감성과학, Vol. 10, No. 2, pp. 147-154. June 2007

심장 질환 진단을 위한 데이터 마이닝 기법

Data Mining Approach for Diagnosing Heart Disease

노기용* · 이헌규**† · 류근호**

Kiyong Noh* · Heon Gyu Lee**† · Keun Ho Ryu**

한국표준과학연구원*

Korea Research Institute of Standards and Science

충북대학교 데이터베이스연구실**

Database/Bioinformatics Laboratory, Chungbuk National University

Abstract : Electrocardiogram(ECG) 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 in the con due to inaccuracy of domestic diagnosis results for a heart disease. This paper proposes ST-segment extraction technique diagnosing heart disease parameter from raw ECG data. As the ST-segment is used for prediction of Coronary Artery Disease, we can predict heart disease using classification approach in data mining technique. We can also predict patient's clinical characterization from patient clinical data.

Key words : ECG, ST-segment, data mining, classification, association rule

요약 : 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 국내 진단 결과의 부정확성 때문에 외국의 진단 알고리즘을 사용하고 있다. 이 논문에서는 원시 심전도 데이터로부터 심장 질환 진단의 파라미터인 ST-segment 추출 방법을 제안한다. ST-segment는 관상동맥 질환 예측에 활용되므로 데이터마이닝의 분류기법을 적용하여 질환을 예측한다. 또한 연관규칙 마이닝을 통해 환자들의 임상 데이터로부터 심장 질환자들의 임상적 특징을 예측한다.

주제어 : 심전도, ST-segment, 데이터마이닝, 분류, 연관규칙

† 교신저자 : 이헌규(충북대학교 데이터베이스연구실)

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

TEL : 016-351-3779, 043-267-2254

FAX : 043-275-2254

1. 서론

심전도(Electrocardiogram: ECG)를 이용한 심장 관련 질환 알고리즘에 대한 연구가 지난 수년 동안 많이 진행되어 왔다. 심전도란 심장의 상태를 비관혈적으로 진단하는 매우 중요한 수단으로 활용되며, 진폭의 수와 주파수를 이용한 생체 전위 신호 중의 하나이다[3]. 이 논문에서는 데이터마이닝 기술을 적용하여 심혈관계 질환자들의 임상정보로부터 질환에 영향을 주는 속성들의 연관관계를 분석하고, 원시 심전도 신호에서 심장 질환 진단의 중요 파라미터인 ST-segment를 추출한다. 추출된 ST-segment는 허혈성 심장 질환, 확장성 심근성, 비후성 심근증 진단에 활용되므로 데이터마이닝 기술 중에서 분류 기법들을 사용하여 질환 예측을 한다.

심장 질환자들의 임상정보와 심전도로부터 데이터마이닝 기술을 적용하기 위해 이 논문에서는 원시 데이터의 전처리 과정에서부터 심장 질환의 특징 분석과 자동 진단을 위한 연관규칙과 분류기법을 제안하며, 세부 내용은 다음과 같다.

- 심장 질환자들의 임상 정보와 심전도 데이터를 수집하여 분류한다. 심전도 데이터는 특징 벡터 추출을 위해 ST-segment의 경사와 면적을 추출하여 질환 진단을 위한 파라미터로 사용한다.
- 임상 데이터에 연관규칙(association rule)을 적용하여 환자들의 임상적 속성들의 모든 연관규칙을 추출한다.
- ST-segment의 특징 벡터를 이용하여 관상동맥 질환에 대해, 협심증(AP: Angina Pectoris) 환자, 급성관동맥 증후군(ACS: Acute Coronary Syndrome) 또는 Normal people로 진단을 위한 분류 기법을 적용하여 평가한다. 적용된 분류 기법으로는 Weka[4]의 결정트리, 베이지안 분류 그리고 연관적 분류 알고리즘들을 적용해 그 결과를 분석한다.

