J. Biomed. Eng. Res : 40-49, 2010

호흡곤란환자의 입-퇴원 분석을 위한 규칙가중치 기반 퍼지 분류모델

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

[1]계명대학교 생체정보기술개발사업단,
[2]계명대학교 의과대학 의료정보학교실,
[3]계명대학교 동산병원 심장내과

Rule Weight-Based Fuzzy Classification Model for Analyzing Admission-Discharge of Dyspnea Patients

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

[1]*Biomedical Informatics Technology Center, Keimyung Univ.,*
[2]*Dept. of Medical Informatics, School of Medicine, Keimyung Univ.,*
[3]*Interventional Cardiology Dept. of Internal Medicine, Keimyung Univ.,*
(Received August 3 2009. Accepted February 19 2010)

Abstract

A rule weight-based fuzzy classification model is proposed to analyze the patterns of admission-discharge of patients as a previous research for differential diagnosis of dyspnea. The proposed model is automatically generated from a labeled data set, supervised learning strategy, using three procedure methodology: i) select fuzzy partition regions from spatial distribution of data; ii) generate fuzzy membership functions from the selected partition regions; and iii) extract a set of candidate rules and resolve a conflict problem among the candidate rules. The effectiveness of the proposed fuzzy classification model was demonstrated by comparing the experimental results for the dyspnea patients' data set with 11 features selected from 55 features by clinicians with those obtained using the conventional classification methods, such as standard fuzzy classifier without rule weights, C4.5, QDA, kNN, and SVMs.

Key words : Fuzzy Classification Model, Rule Weight, Rule Generation, Dyspnea Patient

Ⅰ. 서 론

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 하나로서 심장질환, 기도 폐쇄 질환, 폐색전증을 포함해 다수의 구별하기 어려운 원인들이 하나 혹은 복합적으로 작용해 나타난다[1]. 응급실에 호흡곤란을 주호소로 내원한 환자는 크게 두 가지 원인 질환으로 구분할 수 있는데 첫 번째는 좌심실 부전, 폐부종, 울혈성 심부전 등이 원인인 심인성 질환이고, 두 번째는 천식, 만성 폐쇄성 폐질환, 폐렴, 폐암 등의 원인인 폐인성 질환이다[2]. 호흡곤란을 유발하는 잠재적인 요소가 많은 이들 질환들을 조기에 발견해서 적절한 조치를 취하지 않는다면 생명을 위협하는 심각한 위험을 초래할 수 있다[3]. 이러한 호흡곤란의 원인질환들은 짧은 시간의 문진으로 진단을 하기가 어렵기 때문에 피검사나 흉부 방사선 검사 등을 이용하여 진단하게 되고, 이들을 감별하고 분석하기 위한 연구들이 진행되었다[4,5]. 그러나 효과적인 치료를 위해 임상전문가들은 중요한 검사항목들을 분석하고 판별하는데 많은 시간을 투자하고 있으며, 이러한 점들을 개선시키기 위한 연구가 필요하다.

최근 수십 년 동안 다양한 소프트컴퓨팅 기술이 개발되었고, 그 중에서 퍼지이론[6]은 의사결정지원시스템과 임상분야, 특히 임상검사 자동분석[7], 퍼지추론에 의한 요분석 검사결과 평가[8], 의학영상처리[9-11] 등과 같은 다양한 분야에서 전문가의 경험을 지식으로 표현하고 부정확하고 불명확한 정보를 처리하기 위한 도구로서 사용되어져 왔다. 또한 최근에는 전문가의 지식을 고려하지 않고 주어진 데이터로부터 시스템을 설계하고 규칙을 자동생성

Corresponding Author : 박희준
(704-701) 대구시 달서구 달구벌대로 2800 계명대학교 성서캠퍼스
의과대학 의료정보학교실
Tel : +82-53-580-3740 / Fax : +82-53-580-3746
E-mail : hjpark@dsmc.or.kr
본 연구는 지식경제부 지방기술혁신사업(RTI04-01-01)지원으로 수행되었음

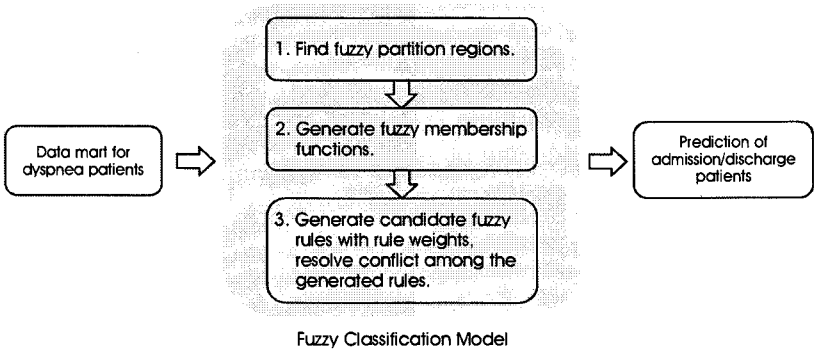

그림 1. 제안된 퍼지 분류모델의 전체적인 구조
Fig. 1. Overall structure of the proposed fuzzy classification model

하기 위한 연구가 활발히 이루지고 있다[12-16]. 그러나 이들 방법들은 초기의 퍼지분할 구간을 발견적인(heuristic) 방법으로 결정하기 때문에 유전자알고리즘(genetic algorithm)이나 정보량(entropy)과 같은 개념을 이용하여 분할구간을 지속적으로 조정해야만 한다. 따라서 이러한 계산의 복잡성을 줄이기 위한 보다 단순하고 효과적인 방법이 필요하다.

본 연구에서는 이러한 제약점을 줄이기 위한 방법으로 퍼지분할 방법[17]을 이용한 규칙가중치 기반 퍼지 분류모델을 제안한다. 제안된 모델은 응급실에 호흡곤란으로 내원한 환자들의 원인질환을 진단하기 위한 선행연구로서 호흡곤란환자의 입・퇴원을 분석하기 위한 모델로서 사용되었고, 실험에서는 제안된 모델의 효과성을 보이기 위해서 기존의 5가지 분류기법들과의 비교실험을 수행하였다.

II. 제안된 퍼지 분류모델

1. 퍼지 분류모델

응급실 호흡곤란환자의 입・퇴원 유무에 대한 패턴분석을 위해 제안된 퍼지 분류모델의 전체적인 구조는 그림 1에서와 같이 3가지 모듈로 구성된다.

