Journal of Biomedical Engineering Research 32: 134-143 (2011)

학 술 논 문

데이터 이산화와 러프 근사화 기술에 기반한 중요 임상검사항목의 추출방법: 담낭 및 담석증 질환의 감별진단에의 응용

손창식[1] · 김민수[2] · 서석태[2] · 조윤경[3] · 김윤년[1,3]

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

Extraction Method of Significant Clinical Tests Based on Data Discretization and Rough Set Approximation Techniques: Application to Differential Diagnosis of Cholecystitis and Cholelithiasis Diseases

Chang-Sik Son[1], Min-Soo Kim[2], Suk-Tae Seo[2], Yun-Kyeong Cho[3] and Yoon-Nyun Kim[1,3]

[1]*Dept. of Medical Informatics, School of Medicine, Keimyung Univ.*
[2]*Biomedical Information Technology Center, School of Medicine, Keimyung Univ.*
[3]*Dept. of Internal Medicine, School of Medicine, Keimyung Univ.*
(Received December 10, 2010. Accepted March 21, 2011)

Abstract: The selection of meaningful clinical tests and its reference values from a high-dimensional clinical data with imbalanced class distribution, one class is represented by a large number of examples while the other is represented by only a few, is an important issue for differential diagnosis between similar diseases, but difficult. For this purpose, this study introduces methods based on the concepts of both discernibility matrix and function in rough set theory (RST) with two discretization approaches, equal width and frequency discretization. Here these discretization approaches are used to define the reference values for clinical tests, and the discernibility matrix and function are used to extract a subset of significant clinical tests from the translated nominal attribute values. To show its applicability in the differential diagnosis problem, we have applied it to extract the significant clinical tests and its reference values between normal (N = 351) and abnormal group (N = 101) with either cholecystitis or cholelithiasis disease. In addition, we investigated not only the selected significant clinical tests and the variations of its reference values, but also the average predictive accuracies on four evaluation criteria, i.e., accuracy, sensitivity, specificity, and geometric mean, during 10-fold cross validation. From the experimental results, we confirmed that two discretization approaches based rough set approximation methods with relative frequency give better results than those with absolute frequency, in the evaluation criteria (i.e., average geometric mean). Thus it shows that the prediction model using relative frequency can be used effectively in classification and prediction problems of the clinical data with imbalanced class distribution.

Key words: Data Discretization, Rough Set, Cholecystitis, Cholelithiasis, Differential Diagnosis

Corresponding Author : 김윤년
(704-701) 대구시 달서구 달구벌대로 2800 계명대학교 성서캠퍼스 의과대학 의료정보학교실
TEL: +82-53-580-3740 / FAX: +82-53-580-3745
E-mail: ynkim@dsmc.or.kr
본 연구는 지식경제부 지방기술혁신사업(RTI04-01-01)지원으로 수행되었습니다.

I. 서 론

임상검사의 참고치는 질환 진단 및 환자의 예후를 판단하는데 있어서 중요한 기준이 되며 올바른 참고치의 설정은 매우 중요하다. 일반적으로 의료기관에서 사용되고 있는 참고

치는 검사 시약이나 기기를 공급하는 회사에서 제시하는 기준을 참고하여 개개의 검사실에서 다시 산출하여 사용하거나, 검사실에 적합한 참고치를 도출하여 사용하도록 권고하고 있다[1]. 그러나 의료기관마다 일부 항목에서는 검사기기, 사용방법 및 시약 등이 같은 데도 불구하고 참고치의 기준이 상이한 경우가 있으며, 국외에서 보고된 기준을 수정하여 적용하는 경우도 있다. 이로 인해 판독의 통일성 결여와 결과의 신뢰성 등이 문제가 될 수 있으며, 다양한 질환에 대한 임상검사의 참고치 뿐만 아니라 한국인에게 적합한 참고치를 획득할 수 없다는 제약점을 가진다.

차은종 외 6인[2]은 간기능 검사결과의 자동분석을 위해 전문가에 의해서 선별된 임상검사 6종목을 근거로 퍼지 이론(fuzzy theory)를 적용한 분석 방법을 제안하였고, 이갑노 외 6인[1]은 2000년 10월에서 2001년 9월까지 15개 지역 중앙검진센터 검사실에서 수집된 자료 중 56종목의 검사결과를 선별하여 통계적 검정 법(Shapiro-Wilk 검정과 Kolmogorov 검정)을 이용하여 이들 검사항목의 참고치를 결정하였다. 그리고 손창식 외 5인은 2006년 7월에서 2007년 6월까지 응급실에 호흡곤란으로 내원한 환자들의 의무기록을 대상으로 입·퇴원의 유무를 결정하기 위한 가변 임계값 기반 특징선택 방법[3]과 규칙가중치 기반 퍼지 분류 모델[4]을 제안하였다. 이들 대부분의 선행연구 및 방법들은 임상적 측면에서 보다 신뢰할 수 있는 결과를 제공하기 위해서, 수집된 임상검사 결과들로부터 전문가에 의해서 선별된 항목만을 사용하여 중요 검사항목을 추출하고 참고치를 결정하고 있다. 그러나 수집된 대부분의 임상 데이터 (clinical raw data)의 경우 이상치(outlier)와 같은 잡음뿐만 아니라, 질환 별 사례의 수가 서로 다른 비율을 나타내는 클래스 불균형 분포(imbalanced class distribution)를 가지기 때문에, 데이터의 전처리 과정을 고려하지 않고 분석할 경우 임상적으로 의미 있는 결과를 도출하기란 어렵다.

따라서 본 연구에서는 이러한 문제점을 고려하기 위해서 클래스 불균형 분포를 가진 담낭 및 담석증 질환을 가진 환자들의 임상검사 결과로부터 진단 혹은 감별진단 시 중요검사항목과 참고치를 추출할 수 있는 데이터 이산화 기반 러프 근사화 방법에 대해서 논의하고, 실험결과를 통해 그 효과성을 보이고자 한다.