논문의 효과적인 이해를 위해서 이 논문의 구성은 다음과 같다. 먼저 관련 연구로써 2장에서는 심전도의 ST-segment를 이용한 심장 질환 패턴에 대해 기술하고 기존의 연관규칙과 분류 기법에 대해 설명한다. 3장에서는 심장 질환 분류 분석을 위한 임상 데이터와 ST-segment 특징 벡터들에 대한 전처리 과정으로 데이터의 이산화 및 정규화 과정을 설명한다. 4장에서는 전처리된 임상 데이터의 임상적 속성들의 상관성 분석을 위해 연관규칙 마이닝 적용을 하며, 심혈관계 질환 진단을 예측하기 위한 분류 기법의 적용 및 그 결과를 분석한다. 마지막으로 5장에서는 데이터마이닝 기술을 이용한 심장 질환 진단에 대한 논문의 결론을 맺는다.

2. 관련연구

이 절에서는 심혈관계 질환자의 특징 분석 및 진단을 위한 선행연구로써 심전도 신호와 ST-segment, 그리고 연관규칙과 분류 기법들에 대한 관련연구를 설명한다.

2.1 심전도의 ST-segment를 이용한 심장질환 패턴

심혈관계 질환은 다인자성 질환으로 여러 가지 변이가 복합적으로 질병발생과 진전에 영향을 미치며, 심혈관 질환의 위험요인으로 알려진 비만, 흡연, 식이요인 등의 다양한 환경적 요인과 상호작용에 의해 질병에 영향을 미친다. 심혈관계 질환, 특히 동맥경화의 진행으로 인한 허혈성 심장질환의 발생빈도는 서구 국가뿐 아니라 우리나라에서도 날로 증가하고 있으며 단일 질환군으로 전 국민 의료비의 11%를 차지하여 국가경제에 큰 영향을 미친다. 또한 심혈관계 질환은 발병 후 심각한 합병증 및 지속적인 치료가 요구되는 질환으로 질병의 예방이 중요하다. 고혈압 및 동맥경화성 질환 등 심혈관계 질환은 생활 습관이나 환경적인 영향과 함께 유전적 요인에 의해 질병 발생률에 차이를 보이고 있어 유전적 위험 요인의 규명으로 고위험군을 예측하고 이들에 대한 교육 및 환

경적 요인의 조절을 통한 질병 발생의 예방이 중요하게 인식되고 있다.

이러한 심혈관계 질환의 조기 발견과 예측을 위해 심전도는 심장의 상태를 비관혈적으로 진단하는 매우 중요한 수단으로 진폭의 수와 주파수를 이용하는 생체 신호 중의 하나이다. 국내 심전도 시스템에서는 심장질환 환자의 심장상태를 감시할 수 있는 홀터 시스템 사용이 늘어나면서 그 중요성도 높아지고 있으며, 그 외 12채널 진단 심전계, 스트레스 심전계 등의 심장관련 진단기기에 대한 연구가 활발히 진행되고 있다[3].

심장의 전기적 활성단계는 크게 심방 탈분극, 심실 탈분극, 심실 재분극 시기로 나뉘며, 이러한 각 단계는 다음 그림 1과 같이 P, QRS, T파라고 불리는 몇 개의 파의 형태로 구성된다. 이러한 파들은 표준 형태를 갖추어야 심장의 전기적 활성이 정상이라고 볼 수 있다. 심전도의 ST-segment는 elevation 또는 depression 되는 episode를 띠게 된다. 그림 1은 심전도 데이터에서 ST-segment, R-R간격, QRS complex, J point 등을 표현한 것이다[5, 6, 7, 8].

