Regular paper
Journal of the KIECS. pp. 547-552, vol. 14, no. 3, Jun. 30. 2019, t. 95, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2019.14.3.547

빅데이터의 정규화 전처리과정이 기계학습의 성능에 미치는 영향

조준모*

Effectiveness of Normalization Pre-Processing of Big Data to the Machine Learning Performance

Jun-Mo Jo*

요 약

최근, 빅데이터 분야에서는 빅 데이터의 양적 팽창이 주요 이슈로 떠오르고 있다. 더군다나 이러한 빅데이터는 기계학습의 입력값으로 사용되어지고 있으며 이들의 성능을 향상시키기 위해 정규화 전처리가 필요하다. 이러한 성능은 빅데이터 컬럼의 범위나 정규화 전처리 방식에 따라 크게 좌우된다. 본 논문에서는 다양한 종류의 정규화 전처리 방식과 빅데이터 컬럼의 범위를 조절하면서 서포트벡터머신(SVM)의 기계학습방식에 적용함으로써 더욱 효과적인 정규화 전처리 방식을 파악하고자 하였다. 이를 위하여 파이썬언어와 주피터 노트북 환경에서 기계학습을 수행하고 분석하였다.

ABSTRACT

Recently, the massive growth in the scale of data has been observed as a major issue in the Big Data. Furthermore, the Big Data should be preprocessed for normalization to get a high performance of the Machine learning since the Big Data is also an input of Machine Learning. The performance varies by many factors such as the scope of the columns in a Big Data or the methods of normalization preprocessing. In this paper, the various types of normalization preprocessing methods and the scopes of the Big Data columns will be applied to the SVM(: Support Vector Machine) as a Machine Learning method to get the efficient environment for the normalization preprocessing. The Machine Learning experiment has been programmed in Python and the Jupyter Notebook.

키워드

AI, Machine Learning, BigData, Preprocessing, Performance Evaluation
인공 지능, 기계 학습, 빅 데이터, 전처리, 성능 평가

Ⅰ. Introduction

The accuracy of convolutional neural networks (CNNs) has been continuously improving, but the computational cost of these networks also increases significantly. For example, the very deep VGG models, which have witnessed great success in a wide range of recognition tasks are substantially

* 교신저자 : 동명대학교 전자공학과
• 접 수 일 : 2019. 05. 17
• 수정완료일 : 2019. 05. 31
• 게재확정일 : 2019. 06. 15
• Received : May. 17, 2019, Revised : May. 31, 2019, Accepted : Jun. 15, 2019
• Corresponding Author : Jun-Mo Jo
Dept. Electronic Engineering, TongMyong University,
Email : jun@tu.ac.kr

slower than earlier models. Real world systems may suffer from the low speed of these networks. A loud service needs to process thousands of new requests per seconds. Portable devices such as phones and tablets may not afford slow models and semantic segmentation need to apply these models on many higher resolution images. It is very important to accelerate test-time performance of CNNs[1].

Machine learning has made possible the concept of self-driving cars, automatic intelligent web search, user based speech recognition software, personalized marketing and so on. So the world wide research is underway in many universities, companies and research facilities. The researches are related to the supervised and unsupervised learning methods as well as the field of the deep learning. Some methods classifies network of given patterns is a form of learning from observation. Such observation can define a new class or assign a new class to an existing class. This classification facilitates new theories and knowledge that is embedded in the input patterns. Learning behavior of the neural network model enhances the classification properties. The supervised and unsupervised and investigated its properties in the classification of post graduate students according to their performance during the admission period[2].

Introduction of cognitive reasoning into a conventional computer can solve problems by example mapping like pattern recognition, classification and forecasting. Artificial Neural Networks provides these types of models. These are essentially mathematical models describing a function. ANN is characterized by three types of parameters such as interconnection property as feed forward network and recurrent network. And the application function as a classification model. Finally, a learning rule such as supervised, unsupervised, and the reinforcement methods[3-4].

In this paper, the various known types of the normalization preprocessing methods and the importance of the scope or selection of the columns in the Big Data will be elaborated in the section II. Then the Machine Learning algorithm such as SVM and the impact of a various normalization methods for enhancing the performance of the specific Machine Learning will be explained in the section III. Then in the section IV, the result of the training will be compared and analyzed for the more efficient normalization methods. Finally, the conclusion is made in section V.

II. Normalization Methods

Normalization is a preprocessing method. First of all, the preprocessing is not only a technique that is used to convert the raw data into a clean data set, but also enhancing the performance of the Machine Learning. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis and the Machine Learning.

