The Korean Journal of Applied Statistics (2019) DOI: https://doi.org/10.5351/KJAS.2019.32.5.671
32(5), 671–681

Prediction of arrhythmia using multivariate time series data

Minhai Lee[a] · Hohsuk Noh[b,1]

[a]Department of Statistics, Sookmyung Women's University;
[b]Department of Statistics, The Research Institute of Natural Sciences,
Sookmyung Women's University

(Received July 25, 2019; Revised August 20, 2019; Accepted August 26, 2019)

Abstract

Studies on predicting arrhythmia using machine learning have been actively conducted with increasing number of arrhythmia patients. Existing studies have predicted arrhythmia based on multivariate data of feature variables extracted from RR interval data at a specific time point. In this study, we consider that the pattern of the heart state changes with time can be important information for the arrhythmia prediction. Therefore, we investigate the usefulness of predicting the arrhythmia with multivariate time series data obtained by extracting and accumulating the multivariate vectors of the feature variables at various time points. When considering 1-nearest neighbor classification method and its ensemble for comparison, it is confirmed that the multivariate time series data based method can have better classification performance than the multivariate data based method if we select an appropriate time series distance function.

Keywords: arrhythmia prediction, multivariate time series, 1-nearest neighbor, time series distance function, ventricular tachycardia

1. 서론

부정맥은 심장의 규칙적인 수축이 이루어지지 않고, 심장 박동이 불규칙하게 비정상적으로 빨라지거나 늦어지는 현상을 말한다. 일반적으로 심방성 빈맥이나 다른 빈맥보다도 더 위험하다고 알려진 심실빈맥(ventricular tachycardia; VT)은 갑작스런 심장마비를 일으켜 목숨을 앗아갈 수 있는 악성 부정맥이다. 이러한 심실빈맥을 예측하여 실신이나 돌연사를 예방하기 위한 노력의 일환으로 머신러닝 기법을 활용하여 그 발생을 예측하는 것에 대한 연구가 활발히 진행되어왔다 (Lee 등, 2016). 기존 연구들은 정상적인 심장리듬과 심실빈맥 발생 직전의 심장리듬으로부터 각각 심장 상태를 잘 나타내는 특징변수들의 다변량 벡터를 추출하고 이에 다변량 분류(classification)기법을 적용함으로써 심실빈맥 발생을 예측하였다. 본 연구에서는 심장 상태가 시간에 따라 변해가는 패턴도 부정맥 예측에 중요한 정보가 될 수

This research was supported by the 'Basic science research program' through the National Research Foundation of Korea funded by the Ministry of Education (NRF-2017R1D1A1A09000804).
[1]Corresponding author: Department of Statistics, Sookmyung Women's University, Cheongpa-ro 47-gil 100, Yongsan-gu, Seoul 04310, Korea. E-mail: word5810@gmail.com

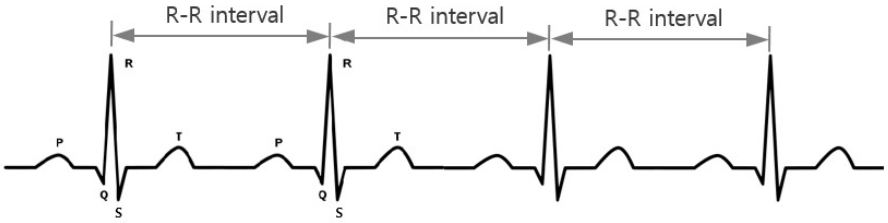

Figure 2.1. Example of PQRST cycles and RR intervals on electrocardiogram.

있다고 생각하여 일정한 시간 간격을 두고 심장 상태를 나타내는 특징변수의 다변량 벡터를 추출하여 쌓음으로써 얻어지는 다변량 시계열 데이터로 부정맥을 예측하는 것의 유용성에 대해 살펴보고자 한다.

본 논문의 구성은 다음과 같다. 2절에서는 분석에 사용되는 심박동 데이터에 대해 먼저 설명하고 그 데이터로부터 심장 상태를 파악하기 위해 추출하는 특징변수에 대해 설명하였다. 3절에서는 2절에서 설명한 데이터로부터 심실빈맥 발생예측을 위해 다변량 데이터와 다변량 시계열 데이터를 어떻게 생성하는지 설명하고 4절에서는 각각의 데이터에 기반한 분류 방법의 성능을 비교하였다. 5절에서는 결론과 추후 연구방향을 제시하였다.

2. 심박동 데이터 및 특징변수 설명

심박동 데이터의 대표적인 유형인 심전도(electrocardiogram; ECG)는 심장활동에 의해 국소적으로 발생하는 전기 변화를 기록한 그림이다. 심전도상에서 심방의 흥분에 의해서 생긴 파형을 P파라 하고, P파에 이어지는 최초의 하향파를 Q파, 최초의 상향파를 R파, R에 이어지는 하향파를 S파라 한다. 박동 간격(inter-beat interval; IBI)은 연속적인 R파의 시간 차를 말하며, RR 간격(RR interval)이라고도 한다. Figure 2.1은 심전도상에서 PQRST 주기와 RR 간격을 나타낸 예시이다.

