Shapelet을 이용한 시계열 패턴 분류

백한솔*, 사재원, 김희곤, 정용화, 박대희
*고려대학교 컴퓨터정보학과
e-mail:hansol100@korea.ac.kr

Classification of Time Series Patterns using Shapelet

*Hansol Baek, Jaewon Sa, Heegon Kim, Yongwha Chung, and Daihee Park
*Dept. of Computer and Information Science, Korea University

요 약

기술의 발전에 따라 소형 디바이스에서도 데이터를 수집하고 전송하는 것이 가능해졌다. 따라서 최근 IoT와 헬스케어가 부각되고 있으며 여기서 발생한 데이터에 대한 많은 연구가 진행되고 있다. 그 중에서도 헬스케어 장비에 내장된 심전도 센서를 이용하여 시계열 데이터를 수집할 수 있고, 여기서 수집한 데이터는 부정맥 등의 심장질환 진단의 중요한 지표로서 사용될 수 있다. 시계열 데이터는 시계열 분석 방법을 사용하여 정상 패턴과 비정상 패턴으로 분류할 수 있지만, 대량의 시계열 분석 방법은 수행시간이 많이 소요되기 때문에 이를 단축 할 필요성이 있다. 본 논문에서는 시계열 데이터 분석 기법 중 하나인 Shapelet을 사용하여 심전도 데이터의 패턴을 정상 및 비정상으로 분류하였고, 병렬처리 기법을 적용하여 수행시간을 단축하였다. 실험 결과, 각각의 심전도 데이터는 87%의 정확도로 분류되었고, Shapelets을 탐색하는 구간의 병렬처리를 통하여 수행 시간이 약 60%로 감소하였음을 확인하였다.

1. 서론

최근 사물 인터넷(Internet of Things, IoT)은 가전제품, 스마트 홈, 스마트 카, 헬스케어 등 다양한 분야에서 네트워크를 통해 각 센서로부터 수집된 정보를 공유할 수 있다. 특히, 헬스케어 분야에서 스마트 워치, 스마트 밴드 등의 디바이스는 심전도 센서를 부착하여 실시간으로 사용자의 상태를 확인할 수 있다[1]. 심전도(Electrocardiogram, ECG) 데이터는 심장이 활동하며 발생한 전기적 자극을 시간에 따라 기록한 데이터이며 ECG 분석[2-12]을 통해 부정맥과 같은 심장 질환을 조기에 진단할 수 있다. 심전도는 P, Q, R, S, T 5개의 파형으로 나타나며 그 중 QRS 콤플렉스와 R 피크를 이용하여 분류하는 연구[10]가 보고되고 있다.

심전도는 시계열 데이터[13]의 형식이기 때문에 시계열 데이터 분석 기법 중 하나인 Shapelet[14]을 사용하면 심전도의 상태를 정상 및 비정상으로 판단할 수 있다. 시계열 분석 기법으로 얻은 Shapelet은 시계열 데이터에서 여러 클래스간의 거리를 최대로 하는 부분 시퀀스를 의미한다. Shapelet을 이용한 분류는 해당 Shapelet과 새로 입력된 시계열 데이터 간의 거리를 구하여 임계값을 기준으로 그 이하면 정상, 그 이상이면 비정상으로 분류한다. 각 시계열 데이터를 구분 할 수 있는 Shapelet을 찾기 위한 학습 시간이 오래 걸리는 단점이 있지만, 구분이 가능한 Shapelet이 찾아진 이후에는 분류를 위한 부분에서는 필요한 연산이 적기 때문에 정상과 비정상을 빠른 시간 안에 분류 할 수 있는 장점이 있다.

본 논문에서는 시계열 데이터 셋으로부터 Shapelet을 찾고, 이 과정에 병렬처리 기법을 적용하여 수행 시간을 감소시키는 방법을 제안하였다. 또한, 탐색된 Shapelet을 이용하여 정상 및 비정상 클래스로 분류된 데이터들의 정확도를 확인하였다.

