심전도 패턴 판별을 위한 빈발 패턴 베이지안 분류

노 기 용[†] · 김 원 식[†] · 이 헌 규[††] · 이 상 태[†] · 류 근 호[†††]

요 약

심장의 활동을 기록한 심전도는 심장의 상태에 대한 가치 있는 임상 정보를 제공한다. 지금까지 심전도를 이용한 심장 질환 진단 알고리즘에 대한 많은 연구가 진행되어 왔으나, 심장 질환에 대한 진단 결과의 부 정확성으로 인해 심전계에서는 외국의 진단 알고리즘을 사용하고 있다. 이 논문에서는 심전도 데이터의 수집에서부터 전 처리 과정 그리고 데이터마이닝을 이용한 심장 질환 패턴 분류 기법을 제안한다. 이 패턴 분류 기법은 빈발 패턴 베이지안이며 기존의 나이브 베이지안과 빈발 패턴 마이닝의 통합이다. 빈발 패턴 베이지안은 훈련단계에서 탐사된 빈발 패턴들을 사용하여 Product Approximation 구성하므로써 클래스 조건 독립 가정을 가진 나이브 베이지안의 단점을 해결한다.

Frequent Pattern Bayesian Classification for ECG Pattern Diagnosis

Gi Yeong Noh[†] · Wuon Shik Kim[†] · Hun Gyu Lee[††]
Sang Tae Lee[†] · Keun Ho Ryu[†††]

ABSTRACT

Electrocardiogram being the recording of the heart's electrical activity provides valuable clinical information about heart's status. Many researches have been pursued for heart disease diagnosis using ECG so far. However, electrocardio-graph uses foreign diagnosis algorithm due to inaccuracy of diagnosis results for a heart disease. This paper suggests ECG data collection, data preprocessing and heart disease pattern classification using data mining. This classification technique is the FB(Frequent pattern Bayesian) classifier and is a combination of two data mining problems, naive bayesian and frequent pattern mining. FB uses Product Approximation construction that uses the discovered frequent patterns. Therefore, this method overcomes weakness of naive bayesian which makes the assumption of class conditional independence.

키워드 : 데이터마이닝(Data Mining), 빈발 패턴 베이지안 분류(Frequent Pattern Bayesian Classification), 빈발 패턴 마이닝(Frequent Pattern Mining), 베이지안 분류(Bayesian Classification), 심전도 패턴 판별(ECG Pattern Diagnosis)

1. 서 론

심전도(ECG : electrocardiogram)는 심장의 상태를 비관혈적(non-invasive)으로 진단하는 매우 중요한 수단으로 활용되며, 진폭은 수 mV이고 주파수는 250Hz이내의 생체전위 신호 중 하나이다.

심전도를 이용한 심장 질환 진단 알고리즘에 대한 많은 연구가 지난 수년 동안 국내에서 진행되어 왔다. 그러나 심장 질환에 대한 진단 결과의 부 정확성으로 인해 대부분의 심전계에서는 외국의 진단 알고리즘을 사용하고 있다. 이러한 이유로는 심장 질환별 심전도와 이에 대한 임상의의 진단을 저장한 데이터베이스의 부재를 들 수 있고 외국 환자들을 대상으로 한 데이터베이스를 사용하여 질환 진단 알고리즘들을 개발하였기 때문이다. 따라서 이 논문에서는 심전도 데이터의 수집에서부터 전처리 과정 그리고 심장질환을 판별하기 위한 통계적 마이닝 기법인 베이지안 분류를 확장한 심전도 패턴 판별 기법을 제안하며 세부 내용은 다음과 같다.

- 심전도 데이터 추출로서 잠정적인 심전도 데이터 포맷을 결정하여 한국인의 심전도 데이터를 추출하고, 심전도 패턴 분석 및 성능 평가를 위하여 외국의 ST-T 데이터베이스로부터 병력 심전도 데이터를 획득한다.
- 데이터의 전처리 단계로써 심전도 신호의 원신호 왜곡을 최소화하기 위한 웨이블릿 모함수를 결정하였고 노이즈, 왜곡된 부분을 시각적으로 수정한다.
- 데이터 검출 단계로 심전도 시그널에서의 QRS 및 R-peak 발견하여 심근허혈 진단에 필요한 세그먼트들을 추출한다.
- 특징 벡터 추출로서 심전도 신호의 정확한 분류를 목적으로 ST-segment의 slope와 면적을 추출하여 특징 벡터로 사용하고, 심전도 데이터의 좀더 정확한 분류를 위해

※ 이 연구는 2003년도 한국과학기술평가원 및 과학기술부 RRC(청주대 ICRC)의 지원으로 수행되었음.
† 정 회 원 : 한국표준과학연구원
†† 준 회 원 : 한국표준과학연구원 위촉연구원
††† 종신회원 : 충북대학교 전기전자및컴퓨터공학부 교수
논문접수 : 2004년 5월 11일, 심사완료 : 2004년 7월 1일

특징벡터를 3개의 그룹(D_1, D_2, D_3)으로 구성한다.

- 심전도 데이터의 허혈/정상을 분류하기 위한 분류 알고리즘을 제안하며 이 알고리즘을 이용하여 심전도 데이터를 분류한다. 제안한 분류 기법은 기존의 나이브 베이지안 분류기에 데이터 속성들의 종속성을 고려하기 위해 빈발 패턴 마이닝을 이용한다. 따라서 나이브 베이지안이 가지는 클래스 조건 독립 가정의 단점을 극복한다. 또한 빈발 패턴 마이닝 시 성능의 향상을 위해 후보 생성이 없는 FP-growth 방법을 확장한 알고리즘을 설계하여 적용한다.

이 논문의 구성은 다음과 같다. 제 2장에서는 심전도의 ST-segment를 이용한 심장 질환 판별에 대해 기술하고 기존의 빈발 패턴 마이닝과 베이지안 분류 기법에 대해 분석한다. 제 3장에서는 심장 질환 패턴 분석을 위한 데이터 전처리 과정을 설명하고 제 4장에서는 심근 허혈 질환 분류를 위한 Frequent pattern Bayes 분류 알고리즘을 제안한다. 제 5장에서는 제안된 심전도 패턴 분류 알고리즘에 대해 실험 및 평가를 기술하고 제 6장에서 결론을 맺는다.

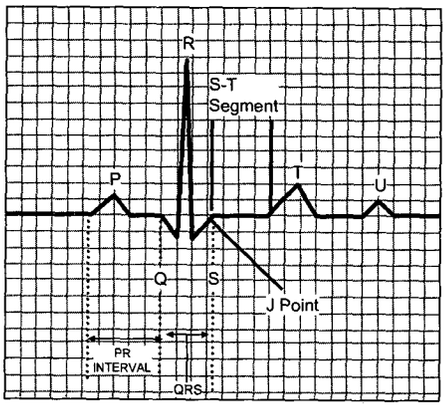
2. 관련 연구

2.1 ECG의 ST-segment를 이용한 심장 질환 패턴 판별

심전도(ECG : electrocardiogram)는 심장의 상태를 비관혈적(non-invasive)으로 진단하는 매우 중요한 수단으로 활용되며, 진폭은 수 ㎷이고 주파수는 250Hz이내의 생체전위 신호 중 하나이다. 국내의 심전도 시스템의 설계와 신호처리에 관한 연구는 1980년 초부터 시작되었다. 대략 10년 동안의 기반기술 축적으로 1990년 초반부터 본격적인 심전계의 개발에 들어갔으며, 현재 12채널 진단 심전계, 홀터 심전계, 스트레스 심전계, 그리고 환자감시 장치 등의 심장관련 진단기기에 대한 연구가 활발히 이루어지고 있다[1].

