데이터의 공간적 분포를 고려한 퍼지 이산화와 특징선택에의 응용

Fuzzy discretization with spatial distribution of data and Its application to feature selection

손창식* · 신아미* · 이인희* · 박희준** · 박형섭*** · 김윤년***

Chang-Sik Son, A-Mi Shin, In-Hee Lee, Hee-Joon Park, Hyoungh-Seob Park and
Yoon-Nyun Kim

* 계명대학교 의과대학 의료정보학교실

** 계명대학교 의과대학 의용공학과

*** 계명대학교 의과대학 내과학(심장내과) 교실

요 약

임상 데이터마이닝에서 최적의 특징 집합을 선택하는 것은 주어진 데이터로부터 생성된 모델의 복잡성을 줄일 뿐만 아니라 유용성을 향상시키는 데에 매우 중요하고, 선택된 특징들의 임계값은 질병의 감별진단을 위해 임상 전문가의 결정기준으로 사용된다. 본 논문에서는 데이터의 공간적인 분포, 즉 중첩영역에서 중복 속성값을 포함하는 데이터의 분리성 정도를 평가함으로써 연속형 속성을 가진 데이터에 대한 퍼지 이산화기법을 제안한다. 제안된 방법에서 중복 속성값의 가중치 평균값은 각 특징의 임계값(즉 경계값)을 결정하기 위해서 사용되었고, 러프집합은 전체 특징들 중에서 중요특징들의 집합을 선택하기 위해서 이용하였다. 제안된 방법의 타당성을 검증하기 위해 호흡곤란을 주소로 내원한 668명의 환자 데이터를 근거로 3가지 이산화방법과 제안된 이산화방법에 대한 실험을 수행하였다. 실험결과, 퍼지분할을 기반으로 한 이산화방법이 하드분할을 기반으로 한 이산화방법에 비해서 평균 분류정확도와 G-mean 성능에서 보다 좋은 결과를 제공함을 확인하였다.

키워드 : 연속형 속성, 분리성, 이산화, 특징선택

Abstract

In clinical data minig, choosing the optimal subset of features is such important, not only to reduce the computational complexity but also to improve the usefulness of the model constructed from the given data. Moreover the threshold values (i.e., cut-off points) of selected features are used in a clinical decision criteria of experts for differential diagnosis of diseases. In this paper, we propose a fuzzy discretization approach, which is evaluated by measuring the degree of separation of redundant attribute values in overlapping region, based on spatial distribution of data with continuous attributes. The weighted average of the redundant attribute values is then used to determine the threshold value for each feature and rough set theory is utilized to select a subset of relevant features from the overall features. To verify the validity of the proposed method, we compared experimental results, which applied to classification problem using 668 patients with a chief complaint of dyspnea, based on three discretization methods (i.e., equal-width, equal-frequency, and entropy-based) and proposed discretization method. From the experimental results, we confirm that the discretization methods with fuzzy partition give better results in two evaluation measures, average classification accuracy and G-mean, than those with hard partition.

Key Words : Continuous attribute, Separation, Discretization, Feature selection

1. 서 론

접수일자 : 2009년 10월 21일

완료일자 : 2010년 3월 17일

† Corresponding author

본 연구는 지식경제부 지방기술혁신사업(RTI04-01-01)

지원으로 수행되었음.

임상 데이터베이스는 환자의 임상적 소견들에 관한 방대한 정보들이 축적되어 있고 수집된 데이터에는 정량적·정성적인 속성들이 혼합된 형태를 가지고 있으며, 중복 또는 누락된 속성 값 등의 불완전 요소가 포함되어 있다. 이러한 불완전하고 불명확한 임상 데이터로부터 노이즈와 결측치를 처리하고, 이산화(discretization)와 특징선택과 같은 데이터 변환과정을 통해 데이터에 포함된 유용한 특징을 선택할 수 있어야한다. 또한 마이닝된 지식이 최적화된 결과를 제공한다 할지라도 임상 전문가들의 의학적인 지식 (즉 관

런 특징으로 선택된 속성들과 선택된 속성들의 경계값)을 근거로 해석 가능하고 이해할 수 있어야만 한다[1]. 이것은 임상 의사결정지원시스템(clinical decision support system)과 임상 전문가시스템(clinical expert system)에서 지식을 획득하거나 발견하기 위해서 반드시 수행되어야 하는 데이터 정제(data cleaning)과정이라 볼 수 있다.