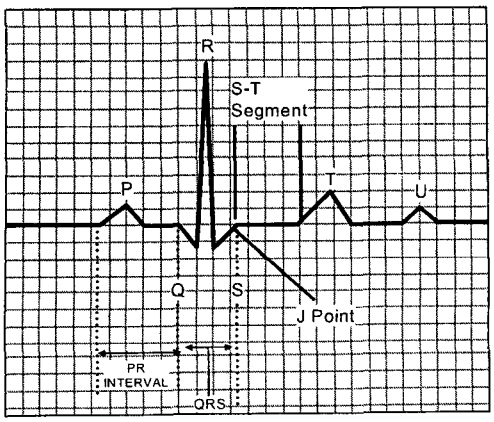

그림 1. 심전도 파형의 구성요소

2.2 연관규칙 마이닝

연관규칙이란 "어떤 사건이 일어나면 다른 사건이 일어난다"와 같은 연관성을 이야기한다[9, 10]. 주어진 트랜잭션의 집합에 대해서 연관규칙은 "A→B"로 표시할 수 있다. 이때 전체 항목들의 집합 A는 결론 항목들의 집합 B를 야기한다고 정의 할 수 있다. 연관규칙의 기본 개념은 아이템(item set)집합, 즉 트랜잭션의 집합이 있고 이 트랜잭션 집합을 I라고 표현했을 경우 공집합이 아닌 항목 집합 X, Y에 대해서 X⊂I, Y⊂I일 경우 가정 X라는 사건이 발생했을 때 결과 Y라는 사건이 발생한다라는 것을 말한다. 예를 들면 "편의점에서 목요일에 기저귀를 구매하는 고객들은 맥주도 동시에 구매 한다"는 연관성이 있음을 연관규칙 탐사의 결과로 얻어 낼 수 있는 것이다. 편의점에서 각 고객이 구매하는 물품들의 집합을 한 트랜잭션이라고 하고 이런 물품 구매에 관련한 트랜잭션이 일정한 기간 동안 데이터베이스에 저장돼 있다면 기저귀를 사는 사람이 맥주를 구매하는 연관성을 연관규칙으로 다음과 같이 표현 할 수 있다. 여기에서 의미하는 2%의 지지도(support)는 주어진 트랜잭션들 중에서 2%가 기저귀와 맥주를 동시에 구입한다는 것이고 30%의 신뢰도(confidence)라는 것은 기저귀를 사는 고객들 중에서 30%가 맥주를 산다는 것을 의미한다. 지지도와 신뢰도의 대한 정의는 식(1)과 식(2)이다.

$$Support(X,\ Y) = \frac{Count(X \cap Y)}{\text{전체 트랜잭션의 개수}} \quad \text{식(1)}$$

$$Confidence(X,\ Y) = \frac{Count(X \cap Y)}{Count(X)} \quad \text{식(2)}$$

연관규칙 탐사에서는 사용자가 지지도와 신뢰도의 값을 적절하게 명시해 트랜잭션 데이터베이스에서 사용자가 정한 정도의 모든 상호 연관성을 발견해 낼 수 있고, 구매 데이터베이스와 같은 성격의 데이터베이스에서 고객들의 구매 패턴을 찾을 수 있다.

2.3 분류 기법 마이닝

분류란 중요한 데이터 클래스를 설명하는 모형을 생성하거나 미래 데이터의 경향을 예측하고자 할 때 사

용되는 기법이다. 분류 기법에 대한 연구는 통계, 신경망, 결정트리 등의 분야에서 연구되었으며 의료진단 예측 수행, 선택적 마케팅 분야에서 응용된다.

SVM(Support Vector Machine)은 1998년 Vapnik에 의해 제안된 학습이론으로 분류 문제를 해결하기 위해서 최적의 분리 경계면인 hyperplane을 제공한다[9]. 최근, SVM이 주목 받는 이유로는 첫째, 이론적 근거에 기반하므로 결과 해석이 용이하고, 둘째, 실제 응용에 있어서 신경망 수준의 높은 성과를 내며, 셋째, 적은 학습자료만으로 신속하게 분류 학습을 수행할 수 있기 때문이다. 또한 SVM은 기존의 학습 알고리즘이 경험적 위험 최소화 원칙(empirical risk minimization)을 구현하는 것인데 비해, 구조적 위험 최소화 원칙(structural risk minimization)에 기반 하므로 overfitting을 피할 수 있다. 이 논문에서는 SVM의 커널 함수로서 가우시안 RBF(Radial Basis Function)를 사용한다.

베이지안 분류기[12, 13]는 통계적 분류기이다. 이것은 주어진 샘플이 특정 클래스에 속할 확률과 같이 어떤 항목이 특정 클래스에 속할 확률을 예측한다. 나이브 베이지안 분류기[14]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 분류기보다 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.