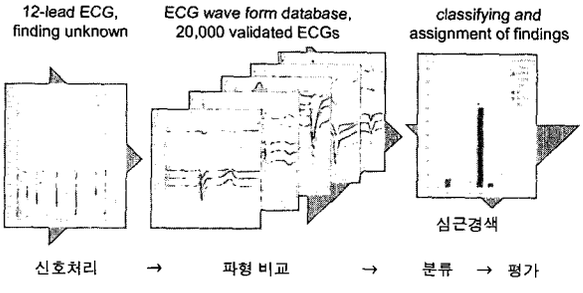
돌연사를 일으키는 대표적 심장질환은 허혈성 심장질환, 확장성 심근증, 비후성 심근증이 대부분을 차지하며 특히 허혈성 심장 질환이 돌연사의 80%를 차지하므로 이 질환의 예방 및 조기진단이 중요하다. 허혈성 심장질환의 증세로서 협심증과 심근경색증이 있는데, 심전도의 ST-segment elevation 또는 depression되는 episode를 띄게 된다. (그림 1)은 심전도 데이터에서 ST-segment, RR 간격, QRS complex, J point 등을 표현한 것이다.

그러나 심근허혈, 심근경색에 대해서는 아직도 오진이 자주 있는 상황이다. 이 ST-segment 1Hz미만의 주파수 대역을 가지고 있으므로 저주파신호의 기저선(baseline) 변동 잡음, 전 주파수 대역에 존재하는 근잡음(muscle artifact)과 같은 주파수 대역에 존재하므로 정확한 잡음을 제거하지 못하면 신호의 왜곡이 발생하므로 오진을 하게 된다. 기저선 변동 잡음 제거를 위한 전처리 필터로는 현재 스플라인 보간법, FIR 필터, 적응필터, 신경회로망, 웨이블릿 변환 방법들로 신호의 왜곡을 최소화하여 기저선 변동 잡음을 제거한다[2, 3].



(그림 1) 심전도 파형의 구성요소

심전계의 심장 질환 판별은 입력/출력 관계의 타당성을 테스트하는 새로운 접근 방법이 요구되며, 이러한 방법으로서 심전도 전문가에 의하여 진단이 알려진 저장되어 있는 심전도 신호 집합을 입력시키고 전문가의 진단과 심전계에 의하여 제안된 진단 결과를 비교하는 것이다[4]. (그림 2)는 심전도 데이터를 이용한 심장 질환 판별 프레임워크를 나타낸 것이다. 알려지지 않은 특정인의 심전도 데이터를 국내외 데이터베이스에 저장된 데이터 집합을 이용하여 질환 유무에 대한 파형 분석을 거친다. 이러한 분석을 거친 후 특정인의 심전도 파형이 정상인지 혹은 질병을 가지고 있는지를 평가한다. 심전도의 파형 분류를 위해서는 주로 신경망 알고리즘을 이용한다[3, 5, 6].



(그림 2) 심장 질환 검출을 위한 심전도 판별 프레임 워크

2.2 빈발 패턴 마이닝과 베이지안 분류

빈발 패턴 마이닝은 빈발한 항목들의 완전한 집합을 찾아내는 일련의 과정을 말하며, 연관규칙, 상관분석, 순차패턴 등에서 핵심 요소로 사용된다. 빈발 패턴 마이닝을 위한 알고리즘 중 대표적인 것이 Apriori[7]이다. 이 방법은 연관규칙에 대한 빈발 패턴을 찾아내는 데 유용하며 후보 집합 생성 및 검사법은 후보 집합의 크기를 현저히 축소시켜 좋은 결과를 도출한다. 그러나 Apriori는 상당한 크기의 후보 집

합 생성이 필요하며 반복적인 데이터베이스 스캔과 대규모 후보 집합에 대한 패턴 매칭 검사가 필요하다는 단점을 가진다. FP-growth[8]는 후보 생성 없이 완전한 빈발 패턴 집합을 발견하는 알고리즘으로 분할-정복 기법을 적용한다. 이 방법은 먼저 빈발 항목을 가지는 데이터베이스를 빈발 패턴 트리로 압축한다. 이때 항목간의 연관정보는 손실이 없다. 그런 후에 압축된 데이터베이스를 하나의 빈발 항목에 대하여 관련된 조건 데이터베이스의 집합으로 분할하고 이와 같은 분할된 개개의 데이터베이스에 대해서 개별적으로 마이닝을 수행한다.

분류란 중요한 데이터 클래스를 설명하는 모형을 생성하거나 미래 데이터의 경향을 예측하고자 할 때 사용되는 분석 기법이다. 분류 기법에 대한 연구는 통계, 신경망, 결정 트리 등의 분야에서 연구되었으며 의료진단 예측 수행, 선택적 마케팅 분야에서 응용된다.

베이지안 분류기[9, 10]는 통계적 분류기이다. 이것은 주어진 샘플이 특정 클래스에 속할 확률과 같이 어떤 항목이 특정 클래스에 속할 확률을 예측한다. 나이브 베이지안 분류기[11]는 주어진 클래스의 한 속성 값이 다른 속성의 값과 서로 독립이라는 것을 가정한 베이지안 분류기이다. 이 가정을 클래스 조건 독립이라고 하며 계산과정을 간단하게 한다. 만약 이 가정이 사실인 경우 나이브 베이지안은 다른 분류기 보다 더 높은 정확성을 가진다. 그러나 실제 데이터의 변수들 사이에는 종속성을 포함하며 이로 인해 조건 독립 가정에 따른 부정확성과 가용 확률 데이터의 부족으로 분류기의 성능이 저하된다.

3. 심전도 데이터의 전처리

이 절에서는 심근 허혈 질환의 분류를 위한 전처리 단계로서 European ST-T 데이터베이스의 심전도 데이터와 병원환자를 대상으로 심전도를 측정한 데이터를 적용하며 다음과 같은 전처리 단계를 거쳐 특징 벡터를 추출한다.

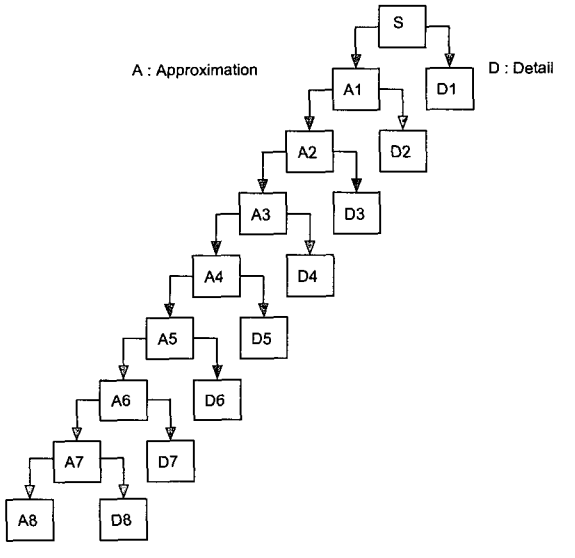
3.1 심전도 신호의 기거선 제거를 위한 웨이블릿 변환

웨이블릿 변환은 주파수 대역에 따라 주파수와 시간영역의 해상도를 달리 할 수 있는 장점이 있으므로 심전도 신호와 같은 비정상적 신호분석에 우수하다. 따라서 심전도 데이터에 대해 이산 웨이블릿 변환의 다해상도 분석법을 이용하여 잡음의 주파수 대역을 제거하는 방법으로 필터를 구현한다. 250 샘플로 구성된 European ST-T 데이터베이스의 데이터에 대하여 다해상도 웨이블릿 변환과 주파수 분할 관계를 보면 <표 1>과 같다.