특히 여러 형태의 데이터 속성들 중에서 연속형 속성값을 가진 데이터의 이산화는 정보의 손실을 최소화하고 미지의 데이터에 대한 범주의 일반성을 최대화할 수 있는 이산화 구간을 찾는 것이 주된 목적이며[2], 3가지 이산화기법들이 보편적으로 사용되어지고 있다: I) Equal-width (distance) 분할; II) Equal-frequency (depth) 분할; III) Entropy-based 분할. 이들 3가지 기법들 중에서 equal-width와 equal-frequency 분할 기법[3]은 클래스 레이블에 대한 데이터의 분포를 고려하지 않기 때문에 비교사적인(unsupervised) 분할기법에 해당되고, 반면에 entropy based 분할기법[4]은 교사적인(supervised) 분할기법이라 할 수 있다. 하지만 equal-width 분할은 속성의 전체 도메인을 일정한 간격 ($d = (B - A)/k$, 여기서 A 와 B 는 해당 속성의 최소값과 최대값, k 는 분할 수)으로 분할하기 때문에 클래스 레이블에 대한 데이터의 분포를 잘 반영하지 못하는 제약점을 가지고, equal-frequency 분할은 이산화된 구간내의 속성값의 빈도수가 거의 동일한 지점을 기준으로 특징공간을 분할하기 때문에 동일한 속성값이 밀집되어 있는 경우에는 이산화된 구간이 특정 공간에서 세분화되는 형태를 가지며, entropy-based 분할은 초기에 생성된 분할구간을 병합하는 과정에서 추가적인 매개변수들을 필요로 하고 이들 매개변수에 의해 성능이 좌우된다. 또한 이들 방법 뿐만 아니라 다른 이산화 방법(data distribution-based[2], chi-square[5], clustering-based[6,7], hierarchical 분할[8])들은 구간분할의 최적화에 중점을 두고 있기 때문에 결과의 신뢰성에 중점을 둔 임상에서는 받아들여질 가능성이 낮다고 볼 수 있다.

따라서 본 논문에서는 이러한 점들을 고려하기 위해서 데이터의 공간적인 분포를 고려한 퍼지 이산화방법을 제안하고 러프집합에 기반을 둔 특징선택문제로 그 응용 가능성을 보이고자 한다. 본 논문의 구성은 다음과 같다. 2장에서는 특징선택을 위한 러프집합[9,10]의 기본 개념을 소개하고, 3장에서는 연속형 속성값을 가진 데이터에서 속성값들의 공간적인 분포를 고려한 퍼지 이산화방법과 러프집합에서 특징선택문제에 대해서 논의한다. 4장에서는 제안된 방법의 효과성과 신뢰성을 보이기 위해 A 광역시에 소재한 A 의료원에 호흡곤란으로 내원한 668명의 환자를 대상으로 한 실험결과를 비교하고, 결론 및 향후연구과제는 5장에서 논의한다.

2. 특징선택을 위한 러프집합

[정의1] $U = \{x_1, x_2, \dots, x_n\}$ 는 오브젝트들의 유한집합, $A = \{a_1, a_2, \dots, a_m\}$ 는 속성(특징)들의 유한집합이라고 할 때, 정보 시스템(information system)은 식 (1)과 같다.

$$IS = (U, A) \text{ or } IS = (U, C, D, V, f) \quad (1)$$

식 (1)에서 C 는 조건 속성(condition attribute), D 는 의사결정 속성(decision attribute)을 의미하고, 각 속성 $a \in A$

는 정보함수(information function) $f_a: U \rightarrow V_a$ 에 의해서 정의된다. 여기서 V_a 는 속성 a 의 도메인, 즉 속성 a 의 값들의 집합을 나타낸다. 예를 들어 표 1과 같은 결정 테이블이 있다고 가정할 때, 정보 시스템 IS는 $U = \{x_1, x_2, \dots, x_8\}$, $A = \{a, b, c, d\}$, $C = \{a, b, c\}$, $D = \{d\}$, $V_a = V_b = V_c = V_d = \{Low, Medium, High\}$ 로 나타낼 수 있다.

[정의 2] 속성들의 모든 집합 $P \subset A$ 에 대해서, 속성 P 의 식별불능 관계(indiscernibility relation) $IND(P)$ 와 동치류(equivalence class)는 식 (2), (3)와 같다.

$$IND(P) = \{(x, y) \in U^2 \mid \forall a \in P, a(x) = a(y)\} \quad (2)$$

식 (2)에서 $a(x)$ 는 오브젝트 x 에 대한 속성 a 의 값, 즉 속성값을 나타낸다.

$$U/IND(P) = \{U/IND(\{a\}) \mid a \in P\},$$

$$A \otimes B = \{X \cap Y \mid X \in A, Y \in B, X \cap Y \neq \emptyset\} \quad (3)$$

식 (3)에서 $\otimes$ 는 속성들의 부분집합(subset)에 대한 카티션 곱을 나타낸다. 표 1에서 $P = \{a, b\}$ 라고 가정할 때 $U/IND(P) = \{\{x_1, x_3, x_5\}, \{x_2, x_4, x_8\}, \{x_6\}, \{x_7\}\}$ 이고, 이것은 속성 집합 P 에서 오브젝트 집합 U 의 기본집합(elementary set)을 고려하는 것을 의미한다.