본 논문에서 사용하는 심박동 데이터는 Au-Yeung 등 (2018)에서 사용된 데이터이다. 데이터는 6,660개의 정상 심장리듬으로부터 얻어진 RR 간격 자료와 230개의 심실빈맥이 나타나기 직전까지의 심장리듬으로부터 얻어진 RR 간격 자료로 구성되어 있다. 심실빈맥이 나타나기 직전의 RR 간격 자료의 끝은 심실빈맥이 발생한 시점이다. 각각의 RR 간격 자료는 25분에서 30분 분량의 심장 박동 자료이며 최대 2,048개의 RR 간격으로 이루어져 있다. 본 논문에서 사용한 RR 간격 자료는 Au-Yeung 등 (2018)이 분석에 사용한 RR 간격 자료와 기본적으로 동일하나 premature ventricular contraction과 compensatory pause에 해당하는 박동부분이 제거되지 않은 점에서만 다르다.

본 연구에서는 기존의 연구들과 동일하게 심실빈맥 발생을 예측하는데 있어 RR 간격 자체를 이용하지 않고 그로부터 심장 상태의 특징을 나타내 줄 수 있는 특징변수(feature)를 추출하여 사용하였다. 본 논문에서 RR 간격 자료에서 추출한 52개의 특징변수는 기존의 심장관련 머신러닝 연구들에서 자주 사용하는 변수들이며 그에 대한 자세한 설명은 아래에 주어져 있다. 처음 41개의 특징변수는 RR 간격 자체(RR(i), $i = 1, \ldots, N$, N은 주어진 심장리듬의 RR 간격의 수)를 이용하여 생성한 변수이며 나머지 11개의 특징변수는 차분한 RR 간격(δRR(i) = RR(i) − RR($i-1$), $i = 2, \ldots, N$)을 이용하여 생성한 변수이다.

1) feature 1–11: RR 간격에 대해 기본적인 통계값을 계산하는 변수들로 구성되어 있다. 계산하는 통계값은 RR 간격의 변동계수, 평균값, 중앙값, 최솟값, 최댓값, 왜도(skewness), 첨도(kurtosis), 최댓값과 최솟값 차이, 분산, RR 간격의 normalized RMSSD(root mean square of successive differ-

ences) [RR 간격의 평균으로 표준화], RR 간격의 중앙값 절대 편차(median absolute deviation)이다.

2) feature 12–23: RR 간격 자료에 대한 다양한 entropy를 계산하였다. feature 12는 Shannon entropy를 계산한 것이고 feature 13은 Lake 등 (2002)에 제시되어 있는 approximate entropy인 $\mathrm{ApEn}(m, r, N)$를 $m = 1$, $r = 0.2$일 때 계산한 것이다 (N은 RR 간격의 수). feature 14–18은 Lake 등 (2002)에 제시되어 있는 sample entropy인 $\mathrm{SampEn}(m, r, N)$를 $r = 0.2$이고 $m = 0, 1, 2, 3, 4$에 대해 각각 계산한 것이다. feature 19–23은 feature 14–18을 5차원 벡터로 인식하였을 때 값이 ∞인 원소의 개수, 최솟값, 최댓값, 5차원 벡터상의 최솟값의 위치, 최댓값의 위치를 계산한 것이다.

3) feature 24, 25: RR 간격에 대해 Hjorth mobility와 Hjorth complexity를 계산하였다.

4) feature 26–28: RR 간격에 대해 kernel smoothing 방법을 이용하여 확률밀도함수를 계산하고 RR 간격 자료의 범위상의 적절한 grid위의 점들에 대한 확률밀도함수값를 구해 그 첨도와 왜도, peak의 수를 계산하였다.

5) feature 29–30: 4)에서 구한 RR 간격의 확률밀도함수의 peak가 여러 개인 경우 RR 간격의 확률밀도함수의 peak가 여러 개인 경우 가장 높은 peak를 가지는 위치와 다른 peak를 가지는 위치들사이 거리의 최댓값과 최솟값를 계산하였다. peak가 한 개인 경우는 두 feature 모두 0으로 계산하였다.

6) feature 31–36: Lomb-Scale 방법을 이용하여 RR 간격 자료에 대한 power spectral density (PSD)를 추정하고 주파수 구간별 power 및 power ratio를 구하였다. 본 논문에서 고려한 주파수 구간은 VLF = [0.0033 0.04], LF = [0.04 0.15], HF = [0.15 0.4] (단위 Hz)이며 PSD 추정치를 이용하여 pVLF, pLF, pHF, pVLF/pLF, pVLF/pHF, pLF/pHF를 계산하였다.

