심장 질환 진단을 위한 베이지안 분류 기법

손호선*, 이헌규*, 조경환*, 류근호*, 노기용**
*충북대학교 전자계산학과
**한국표준과학연구원
e-mail : shon0621@dblab.chungbuk.ac.kr

Bayesian Classification Method for Diagnosing Heart Disease

Ho Sun Shon*, Heon Gyu Lee*, Kyung Hwan Cho*, Keun Ho Ryu*, Ki yong Noh**
*Dept. of Computer Science ChungBuk University
**Korea Research Institute of Standards and Science

요 약

심전도는 각종심장질환 들을 예측하는데 널리 사용되고 있다. 이러한 심전도에서 ST-분절은 허혈성 심장 질환, 확장성 심근성, 비후성 심근증 등을 예측하는데 이용되고 있다. 이 논문에서는 환자들의 임상 정보와 심전도로부터 심장 질환 예측을 위한 중요 파라미터인 ST-분절을 추출하였다. 그리고 이러한 추출된 데이터 분석을 위해서 데이터마이닝 기법을 적용한다. 데이터마이닝의 분류 알고리즘인 베이지안 네트워크를 적용 심장 질환을 효율적으로 분류하기 위한 방법을 제시 하였다.

1. 서론

심전도(Electrocardiogram: ECG)를 이용한 심장 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. 이 논문에서는 임상 정보와 심전도 데이터를 이용해 ST-분절의 특징 파라미터를 추출하고 심전도 데이터의 허혈/정상을 분류하기 위한 분류 알고리즘 기법으로는 나이브 베이지안과 베이지안 네트워크를 이용 하였다.

따라서 이 논문에서는 임상 정보와 심전도 데이터로부터 심장 질환 예측을 위한 ST-분절을 추출하고, 베이지안 분류기에 적용 가능하도록 엔트로피 기반의 이산화와 정규화 등의 전처리 작업을 수행하였다. 나이브 베이지안과 베이지안 네트워크을 이용하여 정상인과 CAD 를 분류하고, 이력 데이터를 통해 심장질환을 예측을 할 수 있다.

이 논문은 2006 년도 교육인적자원부 지방연구중심 대학육성 사업의 지원에 의하여 연구 되었음

2. 관련 연구

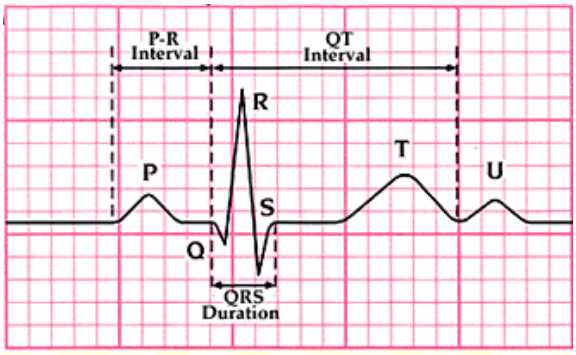
2.1 ECG 의 ST-분절을 이용한 심장 질환 패턴

돌연사를 일으키는 대표적 심장질환은 허혈성 심장 질환, 확장성 심근성, 비후성 심근증이 대부분을 차지하며, 특히 허혈성 심장 질환이 돌연사의 80%를 차지하므로 이 질환의 예방 및 조기 진단이 중요하다. 허혈성 심장 질환의 증세로서 협심증과 심근경색증이 있는데, 심전도의 ST 분절이 elevation 또는 depression 되는 episode 를 띄게 된다. (그림 1)은 심전도 데이터에서 ST-분절, RR 간격, QRS complex, J point 등을 표현한 것이다. [1][4]

2.2 분류 기법

분류란 어떤 새로운 사물이나 대상의 특징을 파악하여 미리 정의되어 있는 클래스에 따라 어느 한 범주에 할당하는 것이다. 이러한 목적을 위하여 대부분의 분류 대상들은 기존의 데이터베이스를 토대로 새

로운 중요한 데이터 클래스를 설명하는 모형을 생성하거나 미래 데이터의 경향을 예측할 때 사용되는 분



(그림 1) 심전도 파형의 구성 요소

석 기법이다. 분류 기법에 대한 연구는 통계, 신경망, 의사결정트리, 사례기반추출 등의 분야에서 연구 되었으며, 의료 진단 예측 수행, 마케팅분야 등 여러 분야에 응용되고 있다.