표 1. 결정 테이블

Table 1. A decision table

U/A	a	b	c	d
x_1	Medium	Low	High	Low
x_2	Low	Medium	Medium	High
x_3	Medium	Low	High	Low
x_4	Low	Medium	Medium	High
x_5	Medium	Low	High	Medium
x_6	High	High	Low	Medium
x_7	High	Medium	Medium	High
x_8	Low	Medium	Medium	Medium

[정의 3] 전체집합 U 에 대한 원소들의 부분집합을 X 라 할 때, 속성 공간 $P (P \subset A)$ 에서 X 의 하한(lower)과 상한 근사화(upper approximation)는 식 (4)와 같다.

$$\begin{aligned} PX_* &= \{x \mid U/IND(P) \subseteq X\}, \\ PX^* &= \{x \mid U/IND(P) \cap X \neq \emptyset\}, \end{aligned} \quad (4)$$

[정의 4] 전체집합 U 에서 식별불능 관계를 나타내는 속성들의 집합 P 와 Q 가 주어졌을 때, 긍정(positive), 부정(negative), 경계영역(boundary region)은 식 (5)와 같다.

$$\begin{aligned} POS_P(Q) &= \bigcup_{X \in U/IND(Q)} PX_*, \\ NEG_P(Q) &= U - \bigcup_{X \in U/IND(Q)} PX^*, \\ BND_P(Q) &= \bigcup_{X \in U/IND(Q)} PX^* - \bigcup_{X \in U/IND(Q)} PX_*, \end{aligned} \quad (5)$$

식 (5)에서 $POS_P(Q)$ 는 속성 집합 Q 의 동치류에 완전히 포함된 오브젝트들의 집합을 의미하고, $NEG_P(Q)$ 는 전체 집합에서 속성 집합 Q 의 동치류에 부분적으로 포함된 오브젝트들을 제외한 집합을 나타낸다. 만약 $P = \{a, b\}$ 이고

$Q = \{d\}$ 일 때 긍정영역, 부정영역, 경계영역은 식 (5)에 의해서 다음과 같이 나타낼 수 있다; $POS_P(Q) = \{x_6, x_7\}$, $NEG_P(Q) = \emptyset$, $BND_P(Q) = \{x_1, x_2, x_3, x_4, x_5, x_8\}$.

[정의 5] 조건부 속성 집합 P 와 의사결정부 속성 집합 Q 가 주어졌을 때, P 에 대한 Q 의 의존성 정도(dependency degree)와 속성 집합 P 에서 임의의 속성 $a \in P$ 의 중요성 정도(significance degree)는 식 (6), (7)과 같이 정의된다.

$$\gamma_P(Q) = \frac{|POS_P(Q)|}{|U|}, \gamma_P(Q) \in [0, 1] \quad (6)$$

$$\sigma_P(Q, a) = \gamma_P(Q) - \gamma_{P-\{a\}}(Q), \sigma_P(Q, a) \in [0, 1] \quad (7)$$

식 (6)에서 $| \cdot |$ 는 카디널리티(cardinality)를 의미하고, 의존성 정도가 1이면 Q 는 P 에 완전히 의존하고, 반면에 0이면 의존하지 않음을 나타낸다. 또한 식 (7)에서 속성 집합 P 에 대한 의존성과 임의의 속성 a 를 제거한 속성 집합의 의존성 정도가 같다면, 속성 a 는 불필요(dispensable)하다고 하고, 그렇지 않으면 필요 불가결하다고 한다. 따라서 식 (7)에서 $\sigma_P(Q, a) = 0$ 인 조건부 속성들의 집합을 리덕트 속성(reduct attribute)이라 하고, 이들 리덕트 속성들 간의 교집합인 속성들을 코어 속성(core attribute)이라 한다.

3. 연속형 속성의 이산화 방법

3.1 이산화 방법

만약 데이터 $\mathbf{X} = \{x_i | x_i = (x_{i1}, x_{i2}, \dots, x_{im}), i = 1, 2, \dots, m\}$ 가 m 개의 인스턴스들을 포함하는 n 차원 벡터이고, 출력 $C = \{c_q | q = 1, 2, \dots, Q\}$ 은 Q 개의 클래스로 이루어진 집합이라고 가정하자. 이때 속성값들의 공간적인 분포를 고려한 이산화 방법은 다음과 같은 과정으로 결정된다.

[단계 1] j 번째 속성 $a_j = \{x_{1j}, x_{2j}, \dots, x_{mj}\}$ 의 도메인 d_j , $d_j = [\min(a_j), \max(a_j)]$ 과 임의의 클래스에 대응된 j 번째 속성의 내부구간 $I_{jq} = [I_{jq}^L, I_{jq}^U]$, $I_{jq} \in d_j$ 들을 추출한다. I_{jq}^L 과 I_{jq}^U 는 j 번째 속성에서 q 번째 클래스에 속하는 속성값들의 최소와 최대값을 나타낸다.

[단계 2] 단계 1에서 추출된 내부구간들 사이에서 중첩영역(overlapping region) $O_j = [o_j^L, o_j^U]$ 을 찾는다. o_j^L 과 o_j^U 는 모든 중첩영역들 사이에서 하한과 상한경계를 나타낸다 (그림 1 참조).