AHA(American Heart Association)에서는 ST-segment의 왜곡을 방지하기 위해 차단주파수로서 0.8Hz를 추천하였다[1]. 이를 바탕으로 실험의 기저선 구간을 웨이블릿 변환의 approximation level 8(A8)으로 정의하였다.

〈표 1〉 샘플링 주파수와 웨이블릿 변환 레벨에 의한 주파수 분할 관계

Level	fs = 250Hz	
	저주파[Hz]	고주파[Hz]
-1	0~60.3	60.3~125
-2	0~30.2	30.2~60.3
-3	0~15.1	15.1~30.2
-4	0~7.6	7.6~15.2
-5	0~3.8	3.8~7.6
-6	0~1.9	1.9~3.8
-7	0~0.8	0.8~1.9
-8	0~0.4	0.4~0.8



(그림 3) Dyadic Tree 구조의 다해상도 웨이블릿 변환

S를 심전도 + 잡음이라 하였을 때, 다음과 같이 정의한다.

$$S = D1 + D2 + D3 + D4 + D5 + D6 + D7 + D8 + A8 \quad (1)$$

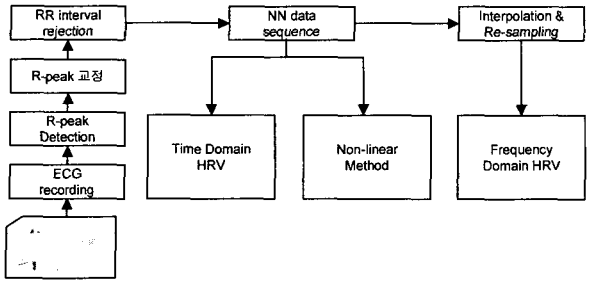
S′을 심전도 - 잡음이라 하였을 때, 다음과 같이 정의한다.

$$S' = D1 + D2 + D3 + D4 + D5 + D6 + D7 + D8 \quad (2)$$

식 (1)에서 기저선 대역으로 정의한 approxi mation level 8(A8)의 신호를 제거한 후 식 (2)과 같이 level 1에서 level 8까지의 detail 신호를 재구성하여 필터를 구성한다.

3.2 R-peak 및 QRS complex 검출

심전도 R-peak 및 QRS complex 검출 과정을 위하여 이 논문에서는 (주)Bionet의 심전계(모델 : CardioCare3000)를 이용하여 심장질환 환자의 심전도를 측정하여 (그림 4)와 같은 과정을 통하여 심박동변이도(heart rate variability : HRV)를 분석한다.



(그림 4) HRV 분석을 위한 단계별 요약

심박동변이도 분석에서 심전도의 QRS complex를 정확히 검출하는 것뿐만 아니라, 검출된 R-peak 중에서 부정맥으로 인한 ectopic beat를 심장전문의가 찾아서 컴퓨터상에서 수작업으로 제거하는 편집이 매우 중요한데, CardioCare3000과 연동되는 심박동변이도 분석용 프로그램에는 이 편집 작업을 할 수 있는 부분이 없어서 이 연구를 통하여 QRS complex를 정확히 검출한 뒤 수작업으로 ectopic beat를 제거할 수 있도록 편집 기능이 추가된 QRS complex 검출 프로그램을 개발하였다. CardioCare3000 장비는 sampling rate가 500Hz이므로 심박동변이도의 오차를 최소화하기 위해 sampling rate를 2,048Hz로 re-sampling하여 분석 하였다. 이렇게 re-sampling 되어진 심전도 신호를 다양한 QRS complex 검출 알고리즘(Pan과 Tomkins가 제안한 방법, Wavelet을 이용한 방법, CIC filter를 이용한 방법)을 적용하여 QRS complex를 검출하였다. 검출 되어진 QRS complex의 R-peak를 심장전문의들이 <표 2>를 기준으로 수작업으로 편집할 수 있도록 하였다.

<표 2> RR 간격 스퀀스 편집 대상

편 집 대 상	
1	RR 간격이 그 이외의 정상 RR 간격에 비하여 0.8배 미만이거나 1.2배를 초과한 경우
2	RR 간격이 150ms 미만이거나 5000ms를 초과한 경우
3	fusion, idio-ventricular, faced fusion, dual paced, atrial paced, ventricular paced beat

편집대상 beat와 그 앞과 뒤의 beats(전체 3개 beats)를 위의 표와 같이 제거하여 normal R-peak 로서 256 샘플을 추출한다(NN data sequence). 주파수 영역 해석을 위해서는 동일 시간 간격으로 re-sampling 해주어야 하는데, 이를 위해 Cubic-Spline Interpolation을 하여 4Hz로 re-sampling 하였고, DC와 저주파 성분을 제거하기 위해 선형성분을 제거 하며, 그 후 FFT와 Burg AR(AutoreRressive) model을 이용하여 주파수 영역에서 심박동변이도를 분석한다.

3.3 ST-segment 특징 벡터 추출

심전도의 전처리 과정 후 심전도의 QRS complex를 웨이블릿 변환을 응용하여 검출한 후 심근허혈 진단의 중요 파라미터인 ST-세그먼트의 특징 벡터를 추출한다. QRS complex는 5~30Hz의 주파수 성분을 갖기 때문에 웨이블릿 특성을 이용하여 5~30Hz을 추출하여, 변화하는 심전도 파형에 적응적인 문턱치 알고리즘을 적용하여 정확히 QRS를 검출한다. QRS complex 검출 후 R-peak를 검출하여 ST-세그먼트의 시작점인 J point는 RR간격이 600ms보다 클 경우는 J point = R + 60ms, 작을 경우는 J point = R + 40ms로 정의한다. 또한, ST60과 ST80을 특징벡터로 사용하였는데, RR간격이 600ms 보다 크면 ST60은 R + 120ms로 하고 ST80은 R + 140ms를 사용하는 반면에, 600ms 보다 작으면 ST60은 R + 100ms로 하고 ST80은 R + 120ms로 사용한다. ST-segment의 slope와 area도 추출하여 특징벡터로 사용하여, 제안된 분류 알고리즘의 입력을 위해 각 벡터 값을 질환 진단 기준에 의해 이산화 한다. ST80(ST60)이 0.08보다 작고 slope 65°이상이거나 ST80(ST60)이 0.08 이상일 경우 이상 진단, 그렇지 않은 경우 정상으로 하였다. 또한 심전도 신호의 정확한 분류를 목적으로 특징벡터를 3가지 그룹으로 구성하였다.

$D_1 = [J, ST80, Slope, Area]$, $D_2 = [J, Slope, ST80]$,
$D_3 = [Slope, Area, ST80]$

4. 심전도 패턴 분류를 위한 FB 분류

전처리 단계를 거친 ST-segment 특징 벡터를 해당 클래스(정상/허혈)로 분류하기 위해 데이터마이닝의 분류 기법을 적용한다.