II. 재료 및 방법

1. 연구대상

D광역시에 소재한 D의료원에 2006년 7월에서 2007년 6월 사이에 복통을 주증상으로 응급실에 내원한 환자 1,103명의 의무기록 중에서 국제질병사인(International Classification of Diseases, ICD-10)[5]분류코드에서 담낭 및 담석증으로 진단된 환자 101명(급성 담낭염을 동반한 쓸개(담낭)의 결석(K80.0), 기타 담낭염을 동반한 쓸개(담낭)의 결석(K80.1), 담관염을 동반한 쓸개관(담관) 결석(K80.3), 급성 담낭염(K81.0), 담관염(K83.0))과 복통을 주호소로 내원하였으나 임상검사 소견상 별 다른 특이점이 없어 퇴원된 환자 351명을 대상으로 하였다.

또한 대상환자들의 자료 추출 시 인적사항을 제외한 등록번호, 성별, 나이, 응급실 내원일자 및 시간, 진료결과, 입원 시 진단, 초기 검사 항목 등의 정보를 수집하였다. 수집된 자료 중에서 초기 검사항목으로는 52가지 검사항목, 15가지 전혈구 검사(common blood cell & differential count, CBC & diff. count), 프로트롬빈 시간(prothrombin time, PT), 활성화 부분 트롬보플라스틴 시간(activated partial thromboplastin, APTT), 3가지 혈청 전해질 검사(serum electrolytes), 13가지 입원환자에 대한 검사(routine admission), 혈청아밀라제(amylase), 리파아제(lipase), 17가지 소변검사가 있었으며, 복통의 통증 부위를 나타내는 주증상(chief complaints)도 하나의 독립변수로 고려하였다. 또한 담낭 및 담석증으로 진단 받은 환자 101명 (9.16%)과 퇴원한 환자 351명 (31.8%)의 주증상으로는 복통(abdominal), 배꼽주위의 통증(periumbilical), 우측 하복부 통증(right lower quadrant), 좌측 하복부 통증(left lower quadrant), 하복부 통증(lower abdominal), 좌측 상복부 통증(left upper quadrant), 우측 상복부 통증(right upper quadrant), 상복부 통증(upper abdominal)과 같이 8가지로 조사되었다.

2. 연구방법

본 연구에서는 담낭 및 담석증 질환을 가진 환자 군과 복통을 주호소로 내원하였으나 임상검사 소견상 별 다른 특이점이 없어 퇴원한 환자군들 간에 감별진단을 위한 중요 검사항목들과 임상적 참고치(clinical reference)를 추출하기 위해서 2가지 데이터 이산화 방법을 기반으로 한 러프 근사화 방법을 이용하였다. 그림 1은 본 연구에서 사용된 방법의 과정을 나타내고, 각 절에서 이들 단계에 대한 세부적인 사항을 설명하였다.

(1) 이산화 방법

일반적으로 데이터 이산화의 목적은 정보의 손실을 최소화하고 미지의 데이터에 대한 범주의 일반성을 최대화할 수 있는 구간을 찾는 것이며, 가장 대표적인 이산화 방법으로는 등 간격 이산화(equal width discretization)와 등 빈번 이산화(equal frequency discretization) 방법으로 구분된다.

등 간격 이산화 방법[6]은 각 속성에 대해서 입력공간의 전체 도메인을 동일한 등 간격 $d = (\alpha_{max} - \alpha_{min})/k$ (여기서 α_{max}와 α_{min} 은 각각 임의의 속성에서 속성값들의 최대값과 최소

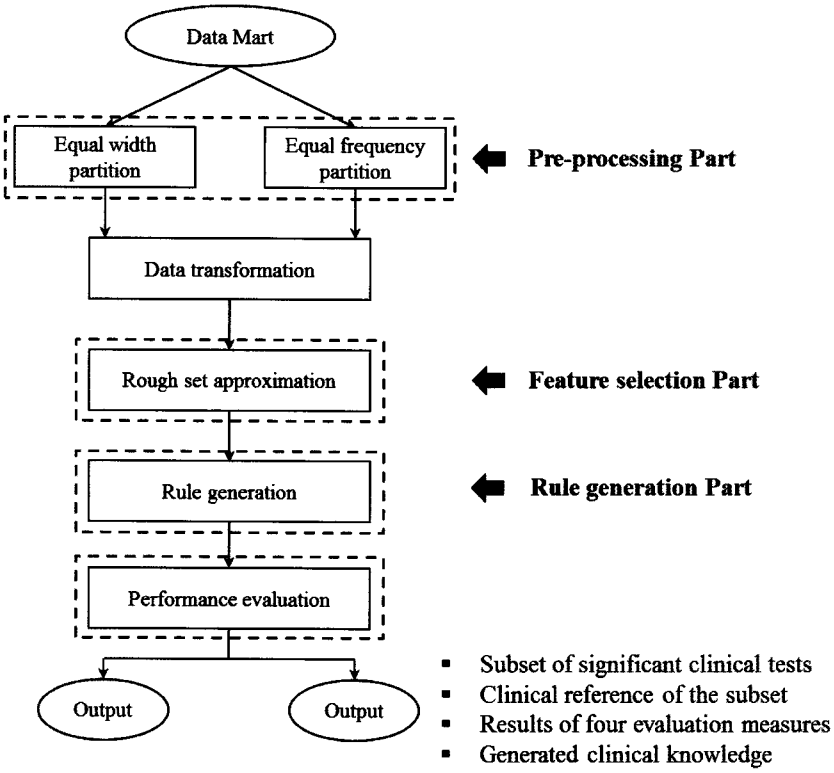

그림 1. 시스템 흐름도
Fig. 1. System flow chart

값, k는 분할된 구간의 수)으로 분할하는 것을 의미한다. 만약 임의의 속성에 대해서 분할된 구간의 폭(width) 혹은 거리(distance)가 d라고 할 때, i번째 구간에서의 경계와 이를 기준으로 특징공간을 분할하기 위한 소속함수(membership function)는 다음과 같이 정의된다:

$$a[i]_{\min} = a_{\min} + d \times i$$
$$a[i]_{\max} = a_{\min} + d \times (i+1),\ i = 0,1,...,k-1 \quad (1)$$

$$MF_i(x) = \begin{cases} 1, & \text{if } a[i]_{\min} \le x < a[i]_{\max} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

