DOI: 10.3745/KIPSTD.2009.16-D.1.11

심혈관계 질환 진단을 위한 복합 진단 지표와 출현 패턴 기반의 분류 기법

이 헌 규[†] · 노 기 용[††] · 류 근 호[†††] · 정 두 영[††††]

요 약

심혈관계 질환의 진단 위해서 복합 진단 지표를 이용한 출현 패턴 기반의 분류 기법을 제안하였다. 복합 진단 지표 적용을 위해서 심박동변이도의 선형/비선형적 특징들을 세 가지 누운 자세에 대해 분석하였고 ST-segments로부터 4개의 진단 지표를 추출하였다. 이 논문에서는 질환 진단을 위해서 필수 출현 패턴을 이용한 분류 모델을 제안하였다. 이 분류 기법은 환자 그룹의 질환 패턴들을 발견하며, 이러한 출현 패턴은 심혈관계 질환 환자들에서는 빈발하지만 정상인 그룹에서는 빈발하지 않는 패턴들이다. 제안된 분류 알고리즘의 평가를 위해서 120명의 협심증(AP: angina pectrois) 환자, 13명의 급성관상동맥증후군(ACS: acute coronary syndrome) 환자 그리고 128명의 정상인 데이터를 사용하였다. 실험 결과 복합 지표를 사용하였을 때, 세 그룹의 분류에 대한 정확도는 약 88.3%였다.

키워드 : 심혈관계 질환, 출현 패턴 마이닝, 분류, 심박동변이도, ST-segments

Multi-parametric Diagnosis Indexes and Emerging Pattern based Classification Technique for Diagnosing Cardiovascular Disease

Heon Gyu Lee[†] · Ki Yong Noh[††] · Keun Ho Ryu[†††] · Doo-Young Jung[††††]

ABSTRACT

In order to diagnose cardiovascular disease, we proposed EP-based(emerging pattern- based) classification technique using multi-parametric diagnosis indexes. We analyzed linear/nonlinear features of HRV for three recumbent postures and extracted four diagnosis indexes from ST-segments to apply the multi-parametric diagnosis indexes. In this paper, classification model using essential emerging patterns for diagnosing disease was applied. This classification technique discovers disease patterns of patient group and these emerging patterns are frequent in patients with cardiovascular disease but are not frequent in the normal group. To evaluate proposed classification algorithm, 120 patients with AP (angina pectrois), 13 patients with ACS(acute coronary syndrome) and 128 normal people data were used. As a result of classification, when multi-parametric indexes were used, the percent accuracy in classifying three groups was turned out to be about 88.3%.

Keywords : Cardiovascular Disease, Emerging Pattern Mining, Classification, Heart Rate Variability, ST-segments

1. 서 론

최근 심혈관계 질환의 진료 수요와 심혈관계 질환에 의한 한국인의 사망자 수가 급격히 증가함에 따라 조기 진단 및 진단의 신뢰성은 사회적으로 매우 중요한 문제로 인식되고 있다. 2006년 한국인 사망 및 그 원인 통계 결과에 따르면 [1] 심뇌혈관질환(심장질환 및 뇌혈관질환)에 의한 사망자 수는 총 243,934명의 사망자 중 66,594명으로 27.3%를 차지하며 10년 전에 비해 2배 이상 급증한 것이고, 동맥경화로 인한 혈관 질환(뇌경색, 뇌졸중 등)을 포함할 경우 10배 증가 되었다. 서구화된 한국의 식생활 패턴이 심혈관계 질환의 주요 원인으로 보고되고 있으며, 특히 젊은 연령 층의 환자가 현저하게 증가하고 있다. 심혈관계 질환의 조기 발견을 위해 심전도 (ECG: electrocardiogram)는 심장의 상태를 비관혈적으로 진단하는 매우 중요한 생체신호 중의 하나이며 원시 신호로부터 ST-segment와 심박동변이도의 선형, 비선형 특징(feature) 등과 같은 다양한 진단 지표들을 추출할 수 있다. 심혈관계 질환 중 심근허혈은 심장 근육에 일시적으로 산소가 부족하여 심장이 활동하지 못하는 것이며 허혈의 대표적인 증상이 협심증(AP: angina pectrois)이다. 심근허혈은 심장의 관상동맥 중 일부에 협착이 생겨 혈관이

† 정 회 원 : 한국전자통신연구원 우정기술센터 연구원
†† 정 회 원 : 한국표준과학연구원 선임연구원
††† 종신회원 : 충북대학교 전기전자컴퓨터공학부 교수(교신저자)
†††† 정 회 원 : 충북대학교 전기전자컴퓨터공학부 교수

논문접수 : 2008년 9월 9일
수 정 일 : 1차 2008년 11월 25일, 2차 2009년 1월 12일
심사완료 : 2009년 1월 13일

좁아지는 관상동맥질환(CAD: coronary artery disease)이 주요인이며, 이러한 심근허혈의 진단 지표로 심전도 신호의 ST-segments가 사용되는데, segments의 상승 또는 하강 여부를 조사함으로서 질환을 진단한다. 예를 들어 급성관동맥증후군이 의심되는 환자의 경우, 흉통과 함께 ST-segment의 상승이 있으면 심근경색으로 진단 한다[2, 3]. 한편 ST-segments와 상호보완적으로 심박동변이도의 다양한 선형/비선형적 특징이 진단 정확성을 높이기 위한 지표로 연구되고 있다[4, 5]. 심혈관계의 동적 안정성은 내적 또는 외적 자극에 심장박동 등이 재빨리 반응함으로서 자동적으로 조절되어 달성되며 심장박동은 이러한 자극에 복합적 방식으로 반응하여 변화하는데 교감신경계와 부교감신경계의 두 자율적 시스템에 의해 집중적으로 자극 된다. 부교감신경계의 활성화는 심장박동을 느리게 하고 교감신경계 활성화는 수축성의 증가와 함께 심장박동의 속도를 증가시킨다. 신경전달체 기능의 이러한 차이에 의해 자율신경계의 두 시스템은 서로 다른 주파수로 동작하며, 심장박동의 변동이 교감신계 변화에 지배적인지 부교감신경계에 관련되어 있는지 알 수 있도록 해준다[6]. 이와 같은 심장박동의 변동성을 심박동변이도(HRV: heart rate variability)라 하며, 교감 및 부교감 신경의 활성도를 심박동변이도의 선형적 비선형적 특징 분석을 통해 양적으로 평가할 수 있다[4]. 예를 들어 관상동맥질환을 가진 환자의 심박동변이도를 분석해 보면 자율신경계의 조절작용이 저하되며[4], 급성 심근경색의 경우 사망 위험도는 자율신경계가 활발히 작용할 때 감소된다[7]. 따라서 ST-segments와 함께 심박동변이도는 심혈관계 질환의 중요한 진단 지표가 될 수 있다. 또한 심전도 측정 시에 누운 자세 변화가 교감 및 부교감 신경계의 활성도를 나타내는 심박동변이도 특징들에 유의한 영향을 미치는 것으로 보고되고 있다[8], [9]. 즉, 중증의 심혈관계 질환 환자 일수록 똑바로 눕는 것은 떨어진 부교감 신경계의 활성도를 더욱 떨어뜨려 악화시킨다. 반면에 우측으로 누운 자세에서는 부교감 신경계의 활성도가 증가하여 손상된 부교감신경계의 활성도를 회복시킬 수 있게 해준다. 따라서 심혈관계 질환(협심증, 급성관동맥증후군) 진단을 위한 복합 지표로서 ST-segment의 파라미터들과 세 가지 누운 자세에서 분석된 심박동변이도의 선형, 비선형적 특징들을 모두 적용한다.

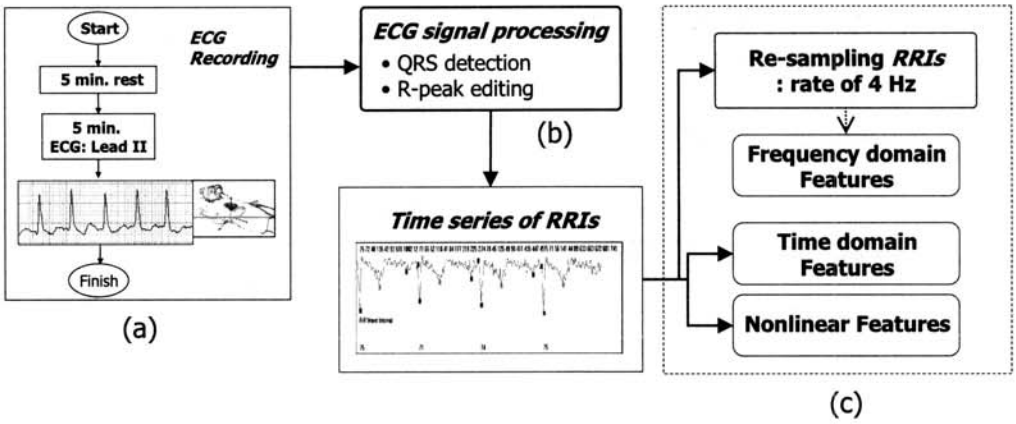
최근 신경망[10], 유전자 알고리즘[11] 및 규칙-기반 분류모델[12]을 이용한 심근허혈 심전도 신호를 분류한 연구가 진행되었다. 그러나 기존 연구들에서는 심근허혈과 정상인군에 대한 두 클래스 분류 결과만을 평가하였고, PhysioNet[13]의 ST-T database의 ST-segments만을 진단 지표로 사용하여 협심증이나 급성관동맥증후군과 같은 세부 관상동맥 질환의 진단에 적용할 수 없는 한계를 가지고 있다. 또한 [4]에서는 이 연구에 앞서 처음으로 심박동변이도 분석을 통해 선형/비선형적 특징을 추출하고 고차원 데이터의 분류에 적합한 출현 패턴 마이닝 개념을 적용하여 심혈관계 질환 진단을 위한 분류 기법을 적용하였다. 그러나 단순한 CAEP[14] 알고리즘을 적용함으로서 이 알고리즘이 가진 불필요한 많은 중복된 패턴을 생성한다는 것과 클래스 분류에 악영향을 미치는 잡음(noisy) 패턴을 그대로 분류 모델에 사용하는 단점을 극복하지 못하였고, 단지 정상인과 관상동맥 질환자의 두 그룹만을 분류하는 데 그쳤다. 따라서 이 논문에서는 자동화된 심혈관계 질환 진단 방법으로서 첫째, ST-segments와 심박동변이도 특징을 포함한 복합적인 진단 지표들을 처음으로 분류 모델에 적용한다. 둘째, 기존 분류 기법들에 비해 더 정확하고 효율적인 분류 모델 생성을 목표로 하며, 이를 위해서 분류 모델 생성에 필수적인 출현 패턴(EPs: emerging patterns) 만을 적용하기 위한 트리 구조의 알고리즘과 통계적 기법의 중복 및 잡음 패턴 제거 알고리즘을 제안한다. 심혈관계 질환 진단을 위한 출현 패턴 발견 및 분류 기법에 대한 전체 수행 단계는 다음과 같다.

