컨벌루션 신경망과 변종데이터를 이용한 시계열 패턴 인식

안명호* · 류미현**

*주) MHR

Convolutional Neural Network and Data Mutation for Time Series Pattern Recognition

Myong-ho Ahn* · Mi-hyeon Ryoo**

*MHR Inc

E-mail : mhr.james@gmail.com

요 약

TSC(Time Series Classification)은 시계열데이터를 패턴에 따라 분류하는 것으로, 시계열이 매우 흔한 데이터형태이고, 또한 활용도가 높기 때문에 오랜 시간동안 Data Mining 과 Machine Learning 분야의 주요한 이슈였다. 전통적인 방법에서는 Distance와 Dictionary 기반의 방법들을 많이 활용하였으나, Time Scale과 Random Noise의 문제로 인해 분류의 정확도가 제한되었다. 본 논문에서는 Deep Learning의 CNN(Convolutional Neural Network)과 변종데이터(Data Mutation)을 이용해 정확도를 향상시킨 방법을 제시한다. CNN은 이미지분야에서 이미 검증된 신경망 모델로써 시계열데이터의 특성을 나타내는 Feature를 인식하는데 효과적으로 활용할 수 있고, 변종데이터는 하나의 데이터를 다양한 방식으로 변종을 만들어 CNN이 특정 패턴의 가능한 변형에 대해서도 학습할 수 있도록 데이터를 제공한다. 제시한 방식은 기존의 방식보다 우수한 정확도를 보여준다.

ABSTRACT

TSC means classifying time series data based on pattern. Time series data is quite common data type and it has high potential in many fields, so data mining and machine learning have paid attention for long time. In traditional approach, distance and dictionary based methods are quite popular. but due to time scale and random noise problems, it has clear limitation.

In this paper, we propose a novel approach to deal with these problems with CNN and data mutation. CNN is regarded as proven neural network model in image recognition, and could be applied to time series pattern recognition by extracting pattern. Data mutation is a way to generate mutated data with different methods to make CNN more robust and solid. The proposed method shows better performance than traditional approach.

키워드

CNN, Time Series, Classification, Data, Pattern, Deep Learning

Ⅰ. 서 론

시계열데이터분류(Time Series Classification)은 데이터마이닝과 기계학습 분야에서 인기있는 주제로 비교적 오랜 시간동안 많은 사람들의 꾸준한 관심을 불러일으키기고 있는 분야이다.

TSC는 시간에 따라 데이터가 변하는 데이터를 특성에 따라 분류하는 것으로 과학,사회,경제,의료 등 다양한 분야에서 관찰되기 때문에 활용범위가 넓다고 할 수 있다.

예를 들어 서버나 데이터센터의 전력사용량을 예측하거나, 전자기파의 에러검출, 주가예측 그리고, ECG 데이터를 이용한 심장병 발생예측등 많은 곳에서 적용되고 있는 것이 현실이다.

하지만 TSC는 시계열 데이터를 다루기 때문에 데이터의 패턴이 데이터들간의 상관관계뿐만 아니라 시간에 따른 순서와도 밀접한 관계가 있기 때문에 비시계열데이터에 비해 패턴을 찾고, 분류하는 것이 쉽지 않은 것이 사실이다.

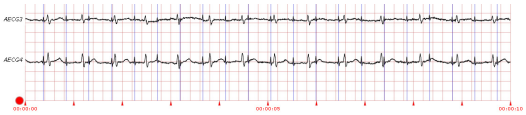

그림 1. TSC 예시

정확도 높은 TSC를 구현하는데는 크게 2가지의 기술적 문제가 있다.

첫번째는 시계열데이터의 불규칙한 시간간격(Time Scale)이다. 예를 들어 서버의 에너지 사용량 시계열 데이터의 패턴이 있을때, 해당 패턴들이 모두 동일한 시간간격내에 존재하지 않고 서로 다른 시간간격내에 존재하기 때문에, 패턴을 인식하기 위해서는 동일한 패턴의 데이터들을 적정한 시간간격으로 변화시키거나 혹은 이러한 시간간격을 무시하고 특징(Feature)를 찾아내는 등의 방법이 필요하게 된다.