2. 관련 연구

ECG 패턴을 이용한 다양한 연구가 진행되고 있다[9-12]. 예를 들어, ECG 패턴을 이용한 운전자의 인지 부하 평가 방법[11]은 운전자의 신체에 ECG 센서를 부착하고 여러 상황에서 개인별 ECG 변화 수치가 적정 수치가 되면 운전하기 적합한 상태로 판단하고, 너무 높거나 낮은 수치라면 인지 부하 상태로 판단하게 된다. 조기 수축 부정맥을 분류하기 위에 ECG를 도입하여 템플릿 문턱치와 RR 간격을 적용하여 분류를 수행하는 연구[12]도 진행되었다. 부정맥은 ECG를 통해 즉각적으로 알 수 있는 대표적인 질환이다. 부정맥이 발생하면 곧바로 심장 박동, 맥박에 영향을 주기 때문에 ECG를 이용하면 빠르게 진단을 내릴 수 있다.

위와 같은 기존 연구들은 각각 사람이 패턴을 파악하고 그에 맞춰 적절한 알고리즘을 수정해야 할 뿐만 아니라 파라미터를 실험적으로 설정해야 하는 단점이 있다. 본 논문에서는 Shapelet을 이용하여 ECG의 특징 부분을 자동으로

빠르게 추출하여 ECG 데이터를 분류하는 방법을 제안한다.

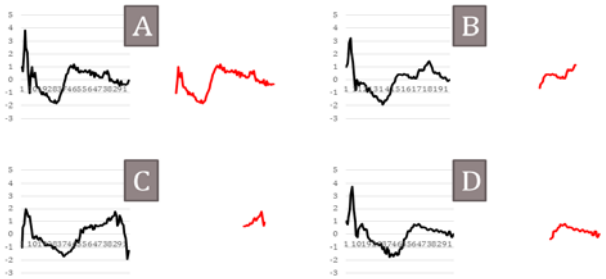
3. 제안 방법

2.1 Shapelet의 특성

Shapelet[14]은 여러 시계열 데이터를 각각의 다른 클래스간의 거리를 최대로 하는 서브 시퀀스이다. 먼저 수식 (1)을 이용하여 분류되기 전의 데이터로부터 엔트로피를 계산한다.

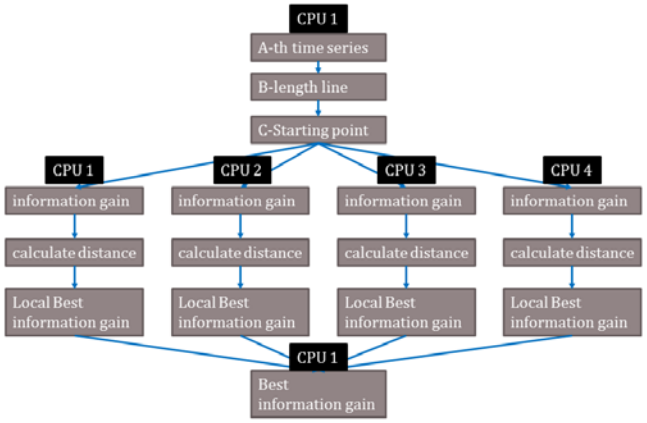
$$I(D) = -p(A)\log(p(A)) - p(B)\log(p(B)) \qquad (1)$$

A, B는 각 클래스의 이름, p는 전체 데이터에서 해당 클래스가 차지하는 비율을 의미한다. 다음으로 Information gain을 계산한다. Information gain은 특정 서브 시퀀스로 분류한 이후의 엔트로피와 분류 전과 후의 엔트로피 차이로 정의 되고, Information gain이 최대로 되는 서브 시퀀스를 Shapelet으로 결정하게 된다. 이 Shaplet으로 분류된 각 데이터들도 위와 같은 과정을 반복적으로 수행하여 새로운 Shapelet을 탐색한다. 이후 Shapelet으로 분류된 데이터가 각각 오직 하나의 클래스의 데이터만을 포함할 때까지 반복한다. 그림 1은 Information gain이 최대로 되는 Shapelet을 탐색한 결과이다. 좌측의 기존 시계열 데이터로부터 우측의 서브 시퀀스를 찾아내었고 이를 이용하여 2.2절에서 트리를 형성하여 분류 수행의 지표로써 활용된다.