A. 퍼지분할 및 소속함수 생성

만약 주어진 데이터 $\mathbb{X}=\{x_i|x_i=(x_{i1},x_{i2},...,x_{in}),\ i=1,2,...,s\}$가 s개의 인스턴스들을 포함하는 n차원 벡터이고, j번째 변수 $a_j=\{x_{1j},x_{2j},...,x_{sj}\},(j=1,2,..,n)$은 s개의 데이터 포인트들로 이루어져 있고, 출력 $C=\{c_k|k=1,2,...,m\}$은 m개의 클래스로 이루어진 집합이라고 하자. 이때 각 속성의 퍼지분할 구간의 선택[17]과 소속함수의 생성은 다음 절차에 의해서 결정한다.

Step 1: j번째 입력 변수의 도메인 $d_j=[\min(a_j),\ \max(a_j)]$과 해당 클래스에 대응하는 j번째 변수의 내부구간 $I_{jk}=[I_{jk}^l,\ I_{jk}^u]$들을 추출한다. 여기서 $I_{jk}^l=\min(a_j)$과 $I_{jk}^u=\max(a_j)$은 j번째 변수에서 k번째 클래스에 속하는 속성 값들의 최소값과 최대값을 나타낸다.

Step 2: 추출된 내부구간들 사이에서 중첩영역 $O_j=[o_j^{L_j},o_j^{U_j}]$을 추출한다. 여기서 $o_j^{L_j}$와 $o_j^{U_j}$는 모든 중첩된 영역들 사이에서 최소값(minimum value)과 최대값(maximum value)을 나타낸다(그림 2 참조).

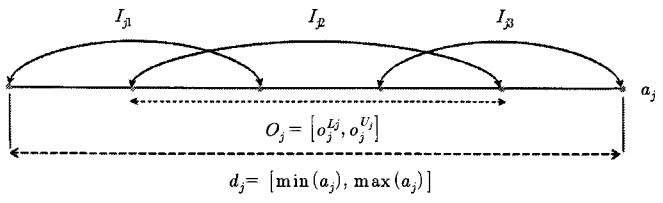

그림 2. 입력변수 a_j에서 중첩영역 O_j
Fig. 2. An overlapped region O_j of the input variable a_j

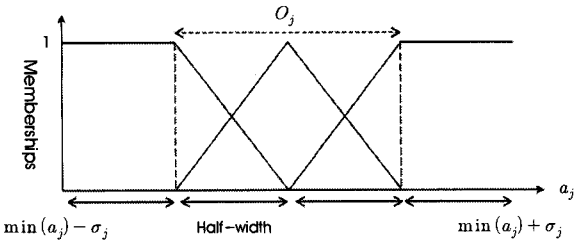

그림 3. 중첩영역에서 정의된 퍼지 소속함수
Fig. 3. Fuzzy membership functions defined on overlapping regions

Step 3: 중첩영역에서 정의될 퍼지 소속함수의 half-width를 계산한다. 만약 변수 a_j의 중첩영역에서 p개의 퍼지분할수로 나누어진다면, 퍼지 소속함수의 half-width를 계산하기 위한 식은 다음과 같이 정의된다(그림 3 참조).

$$W_j = \left(o_j^{U_j} - o_j^{L_j}\right)/(p-1) \quad (1)$$

본 논문에서는 그림 3에서 나타낸 것처럼, j번째 변수의 도메인 $d_j = [\min(a_j), \max(a_j)]$을 $d_j = [\min(a_j) - \sigma, \max(a_j) + \sigma]$로 확장함으로써 추가적인 입력 인스턴스의 매핑공간을 확보하였다. 또한 퍼지분할 수 p의 값은 통상적으로 클래스 레이블의 수와 동일한 값으로 정의하여 쓰는 것이 일반적이나, 제안된 모델에서는 Ishibuchi[12]와 Mansoori[13]의 방법에서 보여주는 것처럼 퍼지분할의 수에 따라 규칙의 수와 분류 정확도를 나타내기 위해서 2개에서 6개까지 조정하여 분석하였다.

B. 후보규칙 생성

만약 다음과 같은 형태의 퍼지 if-then 규칙이 있다고 가정하자.

$$R_j : If\ a_1\ is\ A_{j1},\ a_2\ is\ A_{j2}, \ldots, a_n\ is\ A_{jn}$$
$$Then\ Class\ is\ c_k^j\ with\ cf_j,\ (j = 1, 2, \ldots, M)$$

여기서 $a_1, a_2, \ldots, a_n$와 $Class$은 각각 n개의 입력변수(antecedent variable)와 한 개의 출력변수(output or consequent variable)를 의미하고, $A_{j1}, A_{j2}, \ldots, A_{jn}$은 각 입력변수에 대한 언어적인 값(linguistic value)을 나타낸다. 또한 $c_k^j (k = 1, \ldots, m)$는 j번째 규칙 R_j에 대응된 출력변수 값(즉 클래스 레이블)을 의미하고, cf_j ($cf_j \in [0, 1]$)는 대응된 규칙의 확신도(certainty factor)를 나타낸다. 다음은 주어진 데이터로부터 후보규칙을 생성하기 위한 과정을 보여준다.

Step 1: 주어진 데이터의 클래스 레이블을 근거로 입력변수의 값들을 오름차순으로 정렬한다.

Step 2: i번째 입력벡터 $\mathrm{x}_i = (x_{i1}, x_{i2}, \cdots x_{in})$, $\mathrm{x}_i \in c_k$에 대해서 입력부 적합도(compatibility grade)를 계산한다.

$$\mu_j^k(\mathrm{x}_i) = \mu_{j1}(x_{i1}) \times \mu_{j2}(x_{i2}) \times \cdots \times \mu_{jn}(x_{in}) \quad (2)$$

여기서 $\mu_{ji}(\bullet)$은 각 퍼지 소속함수 A_{ji}에 대한 소속정도(membership grade)를 나타낸다.