이론적으로 다른 모든 분류 기법와 비교하면 베이지안 분류는 최소 오류율을 갖는다. 또한 베이지안 분류기는 베이즈 이론을 쓰지 않는 다른 분류기와 달리 이론적 근거를 제공하는 면에서 유용하다. 예를 들어 특정 가정 하에서 많은 신경망이나 곡선적합(curve-fitting) 알고리즘들이 나이브 베이지안 분류기와 같이 최대 사후(maximum posteriori) 가설을 제공한다는 것을 보여준다. 그러나 클래스 조건 독립성이라는 가정에 따른 부정확성과 가용 확률 데이터의 부족으로 인하여 결과가 항상 정확하지는 않다. 실제로 대부분 데이터들의 속성들 사이에는 종속성이 존재한다. 따라서 데이터 속성들 사이의 종속성을 고려한 베이지안 분류기가 필요하며 이를 위해 빈발 패턴(또는 빈발 항목집합)을 이용한 베이지안 분류 기법을 제안한다. 이는 나이브 베이지안이 가진 단점(분류모델 생성 시 단 하나의 속성(항목)만을 적용)을 보완한 분류기를 생성한다.

4.1 빈발 패턴을 이용한 베이지안 분류(Frequent Bayesian)

FB(Frequent Bayesian)는 각 속성들의 빈발한 임의의 길이를 가지는 패턴 집합을 사용하여 근사(approximation)하는 것으로 NB(Naive Bayesian)의 확장이다. FB를 이용한

분류는 다음의 두 단계로 구성된다.

① 훈련단계 : 미리 정의된 클래스를 갖는 훈련 집합에서 클래스 라벨에 대한 지지도를 가지는 모든 빈발 패턴들의 집합을 추출 한다. (효율적인 빈발 패턴 탐사를 위해 FP-growth 방법을 확장하여 이용한다.)

② 분류단계 : 클래스 라벨이 없는 속성-값의 벡터 $A=\{a_1, a_2, \cdots, a_n\}$을 확률 $P(c_i \mid A)$이 최대인 클래스 c_i에 할당 한다($c_i \in C=\{c_1, c_2, \cdots, c_m\}, 1 \le i \le m$).

확률 $P(c_i \mid A)$은 서로 다른 approximation들을 사용하는 것으로 추정할 수 있다[12]. 각 approximation은 속성들에 대해 서로 다른 조건 독립 가정을 나타낸다. 예를 들어 $P(a_1a_2a_3a_4c_i)$은 확률의 곱 $P(a_1a_2c_i)P(a_3a_4 \mid a_1c_i)$ 또는 $P(a_1a_2c_i)P(a_4 \mid a_2c_i)P(a_3 \mid a_1a_4c_i)$으로 계산되며, 이 때, 두 확률을 "*Product Approximation*"이라고 한다. 여기서 특정 Product Approximation의 선택은 학습 단계에서 생성된 빈발 패턴들을 선택하는 것과 동등하다. Product Approximation의 선택을 위한 패턴들의 선택 전략은 다음의 3가지 원리를 이용한다.

① 선택 될 패턴들은 빈발해야만 하고 그러한 패턴을 "*frequent pattern*" 또는 "*frequent itemset*"이라고 부른다.

② 가능한 많은 빈발 패턴들이 사용되어야 한다.

③ 각각의 패턴은 가능한 긴 패턴들을 선택한다.

만약 FB가 product approximation을 선택하기 위해 1-항목의 패턴들로만 구성된다면, 그것은 NB와 같다.

4.2 빈발 패턴 탐사를 위한 CFP-growth

FB의 훈련단계에서는 빈발 패턴 집합들의 생성을 위해 2장에서 소개된 FP-growth 알고리즘을 클래스 라벨의 분포를 고려하기 위해서 CFP(Class Frequent Pattern)-growth로 확장한다.

훈련데이터의 집합을 D, 항목들의 집합(패턴)을 A라 한다. 그리고 c_i를 클래스 라벨이라고 한다.

[정의 1] 패턴 A의 클래스 지지도 (Class support) : D에서 A와 c_i를 포함하는 수이며, 데이터 집합 D에서의 클래스에 대한 패턴 A의 클래스 지지도(Sup_i)는 다음과 같다.

$$A.sup_i = \frac{count(A, c_i)}{|D|} = P(A, c_i) \quad \blacksquare$$

[정의 2] 패턴 A의 지지도(Support) : 데이터 집합 D에서 A의 지지도는 모든 클래스 c_i에 대한 $A.sup_i$들의 합이다.

$$A.sup = \frac{\sum_i count(A, c_i)}{|D|} = P(A) \quad \blacksquare$$

4.2.1 CFP-tree 생성 알고리즘

클래스 라벨을 가진 훈련데이터에 대한 빈발 패턴 탐사를 위한 FP-tree 알고리즘의 확장은 다음과 같다.

① CFP-tree는 각 노드 항목들의 카운트 값과 항목들의 클래스 지지도 카운트를 유지하기 위해 해더 테이블에 클래스 지지도 카운트를 추가한다.

② 트리에 삽입되는 모든 패턴들은 항목들의 카운트와 해당 클래스에 대한 카운트 값이 추가된다.

(그림 5)와 (그림 6)은 FP-tree를 확장한 CFP-tree 구성 알고리즘이다.

Input : (1) training data set D ; (2) minimum support *min_sup*.
Output : ***CFP-tree*** corresponding to D and satisfying *min_sup*.
Method :
(1) Scan D once and collect the set of frequent items F and their supports.
Sort F in support descending order as L, the *list* of frequent items.
If several items have the same support, and their names are numbers, sort the items in ascending order of their names.
(2) Create the root R of a new **CFP-tree** and label it as "null". Create frequent-item header table with $|F|$ entries. Set all head of node-link pointers to *null*.
(3) **for** each transaction data $d \in D$ **do** { // Read D the second time.
(4) c_d = class of transaction data d ;
(5) Select only frequent items of d into a record P ;
(6) Sort P in the order of L ;
(7) Call *insert_tree* (P, c_d, R) ;
(8) }

(그림 5) CFP-tree 생성 알고리즘

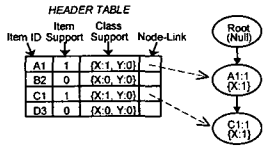
```
Procedure insert_tree (P, c, R) {
    Let P = [p | P-p], where p is the first element of P, and
    P-p is the remaining list.
 1) if R has a child N such that N.item_name=p then
 2)   N.count = N.count + 1 ;
 3)   N.count (c ) = N.count (c ) + 1 ;
 4) else {
 5)       create a new node N ;
 6)       for all classes c i do
 7)           if c i = c then
 8)             N.count = 1 and N.count (c i ) = 1 ;
 9)           else
10)             N.count = 0 and N.count (c i) = 0 ;
11)       N.item_name = p ;
12)       N.parent = R ;
          // Make node-link of p point to the first item with the
             same item_name.
13)       N.node-link = H (p ).head ;
14)       H(p).head = N ;
15) }
    // Increase by 1 count of item p in frequent-item header
       table H.
16) H (p).count = H (p).count + 1 ;
17) H ( p).count (c i) = H (p).count(c i) + 1 ;
18) if P-p ≠ φ then
19)   Call insert_tree(P-p, c, N) recursively.
 }
```

(그림 6) insert_tree() 프로시저

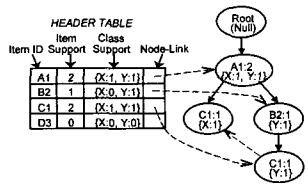
예를 들어 <표 3>이 훈련데이터 집합이고 최소지지도가 2일 경우, CPF-tree의 구성은 다음과 같다.

