activated partial thromboplastin time, aPTT), 혈청 전해질(serum electrolytes), 입원환자에 대한 기본 검사(routine admission), 혈청 아밀라제, 동맥혈 가스 분석(blood pH and gas), 리파아제, CK-MB, Troponin I, CK, LDH, CRP, Fibrinogen, Ca^{2+}, Mg^{2+}, Pro-BNP가 있었다. 수집된 자료 중 타병원으로 전원된 환자, DOA (death on arrival), CPR (Cardio-Pulmonary Resuscitation) 후 혹은 DNR (Do Not Resuscitate)로 사망한 환자, 자의퇴원 혹은 미상의 기타 환자, 의무기록이 불완전한 경우를 제외한 총 668명의 환자(입원환자 500명, 퇴원환자 168명)에 대한 데이터를 분석하였다.

2. 특징선택

만약 임의의 데이터 $X=\{\chi_i | \chi_i=(\chi_{i1}, \chi_{i2}, \cdots, \chi_{in}), i=1,2,\cdots,s\}$가 s개의 인스턴스들을 포함하는 n차원 벡터이고, j번째 속성(즉 특징) $\alpha_j=\{\chi_{1j}, \chi_{2j}, \cdots, \chi_{sj}\}$, $(j=1,\cdots,n)$은 s개의 데이터 포인터들로 이루어져 있고,

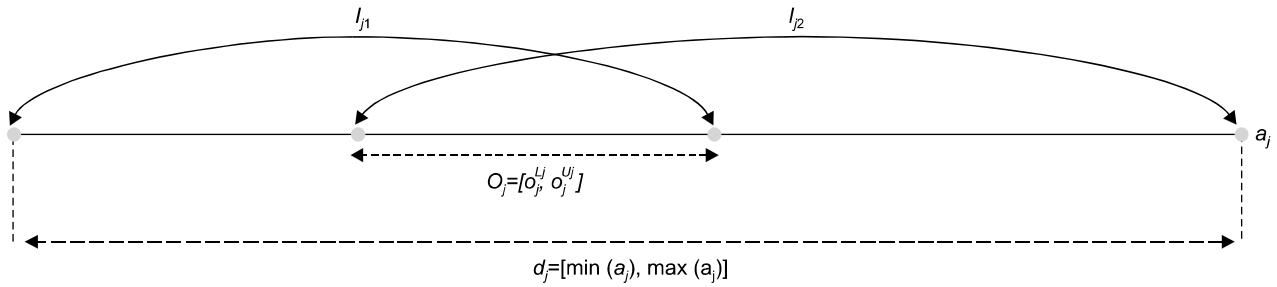

Figure 1. An overlapping O_j region of the input attribute α_j

출력 $C=\{c_k|k=1,2,\cdots,m\}$은 m개의 클래스로 이루어진 집합이라고 가정하자. 이때 속성 값들의 공간적인 분포를 고려한 중요특징의 선택은 다음 과정으로 결정된다.

Step 1: j번째 속성 α_j의 도메인 d_j=[min(α_j), max (α_j)]과 임의의 클래스에 대응하는 j번째 속성의 속성값들의 내부구간 $I_{jk}=[I_{jk}^l, I_{jk}^u]$, $I_{jk}\in d_j$들을 추출한다. 여기서 I_{jk}^l와 I_{jk}^u는 j번째 속성에서 k번째 클래스에 속하는 속성 값들의 최소값과 최대값을 나타낸다.

Step 2: 내부구간들 사이에서 중첩영역(overlapping region) $O_j=[o_j^{L_j}, o_j^{U_j}]$을 찾는다.$^k_j$여기서 $o_j^{L_j}$와 $o_j^{U_j}$는 모든 중첩된 영역들 사이에서 하한(lower)과 상한경계(upper bound)을 나타낸다(Fig. 1).