Step 3: Step 2에 의해서 계산된 입력부 적합도들 중에서 최대 적합도를 가지는 규칙패턴을 i번째 입력벡터 x_i의 후보규칙으로 할당한다.

$$\mu_i^{k^*}(\mathrm{x}_i) = \arg\max_j \left\{\mu_j^k(\mathrm{x}_i)\right\},\ j = 1, 2, \cdots, M \quad (3)$$

여기서 $\mu_i^k(\mathrm{x}_i)$은 입력벡터 x_i의 규칙패턴들을, $\arg\max_j \left\{\mu_j^k(\mathrm{x}_i)\right\}$은 규칙패턴들 중에서 최대 적합도를 가지는 규칙의 인덱스를 나타낸다. 또한 식 (3)으로부터 얻은 인덱스에 대응된 규칙의 확신도를 계산하기 위해서 식 (4)를 이용하였다.

$$cf_j = \max\left\{\mu_j^k(\mathrm{x}_i)\right\} \quad (4)$$

그러나 식 (3)과 (4)로부터 생성된 후보규칙들은 서로 다른 클래스 레이블에 대해서 하나 이상의 동일한 후보규칙을 포함할 수 있다. 예를 들어, 식 (3)과 (4)로부터 생성된 다음 4개의 후보규칙이 있다고 가정하자.

$$R_1 : If\ x_1\ is\ A_1(0.2)\ \text{and}\ x_2\ is\ B_2(0.4)$$
$$Then\ y\ is\ c_1\ with\ cf_1 = 1.0,$$
$$R_2 : If\ x_1\ is\ A_2(0.8)\ \text{and}\ x_2\ is\ B_1(0.6)$$
$$Then\ y\ is\ c_1\ with\ cf_2 = 1.0,$$
$$R_3 : If\ x_1\ is\ A_2(0.8)\ \text{and}\ x_2\ is\ B_1(0.6)$$
$$Then\ y\ is\ c_2\ with\ cf_3 = 1.0,$$
$$R_4 : If\ x_1\ is\ A_2(0.8)\ \text{and}\ x_2\ is\ B_2(0.4)$$
$$Then\ y\ is\ c_2\ with\ cf_4 = 1.0$$

위의 규칙에서 클래스 c_1과 c_2의 후보규칙 중에서 규칙 R_2와 R_3에서는 동일한 규칙패턴 (즉 $If\ x_1\ is\ A_2\ \text{and}\ x_2\ is\ B_1$ $Then\ y\ is\ c_1 \text{or} c_2$)에 대해서 동일한 소속정도를 가지므로 서로 충돌하는 문제가 발생된다. Ishibuchi[12]와 Mansoori[13]는 이러한 문제를 최소화하기 위해서 각 입력벡터에 대해서 단지 하나의 승자규칙(single winner rule)만을 선택하는 방법을 사용하였다. 그러나 이러한 방법은 충돌된 규칙을 고려하지 않기 때문에 규칙공간의 홀(hole) 영역을 최소화하기 위해서 규칙의 가중치를 고려하거나 퍼지 소속함수를 지속적으로 조정(tuning)하여야만 하고, 최악의 경우 입력벡터 x_i로부터 얻은 모든 규칙패턴들이 서로 충돌될 때 승자규칙을 얻을 수 없는 제약점을 가진다. 제안된 모델에서는 이러한 문제점을 보다 단순한 방법으로 개선시키기 위해서, 각 입력벡터로부터 얻은 승자 후보규칙의 빈도수(frequency)와 식 (4)에 의해서 계산된 승자 후보규칙의 최대 적합도를 이용하였다.

표 1에서 $\max(R_j^* \Rightarrow c_p)$와 $\max(R_j^* \Rightarrow c_q)$은 식 (4)로부터 계산된 2개의 후보규칙 c_p와 c_q의 최대 적합도를, $fre(R_j^* \Rightarrow c_p)$와

표 1. 충돌된 후보규칙들을 분해하기 위한 2가지 기준
Table 1. Two criteria to resolve the conflicted candidate rules

Freq[b] \ MCG[a]	$\max(R_j^* \Rightarrow c_p) > \max(R_j^* \Rightarrow c_q)$	$\max(R_j^* \Rightarrow c_p) = \max(R_j^* \Rightarrow c_q)$	$\max(R_j^* \Rightarrow c_p) < \max(R_j^* \Rightarrow c_q)$
$fre(R_j^* \Rightarrow c_p) > fre(R_j^* \Rightarrow c_q)$	$R_j^* \Rightarrow c_p$	$R_j^* \Rightarrow c_p$	$R_j^* \Rightarrow c_p$
$fre(R_j^* \Rightarrow c_p) = fre(R_j^* \Rightarrow c_q)$	$R_j^* \Rightarrow c_p$	$R_j^* \Rightarrow c_p$ and $R_j^* \Rightarrow c_q$	$R_j^* \Rightarrow c_q$
$fre(R_j^* \Rightarrow c_p) < fre(R_j^* \Rightarrow c_q)$	$R_j^* \Rightarrow c_q$	$R_j^* \Rightarrow c_q$	$R_j^* \Rightarrow c_q$

[a] MCG(Maximum Compatibility Grade): 생성된 후보규칙의 최대 적합도.
[b] Freq(Frequency): 후보규칙의 발생 빈도수.

$fre(R_j^* \Rightarrow c_q)$는 후보규칙 c_p와 c_q에 대응된 빈도수를 나타낸다. 표 1에서 보여주는 것처럼, 제안된 모델에서는 발생 빈도가 높은 후보규칙이 다음에 발생할 가능성이 높다는 국부성(locality) 원리의 가정 하에 후보규칙의 빈도수에 보다 큰 가중치를 부여하였고, 최대 적합도와 빈도수가 동일한 충돌규칙들이 발생할 경우에는 규칙공간 내에서 홀 영역을 최소화하기 위해 이들 규칙패턴들을 대응하는 모든 클래스의 후보규칙으로 할당하였다.

Ⅲ. 실험 및 결과

실험에서는 D광역시에 소재한 D의료원에 2006년 7월에서 2007년 6월 사이에 호흡곤란을 주호소로 응급실에 내원한 환자 1,129명의 의무기록을 대상으로 하였다. 대상자의 인적사항을 제외한 등록번호, 성별, 나이, 응급실 내원일자 및 시간, 진료결과, 입원 시 진단, 초기 검사 항목 등의 자료를 데이터웨어하우스(data warehouse)에서 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 타 병원으로 전원된 환자, DOA(death on arrival), CPR(Cardio-

표 2. 데이터 셋의 특성
Table 2. Dataset's feature

#	Variable[a]	Unit	Range[b]	Mean±SD
1	WBC	$\times 10^3$/uL	[0, 345.3]	11.389±13.547
2	PLT	$\times 10^3$/uL	[0, 1,105]	270.388±123.267
3	Cl^-	mmol/L	[72, 134]	104.395±6.926
4	AST	U/L	[0, 3,654]	71.803±239.088
5	ALT	U/L	[0, 2,481]	43.354±129.436
6	pCO_2	mmHg	[8.3, 98.5]	39.603±13.560
7	pO_2	mmHg	[35.9, 354]	80.967±23.316
8	O_2SAT	%	[55.9, 99.9]	96.015±3.612
9	LDH	U/L	[0, 8,178]	648.310±524.888
10	Ca^{2+}	mEq/L	[0, 3.23]	2.049±0.656
11	Mg^{2+}	mg/dL	[0, 5.2]	1.944±0.800

[a] WBC: White blood cell; PLT: Platelet count; Cl^-: Chloride; AST: Aspartate transaminase; ALT: Alanine transaminase; pCO_2: Pressure of carbon dioxide; pO_2: Pressure of oxygen; O_2SAT: Oxygen saturation; LDH: Lactate de-hydrogenase; Ca^{2+}: Calcium; Mg^{2+}: Magnesium
[b] $[\alpha, \beta]$: 각 속성의 데이터 셋의 범위 (여기서 α는 최소값, β는 최대값을 의미).