베이지안 네트워크(Bayesian network)는 의학 도메인에서 질병의 진단이나 예측 문제를 해결하기 위해 많이 사용되고 있으며, 또한 좋은 성능을 보여 왔다.

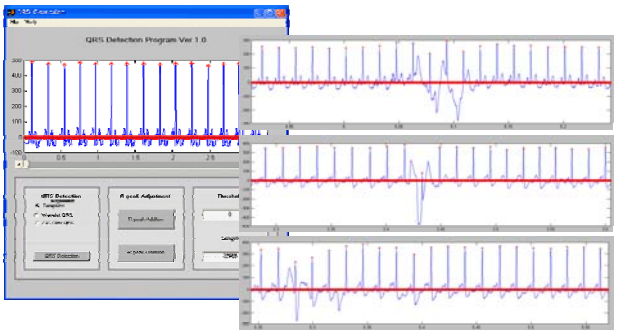
장점으로는 분류나 예측 문제를 위해 많이 사용되고 있는 신경망의 블랙박스 분류기와 비교해 도메인 지식을 적용하기 쉬우며, 결과의 분석이 가능하며, 단점으로는 입력으로 연속 값이 아닌 범위가 정해진 상태 값을 사용함으로써 정확도 면에서 문제가 생길 수 있으며, 노드 수가 많아 지면 실험 시간이 오래 걸린다. 하지만 의학 지식을 적용하여 분석 등이 가능한 의학 도메인에서 도메인 지식 가능성이나 원인 분석이 가능하다는 특성은 큰 장점이 된다. 따라서 이 논문에서는 위와 같은 장점을 지닌 나이브 베이지안(Naïve Bayesian)과 베이지안 네트워크(Bayesian network) 분류기를 사용하였다.

3. ECG, 임상데이터의 전처리

이 절에서는 심근 허혈 질환의 분류를 위한 전처리 단계로서 심전도 데이터를 이용하여 ST-분절 특성 파라미터를 추출하고, 분석을 위해 전처리 작업을 한다.

3.1 ST-분절 특징 파라미터 추출

R-Peak 와 QRS Complex 검출 프로그램을 Tompkins 알고리즘을 이용하여 MATLAB 으로 개발되었으며, 다음 (그림 2) 에서 그 예를 볼 수 있다.[2][3][4]



(그림 2) R-Peak 와 QRS Complex 검출 프로그램

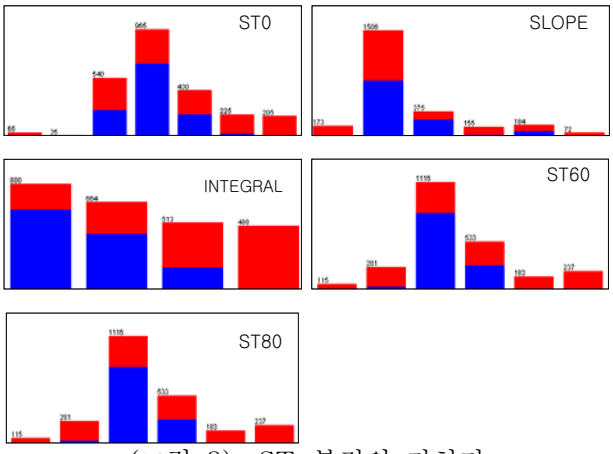
ST 분절 구간 설정은 RR 간격이 600ms 보다 클 경우는 R-peak 에서 60ms 동안으로 하였다. ST 분절의 특징을 추출하기 위하여 RR 간격이 600ms 보다 클 경우 또는 그렇지 않을 경우에 대하여, ST0(ST 분절이 시작되는 지점), ST80(R+140 또는 R+120 ms 지점에서의 진폭), ST-분절의 slope, ST 분절의 면적(ST 분절의 구간에서 ST-분절 isoelectric level 로 둘러 싸여진 면적)을 ST-분절의 특징을 나타내는 파라미터의 후보로 설정하였다.