Step 3: 중첩영역 내에서 서로 다른 클래스들이 동일한 속성 값들을 가지지 않을 때 속성 값들의 빈도수를 계산하고, 해당 속성의 적합도(fitness degree)를 평가한다.

$$H_j = \sum_{k=1}^{m} h_j^k, \quad where\ h_j^k = \frac{t}{s} \tag{1}$$

여기서 t_j^k는 j번째 속성에서 k번째 클래스에 포함된 유일한 속성 값들의 빈도수를 의미하고, s는 인스턴스들의 수를 나타낸다. 이때 h_j^k는 중첩영역에서 k번째 클래스에 속한 속성 값들의 상대적인 분리가능성 정도를 나타내고, $h_j^k\in[0,1]$의 값을 가진다.

예를 들어, 전체 인스턴스의 수가 20일 때 j번째 속성의 중첩영역에서 다음과 같은 클래스별 속성 값들의 분포가 존재한다고 가정하자.

이때 j번째 속성의 적합도는 식 (1)에 의해서 클래스 1과 클래스 2의 분리가능성 정도의 합으로 나타낼 수 있다: $H_j=h_j^1+h_j^2=0.2+0.2=0.4$. 따라서 j번째 속성의 최대 분리가능성 정도는 0.4라는 것을 알 수 있으며, 이러한 방법으로 각 속성의 분리가능성 정도를 평가함으로써 각 속성의 중요성 정도를 순위화(ranking)할 수 있다.

Step 4: 각 속성의 분할경계(즉 cut-off 값)을 결정하기 위해서 중첩영역 내에서 중복된 속성 값들의 가중치 평균값(weighted average value), 즉 무게중심 값을 계산한다.

$$B_j=\frac{\sum_{i=1}^{s} n_{ij}\times x_{ij}}{\sum_{i=1}^{s} n_{ij}},\ \text{for}\ \chi_{ij}\in O_j \tag{2}$$

여기서 χ_{ij}는 j번째 속성에서 i번째 인스턴스의 속성 값을 의미하고, n_{ij}는 중복된 속성값들의 빈도수를 나타낸다. 예를 들어 Step 3의 예제에서 서로 다른 2개의 클래스에 대한 중복된 속성 값 1.1, 1.3, 1.4의 빈

α_j	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8
t_j^1	1	2	1	1	2	2	0	0	0
t_j^2	0	1	0	2	1	0	1	2	1
h_j^1	1/20	0	1/20	0	0	2/20	0	0	0
h_j^2	0	0	0	0	0	0	1/20	2/20	1/20

도수는 각각 3, 3, 3이므로, 식 (2)에 의해서 계산된 무게중심 값은 11.4/9=1.2667이 된다. 이것은 전체 중첩영역에 대한 무게중심 1.3765의 오분류(7개)에 비해 상대적으로 적은 오분류(6개)을 가진다.

제안된 방법에서는 이러한 방법으로 속성값들의 공간적인 분포를 고려하여 각 속성의 분할경계를 결정하고 결합함으로써 비선형 공간을 생성하였다. 주어진 데이터로부터 규칙패턴을 생성하기 위해서 다음과 같은 형식의 IF～THEN 구조를 사용하였다.

$$R_i : IF\ \ \alpha_1\ is\ A_{i1}\ and\ \ \alpha_2\ is\ A_{i2} \cdots \alpha_n\ is\ A_{in}$$
$$THEN\ Class\ is\ c_k\ with\ freq_i^k$$

여기서 α_j (j=1,2,⋯n)는 j번째 속성을, A_{ik} (i=1,2,⋯, M; k=1,2,⋯,m)는 식 (2)로부터 계산된 cut-off 값을 중심으로 분할된 번째 규칙에서 번째 클래스의 구간을, 는 번째 규칙패턴이 번째 클래스에 대응될 때의 발생빈도수(occurrence frequency)를 나타낸다.