표 3. 입원/퇴원 환자의 통계적인 정보
Table 3. Statistical information for admission/discharge patients

Variable	Admission			Discharge		
	Min	Max	Mean±SD	Min	Max	Mean±SD
WBC	0	345.3	12.056±15.325	2.77	53.47	9.376±4.857
PLT	0	1,105	267.708±128.988	52	676	278.481±103.968
Cl^-	72	134	104.028±7.419	90	124	105.504±5.015
AST	5	3,654	82.599±273.365	0	645	39.209±53.057
ALT	3	2,481	48.159±148.174	0	212	28.847±28.231
pCO_2	8.3	98.5	39.439±13.197	8.9	96	40.095±14.622
pO_2	35.9	354	81.459±24.859	39.4	160.6	79.481±17.836
O2SAT	55.9	99.9	95.938±3.848	75.2	99.7	96.248±2.777
LDH	0	8,178	644.2±508.632	0	5,625	660.658±572.284
Ca^{2+}	0	3.23	2.052±0.655	0	2.93	2.041±0.660
Mg^{2+}	0	5.2	1.945±0.822	0	3.3	1.940±0.731

Pulmonary Resuscitation) 후 혹은 DNR(Do Not Resuscitate)로 사망한 환자, 자의 퇴원 혹은 미상의 기타 환자, 의무기록이 불완전한 경우를 제외한 총 844명의 환자(입원환자 634명, 퇴원환자 210명)에 대한 데이터 셋(55개의 입력변수 중에서 임상전문가에 의해서 선택된 11개 변수)을 이용하였다. 또한 데이터 셋에 포함된 null값은 0의 정수 스케일 값으로 할당하였고(표 2 참조), 실험에서 사용된 모든 알고리즘은 Matlab 2008에서 구현되었다.

A. 표준 퍼지분류기 vs. 규칙가중치를 고려한 퍼지 분류모델

실험 1에서는 규칙가중치를 고려하지 않는 퍼지분류기[17](이하 표준 퍼지분류기)와 제안된 퍼지 분류모델의 성능을 비교하기 위해서 844명의 호흡곤란 환자 데이터 셋 전체를 훈련-실험 데이터로 사용하였을 때의 분류 정확성과 규칙의 수의 변화를 비교하였다. 실험에 앞서 각 입력변수의 퍼지 분할구간을 결정하기 위해서 입·퇴원 유무에 따라 입력변수 값들의 공간적인 분포와 중첩구간을 분석하였고(표 3과 4 참조), 그림 3에서 보여주는 것처럼 추가적인 입력 데이터들의 공간을 확보하기 위해서 표 2의 'Ranges'를 표 4의 'Modified ranges'(즉 [Min-SD, Max+SD])로 변환하였다. 또한 퍼지분할의 수는 표 4의 'Overlapping regions'를 기준으로 출력 클래스의 수와 동일하게 나누었고, 퍼지 소속함수는 식 (1)을 이용하여 그림 4와 같이 생성하였고, 식 (2)-(4)와 표 1의 충돌된 규칙의 분해기준을 근거로 844명의 호흡곤란 환자의 규칙패턴을 추출하였다.

표 5에서 보여주는 것처럼, 추출된 규칙패턴에서 입원환자 634명의 경우 78개, 퇴원환자 210명의 경우 9개의 규칙패턴들을 생성할 수 있었고, 표준 퍼지분류기의 분류 정확도는 75.8294%, 제안된 규칙가중치를 고려한 퍼지 분류모델은 75.5924%임을 볼 수 있었다(표 6 참조). 표 5에서 Low와 High는 퍼지 소속함수의 언어적인 값들을, CF는 식 (4)로부터 계산된 각 규칙의 확신도를, Freq.는 각 규칙에 대응된 발생 빈도수를 의미한다.

표 4. 퍼지분할 수와 중첩영역
Table 4. Number of fuzzy partitions and their overlapped regions for each variable

Variable	Modified range	Overlapped region	Num. fuzzy partition
WBC	[-13.547, 358.847]	[2.77, 53.47]	
PLT	[-123.267, 1,228.267]	[52, 676]	
Cl^-	[65.073, 140.926]	[90, 124]	
AST	[-239.088, 3,893.088]	[5, 645]	
ALT	[-129.436, 2,610,436]	[3, 212]	
pCO_2	[-5.260, 112.060]	[8.9, 96]	2
pO_2	[12.583, 377.316]	[39.4, 160.6]	
O2SAT	[52.287, 103.512]	[75.2, 99.7]	
LDH	[-524.888, 8,702.889]	[0, 5,625]	
Ca^{2+}	[-0.656, 3.886]	[0, 2.93]	
Mg^{2+}	[-0.800, 6.000]	[0, 3.3]	