이에 반해, 등 빈번 이산화 방법[6]은 각 속성의 특징공간을 근사적으로 동일한 샘플수를 가지도록 k개의 부분공간으로 분할하는 방법을 나타낸다. 만약 N개의 속성값을 가진 임의의 속성이 존재하고, 이때 분할된 구간의 크기 s가 N/k이라면, 주어진 속성에서 i번째 구간의 경계와 이산화 함수는 다음과 같이 정의된다:

$$a[i]_{\min} = i \times s,$$
$$a[i]_{\max} = (i+1) \times s,\ i = 0,1,...,k-1 \quad (3)$$

$$MF_i(x) = \begin{cases} 1, & \text{if } a[i]_{\min} \le r[x] < a[i]_{\max},\ i=0,1,\ldots,k-2 \\ 1, & \text{if } a[i]_{\min} \le r[x] \le a[i]_{\max},\ i=k-1 \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

여기서 $r[x]$는 정렬된 속성값 x의 순서(rank)를 의미한다.

그림 2는 분할된 구간의 수 k가 3이라고 가정할 때, 등 간격 이산화 방법과 등 빈번 이산화 방법을 도식화한 예를 보여준다. 그림 2에서 점선으로 표시된 막대 그래프는 해당 속성값의 출현 빈도수를 의미하고, CP1과 CP2는 각 이산화 방법에 의해서 결정된 분할 경계 (즉 하한 경계값과 상한 경계값)을 나타낸다. 그림 2(a)의 등 간격 이산화의 경우, 식 (1)에서 정의된 것처럼 각 구간 'a_{min}~CP1', 'CP1~CP2', 'CP2~a_{max}'는 모두 동일한구간의 폭을 가지는 것을 볼 수 있으며, 그림 2(b)의 등 빈번 이산화의 경우, 전체 속성값의 빈도수 N과 분할 구간 수 k 값에 따라 구간의 간격이 근사적으로 결정됨을 볼 수 있다.

본 연구에서는 이들 2가지 이산화 방법에 의해서 결정된 분할 경계의 하한과 상한 값을 기준으로, 수집된 임상 데이터를 lower abnormal, normal 및 upper abnormal과 같이 3개의 명목형 속성값으로 변환하였다.

(2) 러프 근사화

러프 근사화(rough approximation)는1982년 Pawlak[7]에 의해서 제안된 모호성과 불확실성을 다룰 수 있는 집합의 개념으로, 하한근사(lower approximation)와 상한근사(upper approximation)의 2가지 집합의 근사화 개념을 활용하여 데이터에 존재하는 패턴의 의존성을 고려하여 중요 특징을 추출

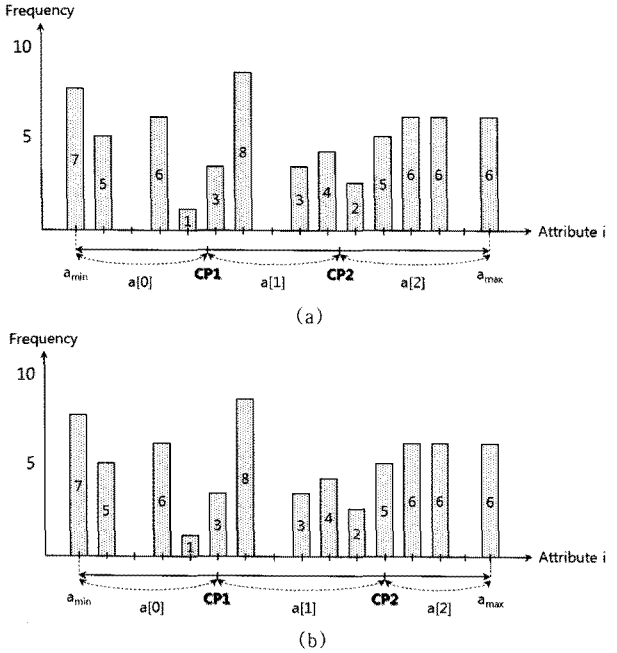

그림 2. 분할 구간의 수가 3일 때, 등 간격 이산화와 등 빈번 이산화의 예. (a) 등 간격 이산화, (b) 등 빈번 이산화

Fig. 2. An example of equal width and frequency discretization when $k = 3$. (a) Equal width discretization, (b) Equal frequency discretization.

하고, 의사결정을 위한 지식을 자동 생성할 수 있다는 특성들 때문에 제어, 진단, 예측문제 등과 같은 다양한 응용분야에서 폭넓게 활용되고 있다. 본 연구에서는 임상 데이터 내에 포함된 코어(core)와 리덕트 (reduct)[7~9] 특성 (즉 임상검사항목)을 추출하기 위해서 Slowinski와 Stefanowski[8]에 의해서 제안된 식별가능행렬(discernibility matrix)과 식별가능함수(discernibility function)의 개념을 이용하였고, 그 개념은 다음과 같이 정의된다:

정보 시스템(information system) $IS = (U,A)$는 오브젝트 또는 사례들의 집합 $U = \{x_1,x_2,...,x_n\}$과 속성들의 집합 A ($A = C \cup D$)으로 구성되고, 이때 C는 조건부 속성(독립변수), D는 의사결정부 속성(종속변수)들의 집합으로 표현될 때, 임의의 속성 $a \in A$의 도메인 V_a는 정보 함수 $f_a: U \rightarrow V_a$에 의해서 정의되고, 식별가능행렬 DM은 식 (5)에 의해서 $n \times n$행렬로 표

표 1. 정보 시스템의 예

Table 1. An example of information system (IS)

Objects (Cases)	Conditional attributes				Decision attribute
	WBC	PLT	LYMP	ALP	DX
1	Upper abnormal	Upper abnormal	Normal	Upper abnormal	K80.0
2	Upper abnormal	Normal	Normal	Upper abnormal	K80.0
3	Normal	Normal	Normal	Normal	No
4	Upper abnormal	Lower abnormal	Upper abnormal	Normal	K80.0
5	Lower abnormal	Normal	Normal	Upper abnormal	No

현된다.

$$DM(c_{ij}) = \begin{cases} \varnothing & \text{if } D(x_i) = D(x_j) \\ \{a \in A : a(x_i) \neq a(x_j)\} & \text{otherwise} \end{cases} \quad (5)$$

$D(x_i)$와 $D(x_j)$는 의사결정부 속성에서 i와 j번째 오브젝트들의 속성값을 의미하고, $a(x_i)$와 $a(x_j)$는 각각 임의의 조건부 속성 a에서 i와 j번째 오브젝트들의 속성값을 의미한다. 이때 $c_{ij}(i \neq j; i,j = 1,2,...,n)$는 i와 j번째 오브젝트들을 구별 가능하게 하는 모든 속성들을 나타낸다.