그림 1. 중첩영역

Fig. 1. An example of overlapping region

[단계 3] 각 속성의 중첩영역에서 클래스 간 중복된 속성값의 분리성 정도(separation degree)로부터 해당 속성의 임계값(cut-off value)을 추출한다.

$$cp_j = \frac{\sum_{i=o_j^L}^{o_j^U} n_{ij} \cdot x_{ij}}{\sum_{i=o_j^L}^{o_j^U} n_{ij}}, \quad (8)$$

where $n_{ij} = h(x_{ij})$, $x_{ij} \in c_k, c_l$

식 (8)에서 n_{ij} 는 j 번째 속성의 중첩영역에서 속성값 x_{ij} 가 k, l 번째 클래스에 포함될 때의 빈도수를 의미하고, cp_j 는 빈도수에 대한 가중치 평균값 (임계값)을 나타낸다. 예를 들어 전체 인스턴스의 수가 30개일 때, j 번째 속성의 중첩영역에서 다음과 같이 클래스별 속성값들의 분포가 존재한다고 가정하면, 식 (8)에 의해서 j 번째 속성의 임계값은 2.2533이 된다.

a_j	2.0	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8
c_1	1	5	1	1	4	2	0	0	0
c_2	0	1	0	3	1	0	1	2	1
n_{ij}	-	6	-	4	5	-	-	-	-

반면에 전체 중첩영역의 분리성 정도를 평가한 경우에 임계값은 2.3391이 되고, 중첩영역에서 중복된 속성값의 분리성 정도를 평가한 것에 비해 상대적으로 높은 오분류 정도를 가진다 (전체 중첩영역을 고려한 경우 10개, 중첩영역에서 중복된 속성값을 고려한 경우 8개를 오분류).

3.2 규칙생성 및 특징선택

다음과 같은 형태의 퍼지 If ~ Then 규칙을 가정하자.

$R_k: \text{If } a_1 \text{ is } A_{k1}, a_2 \text{ is } A_{k2}, \dots, a_n \text{ is } A_{kn} \text{ Then } c_q$

여기서 $a_1, \dots, a_n$ 은 n 개의 입력속성, A_{kj} ($k = 1, \dots, N$; $j = 1, \dots, n$)는 k 번째 규칙에서 j 번째 입력속성의 퍼지 소속함수 (그림 2 참조), N 은 규칙의 수, c_q ($q = 1, \dots, Q$)는 클래스 레이블을 나타낸다.

그림 2. 퍼지 소속함수

Fig. 2. Fuzzy membership function

그림 2에서 cp_j 는 식 (8)로부터 계산된 j 번째 속성의 임계값, σ_j 는 표준편차 값을 의미하고, 퍼지 소속함수의 half-width는 cp_j 를 중심으로 $\pm \sigma_j$ 를 가진다. 임의의 입력 패턴 $\mathbf{x}_i$ 에 대한 후보규칙(candidate rule)은 식 (9)을 이용하여 최대 적합도를 가진 하나의 승자규칙(winner rule)[11]을 추출하고, 식 (9)로부터 생성된 후보규칙들 사이에서 커플링 문제 (즉 서로 다른 출력부 클래스를 가진 후보규칙들 중에서 동일한 입력부를 가진 규칙)는 참고문헌

[12]의 방법 (표 2 참조)을 이용하였다.

$$\mu_k^*(\mathbf{x}) = \max \{ \mu_k(\mathbf{x}) \}, k = 1, \dots, N$$

$$\text{where } \mu_k(\mathbf{x}) = \mu_{k1}(x_{i1}) \times \mu_{k2}(x_{i2}) \times \dots \times \mu_{kn}(x_{in}) \quad (9)$$

표 2에서 α 와 β 는 $\text{freq}(R_k^* \Rightarrow c_p)$ 와 $\text{freq}(R_k^* \Rightarrow c_q)$ 을 의미하고, 이것은 후보규칙의 출력부 클래스가 각각 p 번째와 q 번째 클래스에 대응될 때의 빈도수를 나타낸다. 또한 γ 와 δ 는 $\max(R_k^* \Rightarrow c_p)$ 와 $\max(R_k^* \Rightarrow c_q)$ 을 의미하고, 생성된 후보규칙의 출력부 클래스가 p 번째와 q 번째 클래스에 대응될 때 후보규칙들에 대한 최대 퍼지 적합도를 나타낸다.

표 2. 커플링 문제를 해결하기 위한 기준

Table 2. Two criteria to resolve the coupling problem

Max. degrees Frequencies	$\gamma > \delta$	$\gamma = \delta$	$\gamma < \delta$
$\alpha > \beta$	c_p	c_p	c_p
$\alpha = \beta$	c_p	NA	c_q
$\alpha < \beta$	c_q	c_q	c_q

본 연구에서는 표 2의 기준으로부터 생성된 퍼지 If ~ Then 규칙에서, 입력부 속성에 대한 중요특징들을 선택하기 위해서 식 (7)의 중요성 정도가 $\sigma_p(Q, a) = 0$ 또는 $\gamma_p(Q) = \gamma_{p-\{a\}}(Q)$ 인 조건을 만족하는 리덕트 속성들의 집합을 추출하여 최종적인 규칙으로 사용하였다.

4. 실험 및 결과

실험에서는 제안된 방법의 효과성을 보이기 위해서 하드분할을 기반으로 한 기존의 3가지 이산화방법과 퍼지분할과 러프집합을 결합한 3가지 이산화방법들의 10-fold 교차검증 (훈련: 90%, 실험: 10%)의 결과들을 비교하였다. 또한 10-fold 교차검증 동안에 각 fold의 평가는 식 (10)과 (11)의 분류 정확성과 G-mean 척도[13]를 근거로 평가하였다 (표 3 참조).

