

시계열 데이터를 분류를 위한 기계 학습 모델 설계 및 성능 비교

김영송, 양성민, 조서형

skins346@naver.com, ysm0622@gmail.com, westbro00@naver.com

Design and Performance Comparison of Machine Learning Model for Time Series Data Classification

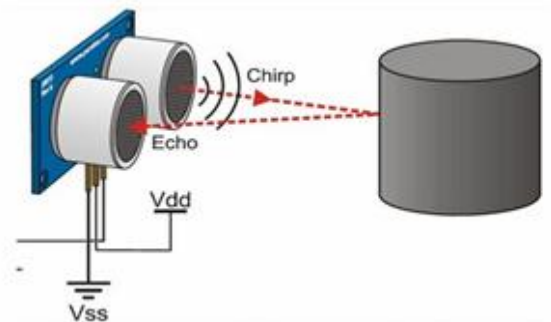
Kim Young Song, Yang Sung Min, Cho Seo Hyung

요약

사물 인터넷 시대가 도래함에 따라 웨어러블 디바이스의 개수가 기하급수적으로 증가하고 있다. 그에 상응하여 웨어러블 디바이스의 센서로부터 대규모의 시계열 데이터가 발생한다. 시계열 데이터란 일정 기간에 대해 시간의 함수로 표현되는 데이터를 말한다. 시계열 데이터는 과거의 시계열 데이터 분석을 통해 앞으로의 사건을 예측한다는 점에서 중요한 의미를 가진다. 본 논문에서는 이런 시계열 데이터의 처리 및 예측을 위해 SVM(Support Vector Machine), RF(Random Forest), MLP(Multilayer Perceptron)를 활용하여 세 가지 학습모델을 설계한다. 기계학습 방법을 실제 사물인터넷 시스템에 활용해보는 사례 연구를 수행하여 위 세 가지 학습모델의 성능 평가를 실시하였다.

I. 서론

최근 사물 인터넷 시대의 도래와 함께 다양한 웨어러블 디바이스가 등장하고 있다. 또한 이러한 웨어러블 디바이스의 개수는 기하 급수적으로 증가하고 있다. 여기서 우리가 주목해야할 점은 이런 디바이스의 증가와 함께 대규모 시계열 데이터도 함께 증가한다는 것이다. 시계열 데이터란 일정 기간에 대해 시간의 함수로 표현되는 데이터를 말한다. 시계열 데이터는 과거의 시계열 데이터 분석을 통해 앞으로의 사건을 예측 할 수 있다는 점에서 중요한 의미를 가진다. 시계열 데이터 예측은 주로 학습 데이터를 이용해 학습 모델을 생성하여 입력 데이터를 특정 시간의 값으로 예측하는 방식을 사용한다. 이런 시계열 데이터를 처리하는 방법으로 지금까지 다양한 기술들이 연구되고 소개되었다.[1,2] 대표적인 기계 학습 방법으로는 SVM(Support Vector Machine), Decision Tree, Hidden Markov model, Naive Bayes Classifier 등을 비롯하여 최근에 다시 각광받고 있는 심층 신경망(Deep Neural Network) 등이 있다.[3,4] 본 논문에서는 시계열 처리의 다양한 방법들 중 SVM(Support Vector Machine), RF(Random Forest), MLP(MultiLayer Perceptron) 총 세 가지 방법을 활용하여 시계열 데이터처리 성능을 비교 분석한다. 시계열 데이터를 처리하고 분류해 보는 사례연구를 수행하였고, 이를 통해 위 세 가지 방법들의 시계열 데이터처리 성능을 비교 분석한다.



(그림 1) 초음파 센서(HC-SR04) 측정 원리

II. 시계열 데이터

II-1. 데이터 수집

본 논문에서 시계열 데이터는 저비용의 초음파 거리 측정 센서(HC-SR04)에서 발생한 데이터를 사용한다. 초음파 거리 측정 센서는 사람의 귀에 들리지 않을 정도의 높은 주파수(약 20 kHz 이상)의 소리를 발산하여 음파가 물체에 부딪혀 되돌아올 때까지의 시간을 측정함으로써 물체까지의 거리를 측정한다. 음파는 15℃에서 약 340 m/s의 속도를 가지며 이에 따라 ($V=340\text{ m/s}$, t =초음파 발산(ping) 후 수신(echo)되기 까지 시간) 공식에 의해 물체까지의 거리(L)를 측정 할 수 있다.

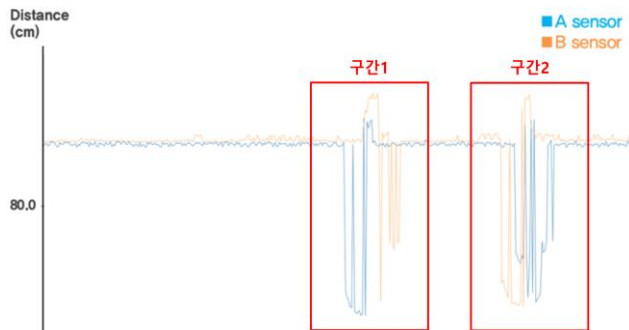
<표 1> 초음파 센서(HC-SR04) 모듈 사양

Electrical Parameters	HC-SR04 Ultrasonic Module
Operating Voltage	DC-5V
Operating Current	15mA
Operating Frequency	40KHZ
Farthest Range	4m
Nearest Range	2cm
Measuring Angle	15 Degree
Input Trigger Signal	1Dus TTL pulse
Output Echo Signal	Output TTL level signal, proportional with range
Dimensions	45*20*15mm



(그림 2) 초음파 거리 측정 센서 설치 구조

위 원리를 이용하여 (그림 2)와 같이 초음파 거리 측정 센서를 출입문에 병렬로 설치하여 사람이 문을 통과할 때 발생하는 시계열 데이터를 실험에 활용한다. 초음파 거리 측정 센서로부터 발생하는 데이터는 (그림 3)과 같은 그래프를 형성한다.