식 (5)에 의해서 표현된 식별가능행렬의 원소들 중에서 유일하게 한 가지 속성으로만 구성된 원소들은 코어 특성으로 정의되고, 그렇지 않은 원소들은 리덕트 특성으로 정의된다. 여기서 코어 특성은 의사결정부 속성값 (즉 종속변수의 값)을 식별하는데 있어서 반드시 필요한 조건부 속성들을, 리덕트 특성은 한 개 이상의 코어 특성을 포함한 모든 조건부 속성들을 의미한다.

예를 들어 등 간격 이산화 방법에 의해서 명목형 속성값으로 변환된 정보 시스템이 표 1과 같다고 가정하자. 조건부 속성 WBC, PLT, LYMP, ALP는 각각 백혈구 수(white blood cell), 혈소판 개수(platelet), 림프구(lymphocyte), 알카라인포스파타제(alkaline phosphatase)를 의미하고, 이들 4가지 검사항목들은 식 (5)에서 정의된 조건부 속성들의 집합 A를 나

표 2. 식별가능행렬의 예

Table 2. An example of discernibility matrix

Objects (Cases)	1	2	3	4	5
1	X	-	WBC, PLT, ALP	-	WBC, PLT
2	-	X	WBC, ALP	-	WBC
3	WBC, PLT, ALP	WBC, ALP	X	WBC, PLT, LYMP	-
4	-	-	WBC, PLT, LYMP	X	WBC, PLT, LYMP, ALP
5	WBC, PLT	WBC	-	WBC, PLT, LYMP, ALP	X

타낸다. 의사결정부 속성에서 DX는 ICD-10 질병분류코드로 분류된 진단명을 나타내고, No는 복통을 주호소로 내원하였으나 임상검사 소견상 별 다른 특이점 없어 퇴원한 환자를 의미한다.

표 2는 식 (5)의 정의에 따라 생성된 식별가능행렬의 예를 보여준다. 표 2의 가로와 세로축에서 1~5는 해당 오브젝트들을 의미하고, 기호 'X'로 표시된 대각행렬 원소는 동일한 오브젝트들 간의 연산을 통해 제거된 원소들을 나타내며, 기호 '-'로 표시된 원소는 동일한 의사결정부 속성값을 가진 동치류(equivalence class)들 간의 연산을 통해 제거된 원소들을 의미한다. 또한 나머지 원소들은 오브젝트들 간에 조건부 속성값이 서로 다른 속성들을 나타낸다. 표 2의 식별가능행렬에서 5가지 부분집합 {WBC, PLT, ALP}, {WBC, PLT}, {WBC, ALP}, {WBC, PLT, LYMP}, {WBC, PLT, LYMP, ALP}은 리덕트특성을, 이들 특성들 사이에서 공통원소인 WBC 항목은 코어 특성을 나타내며, 정보 시스템에서 필요불가결한 요소를 의미한다.

본 연구에서는 두 군들 사이에 감별진단을 위한 중요 검사항목을 추출하기 위해서, 식 (5)에 의해서 생성된 식별가능행렬을 근거로 식 (6)의 식별가능함수 $DF(A)$을적용하였다.

$$DF(A)=\prod(\sum DM(c_{ij}) : c_{ij}\in U^2 \ and \ DM(c_{ij})\neq\varnothing) \quad (6)$$

$\sum DM(c_{ij})$ 는 식별가능원소 c_{ij}의 후보 속성들 간에 논리합(logical disjunction) 연산을, $\Pi(\cdot)$는 서로 다른 원소들 간의 논리곱(logical conjunction) 연산을 나타내고, 이러한 방법으로 생성된 식별가능함수는 논리곱 정규형(conjunctive normal form, CNF)의 형태를 가지며, 불 대수(Boolean algebra)의 2가지 일반 공리 즉 분배법칙(distribution law)과 흡수법칙(absorption law)에 의해서 간소화될 수 있다. 예를 들어, 표 2의 식별가능행렬은 식 (6)을 적용함으로써 다음과 같은 리덕트를 추출할 수 있다:

DF(A) = (WBC + PLT +ALP)·(WBC + PLT)·(WBC + ALP)·WBC·(WBC + PLT + LYMP)·(WBC + PLT + LYMP + ALP) = WBC

즉

①: (WBC + PLT + ALP)·(WBC + PLT) = WBC + PLT
②: (WBC + PLT)·WBC = WBC,
③: (WBC + ALP)·WBC = WBC,
④: (WBC + PLT + LYMP)·WBC = WBC,
⑤: (WBC + PLT + LYMP + ALP)·WBC = WBC

하지만 표 1에서 나타낸 정보 시스템에서 조건부 속성의 수가 증가할수록 생성 가능한 후보 리덕트들의 수도 증가하게 됨으로, 최적의 리덕트를 추출하기 위한 탐색과정은 비결정 난해(nondeterministic polynomial time hard, NP-hard) 문제를 가지게 된다. 따라서 본 연구에서는 이러한 문제점을 발견적인 전략(heuristic strategy)으로 해결하기 위해서 Johnson reducer 알고리즘[10]을 이용하여최소 길이를 가진 한 개의 리덕트 집합만을 추출하였고, 이를 감별진단을 위한 중요 검사항목으로 이용하였다.

(3) 지식생성

식별가능함수에 의해서 선택된 리덕트 (예, WBC)는 전체 조건부 속성을 대표할 수 있는 속성이므로, 이 리덕트속성을 이용하여 표 1의 정보 시스템을 표 3과 같이 수정될 수 있으며, 이를 근거로 다음의 3가지 지식을 생성하였다:

R1: If WBC is Upper abnormal Then DX is K80.0 (support: 3),

R2: If WBC is Normal Then DX is No (support: 1),

R3: If WBC is Lower abnormal Then DX is No (support: 1).