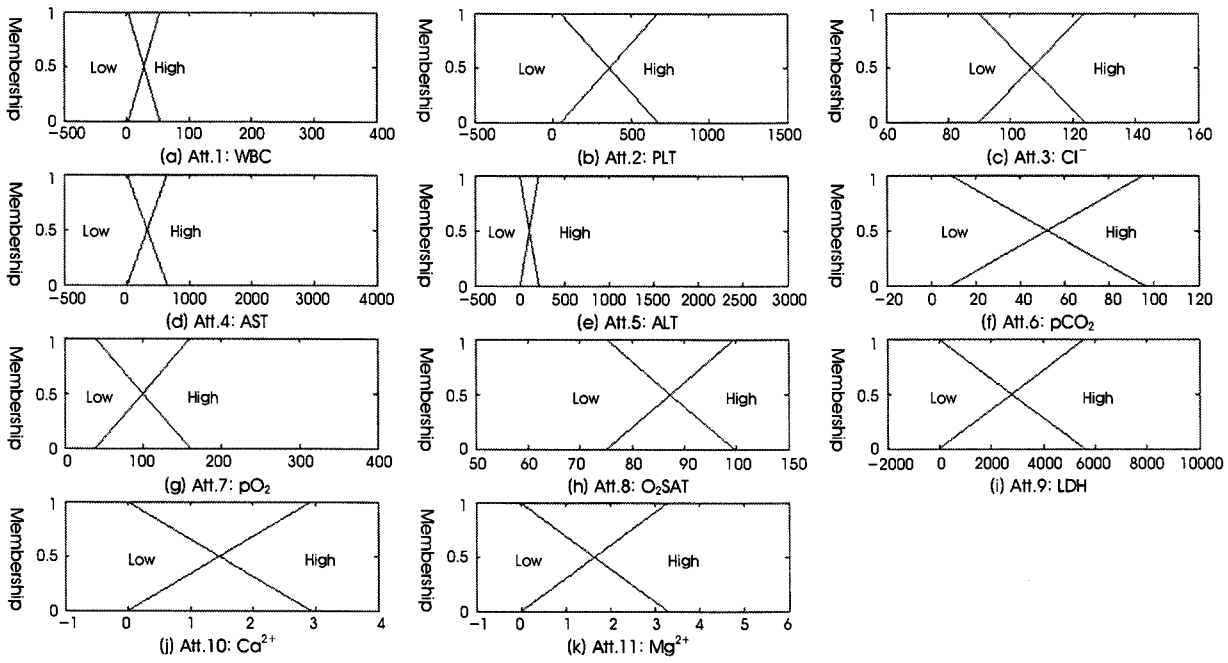

그림 4. 입력부 퍼지 소속함수

Fig. 4. Fuzzy membership functions of antecedent part for dyspnea patients

표 5. 생성된 퍼지규칙과 규칙의 확신도

Table 5. Generated fuzzy rules and their certainty factors

#	Antecedent variable											Class	CF[a]	Freq.[b]
	WBC	PLT	Cl^-	AST	ALT	pCO_2	pO_2	O2SAT	LDH	Ca^{2+}	Mg^{2+}			
1	Low	Low	Low	Low	Low	Low	Low	Low	Low	High	Low		1	1
2	Low	Low	Low	Low	Low	Low	Low	Low	Low	High	High		0.1142	1
3	Low	Low	Low	Low	Low	Low	Low	High	Low	Low	Low		0.1458	8
4	Low	Low	Low	Low	Low	Low	Low	High	Low	Low	High		0.1389	2
5	Low	Low	Low	Low	Low	Low	Low	High	Low	High	Low		0.0474	2
6	**Low**	**Low**	**Low**	**Low**	**Low**	**Low**	**Low**	**High**	**Low**	**High**	**High**		**0.1956**	**157**
7	Low	Low	Low	Low	Low	Low	High	Low	Low	Low	High		0.0502	1
8	Low	Low	Low	Low	Low	Low	High	High	Low	Low	Low		0.2127	4
9	Low	Low	Low	Low	Low	Low	High	High	Low	Low	High		0.0821	3
10	Low	Low	Low	Low	Low	Low	High	High	Low	High	Low		0.1242	2
11	Low	Low	Low	Low	Low	Low	High	High	Low	High	High		0.1329	11
12	Low	Low	Low	Low	Low	High	Low	Low	Low	High	Low		0.0327	2
13	Low	Low	Low	Low	Low	High	Low	Low	Low	High	High		0.0749	1
14	Low	Low	Low	Low	Low	High	Low	High	Low	Low	High	Admission	0.0568	1
15	Low	Low	Low	Low	Low	High	Low	High	Low	High	Low		0.0965	5
16	Low	Low	Low	Low	Low	High	Low	High	Low	High	High		0.1096	32
17	Low	Low	Low	Low	Low	High	Low	High	High	High	High		0.0249	1
18	Low	Low	Low	Low	Low	High	High	High	Low	High	High		0.1157	1
19	Low	Low	Low	Low	High	Low	Low	High	Low	High	Low		0.0646	3
20	Low	Low	Low	Low	High	Low	Low	High	Low	High	High		0.0709	4
21	Low	Low	Low	Low	High	Low	High	High	Low	Low	High		0.0571	1
22	Low	Low	Low	High	Low	Low	Low	High	Low	High	Low		0.1142	1
23	Low	Low	Low	High	High	Low	Low	High	Low	High	High		0.1458	8
24	Low	Low	Low	High	High	Low	High	High	Low	High	High		0.1389	2
25	Low	Low	Low	High	High	High	Low	High	Low	High	High		0.0474	2
26	Low	Low	High	Low	Low	Low	Low	Low	Low	High	Low		0.0669	2
27	Low	Low	High	Low	Low	Low	Low	Low	Low	High	High		0.0590	2