(그림 1) Time Series Data and Shapelet

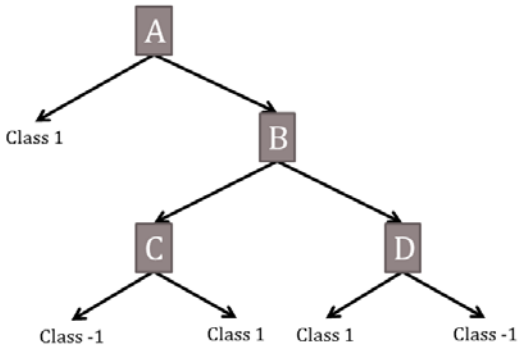
이 과정은 모든 시계열 데이터에 대하여 전수조사가 수행되기 때문에 수행 시간이 오래 걸리는 문제점이 있고, 본 논문에서는 멀티코어 CPU를 활용한 병렬처리 기법을 적용하여 수행 시간을 단축한다. 그림 2는 병렬처리 과정에 대하여 CPU의 각 코어에 작업량을 분배한 그림이다. 모든 과정에서 병렬처리를 수행할 경우, 앞서 계산한 값을 할당하는 과정에서 경쟁 상태가 발생하여 순차처리와는 다른 결과를 얻게 된다. 이러한 경쟁 상태를 피하기 위하여 각각의 코어에서 할당받은 쓰레드는 데이터 중 최대의 information gain을 구하고 추후 마스터 쓰레드에서 각각의 최댓값들을 비교하여 가장 큰 값을 할당한다.



(그림 2) 4-코어 CPU상에서의 병렬처리 예시

2.2 Shapelet을 이용한 데이터 분류

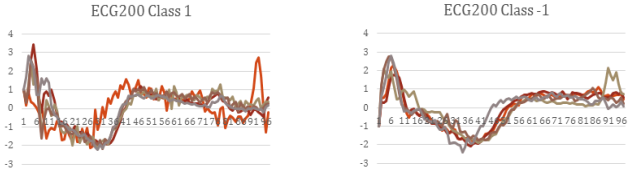
Shapelet이 추출되면 그림 3과 같은 트리가 구성된다. 트리의 루트에 분류 할 데이터를 대입하면 각 노드에서 Shapelet과의 거리를 계산하고 임계 값 이하이면 왼쪽으로, 그 이상이면 오른쪽으로 진행하고 리프에 도달하면 해당 클래스를 최종적으로 분류한다.



(그림 3) Classification Decision Tree

4. 실험 결과

본 논문에서의 실험은 Intel Core i5-4690(4-코어) 3.5GHz, 8GB RAM, Windows 10 의 환경에서 수행되었다. 본 실험의 데이터는 UCR[15]에서 제공하는 ECG 데이터를 이용하였다. Shapelet을 구하기 위하여 100개의 ECG 데이터 셋을 사용하여 학습하였고, 또 다른 100개의 ECG 데이터 셋을 사용하여 정확도를 계산하였다.



(그림 4) Train Data Set

그림 4와 같이 분류된 2개의 클래스 데이터를 이용하여 Shapelet을 탐색하였다. 탐색한 Shapelet에 대한 실험은 학

습에 사용한 데이터를 제외한 100개의 데이터를 분류 하도록 진행하였다. Shapelet을 탐색할 때의 순차처리 수행 시간이 63.55초, 제안한 병렬처리 수행 시간이 38.16초로, 기존의 순차처리 수행 시간과 비교하여 병렬처리 수행 시간이 약 60%로 감소되었음을 확인하였다(표 1 참조). 한편, 데이터 하나당 분류에 걸린 수행 시간은 0.8ms이였고, 분류를 위한 총 수행 시간은 84.151ms로 측정되어 빠르게 분류할 수 있음을 알 수 있었다. 마지막으로, 표 2의 실험 결과와 같이 전체 100개의 데이터를 87%의 정확도로 분류함을 확인하였다.

(표 1) Execution Time and speedup

	Execution Time (sec)	Speedup
Sequential method	63.55	1.67
Proposed method	38.16	

(표 2) Result of Shapelet Classification

Class Type	Correct classification	Failed classification	Accuracy (%)	Execution time (ms)
Class - 1	28	8	77.77	84.151
Class 1	59	5	92.18	