생성된 지식에서 'support'는 동일한 지식패턴을 가진 오브젝트들의 수를 나타내고, 지식의 'support' 값이 크면 클수록 임상적으로 아주 중요한 의미를 가진 지식일 가능성이 높다. 본 연구에서는 이러한 방법으로 10-fold 교차검증 동안에 각 fold의 훈련 데이터로부터 지식을 생성하였고, 실험 데이터의 출력값을 예측하기 위해서 2가지 방법 (절대빈도와 상대빈도)을 이용하였다.

만약 훈련 데이터로부터 결정된 중요 검사항목들과 이들의 참고치를 기준으로 임의의 실험 데이터를 명목화 하였을 때, 다음과 같은 지식 패턴이 생성되었다고 가정하자.

Test_R: If WBC is Upper abnormal Then Target_DX is K80.0,

여기서 'Target_DX'는 실험 데이터의 목표 출력값 (즉 출력 레이블) 의미한다.

훈련 데이터로부터 생성된 3가지 지식패턴과 비교해 볼 때, 대응된 지식의 절대빈도를 기준으로 출력값을 예측한 경우, Test_R은 R1의 지식에만 대응됨으로 실제 출력값은 K80.0

표 3. 수정된 정보 시스템의 예
Table 3. An example of modified information system

Objects (Cases)	Conditional attribute	Decision attribute
	WBC	DX
1	Upper abnormal	K80.0
2	Upper abnormal	K80.0
3	Normal	No
4	Upper abnormal	K80.0
5	Lower abnormal	No

표 4. 10-fold 교차검증 동안에 선택된 중요검사 항목 (등 간격 이산화 기반 특징선택)
Table 4. Significance clinical tests selected during 10-fold cross validation (equal width discretization based feature selection)

Selected clinical tests	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10	Freq.[a]
Chief Complaints	●	●	●	●	●	●	●	●	●	●	10
U_O.B		●	●	●	●	●	●	●	●	●	9
U_Glucose	●		●	●	●	●	●	●	●	●	9
U_Ketone			●								1
U_Urobilinogen	●	●	●	●	●	●	●	●	●	●	10
U_Bilirubin	●	●	●	●	●	●	●		●	●	10
U_WBC1		●		●				●		●	4
U_RBC	●	●	●	●	●	●	●	●	●	●	10
U_WBC2	●		●	●	●	●	●	●	●	●	9
U_Ep. Cell	●	●	●	●	●	●	●	●	●	●	10
U_Other								●			1
U_Crystal			●					●			2
CBC_WBC	●	●	●	●	●	●	●	●	●	●	10
CBC_PLT	●	●	●	●	●	●	●	●	●	●	10
CBC_NEUT								●			1
CBC_LYMP	●	●	●	●	●	●	●	●	●	●	10
CBC_MONO	●	●	●	●	●	●	●	●	●	●	10
CBC_LUC	●										1
PT	●	●	●		●	●	●		●		7
R/A_Inorganic Phosphorus	●	●	●	●	●	●	●	●	●	●	10
R/A_Glucose	●	●	●	●	●	●	●	●	●	●	10
R/A_Cholesterol(T)	●	●	●	●	●	●	●	●	●	●	10
R/A_ALP	●	●	●	●	●	●	●	●	●	●	10
Num. of clinicaltests	17	16	19	17	17	17	17	19	17	17	173

[a]Freq.: Frequency,
O.B.: Occult Blood, WBC1: White Blood Cell1, RBC: Red Blood Cell, WBC2: White Blood Cell2, Ep.Cell: Epithelial Cell, WBC: White Blood Cell, PLT: Platelet Count, NEUT: Neutrophil, LYMP: Lymphocyte, MONO: Monocyte, LUC: Large Unstained Cell, PT: Prothrombin Time, ALP: Alkaline Phosphatase

이 되고, 이때의 각 출력 레이블 K80.0과 No의 빈도수는 각각 1과 0이 된다. 반면에 대응된 지식의 상대빈도를 기준으로 예측한 경우, Test_R의 실제 출력값은 K80.0이 되고, 이때의 각 출력 레이블 K80.0과 No의 빈도수는 각각 3과 0이 된다.

(4) 성능평가

각각의 이산화 방법과 러프 근사화 개념을 결합한 예측모형의 성능을 평가하기 위해서, 식 (7)과 (8)에서 정의된 4가지 성능평가 척도, 정확도(accuracy), 민감도(sensitivity), 특이도(specificity) 및 기하평균(geometric mean)을 이용하였다. 특히 기하평균은 클래스 불균형 분포를 가진 데이터-셋의 분류 및 예측성능을 평가하기 위한 척도로 사용되었다[11,12].

$$\text{Accuracy} = (TP + TN)/(TP + FN + TN + FP) \quad (7)$$

$$\text{Geometric mean} = \sqrt{\text{sensitivity} \times \text{specificity}},$$
$$\text{where sensitivity} = TP/(TP+FN),$$
$$\text{specificity} = TN/(TN+FP) \quad (8)$$

식 (7)에서 *TP*, *FN*, *TN*, *FP*는 각각 true positive, false negative, true negative, false positive를 의미하고, 식 (8)에서 민감도와 특이도는 각각 true positive rate와 true negative rate를 나타낸다.