〈표 3〉 훈련데이터 집합에 대한 삽입될 패턴 예

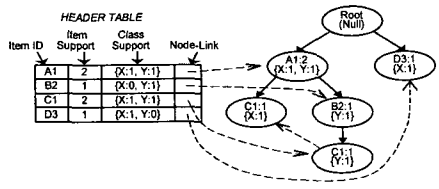
R-ID	A	B	C	D	CLASS	INSERT PATTERNS
1	A1	B1	C1	D1	X	(A1, C1) : X
2	A1	B2	C1	D2	Y	(A1, B2, C1) : Y
3	A2	B3	C2	D3	X	(D3) : X
4	A1	B2	C3	D3	Y	(A1, B2, D3) : Y
5	A1	B2	C1	D3	Y	(A1, B2, C1, D3) : Y



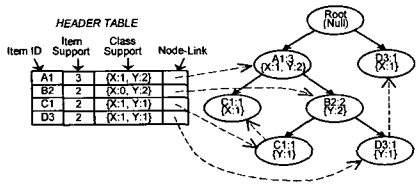
(a) Inserting pattern (A1, C1) : X



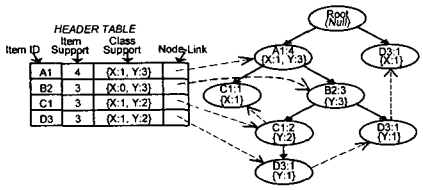
(b) Inserting pattern (A1, B2, C1) : Y



(c) Inserting pattern(D3) : X



(d) Inserting pattern(A1, B2, D3) : Y



(e) Inserting pattern(A1, B2, C1, D3) : Y

(그림 7) CPF-tree 구성 예

첫 번째로 최소지지도를 만족하는 빈발 1-항목들을 카운트를 기준으로 내림차순 정렬한다. 결과 집합은 L = [A 1 : 4, B2 : 3, C1 : 3, D3 : 3]이고 <표 3>의 INSERT PATTERNS에 트리에 삽입될 항목들의 순서와 클래스를 L의 기준에 따라 나타낸다. <표 3>의 첫 번째 삽입될 항목집합 {(A1, C1) : X}는 스캔을 통해서 두 노드 <A1 : 1(X : 1), C1 : 1(X : 1)>를 가지는 트리의 첫 번째 가지를 생성하고 항목에 대한 카운트와 클래스에 대한 카운트를 1로 할당한다((그림 7)(a)). 두 번째 삽입될 항목집합, {(A1, B2, C1) : Y}는 (그림 7)(b)와 같이 A1의 항목 카운트를 1증가 시키고 B2, C1에 대한 새로운 두 노드를 생성한다. 클래스 지지도에 대해 이 항목집합은 Y 클래스를 가지므로 클래스 Y에 대한 카운트를 각각 1로 할당한다.

세 번째 항목집합 {(D3) : X}는 루트에서 새로운 가지를 생성하고 D3에 대한 카운트, 클래스 X에 대한 카운트를 각각 1로 할당한다. 같은 방식으로 네 번째와 마지막 항목집합에 대해 위의 방법과 같이 빈발 CFP-tree를 생성한다. <표 3>의 훈련데이터에 대한 트리의 구성은 (그림 7)과 같다.

4.2.2 CFP-tree로부터의 빈발 패턴 마이닝

CFP-tree로부터의 빈발 패턴들의 생성은 FP-tree와 유사한 방식으로 분할-정복 기법을 적용하지만 차이점은 빈발 패턴 생성 시 각 클래스에 대한 빈발 패턴의 지지도를 유지한다는 것이다. (그림 7)(e) 트리로부터 빈발 패턴 탐사 과정의 예를 들면, 먼저 트리의 리프 노드인 D3에 대해 생성되는 조건부 패턴 베이스는 [A1, B2, C1 : 1(Y : 1)], [A1, B2 : 1(Y : 1)]이 된다. 여기서 C1의 카운트는 1이고 최소지지도를 만족 못하므로 삭제된다. 따라서 패턴 베이스로부터 조건부 CFP-tree는 <A1 : 2(Y : 2), B2 : 2(Y : 2)>이 되며, 빈발 패턴 집합으로 {A1 D3 : 2(Y : 2), B2 D3 : 2(Y : 2), A1 B2 D3 : 2(Y : 2)}을 생성한다. 노드 C1에 대한 조건부 패턴 베이스는 [A1 : 1(X : 1)], [A1, B2 : 2(Y : 2)]이 되고 조건부 CFP-tree는 <A1 : 3(X : 1, Y : 2), B2 : 2(Y : 2)>이 된다. 따라서 생성되는 빈발 패턴은 {A1 C1 : 3(X : 1, Y : 2), B2 C1 : 2(Y : 2), A1 B2 C1 : 2(Y : 2)}이다. 위의 분석 방법과 동일하게 B2에 의한 빈발 패턴은 {A1 B2 : 3(Y : 3)} 된다. (그림 8)은 빈발 패턴 탐사를 위한 CFP-growth 알고리즘이다.

Input : (1) CFP-tree Tree for training data D ;
(2) minimum support *min_sup*
Output : a complete set of frequent patterns *F* and their class count *F.count*$_i$
Method : Call *CFP-growth* (*Tree*, *null*), where null is the initial frequent suffix.

Procedure CFP-growth(*Tree, b*)
1) *F* = { *all 1-itemsets* } ;
2) **if** *Tree* contains a single path *P* **then** {
3) **for each** combination *a* of the node in *P* **do** {
4) generate pattern $p = a \cup b$;
5) *sup* (*p*) = minimum support of nodes in *a* ;
6) $F = F \cup p$;
7) }
8) }
9) **else**
10) **for each** a_i in the header of Tree starting from the least frequent **do** {
11) generate pattern $a = a_i \cup b$;
12) *sup* (*a*) = *sup*(a_i) ;

```
13)    F = F ∪ a ;
14)    construct a's conditional pattern base ;
15)    construct a's conditional CFP-tree Tree_a ;
16)    if Tree_a ≠ ∅ then
17)       call CFP-growth(Tree_a, a) ;
18) }
 }
```

(그림 8) CFP-growth() 프로시저

4.3 빈발 패턴에 의한 분류 단계

4.2에서 탐사된 모든 빈발 패턴의 집합 F에서 속성들 사이에 유용한 정보를 가진 패턴(long itemsets)들의 선택하여 새로운 case d를 분류한다. 이 때 선택되어진 빈발 항목 집합들은 case d의 부분 패턴이다(단 F의 1-itemsets은 빈발하거나 또는 빈발하지 않은 항목집합이다. 왜냐하면 탐사된 2-itemsets 이상의 빈발 패턴들이 d의 모든 항목들에 매치되지 않을 경우 1-itemset을 이용하며 이것은 Naive Bayes와 동일하다.).

[정의 3] 빈발 패턴 집합 F의 경계(border), B는 d의 부분 패턴들로 F안에 존재하는 가장 긴 항목 집합들로 구성된다. ■

분류를 위한 패턴 선택으로는 단지 경계 B의 패턴들만을 이용한다. 예를 들어 5가지의 속성을 case, $d = \{A1, B2, C1, D3, E2\}$을 분류한다고 가정하면, 빈발 패턴들의 집합 F는 이미 훈련 단계에서 모두 생성된다((그림 9)). 또한 (그림 10)에는 F에서 패턴 A의 경계 B를 나타낸다.