이러한 클래스 별 규칙의 발생 빈도수 $freq_i^k$을 근거로 클래스들 간 규칙의 충돌문제(conflict problem)를 다음과 같은 기준으로 분해하였다.

$$R_i^* = \begin{cases} c_i & if\ freq_i^k(R_i) > freq_i^k(R_j) \\ c_j & else\ if\ freq_i^k(R_i) < freq_i^k(R_j) \\ NA & otherwise \end{cases} \quad (3)$$

여기서 R_i^*는 식 (3)에 의해서 새롭게 정의된 i번째 규칙패턴 R_i의 출력을 나타내고, NA는 동일한 규칙패턴이 서로 다른 클래스에서 동일한 발생 빈도수를 가질 때, 어떠한 클래스의 규칙패턴으로도 정의할 수 없는 규칙을 의미한다.

III. 결과

본 연구에서는 응급실 호흡곤란으로 내원한 1,129명의 환자 중 입원환자 500명, 퇴원환자 168명에 대한 데이터 셋, 즉 전체 55개의 입력속성들 중 임상전문가에 의해서 결정된 11개의 속성만을 사용하여 각 속성의 적합도를 근거로 가변 임계값(variable threshold) α, $\alpha \in [0, 1]$의 변화에 따른 특징선택의 변화와 분류정확도를 분석하였다.

1. 전체 데이터 집합을 훈련-실험 데이터로 사용

실험 1에서는 전체 데이터 집합을 훈련-실험 데이터로 사용한 경우에 가변 임계값의 변화에 따라 선택된 특징들의 변화와 분류정확도를 비교하였다. 각 특징들의 통계적인 정보는 Table 1과 같고, 입・퇴원 환자에 대한 각 속성(즉 검사항목)의 중첩영역은 *WBC* =[2.77, 53.47], *PLT* : [52, 676], Cl^- : [90, 124], *AST* : [10, 645], *ALT* : [5, 212], PCO_2 : [9.5, 96], PO_2 : [39.4, 134], O_2SAT : [87.8, 99.7], *LDH* : [268, 5,006], Ca^{2+} : [1.57, 2.93], Mg^{2+} : [1.1, 3.3]이었다(Steps 1과 2 참조). 또한 식 (1)을 근거로 각 속성의 적합도를 평가한 결과, *WBC* (0.8114), *LDH* (0.732), PO_2 (0.6751), *PLT* (0.5734), PCO_2 (0.5015), *AST* (0.1662), *ALT* (0.1332), Ca^{2+} (0.1108), O_2SAT (0.0883), Cl^- (0.0254), Mg^{2+} (0.0150) 순으로 나타났다. 이것은 속성 *WBC*가 속성 Mg^{2+}에 비해 상대적으로 부분적인 선형분리 가능성 정도가 높다는 것을 의미하고, 11개의 속성들 중에서 *WBC* 속성 값의 분포가 전체 공간적인 분포에 미치는 영향이 크다는 것을 보여준다. Table 2에서 6까지는 가변 임계값 α을 0.1, 0.2, 0.6, 0.7, 0.8으로 변

Table 1. Dataset's feature

Feature*	Unit	Min	Max	Mean±SD
WBC	$\times 10^3/\mu l$	0.11	75.9	11.0196±6.3942
PLT	$\times 10^3/\mu l$	23	1,105	270.6856±120.0240
Cl^-	*mmol/L*	72	134	104.2455±6.9351
AST	*U/L*	5	3,321	73.5195±227.1527
ALT	*U/L*	3	2,481	46.4416±143.2200
PCO_2	*mmHg*	8.3	98.5	39.8760±13.5689
PO_2	*mmHg*	35.9	354	80.1507±22.0636
O_2SAT	%	59	99.9	96.1350±3.1972
LDH	*U/L*	152	8,178	688.5834±509.8108
Ca^{2+}	*mEq/L*	1.25	3.2	2.2451±0.1706
Mg^{2+}	*mg/dl*	0.3	4.1	2.2054±0.3770

* *WBC*: White blood cell
PLT: Platelet count
Cl^-: Chloride
AST: Aspartate transaminase
ALT: Alanine transminase
PCO_2: Pressure of carbon dioxide
PO_2: Pressure of oxygen
O_2SAT: Oxygen saturation
LDH: Lactate dehydrogenase
Ca^{2+}: Calcium
Mg^{2+}: Magnesium