There are several well-known normalization methods such as Simple Feature Scaling, Min-Max, Z-score and so on. Firstly, the Simple Feature Scaling method is typically done via the following equation (1):

$$X_{norm} = \frac{X_{old}}{X_{\max}} \quad \cdots(1)$$

In the Min-Max method, the data is scaled to a fixed range, usually 0 to 1, the cost of having this bounded range is that it will end up with smaller standard deviations. A Min-Max scaling equation is as following equation (2):

$$X_{norm} = \frac{X_{old} - X_{\min}}{X_{\max} - X_{\min}} \quad \cdots(2)$$

The result of the Z-score normalization is that the features will be rescaled so that they will be the properties of a standard normal distribution with

$$\mu = 0 \quad \text{and} \quad \sigma = 1$$

where μ is the mean and σ is the standard deviation from the mean. The standard scores of the samples are calculated as following equation (3):

$$z = \frac{X - \mu}{\sigma} \quad \cdots(3)$$

The three normalization methods were applied to the SVM for the performance comparison and the result did not varies too much with this dataset, so the Min-Max normalization method is used in this paper.

III. Selecting Columns for Normalization

3.1 Support Vector Machine(SVM)

The Support Vector Machine(SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data as a supervised learning, the algorithm outputs an optimal hyperplane which categorizes new examples.

The related study on a deep learning, a decomposition method can be used as a learning method shown in Fig. 1. a filter is used for responding at a pixel of a layer approximately lies on a low-rank subspace. A resulting low-rank decomposition reduces time complexity. To find the approximate low-rank subspace, it minimizes the reconstruction error of the responses.

Machine Learning methods requires the fine tuning of the parameters and also feasible number of the data set. Therefore, choosing the best performance of the learning algorithm is important in the real world. And also for a particular data set does not guarantee the precision and accuracy for another set of data whose attributes are logically different from the other[4].

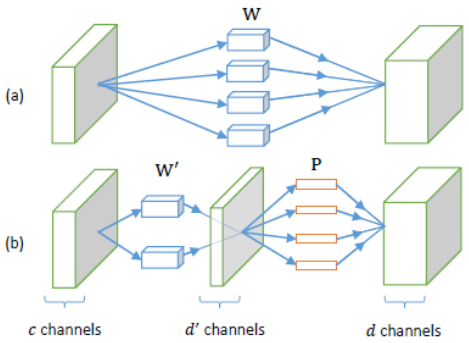

Fig. 1 Illustration of the decomposition. (a) An original layer with complexity O(dk2c). (b) An approximated layer with complexity reduced to O(d0k2c) + O(dd0) [1]

3.2 Selection of Columns in Big Data

A series of a distribution of Certain data could affects the performance of a machine learning. Most of the Big data as an input of a machine learning is a table consists of rows and columns. Mostly, the data in a column is in certain range and it is commonly different from others. In order to normalize the column data, we have to figure out the patters or scopes of the data. Therefore, the selection of the column(s) is very important.

The machine learning is to distinguish classes of the input data in order to predict an accurate result. The related study, for instance, the presence of full knowledge of the underlying probabilities, the Bayes decision theory gives optimal error rates[5-7].

The best performance is decided by not only the recognition ratio but also the time of the simulation. However, the key question when dealing with ML classification is not whether a learning algorithm is superior to others, but under which conditions a particular method can significantly outperform others on a given application problem. Meta learning is moving in this direction, trying to find functions that map data sets to algorithm performance. After a better understanding of the strengths and limitations of

each method, the possibility of integrating two or more algorithms together to solve a problem should be investigated. The objective is to utilize the strengths of one method to complement the weaknesses of another. If we are only interested in the best possible classification accuracy, it might be difficult or impossible to find a single classifier that performs as well as a good ensemble of classifiers[4-5].

Unlike the unsupervised learning, the supervised learning is a method by which you can use labeled training data to train a function that can be generalized for new examples. The training involves a critic that can indicate when the function is correct or not. There are various kinds of the methods are exist, such as the decision tree classifier, KNeighbors classifier, and the support vector machine(SVM) and so on[8-10].

IV. Result and Analysis

To train and to predict with the normalized data, the fit(), predict(), and closest() methods supported in Scikit-Learn are used. The Table 1 shows the utilities used in the program.

Table 1. Scikit-learn utilities used in training

```
from sklearn import datasets
from sklearn.datasets import load_wine
from sklearn.model_selection import
    train_test_split
from sklearn.metrics import accuracy_score
```

Four columns(0, 3, 9, 12) in the Wine dataset are selected to be normalized after calculating the standard deviation of each columns in Table 2. The columns are selected by the standard deviation value for the training in SVM. The columns are selected by the level of the standard deviation. The original raw data of the columns in the Wine data are shown in Fig. 2 below.

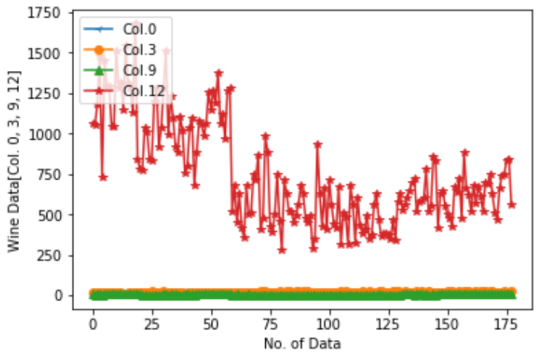

Fig. 2 Result of the normalization(0, 3, 9, 12)

The data in the column 12(Col. 12) are exceedingly high compare to the others. However, after the normalization process, the Col. 12 is the lowest among all. In this case, the Min-Max normalization method is used for the result as shown in Fig. 3. The result of the graph is programmed in Python.