표 3. 혼동행렬

Table 3. Confusion matrix

Actual output	Positive Negative	Predicted output	
		Positive	Negative
		<i>TP</i>	<i>FN</i>
		<i>FP</i>	<i>TN</i>

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP} \quad (10)$$

$$G\text{-mean} = \sqrt{\text{sensitivity} \times \text{specificity}},$$

$$\text{where } \text{sensitivity} = \frac{TP}{TP + FN}, \text{ specificity} = \frac{TN}{TN + FP} \quad (11)$$

식 (10)에서 TP, FN, TN, FP는 각각 true positive, false negative, true negative, false positive를 의미하고, 식 (11)에서 민감도(sensitivity)와 특이도(specificity)는 true positive rate와 true negative rate를 나타내고, 이 척

도는 불균형 분포(imbalanced distribution)를 가진 데이터의 성능평가를 위한 척도로서 주로 사용된다. 실험에서 사용된 데이터-셋은 A 광역시에 소재한 A 의료원에 2006년 7월에서 2007년 6월 사이에 호흡곤란으로 내원한 환자 1,129명의 의무기록을 근거로 대상자의 인적사항을 제외한 등록번호, 성별, 나이, 응급실 내원일자 및 시간, 진료결과, 입원 시 진단, 초기 검사 항목 등의 자료를 데이터웨어하우스에서 추출하였다. 추출된 자료 중 타 병원으로 전원된 환자, DOA (Death on arrival), CPR(Cardio-pulmonary resuscitation) 후 혹은 DNR(Do not resuscitate)로 사망한 환자, 자의 퇴원 혹은 미상의 기타 환자, 의무기록이 불완전한 경우를 제외한 55개의 특징들 중에서 감별진단(differential diagnosis) 시 임상전문가에 의해서 중요하다고 판단되는 11개의 특징들로 구성된 668명의 환자 (즉 입원환자 500명, 퇴원환자 168명)에 대한 데이터-셋을 이용하였다. 그림 3과 표 4는 사용된 데이터-셋의 공간적인 분포와 특징을 나타낸다. 그림 3에서 x축은 속성값, y축은 속성값에 대한 빈도수, admission과 discharge는 각각 입원환자와 퇴원환자 군에 대한 속성값의 분포를 보여준다. 그림에서 볼 수 있듯이, 전체 11개의 속성에 대해서 퇴원환자 군의 분포가 입원환자 군의 분포에 완전히 포함된 형태를 나타낸다.

표 4. 데이터-셋의 특징

Table 4. Dataset's feature

Feature	Unit	Min	Max	Mean $\pm$ σ
<i>WBC</i>	$\times 10^3/uL$	0.11	75.9	11.02 $\pm$ 6.39
<i>PLT</i>	$\times 10^3/uL$	23	1,105	270.69 $\pm$ 120.02
<i>Cl⁻</i>	<i>mmol/L</i>	72	134	104.25 $\pm$ 6.94
<i>AST</i>	<i>U/L</i>	5	3,321	73.52 $\pm$ 227.15
<i>ALT</i>	<i>U/L</i>	3	2,481	46.44 $\pm$ 143.22
<i>pCO₂</i>	<i>mmHg</i>	8.3	98.5	39.88 $\pm$ 13.57
<i>pO₂</i>	<i>mmHg</i>	35.9	354	80.15 $\pm$ 22.06
<i>O₂SAT</i>	%	59	99.9	96.14 $\pm$ 3.20
<i>LDH</i>	<i>U/L</i>	152	8,178	688.58 $\pm$ 509.81
<i>Ca²⁺</i>	<i>mEq/L</i>	1.25	3.2	2.25 $\pm$ 0.17
<i>Mg²⁺</i>	<i>mg/dL</i>	0.3	4.1	2.21 $\pm$ 0.38

WBC: 백혈구(White blood cell); *PLT*: 혈소판 수 (Platelet count); *Cl⁻*: 염소농도(Chloride); *AST*: 아스파르트레이트 아미노전이효소(Aspartate aminotransaminase); *ALT*: 알라닌 아미노전이효소(Alanine amino-transaminase); *PCO₂*: 이산화탄소 압(Pressure of carbon dioxide); *PO₂*: 산소 압(Pressure of oxygen); *O₂SAT*: 산소포화도(Oxygen saturation); *LDH*: 젖산 탈수소효소(Lactate dehydrogenase); *Ca²⁺*: 칼슘(Calcium); *Mg²⁺*: 마그네슘(Magnesium).

4.1 하드분할을 기반으로 한 이산화방법들과의 비교

하드분할을 기반으로 10-fold 교차검증에서는 러프집합의 특징선택 문제와 커플링 문제를 해결하기 위한 기준을 고려하지 않았을 때, 기존의 3가지 이산화방법들과 하드분할을 기반으로 한 제안된 방법의 결과를 비교하였고, 그 결과는 표 5와 같다.

