베이지안 네트워크에 기반한 심전도 데이터의 정확도 향상에 관한연구

이현주*, 신동일**, 신동규***
세종대학교 컴퓨터공학과
e-mail: nedkelly@gce.sejong.ac.kr, {dshin, shindk}@sejong.ac.kr

Research on improving correctness of cardiac disorder data based on Bayesian Network

Hyun-Ju Lee* , Dong-Il Shin**, Dong-Kyoo Shin***
*Dept of Computer Engineering, Sejong University

요 약

심전도 데이터는 일반적으로 분류기를 사용한 실험이 많으며, QRS-Complex와 R-R interval 간격을 추출하여 실험한다. 본 연구에서는 R-R interval을 추출하였다. 그리고 R-R interval 데이터와 HRV 데이터를 구성하였고, 베이지안 네트워크 분류기를 사용하여 정확도를 도출하였다. 심장관련 데이터는 심전도 뿐 아니라 심장병 데이터도 있는데 심전도 데이터와 같이 분류실험을 시행하여 정확도를 도출하였다. 그리고 베이지안 네트워크분류기의 정확도를 분석하기 위해 타 논문의 실험결과와 비교하였다. 타 논문과 본 연구의 결과를 비교해보니 베이지안 네트워크가 타 결과에 비해서 정확도 도출이 우수하였다.

1. 서론

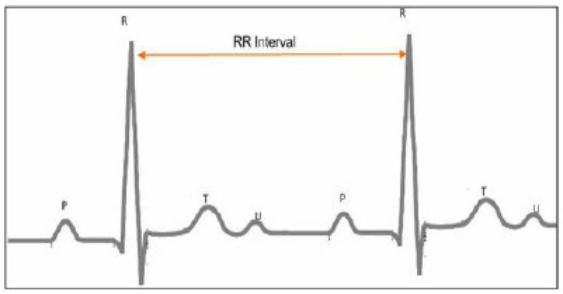
심전도는 심장이 박동을 할때 나타나는 전기적인 신호로 심장상태와 질환을 알아볼 수 있는 중요한 자료로 쓰인다[1]. 심전도 신호로 탐지할 수 있는 질환 중에는 부정맥이 있으며, 이는 MIT-BIH Arrhythmia Database[2]를 포함한 공개된 데이터가 존재하며, 이를 실험한 사례가 많다. 심전도는 P, Q, R, S, T 다섯 개의 파형으로 구성되어 있으며, QRS-Complex, R-R interval 두 간격으로 추출하한 실험이 일반적이다. 본 연구에서는 R-R interval을 추출하여 베이지안 네트워크 분류기로 실험을 하였다. 실험결과로는 정확도를 도출하였다. 그리고 분류기 성능의 비교를 위해 동일한 데이터로 실험한 타 논문의 결과와 비교하였다. 심장과 관련된 데이터 중에는 심전도 외에도 심장병 데이터가 존재하는데 UCI Machine Learning Repository[3]는 여러 종류의 데이터가 오픈된 사이트로 네 개의 심장병 데이터가 있다. 본 연구에서는 네 가지 중 Statlog(Heart) 데이터를 선택하여 베이지안 분류기로 실험하였고, 심전도와 마찬가지로 동등한 데이터를 사용한 타 논문의 결과와 정확도를 비교하였다.

2. 이론적 배경

2.1 데이터 수집

MIT-BIH Arrhythmia Database는 1975년 이래로 Boston's Israel Hospital과 MIT의 지원을 받아서 부정맥의 분석과 관련된 주제들을 연구한 데이터이다[2]. 총 48개의 데이터로 구성되어 있으며, 데이터의 기록은 시간과 채널당 360 샘플에서 디지털화한 기록이다. Statlog (Heart)는 UCI Machine Learning Repository(Center for Machine Learning and Intelligent Systems)[3]에서 공개한 데이터로 총 14개의 속성으로 구성되어 있다.

