제26회 한국정보처리학회 추계학술발표대회 논문집 제13권 제2호 (2006. 11)

관상동맥질환 진단을 위한 데이터마이닝 기법

박홍규, 이헌규, 류근호
충북대학교 전기전자컴퓨터 공학부
e-mail:{khpark1980, hglee, khryu@somewhere.sck.ac.kr

Data Mining Approach for Diagnosing Cardiovascular Disease*1)

Hong Kyu Park, Heon Gyu Lee, Keun Ho Ryu
School of Electrical & Computer Engineering,
Chungbuk National University

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 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.

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 중 하나이다[1]. 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 적용하기위해 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. 원시 심전도 데이터는 특징 벡터 추출을 위해 ST-segment의 경사와 면적을 추출하여 질환진단을 위한 파라미터로 사용한다.
- 임상 데이터에 연관규칙을 적용하여 환자들의 임상적 속성들의 모든 연관규칙을 추출한다.
- ST-segment의 특징 벡터를 이용하여 CAD (Coronary Artery Disease) 또는 Normal people로 진단을 위한 분류 기법을 적용하여 평가한다. 적용된 분류 기법으로는 Java Weka[2]의 결정트리, 베이지안 분류 그리고 연관적 분류 알고리즘들을 적용해 그 결과를 분석한다.

이 논문의 구성은 다음과 같다. 먼저 2장에서는 심전도의 ST-segment를 이용한 심장 질환 패턴에 대해 기술하고 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.

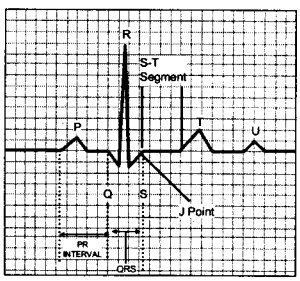
1) 이 논문은 2006년도 교육인적자원부 지방연구중심대학 육성사업의 지원에 의하여 연구되었음

2. 심전도의 ST-분절을 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 환경적 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. 국내에서의 심전도 시스템은 심장질환 환자의 심장상태를 감시하기 위한 홀터 시스템 그리고 환자 감시장치의 사용이 늘어나면서 그 중요성도 높아지고 있다. 그 외 12채널 진단 심전계, 스트레스 심전계 등의 심장관련 진단기기에 대한 연구가 활발히 진행되고 있다[1].

심장의 전기적 활성단계는 크게 심방 탈분극, 심실 탈분극, 심실 재분극 시기로 나뉘며, 이러한 각 단계는 다음 (그림 1)과 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 띄게 된다. (그림 1)은 심전도 데이터에서 ST-segment, R-R간격, QRS complex, J point 등을 표현한 것이다[3][4][5][6].



(그림 1) 심전도 파형의 구성요소

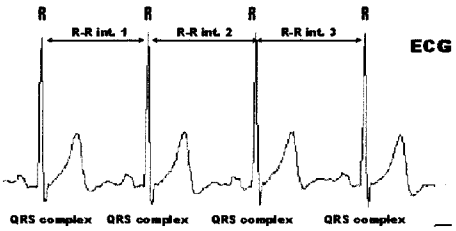
3. 임상정보 및 심전도 데이터의 전처리

이 절에서는 환자의 임상 데이터와 원시 심전도로부터 ST-segment 특징벡터들의 추출 등의 데이터 전처리 과정에 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 2)와 같이 추출하였다[1][2].

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 특징벡터로 사용하였는데 R-R간격이 600ms 보다 크면 ST60은 R+120ms로 허고 ST80은 R+140ms를 사용하며 만약 600ms 보다 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[14, 15].



(그림 2) R-Peak와 QRS Complex의 검출

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

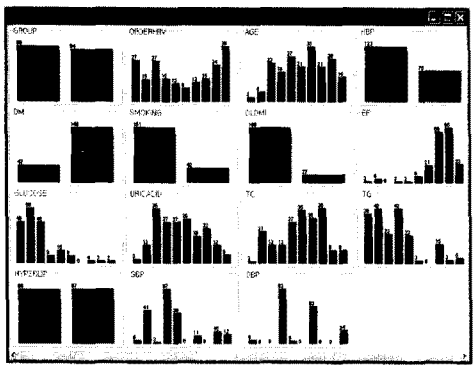
<표 1> 연관규칙 적용을 위한 임상 정보 리스트

환자의 임상 정보 리스트
Age, Hyper Blood Pressure, Diabetes Mellitus, Smoking, Old Myocardial Infarction, Ejection Fraction, Blood Glucose, Total Cholesterol, Triglyceride, Hyperlipidemia, Systolic Blood Pressure, Diastolic Blood Pressure

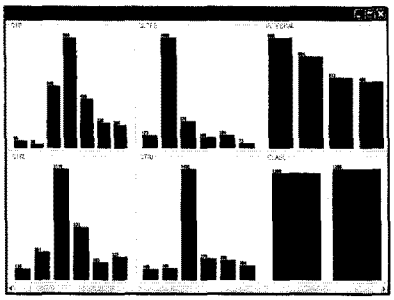
3.2 임상 데이터와 심전도 특징 벡터의 전처리

심장 질환 환자들의 임상적 속성들의 연관성 탐사

를 위한 연관규칙의 적용과 ST-segment 벡터로부터의 질환 지단을 위해서는 이산화와 정규화가 필요하다. 일반적으로 이산화를 하기 위한 알고리즘은 엔트로피 척도를 사용한다. 엔트로피 기반 척도의 결과는 이산화가 되는데 이러한 이산화를 통해 그 속성의 수치 개념 계층이 형성 된다[16]. (그림 3)은 Java Weka 프로그램의 임상정보 및 ST-segment에 대한 엔트로피 기반 이산화 결과이다.