<표 1> 특징 선택의 결과

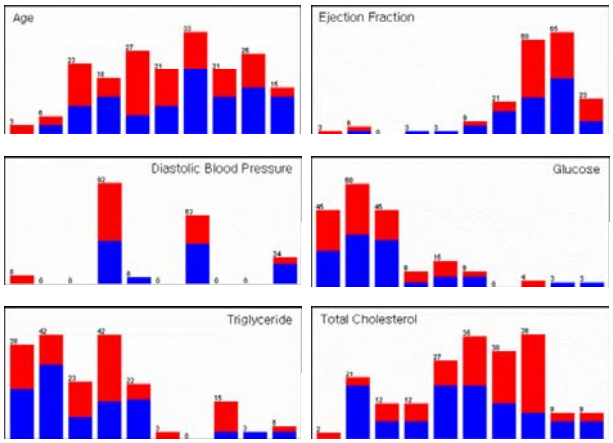
Data	특징 파라미터
ST-Segment	STO, SLOPE, INTEGER, ST60, SR80
Clinical Info	Age, Hyper Blood Pressure, Diabetes Mellitus, Smoking, Old Myocardial Infarction, Ejection Fraction, Blood Glucose, Total Cholesterol, Triglyceride, Hyperlipidemia, Systolic Blood Pressure, Diastolic Blood Pressure

3.2 심전도 데이터의 전처리

심전도 데이터들은 진폭의 수와 주파수를 이용한 생체 전위 신호 이므로 모두 연속적인 데이터들로 이루어져 있다. 그러므로 다음 (그림 3)과 (그림 4)는 심전도 데이터에 대한 전처리 과정을 통해 나온 결과들이다.



(그림 3) ST-분절의 전처리



(그림 4) 연속 변수에 대한 이산화

4. 심전도 데이터의 분류

전처리 단계를 거친 ST-분절 특징 파라미터를 해당 클래스(정상/허혈)로 분류하기 위한 데이터 마이닝의 분류 기법을 적용한다. 실험을 위해 사용한 분류기는 나이브 베이지안과 베이지안 분류기를 이용하여 심전도 데이터를 분류 하였다.

나이브 베이지안 분류기는 클래스의 조건 독립성이라는 가정하에 따른 부정확성과 가용 확률 데이터의 부족으로 인하여 결과가 항상 정확하지는 않다. 실제로 대부분의 데이터에 대한 종속성이 존재하므로 베이지안 분류기를 이용하여 비교 분류 하였다.

4.1 나이브 베이지안(Naïve Bayesian) 분류

나이브 베이지안(Naïve Bayesian) 분류기는 베이지안 분류 모형에서 각 입력 변수들 간에 독립을 가정하여 사후확률을 계산한다. 사후확률은 다음 식(1)과 같은 베이스 정리에 의해 구해진다.[7]

$$P(C_i) = \frac{P(X \mid C_i)P(C_i)}{P(X)} \cong P(X \mid C_i)P(C) \qquad (1)$$

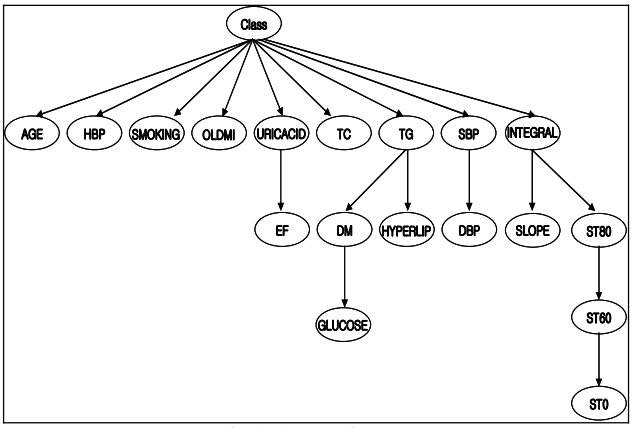
위의 식(1) 에서 X 는 입력벡터, $(x_1, ..., x_p)$ 를 나타내고, C_i는 목표변수의 k 개의 클래스, $(C_i, \cdots, C_k)$ 중에서 i 번째 클래스를 나타낸다. 위의 식에서 입력벡터의 차원 p 가 커지면 속성들간의 연관성이 존재할 수 있기 때문에 $P(X \mid C_i)$ 는 다음과 같이 간단히 계산된다.

