A Deep Insight method with Morphological Analysis

Toyoaki Tomimoka
Graduate School of Engineering, Miyazaki University,
Japan, hm18026@student.miyazaki-u.ac.jp

Satoshi Ikeda*
Department of Computer Science and System Engineering, Miyazaki University,
Japan, bisu@cs.miyazaki-u.ac.jp

Makoto Sakamoto
Department of Computer Science and System Engineering, Miyazaki University,
Japan, saka2000@cc.miyazaki-u.ac.jp

Takao Ito
Graduate School of Engineering, Hiroshima University,
Japan, itotakao@hiroshima-u.ac.jp

**Corresponding Author*

Abstract

DeepInsight method, which interprets non-image data as image data, has garnered significant attention; however, a universal approach for this method has yet to be standardized. Our objective was to formulate a comprehensive method for DeepInsight, encompassing three key steps: strategic variable placement, effective feature extraction, and optimal model construction. Consequently, we successfully developed a model capable of predicting the Nikkei Stock Average with an accuracy of approximately 60%, marking a significant stride towards establishing the versatility of the DeepInsight method.

Keywords: DeepInsight method, t-SNE, Convolutional Neural Network, Morphological analysis.

1. Introduction

1.1 Research background

Machine learning faces challenges when dealing with high-dimensional data, known as the curse of dimensionality [1]. This issue arises because as the number of dimensions increases, the potential combinations of data grow exponentially, leading to computational challenges and inadequate learning results. To address this, two approaches can be employed:

1) Dimensionality Reduction: One strategy is to reduce dimensionality while preserving information. t-SNE [2] is a common technique that expresses data distances as conditional probabilities, keeping similar high dimensional data close in lower dimensions.

2) Feature Combination Restrictions: Another approach involves limiting combinations of features. Deep Learning [3] achieves this by dividing the problem into smaller regions, but it is most effective with data exhibiting strong pixel correlations, such as images. DeepInsight method [4], has emerged to extend deep learning to non-image data, although a standardized procedure for its application is yet to be established.

With the rise of social media, text mining utilizing platforms like SNS has gained prominence. Predicting stock prices has seen success by leveraging emotions expressed on SNS. Researchers are exploring deep learning analysis of news articles to build more accurate models, especially as sentiment analysis on platforms like

SNS has proven effective. The broader adoption of these techniques holds promise for improving predictions and insights across various domains.

1.2 Research purpose

The aim of this research is to transform the three stages involving proper variable placement, feature extraction, and model construction into a versatile DeepInsight method. By utilizing the proposed method, we aim to accurately predict tomorrow's stock price movements and validate the effectiveness of the approach.

2. Research method

2.1 Summary of t-SNE

The t-SNE is a dimension reduction algorithm designed to reduce high-dimensional data to two or three dimensions. The algorithm represents the proximity between points x_i and x_j on a high dimension as a joint probability distribution p_{ij}:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n} \; (i \neq j), \quad p_{ii} = 0.$$

$$\text{Here } \; p_{j|i} = \frac{\exp\left(\frac{-\left|\left|x_i - x_j\right|\right|^2}{2\sigma_i{}^2}\right)}{\sum_{k \neq i} \exp\left(\frac{-\left|\left|x_i - x_k\right|\right|^2}{2\sigma_i{}^2}\right)} \; \text{for } \; i \neq j.$$

And the proximity between points y_i and y_j in a low dimension as a joint probability distribution q_{ij}:

$$q_{ij} = \frac{\left(1 + \left|\left|y_i - y_j\right|\right|^2\right)^{-1}}{\sum_{k \neq i} \left(1 + \left|\left|y_i - y_j\right|\right|^2\right)^{-1}} \; (i \neq j).$$

The objective is to match the distance relationships between data points in high dimensions with those in low dimensions after dimension reduction. This is achieved by setting set $p_{i|j} = q_{i|j}$, and the Kullback-Leibler (KL) divergence is employed to measure the distance between $p_{i|j}$ and $q_{i|j}$ with the goal of minimizing the loss function:

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

Finally, stochastic gradient descent is used to minimize the loss function. The gradient is given by

$$\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ji} - q_{ji})(y_i - y_j)\left(1 + \left|\left|y_i - y_j\right|\right|^2\right)^{-1}$$

Using this gradient, is gradually updated using the renewal formula:

$$Y^{(t)} = Y^{(t-1)} + \eta \frac{\delta C}{\delta Y} + \alpha\left(Y^{(t-1)} - Y^{(t-2)}\right).$$

Here, t denotes the iteration, η is the learning rate, and α is the momentum term.

2.2 Learning Image Data with t-SNE using CNN

A Convolutional Neural Network (CNN) comprises fully connected, convolution, and pooling layers. In the fully connected layer, units are interconnected, with weights depicted in Fig. 2.1. Inputs and outputs to each unit are one-dimensional vectors, and the output unit's value is determined by multiplying input values by connection weights and adding bias.