7) feature 37–39: Park 등 (2009)와 같이 RR 간격의 Poincare plot으로부터 세 개의 feature(RR 간격의 mean stepping increment, RR 간격의 Poincare plot에서 주대각선 주위로의 퍼진 정도와 Poincare plot상의 점들이 형성하는 cluster의 개수)를 계산하였다. 세 번째 feature인 cluster 개수를 계산할 때는 Ester 등 (1996)의 DBSCAN algorithm을 사용하였다.

8) feature 40, 41: Shieh 등 (2010)에서와 같이 short term detrended fluctuation analysis를 실시하여 feature를 계산하되 절편과 기울기 모두를 feature로 사용하였다.

9) feature 42–46: Sarkar 등 (2008)에서와 같이 차분한 RR 간격의 Lorenz plot를 그리고 그로부터 심장 상태의 특징을 나타내는 OriginCount, IrrEv, PACEv, DensityEv, AniEv를 계산하였다. 각 변수의 정의와 계산 방법은 Sarkar 등 (2008)에 나와 있다.

10) feature 47, 48: 차분한 RR 간격으로부터 kernel smoothing 방법을 이용하여 확률밀도함수를 구하고 차분한 RR 간격 범위상의 적절한 grid위의 점에서 확률밀도함수값을 구하여 그 첨도와 왜도를 계산하였다.

11) feature 49, 50: 차분한 RR 간격의 절댓값이 0.2를 넘는 비율, 0.5를 넘는 비율를 계산하였다.

12) feature 51, 52: 차분한 RR 간격의 표준편차, 변동계수를 계산하였다.

3. 심실빈맥 예측을 위한 데이터 생성

본 논문에서는 기존의 다변량 특징벡터의 모임인 데이터 행렬에 기반한 심실빈맥 예측과 다변량 시계열 데이터의 모임인 데이터 텐서에 기반한 심실빈맥 예측의 성능을 비교하고자 한다. 이를 위해 본 절에서는 RR 간격 자료로부터 다변량 데이터와 다변량 시계열 데이터를 각각 어떻게 생성하였는지를 설명하고자 한다.

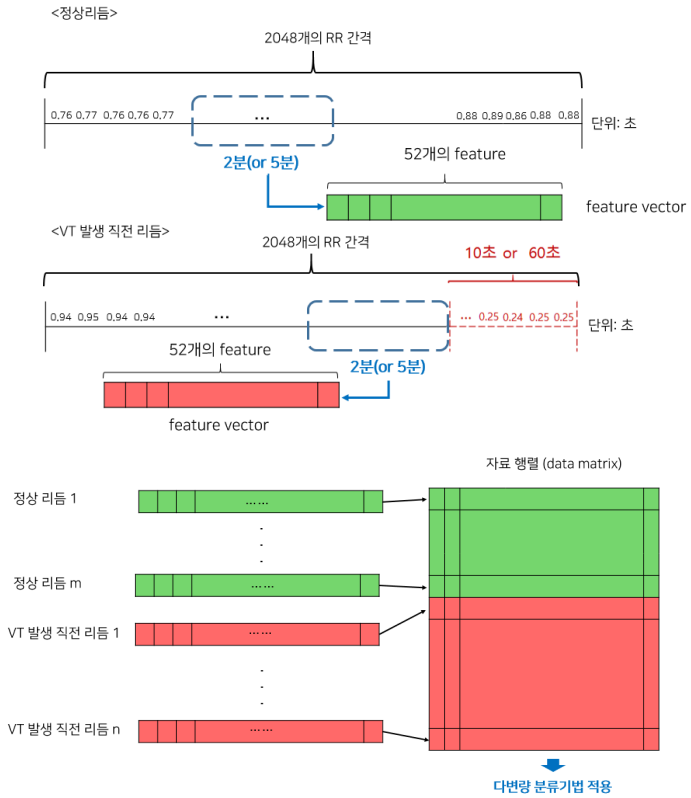

Figure 3.1. Data matrix construction for arrhythmia prediction by accumulating the feature vectors.