화시켰을 때에 선택된 속성들과 cut-off 값, 각 속성의 적합도, 입・퇴원 환자에 대한 규칙 수의 변화와 분류 정확도를 보여준다. 이들 결과에서 볼 수 있듯이, 가변 임계값 α의 값이 증가할수록 선택된 특징들과 규칙의 수는 점차적으로 감소하는 경향을 보였고, 분류 정확도는 56.2874%에서 74.8503%까지 증가됨과 동시에 수렴되는 특성을 나타내었다. 또한 각 검사항목들에 대한 동산 의료원의 임상지침(*WBC* : [5.2, 12.4], *PLT*: [130, 400], Cl^- : [95, 108], *AST* : [13, 36], *ALT* : [5, 44], PCO_2 : [32, 48], PO_2 : [83, 108], O_2SAT : [95, 99], *LHD* : [211, 423], Ca^{2+} : [1.2, 3.2], Mg^{2+} : [1.5, 2.7])과 비교하였을 때 는 정상치 범위에서 다소 높게, PO_2는 다소 낮게 평가되었고, 나머지 검사항목들의 cut-off 값은 모두 정상치 범위에서 결정됨을 볼 수 있었다(Table 2).

Table 2. Classification accuracy and selected features when the threshold (α=0.1)

Selected feature	Cut-off	Fitness	Num. rules in admission	Num. rules in discharge	Num. total	Accuracy (%)
WBC	9.0914	0.8114	179	20	199	56.2874
LDH	550.0329	0.7320				
PO_2	80.4923	0.6751				
PLT	247.9383	0.5734				
PCO_2	35.9192	0.5015				
AST	31.7305	0.1662				
ALT	20.4743	0.1332				
Ca^{2+}	2.2395	0.1108				

Table 3. Classification accuracy and selected features when the threshold (α=0.2)

Selected feature	Cut-off	Fitness	Num. rules in admission	Num. rules in discharge	Num. total	Accuracy (%)
WBC	9.0914	0.8114	32	0	32	74.8503
LDH	550.0329	0.7320				
PO_2	80.4923	0.6751				
PLT	247.9383	0.5734				
PCO_2	35.9192	0.5015				

Table 4. Classification accuracy and selected features when the threshold (α=0.6)

Selected feature	Cut-off	Fitness	Num. rules in admission	Num. rules in discharge	Num. total	Accuracy (%)
WBC	9.0914	0.8114	8	0	8	74.8503
LDH	550.0329	0.7320				
PO_2	80.4923	0.6751				

2. 10-fold 교차검정

실험 2에서는 제안된 방법의 타당성을 보이기 위해서 전체 데이터 집합에 대한 10-fold 교차검정분석(훈련: 90%, 실험: 10%)과 기존의 7가지 분류기 C4.5[12], kNN (k-nearest neighbor)[13], LDA (linear discriminant analysis)[14], QDA (Quadratic discriminant analysis)[15], 3가지 형태의 커널함수(polynomial, sigmoid, rbf)을 기반으로 한 SVM (support vector machine)[16]과 비교하

Table 5. Classification accuracy and selected features when the threshold (α=0.7)

Selected feature	Cut-off	Fitness	Num. rules in admission	Num. rules in discharge	Num. total	Accuracy (%)
WBC	9.0914	0.8114	4	0	4	74.8503
LDH	550.0329	0.7320				

Table 6. Classification accuracy and selected features when the threshold (α=0.8)

Selected feature	Cut-off	Fitness	Num. rules in admission	Num. rules in discharge	Num. total	Accuracy (%)
WBC	9.0914	0.8114	2	0	2	74.8503

Table 7. Results of 10-fold cross validation when the threshold (α=0.5)