2.2 심전도 데이터 특징추출

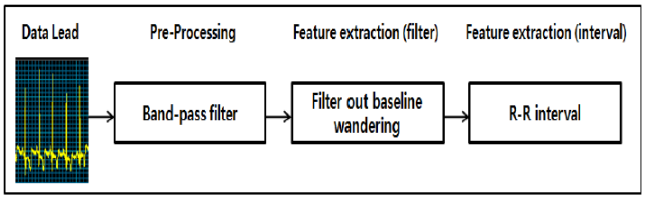


(그림 1) R-R interval

R-R interval은 R파의 한 피크(Peak)에서 그 다음에 측정되는 피크까지의 시간을 의미하며, 각각의 R-R interval은 한 번의 cardiac cycle을 나타낸다. R-R interval은 연속된 시간의 형태로 반복하여 지속적으로 발생하는데, 심전도 신호에서 QRS 검출기

본 연구는 서울시 전략산업 지원사업(SS110008)의 지원에 의해 수행되었음

를 적용할 때 R-R interval의 Sequence가 변형된다[4]. 그림 1은 R-R interval의 형태를 나타내었다[5].



(그림 2) R-R interval 특징 추출

그림 2는 R-R interval의 특징추출과정을 나타낸 것이며, 전처리과정에서는 필터를 사용하여 데이터의 기저선 잡음을 제거한다. 본 연구에서는 NI Labview(National Instrument Labview)[5]에서 제공되는 NI Biomedical Stqartup Kit3.0으로 추출작업을 하였으므로 Kit에서 제공하는 대역통과필터(Band-pass Filter)를 사용하였다. 대역통과필터는 single-filter안에서 low-pass와 high-pass를 조합하여 잡음을 걸러내도록 디자인된 필터이다.

본 연구에서는 R-R interval을 추출하여 R-R interval 특징데이터와 HRV(Heart Rate Variability) 데이터를 구성하여 실험하였다. 심장병 데이터(Statlog(Heart))는 14개의 속성을 가진 데이터로 속성은 “Age”,“Sex”, “CP”, “Trestbps”, “Chol”, ″Fbs″, “Restecg”, “Thalach”,“Exang”, “Oldpeak”, “Slope”, “Ca”, “Thal”, “Num”으로 구성되어있다.

3. 베이지안 네트워크(Bayesian Network)

베이지안 네트워크(Beyesian Network)는 그래픽 기반 모델(Graphical Model)로서 기본적으로 여러 가지 변수들의 결합 확률 분포를 표현하는 수단으로 사용된다. 특히 변수들 간의 관계가 지역적인 경우, 즉 각 변수가 적은 수의 다른 변수들로부터 영향을 받는 경우에 특히 효과적이다. 또한 고차원의 데이터로부터 데이터에 맞는 구조를 모델링 하고,간단한 구조로 표현할 수 있는 장점이 있다. 따라서 연구하고자 하는 해당 데이터에 대한 사전 지식이 부족하더라도, 결합 확률 분포를 활용하여 각 변수들 간의 관계를 밝히는데 있어 매우 유용하게 쓰인다. 베이지안 네트워크는 의사결정 시스템이나, 의료 진단 시스템, 기상예측, 항공우주, 생물학 등의 여러 분야에 걸쳐 각 분야에 속한 특정 도메인에서의 데이터분석 및 예측을 위해 사용된다[6].