III. 실험 결과

1. 등 간격 이산화 기반 러프 근사화

표 4은 10-fold 교차검증 동안에 등 간격 이산화 방법으로 특징공간을 3개의 부분공간으로 분할한 후 식별가능행렬과 함수를 사용하였을 때, 각 fold에서 선택된 중요 검사항목들과 이때의 검사항목들의 발생빈도 수(Num. of clinical tests)를

표 5. 10가지 검사항목들의 참고치의 변화
Table 5. Variations of reference values for ten clinical tests (equal width discretization based feature selection)

Clinical tests	Unit	Reference		D Medical Center	
		Lower	Upper	Lower	Upper
U_Urobilinogen	E.U/dL	3.97	7.93	0.2	1.0
CBC_WBC	$\times 10^3$ μL	[8.72, 9.17]	[17.45, 18.33]	5.2	12.4
CBC_PLT	$\times 10^3$ μL	[185, 219]	[370, 438]	130	400
CBC_LYMP	%	[17.70, 18.43]	[35.40, 36.87]	19	48
CBC_MONO	%	5.23	10.47	3.4	9.0
PT	sec	[10.97, 12.73]	[21.93, 25.47]	10	14
R/A_Inorganic Phosphorus	mg/dL	[1.77, 2.57]	[3.53, 5.13]	2.5	4.5
R/A_Glucose	mg/dL	[153.33, 189.00]	[306.67, 378.00]	75	115
R/A_Cholesterol(T)	mg/dL	[116.67, 130.00]	[233.33, 260.00]	120	220
R/A_ALP	U/L	[201.67, 240.67]	[403.33, 481.33]	40	122

보여준다. 표 4에서 'U_'는 소변검사 항목, 'CBC_'는 전혈구 검사항목, 'R/A_'는 입원환자에 대한 기본검사를 나타낸다.

실험결과에서 볼 수 있듯이, 10-fold 교차검증 동안에 5회 이상 발생한 검사항목으로는 소변검사 항목에서 잠혈(occult blood, O.B.), 글루코스(glucose), 유로빌로노겐(urobilinogen), 빌리루빈(bilirubin), 적혈구 수(red blood cell, RBC), 백혈구 수(white blood cell2, WBC2), 상피세포(epithelial cell, Ep. Cell)이었고, 혈액검사에서는 백혈구 수(white blood cell, WBC), 혈소판 수(platelet count, PLT), 림프구(lymphocyte, LYMP), 단핵구(monocyte, MONO), 프로트롬빈 시간(prothrombin time, PT), 무기 인(inorganic phosphorus), 글루코스(glucose), 총 콜레스테롤(total cholesterol, Cholesterol(T)), 알카라인포스파타제(alkaline phosphatase, ALP)와 주증상으로 나타났다. 이들 항목들 가운데 발생빈도 수가 높은 항목일수록 담낭 및 담석증 질환을 가진 환자군을 진단하는데 있어서 중요한 검사항목일 가능성이 크므로, 교차검증 동안에 이들 17가지 검사항목들 가운데 연속형 속성값을 가지는 10가지 항목들의 참고치의 변화를 분석하였다.

표 5의 실험 결과에서 볼 수 있듯이 소변검사에서는 유로빌로노겐, 혈액검사에서는 단핵구가 교차검증 동안에 참고치의 하한과 상한값에서 일정한 결과값를 제공하였다. 표 6의 성능평가 결과에서는 10-fold 교차검증 동안에 각 fold의 훈련 데이터로부터 생성된 지식을 바탕으로, 실험 데이터의 두 그룹(담낭 및 담석증 질환군과 별 다른 특이점이 없어 퇴원한 환자군)에 대응된 지식의 절대빈도를 기준으로 예측한 방법이 상대빈도를 기준으로 예측한 방법에 비해 평균 예측 정확성 (평균 2.18% 향상)과 특이도 (평균 10.22%)에서 보다 좋은 결과를 제공하였다. 반면에 상대빈도를 기준으로 예측한 방법의 경우에는 절대빈도를 기준으로 예측한 방법에 비해 보다 좋은 평균 민감도 (평균 25.73% 향상)와 기하평균 (평균 20.03% 향상)을 보여주었다. 그리고 'Avg. number of clinical tests'는 표 4에서 10-fold 교차검증 동안에 각 fold에서 선택된 중요 검사항목들의 평균 수를 의미하고, 'Avg. number of generated knowledge'는 II-3의 지식생성 방법을 이용하였을 때, 각 fold의 훈련 데이터로부터 생성된 지식의 평균 수를 나타낸다.

표 6. 10-fold 교차검증 동안에 4가지 성능 평가 척도의 비교 결과
Table 6. Comparison results of four performance evaluation criteria during 10-fold cross validation (equal width discretization based feature selection)

Performance	Absolute (%)	Relative (%)
Avg. accuracy	77.93 ± 5.65	75.75 ± 8.91
Avg. sensitivity	24.91 ± 15.91	50.64 ± 13.26
Avg. specificity	93.19 ± 4.97	82.97 ± 10.69
Avg. geometric mean	44.20 ± 20.37	64.23 ± 9.27
Avg. number of clinical tests	17.3	
Avg. number of generated knowledge	205.6	

2. 등 빈번 이산화 기반 러프 근사화

이전 실험에서와 동일한 방법으로 10-fold 교차검증 동안에 각 fold에서 선택된 중요 검사항목들과 5회 이상의 발생빈도를 나타낸 항목들의 참고치와 4가지 성능 평가결과를 비교하였다. 표 7의 결과에서처럼, 주증상을 포함한 전체 26가지 항목들 가운데 8가지 항목이 교차검증 동안에 가장 많은 발생빈도를 보였으며, 이들 가운데 주증상, 백혈구 수, 활성화 부분 트롬보플라스틴 시간과 알카라인포스파타제가 가장 두드러진 특징을 보였다. 또한 이들 8가지 검사항목들 중에서 연속형 속성값을 가지는 6가지 항목들의 참고치의 변화를 분석

표 7. 10-fold 교차검증 동안에 선택된 중요검사 항목
Table 7. Significance clinical tests selected during 10-fold cross validation (equal frequency discretization based feature selection)

Selected clinical tests	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10	Freq.[a]
Chief Complaints	●	●	●	●	●	●	●	●	●	●	10
U_Ketone										●	1
U_pH	●		●	●	●				●	●	6
U_S.G.				●		●					2
U_Glucose								●	●		2
U_Urobilinogen					●						1
U_WBC1								●			1
U_RBC	●		●			●	●		●	●	6
U_WBC2								●			1
U_Ep.Cell	●				●						2
CBC_WBC	●	●	●	●	●	●	●		●	●	9
CBC_RBC				●				●		●	3
CBC_HGB	●	●									2
CBC_MCV					●	●	●		●		4
CBC_HCT			●		●		●				3
CBC_PLT	●	●	●	●			●	●		●	7
CBC_NEUT				●				●			2
CBC_MPV	●										1
CBC_LUC						●					1
PT		●									1
APTT	●	●	●	●	●	●	●	●	●	●	10
S/E_Cl						●			●		2
R/A_Glucose					●	●					2
R/A_Cholesterol(T)		●									1
R/A_ALP	●	●	●	●	●	●	●	●	●	●	10
Amylase		●	●	●			●	●	●	●	7
Num. of clinical tests	10	9	9	10	10	10	9	10	10	10	97