Fold	Train (%)		Test (%)	Fold	Train (%)		Test (%)
k=1	74.5424	(32)	74.6269	*k*=6	74.8752	(32)	74.6269
k=2	74.8752	(32)	74.6269	*k*=7	74.8752	(31)	74.6269
k=3	74.8752	(31)	74.6269	*k*=8	73.7105	(32)	71.6418
k=4	74.8752	(31)	74.6269	*k*=9	72.7575	(31)	75.7576
k=5	74.5424	(32)	74.6269	*k*=10	74.7508	(32)	75.7576
Total			Avg. train: 74.4680,	Avg. test: 74.5545,	Num. rule: 31.6		

Table 8. Comparison results between the proposed method and the conventional methods (10-fold cross validation)

Method		Avg. train (%)	Avg. test (%)
Decision tree*	C4.5	78.0433	70.9408
Statistical classifiers†	*k*NN (*k*=1)	68.4959	68.7065
	*k*NN (*k*=2)	73.6860	73.3492
	*k*NN (*k*=3)	70.1428	69.3012
	LDA	74.6507	74.5545
	QDA	48.5199	45.9362
SVMs‡	Polynomial	25.1497	25.1470
	Sigmoid	74.8503	74.8530
	RBF	100	74.8530
Proposed method	α =0.5	74.4680	74.5545

* Experiment condition: 1) Confidence: 0.25, 2) Number of leafs: 2
† Experiment condition: 1) *k*=1-3, 2) Measure: Euclidean distance
‡ Experiment condition: 1) Kernel type: Polynomial, Sigmoid, and RBF functions, 2) eps: 0.001, 3) d (degree): 10, 4) g (gamma): 1.0, 5) r (coef0): 1.0, 6) n (nu): 0.5, 7) epsilon: 1.0, 8) h (shrinking): 0

였다. Table 7은 임상 전문가에 의해서 결정된 임계값일 때 10-fold 교차검정 결과를 보여주고, '( )'의 값은 각 fold에서 생성된 전체 규칙의 수를 나타낸다. 실험결과, 10-fold 교차검정 동안에 훈련과 실험 데이터의 평균 분류정확도는 각각 74.4680%, 74.5545%이고, 생성된 평균 규칙의 수는 31.6개였다. 각 fold에서 중요특징으로는 실험 1에서처럼 *WBC, LDH, PO_2, PLT, PCO_2* 순위로 선택되었고, 교차검증 동안에 각 속성의 cut-off 값들의 변화된 범위를 분석한 결과 *WBC* (8.8771-9.4402), *LDH* (532.2919-560.1838), *PO_2* (79.3269- 81.9370), *PLT* (243.4908-253.0885), *PCO_2* (35.2952- 36.3225)이었으며, 실험 1의 결과에서처럼 *LDH*는 정상치 범위보다 다소 높게, *PO_2*는 다소 낮게 평가됨을 볼 수 있었다. 또한 기존의 7가지 분류기의 10-fold 교차검정결과를 비교해 볼 때 우수한 성능을 보이는 sigmoid, rbf 커널함수를 기반으로 한 SVM, LDA와 유사한 성능을 보였다(Table 8).