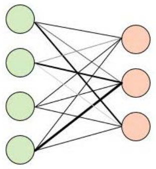
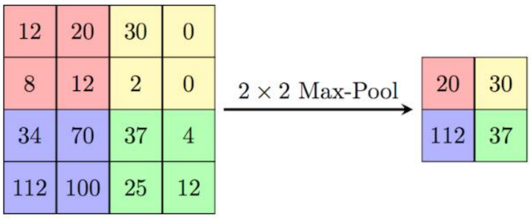


Fig. 2.1 Fully Connected layer weights

In the Convolutional layer, the outcomes of convolution operations for each filter are generated and serve as one unit in the subsequent layer. To derive the output, akin to the Fully Connected layer, compute the inner product for each filter, incorporate the bias, and apply the activation function. The Pooling layer condenses information to convert the input data into a more manageable format. Fig. 2.2 illustrates Max-pooling, wherein the maximum value is selected from color-coded small areas.

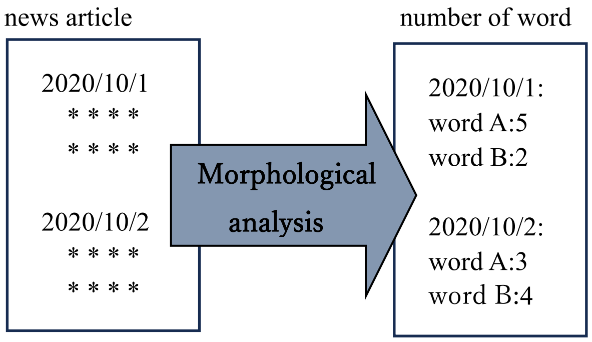


12	20	30	0
8	12	2	0
34	70	37	4
112	100	25	12

20	30
112	37

Fig. 2.2 Max-pooling

2.3 Morphological analysis



news article

2020/10/1
＊＊＊＊
＊＊＊＊

2020/10/2
＊＊＊＊
＊＊＊＊

Morphological analysis

number of word

2020/10/1:
word A:5
word B:2

2020/10/2:
word A:3
word B:4

Fig. 2.3 Example of Morphological analysis

Morphological analysis dissects sentences into morphemes, assigning parts of speech. Using a wordbook, the least-cost method selects optimal words based on part-of-speech connection and word appearance. This process is applied to news articles to count occurrences of words essential for stock price prediction, as illustrated in Fig. 2.3.

3. Experiment

3.1 Create input data

In the experiment, two datasets were used:

1) Daily stock prices of the 225 stocks constituting the Nikkei Stock Average.
2) Word count in news articles.

For 2), news articles related to stocks from the Kabutan website [5] during the period from 2020 to 2021 were used. Morphological analysis was applied to the news articles to count the occurrences of words considered effective in predicting stock prices (e.g., "weak yen"). A specialized dictionary (with 563 registered words) tailored for stock price prediction was employed for word counting. The data for stock prices and word counts in news articles were consolidated into a single CSV datasheet (Table 3.1).

Table3.1 CSV datasheet

	10/1	10/2	・・	10/10
company1	a1	b1	・・	j1
company2	a2	b2	・・	j2
・・	・・	・・	・・	・・
company225	a225	b225	・・	j225
word1	A1	B1	・・	J1
word2	A2	B2	・・	J2
・・	・・	・・	・・	・・

The first row of Table 3.1 represents the dates for which data was obtained. In this study, one sheet was created for a continuous 10-day period. Lowercase letters (a1, a2, etc.) represent daily stock prices for individual companies, while uppercase letters (A1, A2, etc.) represent the word count for each date. A total of 222 sheets of data were created for the entire data collection period. Using t-SNE, the data from each of the 225 sheets were transformed into image data composed of 783 points (225 for stock prices + 563 for words), as illustrated in Fig. 3.1. These images will be used as training and testing data for the Convolutional Neural Network (CNN).

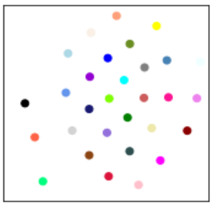


Fig. 3.1 An example of t-SNE transformation on CSV data

Each generated image is labeled with 0, 1, or 2. Here, 0 signifies that the next day's Nikkei Stock Average will remain unchanged (range from -100 yen to +100 yen), 1 indicates an increase (range from more than +100 yen), and 2 indicates a decrease (range from less than -100 yen).

3.2 Data processing

Firstly, split all the data into training and test sets. Allocate 80% of the total data to the training set and 20% to the test set to ensure unbiased data representation. Subsequently, normalize the input image data. Given that color values are typically expressed as {0, 1, ..., 255}, normalizing the data to the range [0, 1] can be achieved by dividing each value by 255. Finally, apply a one-hot vector transformation to the label data of the input images, converting them into a binary class matrix.

3.3 Evaluation method

Display the accuracy rates for each training and test dataset to assess the progress of learning and whether DeepInsight method is effective for prediction.