$$P(X \mid C_i) = \prod_{j=1}^{p} P(x_i \mid C_i) \qquad (2)$$

위의 식(2)의 확률 구조를 가우시안 분포(Gaussian Distribution)로 하였는데 이는 대부분 데이터들의 평균이 표본수의 증가에 따라 가우시안 분포로 접근(Convergence)하는 중심극한정리(Central Limit Theorem)에 의하기 때문이다. 하지만 가우시안 분포를 사용하지 않고 다른 특정 분포를 사용할 수 있으며 이 경우 사후확률 계산에 있어서 더 고급 베이지안 계산 기법(Advanced Bayesian Computing)을 요구하게 된다.

4.2 베이지안 네트워크(Bayesian network) 분류

나이브 베이지안(Naïve Bayesian)의 문제점 즉 주어진 클래스의 한 속성값이 다른 속성값과 상호 독립임을 가정하지 않고 베이지안 네트워크는 속성의 부분집합이 가지는 종속 관계를 표현할 수 있는 그래프 모델을 이용한다. 베이지안 네크워크(Bayesian network)는 최근 복잡한 도메인에서 불확실성을 해결하기 위한 강력한 데이터 마이닝 기법으로 사용되고 있다. 즉 결합확률분포(Joint Probability Distribution)를 이용하는 모델로 도메인 지식을 쉽게 반영할 수 있는 장점을 가지며, 방향성 비순환 그래프(Directed Acyclic Graph)의 형태를 취한다. 이 그래프에서 노드는 변수를, 노드 간의 연결은 확률적인 종속 관계를 나타낸다. 실제 심전도 데이터에 대한 베이지안 네트워크 구축 모형은 다음 (그림 5)와 같다. 예를들어, 다음 네트워크 모델에서 변수 INTEGER 는 SLOPE 와 ST80 에 영향을 주며 ST60 은 다시 ST80 에 종속되어 있으며, 다시 ST0 은 ST60 에 영향을 받아 조건부 확률값이 계산된다.[6][8][9]



(그림 5) 베이지안 네트워크 모델

이 논문에서 사용되는 나이브 베이지안(Naïve Bayesian)과 베이지안 네트워크(Bayesian network) 분류기는 베이즈 정리를 이용하여 개체 X 를 다음의 조건을 만족하는 C_i에 할당한다.

$$P(X|C_i)P(C_i) \rangle P(X|C_j)P(C_j) \quad for 1 \le j \le k, i \ne j \qquad (3)$$

즉, X 는 k 개의 클래스 중에서 가장 큰 사후 확률값을 갖게 되는 집단으로 할당된다.

5. 실험 평가

이 논문에서 제안한 알고리즘들에 대한 시스템 구현 환경은 다음과 같다.

ST-분절 파라미터 추출은 MATLAB 상에서 구현 하였다. 실험에 사용된 데이터는 심장질환자(CAD) 290 명과 정상인 370 명을 이용하였다.

실제 실험 평가는 평균절대오차(MAE), 평균자승오차제곱근(RMSE)과 F-Measure 를 이용하여 분석 하였으며, 다음 <표 2>와 <표 3>은 나이브 베이지안과 베이지안 네트워크를 이용한 분류 결과를 보여 주고 있다.

<표 2> 나이브 베이지안의 분류 결과

Mean Absolute Error(MAE) = 0.2525					
Root Mean Squared Error(RMSE) = 0.4374					
TP- Rate	FP- Rate	Precision	Recall	F-Measure	Class
0.905	0.382	0.695	0.905	0.786	Nor
0.618	0.095	0.871	0.618	0.723	CAD

<표 3> 베이지안 네트워크의 분류 결과

Mean Absolute Error(MAE) = 0.2388					
Root Mean Squared Error(RMSE) = 0.4201					
TP- Rate	FP- Rate	Precision	Recall	F-Measure	Class
0.898	0.333	0.722	0.898	0.801	Nor
0.667	0.102	0.872	0.667	0.756	CAD