베이지안 네트워크(Bayesian network)는 최근 복잡한 도메인에서 불확실성을 해결하기 위한 강력한 데이터 마이닝 방법으로 부각되고 있다. 베이지안 네트워크는 결합 확률 분포를 이용하는 모델로 도메인 지식을 쉽게 반영할 수 있는 장점을 가지며 방향성 비순환 그래프의 형태를 취한다. 이 그래프에서 노드는 변수를, 노드간의 연결은 확률적인 종속관계를 의미한다. 따라서 베이지안 네트워크는 분류 문제를 속성 노드와 결과 노드간의 확률관계로 가정하며, 이로 인해 여러 가지 장점을 가진다[13].

의사결정트리[15, 16]는 데이터의 분류에 많이 사용되고 어떤 결과가 일어났을 때 그 결과가 일어나기 위해서 왜 그런 현상이 나타났는지를 설명한다. 결정트리는 훈련 샘플 데이터의 단일 노드로 시작하여, 샘플이 모두 같은 클래스라면, 노드는 잎이 되고, 해당 클래스로 분류한다. 그렇지 않으면 정보 이득(information gain)이라는 엔트로피 기반 척도를 사용하여 샘플들을 각각의 클래스로 가장 잘 분리하는 속성을 선택하기 위해 휴리스틱한 방법을 사용한다. 그러나 의사결정트리 또한 나이브 베이지안(naive bayesian)과 유사하게 항목들을 독립적으로 고려하며, 한번에 하나의 변수만을 조사하는 제약 사항을 가지고 있다. 따라서 이러한 단점을 연관적 분류 및 베이지안 네트워크를 통해 해결할 수 있다.

연관적 분류란 서로 독립적인 연관규칙과 분류규칙을 일부분 통합시킨 새로운 분류 방식으로 클래스를 예측하기 위해 연관규칙을 사용한다. 연관적 분류는 결정트리에 비해, 기존의 결정트리가 데이터 객체의 분류를 위해 단지 수백 개의 속성들만을 조작하는 것으로 제한되는 반면, 연관적 분류는 수천의 속성 차원을 조작할 수 있다는 것이고 Naive Bayes와 유사하게 항목들을 독립적으로 고려할 수 있다. 또한 탐사된 규칙은 단순성(*simplicity*)을 가지므로 사용자들은 규칙을 쉽게 이해할 수 있다.

3. 임상정보 및 심전도 데이터의 전처리

이 절에서는 임상 데이터와 원시 심전도로부터 ST-segment 특징벡터들의 추출에 대해 기술한다.

3.1 ST-segment 특징 벡터 추출

ST-segment 벡터들의 추출하기 위해서는 먼저, R-Peak와 QRS Complex 검출 프로그램을 Tompkins 알고리즘을 이용하여 그림 2와 같이 추출하였다[3,

4]. QRS complex는 5-30Hz의 주파수 성분을 갖기 때문에 변화하는 심전도 파형에 적응적인 문턱치 알고리즘을 적용하여 정확히 QRS를 검출한다. QRS complex 검출 후 R-peak를 검출하여 ST-segment의 시작점인 J point는 R-R간격이 600ms보다 클 경우는 J point=R+60ms, 작을 경우는 J point=R+40ms로 정의 한다. 또한 ST60과 ST80을 특징 벡터로 사용하였는데 R-R간격이 600ms보다 크면 ST60은 R+120ms로 하고 ST80은 R+140ms를 사용하며, 만약 600ms보다 작으면 ST60은 R+100ms로 하고 ST80은 R+120ms로 사용한다. 추가적으로, ST-segment의 기울기(경사)와 면적도도 추출하여 특징 벡터로 사용하며, 최종 분류기법 알고리즘의 입력 벡터 집합은 D={ST0, SLOPE, INTEGER, ST60, ST80}이다[1, 2].

연관규칙 적용을 위한 심장 질환자들의 임상 정보로는 표 1의 정보를 사용한다.