그림 3. 데이터-셋의 공간적인 분포
Fig. 3. Spatial distribution of dataset

표 5. 10-fold 교차검증 결과

Table 5. Comparison results (10-fold cross validation)

	E.W	E.F	Ent.	Pro-posed
평균 규칙 수	24	481	214.8	480.5
평균 정확도	0.0749	0.3067	0.3354	0.2993
평균 G-mean	0.0000	0.2236	0.1652	0.2195

E.W: Equal-width(distance) partition; E.F: Equal-frequency(depth) partition; Ent.: Entropy-based partition

표 5에서 ‘평균 규칙 수’는 임상 전문가에 의해서 선택된 11개의 특징을 이용하였을 때 생성된 규칙의 수를, ‘평균 정확도’와 ‘평균 G-mean’은 10-fold 교차검증 동안에 각 fold에서 식 (10)과 (11)을 이용하여 평가된 결과값을 보여준다. 표 5의 결과에서 보여주듯이, 훈련 데이터로부터 생성된 ‘평균 규칙의 수’는 equal-width를 기반으로 한 이산화방법이 가장 적은 평균 규칙의 수를 제공하였지만, 실험 데이터에 대한 ‘평균 정확도’는 가장 낮은 성능을 제공하였고, ‘평균 G-mean 값’은 입원환자(admission)와 퇴원환자(discharge)군 중에서 데이터의 빈도수가 상대적으로 많은 입원환자군으로만 분류된 결과를 보여주었다. 반면에 equal-frequency를 기반으로 한 이산화방법과 식 (8)을 이용하여 특징공간을 하드 분할한 제안된 이산화방법이 ‘평균 G-mean 값’에서 equal-width와 entropy-based 이산화방법들에 비해 상대적으로 좋은 성능을 제공함을 볼 수 있었고, 사용된 11가지 혈액검사(blood test) 항목들에 대한 임상적 정상치 범위(clinical reference range)[14] (즉 WBC: [5.2, 12.4], PLT: [130, 400], Cl⁻: [95, 108], AST: [13, 36], ALT: [5, 44], pCO₂: [32, 48], pO₂: [83, 108], O₂SAT: [95, 99], LDH: [211, 423], Ca²⁺: [1.2, 3.2], Mg²⁺: [1.5, 2.7])와 10-fold 동안에 각 방법에서 결정된 임계값의 변화된 범위(variation)로부터 다음과 같은 결과를 얻을 수 있었다: equal-width 기반 이산화방법은 Cl⁻, Ca²⁺, Mg²⁺를 제외한 나머지 8개의 검사항목에서, entropy-based 이산화방법은 WBC, pO₂, LDH에서, equal-frequency와 제안된 방법은 pO₂와 LDH에서 정상치 범위와 비교하였을 때 다소 벗어난 결과를 제공하였다.

다음은 각 이산화방법에서 10-fold 교차검증 동안에 결정된 임계값의 변화된 범위를 보여준다: I) equal-width 방법: WBC: [26.79, 38.13], PLT: [388, 573.5], Cl⁻: [99.5, 106], AST: [831, 1,664], ALT: [511, 1,242.5], pCO₂: [52.15, 54], pO₂: [138.45, 196.55], O₂SAT: [79.45, 90.45], LDH: [2,585, 4,223], Ca²⁺: [2.09, 2.315], Mg²⁺: [2.1, 2.6]; II) equal-frequency 방법: WBC: [9.44, 9.57], PLT: [252, 255], Cl⁻: [105, 105], AST: [31, 32], ALT: [21, 22], pCO₂: [36.4, 36.9], pO₂: [77.5, 78.8], O₂SAT: [96.7, 96.8], LDH: [574, 592], Ca²⁺: [2.24, 2.25], Mg²⁺: [2.2, 2.2]; III) entropy-based 방법: WBC: [4.26, 22.55], PLT: [169, 358], Cl⁻: [95, 111], AST: [19, 39], ALT: [10, 45], pCO₂: [23.3, 67.9], pO₂: [56.4, 125], O₂SAT: [97.2, 99.4], LDH: [668, 1,008], Ca²⁺: [1.8, 2.37], Mg²⁺: [1.8, 2.9]; IV) 제안된 방법: WBC: [8.74, 9.33], PLT: [245.72, 254.40], Cl⁻: [104.52, 104.89], AST: [31.28, 32.66], ALT: [21.01, 21.63], pCO₂: [35.45, 36.36], pO₂: [79.36, 81.98], O₂SAT: [96.60, 96.75], LDH: [535.76, 559.22], Ca²⁺: [2.23, 2.25], Mg²⁺: [2.16,

2.19].

4.2 퍼지분할을 기반으로 한 이산화방법들과의 비교

본 절에서는 이전의 실험에서 결정된 각 이산화방법들의 임계값을 중심으로 퍼지 소속함수의 half-width를 σ_j ($j=1,\dots,11$)로 설정하였을 때 10-fold 교차검증 결과를 비교하였고 (표 6 참조), 성능에 대한 평가 기준은 이전의 실험에서와 동일한 기준을 사용하였다. 표 6에서 ‘특징선택 전 평균 규칙 수’는 전체 11개의 속성을 퍼지이산화 하였을 때 생성된 평균 퍼지규칙의 수, ‘특징선택 후 평균 규칙 수’는 중요특징 (즉 리덕트) 집합을 선택한 후 생성된 평균 퍼지 규칙의 수, ‘평균 리덕트 속성 수’는 [정의 5]를 근거로 각 fold에서 중요특징으로 선택된 속성들의 평균 수를 의미하고, ‘불필요한 규칙제거 후 규칙 수’는 특징선택 후 러프집합의 식별가능행렬과 식별가능함수[15, 16]를 적용하였을 때의 규칙 수를 나타낸다.