(1) 복합 진단 지표 추출 : 세 가지 누운 자세에 대한 심전도 신호로부터 ST-segment의 4가지 특징 벡터(*J* point, *J80* point, *Slope*, *Area*)를 추출하고, 각 자세에 대한 심박동변이도 분석을 수행하여 선형적, 비선형적 특징들을 추출한다.
(2) 데이터 이산화 및 특징 벡터 선택 : 연속적인 실수값인 진단 지표에 대한 이산화 작업 및 p-value를 이용한 목표 클래스에 대한 특징 벡터 선택을 수행한다.
(3) 심혈관계 질환 진단을 위한 출현 패턴 분류 : 목표 클래스에 대해 높은 발생빈도를 갖는 출현 패턴들을 발견하고 동시에 효율적인 패턴 발견 및 분류 모델 생성 시 필요한 패턴의 빠른 접근을 위한 트리 구조의 출현 패턴 탐사 알고리즘을 제안한다. 또한 불필요한 패턴 및 중복 패턴 제거를 위한 통계적 유의성 검정 테스트인 chi-square(χ^2) 기법을 적용한다.
(4) 분류 모델의 평가 : 261명의 심전도 데이터를 이용하여 분류기를 생성하고 교차검증을 수행, 제안 모델의 성능을 평가하며, 각 자세별 독립적인 진단 지표 적용과 제안된 복합 진단 지표와의 진단 정확성 및 유용성을 비교한다.

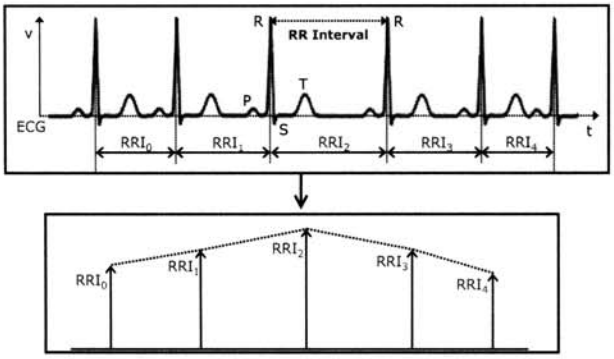
논문의 효과적인 이해를 위해서 논문의 구성은 다음과 같이 구성하였다. 2장에서는 심전도 데이터로부터 복합 진단 지표 추출과정을 기술하고 3장에서는 데이터 전처리 단계로서 연속적 실수 값의 이산화와 특징 벡터 선택 과정을 설명한다. 4장에서는 기존 출현 패턴 마이닝 기법을 확장한 분류기 생성 알고리즘을 기술한다. 제안한 질환 진단 분류 기법에 대한 실험 및 결과 분석은 5장에 기술한다. 마지막으로 6장에서는 이 논문에 대한 전체적인 결론을 맺는다.

2. 복합 진단 지표의 추출

이 장에서는 협심증, 급성관상동맥증후군과 같은 심혈관계 질환의 진단 지표인 ST-segments와 HRV의 선형 및 비선형 지표 추출 과정을 기술하며 (그림 1)의 단계를 가진다.



(그림 1) 복합 진단 지표 (심박동변이도, ST-segments) 추출 과정



(그림 2) *RRI* 시계열 데이터 생성

2.1 ECG 데이터 분석

심전도 신호의 R-peak 및 QRS complex 검출 과정(그림 1-(b))을 위하여 이 논문에서는 심전계를 이용하여 심장질환 환자의 심전도를 측정하였다. (그림 1-(a))와 세 가지 누운 자세 (반듯이 누운 자세, 오른쪽으로 누운 자세, 왼쪽으로 누운 자세)에 대해서 Lead II 채널을 이용하여 각각 5분 동안의 심전도를 측정한다. 첫 번째 자세에 대한 심전도 측정 후에 환자는 5분의 휴식을 취한 뒤, 다음 자세에 대해 심전도를 측정하였다. 이렇게 측정된 심전도 신호의 샘플링 주파수는 500Hz이며, ectopic beats와 artifacts는 제거된다. ST-segments는 심전도 파형에서 QRS complex를 검출함으로써 추출할 수 있다. 심박동변이도 분석을 위해서는 Thomkin's 알고리즘[15]을 이용하여 심전도 신호의 모든 *R*파 사이의 간격(*RRIs*: *RR* intervals)을 구하며, (그림 2)와 같이 시계열 데이터를 생성한다. 또한 *RRIs* 시계열 데이터는 선형적 분석 방법인 주파수 영역에서의 지표 추출을 위해서 4Hz의 비율로 재 샘플링을 수행한다 (그림 1-(c)).

2.2 ST-segments 진단 지표 추출

QRS complex는 5~30Hz의 주파수 성분을 갖기 때문에 웨이블릿 특성을 이용하여 5~30Hz을 추출하여, 정확히 QRS를 검출한다(그림 1-(c)). QRS complex 검출 후 R-peak를 검출하여 ST-segment 시작점인 *J* point와 *J* point로부터 80ms 떨어진 지점인 *J80* point 추출하며, ST-segment의 *slope*와 *area*를 특징 벡터로 사용한다. 진단 지표로 추출된 4가지의 ST-segments는 (그림 3)과 같다[13].

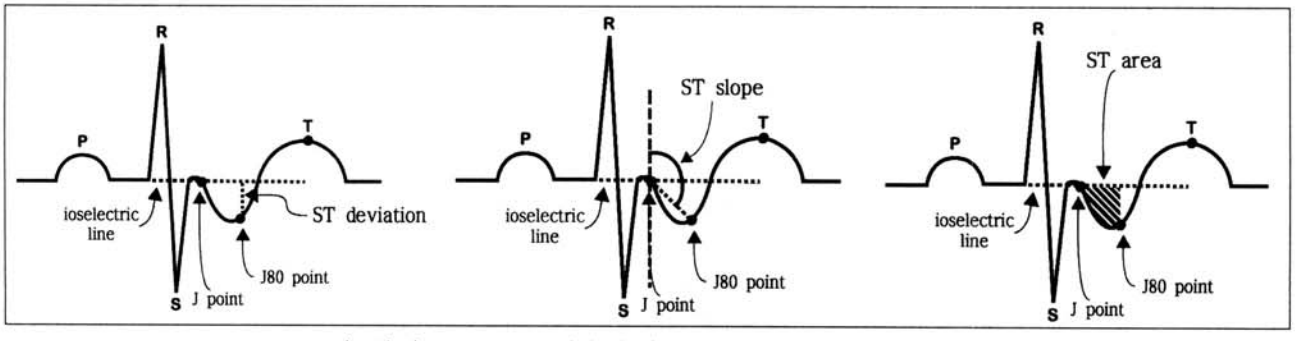
2.3 심박동변이도 분석을 통한 선형/비선형 진단 지표 추출

세 가지 누운 자세에 대한 *RRIs*로부터 주파수 및 시간 영역의 선형적 진단 지표와 HRV의 비선형적 특징의 진단 지표를 추출한다(그림 1-(c)).

2.3.1 시간 영역에서의 선형적 진단 지표 [8]

- *SDRR* [ms] : *RR* 간격의 표준편차
- *SDSD* [ms] : 인접한 *RR* 간격사이의 차이값의 표준편차
- *RRm* [ms] : 모든 *RR* 간격의 평균

2.3.2 전력스펙트럼밀도 (PSD: power spectral density)



(그림 3) ST-segment 진단 지표(J point, J80 point, Slope, Area)

분석을 통한 주파수 영역에서의 선형적 진단 지표[8, 16]

- *TP* (total power) [ms^2] : 0.4 Hz이하의 *RR* 간격들의 주파수 영역에서의 총 파워
- *VLF* (very low frequency) [ms^2] : 0.04 Hz 이하에서의 파워
- *LF* (low frequency) [ms^2] : 0.04 - 0.15 Hz 에서의 파워
- *nLF* [nu] : *LF*의 정규격화 값 ($nLF=\frac{LF-VLF}{TP}\times 100$)
- *HF* [ms^2] : 0.15 - 0.4 Hz 에서의 파워
- *nHF* [nu] : *HF*의 정규격화 값 ($nHF=\frac{HF-VLF}{TP}\times 100$)
- *LF/HF* : *LF/HF* 비율

7가지의 주파수 영역에서의 진단 지표 중에서 전체 자율신경계의 상대적인 교감신경계 활성률을 반영하는 *nLF*와 부교감신경계의 활성률을 반영하는 *nHF* 그리고 교감-부교감신경계의 균형을 반영하는 지표인 *LF/HF*만을 사용한다.

2.3.3 Poincare plot 비선형 지표[17]

재구성된 *RRIs* 시계열 데이터에 대해, *RRI(n)*을 *x*축으로 하고 *RRI(n+1)*을 *y* 좌표로 플롯팅하여 Poincare plot 비선형 지표를 추출한다. 평면상의 점(*x,y*)은 타원 형태로 분포되는데 두 축에 대한 표준편차(*SD*: standard deviation)를 이용한다. 타원의 폭 방향의 표준편차를 나타내는 *SD1*(width of poincare plot)과 타원의 길이 방향을 나타내는 *SD2*(length of poincare plot)를 구한다. *SD1*과 *SD2*의 관계를 나타내는 *SD1SD2*(*SD1*×*SD2*) 및 교감신경계 지표로서 *SD2*와 *SD1*의 비율인 *SD2/SD1*을 계산한다.