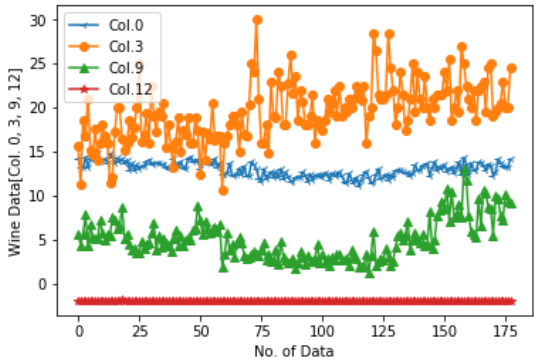

Fig. 3 Result of the normalization(0, 3, 9, 12)

We can see that the Col. 12 has shrank the most. Finally, the standard deviations and the training accuracies of the selected columns are shown in Table 2.

Table 2. Std. Dev. and accuracy result of columns

n^{th} Column	Standard Deviation	Prediction Accuracy(%)
0	0.8	70
3	3.3	78
9	2.3	74
12	314	94

According to the Table 2 and the Fig. 4, there is a relation between standard deviation of columns and the prediction accuracy.

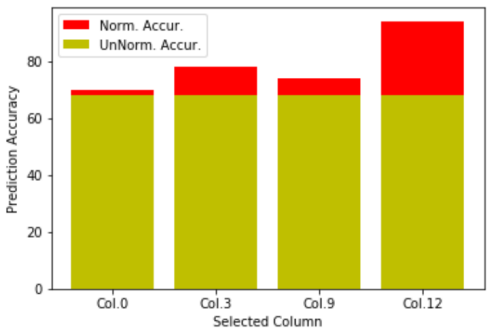

Fig. 4 Enhancement by the normalization

The higher the standard deviation, the higher the prediction accuracy. If all the columns are to be preprocessed, the result should be the best of all, however it takes great amount of expanse with the Big data then we need to select the most efficient column(s) what we deal with.

V. Conclusion

A series of a distribution of Certain data could affects the performance of a machine learning. The standard deviation could be a clue of an efficient performance. So I have calculate all the columns of the dataset to select a significant column to normalize. The four columns are selected to be normalized for comparing the prediction accuracy. The result showed that the higher the standard deviation, the higher the prediction accuracy. The normalization of the column 12 in Wine dataset affects the most of all since it has the highest standard deviation. The normalization for the deep learning using the Tensorflow will be the next study.

Reference

[1] X. Zhang, J. Zou, K. He, and J. Sun," Accelerating very deep convolutional networks for classification and detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 38, 2015, pp. 1943-1955.

[2] R. Sathya, and A. Annamma, "Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification," *IJARAI*, vol. 2, no. 2, 2013, pp.34-38.

[3] R. Sathya and A. Abraham, "Unsupervised Control Paradigm for Performance Evaluation," *International Journal of Computer Application*, vol. 44, no. 20, 2012, pp. 27-31.

[4] X. C. Yin, X. Yin, K. Huang, and H. W. Hao, "Robust text detection in natural scene images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, 2014, pp. 970-983.

[5] N. Kim and Y. Bae, "Status Diagnosis of Pump and Motor Applying K-Nearest Neighbors," *J. of the Korea Institute of Electronic Communication Science*, vol. 13, no. 6, 2018, pp. 1249-1255.

[6] J. M. Keller, M. R. Gray, and J. A. Givens, "A Fuzzy K-Nearest Neighbor Algorithm," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 15, no. 4, 1985, pp. 581-585.

[7] S. Bang, "Implementation of Image based Fire Detection System Using Convolution Neural Network," *J. of the Korea Institute of Electronic Communication Science*, vol. 12, no. 2, 2017, pp. 331-336.

[8] Y. Kim, S. Park, and D. Kim, "Research on Robust Face Recognition against Lighting

Variation using CNN," *J. of the Korea Institute of Electronic Communication Science*, vol. 12, no. 2, 2017, pp. 325-330.

[9] C. Jung, R. Jang, D. Nyang, and K. Lee "A Study of User Behavior Recognition-Based PIN Entry Using Machine Learning Technique," *Korea Information Processing Society review, computer and communication systems*, vol. 7, no. 5, 2018, pp. 127-136.

[10] G. Lee, H. Ha, H. Hong, and H. Kim "Exploratory Research on Automating the Analysis of Scientific Argumentation Using Machine Learning," *J. of the Korean Association for Science Education*, vol. 38, no. 2, 2018, pp. 219-234.

저자 소개

조준모(Jun-Mo Jo)

1991년 아이오아주립대학교 컴퓨터과학과 졸업 (공학사)
1995년 경북대학교 대학원 컴퓨터공학과 졸업(공학석사)
2004년 경북대학교 대학원 컴퓨터공학과 졸업(공학박사)
1998년~현재 동명대학교 전자공학과 교수
※ 관심분야 : 이동통신, 뇌파통신, 인공지능