#	Antecedent variable											Class	CF[a]	Freq.[b]
	WBC	PLT	Cl^-	AST	ALT	pCO_2	pO_2	O2SAT	LDH	Ca^{2+}	Mg^{2+}			
28	Low	Low	High	Low	Low	Low	Low	High	Low	Low	Low		0.0977	2
29	Low	Low	High	Low	Low	Low	Low	High	Low	Low	High		0.1248	12
30	Low	Low	High	Low	Low	Low	Low	High	Low	High	Low		0.2232	23
31	**Low**	**Low**	**High**	**Low**	**Low**	**Low**	**Low**	**High**	**Low**	**High**	**High**		**0.1979**	**118**
32	Low	Low	High	Low	Low	Low	Low	High	High	High	Low		0.0358	1
33	Low	Low	High	Low	Low	Low	High	High	Low	Low	Low		0.1528	1
34	Low	Low	High	Low	Low	Low	High	High	Low	Low	High		0.2010	2
35	Low	Low	High	Low	Low	Low	High	High	Low	High	Low		0.0628	1
36	Low	Low	High	Low	Low	Low	High	High	Low	High	High		0.1366	15
37	Low	Low	High	Low	Low	High	Low	High	Low	Low	High		0.0651	2
38	Low	Low	High	Low	Low	High	Low	High	Low	High	Low		0.1291	3
39	Low	Low	High	Low	Low	High	Low	High	Low	High	High		0.0822	18
40	Low	Low	High	Low	High	Low	Low	Low	Low	High	High		0.0205	1
41	Low	Low	High	Low	High	Low	Low	High	Low	High	High		0.0402	3
42	Low	Low	High	High	Low	Low	Low	Low	Low	High	High		0.0477	1
43	Low	Low	High	High	Low	Low	Low	High	Low	High	High		0.0825	1
44	Low	Low	High	High	High	Low	Low	Low	Low	High	Low		0.0593	2
45	Low	Low	High	High	High	Low	Low	High	Low	High	High		0.1028	4
46	Low	Low	High	High	High	High	Low	High	Low	High	High	Admission	0.1356	1
47	Low	High	Low	Low	Low	Low	Low	Low	Low	High	High		0.0423	1
48	Low	High	Low	Low	Low	Low	Low	High	Low	Low	High		0.1159	4
49	Low	High	Low	Low	Low	Low	Low	High	Low	High	Low		0.0870	3
50	Low	High	Low	Low	Low	Low	Low	High	Low	High	High		0.1548	40
51	Low	High	Low	Low	Low	Low	Low	High	High	High	Low		0.0553	1
52	Low	High	Low	Low	Low	Low	High	High	Low	Low	High		0.1082	3
53	Low	High	Low	Low	Low	Low	High	High	Low	High	Low		0.0819	1
54	Low	High	Low	Low	Low	Low	High	High	Low	High	High		0.1661	6
55	Low	High	Low	Low	Low	High	Low	High	Low	Low	High		0.1131	1
56	Low	High	Low	Low	Low	High	Low	High	Low	High	Low		0.0893	1
57	Low	High	Low	Low	Low	High	Low	High	Low	High	High		0.0715	7
58	Low	High	Low	Low	Low	High	High	High	Low	High	High		0.0534	1
59	Low	High	Low	Low	High	High	Low	High	Low	High	High		0.0723	2
60	Low	High	High	Low	Low	Low	Low	High	Low	Low	High		0.0740	2
61	Low	High	High	Low	Low	Low	Low	High	Low	High	Low		0.0804	3
62	Low	High	High	Low	Low	Low	Low	High	Low	High	High		0.0885	22
63	Low	High	High	Low	Low	Low	High	Low	Low	High	High		0.0826	1
64	Low	High	High	Low	Low	Low	High	High	Low	Low	High		0.1425	2
65	Low	High	High	Low	Low	High	Low	High	Low	High	Low		0.0510	2
66	High	Low	Low	Low	Low	Low	Low	High	Low	High	Low		0.1628	1
67	High	Low	Low	Low	Low	Low	Low	High	Low	High	High		0.0395	2
68	High	Low	Low	Low	Low	High	Low	Low	Low	High	Low		0.0846	1
69	High	Low	Low	Low	Low	High	Low	High	Low	High	High		0.0565	1
70	High	Low	High	Low	Low	Low	Low	High	Low	High	High		0.0797	3
71	High	Low	High	High	Low	Low	Low	High	Low	High	High		0.0564	1
72	High	High	Low	Low	Low	Low	Low	High	Low	Low	High		0.1076	1
73	High	High	Low	Low	Low	Low	Low	High	Low	High	Low		0.1433	1
74	High	High	Low	Low	Low	Low	Low	High	Low	High	High		0.0780	3
75	High	High	Low	Low	Low	Low	High	High	Low	High	High		0.0912	1
76	High	High	Low	Low	Low	High	Low	High	Low	High	Low		0.0687	1
77	High	High	Low	High	High	Low	Low	High	Low	High	High		0.0467	1
78	High	High	High	Low	Low	Low	Low	High	Low	Low	High		0.0916	1
79	Low	Low	Low	Low	Low	Low	Low	High	High	High	High		0.0774	1
80	Low	Low	Low	Low	Low	High	Low	High	High	High	Low		0.1167	1
81	Low	Low	Low	Low	High	Low	Low	High	Low	Low	High		0.0309	1
82	Low	High	Low	Low	Low	Low	Low	Low	Low	High	Low	Discharge	0.0398	1
83	Low	High	Low	Low	Low	Low	Low	High	Low	Low	Low		0.1577	1
84	Low	High	Low	Low	Low	Low	High	High	Low	Low	Low		0.0812	1
85	Low	High	High	Low	Low	High	Low	High	Low	High	High		0.0763	2
86	Low	High	High	Low	High	Low	Low	High	Low	Low	High		0.0288	1
87	Low	High	High	Low	High	Low	Low	High	Low	High	High		0.0356	1

[a] CF(Certainty factor): 규칙의 확신도
[b] Freq(Frequency): 발생 빈도수

표 6. 표준 퍼지분류기와 제안된 모델과의 비교결과
Table 6. Comparison results between standard fuzzy classifier and proposed model

Method	Num. initial candidate rule		Num. fuzzy rules after removing redundant rule		Num. final fuzzy rule		Classification accuracy (%)		
	C_1[a]	C_2[b]	C_1	C_2	C_1	C_2	C_1	C_2	Total
Standard fuzzy classifier	634	210	81	37	78	9	99.369	4.762	75.829
Proposed model							99.843	2.381	75.592

[a] C_1: 입원환자에 대한 레이블
[b] C_2: 퇴원환자에 대한 레이블

표 6에서 'Num. initial candidate rule'은 식 (2)-(3)으로부터 생성된 클래스 레이블별 후보 규칙의 수를, 'Num. rules after removing redundant rule'은 중복된 규칙패턴들을 제거한 후 생성된 퍼지 규칙의 수를 의미한다. 또한 'Num. final fuzzy rule'은 식 (4)와 표 1의 충돌규칙의 분해기준을 근거로 생성된 최종 퍼지 규칙의 수를 나타내고, 'Classification accuracy'는 입·퇴원 환자 각각의 분류 정확성과 전체 분류 정확성을 보여준다. 그 결과 제안된 퍼지모델이 표준 퍼지분류기와 거의 유사한 성능을 제공함을 볼 수 있었고, 입원환자의 경우 생성된 규칙패턴들 중에서 6, 31번째 규칙패턴에서 가장 많은 발생빈도를 나타내었고, 퇴원환자의 경우에는 출현빈도가 높은 규칙패턴이 존재하지 않음을 볼 수 있었다(표 5 참조). 이러한 결과는 전체 데이터 셋에서 입원환자의 비율이 퇴원환자의 비율에 비해 상대적으로 높은 분포를 가지기 때문이다.

Rule 6: If WBC is Low(<=53.47) and PLT is Low(<=676) and Cl^- is **Low**(<=124) and AST is Low(645) and ALT is Low(<=212) and pCO_2 is Low(<=96) and pO_2 is Low(<=160.6) and O_2SAT is High(>=75.2) and LDH is Low(<=5,625) and Ca^{2+} is High(>=0) and Mg^{2+} is High(>=0) Then Class is Admission with CF is 0.1956,

Rule 31: If WBC is Low(<=53.47) and PLT is Low(<=676) and Cl^- is **High**(>=90) and AST is Low(645) and ALT is Low(<=212) and pCO_2 is Low(<=96) and pO_2 is Low(<=160.6) and O_2SAT is High(>=75.2) and LDH is Low(<=5,625) and Ca^{2+} is High(>=0) and Mg^{2+} is High(>=0) Then Class is Admission with CF is 0.1979.

