지역 군집화를 위한 CNN-GRU 기반 다변량 시계열 데이터의 특성 추출

김진아*, 이지훈**, 최동욱**, 문남미**
*호서대학교 컴퓨터공학과
**호서대학교 컴퓨터정보공학부
e-mail:jina9406@gmail.com

Feature Extraction of CNN-GRU based Multivariate Time Series Data for Regional Clustering

Jinah Kim*, Ji-Hoon Lee**, Dong-Wook Choi**, Nammee Moon**
*Dept of Computer Engineering, Hoseo University
**Division of Computer Information Engineering, Hoseo University

요 약

시간의 연속성을 갖는 데이터에 대한 군집화 관련 연구는 주로 통계 분석을 통해 이뤄지기 때문에 데이터의 특성을 온전히 반영하지 못한다. 본 논문에서는 다변량 데이터에서의 군집화 방법을 위하여 변수별로 시간에 따른 변화와 특징을 추출하기 위한 CNN-GRU(Convolutional Neural Network - Gated Recurrent Unit) 기반의 신경망 모델을 제안한다. CNN을 활용하여 변수별로 갖는 특성을 파악하고자 하였으며, GRU을 통해 전체 시간에 따른 소비 추세를 도출하고자 하였다. 지역별로 업종에 따라 사용된 2년 치의 실제 카드 데이터를 활용하였으며, 유사한 소비 추세를 보이는 지역을 군집화하는데 이를 적용하였다. 결과적으로, 다변량 시계열 데이터를 통해 전체적인 흐름을 반영하여 패턴화했다는 점에서 의의를 갖는다.

1. 서론

시계열 데이터는 추세나 계절성 등에 따라 불규칙한 변동을 가지는 데이터로 비선형적인 특징을 가진다. 그리하여 VAR(Vector Autoregression Model), ARIMA(Auto - Regressive integrated Moving Average Model) 와 같은 통계 모델 기반에서 LSTM(Long Short-Term Memory models), GRU(Gated Recurrent Unit)과 같은 신경망 기반으로 주로 예측, 규칙 발견, 분류, 군집화 등 다양한 연구가 이뤄지고 있다[1-4].

그중에서 시계열 데이터 기반의 군집화 연구는 일반적으로 유사한 시간 영역에 대한 군집화가 많이 이뤄지고 있다. 그러나, 다변량 시계열 데이터의 경우에는 데이터에 영향을 끼치는 여러 다양한 변수들을 포함하기 때문에 이를 고려해 변수에 따라 다양하게 군집화가 가능하다. 이때 가장 중요한 것은 변수별 시계열의 특징을 추출하는 것이다.

본 연구에서는 다변량 시계열 데이터 기반의 군집화에 대하여 다루며, 소비 데이터를 기반으로 지역 군집화에 적용하고자 한다. 이를 위해 실제 2년 치의 카드 데이터를 활용하며, 변수별 시계열 특징을 CNN-GRU 신경망을 통하여 추출하는 방법을 제안한다. 최종적으로, 추출된 특징을 통해 다변량 시계열 데이터의 추세 경향을 파악할 수 있다.

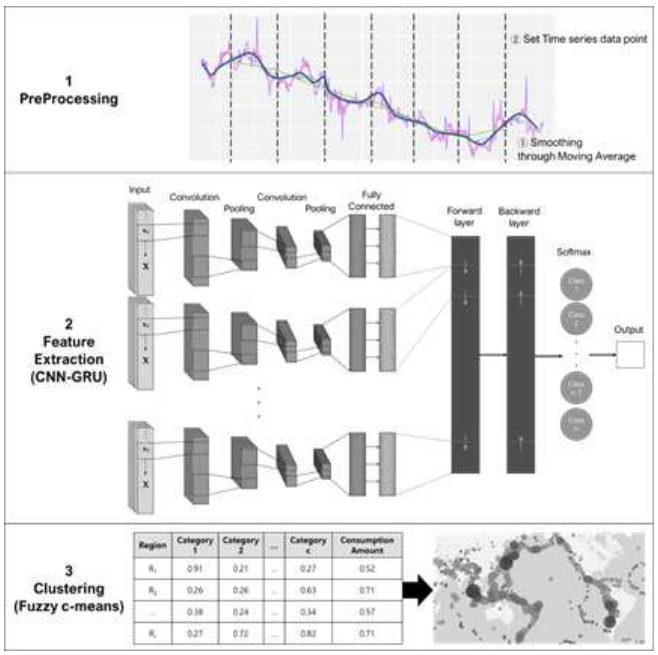
2. 다변량 시계열 데이터 기반의 군집화 방법

제안하는 다변량 시계열 데이터로부터의 특징 추출을 통한 군집화 방법의 과정은 (그림 1)과 같다. 이를 소비 분야에 적용하여 지역별 소비 업종에 따른 추세에 대해 특징을 추출함으로써 소비 금액과 추세가 유사한 지역들을 군집화하는 것이 목적이다. 업종별 소비 금액 뿐만 아니라 시간에 따른 추세를 시퀀스화하기 때문에 기존의 군집화 방법과 차이를 갖는다.