3.1. 다변량 데이터 생성

보통 부정맥 예측을 위해서는 비교적 짧은 시간 분량의 심박동 데이터에 변화하는 심장특성이 잘 반영된다는 인식하에 2분이나 5분 분량의 심박동 자료를 이용하는 경우가 많이 있다. 이를 위해서 분석에 사용하는 25분에서 30분 분량의 정상리듬의 데이터에서 해당 분량의 심박동 자료를 선택하는 과정이 필요하다. 정상리듬의 RR 간격 데이터는 데이터의 어느 부분이나 정상리듬의 특성을 가진다고 할 수 있다. 따라서 각 정상리듬의 RR 간격 자료 데이터로부터 임의의 2분 또는 5분 분량의 RR 간격 데이터를 선택하여 2절에서 설명한 특징변수를 추출하여 특징변수 벡터를 생성하였다. 심실빈맥 발생 직전 리듬의 RR 간격 데이터는 본 연구에서 10초 또는 60초 전에 부정맥을 예측하려는 예측 목적을 고려하여 먼저 심실빈맥 발생 시점 앞의 10초 또는 60초 분량의 데이터를 제거하였다. 그리고 그 앞의 2분 또는 5분 분량의 RR 간격 데이터로부터 특징변수 벡터를 생성하였다. 이렇게 얻어진 정상리듬과 심실빈백 발생 직전 리듬의 특징변수 벡터들을 모으면 심실빈맥 예측을 위한 2차원 데이터 행렬을 얻을 수 있다. 이 데이터 행렬에 기반하여 다변량 분류 기법들을 훈련(training) 시킴으로써 심실빈맥을 예측할 수 있다. 심실빈맥 예측을 위한 다변량 데이터 생성과정을 요약하면 Figure 3.1과 같다.

3.2. 다변량 시계열 데이터 생성

심실빈맥 예측을 위한 다변량 시계열자료 행렬 생성은 다음과 같은 방식으로 이루어졌다. 정상리듬의

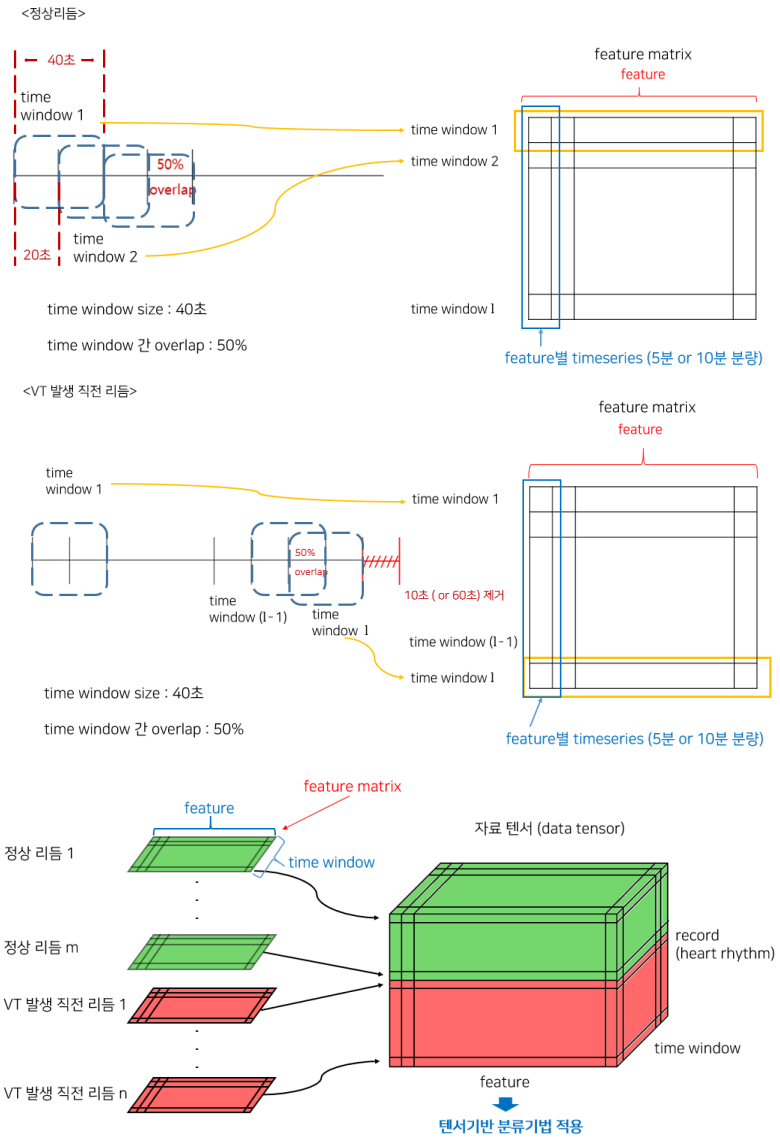

Figure 3.2. Data tensor construction for arrhythmia prediction by accumulating the feature matrices

RR 간격 자료의 시작점에서 40초 길이의 time window로부터 2절에서 설명된 52개의 특징변수들로 구성된 벡터를 추출하였다. 그리고 time window를 직전의 time window와 50%씩 겹치면서 이동하여 특징변수 벡터를 계속 추출하여 쌓아나감으로써 특징변수 행렬(feature matrix)을 만들었다. 이때 특징변수별 시계열 데이터는 총 사용된 RR 간격 자료의 분량이 5분이나 10분 분량이 되도록 하였다. 심실빈맥 발생 직전 리듬의 RR 간격 자료에서는 다변량 자료 생성에서와 같이 심실빈맥 예측이라는 목적을 생각하여 데이터 끝인 심실빈맥 발생 시점으로부터 10초나 60초까지 데이터를 제거하고 정상리듬의 RR 간격 자료와 똑같은 과정을 거쳐서 특징변수 행렬을 만들었다. 40초의 time window들이 50%씩

