

심전도 신호의 양자화 방법에 대한 심혈관 질환 분류시스템의 성능에 관한 연구

이희창, 윤태영, 여채운, 오현영, 지예빈, 심성우, 강대성
인제대학교 헬스케어 IT 학과

gmlckd1254@oasis.inje.ac.kr, yty20173270@oasis.inje.ac.kr, ducodbs0516@oasis.inje.ac.kr,
hy1013@oasis.inje.ac.kr, librenoblejyb@oasis.inje.ac.kr, mancity12@oasis.inje.ac.kr,
mpkds@inje.ac.kr

A Study on the Performance of the Cardiovascular Disease Classification System for the Quantization of ECG Signals

Heechang Lee, Tae Young Yoon, Chaeyun Yeo, HyeonYeong Oh, Yebin Ji, Seongwoo Sim, Daesung Kang
Dept. of Healthcare IT, Inje Univ.

요 약

본 논문은 1 차원 신호의 형태학적 정보를 가지는 GLCM(gray level co-occurrence matrix), GLRLM(gray level run-length matrix) 행렬에서 특징값을 추출할 때 선행되는 양자화의 방법이 기계학습 기반의 심혈관 질환 분류 시스템에 어떠한 영향을 주는지에 관한 연구이다. 양자화 방법으로는 균일 양자화, 로이드-맥스 양자화, 히스토그램 평활화 양자화, 사분위수 양자화 방법 등을 적용하였고, 기계학습 방법은 결정 트리, 로지스틱 회귀, 랜덤 포레스트를 활용하였다. 전반적으로 로지스틱 회귀 방법이 다른 기계학습 알고리즘 보다 성능이 좋았고, 히스토그램 평활화 양자화 방법과 결합하였을 때 가장 좋은 성능을 보이는 것을 확인하였다.

I. 서 론

기계학습을 통하여 심질환 관련 심전도 신호 분류에 사용되는 방법은 크게 두 가지가 있다. 심전도 신호에서 특징값을 직접 추출해서 기계학습 알고리즘에 적용하는 방법과 딥러닝을 활용하여 특징값을 자동으로 추출하는 End-to-End 방식으로 심질환을 분류하는 방법이다.[1] 특징값을 직접 추출할 때 자주 사용되는 방법으로는 심전도 신호의 피크 값 간의 거리, 푸리에나 웨이블릿 변환에서 얻은 계수 값 등이 있다.[2] 본 연구에서는 최근 연구되는 새로운 방법으로, 심전도 신호의 형태학적 정보를 가지는 GLCM, GLRLM 행렬을 생성하여 특징값을 추출하였다. 일반적으로 GLCM 과 GLRLM 행렬을 생성할 때는 양자화가 선행되어야 한다. 이미지와 같은 2D 데이터에서 양자화의 영향에 관한 연구가 진행되었지만, 아직까지 심전도 신호와 같은 1D 신호에서는 관련 연구가 거의 없다.[3] 이에 관련하여 본 논문은 1D 신호에서 GLCM 과 GLRLM 행렬 생성시 선행되는 양자화 방법이 심질환 분류에 미치는 영향을 다룬다.

II. 본론

본 연구에서 사용한 심전도 신호는 최근 Scientific Data 저널에 공개된 오픈데이터로 10,646 명의 심전도를 12 리드로 측정된 데이터셋이다.[4] 버터워스 저역 통과

필터, LOESS 평활기, NLM(non local means) 등과 같은 전처리를 하였고, 전처리 중에 손상된 58 명의 신호를 제외한 10,588 명의 심전도 신호를 실험에 사용하였다. 심전도 데이터는 11 개의 카테고리로 구성되어 있는데, 데이터 불균형을 고려하여 11 개의 심질환 카테고리를 4 개의 카테고리로 병합하여 사용하였다.