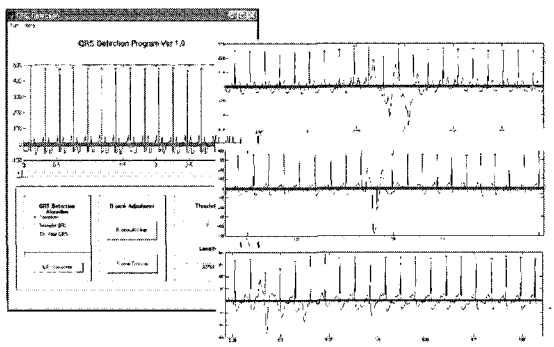

그림 2. R-peak와 QRS Complex의 검출

표 1. 연관규칙 적용을 위한 임상정보 리스트

속성	데이터 형식	설명
Gender	Boolean (M/F)	성별(남/여)
Age	Nameric	나이
Hyper Blood Pressure	Boolean (Yes/No)	고혈압(유/무)
Diabetes Millitus	Boolean (Yes/No)	당뇨병(유/무)
Smoking	Boolean (Yes/No)	흡연(유/무)
Old Myocardial Infarction	Boolean (Yes/No)	과거 심근경색 병력(유/무)
Ejection Fraction	Numeric	심실 펌프에 의해 방출된 혈액 양의 심실 확장 말기 용적에 대한 비율
Blood Glucose	Numeric	혈당
Total Cholesterol	Numeric	총 콜레스테롤
Triglyceride	Numeric	중성지방
Systolic Blood Pressure	Numeric	심상수축시의 혈압
Diastolic Blood Pressure	Numeric	심장확장시회 혈압
Hyperlipidemia	Boolean (Yes/No)	고지혈증(유/무)

3.2 임상 데이터와 심전도 특징 벡터의 전처리

심장 질환 환자들의 임상적 속성들의 연관성 탐사를 위한 연관규칙의 적용과 ST-segment 벡터로부터의 질환 진단을 위해서는 이산화와 정규화가 필요하다. 이산화란 연속적인 실수값의 데이터를 클래스의 분포를 고려하여 순환적으로 분할한다.

일반적으로 이산화를 하기 위한 알고리즘은 엔트로피 기반 척도를 사용한다. 엔트로피 기반 척도의 결과는 이산화가 되는데 이러한 이산화를 통해 그 속성의 수치 개념 계층이 형성된다[17]. 그림 3은 Weka 프로그램의 임상정보 및 ST-segment에 대한 엔트로피 기반 이산화 결과를 보여 주고, 그림 4는 임상정보 데이터에 대해 이산화된 각 속성에 대한 정수형 값으로 사상(mapping)시킨 정규화 결과이다.