여기서 각 규칙의 괄호안의 값들은 그림 4에서 나타낸 언어적인 값들의 하한과 상한 경계값(lower and upper limits), 즉 cut-off 값을 나타낸다.

실험 2에서는 표준 퍼지분류기와 제안된 모델과의 보다 객관적인 비교를 위해서 10-fold 교차검증의 결과를 비교하였다. 각 fold의 퍼지 분할구간은 랜덤으로 추출된 훈련데이터 셋에서 중첩구간을 추출하여 생성하였다. 또한 중첩구간 내에서 퍼지분할 개수의 영향을 살펴보기 위해서 퍼지분할의 수를 2개에서 6개까지 조정하여 퍼지규칙의 수와 분류 정확도를 비교하였다. 표 7의 결과에서 보여주듯이, 훈련데이터의 평균 분류정확도는 퍼지분할의 수가 증가할수록 제안된 모델과 표준 퍼지분류기의 성능이 모두 증가됨을 볼 수 있었으나, 실험 데이터 셋에서는 점차적으로 감소하는 경향을 보였다. 게다가 훈련 데이터 셋의 분류 정확성에서는 제안된 모델이 표준 퍼지분류기의 성능에 비해 다소 떨어졌으나 실험 데이터 셋에서는 개선된 성능(즉 $p=2$일 때 1.4215%, $p=3$일 때 2.0210%, $p=4$일 때 1.7829%, $p=6$일 때 1.1807%)을 보였다. 또한 규칙 수의 변화에서는 표준 퍼지분류기와 제안된 모델에서 모두 동일한 규칙 수의 분포를 보였다.

표 7. 표준 퍼지분류기와 제안된 모델과의 비교 결과 (10-fold 교차검증)
Table 7. Comparison results between standard fuzzy classifier and proposed model (10-fold cross validation)

Method / Partition	Standard fuzzy classifier[12]			Proposed model		
	Avg. rule	Train acc. (%)	Test acc. (%)	Avg. rule	Train acc. (%)	Test acc. (%)
$p=2$	88.7	76.132	**72.156**	88.7	75.579	**73.577**
$p=3$	335.3	81.977	66.585	335.3	75.856	68.606
$p=4$	456.2	86.782	61.833	456.2	79.239	63.616
$p=5$	642	94.329	54.015	642	83.478	52.479
$p=6$	727.7	98.275	50.462	727.7	89.336	51.643

표 8. 4가지 분류기의 실험 결과 (10-fold 교차검증)
Table 8. Experimental results for four types of conventional classifier (10-fold cross validation)

Conventional method	Experimental parameter	Test accuracy (%)
C4.5[13]	Confidence: 0.3 Instances per leaf: 2	73.454
QDA[14]	Default	42.647
kNN[15]	k: 2 Distance measure: Euclidean	71.450
SVM[16]	Polynomial function	24.706
	Sigmoid function	75.118

B. 기존의 패턴분류방법들과의 비교

제안된 모델과 기존의 패턴분류방법들과의 분류성능을 비교하기 위해서 다음과 같은 3가지 형태의 분류기와의 10-fold 교차검증을 실험하였다: i) 의사결정트리 C4.5[18], ii) 통계적 분류기 QDA(quadratic discriminant analysis)[19]와 kNN(k-nearest neighbor)[20], iii) SVM (support vector machine)[21].

각 분류기의 분류성능을 평가하기 위해서 사용된 실험 매개변수는 C4.5의 경우 리프노드 당 인스턴스의 수를 2개로, 신뢰도(confidence)는 0.3으로 설정하였고, kNN의 경우 k의 값은 2로, 거리척도 방법으로는 유클리디안 척도법을 사용하였다. 또한 QDA의 경우 디폴트 파라미터를 이용하였고, SVM의 경우에는 2가지 커널함수 (polynomial과 sigmoid functions)를 이용하여 평가하였다. 표 7과 8의 10-fold 교차검증의 결과로부터 sigmoid 커널함수를 기반으로 한 SVM이 가장 좋은 분류 정확성을 보였고, 그 다음으로 제안된 모델($p = 2$), C4.5, 표준 퍼지 분류기 ($p = 2$), kNN, QDA, polynomial 커널함수를 기반으로 한 SVM 순으로 나타났다. 하지만 SVM의 경우 선형 또는 비선형변환과정을 통해 주어진 데이터 셋을 최적으로 분류할 수 있는 초평면(hyperplane)과 마진(margin)을 결정하는데 사용되는 여러 매개변수들이 있으며, 데이터에 의존적으로 이들의 최적 값을 찾기란 매우 어려운 문제이다. 또한 실험결과에서 보여주듯이, 분류성능은 사용된 커널함수의 형태에 따라 직접적인 영향을 받기 때문에 임상에서 받아들여질 가능성이 낮다고 볼 수 있다. 따라서 이러한 점으로 미루어 볼 때 제안된 퍼지 분류모델이 비교된 여러 분류기들보다 좋은 결과를 제공함을 볼 수 있다.

Ⅳ. 결론 및 토의

본 연구에서는 응급실에 호흡곤란으로 내원한 환자들의 입·퇴원에 관한 패턴분석을 위해서 규칙가중치를 기반으로 한 퍼지 분류모델을 제안하였다. 제안된 모델에서는 입·퇴원 환자의 패턴을 분석하기 위해서 주어진 데이터 셋으로부터 퍼지 분할영역과 소속함수를 자동으로 생성하고 후보규칙들을 생성하는 방법을 제시하였고, 초기에 생성된 후보규칙들 간의 충돌된 규칙들을 분해하기 위한 방법도 논의하였다.