지역 군집화를 위해 먼저, 전처리를 통해 데이터의 노이즈를 제거한다. 이때 MA(Moving Average)를 사용하는데 MA는 데이터에 포함된 노이즈를 제거하여 추세를 보여주는 가장 기본적인 방법이다. 본 연구에서 활용하는 데이터는 일별 데이터이기 때문에 일주일 7일을 주기로 설정하였다. 또한, 업종별 소비 금액이 차이가 심하므로 0에서 1 사이의 값으로 값을 조정한 다음, 특징 추출을 위하여 일정 시간 간격마다 데이터 포인트를 설정하여 전체 시계열 데이터로부터 서브 시퀀스를 생성한다. 각 서브 시퀀스는 특징 추출을 위해 신경망의 입력 값으로 활용된다.

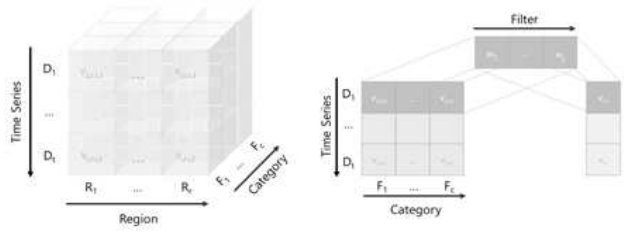
다음으로, 각 서브 시퀀스에 대하여 CNN과 GRU을 결합해 지역별 소비 추세에 대해 특징을 추출한다. CNN을 결합하여 사용하는 이유는 자동으로 특징을 추출하는 데 있어 좋은 성능을 보이기 때문이다. 특히, 데이터의 변형에도 강하기 때문에 불규칙적인 특성을 갖는 시계열 데이터의 특징을 추출하는데 이용될 수 있다[5].

특징 추출이 완료되면, 각 지역은 소비 업종별 소비 추세에 따른 특징 추출 결과를 가지게 되며, 업종별로 유사도를 계산해 유사한 지역끼리 Fuzzy C-means 기반의 군집화를 진행한다. 지역별로 여러 업종에 대한 소비 추세가 여러 군집의 성향을 가질 가능성이 있어 중첩된 군집 결과를 얻는다.



(그림 1) 다변량 시계열 데이터로부터
특징 추출을 통한 군집화 과정

3. CNN-GRU기반 다변량 시계열 데이터 특징 추출



(그림 2) 특징 추출을 위한 CNN-GRU 모델의
입력 데이터 형태

특징 추출을 위해 CNN을 먼저 진행하여 서브 시퀀스의 특성을 업종별로 추출한다. 이때, 입력 데이터는 (그림 2)와 같이 D X R X F 3차원 형태의 뉴런으로 구성된다. 여기서 D는 시계열의 길이를 의미하며, R은 지역, F는 소비 업종을 의미한다. 다음으로, convolutional layer를 통해 특징을 검출하여 feature map을 출력한다. 그리고 특성 데이터 크기를 줄이기 위해 Max Pooling layer를 사용한다. 이 과정을 여러 번 반복하며, 마지막으로 Flatten 과 Dense Layer로 데이터의 크기와 차원을 낮추고, 입출력 모두를 연결한다. 이로써 다시 각 서브 시퀀스마다 특징을 연결하여 시퀀스화 한다.

이 결과를 다시 종합하여 GRU을 통해 추세 패턴을 분류함으로써 각 서브 시퀀스마다 추세 흐름을 파악한다. 또한, Many to one 형태로, 최종적으로 출력계층에서 두 개의 은닉계층으로부터 값을 받아 해당 업종의 추세 흐름을 분류한다.

4. 결론

본 논문에서는 다변량 시계열 데이터를 갖는 관측지들에 대하여 군집화를 위해 CNN-GRU 기반의 다변량 시계열 데이터의 소비 추세 특징을 추출하는 방법을 제시하였다. 지역별로 업종에 따라 카드 사용된 데이터를 활용하였으며, 유사한 소비 추세를 보이는 지역을 군집화하는데 적용하였다. 다변량 시계열의 전체적인 흐름을 반영하여 패턴화했다는 점에서 기존 연구와 차별성을 갖는다. 본 논문에서 활용한 카드 소비데이터는 사회적인 영향에 따라 변동이 심하다는 특징을 갖는다. 그러나, 시간에 따라 트렌드의 변화를 고려할 만한 변수를 추가하지 않았다는 점에서 한계점이 있다. 향후 연구에서는 SNS나 뉴스 등의 데이터를 활용해 사회적인 영향을 고려하고 이에 따라 지역마다 미치는 영향 정도를 반영할 수 있을 것이다.

ACKNOWLEDGEMENT

이 논문은 2018년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2017R1A2B4008886).

참고문헌

[1] Fu, T. C. “A review on time series data mining” Engineering Applications of Artificial Intelligence 24(1), pp.164-181 (2011)

[2] Lin, T., Guo, T., & Aberer, K. “Hybrid neural networks for learning the trend in time series” In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (No. CONF), pp. 2273-2279 (2017)

[3] Kim, J., & Moon, N. “BiLSTM model based on multivariate time series data in multiple field for forecasting trading area” Journal of Ambient Intelligence and Humanized Computing, pp.1-10 (2019)

[4] Gudelek, M. U., Boluk, S. A., & Ozbayoglu, A. M. “A deep learning based stock trading model with 2-D CNN trend detection” In 2017 IEEE Symposium Series on Computational Intelligence (SSCI), pp.1-8 (2017)

[5] Bai, S., Kolter, J. Z., & Koltun, V. “An empirical evaluation of generic convolutional and recurrent networks for sequence modeling”, arXiv preprint arXiv:1803.01271 (2018)