[a]Freq.: Frequency
S.G.: Specific Gravity, WBC1: White Blood Cell 1, RBC: Red Blood Cell, WBC2: White Blood Cell 2, Ep. Cell: Epithelial Cell, WBC: White Blood Cell, HGB: Hemoglobin Concentration, MCV: Mean Corpuscular Volume, HCT: Hematocrit, PLT: Platelet Count, NEUT: Neutrophil, MPV: Mean Platelet Volume, LUC: Large Unstained Cell, PT: Prothrombin Time, APTT: Activated Partial Thromboplastin Time, Cl: Chloride, ALP: Alkaline Phosphatase

해 본 결과, 표 8에서 제시된 것처럼 소변검사 항목에서 산도(pH)가 참고치의 하한과 상한값에서 일정한 결과값을 보여주었다.

표 7에서 제시된 각 fold의 중요 검사항목들로부터 감별진단 지식을 생성하였으며, 그지식을 근거로 예측성능을 분석한 결과, 상대빈도를 기반으로 한 예측방법이 절대빈도를 기반으로 한 예측방법에 비해서 예측 정확성, 민감도, 기하평균에서 각각 평균적으로 2.67%, 23.73%, 21.39% 향상된 결과를 제공하였다. 또한 각각의 이산화 방법을 기반으로 러프 근사화 모형을 설계하였을 때, 절대빈도를 기준으로 예측한 경우 모든 평가 척도에서 등 빈번 이산화 방법을 기반으로 한 방법이 다소 낮은 성능을 보였고, 상대빈도를 기준으로 예측한 경우 평균 예측 정확성과 특이도에서 다소 높은 성능을 제공하였으며, 기하평균에서는 2가지 방법 모두가 유사한 성능을 보였다 (표 6과 표 9 참조).

3. 담낭 및 담석증 진단을 위한 중요 검사항목과 참조치의 비교

급성 담낭염에 대한 일반적인 임상적 소견으로는 백혈구 수가 증가하고, 담관의 폐쇄를 동반하지 않아도 경미한 혈청 아미노 전이 효소(transaminase), 알카라인포스파타제, 아밀라제, 리파아제의 상승 소견이 있을 수 있으며, 2~4 mg/dL 정도의 총 빌리루빈 상승도 드물지 않게 나타나며, 총 빌리루빈

표 8. 6가지 검사항목들의 참고치의 변화
Table 8. Variations of reference values for six clinical tests (equal frequency discretization based feature selection)

Clinical test	Unit	Reference		D Medical Center	
		Lower	Upper	Lower	Upper
U_pH	-	6	7	5.0	9.0
CBC_WBC	$\times 10^3\,\mu L$	[7.87, 8.02]	[11.92, 12.27]	5.2	12.4
CBC_PLT	$\times 10^3\,\mu L$	[210, 216]	[297, 306]	130	400
APTT	Sec	[26.8, 27.2]	[33.2, 33.6]	20	38
R/A_ALP	U/L	[84, 89]	[151,161]	40	122
Amylase	U/L	[47, 49]	[81, 84]	44	128

표 9. 10-fold 교차검증 동안에 4가지 성능 평가 척도의 비교 결과
Table 9. Comparison results of four performance evaluation criteria during 10-fold cross validation (equal frequency discretization based feature selection)

Performance	Absolute (%)	Relative (%)
Avg. accuracy	76.39 ± 6.68	79.06 ± 7.41
Avg. sensitivity	23.82 ± 19.02	47.55 ± 16.19
Avg. specificity	91.50 ± 6.53	88.10 ± 8.75
Avg. geometric mean	42.30 ± 20.43	63.69 ± 11.44
Avg. number of clinical tests	9.7	
Avg. number of generated knowledge	206.7	

의 수치가 4 mg/dL 이상으로 증가하거나 아밀라제가 현저하게 상승되어 있으면 담관 / 담석의 동반 여부와 함께 담석성 췌장염에 대한 조사가 필요하다[13]. 이러한 점으로 볼 때, 백혈구 수, 아미노 전이 효소, 총 빌리루빈, 알카라인포스파타제, 아밀라제 및 리파제의 검사항목들과 이들 검사상의 수치는 담낭 및 담석증을 진단하는데 있어서 중요한 요인이라 할 수 있다.

이러한 임상적 견해와 비교해 볼 때, 등 간격 이산화를 기반으로 한 러프 근사화 방법에서는 백혈구 수와 알카라인포스파타제가 중요 검사항목으로 포함되어 있으며(표 5 참조), 등 빈번 이산화를 기반으로 한 러프 근사화 방법에서는 백혈구 수, 알카라인포스파타제 및 아밀라제가 중요 검사항목(표 8 참조)으로 포함되어 있다. 그러나 참고치의 변화 결과에서, 등 간격 이산화 방법은 백혈구 수의 경우 상한 값에서, 알카라인포스파타제의 경우 하한과 상한값 모두에서 D의료원 참고치(표 5 참조)와 비교해 볼 때 현저히 높게 결정되었다. 반면에 등 빈번 이산화 방법은 D 의료원의 참고치와 비교해 볼 때, 알카라인포스파타제의 하한값에서 상대적으로 낮게 결정되었고, 백혈구 수와 아밀라제의 경우 임상적으로 신뢰할 수 있는 참고치를 제공하였으며, 다른 임상 검사항목들도 등 간격 이산화 방법에서 제공된 참고치에 비해 좀 더 타당한 결과를 제공하였다.