$(A1), (B2), (C1), (D3), (E2)$
$(A1, B2), (A1, C1), (A1, D3), (A1, E2), (B2, C1),$
$(B2, E2), (C1, D3), (C1, E2), (D3, E2),$
$(A1, B2, C1), (A1, D3, E2)$

(그림 9) 빈발 패턴 집합 F

$(B2, E2), (C1, D3), (C1, E2)$
$(A1, B2, C1), (A1, D3, E2)$

(그림 10) case d에 대한 B

4.3.1 PA의 구성

분류 대상인 case d에 대한 경계 B가 결정되면, FB는 모든 클래스 c_i에 대한 $P(A, c_i)$의 PA(Product Approximation)를 구성하기 위해 B의 빈발 패턴들을 사용한다.

[정의 4] case d가 경계 B의 한 패턴 l의 모든 항목집합을 모두 포함한다면 l은 d에 의해 포함 된다고 정의하고 l을 *covered*라고 한다. ■

PA는 [정의 4]에 의해 case d에 포함되어지는 B의 모든 부분 패턴들의 확률을 곱하는 것으로 구성된다. 다음은 (그림 10)의 경계 패턴들을 이용하여 case, $d = \{A1, B2, C1, D3, E2\}$에 대한 유효한 PA를 구성하는 경우의 예를 나타낸다.

① $(A1, B2, C1), (A1, D3, E2) \rightarrow$
$P(A1, B2, C1, c_i)P(D3, E2|A1, c_i)$

② $(A1, B2, C1), (B2, E2), (A1, D3, E5) \rightarrow$
$P(A1, B2, C1, c_i)P(E2|B2, c_i)P(D3|A1, E2, c_i)$

③ $(A1, B2, C1), (B2, E2), (C1, D3) \rightarrow$
$P(A1, B2, C1, c_i)P(E2|B2, c_i)P(D3|C1, c_i)$

④ $(B2, E2), (C1, E5), (A1, B2, C1), (A1, D3, E2) \rightarrow$
$P(B2, E2, c_i)P(C1|E2, c_i)P(A1|B2, C1, c_i)$
$P(D3|A1, E2, c_i)$

위의 예에서처럼 case d에 대한 PA의 구성은 d의 항목들을 모두 포함하는 부분 패턴들로 이루어진다. PA의 구성을 위한 B의 패턴 l의 선택은 [정의 4.3]의 세 기준에 의해 선택된다. PA 구성을 이용한 분류 방법에 대한 알고리즘은 (그림 11)이고 패턴 선택 조건에 대한 프로시저는 (그림 12)에 나타내었다.

[정의 5] PA의 구성을 위한 패턴 l의 선택 조건

Condition 1) $|l - covered| \geq 1$
Condition 2) $|l_k - covered| \leq |l_j - covered|$
Condition 3) $|l_k| \geq |l_j|$

위의 조건에서 l_j 대신 l_k을 선택한다. ■

Input : The set F of discovered patterns a new instance A
Output : the classification c_i of A

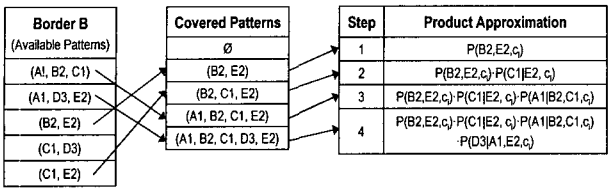
```
1)  COV = ∅ // the subset of A already covered
2)  NOM = ∅ // set of patterns in nominator
3)  DEN = ∅ // set of patterns in denominator
4)  B={l ∈ F | l⊆A and ∄l' ∈ F : l' ⊆ A and l' ⊂ l}
5)  for (k = 1 ; COV ⊂ A ; k++) {
6)      l_k = selectNext(COV, B) ;
7)      NOM = NOM ∪ {l_k} ;
8)      DEN = DEN ∪ {l_k∩COV} ;
9)      COV = COV ∪ l_k ;
10) }
11) Output that class c_i with maximal P(A, c_i) computed as :
        P(A, c_i) = P(c_i) ∏_{l∈NOM} P(l, c_i) / ∏_{l∈DEN} P(h, c_i)
```

(그림 11) 알고리즘 Classify

selectNext(COV, B)
T = { $l \in$ B : | l-covered | $\geq$ 1} ;
Return a patterns $l_k \in$ T such that for all other patterns $l_j \in$ T :
a) | l_k - covered | < | l_j - covered |, or
b) | l_k - covered | = | l_j - covered | and | l_k | > | l_j |

(그림 12) 패턴 선택 조건을 위한 selectNext() 프로시저

예를 들어 case $d=\{A1, B2, C1, D3, E2\}$에 대한 확률 $P(A1, B2, C1, D3, E2, c_i)$의 PA 구성 방법은 (그림 13)과 같다.



(그림 13) 증가적인 PA의 구성 단계

초기 *covered* 항목과 선택된 패턴은 없다. 다음으로 1단계에서 조건 1과 조건 2를 만족하는 패턴들은 $(B2, E2)$, $(C1, D3)$, $(C1, E2)$고 모두 같은 크기의 패턴들이다. 임의적으로 가장 처음에 있는 패턴 $(B2, E2)$을 선택했다고 가정하면, 이 패턴의 두 항목은 *covered*로 된다. 다음으로 가장 적은 *non-covered*를 가지는 패턴 $(C1, D3)$을 선택하고 *covered*에 $C1$을 추가시킨다.

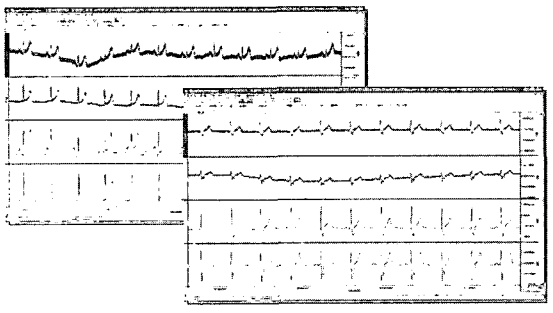
3단계에서 $(C1, E2)$, $(A1, B2, C1)$은 같은 수의 non-covered를 가진다. 그러나 조건 3에 의해 더 많은 항목을 가진 $(A1, B2, C1)$이 선택된다. 같은 분석에 의해 4단계에서는 $(A1, D3, E2)$가 선택되고 *covered patterns*가 case d의 모든 요소를 포함하므로 PA의 구성을 종료한다.

5 실험 및 평가

논문에서 제안한 알고리즘들에 대한 시스템 구현 환경은 다음과 같다.

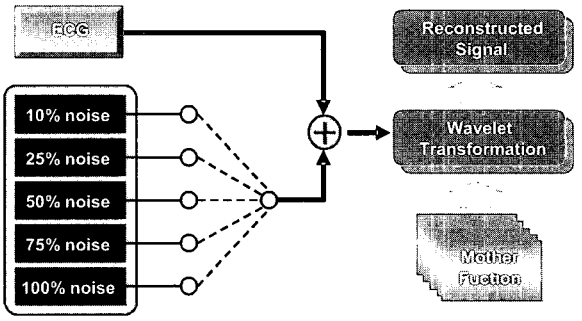
웨이블릿 변환 및 심전도 신호 재생성, R-peak 및 QRS complex 검출 및 수정, ST-segment 벡터 추출들은 MATLAB 상에서 구현 되었다. 심전도 데이터의 저장을 위한 DBMS는 오라클을 사용하였고, 운영체제는 Windows XP를 사용하였다.