표 6의 실험결과에서, equal-width 기반 퍼지이산화 방법이 다른 3가지 이산화방법들에 비해 상대적으로 높은 평균 분류정확도를 보여주지만, 평균 G-mean 값은 가장 좋지 않은 결과를 제공하였고, 반면에 equal frequency와 제안된 퍼지이산화 방법은 equal-width 퍼지이산화 방법에 비해 평균 분류정확도는 조금 낮은 성능을 제공하지만, 평균 G-mean 값은 가장 좋은 성능을 제공함을 볼 수 있었다.

표 6. 10-fold 교차검증 결과

Table 6. Comparison results (10-fold cross validation)

	E.W	E.F	Ent.	Pro-posed
특징선택 전 평균 규칙 수	35.90	442.60	224.40	488.80
특징선택 후 평균 규칙 수	22.50	442.60	222.90	488.80
평균 리덕트 속성 수	5.40	11.00	10.90	11.00
불필요한 규칙제거 후 평균 규칙 수	6.00	305.40	135.60	330.10
평균 정확도	0.7455	0.7365	0.7365	0.7365
평균 G-mean	0.0725	0.4876	0.2931	0.4868

또한 이전 실험에서처럼 각 이산화방법에 대해서 10-fold 교차검증을 수행하였을 때 equal-width 퍼지이산화 방법 (WBC: [26.79, 38.13], PLT: [388, 564], Cl⁻: [99.5, 106], AST: [1,663, 1,663], pCO₂: [52.15, 54], LDH: [4,165, 4,165], Ca²⁺: [2.315, 2.24], Mg²⁺: [2.1, 2.6])을 제외한 나머지 퍼지 이산화방법들은 동일한 임계값의 변화된 범위를 나타내었다.

5. 결론 및 토의

연속형 또는 혼합형 속성을 가진 고차원 데이터-셋으로부터 중요특징을 선택하고, 관련 특징들의 임계값을 결정하는 것은 임상 의사결정시스템을 구축하는데 있어서 매우 중요한 연구주제이다. 본 연구에서는 연속형 속성을 가진 데이터-셋에서 속성값들의 공간적인 분포를 이용한 퍼지 이

산화 방법을 제안하였다. 제안된 방법에서 속성값들의 공간적인 분포는 속성값들의 분리성 정도를 평가하기 위한 기준으로 이용하였고, 러프집합의 개념 중 의존성과 중요성 정도는 주어진 전체 속성들로부터 중요특징들의 하위집합을 추출하기 위해 사용하였다.

실험에서는 제안된 방법의 타당성을 보이기 위해서 A 광역시에 소재한 A 의료원에 2006년 7월에서 2007년 6월 사이에 응급실에 호흡곤란으로 내원한 환자 688명을 대상으로 하였고, 3가지 이산화기법(equal-width, equal-frequency, entropy based 분할)을 기반으로 한 하드분할과 퍼지분할에 러프집합을 결합한 방법들의 결과를 비교하였다. 실험결과, 3가지 이산화방법들을 퍼지이산화로 확장하고 커플링 해결기준을 적용하였을 때 하드분할을 기반으로 한 이산화방법에 비해 보다 좋은 성능을 제공하였고, 평균 분류정확도와 평균 G-mean 성능이 향상됨을 볼 수 있었다. 또한 임상적 측면에서 질병의 감별진단을 위한 판단기준으로 사용되는 임상검사의 변환된 범위를 분석한 결과, equal frequency와 제안된 퍼지 이산화방법을 적용하였을 때 가장 타당한 결과를 제공함을 볼 수 있었다. 그러나 교차검증 동안에 생성된 평균 규칙의 수에서 equal frequency와 제안된 퍼지 이산화방법이 다른 2가지 이산화방법에 비해 다소 많은 수의 규칙을 생성하였다. 향후 연구에는 이러한 제약점을 개선시키기 위한 연구와 혼합형 속성을 가진 데이터셋으로부터의 지식 획득 및 최적화에 대한 연구가 필요할 것으로 판단된다.

참 고 문 헌

- [1] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," *Artificial Intelligence in Medicine*, vol.23, no.1, pp.89-109, 2001.
- [2] 이상훈, 박정은, 오경환, "데이터 분포를 고려한 연속 값 속성의 이산화," *한국퍼지 및 지능시스템 학회 논문지*, 제13권, 4호, pp.391-396, 2003.
- [3] D. Chiu, A. Wong, and B. Cheung, *Information discovery through hierarchical maximum entropy discretization and synthesis*, MIT Press, 1991.
- [4] U.M. Fayyad and K.B. Irani, "Multi-interval discretization of continuous attributes as pre-processing for classification learning," *Proc. 13th International Joint Conference on Artificial Intelligence*, pp.1022-1027, 1993.
- [5] R. Kerber, "Discretization of numerical attributes," *Proc. the 10th National Conference on Artificial Intelligence (AAAI-92)*, pp.123-128, 1992.
- [6] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proc. the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp.281-297, 1967.
- [7] L. Kaufman and P.J. Rousseeuw, *Finding group in data: an introduction to cluster analysis*, John Wiley & Sons, New York, 1990.
- [8] C.-C. Shen and Y.-L. Chen, "A dynamic-programming algorithm for hierarchical discretiza-

tion of continuous attributes," *European Journal of Operational Research*, vol.184, no.2, pp. 636-651, 2008.