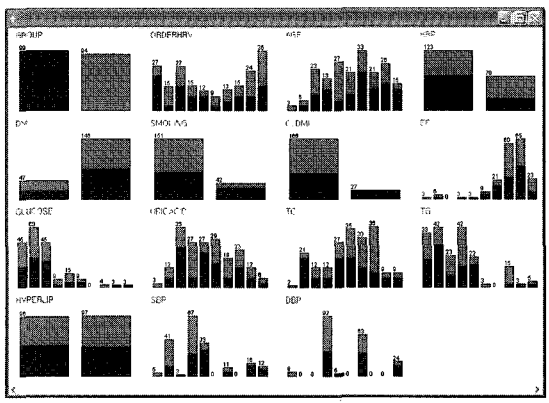

그림 3a. 임상 데이터에 대한 이산화 결과

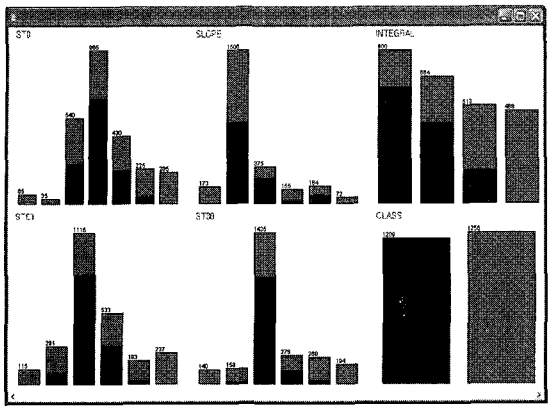

그림 3b. ST-segment 이산화 결과

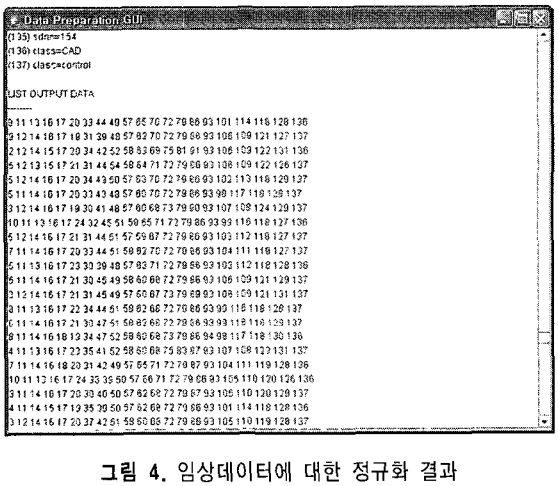

그림 4. 임상데이터에 대한 정규화 결과

4. 임상정보 및 ST-segment 분석 및 질환을 위한 데이터마이닝 기법의 적용

심장 질환 환자들의 임상적 속성들의 연관성 탐사를 위해 기존의 연관규칙 알고리즘 중 LUCS-KDD의 ApprioriT[18] 알고리즘을 적용하여 주어진 임계값(지지도, 신뢰도)에 대한 연관규칙을 추출하였다. 휴리스틱 방식으로 실험에 대한 최적 파라미터를 추정하였으며, 그 파라미터 값으로 지지도는 10%, 신뢰도는 80%로 하였다. 그림 5는 ApprioriT 알고리즘 수행 결과 후의 탐사된 연관 규칙이다.

예를 들어, 그림 5의 연관규칙들 중에서 규칙, R: ⟨16, 21⟩→⟨68⟩ 80.0%의 의미는 "나이가 45~50이고 혈당(blood glucose)이 81~85인 환자들 중에서 심장이완시 혈압(DBP)이 62~65일 가능성은 80%이다"[1] 란 의미이다.

ST-segment를 이용한 질환의 예측을 위해 적용된 분류 기법은 SVM, 의사결정트리로 C4.5, 베이지안 분류기로는 나이브 베이지안, 베이지안 네트워크를 알고리즘을 적용하였다. 마지막으로, 연관적 분류 기법은 CBA[19],cmAR[20] 알고리즘 적용을 하였다.

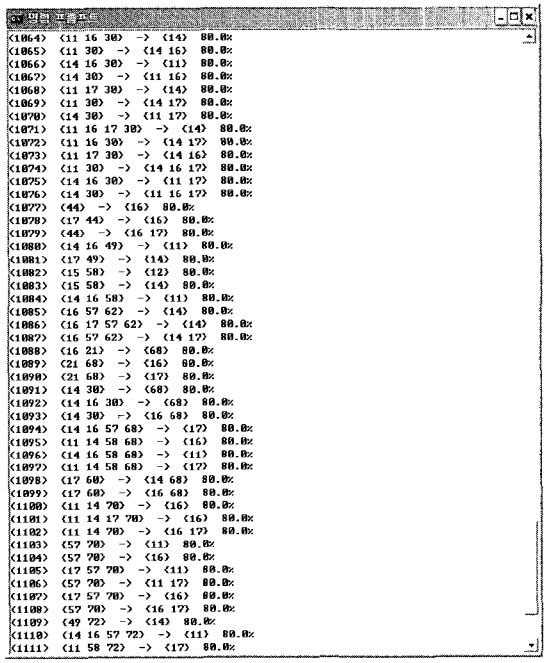

그림 5. ApprioriT 알고리즘 적용 결과

그림 6은 실험에 포함된 분류 기법 중 대표적인 의사결정트리 알고리즘인 C4.5를 적용했을 때, 생성된 트리이고 베이지안 네트워크 모델과 이 모델을 적용한 분류 결과는 그림 7과 같다.

1) 정규화된 속성값 ⟨16, 21, 68⟩의 의미는 ① 16: 나이(45~50), ② 21: 혈당(81~85), ③ 68: 심장이완시혈압(62~65)이다.