2.3.4 근사 엔트로피 비선형 지표[18]

근사 엔트로피는 신호의 복잡성을 나타내는 지수이다. 근사엔트로피는 그 값이 클수록 더 복잡함 내지는 불규칙성을 나타내며, 작은 값일수록 심혈관계 이환 및 병적 생체상태와 관련 있다. *N*개의 *RRIs* 시계열 데이터에서 근사엔트로피를 구하기 위해, 먼저 (식 1)과 같이 지연시간 τ, 매립차원(embedding dimension) *m*으로 매립시킬 수 있다.

$$\overrightarrow{S_i}=[s_i, s_{i+\tau}, \cdots, s_{i+(m-1)\tau}] \qquad \text{(식 1)}$$

매립된 임의의 벡터 점들 $\overrightarrow{S_i}$와 $\overrightarrow{S_j}$ 사이의 거리를 (식 2)와 같이 계산한다.

$$d[\overrightarrow{S_i}, \overrightarrow{S_j}]=\max_{k=1,\cdots,m}[|s_{i+(k-1)\tau}-s_{j+(k-1)\tau}|] \qquad \text{(식 2)}$$

여기서, 데이터 개수 $N\geq i+(m-1)\tau$ 이므로 i의 범위는 $i\leq N-(m-1)\tau$가 된다. 상태공간의 반경을 0에서 1 사이로 규격화 했을 때, 기준반경 r을 $1/r$ 등분하는 크기로 정의하면 기준점 i 각각에 대해서 반경 r 내에 있는 매립된 점의 밀도는 (식 3)으로 나타낼 수 있다.

$$C_i^m(r)=\frac{number\ of\ s_j\ such\ that\ d[\overrightarrow{S_i}, \overrightarrow{S_j}]}{N-m+1} \qquad \text{(식 3)}$$

또한, $\Phi^m(r)$을 $C_i^m(r)$의 log값 평균으로 정의하면 (식 4)와 같이 나타낼 수 있고,

$$\Phi^m(r)=\frac{1}{N-(m-1)\tau}\sum_{i=1}^{N-(m-1)\tau}\ln C_i^m(r) \qquad \text{(식 4)}$$

근사엔트로피 m과 r이 고정된 경우 (식 5)와 같이 정의될 수 있다.

$$ApEn(m, r)=\lim_{N\to\infty}[\Phi^m(r)-\Phi^{m+1}(r)] \qquad \text{(식 5)}$$

만약, 데이터 개수가 *N*개로 고정되는 경우에는 근사엔트로피는 (식 6)으로 정의된다.

$$ApEn(m, r, N)=\Phi^m(r)-\Phi^{m+1}(r) \qquad \text{(식 6)}$$

(식 6)에서 나타나는 것처럼 근사엔트로피는 *m* 차원 매립된 상태 벡터 점들이 *m+1* 차원에서 얼마만큼 벗어나는 정도를 *log* 값으로 나타낸 양으로 시계열이 규칙적이고, 결정론적이면 근사엔트로피는 작은 값을 가지고 불규칙적일 수록 큰 값을 가진다. 근사엔트로피 계산에서 가장 중요한 매개변수는 기준반경 *r*이다. 원칙적으로 무한한 양의 데이터에 대해서는 *r*은 0에 접근하지만, 유한한 양의 데이터의 경우에는 가장 적당한 값을 선택하는 명백한 기준이 없다. [19]에서는 기존의 심장 박동 변이에 적용할 때 시계열의 표준편차의 약 20% 크기로 설정되어야 함이 제안되었으므로 이 논문에서는 임상적인 적용에 적합한 m=2, r=0.2 · *SD*로 정하였다.

2.3.5 허스트 지수 분석(H: hurst exponent analysis)[20]

사람의 건강 상태 및 활동 상태는 인체에 여러 가지 현상들을 통하여 나타난다. 이들을 재구성함으로써 역으로 건강상태를 추정할 수 있게 된다. 즉 심장은 건강 상태에 따라 카오스적 특성이 다르게 나타나는데, 이러한 카오스 현상을 나타내는 심장의 활동 양상을 살펴봄으로서 건강 상태를 구분할 수 있는 것이다. 이에 대표적으로 카오스 현상을 측정할 때 사용하는 측정값이 바로 허스트 지수(H)이다[21]. 허스트 지수, H 값은 다음과 같이 구할 수 있다. 우선, (식 7)과 같이 R-peak의 시간 간격(*RRI*)을 구한다.

$$u(n)=RR(n)=t(R_{n+1})-t(R_n) \qquad \text{(식 7)}$$

다음, 이동평균값(running means)을 구하고, 이에 대한 현재 심박동의 누적편차(accumulated deviation)를 구함으로

써 심장의 동역학적 상태를 이전의 상태와 비교한다.

이동평균값 : $$\bar{u}(n) \equiv \frac{1}{n}\sum_{i=1}^{n} u(i) \quad \text{(식 8)}$$

누적편차 : $$X_{t,n} = \sum_{i=1}^{t} u(i) - \bar{u}(n) \quad \text{(식 9)}$$

이동표준편차($S(n)$:standard deviation)를 기준으로 그 증가하는 비율을 바라보게 되면, 시간이 지남에 따라 나타나는 경향성은 지수값(H)을 통해 나타나게 된다. 즉, 다음과 같이 표현되며, 누적편차의 최대치와 최소치의 차이가 된다.

$$R(n) = \max(X_{t,n}) - \min(X_{t,n}) \quad \text{(식 10)}$$

이 범위는 고려되는 시간에 따라 달라지므로 R을 최초 관측치들의 표준편차 $S(n)$로 나눔으로써 일반화되고, 허스트의 경험법칙에 의해서 (식 8)과 같이 $R(n)/S(n)$ 값이 추정되며, 이를 대수방정식으로 만들면 (식 12)와 같이 쓸 수 있다.

$$R(n)/S(n) = \alpha \times n^{H} \quad \text{(식 11)}$$

$$\log\frac{R(n)}{S(n)} = H\log n + \log\alpha \quad \text{(식 12)}$$

H<0.5라는 의미는 데이터 값들 간에 음의 연관성(negative correlated)을 갖고 있어서 반지속적인(anti-persistence)의 특성을 의미한다. 즉, 연속적으로 나타나는 데이터 값은 이전의 값과는 커다란 차이를 보이며, 이동평균값(running means)을 기준으로 계속적인 진동을 보이 있는 것을 의미한다. 우리가 주의해서 보아야 하는 값은 주로 H>0.5를 만족하는 범위의 값이다. 이는 시계열이 얼마만큼의 기억효과(memory effect)를 갖고, 장기간의 연관성(long-term correlation)을 갖고 있는가를 보는데 그 의미가 있다. 보통 건강한 정상인의 경우, 허스트 지수 값은 0.7 정도의 값을 보인다. 이는 random behavior를 보이는 0.5와 regular behavior를 보이는 1.0의 중간 값에 근접한 값으로서 카오스적인 특성을 보이고 있음을 의미한다.

이 논문에서 제안한 모든 진단 지표들을 <표 1>에 기술하였다.

3. 데이터 이산화 및 특징 선택

심혈관계 질환 진단을 위한 모든 진단 지표는 연속적 실수 값을 가지므로 해당 클래스 별 출현 패턴 탐사가 가능하도록 데이터 이산화가 선행되어야 한다. 이 절에서는 엔트로피 기반 데이터 이산화 및 *p-value*를 적용한 필수 진단 지표 선택 과정을 기술한다.

3.1 데이터 이산화

ST-segments 및 HRV의 모든 진단 지표들은 연속형 속성 값이다. 따라서 목표 클래스에 대한 출현 패턴 마이닝 수행을 위해서는 범주형 속성 값을 갖도록 이산화되어야 한다. 이 절에서는 클래스를 고려하고 구간의 순도를 최대화하는 방식으로 분리점을 배치하는 엔트로피 기반의 이산화 [22]를 적용한다. D를 속성들의 집합과 클래스 라벨에 의해 정의된 데이터 집합이라고 가정한다. 데이터 집합 내의 특정 속성 X의 엔트로피-기반 이산화는 다음과 같다. X의 각 값은 X의 범위를 분할하는 분할점으로 간주될 수 있다. 즉, X의 분할점은 D의 '$X \le$분할점'과 '$X >$분할점'을 만족하는 두 개의 부분 집합으로 분할된다. D의 데이터들을 속성 X에 대해 분할하기 위해, 먼저 기대 정보 요구량(expected information requirement)을 계산하며 (식 13)과 같다.

$$Info_X(D) = \frac{|D_1|}{|D|} Entropy(D_1) + \frac{|D_2|}{|D|} Entropy(D_2) \quad \text{(식 13)}$$

여기서 D_1, D_2는 조건 '$X \le$분할점'과 '$X >$분할점'을 만족하는 데이터들의 집합을 의미하며, $|D|$는 D에 포함된 데이터의 수이다. 주어진 집합에서의 엔트로피 함수는 데이터들의 클래

〈표 1〉 ST-segments 및 HRV의 선형, 비선형 진단 지표

		Diagnosis Index	Description
ST-segments		*J point*	*Edge-detected starting point of ST-segment*
		J80 point	*80 ms after the J point (heart rate ≤ 120 beats/min)*
		Slope	*Slope of the line connecting the J and J80 points*
		Area	*Area between the ECG trace*
Linear parameters	Frequency domain	*nLF*	*Normalized low frequency power*
		nHF	*Normalized High frequency power*
		LF/HF	*The ratio of low- and high-frequency power*
	Time domain	*RRm*	*The mean of RR intervals*
		SDRR	*Standard deviation of all RR intervals*
		SDSD	*Standard deviation of differences between adjacent RR intervals*
Nonlinear parameters		*SD1*	*Standard deviation of the distance of RR(i) from the line y = x in the Poincare*
		SD2	*Standard deviation of the distance of RR(i) from the line y = -x + 2RRmin the Poincare*
		SD2/SD1	*The ratio of SD2 and SD1*
		SD1SD2	*SD1×SD2*
		ApEn	*Approximate Entropy*
		H	*Hurst Exponent*

스 분포에 기반하여 구해진다. m개의 클래스 $C_1, C_2, ..., C_m$가 주어졌을 때, D_1의 엔트로피는 다음과 같다.

$$Entropy(D_1) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad \text{(식 14)}$$

여기서 p_i는 D_1 안에 있는 클래스 C_i의 확률이고, $|D_1|$으로 클래스 C_i의 튜플의 개수를 나누는 방식으로 계산되며, D_2의 엔트로피 값 또한 (식 14)를 이용하여 계산한다.

분할점을 결정하는 과정은 어떤 중단 조건 또는 정지 규칙이 만족될 때까지 구해진 각 분할에 재귀적으로 적용된다. 중간 조건으로는 모든 분할점 대한 최소정보요구량이 특정 기준값 ϵ보다 작다면 분할은 중단된다.

3.2 특징 선택 (feature selection)

대부분의 분류에서 부적절한 특징들의 제거를 위해서 특징 선택을 수행한다. 특징 선택[23] 단계는 특징 순위(ranking)와 선택(selecting) 단계로 구성된다. 선택 알고리즘은 각 특징의 예측 능력과 특징 사이의 중복성(redundancy)을 평가한다. 모든 심혈관계 질환 진단 지표들이 추출된 후에 3.1의 이산화 과정을 거치므로 클래스 및 모든 특징들은 범주형의 값을 가진다. 따라서 속성 선택은 Pearson의 Chi-square에 기반한 p-value를 이용한다.