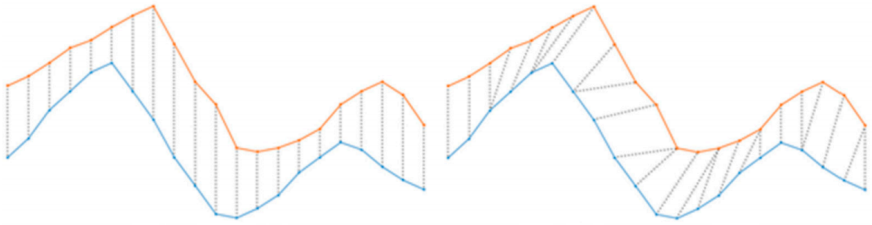

Figure 4.1. Comparison between Euclidean distance and dynamic time warping distance (Csillik *et al.*, 2019).

겹치게 이동하며 특징변수 벡터를 계속 추출하여 쌓아나가되 전체 사용된 RR 간격 자료의 분량이 5분 또는 10분이 되도록 특징변수 행렬을 만들었다. 이 때 특징변수 벡터를 추출하기 위한 마지막 time window의 오른쪽 끝이 심실빈맥 발생 10초 또는 60초전이 되도록 하였다. 이렇게 정상리듬과 심실빈백 발생 직전 리듬의 특징변수 행렬들을 모으면 심실빈맥 예측을 위한 3차원 데이터 텐서(tensor)를 얻을 수 있다. 심실빈맥 예측을 위한 다변량 시계열 데이터 생성과정을 요약하면 Figure 3.2와 같다.

4. 분류 성능 비교

이 절에서는 크게 두 가지 방법으로 심실빈맥 예측에서 다변량 벡터 기반 데이터 행렬을 사용하는 것과 다변량 시계열 기반의 데이터 텐서를 사용하는 것을 비교하였다. 첫 번째는 생성된 데이터 행렬과 데이터 텐서로부터 특징변수별로 데이터를 추출하여 분류 방법을 훈련시켜 분류 성능을 비교하는 것이다. 두 번째는 특징변수별 데이터를 기반으로 훈련된 분류 방법을 앙상블한 것들의 성능을 비교함으로써 다변량 벡터 기반 분류 방법과 다변량 시계열 기반 분류 방법의 성능을 비교하고자 하였다.

4.1. 특징변수별 분류 방법을 기준으로 한 성능 비교

3절에서 얻어낸 두 가지 형태의 자료에서 특징변수별 데이터를 얻어낸 후 분류 방법을 훈련시켜 분류 성능을 비교하는 분석을 하였다. 각 특징변수별 데이터를 기반으로 정상리듬 그룹과 심실빈맥 발생 이전 리듬 그룹을 분류하기 위해 1-nearest neighbor (1-NN) 분류 기법을 사용하였다. 1-NN은 새로운 데이터의 클래스를 기존의 데이터 중 속성이 가장 유사한 데이터를 찾아 그 데이터가 속한 클래스로 예측하는 지도 학습기법이다.

특징변수별 값 데이터간 거리를 계산할 때는 절대값 거리를 사용하였고 시계열 데이터 간 거리를 계산할 때는 R package TSClust에서 제공되는 거리 함수 중 dynamic time warping distance (DTWARP), Euclidean distance (EUCL), correlation-based distance (COR) 총 3가지 거리함수를 고려하였다. 각각의 거리 함수에 대한 설명은 Montero와 Vilar (2014)에서 찾아볼 수 있다. DTWARP 방법은 하나의 시계열과 비교 대상의 시계열이 최대로 일치할 때까지 시계열의 시간 축을 부분적으로 왜곡하거나 변형함으로써 시계열 간의 시간 차이와 연관된 특성이 시계열 간 거리에 최대한 반영되지 않도록 하여 시계열 간 거리를 구하는 방법이다. 이에 비해 EUCL 방법은 각각의 시계열을 다차원 벡터로 보고 단순 벡터간의 거리로 시계열 간 거리를 측정하는 방법이고 COR 방법은 두 벡터간의 상관계수를 구해 거리로 사용하는 방법이다. Figure 4.1은 두 시계열 간의 거리를 EUCL 방법과 DTWARP 방법으로 구하는 것의 차이를 시각화하여 보여주고 있다.