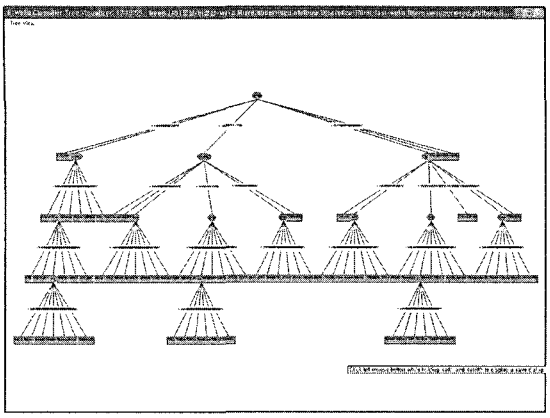

그림 6. C4.5 알고리즘 적용 결과 예제

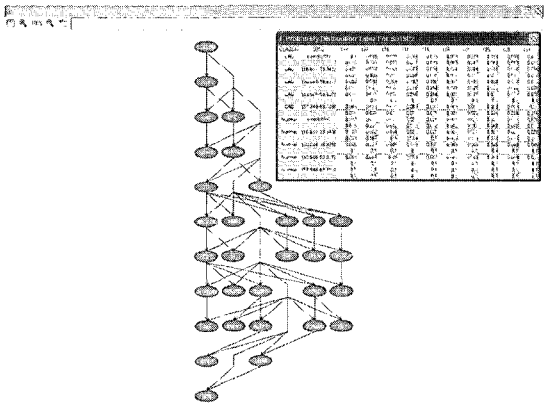

그림 7a. Bayesian Net 모델

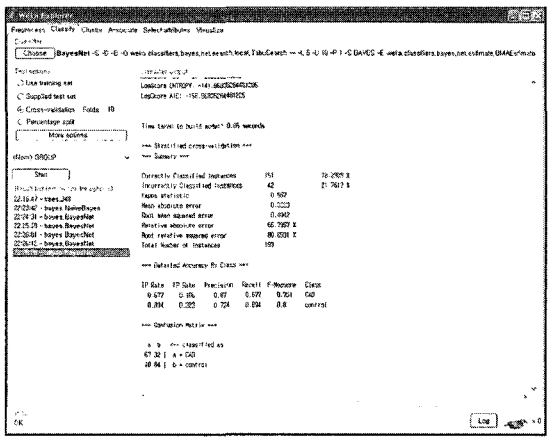

그림 7b. Bayesian Net의 적용 결과 예제

각각 적용된 분류 기법들에 대한 심장 질환 진단의 예측 결과 평가를 위한 지표로는 TP(True Positive)와 FP(False Positive), 그리고 Recall 및 Precision을 이용하였다.

TP의 의미는 분류하고자 하는 클래스를 정확하게 그 클래스로 분류했을 경우를 나타내고 FP는 분류하고자 하는 클래스가 아닌데도 그 클래스로 분류했을 경우의 오류, 즉 양성-오류를 의미한다. Recall과 Precision은 식(3), 식(4)에 의해 얻을 수 있다.

$$Recall = \frac{TP}{TP+FN} \quad \text{식(3)}$$

$$Precision = \frac{TP}{TP+FP} \quad \text{식(4)}$$

각 분류 결과에 대한 성능 평가 비교는 표 2에 요약하였다.

5. 결론

이 논문에서는 심장 질환의 임상적 분석과 자동적인 심혈관계 질환의 진단을 위해 데이터마이닝 기술을 적용하였다. 임상적 분석을 위해서는 연관규칙 마이닝을 적용하였으며, 탐사된 연관규칙들을 통해 환자들의 임상적 연관성을 찾을 수 있었다. 또한 원시 심전도 데이터에서 심장 질환 진단의 중요 파라이터인 ST-segment 특징벡터를 추출하여 기존의 대표적인 분류 기법 알고리즘들을 이용하여 질환을 진단하였다. 그 결과 연관적 분류 기법 중의 하나인 cmAR