(그림 14)는 특정 환자 A, B의 심전도 데이터를 raw 데이터(상위 2개 채널)와 웨이블릿 변환을 이용한 전처리 후의 데이터(아래 2개 채널)를 나타낸 것이다.

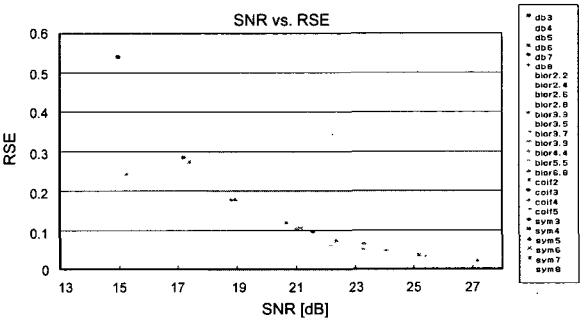


(그림 14) 환자 A, B의 전처리 후의 심전도 데이터

심전도 신호에 적합한 모 웨이블릿 함수의 특성을 평가하기 위해 신호대 잡음비와 재생신호 자승오차를 계산하였다. 전체 적인 시스템 구조는 (그림 15)와 같이 나타내었다. 심전도 신호의 전처리 과정에서 원신호의 왜곡을 최소화하여 기저선을 제거 할 수 있는 웨이블릿 모함수를 결정하기 위하여, European ST-T 데이터베이스의 심전도 신호에 다양한 웨이블릿 모함수를 적용하여 기저선을 제거하였으며, 제거효율을 평가하기 위하여 SNR과 RSE를 계산하였다. 실험결과 가장 우수했던 웨이블릿 모함수는 db8(diff. : 27.12), coif5(diff. : 25.32), sym7(diff. : 25.13)이었으며, diff.(meanSNR - meanRSE)의 값이 23미만으로는 심전도의 진단 파라미터까지 왜곡 시키므로 사용할 수 없다는 것을 알 수 있었다. 이러한 결과 비교를 (그림 16)에 나타내었다.



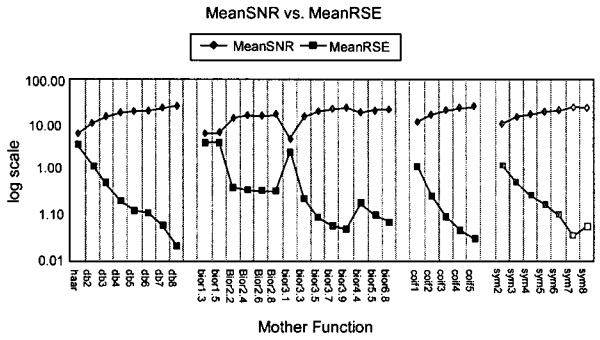
(그림 15) 다양한 웨이블릿 모 함수를 이용한 전처리 과정 구조도



(그림 16) 웨이블릿 모 함수들의 실험 결과 비교

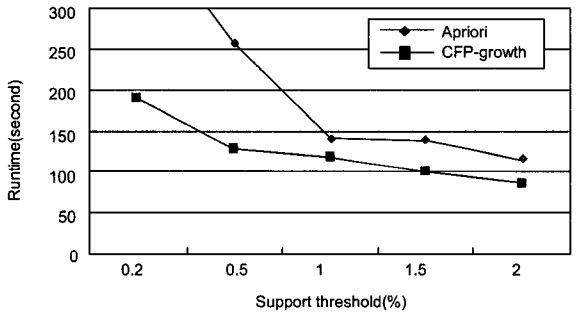
재생신호 자승오차에 대해서도 평가한 결과 역시 신호대 잡음비와 유사하였다. 신호대 잡음비가 비교적 우수했던 db 8, bior3.9, coif5의 웨이블릿 모함수는 재생신호 자승오차에서도 우수한 결과였고, 비교적 취약한 haar, bior1.3, bior1.5, bior3.1 웨이블릿 모함수는 재생신호 자승오차에서도 좋은 결과를 얻지 못하였다.

(그림 17)에서는 각각의 모함수에 대하여 MeanSNR과 MeanRSE의 간격이 가장 멀리 떨어져 있는 것이 심전도 신호처리에 적합한 웨이블릿 모함수이다. 이와 같은 방법으로 본 연구를 통하여 심전도 신호처리에 적합한 웨이블릿 모함수를 확인할 수 있었다.



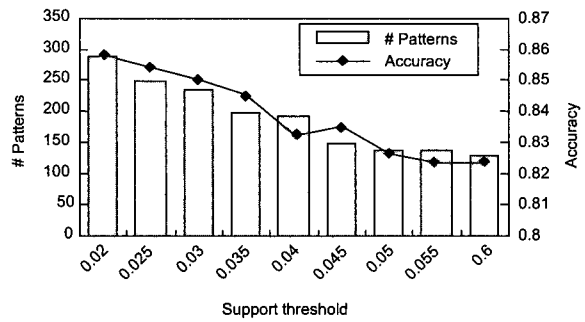
(그림 17) SNR의 평균과 RSE평균의 비교

심근 허혈 질환 분류를 위한 실험에서의 데이터는 European ST-T 데이터베이스의 e0105, e0111을 선택하였으며, European ST-T 데이터베이스로부터 ST-segment에서 임상 소견 중 정상(normal)과 비정상(ST episode를 갖는 데이터)으로 분류하였다. 실험 결과 (그림 18)은 빈발 패턴 탐사를 위해 Apriori와 논문에서 제안된 CFP-growth 알고리즘과의 실행 시간 비교이다. 비교를 위한 데이터 집합은 총 100,000개의 데이터와 최대 빈발 패턴의 길이를 5로 하였다.

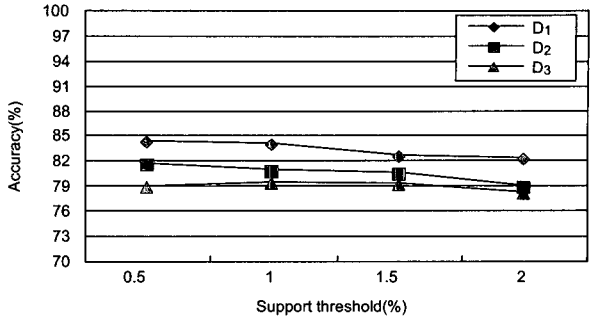


(그림 18) 지지도 변화에 따른 실행 시간

빈발 패턴들은 지지도의 변화에 따라 그 생성 수가 다르게 나타나므로 (그림 19)의 실험에서는 지지도 변하에 따른 빈발 패턴 수의 변화를 실험하였다. 또한 제안된 FB 분류기는 생성된 빈발 패턴에 기반 하여 확률들을 PA의 구성을 통해 이루어지므로 지지도 변화에 따른 분류기의 정확성을



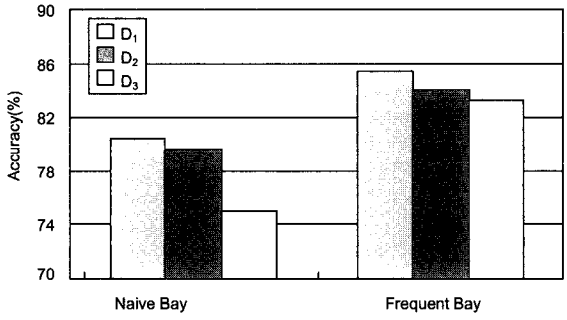
(그림 19) 지지도 변화에 따른 탐사된 빈발 패턴 수와 정확도 비교



(그림 20) 특징 벡터 그룹에 대한 지지도 변화에 따른 정확도 비교

나타내었다. 그 결과 PA 구성을 위한 빈발 패턴들의 수, 즉 가용 확률 데이터의 증가는 분류기의 성능향상에 영향을 미친다는 것을 알 수 있었다. 실험에서 사용된 ST-segments 특징 벡터는 3가지를 사용하였고 (그림 20)에서처럼 D_1 특징 벡터에서 가장 좋은 결과를 얻었다.