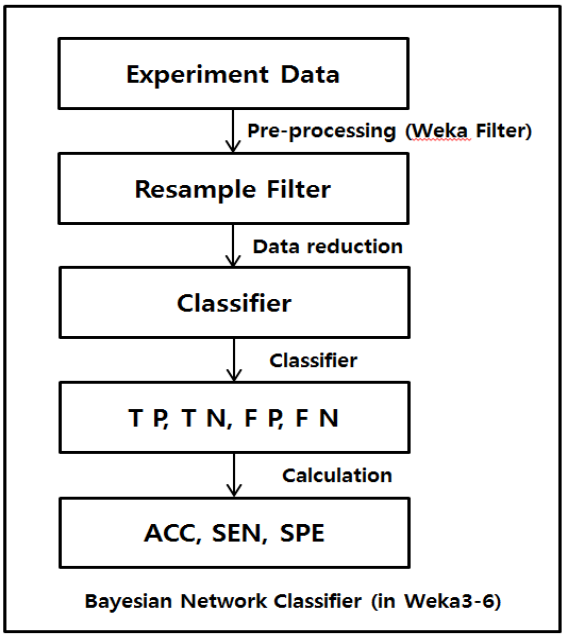
변수들의 유한 집합 X = {X_1,...,X_N}이 있다고 가정할 때, 각 변수가 정수일 경우 X_i는 Val(X_i)의 정수 값 영역에 해당하는 X_i를 갖고, 실수일 경우 그에 해당하는 실수값을 가지게 된다. 베이지안 네트워크에서 중요한 첫 번째 요소는 조건부 독립성(Conditional Independence)으로, 이는 변수 X, Y의 집합이 있을 때,

$$P(X|Y, Z) = P(X|Z) \quad (1)$$

를 만족하면, (X⊥Y|Z)로 표현할 수 있다. 이것은 Z라는 변수가 주어졌을 때 X와 Y사이에는 직접적인 연관성이 없음을 의미한다[6].

4. 실험방법 및 결과

분류기 실험은 Weka의 베이지안 네트워크 분류기를 사용하여 시행하였다. Weka의 버전은 3-6버전을 사용하였다. 실험과정은 그림 3과 같다.



(그림 3) Weka를 활용한 실험과정

추출한 심전도 데이터(R-R interval, HRV)와 Statlog(Heart)데이터를 Weka에서 오픈하면 전처리과정(Pre-processing)을 거치게 된다. 전처리과정에서는 Re-sample filter를 사용하여 데이터를 축소시켜주는 작업을 하였다. 전처리과정을 거친 후 분류기실험을 실행하는데 분류기는 베이지안 네트워크를 선택하였다. 분류실험을 거친 데이터는 결과가 도출되는데 도출된 결과에서 TP(True Positive), TN(True Negative), FP(False Positive), FN(False Negative)를 측정할 수 있다. 측정한 TP, TN, FP, FN으로 ACC(Accuracy: 정확도), SEN(Sensitivity: 민감도), SPE(Specificity: 특이도)를 계산할 수 있다.

4.1 수식

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (3)$$

$$Specificity = \frac{TN}{TN + FP} \quad (4)$$

4.2 실험결과

실험결과는 R-R interval을 실험한 결과와 HRV를 실험한 결과 그리고 Statlog(heart)를 실험한 결과로 총 세 개의 표로 정리하였고, 각 데이터마다 타 논문의 결과를 비교분석하였다.

<표 1> K&L과 BN의 실험결과

Data_no	SEN(%)		SPE(%)		ACC(%)	
	K&L	BN	K&L	BN	K&L	BN
201	96	95.8	39	81.7	65	92.31
202	80	95.6	94	75.4	88	95.1
203	81	92	21	56.7	63	90.6
210	96	96.1	0	89.6	94	96
217	72	90.9	94	86.5	90	90.3
219	96	92.7	64	43.4	91	92
221	92	93.5	50	31.5	65	90.1
Average	87.5	93.8	51.7	66.4	79.4	92.3

표 1의 결과는 R-R interval 데이터를 실험한 결과이다. BN은 베이지안 네트워크(Bayesian Network)를 의미하며, K&L[7]는 K.Tateno의 실험으로 R-R interval을 추출하였고, 본 연구와 동일한 데이터를 사용하였다. 실험결과를 살펴보면, ACC(정확도)부분에 있어서는 베이지안 네트워크가 90%이상의 정확도를 도출하였다. 따라서 베이지안 네트워크의 결과가 K&L보다 정확도 면에서 우수함을 알 수 있다.