분류 성능을 평가하기 위해서 leave-one-out cross-validation (LOOCV)에 근거한 정분류율을 계산하였다. 분류 데이터를 구성할 때 정상리듬이 6,660개이고 심실빈맥 발생 직전 리듬이 230개로 데이터 클

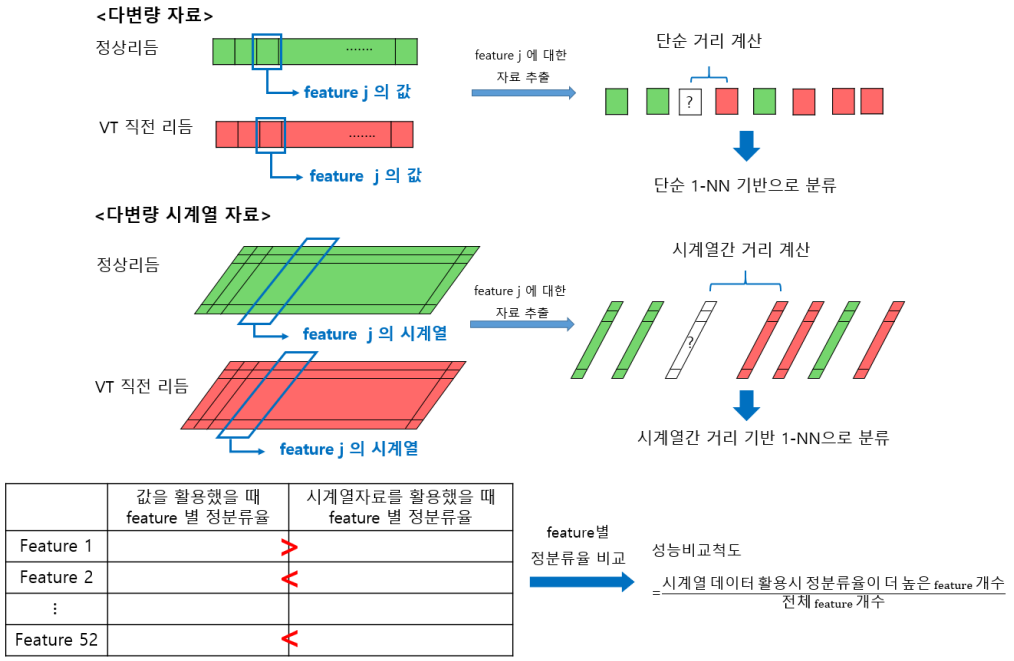

Figure 4.2. Illustration about how to compare arrhythmia prediction performance between the ordinary 1-NN method and the time series 1-NN method.

래스의 불균형이 있어 정상리듬의 수를 심실빈맥 발생 직전 리듬의 수와 동일하게 만들어주는 언더샘플링(undersampling)을 실시하였다. 최종 분류 성능의 평가에 언더샘플링의 sampling variability가 영향을 미치지 않도록 하기 위해 언더샘플링을 100번 반복 후 얻어지는 정분류율의 평균을 최종 분류 결과로 사용하였다. 분석 결과를 토대로 앞에서 설명한 정분류율을 각 특징변수별로 두 가지 방법(값 기반과 시계열 기반)에 대해 구한 후 시계열 기반 1-NN의 정분류율이 값 기반 1-NN의 정분류율보다 높게 나오는 특징변수 개수의 비율을 계산하였다. Figure 4.2는 비율 계산 과정을 그림으로 나타낸 것이고 Table 4.1은 다양한 데이터 생성 조건에서 시계열 기반 1-NN의 정분류율이 값 기반 1-NN의 정분류율보다 높게 나오는 특징변수 개수의 비율을 사용된 시계열 간 거리 함수를 기준으로 정리한 표이다. 표에서 time1은 심실빈맥 예측성능을 확인하기 위해 심실빈맥 발생 직전 리듬에서 발생 시점에서부터 일정 시간 분량만큼 제거하는 시간을 말하며 분석에서는 10초와 60초를 고려하였다. time2는 다변량 데이터를 생성할 때 사용하는 RR 간격 자료의 분량을 말하며 time3는 다변량 시계열 자료를 생성할 때 이용되는 RR 간격 자료의 총 시간분량을 말한다.

Table 4.1를 보면 시계열 간 거리 함수 3가지 중에서 DTWARP로 계산하였을 때 대부분의 데이터 생성 조건에서 시계열 데이터 기반 1-NN 분류 성능 향상이 가장 두드러졌다. 다음으로 시계열 간 거리 함수를 단순거리함수인 EUCL를 사용했을 때에도 어느 정도의 분류 성능 향상은 있었다. 그러나 시계열 간 거리 함수가 COR인 경우에는 오히려 분류 성능 저하를 가져오는 것을 확인 할 수 있다. 단순 Euclidean 거리(EUCL)를 활용할 때보다 DTWARP 거리를 사용할 때 분류 성능의 향상이 더 큰 것과 correlation 기반 거리(COR)를 사용할 때 오히려 분류 성능의 저하를 가져오는 것을 통해 시계열의 특징을 잘 반영하는 적절한 시계열 거리 함수의 선택이 시계열 정보를 활용한 분류 성능의 향상에 중요한 요소임을 확인하였다.

