Variational Autoencoder를 이용한 경제 데이터 증강

최형규 O1 · 김래현 [2] · 강재우 [2,3]
고려대학교 경영학과 [1]; 고려대학교 대학원 컴퓨터·전파통신공학과 [2];
고려대학교 대학원 바이오협동과정 [3]
{imhgchoi, raehyun, kangj}@korea.ac.kr

Economic Data Augmentation with Variational Autoencoder

HyeongKyu Choi O1 · Raehyun Kim [2] · Jaewoo Kang [2,3]
Business School, Korea University [1];
Department of Computer Science and Engineering, Korea University [2];
Interdisciplinary Graduate Program in Bioinformatics, Korea University [3]

Abstract

Data has become key to the Industry 4.0 progress. Considerable amount of data is required for many recently proposed machine learning models. However, in certain domains such as the economic field, it may be difficult to attain sufficient samples of data. With the motivation to alleviate such deficiency, we propose a data augmentation methodology based on the Variational Autoencoder. By applying the method on the currency exchange rate data, we empirically show that the training model's performance and stability is improved when augmented data is added to the train dataset.

1. Introduction

Data plays a crucial role in the Industry 4.0 progress. Rigorous data analysis is required in various fields of study, and a majority of machine learning models need considerable amount of data.

However, attaining sufficient amount of data can be difficult. Also, if the dataset lacks diversity, models trained on the data may overfit or converge to a local minimum. One solution to these problems is data augmentation. Data augmentation refers to the technique of adding artificially perturbed or synthesized data into the original dataset, thereby enhancing data volume as well as diversity. One of the domains where data augmentation is frequently adopted is Computer Vision. Methods such as cropping, rotating, or flipping is applied on the image data to improve performance of the learning model. In addition to such traditional methods, it has been empirically shown that augmenting image data with GAN(Generative Adversarial Network) improves model performance as well[1].

Collecting ample data can become especially difficult for certain fields of study. A representative of such field is the economic domain. In many cases, the population itself is constrained that it only renders limited number of samples. For instance, the source of currency exchange rate data with temporal and national constraints is unique; only one stream of time series exists for a certain nation and time span. Adding other nations' or dates' data will not resolve the problem, as it would only make the dataset non-i.i.d(independently, identically distributed).

Accordingly, we attempt to augment economic data. To this end, the Variational Autoencoder(VAE hereafter) is trained, which is a relatively novel approach in terms of economic data augmentation. Methodology is specified in section 3. Model analysis and verification on the practicality of the augmented data is demonstrated in section 4. Our conclusion and future research are to be discussed in section 5.

2. Related Works

Data Augmentation methods for economic data has consistently been the topic of interest. In a recent research, the Discrete Fourier Transformation technique was utilized to augment stock price data[2]. Here, the stock price time series was decomposed into its high and low frequency domain. After transforming the high frequency domain only, the two were synthesized again to generate an augmented data sample. In another study, the Self-Organizing Map was used to segment stock items into clusters, so that when training a model for one stock item, the other items clustered in the same group could be regarded as augmented data and used for training altogether[3]. Each research has proposed a method to augment economic data, and demonstrated that the methods improved their model's performance.

Data augmentation in domains other than economics has been frequently researched as well. In one study, a CVAE(conditional VAE)-based augmentation model contributed to the performance improvement of speech recognition models[4]. The method holds significance in that it exceled other traditional and GAN-based data augmentation methods.

3. Methodology

3.1 Data

In this research, we set currency exchange rate(CCY hereafter) time series data as the augmentation target. The currency exchange rate is an important indicator of relative purchasing power of each nation.

The CCY data from Yahoo Finance was downloaded using the Python opensource library 'yfinance'. Each time series is the foreign exchange rate represented in Korean Won. CCY data of nations with economic significance were selected, and the date spans from 2004 to 2019. Also, several nations with high portion of missing data were removed. The finally selected CCY's are listed in the appendix.

The dataset is a daily time series of CCY, and each time step consists of its OHLC(Open-High-Low-Close) values. For convenience, only the Close value is used for our experiments. In the preprocessing step, each time series is divided by its corresponding Low value of 2004.01.01 so as to normalize the scale across nations. Furthermore, data instances were sampled with a sliding window so that each sample has a sequence length of 100. The final dataset thus contains 3,871 samples per nation, aggregating to 46,452 samples in total. Data from year 2004 to 2017 were used for training, and the rest from 2018 to 2019 were reserved for model testing and analysis purposes.