두번째는 불규칙적인 노이즈를 이야기 할 수 있다. 측정시점에서의 계측장비나 혹은 상황의 변화로 인해 노이즈가 발생할 수 있고, 측정대상이 어떠한 이유로 인해 엉뚱한 값을 만들어낼 수도 있다. 따라서 이러한 노이즈를 최소화시키고 시계열데이터의 고유한 특징을 가지고 있는 데이터를 추출하거나, 노이즈를 제거하는 방법이 요구된다.

본 논문에서는 CNN과 Data Mutation을 이용하여 이러한 문제를 해결하고 우수한 성능의 시계열데이터 패턴 인식 방법을 제시한다.

II. 본 론

본 논문에서는 CNN을 이용해 시계열 데이터의 패턴을 인식해 분류하는 방법을 제시한다.

핵심 아이디어는 1) 원본데이터를 다양한 방법으로 변환해 다수의 변종데이터를 만들고 이를, 2) CNN을 이용해 다양한 변종데이터의 특성을 학습하도록 해, 앞서 언급한 2가지 문제점을 해결하는 것이다.

CNN을 학습시키는데 있어 하나의 CNN에 모든 종류의 변종데이터를 입력해 학습시키지 않고, 각각의 변종데이터별로 CNN을 학습시키는 방법을 취한다.

즉 n개의 변종데이터가 있다면 CNN의 갯수는 변종데이터수 CNN과 원래데이터를 이용한 CNN으로 총 n+1이 된다.

변종 데이터는 크게 1) Smooth와, 2) Filtering 그리고 3) Sampling의 3가지 방법을 적용해 생성한다. 예를 들어 원본 시계열데이터가 있다면 smoothing을 적용한 데이터, filtering을 적용한 데이터 그리고 sampling을 적용한 데이터로 복수개가 생성되고 이들을 각각의 CNN에 입력데이터로 사용한다.

다수의 CNN을 이용하면 각각의 CNN이 주어진 데이터에서 특성을 추출해 서로 다른 CNN을 가지게 된다. 분류하고 싶은 시계열데이터를 학습된 모든 CNN에 입력하면 각각의 결과가 도출되는데 이들 CNN의 정확도가 다르기 때문에 이를 최적화시킬 수 있는 방법으로 Softmax를 이용한다.

1. CNN(Convolutional Neural Network)

[1]CNN은 Hubel and Wiesel's의 고양이 시신경연구에서 영감을 받아 개발된 MLP(Multi Layer Perceptron)이다. CNN은 다층 신경망 구조를 이용해 주어진 데이터에서 각 레이어별로 Local Feature를 찾고, 이렇게 찾아진 Local Feature들에서 다시 Local Feature를 찾아 전체적으로 Global Feature를 찾아가는 방식이라 할 수 있다. 각각의 레이어별로 Local Feature를 가지고 있기 때문에 데이터가 가지고 있는 Feature를 놓치지 않고 패턴으로 인식할 수 있어 크기가 크고, 복잡한 데이터에도 효과적으로 적용할 수 있다.

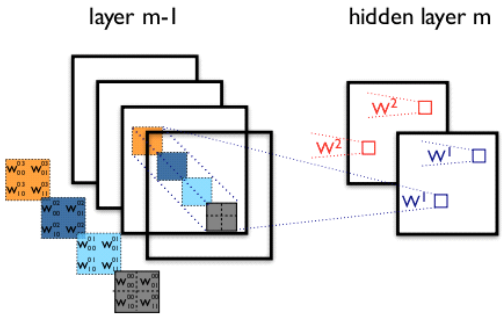

그림 2. Convolutional Neural Network

CNN에서 인식하는 Feature Map은 다음의 수식을 통해 학습된다.

$$h_{ij}^{k} = G((W^{k} * x)_{ij} + b_{k})$$

G는 Activation 함수로 tanh, sigmoid,relu등을 사용할 수 있다. W는 weight, b는 bias를 의미한다. CNN은 다수의 h(f(x))를 가질 수 있기 때문에 데이터가 복잡하거나 크다면 레이어를 늘려 h(f(x))를 복수개로 만들어 사용할 수 있다.