Table 4.1. Comparison of arrhythmia prediction performance between the ordinary 1-NN and the time series 1-NN

Distance function	Time1	Time2	Time3	Ratio
DTWARP	10s	2min	5min	0.731
			10min	0.769
		5min	5min	0.731
			10min	0.808
	60s	2min	5min	0.769
			10min	0.808
		5min	5min	0.731
			10min	0.788
EUCL	10s	2min	5min	0.673
			10min	0.596
		5min	5min	0.750
			10min	0.673
	60s	2min	5min	0.769
			10min	0.750
		5min	5min	0.673
			10min	0.654
COR	10s	2min	5min	0.216
			10min	0.173
		5min	5min	0.288
			10min	0.173
	60s	2min	5min	0.173
			10min	0.231
		5min	5min	0.173
			10min	0.250

1-NN = 1-nearest neighbor; DTWARP = dynamic time warping distance; EUCL = Euclidean distance; COR = correlation-based distance.

4.2. 특징변수별 데이터 기반 분류 방법을 앙상블한 것의 성능 비교

먼저 4.1절에서와 같이 정상리듬의 수를 심실빈맥 발생 직전 리듬의 수만큼 언더샘플링하여 얻어진 데이터를 훈련 데이터 70%, 평가 데이터 30%로 분리하였다. 그 후 훈련 데이터에 기반하여 4.1절에서와 동일한 과정으로 52개의 특징변수 각각의 데이터에 기반한 1-NN 분류 방법을 값 기반과 시계열 기반의 두 가지 형태로 구축하였다. 이 때 특징변수별 시계열 기반 분류 방법을 구축하기 위해서 사용하는 시계열 간 거리함수는 DTWARP만 고려하였다. j번째 특징변수 데이터를 이용하는 1-NN 분류 방법을 $f_j(y_j)$라고 정의했을 때($f_j(y_j) = 0$이면 정상리듬으로 분류, $f_j(y_j) = 1$이면 심실빈맥 발생 직전 리듬으로 분류, y_j는 j번째 특징변수 값 또는 시계열 데이터) 특징변수별 1-NN을 앙상블한 분류 방법으로 다음을 고려하였다.

$$f(\mathbf{y}) = \underset{c\in\{0,1\}}{\operatorname{argmax}} \sum_{j=1}^{52} \alpha_j I(f_j(y_j) = c) \quad \left(\alpha_j \geq 0,\ \sum_{j=1}^{52} \alpha_j = 1,\ \mathbf{y} = (y_1, \ldots, y_{52}) \right), \tag{4.1}$$

여기서 α_j는 훈련 데이터에 기반하여 5-fold cross-validation으로 정분류율이 최적화되는 값으로 결정하였다. 최적화된 α_j에 근거한 앙상블 learner로 평가 데이터를 분류한 후 정분류율을 계산하였다. 최종 성능 비교를 위해서는 언더샘플링을 100번 반복했을 때 다변량 시계열 기반 앙상블 learner의 정분류

Table 4.2. Comparison of arrhythmia prediction performance between the multivariate data based method and the multivariate time series based method

Time1	Time2	Time3	Ratio
10s	2min	5min	0.69
		10min	0.81
	5min	5min	0.75
		10min	0.69
60s	2min	5min	0.70
		10min	0.78
	5min	5min	0.64
		10min	0.73

율이 다변량 기반 앙상블 learner의 정분류율과 같거나 큰 언더샘플링 횟수의 비율을 고려하였다.

Table 4.2은 다양한 데이터 생성 조건에서 위에서 제시된 두 방법의 비교를 위한 세가지 기준값을 정리한 표이다. Table 4.2에서 time1, time2, time3의 정의는 Table 4.1에서와 동일하다. 다양한 데이터 생성 조건하에서 비교하여 보았을 때 시계열 기반 앙상블 learner가 다변량 기반 앙상블 learner의 정분류율을 앞선 언더샘플링의 횟수 비율도 64%에서 81%가 나와 그 성능이 더 우수함을 볼 수 있었다.