표 2. 각 분류 기법에 대한 성능 평가

Classifier	TP	FP	Precision	Recall	Class
SVM	0.625 0.975 0.125	0.021 0.5 0.018	0.909 0.765 0.5	0.625 0.975 0.125	Normal AP ACS
C4.5	0.625 0.8 0.25	0.229 0.292 0.036	0.476 0.821 0.5	0.625 0.8 0.25	Normal AP ACS
Navie Bayesian	0.563 0.875 0.25	0.125 0.417 0.036	0.6 0.778 0.5	0.563 0.875 0.25	Normal AP ACS
Bayesian Network (TAN)	0.625 0.9 0.25	0.063 0.417 0.054	0.769 0.783 0.4	0.625 0.9 0.25	Normal AP ACS
CBA	0.563 0.875 0.125	0.146 0.375 0.054	0.563 0.795 0.25	0.563 0.875 0.125	Normal AP ACS
CMAR	0.9 0.563 0.25	0.542 0.063 0.018	0.735 0.75 0.667	0.9 0.563 0.25	Normal AP ACS

알고리즘이 가장 높은 성능을 보였고, 베이지안 분류, 의사결정트리 순으로 되었다.

참고문헌

[1] 김만선, 김원식, 노기용, 이상태 (2003), 심전도 패턴을 분류하기 위한 신경망 성능 평가, 한국감성과학회 춘계학술대회, 148-153.

[2] 김원식, 노기용, 류근호, 이헌규, 이상태 (2004), 심전도 패턴 판별을 위한 빈발 패턴 베이지안 분류, 정보처리학회논문지 D, 11-D, 5, 1031-1040.

[3] Conumel P. (1990), ECG: Past and Future, Annals NY Aca-demy of Sciences, 601.

[4] Ian H., Witten, Eibe Frank (2005), Data Mining: Practical Machine Learning Tool and Techniques, Morgan Kaufmann Publishers.

[5] Taddei A., Comstantino G., Silipo R. (1995), A system for the detection of ischemic eposodes in ambulatory ECG, Computer in Cardiology. IEEE.

[6] Lehtinen R., Sievänen H., Turjanmaa V., Niemelä K., Malmivuo J. (1997), Effect of ST-segment measurement point on performance of exercise ECG analysis, International Journal of Cardiology 61(3), 239-245.

[7] Viik J. (2003), Importance of Postexercise ECG, International Journal of Bioeletromagnetism, vol.5, no. 1.

[8] Drew, Kirchoff (1999), Multi-lead ST segment monitoring in patients with acute coronary syndromes: A consensus statement for health care professionals, American Journal of Critical, 8(6), 372-386.

[9] Agrawal R., Tomasz (1993), Mining association rules between sets of items in large database, the ACM SIGMOD Conference on Management of Data, 207-216, Washington D.C, USA, May.

[10] Agrawal R., Srikant R. (1994), Fast Algorithms for Mining Association Rules in Large Database, In Proc. of the 1994 International Conference on VLDB, 487-499.

[11] Hearst M., Dumais S., Platt E., Scholkopf B. (1998), Support vector machines, IEEE Intelligent System, 13-4, 18-28.

[12] Han J., Kanmer M. (2000), Data Mining : Concepts and Techniques, Morgan Kamfmann Publishers.

[13] Friedman N., Geiger D., Goldszmidt M. (1997), Bayesian Network Classifiers, Machine Learning, 29, 131-163.

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

[15] Kim H., Loh W. Y. (2001), Classification trees with unbiased multiway splits, JASA 96, 589-604.

[16] Quinlan J. R. (1993), C4.5: Programs for and Neural Networks, Machine Learning, Morgan Kaufman publishers.

[17] Fayyad U. M., Irani K. B. (1993), Multi-Interval discretization of continuous-valued attributes for classification learning, In Proc. of the International Joint Conf. on AI, 1022-1027.

[18] Liverpool unvi. computer science knowledge discovery, http://www.csc.liv.ac.uk/~frans/KDD/

[19] Liu B., Hsu W., Ma Y., (1998), Integrating classification and association rule mining, In Proc. of the 4th International Conference Knowledge Discovery and Data Mining, 4, 89-125.

[20] Li W., Han J., Pei J. (2001), CMAR: Accurate and Efficient Classification Based on Multiple Association Rules, In Proc. 2001 International Conference on Data Mining, 369-376.

원고접수 : 06/07/22
수정접수 : 07/06/01
게재확정 : 07/06/01