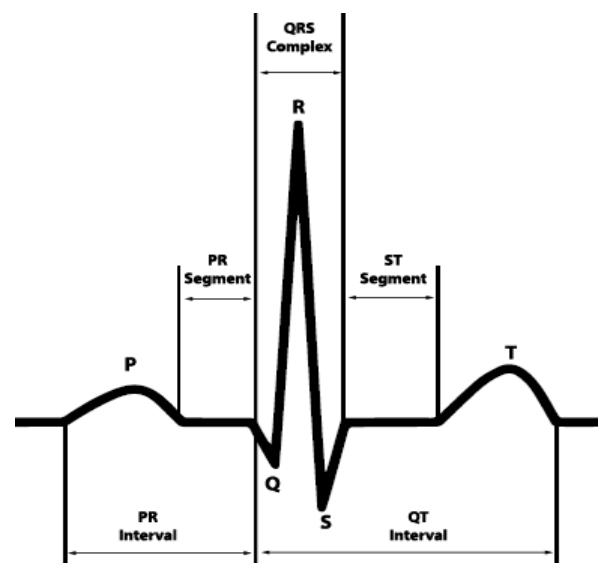


그림 1. 심전도 신호 [4]

데이터를 양자화하는 가장 간단한 방법은 균일 양자화(uniform quantization) 방법이다. 그림 1 과 같은 심전도 신호를 균일하게 양자화하면, R 피크에 비해 상대적으로 좁은 범위의 전압을 가지는 P, Q, S, T 피크 부근의 신호가 동일하거나 거의 같은 값으로 양자화 되는 경우가 생기는 반면 넓은 범위의 전압을 가지는 R 피크 값 주변의 신호가 다양한 값으로 양자화되기도 한다. 심혈관 질환을 분류할 때 R 피크 값도 중요하지만, P, Q, S, T 피크 값 역시 중요한 역할을 할 수 있다.[5] 본 연구에서는 P, Q, S, T 피크 주변에 존재하는 좁은 범위의 전압 신호를 보다 세밀하게 양자화하고, R 피크 부근의 넓은 범위의 전압 신호를 개략적으로 양자화할 수 있는 비균일 양자화 방법으로, 사분위수(quartile), 히스토그램 평활화(histogram equalization) 양자화 방법을 사용하였다. 사분위수 방법은 양자화 레벨의 75%를 사분위수 범위(IQR: 제 1 사분위와 제 3 사분위의 차)에 속하는 전압을 양자화하는데 사용하고, 나머지 25%의 양자화 레벨을 사분위수 범위 밖의 전압을 양자화하는데 사용하는 방법이다. 히스토그램 평활화 방법은 심전도 신호의 히스토그램을 먼저 생성한 후, P, Q, S, T 피크 주변에 밀집되어 있는 히스토그램을 평활화하여 많은 레벨의 양자화 값을 가지게 하고, R 피크 주변은 상대적으로 적은 레벨로 양자화하는 방법이다. 상기의 3 가지 양자화 방법과 대표적인 비균일 양자화 방법인 로이드-맥스(Lloyd-Max) 방법을 이용하여 심전도 신호를 양자화 하였다. 양자화 할 때 사용한 양자화 레벨은 32 단계이고, GLCM 행렬에서 기준점과 이웃하는 샘플의 오프셋 거리는 2 샘플 거리를 선택하였다.

양자화 된 심전도 신호에서 33 개의 특징값을 다음과 같이 추출하였다: 심전도 신호의 히스토그램에서 11 개, GLCM 행렬에서 9 개, GLRLM 행렬에서 13 개의 특징값을 각각 추출하였다. 본 연구에서 사용한 심전도 데이터는 12 리드 데이터이므로 한 명의 심전도 데이터에서 추출한 특징값은 모두 396 개이다.