3.2 Data Augmentation Model

We utilize the VAE for data augmentation. The VAE is a type of generative model that learns the distribution of the prior probability p(z|x), where x is the input data, and generates data by sampling from the distribution[5]. Generally, the VAE's training process is easy and intuitive, and it has edge in its ability to track the likelihood while training. Moreover, the probability density function value of the generated data can be computed with the trained mean vector μ and log-variance vector σ. As we aim to generate many usable data for each specific original data sample, we concluded that the use of a model as VAE would be ideal.

The loss function of the VAE is ELBO(Evidence Lower Bound). It becomes the lower bound of the log likelihood log(p(x)), which the model is trained to maximize. The mathematical expression of ELBO is as follows.

$$\begin{aligned} log(p(x)) &\geq \int log(p(x|z))q_\theta(z|x)dz - \int log\frac{q_\theta(z|x)}{p(z)}q_\theta(z|x)dz \\ &= E_{q_\theta}[log(p(x|z))] - KL(q_\theta(z|x)\,||\,p(z)) = ELBO(\theta) \end{aligned}$$

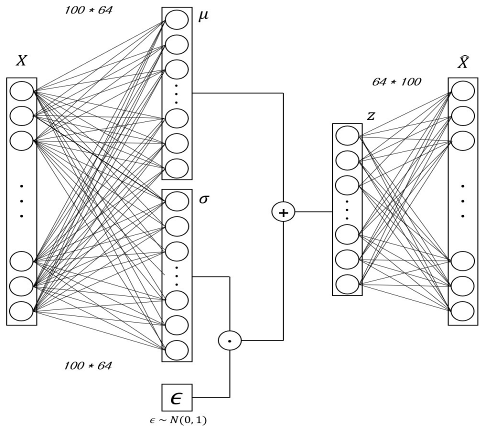
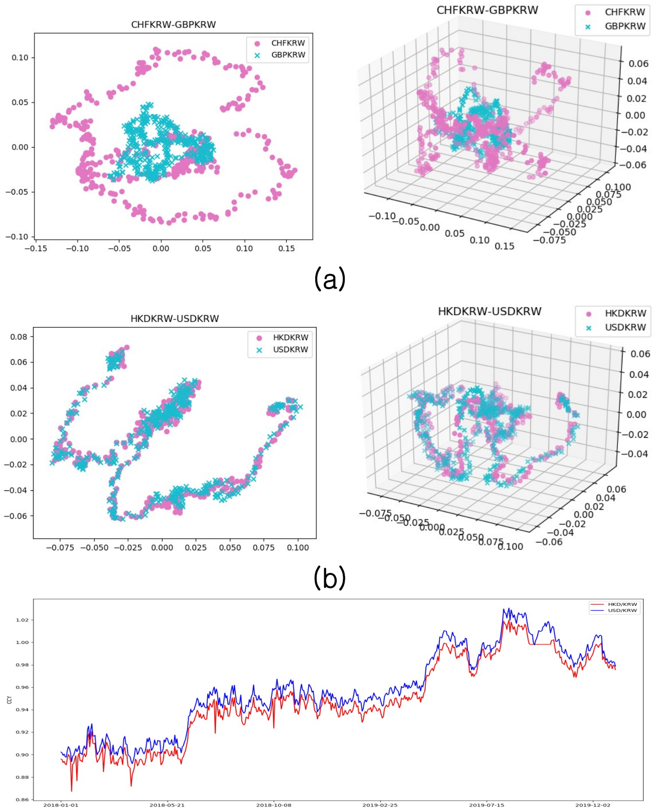


Figure 1. Model Architecture of VAE

Here, x denotes input data, z the sample from the latent space, and θ the model parameters to optimize.

The VAE architecture is demonstrated in Figure 1. We keep the structure simple in order to prevent the model from overfitting on the inherent noise in the CCY data. The model's encoder maps the input data X to 64-dimensional vectors μ and σ. Then, the decoder reconstructs the 100-dimensional augmented data vector from the z vector sampled from the distribution characterized by μ and σ.

The optimization algorithm we use is Adam[6]. We also clip the σ vector value during training, and pre-train the μ vector for stability. Pre-training is done in the way such that the ε value is fixed to 0 for several epochs before training the VAE as usual.