X를 I개의 범주를 갖는 속성이고, Y는 J개의 범주를 갖는 클래스, N은 전체 데이터 수이다. 또한 N_{ij}를 X=i이고 Y=j인 데이터 개수이며, N_i는 $X=i$을 갖는 데이터 수이고 N_j는 $Y=j$, 인 데이터 개수라 할 때, Pearson의 chi-square(χ^2)에 의한 p-value는 확률, $Prob.(\chi_d^2 > X^2)$에 의해 계산되며, 자유도(degree of freedom), $DF=(I-1)(J-1)$이고 X^2은 (식 15)과 같다[23].

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} (N_{ij} - \widehat{N_{ij}})^2 / \widehat{N_{ij}} \ , \ (\widehat{N_{ij}} = N_i N_j / N) \quad \text{(식 15)}$$

이 논문에서는 심혈관계 질환에 대한 복합 진단 지표를 제안하기 위해서 세 가지 자세를 모두 고려한 특징 선택(<표 2>)과 <표 3>과 같이 각 자세별로 구분된 데이터에 대한 특징 선택을 수행하였다.

〈표 2〉 세 가지 누운 자세에서의 선택된 진단 지표 (S: supine, R: right, L: left)

Rank	Selected feature	Relevance score (1-p)	Rank	Selected attribute	Relevance score (1-p)
1	*Area*	1.000	10	*H(R)*	0.965
2	*J*	1.000	11	*SD2/SD1(R)*	0.965
3	*J80*	1.000	12	*nLF(R)*	0.960
4	*SD2(L)*	0.998	13	*nHF(R)*	0.958
5	*SDRR(L)*	0.985	14	*H(S)*	0.955
6	*SD2/SD1(S)*	0.975	15	*ApEn(R)*	0.955
7	*SD2(R)*	0.974	16	*SDRR(S)*	0.954
8	*SDRR(R)*	0.974	17	*SD1SD2(R)*	0.952
9	*RRm(R)*	0.965			

〈표 3〉 각 누운 자세에 대한 진단 지표

Supine recumbent posture			Right recumbent posture			Left recumbent posture		
Rank	Selected attribute	Relevance score(1-p)	Rank	Selected attribute	Relevance score(1-p)	Rank	Selected attribute	Relevance score(1-p)
1	*Area*	1.000	1	*Area*	1.000	1	*Area*	1.000
2	*J*	1.000	2	*J*	1.000	2	*J*	1.000
3	*J80*	1.000	3	*J80*	1.000	3	*J80*	1.000
4	*SD2*	0.998	4	*SD2*	0.990	4	*SD2*	0.999
5	*SDRR*	0.995	5	*SDRR*	0.989	5	*SDRR*	0.998
6	*SD1*	0.992	6	*SD2/SD1*	0.988	6	*SD2/SD1*	0.969
7	*H*	0.955	7	*H*	0.982	7	*SD1SD2*	0.961
8	*LF/HF*	0.954	8	*nHF*	0.963			
9	*SD2/SD1*	0.951	9	*nLF*	0.963			
10	*RRm*	0.950	10	*RRm*	0.963			
			11	*SD1SD2*	0.96			
			12	*LF/HF*	0.955			
			13	*SDSD*	0.95			

4. 출현 패턴 기반 분류 기법

이 절에서는 심혈관계 질환의 정확한 진단을 위한 분류 모델로서 출현 패턴에 기반한 확장된 EP-classifier 알고리즘을 기술한다.

4.1 출현 패턴 마이닝

출현 패턴(emerging pattern)이란 특정 클래스 라벨을 갖는 데이터 집합에서는 높은 발생빈도 (지지도)를 가지며, 반대로 상대 클래스에는 상당히 낮은 발생빈도를 가지는 항목 집합들을 말한다[14]. 즉, 출현 패턴은 두 개의 분할된 데이터 집합을 명확하게 구분해 주는 패턴을 말하는 것으로, 이러한 패턴들은 하나의 데이터 집합에서 다른 클래스를 갖는 데이터 집합 사이에 명확한 차별 점을 가진다. 일반적으로 연관(association) 분석에서 자주 발생하는 패턴과는 달리 출현 패턴은 높은 구별력(discriminating power)으로 분류 문제에 적용되어 더욱 유용하다고 증명되어 있다. 출현패턴에 대한 문제 정의는 다음과 같다.

[정의 1] 성장률 (growth rate) : 두 개의 서로 다른 클래스에 해당되는 두 집합 D_1, D_2에 대해, 패턴 X의 D_1에 대한 D_2의 성장률은 다음과 같이 정의된다[14], [24].

$$GrowthRate(X) = GR(X) = \begin{cases} 0 & If\ sup_1(X)=0\ and\ sup_2(X)=0 \\ \infty & If\ sup_1(X)=0\ and\ sup_2(X)>0 \\ \sup_2/\sup_1 & otherwise \end{cases}$$

여기서, D_1을 배경(background) 데이터 집합, D_2를 목표(target) 데이터 집합이라고 하며, 출현패턴은 배경 데이터로부터 목표 데이터 집합에 대해 높은 성장률을 가지는 패턴을 의미한다. 또한 성장률 임계값 $\rho>1$에 대해서 패턴 X가 $GrowthRate(X) \geq \rho$의 성장률을 가질 때, 패턴 X를 ρ-Emerging Pattern($\rho-EP$)라 한다[14].

[정의 2] 출현패턴 X의 강도(strength)는 (식 16)과 같다[24].

$$strength(X) = \frac{GR(X)}{GR(X)+1} \cdot sup(X) \qquad \text{(식 16)}$$

[정의 3] JEP(jumping emerging pattern) : 점핑 출현 패턴이란 배경 집합 D_1으로부터 목표 집합 D_2에 대해, 성장률(GR)이 무한대(∞)를 갖는 출현 패턴이다. 따라서, 점핑 출현 패턴[24]은 $strength(X) = sup(X)$인 특별한 형태의 출현 패턴이다.

예를 들어 임계값인 최소성장률(*minimum growth rate*), $\rho=2$라 하고, 복합 진단 지표에 대한 환자 및 정상인 두 클래스에 해당되는 출현 패턴은 다음과 같다.

- EP1: {J=(-0.52865~0.47055], J80=(0.0662~0.28495], SD2(R)=(42.525379~∞)}의 값을 가진 패턴이 만약 지지도가 환자군에 대해 2/9, 정상인군에 대해 3/5를 가질 경우, 성장률(GR)=2.7이므로 EP1은 정상인군에 대한 출현패턴이다.
- EP2: {J=(0.47055~∞), Area=(23.9315~∞)}의 지지도가 환자군에 대해 4/9, 정상인군에서 0일 경우, 성장률(GR)=∞이므로 EP2는 환자군에 대한 점핑 출현 패턴이다.
- EP3: {J=(0.47055~∞), Area=(23.9315~∞), SD2/SD1(S)=(2.277184~4.335772], H(R)=(0.710939~0.955847], ApEn(R)=(1.193974~∞), nHF(R)=(0.416069~0.662279], SDRR(R)=(1.004296~1.064343]}은 환자군에서 단 한번 발생되고(1/9), 정상인군에서 0의 지지도를 갖는다면, 성장률(GR)은 ∞이므로 점핑 출현 패턴이 된다.

그러나 EP3의 경우, 비록 점핑 출현 패턴으로 발견되나 목표 데이터 집합인 환자군에서도 단 한번 발생되므로 분류 패턴으로는 적합하지 않는 잡음 패턴이다. 따라서 이러한 잡음 패턴을 제거하기 위해서 연관규칙에서의 최소지지도(minimum support) 임계값을 적용함으로서 점핑 출현 패턴의 단점을 해결한다.

4.2 EP-tree에 의한 필수 출현 패턴 마이닝

$\rho-EP$와 JEPs 모두 성장률 기반의 완전한 출현 패턴 집합들을 분류 모델 생성에 사용하므로 기존의 분류 모델들보다 더 좋은 성능을 가진다. 그러나 너무 많은 패턴들을 생성하므로 패턴 탐사에 드는 시간이 오래 걸리고 특히, 다차원 속성값을 가지는 데이터에 대해서 상당한 탐사 시간을 요구한다. 또한 JEP의 경우 무한대의 성장률을 기반으로 점핑 출현 패턴만을 생성함으로 낮은 지지도를 갖는 출현 패턴이 생성될 가능성이 높다. 이는 잡음 데이터에 민감하게 되어 분류 모델의 성능을 저하 시키는 원인된다. 따라서 다차원 속성을 가지는 심혈관계 진단 지표들로부터 잡음 출현 패턴의 제거 및 효율적인 출현 패턴 마이닝을 위해서 기존의 FP-growth[25] 방법의 빈발항목집합 탐사 알고리즘을 적용한다.

두 개의 클래스, D_1, D_2에 해당하는 데이터 집합을 가정할 경우, D_1, D_2에 나타나는 모든 항목들의 집합은 $I=\{i_1, i_2, ..., i_n\}$이고, 두 조건 $i \in I$, $\{i\} \subset I$을 만족한다. 이때, FP-growth의 빈발한 1-항목 집합의 오름차순 정렬인 F-list 생성을 위해 지지도-비율(support-ratio) 임계값을 사용한다.

[정의 4] 최소지지도 $\delta>0$가 주어질 경우, 항목 i에 대한 지지도-비율(SupRatio(i))은 다음과 같다.

$$SupRatio(i) = SR(i) = \begin{cases} 0 & \text{If } sup_1(i)<\delta \text{ and } sup_2(i)<\delta \\ 1 & \text{If } sup_1(i)>\delta \text{ and } sup_2(i)>\delta, \text{ and } sup_1 = sup_2 \\ \infty & \text{If } sup_1(i)>\delta \text{ and } sup_2(i)=0, \text{ or } sup_1(i)=0 \text{ and } sup_2(i)>\delta \\ \dfrac{sup_2(i)}{sup_1(i)} & \text{If } sup_1(i)>\delta \text{ and } sup_2(i)>\delta, \text{ and } sup_2(i) > sup_1(i) \\ \dfrac{sup_1(i)}{sup_2(i)} & \text{If } sup_1(i)>\delta \text{ and } sup_2(i)>\delta, \text{ and } sup_1(i) > sup_2(i) \end{cases} \qquad \text{(식 17)}$$

위의 (식 17)에서 큰 값의 지지도-비율을 가지는 항목들

〈표 4〉 두 클래스를 갖는 데이터의 예

D_1					D_2				
a		*c*	*d*	*e*	*a*	*b*			
a							*c*		*e*
	b			*e*	*a*	*b*	*c*	*d*	
	b	*c*	*d*	*e*				*d*	*e*

은 더 구별력이 강한 패턴이고 지지도-비율이 ∞인 항목은 점핑 출현 패턴이 된다. 또한 서로 다른 D_1, D_2에서 모두 최소지지도를 만족하지 못하는 패턴일 경우, 지지도-비율은 0이 된다. 따라서 *SR(i)*에 의해 모든 항목들은 순위가 결정될 수 있고, 즉 FP-tree 알고리즘에 적용 가능하게 된다.