2. Ensemble by Softmax

[2]softmax는 일반화된 logistic regression으로 다수의 클래스를 처리하고 싶을 때 사용할 수 있

다. 제안방법에서는 다수의 CNN을 사용하기 때문에, 하나의 데이터에 대해 다수의 CNN에서 패턴인식에 대한 결과값을 내놓는다. Softmax는 다수의 CNN에서 계산한 결과값을 통합해 최종적인 결과를 내놓기 위한 것으로 각 CNN별로 적절한 가중치를 계산해 최종 결과값을 계산한다.

$$P(y=j|X) = \frac{e^{x^T w_j}}{\sum_{k=1}^{K} e^{x^T w_k}}$$

위의 식에서 알 수 있듯이 CNN 각각의 패턴인식에 대한 확률값을 더하고, 이를 개별 CNN으로 나누어 준 것이기 때문에 일종의 정규화 과정이라 생각할 수 도 있다. 제안방법에서는 Softmax를 출력레이어(Output Layer)에 적용해 다수의 CNN의 결과값이 Softmax에 의해 조절되도록 학습시킨다.

3. Data Mutation

변종 데이터를 생성할 때는 원본데이터의 특성을 유지하면서 앞서 설명한 2가지 문제점인 시간간격과 노이즈에 강인하도록 만드는 것이 중요하다. 시간간격 문제를 처리하기 위해서는 변종데이터를 생성하는데 필요한 파라메터인 Window의 값을 다양하게 사용해 Time Scale을 변경한다. 아울러 Up-sampling으로 scale을 늘리고, down-sampling을 이용해 time scale을 줄여 CNN이 특정 패턴의 시간간격의 변화에 민감하지 않도록 학습시킨다. 노이즈에 강인하도록 만들기 위해서는 원본 데이터로부터 해당 데이터의 특징을 잘 표현할 수 있는 특성치(feature)를 추출해야 하는데, 이를 위해 [3]Upsampling, [4] hamming, blackman과 같은 filtering을 적용한다.

이렇게 생성된 변종 데이터들은 원본 데이터를 다양한 시간간격을 포함하고, 이상치와 특이치를 제외한 데이터를 가지고 있어 CNN에 다양한 원본 데이터의 변형에 노출시켜 다각도로 특성을 학습할 수 있도록 한다.

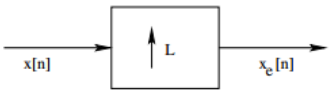

그림 3 Upsampling

III. 관련연구

최근에 시계열데이터 패턴인식에 CNN을 많이 활용하고 있다.

[5]Zheng은 다변량 시계열 데이터를 처리하기 위한 Multi-channel CNN이 제안하였다. 이 방법은 각 채널별로 CNN을 구성하고, 이들로 부터 특성(Feature)를 추출하고, 이렇게 각각의 CNN에서 검출된 특성을 모두 합하여 하나의 CNN을 구성해 시계열데이터 패턴을 인식하는 방법이다.

[6]Zheng은 또 CNN을 이용해 특성지표를 학습해 시계열데이터의 패턴을 인식하는 방법을 제시하였다. 이 방법은 시계열분류에서 전통적으로 사용되는 1-NN(Nearest Neighbour)을 이용해 CNN을 학습시키는 것으로, CNN에 직접 데이터를 입력하지 않고, 1-NN을 거쳐나온 값을 CNN에 입력값으로 사용해 CNN을 학습시키는 방법이다.

[7]Ye와 Keogh가 Shapelet을 제안하였는데 이 방식은 시계열데이터의 패턴을 shapelet이라고 불리는 작은 단위로 나누고 여기에 decision tree를 적용해 패턴을 인식하는데 많은 학자들의 관심을 불러일으켰다. TSC에서는 전통적인 방법으로 거리에 기반한 DTW(Dynamic Time Warping)류의 접근과 Interval기반의 분류인 TSF(Time Series Forest)를 많이 사용하였는데 Shapelet은 기존과 다른 방식이었기 때문으로 이해된다.

[8]Bagnall et al은 메타 앙상블 방식인 COTE(Collective of Transformation Ensembles)를 제안하였다. COTE는 시간, 자기상관, 파워스펙트럼 그리고 Shapelet들을 조합하여 사용하는 방식으로 기존의 DTW나 TSF 계열의 방법들보다 우수한 성능을 보여준다.

IV. 평가 및 검증

1. 베이스라인 수치

제안방법의 우수성을 확인하기 위해 비교할 방법들은 [9]1-NN Euclidean Distance(DTW1), 1-NN Best Warping Window DTW (DTW2), 1-NN DTW, no Warping Window(DTW3) 이다.

베이스라인의 선정기준은 전통적인 시계열데이터 분류방법과 최근에 제안된 의미있는 방법들로 제안방법이 기존방안대비 정확도를 비교하기 위함이다.