<표 2> BN과 NEWFM의 실험결과

Sensitivity		Specificity		Accuracy	
BN	NEW	BN	NEW	BN	NEW
96.29	88.75	73.3	82.28	92.7	86.31

표 2는 HRV데이터를 실험한 결과이다 NEWFM은 장형종[8]의 실험결과로 본 연구과 동일한 데이터를 사용하여 HRV를 도출하였다. NEWFM은 가중 퍼지 소속함수 기반 신경망(Neural Network with Weighted Fuzzy Membership Functions)을 의미한다. 정확도(Accuracy)를 비교해보면 베이지안 네트워크가 NEWFM보다 6.39% 높은 정확도를 도출하였음을 알 수 있다.

<표 3> NB와 BN의 실험결과

Sensitivity		Specificity		Accuracy	
NB	BN	NB	BN	NB	BN
80.29	92.66	86.24	66.82	83.58	89.8

표 3은 심장병 데이터(Statlog(Heart))를 실험한 결과이며, NB는 Naive Bayes을 의미하며, Chau[9]의 실험결과이다. Chau는 Bagging을 활용하여 Naive Bayes를 실험하였다. 본 연구의 베이지안 네트워크와 정확도를 비교해보면 베이지안 네트워크가 6.22% 높은 결과를 도출하였음을 알 수 있다. 따라서 전체 데이터 실험결과를 살펴보면 베이지안 네트워크가 정확도 면에서는 타 실험의 결과보다 우수한 결과를 보였음을 알 수 있다.

5. 결론 및 토의

본 연구는 심전도데이터의 하나인 MIT-BIH Arrhythmia Database와 심장병데이터인 Statlog(heart)를 선택하여 실험하였다. 실험은 분류기를 사용한 정확도향상에 목표를 두었고, 베이지안 네트워크 분류기를 사용하였다. 또한 베이지안 네트워크 분류기의 정확도를 비교하기 위해 동일한 데이터를 활용한 타 논문의 결과와 정확도를 비교분석해 보았다. 실험결과는 베이지안 네트워크가 R-R interval의 경우 90%이상의 결과를 도출하여 K&L의 결과보다 우수하였고, HRV는 타 논문보다 6.39% 높은 결과를 그리고 Statlog(heart)는 타 논문결과보다 6.22% 높은 결과를 도출하였다. 향후에는 심전도 및 심장관련 데이터 외에 다른 종류의 데이터를 사용한 정확도 도출실험과 베이지안 네트워크외에 정확도를 높게 도출할 수 있는 분류기에 관한 연구가 필요하다.

참고문헌

[1] K. S. Park, B. H. Cho, D. H. Lee, S. H. Song, J. S. Lee, Y. J. Chee, I. Y. Kim, and S. I. Kim, “Hierarchical Classification of ECG Beat Using Higher Order Statistics and HermiteModel”, J Kor Soc Med Informatics, Vol.15, pp. 117-131, 2009.

[2] Physiobank(MIT-BIH Arrhythmia Database) : http://physionet.mit.edu/physiobank/database/mitdb/

[3] UCI Repository : http://archive.ics.uci.edu/ml/

[4] G. D. Clifford, F. Azuaje and P. E. McSharry,“ Advanced Methods and Tools for ECG Data Analysis”, pp.101-102, Artech-Hous e, Boston & London, 2006.

[5] NI Biomedical Startup Kit 3.0 : http://decibel.ni.com/content/docs/DOC-12646

[6] 황성철, "효율적인 베이지안 네트워크 학습 방법에 관한 연구", 연세대학교 대학원, 2007.

[7] K. Tateno and L. Glass, “A Method for Detection of Atrial Fibrillation Using RR Intervals”, Computers in Cardiology(IEEE), Vol.27, pp.391-394, 2000.

[8] H. J. Jang and J. S. Lim, "Detection of Arrhythmia Using Heart Rate Variability and A Fuzzy Neural Network", Korean Society for Internet Information (KSIL), Vol.10, pp.107-115, 2009.

[9] T. M. Chau, "Research on effective analysis on physiological signals", Sejong University, 2010.