[정의 5] 패턴의 순위, $\prec$ (pattern ranking) : (식 16)의 지지도-비율에 의해 모든 항목들의 순서를 정의할 수 있다. $i, j \in I$인 두 항목에 대해, *SR(i)>SR(j)*일 경우, i는 j보다 우선하며 $i \prec j$로 표현한다. 또한, *SR(i)=SR(j)*일 경우에는 FP-tree에서와 같이 사전적 순서로 순위가 결정된다. 예를 들어 D_1, D_2에 대한 항목들의 분포는 <표 4>와 같고 각 항목들의 지지도가 두 클래스에 대해 *a=2:2, b=2:2, c=2:2, d=2:2, e=3:2*일 경우, 각 항목의 지지도-비율을 계산하면 *SR(e)=3/2=1.5*가 되고 나머지 *SR(a), SR(b), SR(c), SR(d)*은 모두 2/2=1이 되므로, *e*가 순서상 최우선 되고 나머지 같은 지지도-비율인 항목들은 사전적 순서로 순서가 정해진다. 따라서 모든 1-항목, $I=\{a, b, c, d, e\}$에 대한 순서는 $e \prec a \prec b \prec c \prec d$가 된다.

이 논문에서는 최소지지도(δ) 및 growth rate(ρ)를 만족하고, 중복 패턴이 제거된 필수 출현 패턴(essential emerging patterns) 마이닝 알고리즘을 제안한다.

먼저 분류 모델 생성에 필수적인 출현 패턴을 정의하며, 다음의 4가지 조건을 만족하는 패턴들로 정의된다.

[정의 6] 패턴 X는 다음의 4가지 조건을 만족할 때, 필수 출현 패턴 X로 정의된다.

1. 주어진 임계값 최소지지도 σ에 대해, $sup(X) \geq \sigma$을 만족해야 한다.
2. 주어진 최소성장률 ρ에 대해, $GR(X) \geq \rho$을 만족해야 한다.
3. 패턴 X는 X의 모든 부분 패턴 보다 더 큰 성장률을 가진다.
   $(\forall Y \subset X,\ GR(Y) < GR(X))$.
4. 패턴 X와 임의의 부분 패턴 Y에 대해, chi-square 검정 결과가 주어진 임계값 3.84[1] 보다 클 경우, 패턴 X는 필수 출현 패턴이 될 수 있다.

$$|X|>1 \ \wedge\ (\forall Y \subset X \wedge |Y|=|X|-1 \wedge \chi^2(X, Y) \geq \chi^2_{\alpha=0.05,\ DF=1})$$

[정의 7] EP-tree(emerging pattern tree)의 구조는 다음과 같다.

- 하나의 루트와 자식 노드인 prefix subtree, 해더테이블로 구성된다.
- 트리의 각 노드에서는 item-name, $count_1$, $count_2$, node-link를 가진다. $count_1$은 D_1 트랜잭션에서의 항목의 수이고 $count_2$는 D_2에서의 항목의 수이다.
- 해더테이블의 각 엔트리에는 item-name, head of node-link, $count_1$, $count_2$ 정보를 갖는다. $count_1$, $count_2$는 node-link에 연결된 각 항목들의 모든 개수의 합이다.
- FP-tree와는 다르게 트리의 운행은 top-down 방식이며, EP-tree는 [정의 4]의 오름차순 순서를 유지 한다[2].

FP-tree의 경우, 첫 단계에서 1-항목에 대한 내림차순을 기준으로 트리가 구성되나 제안된 EP-tree에서는 오름차순을 기준으로 구성된다. 모든 항목에 대한 트리 구성 시, 오름차순으로 구성을 하게 되면 트리의 가지가 더 늘어나며, 느린 트리 운행이 되는 단점을 가진다. 그러나 chi-square 검정을 통해 많은 불필요한 패턴의 제거를 EP-growth 단계에서 동시에 수행할 수 있는 장점을 가지며, 중복 패턴이 될 수 있는 후보 패턴을 생성하지 않는 장점을 가진다.

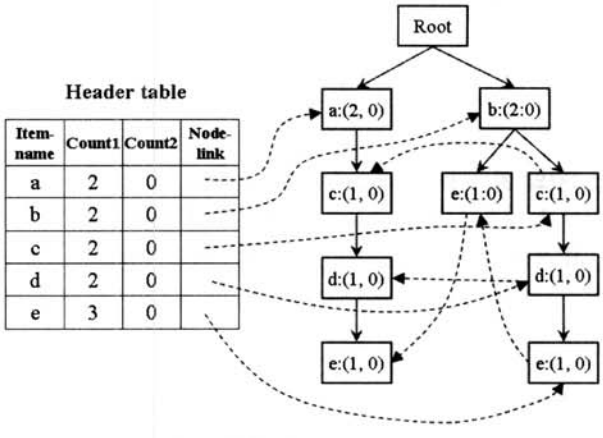
<표 4>의 D_1, D_2에서의 출현 패턴 마이닝을 위한 EP-tree의 구성 단계는 다음과 같다.

1단계 : D_1, D_2의 항목들 중에서 최소지지도, δ을 만족 못하는 모든 항목을 제거한다.

2단계 : [정의 4]의 Pattern ranking에 의해 SupportRatio 기준으로 모든 항목들은 오름차순 정렬된다.

3단계 : <표 4> 트랜잭션의 항목을 오름차순 순서를 고려하여 EP-tree를 구성한다.

<표 4>의 데이터를 $\prec$에 의해 순서화[3] 한 결과는 같은 지지도-비율인 *a, b, c, d*는 그대로 순서를 유지하고 높은 우선순위인 *e*는 오름차순이므로 맨 마지막 순서를 유지한다. 따라서 트리에 삽입되는 항목의 순서는 *a→b→c→d→e* 순서로 삽입된다. (그림 4)는 D_1 데이터 집합의 모든 항목들이 삽입되었을 때까지의 구성된 트리이고 (그림 5)는 D_2 데이

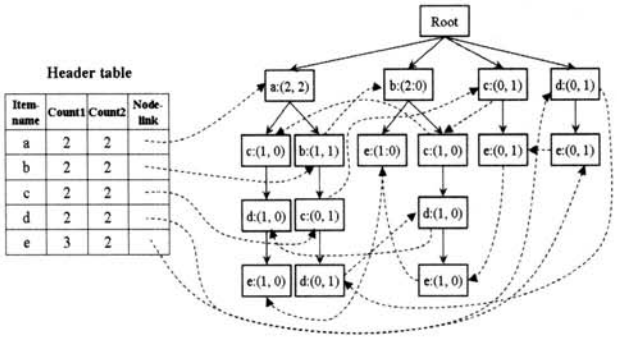


(그림 4) D_1의 항목들(ID 100~400)을 삽입한 트리

1) 신뢰도 95% ($\alpha = 0.05$) 및 자유도 1에서의 카이분포 값이다.

2) 노드 R이 노드 N의 부모노드이고, 항목 i, j가 두 노드에 존재한다면, 두 항목은 오름차순 $i \prec j$를 유지한다.

3) <표 4>의 모든 항목들에 대한 내림차순은 e, a, b, c, d이다.



(그림 5) D_1, D_2 데이터의 모든 항목을 삽입한 완성된 EP-tree

터 집합의 모든 항목까지 삽입한 완성된 트리이다.

EP-tree 구성 알고리즘은 다음과 같다.

4.3 chi-square (χ^2) 검정을 통한 필수 출현 패턴 생성

EP-tree는 기존 FP-tree와는 다르게 최소지지도와 growth rate를 만족하는 출현 패턴을 찾아내며, 노드 링크와 해더테이블로부터 직접적으로 패턴의 임계값들을 조사한다. 트리의 노드 검색은 깊이-우선(depth-first) 탐색으로 진행되며, 패턴-성장(pattern-growth) 방법으로 출현 패턴을 생성한다. 그러나 패턴-성장 과정에서 많은 중복 및 잡음 패턴들이 생성되며, 그 유용성을 검증하기 위해 클래스 분포에 대한 chi-square 검정을 통해 패턴 성장의 진행 여부를 결정한다.

〈표 5〉 A, B 패턴의 데이터 분포에 대한 분할표

count	A={x, y, z}	B={x, y}	$\sum row$
D_1	50	70	120
D_2	25	75	100
$\sum column$	75	145	220

[정의 8] chi-square 테스트 [22] : $|B|>1$, $\forall B \subset A$, $|B|=|A|-1$인 두 패턴에 대해서, chi-square 테스트 결과가 $\chi^2(A,B) \gg \eta$[4]을 만족한다면 A는 서로 다른 클래스 집합에 대해 상당한 분표(발생빈도) 차이를 가지므로 유용한 출현 패턴이 된다. 만약, $\chi^2(A,B)<\eta$일 경우, 두 클래스에 대해 분포 차이가 적다는 의미이며, 두 클래스를 분류하는데 유용하지 못한 패턴이므로 A는 더 이상 성장(growth) 시키지 않는다.

예를 들어 두 패턴이 A={x, y, z}, B={x, y}이고 두 클래스에 대한 분포를 나타내는 분할표가 <표 5> 일 때, $\chi^2(A,B)$는 다음과 같이 계산된다.