실험에서는 제안된 모델의 효과성 및 타당성을 검증하기 위해서 2가지 실험을 하였다: i) 규칙가중치를 고려하지 않은 표준 퍼지 분류기[17]와의 비교실험, ii) 의사결정트리 C4.5[18], 통계적 분류기 QDA[19]와 kNN[20], SVM[21]과의 비교실험. 실험1에서는 844명 호흡곤란환자의 전체 데이터 셋을 훈련-실험 데이터로 사용하였을 때의 규칙패턴과 분류성능을 표준 퍼지분류기와 비교하였고, 제안된 분류모델이 표준 퍼지 분류기에 비해 0.237%정도 저하됨을 볼 수 있었다. 보다 객관적인 검증을 위해서 실험2에서는 전체 데이터 셋을 9대 1비율로 나누어 10-fold 교차검증을 수행하였으며, 제안된 모델($p = 2$)이 표준 퍼지분류기($p = 2$), C4.5, QDA, kNN, polynomial 커널함수를 기반으로 한 SVM에 비해 각각 1.421%, 0.123%, 30.93%, 2.127%, 48.871% 향상됨을 볼 수 있었고, sigmoid 커널함수를 기반으로 한 SVM에 비해 1.541%정도 저하된 분류성능을 보여주었다. 하지만 SVM은 사용된 커널함수의 형태에 따라 현저한 성능의 변화를 보였고, 최적의 마진을 결정하기 위해서 고려되어야하는 매개변수 값들의 결정이 중요한 요인으로 작용하였다. 따라서 이러한 점으로 볼 때 제안된 방법은 추가적인 매개변수를 고려하지 않고 단순히 각 속성의 중첩영역을 추출하고, 추출된 영역으로부터 규칙을 생성한 뒤 충돌된 규칙을 분해함으로써 분류성능을 개선시킬 수 있기 때문에 임상의사결정지원(CDS: clinical decision support)을 위한 도구로 사용될 수 있을 뿐만 아니라 임상검사 항목 (즉 혈액, 요검사 등)들로부터 지식을 추출하거나 분석할 수 있는 도구로 사용될 수 있을 것으로 판단된다.

향후 연구에서는 제안된 모델을 보다 신뢰성 있는 모델로 확장하기 위해서 다음과 같은 몇 가지 연구 i) 이산 혹은 연속형 형태의 데이터 셋에서 최적의 퍼지분할 개수의 선택과 분할된 공간의 병합(merging)에 관한 방법, ii) 입력변수의 증가에 따른 차원의 저주(curse of dimensionality)문제를 해결하기 위한 특징선택(feature selection) 방법과 기존의 neuro-fuzzy, genetic-fuzzy등과 같은 방법들과의 비교, iii) 임상전문가의 견해와 모델의 출력결과와의 차이(즉 생성된 퍼지 규칙에서 얻은 언어적인 값들이 나타내는 cutoff 값들의 신뢰성)를 최소화할 수 있는 방법에 관한 연구가 필요할 것으로 사료된다.

참고문헌

[1] G. Scano and N. Ambrosino, "Pathophysiology of dyspnea," *Lug*, vol. 180, p.131, 2002.

[2] P. Jevon and B. Ewens, "Assessment of a breathless patient," *Nursing*, vol. 15, no. 16, pp.48-55, 2001.

[3] M.B. Parshall, "Psycometric characteristics of dyspnea descriptor ratings in emergency department patients with exacerbation chronic obstruction pulmonary disease," *Research in Nursing & Health*, vol. 25, pp.331-344, 2002.

[4] C.G. Yu, "Differential diagnosis of dyspnea, Tuberculosis and Respiratory Diseases", vol. 55, no. 1, pp.5-14, 2003.

[5] T.H. Kim, "Differential diagnosis and treatment of dyspnea," *Korean J. Med.*, vol. 76, no. 4, pp.425-430, 2009.

[6] L. Zadeh, "Fuzzy sets," *Inform. and Contr.*, vol. 8, pp.338-353, 1965.

[7] E.J. Cha, T.S. Lee, Y.S. Whang, J.W. Kim, S.O. Yang, K.H. Jung, and H.K. Ryu, "Automated clinical test result analysis system-Application to liver function test," *J. Biomed. Eng. Res.*, vol. 14, no. 4, pp.341-348, 1993.

[8] K.R. Jun, S.J. Lee, B.C. Choi, S.H. An, K. Ha, J.Y. Kim, and J.H. Kim, "A study on the development of urine analysis system using strip and evaluation of experimental result by means of fuzzy inference," *J. Biomed. Eng. Res.*, vol. 19, no. 5, pp.477-486, 1998.

[9] J.H. Hwang, K.S. Park, and B.G. Min, "A study on the detection of pulmonary blood vessel using pyramid images and fuzzy theory," *J. Biomed. Eng. Res.*, vol. 12, no. 2, pp.99-105, 1991.

[10] O.K. Yoon, H.S. Kim, D.M. Kwak, B.S. Kim, D.W. Kim, W.M. Pyun, and K.H. Park, "Segmentation of multispectral MRI using fuzzy clustering," *J. Biomed. Eng. Res.*, vol. 21, no. 4, pp.333-338, 2000.

[11] U.C. Yoon, J.W. Hwang, J.S. Kim, J.J. Kim, I.Y. Kim, J.S. Kwon, and S.I. Kim, "Successive fuzzy classification and improved parcellation method for brain analysis," *J. Biomed. Eng. Res.*, vol. 22, no. 5, pp.377-383, 2001.

[12] H. Ishibuchi and T. Nakashima, "Effect of rule weights in fuzzy rule-based classification systems," *IEEE Trans. Fuzzy Syst.*, vol. 9, no. 4, pp.506-515, 2001.

[13] E.G. Mansoori, M.J. Zolghadri, and S.D. Katebi, "A weighting function for improving fuzzy classification systems performance," *Fuzzy Sets and Syst.*, vol. 158, no. 5, pp.583-591, 2007.

[14] H. Ishibuchi, T. Nakashima, and T. Murata, "Three-objective genetics-based machine learning for linguistic rule extraction," *Inform. Sci.*, vol. 136, no. 1-4, pp.109-133, 2001.

[15] Y.C. Tsai, C.H. Cheng, and J.R. Chang, "Entropy-based fuzzy rough classification approach for extracting classification rules," *Experts Syst. with Appl.*, vol. 31, no. 2, pp.436-443, 2006.

[16] Y. Chen, B. Yang, A. Abraham, and L. Peng, "Automatic design of hierarchical takagi-sugeno type fuzzy systems using evolutionary algorithms," *IEEE Trans. Fuzzy Syst.*, vol. 15, no. 3, pp.385-397, 2007.

[17] C.S. Son, H.M. Chung, and S.H. Kwon, "Selection method of fuzzy partitions in fuzzy rule-based classification systems," *J. Korean Institute of Intelligent Syst.*, vol. 18, no. 3, pp.360-366, 2008.

[18] J.R. Quinlan, *C4.5: Programs for machine learning*, Elsevier Science Ltd, 1992.

[19] J.H. Friedman, "Regularized discriminant analysis," *J. American Statistical Association*, vol. 84, no. 405, pp.165-175, 1989.

[20] T.M. Cover and P.E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inform. Theory*, vol. 13, no. 1, pp.21-27, 1967.

[21] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learn.*, vol. 20, no. 3, pp.273-297, 1995.