5. 결론 및 추후 연구방향

최근에 부정맥 환자가 증가하면서 부정맥을 예측하는 연구가 활발하게 진행되고 있다. 기존의 많은 연구들은 특정한 시점의 RR 간격 데이터에서 추출한 특징변수 다변량 데이터에 기반하여 부정맥을 예측하였다. 본 연구에서는 심장 상태가 시간에 따라 변해가는 패턴도 부정맥 예측에 중요한 정보가 될 수 있다고 생각하여 다변량 시계열을 기반으로 생성된 데이터로 심실빈맥을 예측하는 분류 learner를 생성하고 그 분류 성능을 다변량 데이터 기반 learner와 비교하였다. 그 결과 적절한 시계열 거리 함수를 사용하면 시계열 정보로 인한 분류 성능 향상을 기대할 수 있음을 알 수 있었다.

본 연구에서 1-NN 분류 방법과 그것을 앙상블한 분류 방법을 중심으로 다변량 시계열을 기반으로 심실빈맥을 예측하는 것과 다변량 데이터 기반으로 심실빈맥을 예측하는 것을 비교하였다. 추후에 보다 다양한 분류 기법을 적용하여 두 데이터 생성방법을 비교하는 것이 필요하다고 생각된다. 또한 다변량 시계열자료를 활용하는 경우 분류를 위한 데이터가 3차원 텐서형태가 되므로 다양한 텐서기반 분류 learner를 다변량 시계열자료에 적합하는 것도 좋은 연구 주제라고 생각된다.

References

Au-Yeung, W. T. M., Reinhall, P. G., Bardy, G. H., and Brunton, S. L. (2018). Development and validation of warning system of ventricular tachyarrhythmia in patients with heart failure with heart rate variability data, *PLoS ONE*, **13**, e0207215.

Csillik, O., Belgiu, M., Asner, G. P., and Kelly, M. (2019). Object-based time-constrained dynamic time warping classification of crops using sentinel-2, *Remote Sensing*, **11**, 1257.

Ester, M., Kriegel, H. P., Sander, J., and Xu, X. (1996) A density-based algorithm for discovering clusters. In Simoudis, E., Han, J., and Fayyad, U. M. (Eds), *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, (KDD-96), 226–231.

Lake, D. E., Richman, J. S., Griffin, M. P., and Moorman, J. R. (2002). Sample entropy analysis of neonatal heart rate variability, *American Journal of Physiology*, **283**, R789–R797.

Lee, H., Shin, S. Y., Seo, M., Nam, G. B., and Joo, S. (2016). Prediction of ventricular tachycardia one hour before occurrence using artificial neural networks, *Scientific Reports*, **6**, 32390.

Montero, P. and Vilar, J. A. (2014). TSclust: an R package for time series clustering, *Journal of Statistical Software*, **62**, 1–43.

Park, J., Lee, S., and Jeon, M. (2009). Atrial fibrillation detection by heart rate variability in Poincare plot, *BioMedical Engineering Online*, **8**, 1–12.

Sarkar, S., David Ritscher, D., and Mehra, R. (2008). A Detector for a Chronic Implantable Atrial Tachyarrhythmia Monitor, *IEEE Transactions On Biomedical Engineering*, **55**, 1219–1224.

Shieh, J. S., Yeh, R. G., Chen, G. Y., and Kuo, C. D. (2010). Parameter investigation of detrended fluctuation analysis for short-term human heart rate variability, *Journal of Medical and Biological Engineering*, **30**, 277–282.

다변량 시계열 자료를 이용한 부정맥 예측

이민혜[a] · 노호석[b,1]

[a]숙명여자대학교 통계학과; [b]숙명여자대학교 통계학과, 자연과학연구소

(2019년 7월 25일 접수, 2019년 8월 20일 수정, 2019년 8월 26일 채택)

요 약

최근에 부정맥 환자가 증가하면서 머신러닝을 이용한 부정맥을 예측하는 연구가 활발하게 진행되고 있다. 기존의 많은 연구들은 특정한 시점의 RR 간격 데이터에서 추출한 특징변수 다변량 데이터에 기반하여 부정맥을 예측하였다. 본 연구에서는 심장 상태가 시간에 따라 변해가는 패턴도 부정맥 예측에 중요한 정보가 될 수 있다고 생각하여 일정한 시간 간격을 두고 특징변수의 다변량 벡터를 추출하여 쌓음으써 얻어지는 다변량 시계열 데이터로 부정맥을 예측하는 것의 유용성에 대해 살펴보았다. 1-Nearest Neighbor 방법과 그것을 앙상블(ensemble)한 learner를 중심으로 비교했을 경우 시계열의 특징을 고려한 적절한 시계열 거리함수를 선택하여 시계열 정보를 활용한 다변량 시계열 데이터 기반 방법의 분류 성능이 더 좋게 나오는 것을 확인하였다.

주요용어: 부정맥예측, 다변량 시계열, 최근접이웃방법, 시계열 간 거리함수, 심실빈맥

이 논문은 2017년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2017R1D1A1A09000804).

[1]교신저자: (04310) 서울특별시 용산구 청파로47길 100, 숙명여자대학교 통계학과.
E-mail: word5810@gmail.com