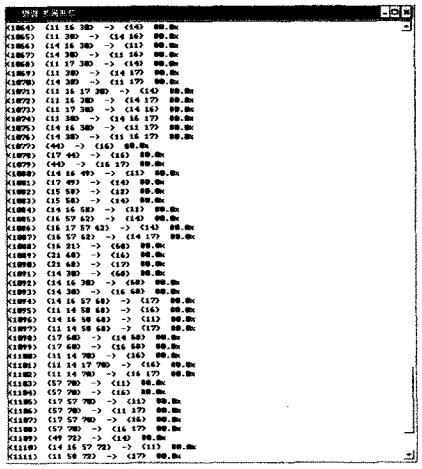
(그림 3.a) 임상 데이터에 대한 이산화 결과



(그림 3.b) ST-segment 이산화 결과

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

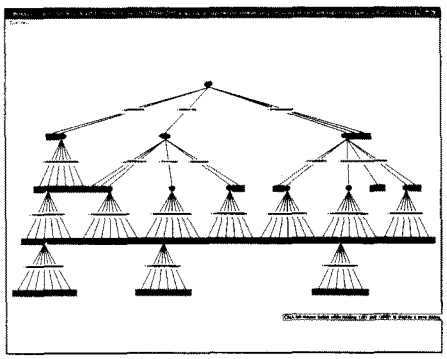
심장 질환 환자들의 임상적 속성들의 연관성 탐사를 위해 기존의 연관규칙 알고리즘 중 AprioriT[17] 알고리즘을 적용하여 주어진 임계값(지지도, 신뢰도)에 대한 연관규칙을 추출하였다. 휴리스틱 방식으로 실험에 대한 최적 파라미터를 추정하였으며, 그 파라미터 값으로 지지도는 10%, 신뢰도는 80%로 하였다. (그림 4)는 Apriori 알고리즘 수행 결과 후의 탐사된 연관 규칙이다. 예를 들어, (그림 5)의 연관규칙들 중에서 규칙, R: <16, 21> --> <68> 80.0%의 정규화된 값 16은 "속성 Age가 45~50이고 21(Glucose)가 81~85인 환자들 중에서 28인 DBP(심장 이완시 혈압)이 62~65일 가능성은 80%이다"란 의미이다.



(그림 4) Apriori 알고리즘 적용 결과

ST-segment를 이용한 심장 질환의 예측을 위해 적용된 분류 기법은 의사결정트리로 C4.5, 베이지안 분류기로는 나이브 베이지안, 베이지안 네트워크를 알고리즘을 적용하였다. 마지막으로, 연관적 분류 기법은 CBA[18], CPAR[19] 알고리즘 적용을 하였다.

(그림 5)는 실험에 포함된 분류 기법 중 C4.5를 적용했을 때의 생성된 트리이다. 나머지 분류 기법에 대해, Java Weka 프로그램과 LUCS-KDD[17] 프로그램을 이용하였다.



(그림 5) C4.5 알고리즘 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, Precision 및 F-Measure를 이용하였다. 각 분류 결과에 대한 성능 평가 비교는 표 2[2]에 요약하였다.

2) NB: Naive Bayesian, BNet: Bayesian Network

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

<표 2> 각 분류 기법에 대한 성능 평가

성능 평가

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
C4.5	0.677	0.106	0.87	0.677	0.761	CAD
	0.894	0.323	0.724	0.894	0.8	control
	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
NB	0.576	0.138	0.814	0.576	0.675	CAD
	0.862	0.424	0.659	0.862	0.747	control
	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
BayNet	0.889	0.128	0.88	0.889	0.884	CAD
	0.872	0.111	0.892	0.872	0.877	control
	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
CBA	0.758	0.16	0.833	0.758	0.794	CAD
	0.84	0.242	0.767	0.84	0.802	control
	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
CPAR	0.939	0.085	0.921	0.939	0.93	CAD
	0.915	0.061	0.935	0.915	0.925	control

참고문헌

[1] P.Conumel, "ECG: Past and Future", Annals NY Aca-demy of Sciences, vol.601, 1990.

[2] Java Weka project, Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/

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

[4] R. Lehtinen, H. Sievänen, V. Turjanmaa, K. Niemelä, J. Malmivuo, "Effect of ST-segment measurement point on performance of exercise ECG analysis", International Journal of Cardiology 61(3), p239-245, 1997.

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

[6] Drew, Kirchoff, "Multi-lead ST segment monitoring in patients with acute coronary syndromes: A consensus statement for health care professionals", American Journal of Critical Care, 8(6), p372-386, 1999.

[7] R. Agrawal, Tomasz Imielinski and Arun Swami, "Mining association rules between sets of items in large database," the SIGMOD Conf. on Manag. of Data, Washington D.C, USA, May. 1993.

[8] R. Agrawal and R. Srikant, "Fast Algorithms Mining Association Rules in Large Database," In Proc. of the 1994 International Conference on VLDB, 1994.

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

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

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

[15] 노기용, 김원식, 이헌규, 이상태, 류근호, "심전도 패턴 판별을 위한 빈발 패턴 베이지안 분류", 정보처리학회논문지 D, 제11-D권 제5호, 2004.

[16] U. M. Fayyad and K. B. Irani, "Multi-Interval discretization of continuous-valued attributes for classification learning", In Proc. of the Interna'l Joint Conf. on AI, 1022-1027, 1993.

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

[18] B. Liu, W. Hsu and Y. Ma, "Integrating classification and association rule mining", In Proc. of the 4th International Conference Knowledge Discovery and Data Mining, 1998.

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