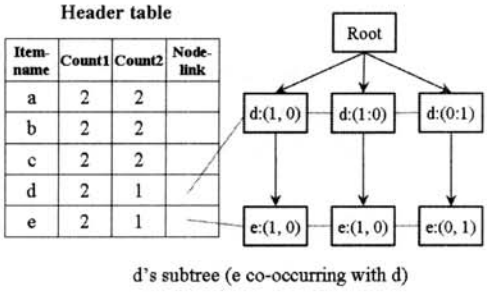
```
input: training dataset D containing two classes (D1, D2), and minimum support δ
output: the EP-tree of D
1. scan D1 and D2, then for each item i, we calculate its support in D1 and D2. Let countD1[i] and countD2[i] denote item i's
support in D1 and D2 respectively.
2.  for each i in D do
3.    if (countD1[i] > |D1|×δ) or (countD2[i] > |D2|×δ) then
4.       add item I into the header table with both counts countD1[i] and countD2[i];
5.    end if
6.  end for
7.  Let L denote the set of all items appearing in header table. Sort L in the ascending order, ≺.
8.  Create the root of a EP-tree.
9.  for each transaction t in D do
10.    select and sort all the L items in t according to the ascending order, ≺;
11.    Let the item list in t be [p|P], where p is the first element and P is the remaining list. Call insert_tree([p|P], root);
12. end for

    Procedure insert_tree([p|P], T)
    input [p|P] list and a subtree of the EP-tree denoted as T.
    output the EP-tree after inserting the new itemset [p|P].
1.  if T has a child node N such that N.items_name=p then
2.       if [p|P]∈D1 then N.countD1 = N.countD1 + 1; end if
3.       if [p|P]∈D2 then N.countD2 = N.countD2 + 1; end if
4.  else
5.       create new node N with N.item_name = p;
6.       if [p|P]∈D1 then N.countD1 = 1; end if
7.       if [p|P]∈D2 then N.countD2 = 1; end if
8.  end if
9.  if P is a nonempty then
10.      call insert_tree(P, N);
11. end if
```

(그림 6) EP-tree 구성 알고리즘

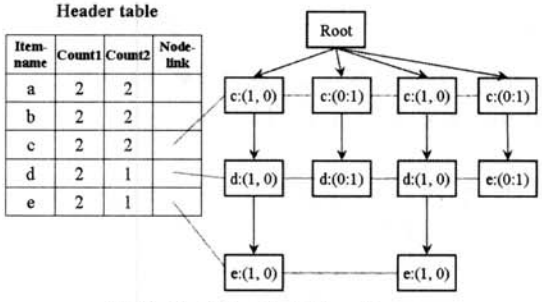
4) r개의 행과 c개의 열을 가진 분할표(contingency table)에 대해서, (r-1)(c-1)인 자유도와 5%의 유의수준하에 χ^2분포표 값 (r=2, c=2일 경우, 자유도는 1이고, 5% 유의수준하에서 기준값 η은 3.84이다.)

〈표 6〉 A, B 패턴의 기대 분할표

count	A={x, y, z}	B={x, y}	$\sum_{row}$
$\sum_{row}$	$75\times\frac{120}{220}\approx 41$	$145\times\frac{120}{220}\approx 79$	120
D_2	$75\times\frac{100}{220}\approx 34$	$145\times\frac{100}{220}\approx 66$	100
$\sum_{column}$	75	145	220



(그림 7) 항목 d에 대한 subtree



(그림 8) 항목 c에 대한 subtree

분할표의 인덱스 {i, j}$\in D_1, D_2\times A, B$에 대해, 두 클래스에 속하게 될 기대 값이 먼저 (식 18)에 의해 계산되며, 전체 기대 분할표는 <표 6>이다.

$$E_{i,j}=\frac{count(j)_{D_1+D_2}(count(A)_i+count(B)_i)}{count(A)_{D_1+D_2}+count(B)_{D_1+D_2}} \quad \text{(식 18)}$$

관찰값(O_{ij}) 및 계산된 기대값(E_{ij}) 분할표로부터 최종 chi-square 테스트(χ^2)를 수행한다.

$$\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \quad \text{(식 19)}$$

계산 결과는 $\chi^2=\frac{(50-41)^2}{41}+\frac{(70-79)^2}{79}+\frac{(25-34)^2}{34}+\frac{(75-66)^2}{66}\approx 6.6>3.84$이며, 5% 유의수준에서 자유도 1일 경우, 기준값 3.84보다 테스트 결과가 크므로 패턴 A, B의 분포는 두 클래스 D_1, D_2에 대해 상당히 다르다고 판단한다. 따라서 패턴 B에 하나의 항목이 추가된 A ($B\subset A$) 패턴은 분류에 유용한 필수 패턴이며 계속 성장된다.

[정의 8]의 chi-square 검정 및 EP-tree로부터 생성되는 출현 패턴은 다음 과정을 거쳐 발견된다. 최소 지지도 δ=1이고 성장률 ρ=2로 가정한다.

- 하위 노드 {e}에 대해, 데이터 집합 D_1, D_2에서의 카운트 값을 해더 테이블로부터 알 수 있고, 최소지지도는 1로 만족한다. 그러나 GR(e) = 1.5<ρ 이므로 출현 패턴이 될 수 없다.
- 항목 {d}는 [{d}:(2:2)의 카운트 값을 가지므로 {d}는 출현 패턴이 아니다. 그러나 {d,e}로 성장할 경우, {d}와 {e}가 동시에 나타나는 항목 d의 subtree를 노드-링크로 탐색하면[{d}:(2:1), {e}:(2:1)이며, GR(d,e)=2이므로 출현 패턴이다.
- 항목 {c}는 [{c}:(2:2)]이므로 출현 패턴이 아니다. 다음으로 c에 대한 subtree에서 {c}와 자식노드 {e} 항목에 대한 카운트 값은 [{c}:(2:2), {e}:(2:1)]로 계산되며, 항목 {c,e}는 GR(c,e)=2인 출현 패턴이다. 또한 자식 노드인 {d}에 대해서, [{c}:(2:2), {d}:(2:1)]이므로 {c,d} 역시 출현 패턴이다. 항목 집합 {c,d}에 {e}를 성장 시키면 [{c}:(2:2), {d}:(2:1), {e}:(2:0)]이다. 따라서 항목 집합 {c,d,e}는 GR(c,d,e) = 2/0 =∞인 점핑 출현 패턴이 된다.
- 항목 {b}는 [{b}:(2:2)]이므로 출현 패턴이 될 수 없고, {b}에 연결된 {e}, {d}, {c}에 대해 확장한다. 먼저 {b} 항목의 자식 노드 {e}를 카운트 하면 [{b}:(2:2), {e}:(2:0)]이 되므로 패턴 {b, e}는 점핑 출현 패턴이다. 다음으로 항목 {d}를 재 카운트 한 결과는 [{b}:(2:2), {d}:(1:1)]이다. 항목 집합 {b,d}와 {b}의 성장률은 모두 1로 같으므로 필수적인 출현 패턴이 될 수 없다. 따라서 {b,d}의 확장은 진행되지 않는다. 실제로 {b,d,e}는 D_1, D_2에 대해 (1:0)의 지지도를 갖는 점핑 출현 패턴이지만, 먼저 탐색된 부분 패턴인 {b,e}:(2:0)가 점핑 출현 패턴이므로 상위 패턴인 {b,d,e}는 필수 출현 패턴이 될 수 없고 이 경우 chi-square 검정을 통해 불필요한 패턴은 모두 제거된다. 마지막으로 항목 {c}의 카운트 결과 [{b}:(2:2), {c}:(1:1)]이다. 이 패턴 또한 항목 {d}와 같은 결과를 보이므로 출현 패턴이 되지 않으며, {b,c}는 더 이상 확장 하지 않는다.
- {a}에 대한 클래스 분포는 [{a}:(2:2)]이므로 {a}는 출현 패턴이 될 수 없고, {e}, {d}, {c}, {b}를 추가하여 확장한다. 먼저 {e}로 확장한 결과는 [{a}:(2:2), {e}:(1:0)]이므로 {a,e}는 점핑 출현 패턴이 된다. 항목 {d}의 추가 결과는

input: EP-tree, minimum support δ, minimum growth rate ρ, and minimum χ^2 value η.
output: the set of essential emerging patterns, F
// Let $I = 1, \ldots, N, (1 < \ldots < N)$.
1. *for each i=N to 1 do*
2. α = {i};
3. *if check_threshold(α) then α into F; //$(\alpha \geq \delta) \wedge (\alpha \geq \rho)$*
4. *subtree(α);*
5. *end for*

Procedure subtree(β)
// $\beta = [\alpha / j]$, α is a prefix of β, and j is the last item of β.
1. *for each item i appears in subtrees of nodes including item j do*
2. *adjust the corresponding node-links and calculate counts of these nodes including i;*
3. *end for*
4. *for each i=N to k+1 do*
5. $\gamma = \beta \cup i$;
6. *if check_threshold(γ) then γ into F;*
7. *if $\chi^2(\gamma, \beta) \gg \eta$ then subtree(γ);*
8. *end for*

(그림 9) χ^2 검정을 통한 필수 출현 패턴 마이닝 알고리즘

[{*a*}:(2:2), {*d*}:(1:1)]이고 항목 {*c*}의 경우도 [{*a*}:(2:2), {*c*}:(1:1)]이므로 두 항목 {*d*}, {*c*}에 대해서는 더 이상 확장하지 않는다. 마지막으로 항목 {*b*}를 추가한 결과 [{*a*}:(2:2), {*b*}:(0:2)]가 되며, {*a,b*}는 점핑 출현 패턴이 된다. 또한 {*a,b*}가 점핑 출현 패턴이 되므로 어떠한 상위 패턴들도 필수 패턴이 되지 않으므로 더 이상 조사되지 않는다.

(그림 9)는 필수 출현 패턴 생성 알고리즘이다.

4.4 출현 패턴에 의한 분류

모든 필수 출현 패턴의 생성 후에 새로운 데이터에 대한 분류는 [14]에서 소개된 score를 계산하여 가장 높은 score 값을 가지는 클래스로 분류하게 된다. 분류를 위한 score 계산식은 다음과 같다.

$$score(s, C) = \sum_{e \subseteq s, e \in E(C)} support_e(e) \cdot \frac{growth\ rate(e)}{growth\ rate(e)+1} \quad \text{(식 18)}$$

여서기 s는 분류될 데이터 인스턴스이고, *E(C)*는 클래스 *C*에서 발견된 필수 출현 패턴이다. 예를 들어 두 클래스 집합 D_1, D_2에 대한 출현 패턴이 각각 D_1={(*a,e*):(50%:25%), (*d,e*):(50%:25%)}, D_2={(*a,d*):(25%:50%)이고, 분류될 데이터가 *s*={*a,d,e*}이라고 가정하면, 두 클래스의 출현 패턴이 *s*를 포함하므로 각각의 score를 계산한다. D_1, D_2에 해당되는 출현 패턴의 score들은 $score_1(s, D_1) = 0.5 \times \frac{2}{2+1} = 0.67$, $score_2(s, D_2) = 0.5 \times \frac{2}{2+1} \approx 0.33$이며, $score_1 > score_2$이므로 *s*는 D_1의 클래스로 분류된다.

5. 실험 평가

이 절에서는 필수 출현 패턴에 의한 분류 기법 알고리즘을 구현한 후에 관상동맥질환자의 제안한 진단 지표 특징들을 추출하여 실험을 통해 알고리즘의 성능을 평가한다.

5.1 데이터 생성

실험에 사용된 환자 데이터는 2004년과 2005년에 표준과학연구원에서 수집된 관상동맥 질환자 261명에 대하여 관동맥조영술을 실시한 후, 관동맥조영술을 통하여 관상동맥에 적어도 50% 이상의 협착이 있는 환자를 관상동맥 질환(CAD: coronary artery disease)으로 분류하고, 협착 정도가 50% 미만을 대조(정상) 그룹으로 하였다. 또한, 관상동맥질환자는 심장전문의들에 의해 다시 협심증(AP: angina pectoris)과 급성관상동맥증후군(ACS: acute coronary syndrome) 집단으로 재분류하였다[4, 26]. 실험에 사용된 환자 데이터의 임상적 특성은 <표 7>와 같다.