5. 결론

ECG 데이터는 현재 IoT 기기를 통해 수집하기 쉬우면서도 심장의 문제에 직결되기 때문에 의미하는 바는 매우 크다고 할 수 있다. 본 논문에서는 ECG 데이터를 Shapelet을 이용하여 분류하였고, 수행 시간을 감소시키기 위하여 병렬처리 기법을 적용하는 방법을 제안하였다. 실험 결과, 각각의 심전도 데이터는 87%의 정확도로 분류되었고, Shapelets을 탐색하는 구간의 병렬처리를 통하여 수행 시간이 약 60%로 감소하였음을 확인하였다. 또한, 100개의 시계열 데이터를 84ms의 수행시간으로 분류가 가능함을 확인하였다. 제안 방법을 연산능력이 일반 PC보다 상대적으로 부족한 임베디드 기기에 적용한다면 데이터를 서버에 보낼 필요 없이 기기 내에서 연산을 수행하고 이상이 발견 될 경우에만 데이터를 전송하는 방법 등으로 응용 할 수 있을 것으로 예상된다.

감사의 글

2016년도 미래창조과학부의 재원으로 한국연구재단의 지원을 받아 수행된 지역신산업선도인력양성사업(2016H1D5A1910730)으로 수행된 연구결과임.

참고문헌

[1] 신광식, 치우리안 야우, 정완영, “무선센서네트워크 기반의 모바일 유비쿼터스 헬스케어 시스템”, 한국해양정보통신학회논문지, 제10권, 제11호, pp.2107-2112, 2006.

[2] 송민, 최진명, 이희영, “24시간 건강 모니터링 시스템을 위한 심전도 신호의 순시 대역폭 추정 및 잡음 제거”, 대한전자공학회 종합 학술 대회 논문집, 2001.

[3] 이석원, 남부희, 장석호, “웨이블릿 변환과 RB 함수를 이용한 심전도 신호의 처리”, 제어계측 자동화 로보틱스 연구회 합동 학술 발표회 논문집, 1999.

[4] 김영섭, 홍성호, 지용석, 이명석, 노학엽, “ECG 분석을 위한 R-R Interval 탐지 시스템”, 한국정보통신설비학회 논문지, 제11권, 제2호, pp.29-33, 2012.

[5] 배정현, 임승주, 김정주, 박성대, 김정도, “ECG 신호에서 단위패턴간 유사도 분석을 이용한 부정맥 분류 알고리즘”, 정보처리학회논문지, 제19권, 제1호, pp.105-112, 2012.

[6] 윤상훈, 강원석, 권형호, “PSO를 활용한 심전도 분류기의 파라미터 최적화”, 대한전자공학회 하계종합학술대회, pp. 876-878, 2013.

[7] 황구연, 신동규, 신동일, “생체신호 분석을 이용한 바이오피드백 인터페이스 설계”, 한국컴퓨터종합학술대회 논문집, 제39권, 제1호(A), pp.337-339, 2012.

[8] 이원섭, 박장운, 김수진, 윤성혜, X. Yang, 이용태, 손준우, 김만호, 유희천, “운전 생체신호 및 운전 수행도 분석 System 개발”, 대한인간공학회지, 제29권, 제1호, pp.47-53, 2010.

[9] 윤상훈, 강원석, “웨이블릿 변환과 서포트 벡터 머신을 이용한 심전도 분류”, 한국HCI학회 학술대회, pp.611-613, 2013.

[10] 조익성, 정종혁, 권혁승, “대상 유형별 ECG 신호의 QRS 패턴을 이용한 부정맥 분류”, 한국정보통신학회논문지, 제19권, 제7호, pp.1728-1736, 2015.

[11] 홍원기, 이원섭, 정기효, 이백희, 박장운, 박수완, 박윤숙, 손준우, 박세권, 유희천, “ECG 기반의 운전자별 인지 부하 평가 방법 개발”, 대한산업공학회지, 제40권, 제3호, pp.325-332, 2014.

[12] 조익성, 조영창, 권혁승, “ECG 패턴 분석과 템플릿 문턱값을 통한 조기수축 부정맥 분류”, 한국정보통신학회논문지, 제20권, 제2호, pp.437-444, 2016.

[13] 이수용, 이경중, “시계열 자료의 데이터마이닝을 위한 패턴 분류 모델 설계 및 성능 비교”, 한국지능시스템학회 논문지, 제21권, 제6호, pp.730-736, 2011.

[14] L. Ye and E. Keogh., “Time Series Shapelets: A New Primitives for Data Mining,” Proc. of the ACM SIGKDD, pp. 947-956, 2009.

[15] The UCR Time Series Classification Archive(2015). URL www.cs.ucr.edu/~eamonn/time_series_data/