(a)

(b)

(c)

Figure 2. Data Latent Space Mapping Examples

4. Experiment Result and Analysis

4.1 Data Augmentation Model Analysis

The training result of a VAE may be assessed by how well each data is mapped to its characteristic latent space. In our study, we show that the train result was successful by demonstrating the principle components of the mean vectors μ for each nation's CCY data in 2D and 3D. In Figure 2 (a), the latent space mapping for Switzerland Franc and England Pound CCY data is displayed. We may observe that a characteristic latent space has been learned for each of the nation's data. On the other hand, we can see in Figure 2 (b) that a very similar latent space was learned for the US Dollar and Hong Kong Dollar. This seems to reflect the fact that the two currency values move in a highly correlated manner (Figure 2 (c)).

4.2 Augmented Data Practicality Verification

In order to test the practical usability of the augmented data, we train a simple Multi-Layer Perceptron using a very small amount of original data. For each original data sample, 100 data instances were generated as augmented data. The model is again trained using both augmented and original data, and its performance is compared with the former case.

The MLP model is trained to predict whether the CCY value for the next day would increase, decrease, or stabilize(|CCY change ratio| ≤ 0.01%), given historical CCY values. Generally, the portion of the three classes are not uniform. Thus, we choose the F1 score as our major performance indicator. Also, in order to render a more robust experiment result, the experiment is iterated 10 times, each using randomly sampled original data. The results are summarized in Chart 1. The use of augmented data improved the mean F1 score by 6.85%p. The p-value of the one-sided t-test for the two is 0.0006. Since this is way smaller than a significance level of 1%, we may safely assume that the F1 score difference is significant. However, the model using augmented data surpasses the random selection model by only a small percentage. Given the low complexity of the trained MLP model, such a result seems natural nonetheless.

Chart 1. MLP Model Training Result Comparison

(%)	Random		Original		Original+Augmented	
Iter.	Acc.	F1	Acc.	F1	Acc.	F1
1	34.14	30.14	44.43	20.51	46.99	**31.14**
2	34.12	30.38	48.42	21.84	47.02	**31.29**
3	32.69	29.29	48.29	22.85	46.95	**31.18**
4	32.87	29.04	48.20	21.68	46.82	**31.16**
5	32.71	29.00	49.01	**33.66**	46.93	31.11
6	32.67	28.95	48.24	21.70	46.86	**31.13**
7	32.62	28.86	48.31	21.72	47.10	**31.36**
8	33.81	29.72	48.46	21.76	47.02	**31.24**
9	32.54	29.06	47.87	**31.51**	46.56	30.89
10	32.69	29.07	48.11	26.31	47.35	**31.52**
Avg	33.06	29.35	47.93	24.35	46.96	**31.20**

When only original data were used, the model frequently converged to a low F1 score around 21. This was the case where the model fell into a local optimum, and only outputted an identical label for every input. However, when augmented data were introduced in training, no such problem occurred. This signifies that the use of augmented data enhanced stability of the learning process as well.

5. Conclusion and Future Work

Our major contribution is that we proposed a method to augment data in a domain where huge datasets are hard to attain, using a relatively novel approach of adopting the VAE. We showed the effectively trained result by visualizing the data latent spaces, and the practicality of the augmented data has been verified through a series of comparative experiments. Thus, we conclude that our data augmentation methodology is valid. However, the limitation of our research is that we could not propose an augmentation model with higher complexity. If trained properly, a deeper neural network may render better performance.

References

[1] J. Wang and L. Perez. "The Effectiveness of Data Augmentation in Image Classification using Deep Learning", *Stanford University research report.* 2017.

[2] X. Teng *et al.* "Enhancing Stock Price Trend Prediction via a Time-Sensitive Data Augmentation Method", *Hindawi.* 2020

[3] J. Zhang *et al.* "Data Augmentation Based Stock Trend Prediction Using Self-organising Map", *International Conference on Neural Information Processing*, 2017

[4] Z. Wu *et al.* "Data Augmentation using Variational Autoencoder for Embedding based Speaker Verification", *Interspeech.* 2019

[5] D. P. Kingma and M. Welling. "Auto-Encoding Variational Bayes", *International Conference on Learning Representations.* 2014

[6] D. Kingma and J. Ba. "A Method for Stochastic Optimization", *International Conference on Learning Representations*, 2015

Appendix

USD/KRW	Won/US dollar	CHF/KRW	Won/Swiss Franc
EUR/KRW	Won/European Euro	HKD/KRW	Won/HongKong Dollar
GBP/KRW	Won/England Pound	SGD/KRW	Won/Singapore Dollar
JPY/KRW	Won/Japan Yen	INR/KRW	Won/India Rupee
AUD/KRW	Won/Australia Dollar	THB/KRW	Won/Thailand Baht
CAD/KRW	Won/Canada Dollar	TWD/KRW	Won/Taiwan Dollar