3.4 Experimental results

The number of training data is 175, and the number of test data is 45. Among the 45 test data, the percentage of '0' was 16%, the percentage of '1' was 53%, and the percentage of '2' was 31%, indicating that the percentage of '0' was relatively small, with '1' accounting for more than half.

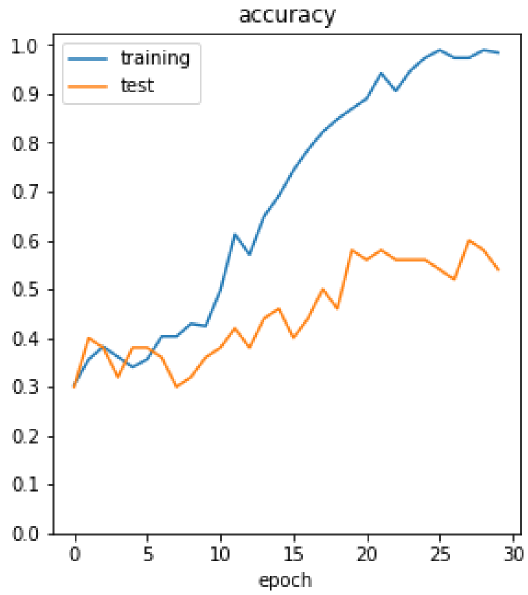


Fig. 3.2 Experimental result

In Fig. 3.2, where the horizontal axis represents the epoch (number of executions) and the vertical axis represents accuracy (correct answer rate), the accuracy rate of the training data reached nearly 100% at 25 epochs. The test data achieved its highest accuracy rate of 59.6% at 27 epochs, but further increasing the number of epochs did not lead to a higher accuracy rate. Additionally, at epoch 10, the accuracy rates of the training data and test data were similar, but thereafter, a significant difference emerged.

Examining the conditional probabilities in Table 3.2, out of 45 test data items, 53% (24 items) were predicted as '1', and among those, 58% (14 items) were correctly predicted as '1'. In contrast, 31% (14 items) were predicted as '2', and among those, 50% were correctly predicted as '2'.

Table3.2 Conditional probability

	0	1	2
0(16%)	14%(1)	71%(5)	14%(1)
1(53%)	16%(4)	58%(14)	25%(6)
2(31%)	14%(2)	36%(5)	50%(7)

4.Consideration

Our experiment aimed to turn the three steps of appropriate variable placement, feature extraction, and appropriate model construction into a general-purpose DeepInsight method. Appropriate variable placement was achieved by using t-SNE to convert effective prediction words obtained through morphological analysis of stock price statistical data and news articles into image data. Feature extraction utilized CNN, an effective method for image recognition in machine learning. Despite struggling with suppressing overfitting, we achieved a maximum accuracy of approximately 60%. Assuming the model was used for actual trading, this resulted in an approximately 11% increase in profits over one month, surpassing the average annual interest rate for stock investments (3-10%).

These results indicate that the "appropriate variable placement," "feature extraction," and "appropriate model construction" approach is effective in DeepInsight. Future challenges include increasing the types of label data and improving overall accuracy.

References

1.R.EBellman, “Dynamic Programming”, Princeton University Press,1957

2.G.E.Hinton, Laurens van Maaten, ”Visualizing Data using t-SNE”,Journal of Machine Learning Research 9 (2008) 2579-2605

3.G.E.Hinton, S.Osindero, Y.W.Teh, ”A Fast Learning Algorithm for Deep Belief Nets”, NeuralComputation 18(7), 2006,4pp.1527-1554

4.Tatsuhiko Tsunoda, ”DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architectu”, Scientific Repotrts. 10.1038/s41598-019-47765-6

5. “株探”

Authors Introduction

Mr. Toyoaki Tomioka

He is a student of the Mechanical and Information Systems Course, graduate school of Engineering, University of Miyazaki.

Prof. Satoshi Ikeda

He received PhD degree from Hiroshima University. He is an associate professor in the Faculty of Engineering, University of Miyazaki. His research interest includes graph theory, probabilistic algorithm, fractal geometry and measure theory.

Prof. Makoto Sakamoto

He is presently a professor in the Faculty of Engineering, University of Miyazaki. His first interests lay in hydrodynamics and time series analysis, especially the directional wave spectrum. He is a theoretical computer scientist, and his current main research interests are automata theory, languages and computation. He is also interested in digital geometry, digital image processing, computer vision, computer graphics, etc.

Prof. Takao Ito

He is Professor of Management of Technology (MOT) in Graduate School of Engineering at Hiroshima University. He is serving concurrently as Professor of Harbin Institute of Technology (Weihai) China. He has published numerous papers in refereed journals and proceedings, particularly in the area of management science, and computer science. He has published more than eight academic books including a book on Network Organizations and Information (Japanese Edition). His current research interests include automata theory, artificial intelligence, systems control, quantitative analysis of inter-firm relationships using graph theory, and engineering approach of organizational structures using complex systems theory.