5.2 성능 평가

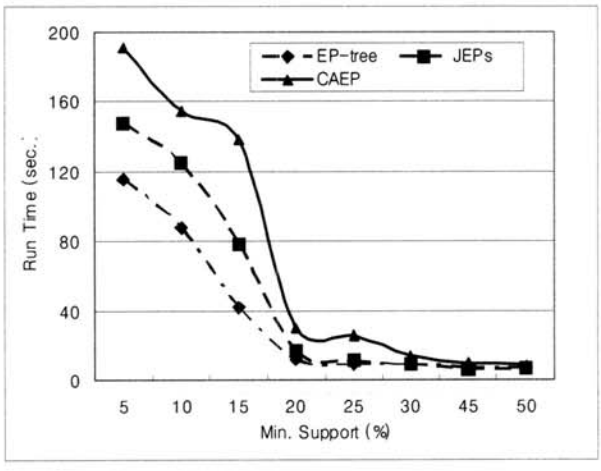
EP-tree 및 chi-square 기반 출현 패턴 생성 알고리즘의 성능평가를 위해 최소지지도 δ, 및 최소 성장률 ρ과 chi-square의 기준값 $\eta \geq 3.84$ 등의 입력파라미터를 고려하여 기존의 출현 패턴 알고리즘들(JEP-classifier, CAEP)[14, 24]과

〈표 7〉 실험에 참가한 환자들의 임상적 특성

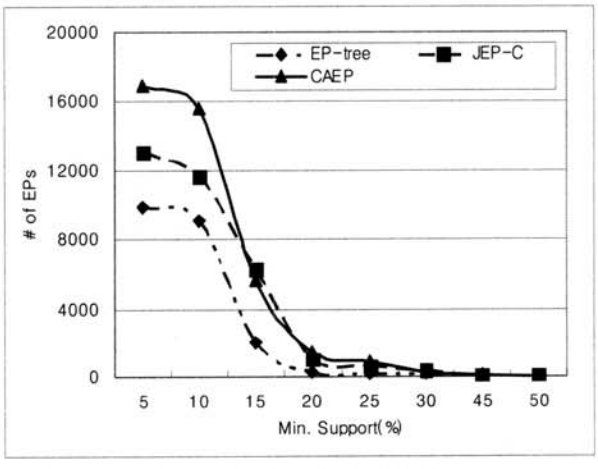
Group	N	Sex (male/female)	Age (years)
Control	128	62/66	53.81 ± 7.23
AP	120	88/32	51.48 ± 8.01
ACS	13	6/7	59.08 ± 9.86

〈표 8〉 출현 패턴 적용을 위한 데이터의 변형

목표 (target) 클래스	배경 (background) 클래스
D_1	D_2+D_3
D_2	D_1+D_3
D_3	D_1+D_2



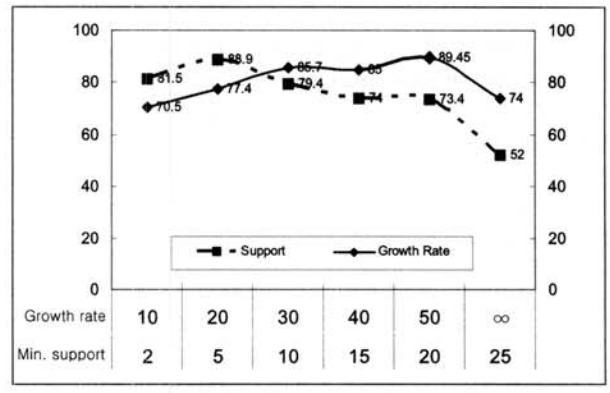
(그림 10) 최소지지도 변화에 따른 실행시간, (JEP: $\rho=\infty$, CAEP, EP-tree: $\rho=8\%$).



(그림 11) 최소지지도 변화에 따른 패턴수(JEP: $\rho=\infty$, CAEP, PF-tree: $\rho=8\%$).

비교 평가한다. 실험 데이터 집합의 크기는 총 261개의 인스턴스이며, <표 4>과 같이 17개의 속성을 가진다. 또한 이 실험에서 사용되는 데이터는 3개의 클래스를 가지므로 출현 패턴 알고리즘들을 적용하기 위해 <표 8>과 같은 방법으로 데이터를 2 클래스 문제로 변형하여 알고리즘을 적용한다.

(그림 10)은 최소 성장률을 8%로 설정 했을 때의 최소지지도 변화에 따른 실행시간 평가이고 (그림 11)은 최소 지지도 변화에 따른 생성되는 출현 패턴 수의 비교이다. (단, JEP-C의 경우, 성장률이 무한대(∞)인 패턴만을 선택하므로 $\rho=\infty$ 으로 설정한다.)



(그림 12) 최소 성장률 및 최소지지도 변화에 따른 정확도 비교

(그림 10)에서 EP-tree는 낮은 최소 지지도(5~25%)에서 JEP 및 CAEP 보다 적은 시간이 걸리나 30% 이후로 JEP가 더 적은 시간을 필요로 한다. 이유는 높은 지지도에서 모든 알고리즘들이 적은 수의 빈발 항목집합을 찾고, JEP의 경우 적은 빈발 항목집합으로부터 성장률이 무한대인 패턴만을 생성하기 때문이며, EP-tree 및 subtree 생성 시간이 더 소요되기 때문이다. 패턴 수 비교의 경우도 유사하게 EP-tree와 JEP가 적은 수의 필수 패턴만을 생성한다. 그러나 JEP의 경우 높은 지지도(30~50%)에서는 전혀 출현 패턴을 생성하지 못한다[5].

출현 패턴 기반 분류 모델은 두 임계값, 최소 지지도 및 최소 성장률에 따라 다른 성능을 보이므로 경험적(heuristic) 방법으로 최적의 파라미터를 설정해야 한다. 따라서 (그림 12)는 최소 성장률 ρ=50% 일 때, 최소 지지도 변화에 따른 분류 정확도 비교와 최소 지지도 δ=5%일 때, 성장률 변화에 따른 정확도 비교이다.

실행 시간 비교와 파라미터 설정을 위한 실험에서는 전체 데이터의 50%의 임의 추출(random sampling)한 데이터만을 사용하였다.

실험 결과 최적 파라미터 설정은 최소 성장률이 50%이고 최소 지지도는 5%이다. 파라미터 임계값들을 선택한 후, 전체 261건의 관상동맥 질환자(AP, ACS) 및 대조(정상인)군 전체 데이터에 대한 정확도 실험을 수행한다.

<표 9>는 이 논문에서 제안한 세가지 누운 자세에서의 선택된 다차원 복합 진단 지표(<표 1>)와 각 자세별 진단 지표(<표 2>)와의 비교를 위한 혼잡 행렬이다. 실험 결과, 제안된 복합 진단 지표의 클래스별 평균 hit rate는 85.33%이고, 독립적으로 자세별 진단 지표를 선택했을 경우는 각각 66%(똑바로), 68%(오른쪽), 76%(왼쪽)이다. 특히 급성관상동맥증후군(ACS)을 협심증(AP)로 잘못 분류하는 경우가 58%(똑바로), 50%(오른쪽), 35%(왼쪽)로 모든 자세를 고려한 진단 지표를 사용했을 때의 6.2%보다 더 많았다.

마지막으로, 제안된 chi-square 검증을 통한 필수 패턴의 사용이 기존의 다른 분류 기법 보다 얼마나 유용한지를 알아

5) EP-tree의 경우 지지도 30%, 45%, 50%에서 35, 8, 7개의 출현 패턴을 생성하며 JEP는 12, 0, 0개를 생성함.

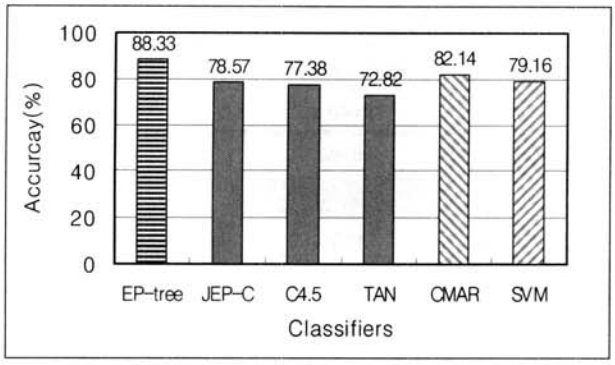
〈표 9〉 복합진단 및 자세별 진단지표에 대한 분류 결과(confusion matrix)

	Posture	Class Label	Predicted Class		
			AP	Control	ACS
Actual Class	All three Postures	AP	92.5%	5.9%	1.6%
		Control	13%	82%	5%
		ACS	6.2%	12.3%	81.5%
	Supine Posture	AP	93%	4%	3%
		Control	25%	75%	0%
		ACS	58%	12%	30%
	Right Posture	AP	87%	7%	6%
		Control	20%	75%	5%
		ACS	50%	8%	42%
	Left Posture	AP	86%	4%	10%
		Control	8%	92%	0%
		ACS	35%	15%	50%

〈표 10〉 분류 모델의 성능 비교 평가

Classifier	Precision	Recall	F-measure	Class
EP-tree(χ^2-test)	0.933	0.925	0.929	AP
	0.781	0.82	0.8	Control
	0.855	0.815	0.835	ACS
JEP Classifier	0.793	0.902	0.844	AP
	0.774	0.6	0.676	Control
	0.762	0.615	0.681	ACS
DT(C4.5)	0.809	0.873	0.881	AP
	0.769	0.750	0.937	Control
	0.579	0.423	0.900	ACS
Bayesian(TAN)	0.748	0.873	0.805	AP
	0.688	0.550	0.611	Control
	0.647	0.423	0.512	ACS
CMAR	0.88	0.863	0.871	AP
	0.822	0.925	0.871	Control
	0.565	0.500	0.531	ACS
SVM(SMO)	0.792	0.931	0.856	AP
	0.811	0.750	0.779	Control
	0.727	0.308	0.432	ACS

보기 위해 성능 비교를 하였다. 비교 대상 분류 기법으로는 *C4.5* 의사결정 트리[27] 및 Bayesian 분류 모델로 *TAN*(tree augmented Naïve Bayes)[28], *RBF*(radial basis function) 커널을 이용한 *SVM* (*SMO*: sequential minimal optimization) [28] 알고리즘과 연관적 분류 기법인 *CMAR*(classification based on multiple class-association rules)[29], 그리고 *JEP*-classifier (jumping emerging patterns)[24] 등이 사용되었다. *CMAR*의 경우 최소 지지도 및 최소 신뢰도 임계값은 각각 0.4, 0.7로 설정하였고 데이터베이스 커버리지는 5% 유의성 기준값과 자유도 1에서의 3.75로 설정하였다. *JEP*-classifier의 경우, 최소 성장률은 무한대, 최소 지지도는 5%로 하였고, *C4.5, TAN*과 *SVM*은 기본(default) 설정을 따랐다. <표 2>



(그림 13) 분류 모델의 정확도 비교

의 복합 진단 지표 데이터에 대해서 모든 분류 모델의 성능 평가는 교차 검증법(10 fold cross-validation)을 적용하였으며, <표 10>과 같이 정밀도(precision), 재현율(recall), F-measure를 평가 지표로 하였다.