위의 <표 2>와 <표 3>은 나이브 베이지안(Naïve Bayesian)과 베이지안 네트워크(Bayesian network) 알고리즘을 이용한 분류 예측값의 정확성 측면에서 성능을 평가하기 위해 평균절대오차(MAE)를 식(4)을 이용하여 구한다.

$$|E| = \frac{\sum_{i=0}^{N} |\varepsilon_i|}{N} \quad (4)$$

나이브 베이지안 분류기에서는 평균절대오차(MAE)가 0.2525 이며, 베이지안 네트워크 분류기에서는 0.2388 의 결과를 얻었으므로 이는 베이지안 네트워크 분류기가 더 정확도가 높다는 것을 알 수 있다. 또한 평균자승오차제곱근(RMSE)은 오차 제곱의 평균에 제곱근을 취한 것으로 표준편차의 정의와 동일하다. 따라서 평균제곱오차제곱근(RMSE)에서도 베이지안 네트워크 분류기가 더 오차가 작음을 알 수 있다.

또한 위의 표에서 정확도(Precision)와 재현율(Recall)에 동등한 중요도를 부여하여 하나의 평가 방법으로 사용하는 F-Measure 는 식(5)를 이용하여 구한다.

$$F = \frac{2 * precision * recall}{precision + recall} \quad (5)$$

즉 F-Measure 는 이용 성능을 평가하기 위해 정확도(Precision)와 재현율(Recall)을 결합한 조화평균이 사용됨을 알 수 있다. 나이브 베이지안(Naïve Bayesian)과 네트워크 베이지안(Bayesian network) 분류기 둘 다에서 정상인이 허혈성 질환의 환자보다 높다는 것을 증명한다.

6. 결론

심근 허혈 질환의 분류를 위해 심전도 데이터와 임상 정보를 이용하여 ST-분절 특성 파라미터를 추출하고, 전처리와 정규화를 수행 하였으며, 데이터들을 이용한 분류 방법으로는 데이터마이닝 알고리즘인 나이브베이지안(Naïve Bayesian)과 베이지안 네트워크(Bayesian network) 분류기를 이용하여 실험하였다. 실험 결과 나이브 베이지안(Naïve Bayesian) 분류기 보다 네트워크 베이지안(Bayesian network) 분류기가 여러 면에서 더 정확함을 알 수 있었다. 즉 평균 절대 오차(MAE) 값과 평균자승오차제곱근(RMSE)에서 베이지안 네트워크 분류기의 값이 더 작은 오차를 가지므로 정확도가 높다. 또한 정확도(Precision)와 재현율(Recall)을 이용한 F-Measure 는 정상인이 허혈성 질환의 환자보다 더 높음을 알 수 있다.

참고문헌

[1] P.Conumel, “ECG: Past and Future”, Annals NY Academy of Sciences, vol.601, 1990.

[2] S.Guzzetti, R.Magatelli, E.Borroni, “Heart rate variability in chronic heart failure”, American Neuroscience, Basic and Clmical, 90, p102-105, 2001.

[3] C.D.Kuo, G.Y.Chen, “Comparison of Three recumbent position on vagal and sympathetic modulation using spectral heart rate variability in patients with coronary artery disease”, American Journal of Cardiology, 81, p392-396, 1998.

[4] F.Jager, G.B.Moody, A.Taddei, R.G.Mark, “performance measures for algorithms to detect transient ischemic ST segment changes”, Computer in Cardiology. IEEE,p 396-372, 1991.

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

[6] I.Kononenko, “Semi-naive Bayesian Classifer”, Proceedings of sixth European Working Session on Learning, Springer-Verlag, p206-219, 1991.

[7] N.Friedman, D.Geiger, and M. Goldszmidt, “Bayesian network classifiers”, Machine learning 29, p131-163, 1997.

[8] W. Lam, W. Iba, K. Thompson, “Learning Bayesian belief networks: an approach based on the MDL principle”, Computational Intelligence 10, p269-293, 1994.

[9] P.Langley, W.Iba, K.Thompson, “An analysis of Bayesian Classifers”, In Proceedings of AAAI-92, p223-228, 1992.