(그림 3) 시간에 따른 두 센서의 거리 그래프

(그림 3)의 구간 1 은 사람의 입장 행동, 구간 2 는 사람의 입장 행동을 나타낸다. 시계열 데이터를 활용하여 사람의 입장 행동과 퇴장 행동을 예측함으로써 사용된 기계 학습의 성능을 비교 분석한다.

II-2. 데이터 전처리

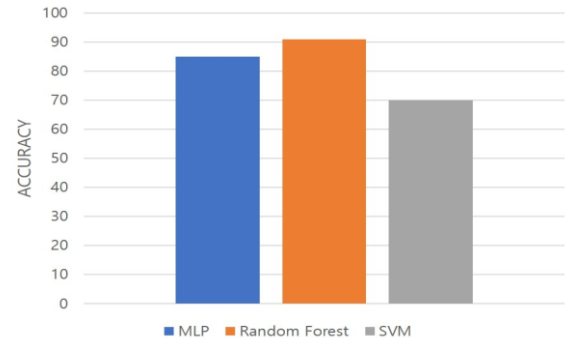
두 개의 초음파 센서를 통해 출입문에 사람이 통과할 때 발생하는 데이터를 사용한다. 먼저 사람이 출입문을 통과할 때 초음파 거리 측정 센서로부터 200Hz 로 데이터를 샘플링 한다. 샘플링 된 데이터 중에서 샘플링 된 시간이 빠른 순서로 100 개의 데이터를 특징(feature)으로 선택하여 100 차원의 데이터로 벡터화한다. 다음 이 벡터 데이터를 입력 데이터로 사용한다. 이때 샘플링 되지 않은 시간의 데이터는 0 으로 처리한다. 마지막으로 출입문 입장과 퇴장을 두 가지 클래스로 구분하였다.

III. 실험 및 측정 결과

Random Forest 는 의사결정트리(decision tree)의 앙상블(ensemble) 학습 기법을 사용하는 모델로서 주어진 데이터로 여러 가지 모델을 학습한 뒤, 그

모델들의 결과를 종합해 사용하여 정확도를 높인다. 동시에 일반화를 위해 Random space method 를 통해 과적합(overfitting)문제를 피하는 학습 기법이다. 여러 개의 Decision tree 가 분류한 최종 예측 결과를 투표(voting) 방식으로 최종 결과를 결정한다. 본 논문에서는 Decision tree 의 개수를 50 개, sampling parameter 는 100 개로 실험을 진행하였다.

MLP(MultiLayer Perceptron)의 은닉층은 시계열 데이터 특성상 시간 축으로 가까운 데이터들 사이의 correlation 이 높기 때문에 높은 correlation 을 갖도록 학습시키기 위해 2 층으로 구성하였으며 MLP 의 경우 비선형적이고 복잡한 분류 문제에 대해 좋은 성능을 나타내기 때문에 비교 실험을 하기에 적절하다.



(그림 4) 분류 성능 결과

MLP, RF, SVM 의 분류 정확도는 (그림 4)와 같다. 초음파 거리 측정 센서 특성상 난반사(diffuse reflection)로 인해 잡음(noise)이 많이 포함된 데이터를 학습하기 때문에 비선형적인 패턴 분류에 대해서는 SVM 이 MLP 와 Random Forest 보다 좋은 성능을 나타내지 못하였다.

IV. 결론

본 논문에서는 세 가지의 기계학습 방법을 이용해 잡음이 많이 포함된 초음파 센서로부터 입력되는 시계열 데이터를 분류하기 위한 비선형 패턴분류 모델을 설계 및 비교 평가하였다. 그 결과 RF 와 MLP 는 SVM 보다 잡음과 Outlier 가 많은 비선형적 데이터에 대하여 높은 성능을 보였다. 본 논문에서 진행한 실험 결과가 향후 또 다른 시계열 데이터를 처리할 때 활용 될 수 있음을 기대한다.

참 고 문 헌

- [1] P Casale, O Pujol, P Radeva, "Human activity recognition from accelerometer data using a wearable device", Pattern Recognition and Image Analysis, 2011.
- [2] A Vahdatpour, M Sarrafzadeh, "Unsupervised discovery of abnormal activity occurrences in multi-dimensional time series, with applications in wearable systems", Proceedings of the 2010 SIAM, 2010
- [3] RJ Rrank, N Davey, "Time series prediction and neural networks", Journal of Intelligent & Robotic System, 2001.
- [4] CLGiles, S Lawrence, AC Tosi, "Noisy time series prediction using recurrent neural networks and grammatical inference", Machine learning, 2001.