양자화에 따른 심혈관 질환 분류를 위해 로지스틱 회귀(logistic regression)를 사용하였다. 분류 모델의 학습을 위해 전체 데이터의 80%를 학습데이터로 사용하였고, 나머지 20%는 테스트데이터로 사용하였다. 학습 모델의 유효성을 검증하기 위해 5 겹 교차검증(5-folded cross validation)을 사용하였다. 분석에 사용한 프로그램은 Python 3.7.3 과 Scikit-learn 0.24.2 이다.[6] 양자화 방법이 심혈관 질환 분류에 미치는 영향을 확인하기 위해서 테스트데이터의 정확도를 표 1 에 기술하였다.

	LR	DT	RF
균일 양자화 방법	87.2	76.7	86.6
로이드-맥스 방법	87.6	78.8	87.5
사분위수 방법	88.1	77.0	84.1
히스토그램 평활화 방법	88.2	76.0	83.4

표 1. 양자화 방법에 각 기계학습 알고리즘의 성능 비교. (LR: 로지스틱 회귀, DT: 결정트리, RF: 랜덤 포레스트)

III. 결론

본 논문에서는 1 차원 심전도 신호의 형태학적 정보를 가지는 GLCM, GLRLM 행렬에서 특징값을 추출할 때 선행되는 양자화 방법이 심혈관 질환 분류시스템의 성능에 미치는 영향에 대해서 알아보았다. 표 1 의 결과에도 나와있듯이 로지스틱 회귀와 네 가지의 양자화 방법과 결합했을 때의 결과가 결정트리나 랜덤 포레스트보다 더 좋은 결과를 보였다. 전체적으로 보면, 로지스틱 회귀와 히스토그램 평활화 방법이 가장 높은 정확도를 보였다.

로지스틱 회귀와 네 가지 양자화 방법의 결과를 보면, P, Q, S, T 주변의 전압값을 세밀하게 양자화하는 사분위수와 히스토그램 평활화 방법이 균일 양자화 방법과 로이드-맥스 방법보다 더 좋은 결과를 보였다. 이는 P, Q, S, T 주변의 전압값이 심혈관 질환의 분류에 중요한 역할을 한다는 것을 간접적으로 보여준다고 볼 수 있다.

비록 본 실험에서 히스토그램 평활화 방법이 가장 좋은 성능을 보였지만, 이 방법이 항상 최선의 방법은 아니다. 본 실험에서는 심혈관 분류기로 로지스틱 회귀, 결정트리, 랜덤 포레스트만 고려하였고, 양자화의 레벨도 32 레벨, GLCM 에서 사용한 오프셋 거리도 2 샘플만 고려하였다. 보다 면밀한 결과를 위해서 SVM(support vector machine), kNN(k-nearest neighbor), XGBoost 등과 같은 다양한 분류기를 사용해야 하고, 양자화 레벨도 32 레벨 뿐만 아니라 16, 64, 128 레벨 등을 고려해야 한다. 그리고 GLCM 의 오프셋 거리도 1 샘플 거리에서부터 10 샘플거리까지 다양하게 바꾸어 가면서 실험이 행해져야 한다.

ACKNOWLEDGMENT

이 성과는 정부(과학기술정보통신부)의 재원으로

한국연구재단의 지원을 받아 수행된 연구임

(No.2020R1G1A1A1102881)

참 고 문 헌

- [1] Parvaneh, S and Rubin, J. "Electrocardiogram Monitoring and Interpretation: From Traditional Machine Learning to Deep Learning, and Their Combination," Computing in Cardiology, 45, 2018.
- [2] Arif M. et al. "Robust electrocardiogram(ECG) beat classification using discrete wavelet transform," Physiological Measurement, 29(5), 2008.
- [3] Lee, G. N. et al. "Effect of quantization on co-occurrence matrix based texture features: An example study in mammography," Proc. SPIE 6144, Medical Imaging, 2006.
- [4] Zheng, J. et al. "A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients," Scientific Data, 7(48), 2020.
- [5] Magnani J.W. et al. "P wave duration is associated with cardiovascular and all-cause mortality outcomes: the National Health and Nutrition Examination Survey," Heart rhythm, 2011.
- [6] <https://scikit-learn.org/stable/index.html>