마지막으로 (그림 21)은 NB와 FB의 성능 비교를 나타내고 ST-segments 데이터에 대한 분류 결과 FB가 더 정확함을 나타낸다.



(그림 21) 데이터에 대한 NB와 FB의 비교

이 실험에서 NB는 클래스 조건 독립을 가지므로 확률 추정을 위해 단지 1-itemsets만을 사용하는 것으로서 NB 분류기를 생성하였다(FB 생성 시 사용된 빈발 패턴이 단지 1-itemsets로만 구성된다면 이것은 NB와 동등하다).

6. 결 론

심혈관계 질환 진단을 위한 심전도 측정은 이미 오래전부터 사용되어져 왔으며 최근에 이르러 심전도 데이터 처리에 IT 기술을 적용하여 보다 빠르고 정확한 심혈관계 질환을 판별하고 있다.

이 논문에서는 심전도 데이터의 수집과 심전도 신호의 원신호 왜곡을 최소화하기 위한 웨이블릿 변환, 심전도 시그널에서의 QRS 및 R-peak 검출과 심근허혈 진단에 필요한 ST-segments 특징벡터를 추출하였다. 그리고 이 특징벡터를 이용하여 심장질환을 판별하기 위한 빈발 패턴 베이지

안 분류 기법을 제안하였다. 제안된 분류 기법은 나이브 베이지안에 데이터가 가진 속성 종속성을 고려하여 빈발 패턴 마이닝을 적용한 분류기를 생성한다. 빈발 패턴 탐사 단계에서는 마이닝의 성능향상을 위해 FP-growth 방법에 노드 클래스 분포를 고려하여 CFP-growth로 확장하였다. 심전도 패턴의 분류 실험 평가에서 심전도 신호에 적합한 모웨이블릿 함수의 특성을 평가하였고, 패턴 분류 결과 빈발 패턴 베이지안이 기존의 나이브 베이지안 분류 기법을 적용한 것보다 정확한 것을 알 수 있었다.

참 고 문 헌

[1] 박광리, "스트레스 심전도의 잡음 제거를 위한 WAF와 WIF의 설계", 연세대학교 의용전자공학과 박사논문, 2000.

[2] 최형민, 김원식, 정광일, 황재호, "웨이브렛 변환을 이용한 심전도의 기저선 제거", 한국감성과학회 춘계학술대회논문집, pp.26-31, 2003.

[3] Vladimir Cherkassky, Steven Kilts, "Myopotential denoising of ECG signals using wavelet thresholding methods," Neural Networks, Vol.14, pp.1129-1137, 2001.

[4] N. Maglaveras, T. Stamkopoulos, K. Diamantaras, C. Pappas, M. Strintzis, "ECG pattern recognition and classification using non-linear transformations and neural networks : A review," International Journal of Medical Informatics, Vol.52, pp.191-208, 1998.

[5] Biju P. Simon and C. Eswaran, "An ECG Classifier Designed Using Modified Decision Based Neural Networks," Computers and Biomedical Research, Vol.30, No.4, pp.257-272, 1997.

[6] 김만선, 김원식, 노기용, 이상태, "심전도 패턴을 분류하기 위한 신경망 성능 평가", 한국감성과학회 춘계학술대회, pp.148-153, 2003.

[7] R. Agrawal and R. Srikant, "Fast Algorithm Mining Association Rules in Large Database," In Proc. of the 1994 Internat'l Conference on VLDB, 1994.

[8] J. Han, J. Pei, Y. Yin, "Mining frequent patterns without candidate generation," In SIGMOD'00, Dallas, TX, May, 2000.

[9] J. Han, M. Kanmer, "Data Mining : Concepts and Techniques," Morgan Kamfmann Publishers, 2000.

[10] N. Friedman, D. Geiger, M. Goldszmidt, "Bayesian Network Classifiers," Machine Learning, 29, pp.131-163, 1997.

[11] P. Domingos, M. Pazzani, "On the optimality of the Simple Bayesian Classifier under Zero-One Loss," Machine Learning, 29, pp.103-130, 1997.

[12] P. M. Lewis, "Approximating Probability Distributions to Reduce Storage Requirements," Information and Control, 2, pp.214-225, 1959.

노 기 용

e-mail : kyno@kriss.re.kr

1981년 충남대학교 물리학과(이학사)

1995년 충남대학교 전산학과(이학석사)

2004년 충북대학교 전산학과(이학박사)

1988년~현재 한국표준과학연구원

관심분야 : DB설계, Image Processing, ATM 등

김 원 식

e-mail : wskim@kriss.re.kr

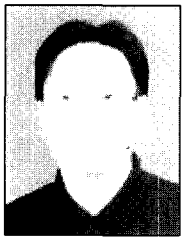
1979년 아주대학교 전자공학과(공학사)

1984년 고려대학교 물리학과(이학석사)

2004년 연세대학교 의공학과(의공학박사)

1984년~현재 한국표준과학연구원

관심분야 : 생체신호 측정, 처리, 해석 등



이 헌 규

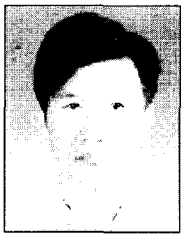
e-mail : hglee@dblab.chungbuk.ac.kr

2002년 경기대학교 전자계산학과(이학사)

2004년 충북대학교 대학원 전산학과 (이학석사)

2004년~현재 한국표준과학연구원 위촉 연구원

관심분야 : 시간 데이터베이스, 시공간 데이터베이스, 데이터 마이닝 등



이 상 태

e-mail : stlee@kriss.re.kr

1977년 아주대학교 전자공학과(공학사)

1992년 전북대학교 전자공학과(공학석사)

1998년 전북대학교 전자공학과(공학박사)

1985년~현재 한국표준과학연구원

관심분야 : 지능망, 광대역통신망, 트래픽 제어 등

류 근 호

e-mail : khryu@dblab.chungbuk.ac.kr

1976년 숭실대학교 전산학과(이학사)

1980년 연세대학교 공학 대학원 전산전공 (공학석사)

1988년 연세대학교 대학원 전산전공 (공학박사)

1976년~1986년 육군 군수 지원사 전산실(ROTC장교), 한국전자통신연구소연구원, 한국 방송대학교 전산학과 조교수

1989년~1991년 University of Arizona, Research Staff (TempIS 연구원, Temporal DB)

1986년~현재 충북대학교 전기전자및컴퓨터공학부 교수

관심분야 : 시간 데이터베이스, 시공간 데이터베이스, Temporal GIS, 지식기반 정보검색 시스템, 데이터 마이닝 및 데이터베이스 보안, 바이오 인포메틱스