- [9] Z. Pawlak, "Rough set," *Int. J. Comput. Inform. Sci.*, vol.11, no.5, pp.341-356, 1982.
- [10] Z. Pawlak, *Rough sets, Theoretical aspects of reasoning about data*, Kluwer, Dordrecht, Netherlands, 1991.
- [11] H. Ishibuchi and T. Nakashima, "Effect of rule weights in fuzzy rule-based classification systems," *IEEE Trans. Fuzzy Systems*, vol.9, no.4, pp.506-515, 2001.
- [12] 손창식, 정환목, 서석태, 권순학, "규칙의 커플링문제를 최소화하기 위한 퍼지-러프 분류방법," *한국 퍼지 및 지능시스템 학회 논문지*, 제17권, 4호, pp.460-465, 2007.
- [13] Y.M. Sun, M.S. Kamel, A.K.C. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognition*, vol.40, no.12, pp.3358-3378, 2007.
- [14] 전시자 외 9인, *성인간호학 (상) 4판*, 현문사, 2005.
- [15] R. Slowinski, *Intelligent decision support, Handbook of applications and advances of the rough sets theory*, Kluwer Academic Publishers, Dordrecht, 1992.
- [16] B. Walczak and D.L. Massart, "Rough set theory," *Chemometrics and Intelligent Laboratory Systems*, vol.47, no.1, pp.1-16, 1999.

저 자 소 개

손창식(Chang-Sik Son)

2000년 : 대구가톨릭대학교 전자정보공학부 (공학사)
 2002년 : 대구가톨릭대학교 전산통계학과 (이학석사)
 2006년 : 대구가톨릭대학교 컴퓨터정보통신공학부 (공학박사)

2007년 ~ 2009년 : 영남대학교 전기공학과 박사 후 연구원
 2009년 ~ 2010년 : 계명대학교 생체정보기술개발사업단 연구교수
 2010년 ~ 현재 : 계명대학교 의과대학 의료정보학교실 연구강사

관심분야 : 데이터마이닝, 영상처리, 임상 의사결정시스템
 Phone : 053-580-3740
 Fax : 053-580-3746
 E-mail : csson@dsmc.or.kr

신아미(A-Mi Shin)

2002년 : 계명대학교 간호학과 (학사)
 2002년 ~ 2005년 : 계명대학교 동산의료원
 간호사
 2007년 : 계명대학교 간호학과 (석사)
 2007년 ~ 현재 : 계명대학교 생체정보기술
 개발사업단 연구원

2009년 ~ 현재 : 계명대학교 의과대학 의료정보학교실 박사과정

관심분야 : 데이터마이닝
 Phone : 053-580-3744
 Fax : 053-580-3745
 E-mail : ami535@kmu.ac.kr

박형섭(Hyung-Seob Park)

2000년 : 계명대학교 의과대학 의학과
 (의학사)
 2004년 : 계명대학교 의과대학 의학과
 (의학석사)
 2008년 ~ 2009년 : 계명대학교 의과대학
 내과학교실 연구강사

2009년 ~ 현재 : 계명대학교 의과대학 내과학교실 전임강사

관심분야 : 내과학(심장내과학), 부정맥, 전기생리학적 검사
 및 전극도자질제술, 영구 심박동기
 Phone : 053-250-7411
 Fax : 053-250-7034
 E-mail : parkgyver@msn.com

이인희(In-Hee Lee)

2000년 : 대구대학교 재활과학대학
 물리치료학과 (학사)
 2000년 ~ 현재 : 계명대학교 동산의료원
 물리치료사
 2005년 : 계명대학교 의학과 (석사)
 2009년 ~ 현재 : 계명대학교 의과대학 의료
 정보학교실 박사과정

관심분야 : 재활복지공학, 신호처리, 정보공학
 Phone : 053-580-3744
 Fax : 053-580-3745
 E-mail : yiinhee@hanmail.net

김윤년(Yoon-Nyun Kim)

1980년 : 경북대학교 의학과 (의학사)
 1983년 : 경북대학교 의학과 (의학석사)
 1987년 : 경북대학교 의학과 (의학박사)
 1984년 ~ 현재 : 계명대학교 의과대학
 내과학교실 교수
 2002년 ~ 현재 : 계명대학교 보건의료정보
 기술연구소 소장

2004년 ~ 현재 : 계명대학교 생체정보기술개발사업단장

관심분야 : 부정맥, 의료정보시스템, u-Healthcare
 Phone : 053-250-7432
 Fax : 053-428-7953
 E-mail : ynkim@dsmc.or.kr

박희준(Hee-Joon Park)

1999년 : 경북대학교 전자공학과(공학사)
 2001년 : 금오공과대학 전자공학과
 (공학석사)
 2006년 : 경북대학교 전자공학과(공학박사)
 2006년 ~ 2010년 : 계명대학교 의과대학
 의료정보학교실 연구강사

2010년 ~ 현재 : 계명대학교 의과대학 의용공학과 전임강사

관심분야 : 생체신호처리, u-Healthcare
 Phone : 053-580-3730
 Fax : 053-580-3731
 E-mail : hjpark@dsmc.or.kr