IV. 결 론

본 연구에서는 클래스 불균형 분포를 가진 고차원의 임상 데이터로부터 질환의 감별진단 시 중요한 검사항목들과 임상적 참고치를 추출할 수 있는 데이터 이산화를 기반으로 한 러프 근사화 방법을 제안하였다. 제안된 방법에서 데이터 이산화 기법은 고차원의 임상 데이터의 분할 경계구간을 결정하기 위한 방법으로 사용되었고, 러프집합의 식별가능행렬과 식별가능함수에 대한 근사화 개념은 명목형으로 변환된 데이터로부터 감별진단 시 중요한 검사항목을 추출하기 위해서 이용되었다. 또한 선택된 중요 검사항목들을 근거로 임상적 판단 지식을 생성할 수 있는 방법에 대해서도 논의하였다.

실험에서는 제안된 방법의 타당성을 보이기 위해서, D광역시에 소재한 D의료원에 2006년 7월에서 2007년 6월 사이에 복통을 주증상으로 응급실에 내원한 환자 1,103명의 의무기록 중에서 담낭 및 담석증 질환으로 진단된 101명의 환자와 검사소견 상 별 다른 특이점이 없어 퇴원된 환자 351명을 대상으로 하였고, 이들 대상환자들의 주증상과 52가지 임상검사 결과들을 분석하였다. 실험결과, 등 간격과 등 빈번 이산화를 기반으로 한 러프 근사화 방법 모두가 실험 데이터를 예측할 때, 절대빈도를 고려한 방법에 비해 상대빈도의 개념을 고려한 방법이 보다 좋은 기하평균 성능을 제공하였다. 그리고 이들 각각의 방법으로부터 선택된 중요 검사항목과 참고치의 임상적 타당성을 평가하기 위해서 참고문헌[13]에서 보고된 결과와 D의료원의 참고치를 비교하였다. 그 결과 등 빈번 이산화를 기반으로 한 근사화 방법이 보다 신뢰할 수 있는 참고치의 결과를 제공하였다.

본 연구에서 논의된 임상검사의 참조치 추출방법은 검사 시약이나 의료기기를 공급하는 회사에서 제시하는 기준을 검사실 단위에서 재산출하는 업무과정을 개선하는데 효과적으로 적용될 수 있으며, 중요 검사항목 선택방법은 질환 진단 혹은 질환의 예후인자를 추출하는 데에 중요한 기술로 사용될 수 있을 것으로 판단된다. 향후 연구에서는 통계적 기법을 기반으로 한 ChiMerge[14]와 Chi2Merge[15], 엔트로피[16], 클

러스터링[17]기법으로부터 얻은 참고치의 비교뿐만 아니라, 클래스 불균형 분포의 비율이 높은 데이터-셋으로부터 중요 검사항목과 참고치를 결정하는 문제로의 확장된 연구가 필요할 것으로 판단된다.

참고문헌

[1] K.N. Lee, J.H. Yoon, Y.H. Choi, H.I. Cho, K.W. Bae, C.H. Yoon, and S.I. Kim, "Standardization of reference values among laboratories of Korean association of health promotion," *J. Lab. Med. & Quality Assuarance*, vol. 24, no. 2, pp. 185-195, 2002.

[2] 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 results analysis system application to liver function test," *J. Biomed. Eng. Res.*, vol. 14, no. 4, pp. 341-348, 1993.

[3] C.S. Son, A.M. Shin, Y.D. Lee, H.J. Park, H.S. Park, and Y.N. Kim, "Variable threshold based feature selection using spatial distribution of data," *J. Kor. Soc. Med. Informatics*, vol. 15, no. 4, pp. 475-481, 2009.

[4] C.S. Son, A.M. Shin, Y.D. Lee, H.S. Park, H.J. Park, and Y.N. Kim, "Rule weight-based fuzzy classification model for analyzing admission-discharge of dyspnea patients," *J. Biomed. Eng. Res.*, vol. 31, no. 1, pp.40-49, 2010.

[5] ICD10 version 2007, http://apps.who.int/classifications/apps/icd/icd10online/

[6] D. Chiu, A. Wong, and B. Cheung, *Information discovery through hierarchical maximum entropy discretization and synthesis*, MIT Press, 1991.

[7] Z. Pawlak, "Rough sets," *Int. J. Comput. Inf. Sci.*, vol. 11, no. 5, pp. 341-356, 1982.

[8] R. Slowinski and J. Stefanowski, "Rough classification in incomplete information systems," *Math.Comput.Modeling*, vol. 12, no. 10-11, pp. 1347-1357, 1989.

[9] Z. Pawlak, *Rough sets: theoretical aspects of reasoning about data*, Kluwer Academic Publisher, Dordrecht, Netherlands, 1991.

[10] R. Jensen and Q. Shen, *Computational intelligence and feature selection: rough and fuzzy approaches*, Wiley-IEEE Press, 2008.

[11] Y.M. Sun, M.S. Kamel, A.K.C. Wong, and Y.Wang, "Cost-sensitive boosting for classification of imbalanced data," *Patt.Recog.*, vol. 40, no. 12, pp. 3358-3378, 2007.

[12] C.S. Son, A.M. Shin, I.H. Lee, H.J. Park, H.S. Park, and Y.N. Kim, "Fuzzy discretization with spatial distribution of data and its application to feature selection,"*J. Kor. Inst. Int. Syst.*, vol. 20, no. 2, pp. 165-172, 2010.

[13] K.S. Yoo, "Diagnosis of gallstone," *Korean J. Med.*, vol. 75, no. 6, pp. 616-623, 2008.

[14] R. Kerber, "ChiMerge: discretization of numeric attributes," *in Proceedings of AAAI-92, Ninth Intpppppl Conf., Artificial Intelligence*, AAAI-Press, pp. 123-128, 1992.

[15] H. Liu and R. Setiono, "Feature selection via discretization of numeric attributes," *IEEE Trans. Knowl.Data Eng.*, vol. 9, no. 4, pp. 642-645, 1997.

[16] U.M. Fayyad and K.B. Irani, "Multi-interval discretization of continuous attributes as preprocessing for classification learning," *in Proceedings of 13th International Joint Conference on Artificial Intelligence*, pp.1022-1027, 1993.

[17] L. Kaufman and P.J. Rousseeuw, "Finding group in data: an introduction to cluster analysis, *John Wiley & Sons*, New York, 1990.