Ⅳ. 고찰

본 연구에서는 각 속성의 중첩영역 내에 포함된 속성 값들의 공간적인 분포에 대한 분할가능성 정도를 평가함으로써 데이터 셋에서 중요특징을 선택할 수 있는 방법을 제안하였고, 분할된 공간으로부터 생성된 규칙패턴들 간의 충돌문제를 최소화할 수 있는 방법에 대해서도 논의하였다. 제안된 방법의 타당성을 보이기 위해서 응급실에 호흡곤란를 주호소로 내원한 668명 환자들의 입·퇴원에 대한 데이터를 2가지 실험을 통해 분석하였다. 실험 1에서 전체 데이터 집합을 훈련-실험 데이터로 사용한 경우에는 가변 임계값의 변화에 따라 선택된 특징 수와 cut-off 값, 규칙 수, 분류정확도의 변화를 나타내었고, 그 결과 5개의 특징 *WBC, LDH, PO_2, PLT, PCO_2*가 나머지 6개의 특징 *AST, ALT, Ca^{2+}, O_2SAT, Cl^-, Mg^{2+}*에 비해 상대적인 중요도가 높은 것으로 평가되었고, 호흡곤란환자의 입·퇴원 결정에 있어서 중요한 영향을 미치는 특징이라는 사실을 알 수 있었다. 또한 실험 2의 10-fold 교차검정 실험에서는 임계값이 0.5일 때 선택된 중요특징 *WBC, LDH, PO_2, PLT, PCO_2*와 각 특징들의 cut-off 값들의 변화된 범위를 나타내었다. 그리고 기존의 7가지 분류방법들 C4.5, kNN, LDA, QDA, 3가지 형태의 커널함수를 기반으로 한 SVM과의 분류정확도를 비교했을 때 가장 좋은 분류성능을 나타낸 sigmoid, rbf 커널함수를 사용한 SVM과 Fisher의 LDA와 유사하거나 동일한 분류성능을 제공함을 볼 수 있었다. 또한 전체 도메인 영역을 고려하지 않고 중첩영역(즉 비선형 분리 가능한 공간)에서 분할 경계를 결정함으로써 신뢰할 수 cut-off 값을 얻을 수 있었으며 보다 효과적으로 특징공간을 분할할 수 있음을 확인하였다. 향후 연구로는 데이터-집합에 의존적으로 최적의 임계값을 결정할 수 있는 방법과 불균형(imbalance) 분포를 가진 데이터-셋으로부터 cut-off 값을 결정하고 중요특징을 선택할 수 있는 방법에 대한 연구가 필요할 것으로 사료된다.

참고문헌

1. Jeven P, Ewens B. Assessment of a breathless patient. Nursing 2001;15(16):48-55.
2. Blum AL, Langley P. Selection of relevant features and examples in machine learning. Artificial Intelligence 1997;97(2):245-271.
3. Kohavi R, John GH. Wrappers for feature subset selection. Artificial Intelligence 1997;92(2):273-324.
4. Dash M, Liu H, Yao J. Dimensionality reduction of unsupervised data. ICTAI, 9th Int Conf. Tools with

Artificial Intelligence (ICTAI '97) 1997:532-539.
5. Steppe JM, Bauer KW, Rogers SK. Integrated feature and architecture selection. IEEE Trans. Neural Networks 1996;7(4):1007-1014.
6. De RK, Pal NR, Pal SK, Feature analysis: neural network and fuzzy set theoretic approaches. Pattern Recognition 1997;30(10):1579-1590.
7. Li RP, Mukaidono M, Turksen IB. A fuzzy neural network for pattern classification and feature selection. Fuzzy Sets and Systems 2002;130(1):101-108.
8. Yang J, Honavar V. Feature subset selection using a genetic algorithm. IEEE Intelligent Systems 1998; 13(2):44-49.
9. Vafaie H, Jong D. Feature space transformation using genetic algorithm. IEEE Trans. Intelligent Systems 1998;13(2):57-65.
10. Tseng LY, Yang SB. Genetic algorithms for clustering, feature selection and classification. Proc IEEE Int Conf Neural Networks 1997;3:1612-1615.
11. Elalami ME. A filter model for feature subset selection based on genetic algorithm. Knowledge-Based Systems 2009;22(5):356-362.
12. Quinlan JR. C4.5: programs for machine learning. San Mateo, Morgan Kaufmann; 1993. pp.109-279.
13. Cover TM, Hart PE. Nearest neighbor pattern classification. IEEE Trans. Information Theory 1967;13(1): 21-27.
14. Fisher RA. The use of multiple measurements in taxonomic problems. Annals of Eugenics 1936;7: 179-188.
15. Friedman JH. Regularized discriminant analysis. J American Statistical Association 1989;84(405):165-175.
16. Cortes C, Vapnik V. Support-vector networks. Machine Learning 1995;20(3):273-297.