2. 평가 데이터셋

실험에 사용할 데이터는 [10]University of riverside에서 제공하는 UCR Time Series Classification Archive 데이터를 사용하였다.

UCR 데이터셋은 커피수요량 예측, 자세예측, 요가예측, 전력사용량 예측 등 실제 현장에서 발생한 데이터를 수집하여 정제한 것으로 연구목적으로 시뮬레이션된 데이터들이 아니기에 신뢰도가 높다. 또한 많은 시계열 데이터 패턴인식 논문에서 실험결과로서 사용하고 있기 때문에 타방법과의 비교를 하기에 적당하다.

3. 평가결과

제안방법과 베이스라인 수치와의 평가결과는 아래의 표에 기술되어 있다. 총 7개의 데이터셋을 비교한 결과 하나의 데이터셋을 제외하고는 제안방법인 CNN-DM이 우수한 결과를 보여주었으며, 어떤 것들은 높은 정확도를 보여주었다. 결과치를 산정하는데 Window,Step 크기등은 각각의 데이터셋에 따라 서로 다른 최적화된 파라메터를 사용했다.

Dataset	DTW1	DTW2	DTW3	CNN-DM
yoga	0.170	0.155	0.164	0.128
Electric Devices	0.450	0.376	0.399	0.341
Beef	0.333	0.333	0.367	0.333
FISH	0.217	0.154	0.177	0.0857
Trace	0.240	0.010	0.000	0.000
ECGFiveDays	0.203	0.203	0.232	0.034
Haptics	0.630	0.588	0.623	0.55

표 1 평가결과

Ⅴ. 결 론

본 논문에서는 CNN과 변종데이터를 이용한 시계열데이터패턴 인식에 대한 방법을 제시하였다. 제시한 방법은 특성인식성능이 우수한 CNN에 원본데이터를 변형한 다양한 데이터로 학습시키고, 이들을 Softmax와 앙상블 기법으로 사용해 패턴인식의 정확도를 높인다.

시계열데이터 패턴인식의 어려움은 시간간격과 노이즈에 기인한데, 변종데이터는 원본데이터를 서로 다른 시간간격으로 변환하고, CNN학습이 용이하도록 Filtering을 거치기 때문에 노이즈는 제거하고, 해당 데이터의 패턴을 잘 보여주는 데이터들을 기반으로 한다. 추후에는 패턴 인식을 위한 학습데이터에 Random Noise를 추가해 에러에 더욱 강인한 패턴 인식방법을 연구해볼 가치가 있다.

참고문헌

1) http://deeplearning.net/tutorial/lenet.html
2) http://en.wikipedia.org/wiki/Softmax_function
3) melodi.ee.washington.edu/courses/ee518/notes/lec9.pdf
4) en.wikipedia.org/wiki/Hamming_distance
5) Y. Zheng, Q. Liu, E. Chen, Y. Ge, and J. L. Zhao. Time series classification using multi-channels deep convolutional neural networks. In Web-Age Information Management, pages 298 – 310. Springer, 2014.
6) Y. Zheng, Q. Liu, E. Chen, J. L. Zhao, L. He, and G. Lv. Convolutional nonlinear neighbourhood components analysis for time series classification. In Advances in Knowledge Discovery and Data Mining, pages 534 – 546. Springer, 2015.
7) Y. Zheng, Q. Liu, E. Chen, J. L. Zhao, L. He, and G. Lv. Convolutional nonlinear neighbourhood components analysis for time series classification. In Advances in Knowledge Discovery and Data Mining, pages 534 – 546. Springer, 2015.
8) A. Bagnall, J. Lines, J. Hills, and A. Bostrom. Time-series classification with COTE: The collective of transformation-based ensembles. IEEE Transactions on Knowledge and Data Engineering, 27:2522 – 2535,
9) D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In KDD workshop, volume 10, pages 359 – 370. Seattle, WA, 1994.
10) Y. Chen, E. Keogh, B. Hu, N. Begum, A. Bagnall,A. Mueen, and G. Batista. The ucr time series classification archive, July 2015. www.cs.ucr.edu/~eamonn/time_series_data

안명호(Ahn Myong Ho)

고려대, KAIST SW 석사
주) MHR 대표이사(현)
*관심분야 : Deep Learning, Cloud Computing

류미현(Ryoo Mi Hyeon)

고려대,동국대 정보보호석사
주) MHR 수석연구원(현)
*관심분야 : Deep Learning, Security