(그림 13)은 각 분류 모델의 정확도를 나타낸다. (그림 13)의 정확도 비교 결과, chi-square 검정을 통한 EP-tree 알고리즘이 다른 분류 모델 보다 좋은 성능(88.33%)을 보였으며, *SVM* 및 *CMAR*의 경우도 *JEP*-classifier, *C4.5* 및 *TAN* 보다는 질환 진단을 위한 적합한 모델이 될 수 있을 정도의 유사한 분류 결과를 보였다.

6. 결 론

이 논문에서는 최근 급격히 증가하고 있는 심혈관계 질환 중 관상동맥질환의 조기 진단 및 정확한 예측을 위해서 심박동변이도 및 ST-segments의 복합 진단 지표를 적용하였다. 특히, 심박동변이도 특징은 세 가지 누운 자세에 대한 선형 및 비선형적 특징을 분석하여 다차원 복합진단 지표를 추출하였다. 또한 이 복합 진단 지표들에서 질환 진단에 필수적인 요소들만을 모델 생성에 이용하고, 효율적인 진단 모델의 생성을 위해서 통계적 기법인 chi-square 검증을 추가한 EP-tree 알고리즘을 제안 하였다. 실험 결과 복합 진단 지표를 EP-tree 분류 알고리즘에 적용 했을 때, 필수 출현 패턴들만을 이용한 EP-tree 분류 모델이 기존의 *JEP*-classifier 및 *C4.5, TAN, SVM, CMAR* 분류기 보다 더 우수한 성능을 보였다.

7. 감사의 글

이 논문은 2007년도 충북대학교 학술연구지원사업의 연구비 지원과 2008년도 정부(과학기술부)의 재원으로 한국과학재단의 지원(R01-2007-000-10926-0) 및 2008년 교육과학기술부 지원 지역거점 연구단 육성사업(충북 BIT 연구중심 대학 육성사업)의 지원으로 수행된 결과입니다.

참 고 문 헌

[1] 통계청 인구동향과, "2006년 사망 및 사망원인통계결과," pp.17-18, 2007.

[2] R. Detrano, D. Mulvihill, K. Lehmann, P. Dubach, A. Colombo, D. McArthur, V. Froelicher, "Exercise-induced ST depression in the diagnosis of coronary artery disease. A meta-analysis," Journal of Circulation, American Heart Association, Vol.80, pp.87-98, 1989.

[3] M. E. Bertrand et al, "Management of acute coronary artery syndrome in patients presenting without persistent ST-segment elevation," European Heart Journal, Vol. 23, pp.1809-1833, 2002.

[4] H. G. Lee, K. Y. Noh, K. H. Ryu, "Mining Biosignal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV," Lecture Notes in Artificial Intelligence, Springer Berlin/Heidelberg, Vol. 4819, Emerging Technologies in Knowledge Discovery and Data Mining, pp.56-66, 2007.

[5] K. H. Ryu, W. S. Kim, H. G. Lee, "A Data Mining Approach and Framework of Intelligent Diagnosis System for Coronary Artery Disease Prediction," The Institute of Electronics, Information and Communication Engineers (IEICE), pp.33-34, 2008.

[6] J. Pumprla, K. Howorka, D. Groves, M. Chester, "Functional assessment of heart rate variability physiological basis and practical applications," Journal of Cardiology, Vol.84, pp.1-14, 2002.

[7] JF. Sneddon, Y. Bashir, "Vagal stimulation after myocardial infarction: accentuating the positive," Journal of American Cardiology, Vol.22, pp.1335-1337, 1993.

[8] W. S. Kim, Y. Z. Yoon, J. H. Bae, K. S. Soh, "Nonlinear characteristics of heart rate time series: influence of three recumbent positions in patients with mild or severe coronary artery disease," Physiological Measurement Vol.26, pp.517-529. 2005.

[9] S. Miyamoto, M. Fujita, K. Tambara, H. Sekiguchi, S. Eiho, K. Hasegawa, et al. "Circadian variation of cardiac autonomic nervous activity is well preserved in patients with mild to moderate chronic heart failure: effect of patient position." Journal of Cardiology, Vol. 93, No.3, pp.247-252. 2004.

[10] C. Papaloukas, D. I. Fotiadis, A. Likas, and L. K. Michalis, "An ischemia detection method based on neural networks and bidirectional associative memories," Journal of Medical Eng. Technologies, Vol.24, pp.167-178, 2002.

[11] Y. Goletsis, C. Papaloukas, D. I. Fotiadis, A. Likas, and L. K. Michalis, "Automatic ischemic beat classification using genetic algorithms and multicriteria decision analysis," IEEE Trans. Biomedical Eng., Vol.51, No. 10, pp.1717-1725, 2004.

[12] C. Papaloukas, D. I. Fotiadis, A. P. Liavas, A. Likas, and L. K. Michalis, "A Knowledge-Based Technique for Automated Detection of Ischemic Episodes in Long Duration Electrocardiograms," Medical & Biological Eng. & comp., Vol.39,1 No.1, pp.105-112, 2001.

[13] "European ST-T database directory," European Society of Cardiology, Pisa, Italy, 1991.

[14] G. Dong, X. Zhang, L. Wong, J. Li, "Classification by aggregating emerging patterns," Proceedings of the. 2nd Int'l Conference on Discovery Science, pp.30-42, 1999.

[15] W. J. Tompkins, "Biomedical digital signal processing," Prentice Hall PTR, Upper Saddle River, New Jersey 07458, 1995.

[16] D. Barnaby, K. Ferric, DT. Kaplan, S. Shah, P. Bijur, EJ. Gallagher, "Heart rate variability in emergency department patients with sepsis," Emerging Medical, Vol.9, pp.661-670, 2002.

[17] M. Brennan, M. palaniswami, P. Kamen, "Do existing measurements of Poincare plot geometry reflect nonlinear features of heart rate variability?" IEEE Trans. on Biomedical Eng., Vol.48, No.11, pp.1342-1347, 2001.

[18] S. M. Pincus, and W. M. Huang, "Approximate entropy: statistical properties and applications," Communication Statist. Theory Meth. Vol.21, pp.3061-3077, 1992.

[19] S. M. Pincus and A. L. Goldberger, "Physiological time-series analysis: what does regularity quantify?" Physiology, Vol.266, H1643, 1994.

[20] TH. Makikallio, T. Ristimae, KE. Airaksinen, CK Peng, AL Goldberger, "Heart rate dynamics in patients with stable angina pectoris and utility of fractal and complexity measures," Journal of Cardiology, Vol.81, pp.27-31, 1998.

[21] N. Kannathal, UR. Acharya, CM. Lim, PK. Sadasivan, "Characterization of EEG-A comparative study," Computer Methods and Programs in Biomedical, Vol.80, No.1, pp.17-23, 2005.

[22] U. Fayyad, K. Irani, "Multi-Interval discretization of continuous-valued attributes for classification learning," Proceedings of the. Int'l Joint Conference. on Artificial Intelligence, pp.1022-1027, 1993.

[23] L. Guyon, A. Elisseeff, "introduction to variable and feature selection," Journal of Machine Learning Research 3, pp.1157-1182, 2003.

[24] G. Dong, J. Li, X. Zhang, "Discovering jumping emerging patterns and experiments on real datasets," Proceedings of the 9th Int'l Database Conference on Heterogeneous and Internet Databases, pp.155-168, 1999.

[25] J. Han, M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann publishers, 2000.

[26] R. M. Bethea, B. S. Duran, T. L. Boullion, "Statistical methods for engineers and scientists," New York : M. Dekker. 1995.

[27] J. Quinlan, "C4.5: Programs for Machine Learning," Morgan Kaufmann San Mateo, 1993.

[28] IH. Witten, E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques," San Mateo, CA: Morgan Kaufmann, 1999.

[29] W. Li, J. Han, J. Pei, "CMAR: Accurate and Efficient Classification Based on Multiple Association Rules," Proceedings of the Int'l Conference Data Mining. Vol.1119, pp.369-376, 2001.

이 헌 규

e-mail : hg_lee@etri.re.kr

2002년 경기대학교 정보과학부(학사)

2004년 충북대학교 대학원 전자계산학과 (이학석사)

2009년 충북대학교 대학원 전자계산학과 (공학박사)

2004년~2006년 한국표준과학연구원 생활계측그룹 위촉연구원

2009년~현 재 한국전자통신연구원 우정기술센터 연구원

관심분야 : 데이터베이스, 데이터마이닝, 실시간 우편물류 운영기술 등

노 기 용

e-mail : kyno@kriss.re.kr

1981년 충남대학교 물리학과(학사)

1995년 충남대학교 대학원 전자계산학과 (이학석사)

2004년 충북대학교 대학원 전자계산학과 (이학박사)

1988년~현 재 한국표준과학연구원 선임연구원

관심분야 : 데이터베이스 설계, ATM, 이미지 처리 등

류 근 호

e-mail : khryu@dblab.chungbuk.ac.kr
1976년 숭실대학교 전산학과(이학사)
1980년 연세대학교 전산전공(공학석사)
1998년 연세대학교 전산전공(공학박사)
1976년~1986년 육군군수 지원사 전산실 (ROTC 장교), 한국전자통신연구원 (연구원), 한국방송통신대학교 전산학과(조교수) 근무
1989년~1991년 Univ. of Arizona Research Staff(TempIS 연구원, Temporal DB)
1986년~현 재 충북대학교 전기전자컴퓨터공학부 교수
관심분야 : 시간 데이터베이스, 시공간 데이터베이스, Temporal GIS, 지식기반 정보검색 시스템, 유비쿼터스컴퓨팅 및 스트림 데이터 처리, 데이터마이닝, 데이터베이스 보안, 바이오인포매틱스 등

정 두 영

e-mail : fiorgeo@chungbuk.ac.kr
2001년 서강대학교 대학원 컴퓨터과학과(공학박사)
1987년~현 재 충북대학교 전기전자컴퓨터공학부 교수
관심분야 : 데이터 통신, 데이터베이스 등