Regular paper

Journal of the KIECS. pp. 701-706, vol. 14, no. 4, Aug. 31. 2019, t. 96, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2019.14.4.701

필기숫자 데이터에 대한 텐서플로우와 사이킷런의 인공지능 지도학습 방식의 성능비교 분석

조준모*

Performance Comparison Analysis of AI Supervised Learning Methods of Tensorflow and Scikit-Learn in the Writing Digit Data

Jun-Mo Jo*

요 약

최근에는 인공지능의 도래로 인하여 수많은 산업과 일반적인 응용에 적용됨으로써 우리의 생활에 큰 영향을 발휘하고 있다. 이러한 분야에 다양한 기계학습의 방식들이 제공되고 있다. 기계학습의 한 종류인 지도학습은 학습의 과정 중에 특징값과 목표값을 입력으로 가진다. 지도학습에도 다양한 종류가 있으며 이들의 성능은 입력데이터인 빅데이터의 특성과 상태에 좌우된다. 따라서, 본 논문에서는 특정한 빅 데이터 세트에 대한 다수의 지도학습 방식들의 성능을 비교하기 위해 텐서플로우(Tensorflow)와 사이킷런(Scikit-Learn)에서 제공하는 대표적인 지도학습의 방식들을 이용하여 파이썬언어와 주피터 노트북 환경에서 시뮬레이션하고 분석하였다.

ABSTRACT

The advent of the AI(: Artificial Intelligence) has applied to many industrial and general applications have havingact on our lives these days. Various types of machine learning methods are supported in this field. The supervised learning method of the machine learning has features and targets as an input in the learning process. There are many supervised learning methods as well and their performance varies depends on the characteristics and states of the big data type as an input data. Therefore, in this paper, in order to compare the performance of the various supervised learning method with a specific big data set, the supervised learning methods supported in the Tensorflow and the Sckit-Learn are simulated and analyzed in the Jupyter Notebook environment with python.

키워드

AI, Machine Learning, Supervised Learning, Tensorflow, Performance Evaluation
인공 지능, 기계 학습, 지도 학습, 텐서플로, 성능 평가

Ⅰ. Introduction

Machine learning is a modern science to work without being explicitly programmed by human. Recently, in Korea, the field of the smart factory using artificial intelligence is emerging to make the efficiency of producing merchandise with fault tolerant system since the decrease of the

* 교신저자 : 동명대학교 전자공학과
• 접 수 일 : 2019. 05. 09
• 수정완료일 : 2019. 06. 27
• 게재확정일 : 2019. 08. 15
• Received : May. 09, 2019, Revised : Jun. 27, 2019, Accepted : Aug. 15, 2019
• Corresponding Author : Jun-Mo Jo
Dept. Electronic Engineering, TongMyong University,
Email : jun@tu.ac.kr

population[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 supervised learning methods will be introduced and described in section II. In the section III, the data set to simulate in the supervised learning methods introduced in section II will be explained and prepared. Then, in the section IV, the simulation of the result will be explained and verified for the best performance among the learning methods. Finally, the conclusion is made in section V.

II. Supervised Learning Methods

Classification is an important in the machine language field to distinguish classes of the input data in order to predict accurate result. In the presence of full knowledge of the underlying probabilities, Bayes decision theory gives optimal error rates[3-4]

Logistic regression is targeted at classification problems with binary or categorical response. Let be the feature vector, corresponding to the base similarity measures.

It employs soft computation to update weights in a planar field of neurons. In other words, a neuron that wins the competition tends to excite the neurons called cooperating neurons in its immediate neighborhood, the h_{ji}, more than those far away from it. The lateral distance between the winning neuron and the excited neuron tends to increase with the neighborhood decaying to zero. In addition, the size of topological neighborhood which is centered around winning neuron shrinks with time. Weight updates the w take place in a cluster of neurons in the neighborhood of the winner as the following equation[3-5].

$$W_{old} = W + \eta h(X - W_{old})$$

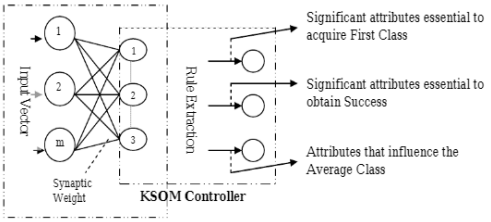

Fig. 1 Architecture of KSOM controller[3]

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. The best performance is decided by not only the recognition ratio but also the time of the simulation. 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. 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[6-8].

III. Training of Hand Written Digit Dataset

3.1 Hand Written Digit Dataset

Handwritten character recognition is one of the important issues in pattern recognition area. The digit recognition field includes in postal mail sorting, bank check processing, form data entry and so on. The heart of the problem lies within the ability to develop an efficient algorithm that can recognize hand written digits and which is submitted by users by the way of a scanner, tablet, and other digital devices[9-11].

The MNIST(: Modified National Institute of Standards and Technology) database is a large database of handwritten digits that is commonly used for training various image processing systems shown as Fig. 2.

Fig. 2 Example of the handwritten digit data

The database is also widely used for training and testing in the field of machine learning. It was created by "re-mixing" the samples from NIST's original datasets. We can download the dataset and the data is shown as the Table 1.

Table 1. Contents of the dataset

Name of the Files	Contents	No. of Data
train-images-idx3-ubyte.gz	Learning Image	60,000개
train-labels-idx1-ubyte.gz	Learning Label	
t10k-images-idx3-ubyte.gz	Test Image	10,000개
t10k-labels-idx1-ubyte.gz	Test Label	

3.2 Machine Learning with the Dataset

There are several machine learning algorithms supported from the Tensorflow and the Scikit-Learn. The following selected methods will be used for training and comparing their performance with the hand written dataset explained earlier - the DecisionTree Classifier,

KNeighbors classifier, the Euclidean Distance Classifier, and the Support Vector Machine.

The Decision Tree Classifier is a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

The KNeighbors classifier is provides functionality for unsupervised and supervised neighbors based learning methods. Unsupervised nearest neighbors is the foundation of many other learning methods. Supervised neighbors-based learning comes in two flavors which are the classification for data with discrete labels, and the regression for data with continuous labels.

The Euclidean Distance Classifier uses the "ordinary" straight line distance between two points in Euclidean space to classify the training data and categorizes the classes of the input data.

The Support Vector Machine 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.

These training methods will be applied to the hand written dataset from MNIST.

IV. Training Result and Analysis

Each training methods of the supervised learning shown previous chapter are applied to the hand written dataset for the training and comparing the performance of each methods.

For instance, in order to do the training of the Euclidean Distance Classifier, a class for a fit(), predict(), and closest() methods are programmed and tested with the dataset. The performance of many training methods with the different quantity of the data could be different. So, there will be three types of experiment such as very small data, small data, and very big data. These three categories will be applied to each of the training methods explained above. For instance, the Table 1 shows the Euclidean distance method programmed in python. There are three methods for training the input data.

Table 1. Euclidean distance method

```
from scipy.spatial import distance
def euc(a,b) :
    return distance.euclidean(a, b)
class eucKNN():
    def fit(self, X_train, y_train):
    def predict(self, X_test):
    def closest(self, row):
```

The result of the training is shown in Table 2. The 'Tree' is the DecisionTree Classifier, the 'Kneighbor' is the KNeighbors classifier, 'Euc' is the Euclidean Distance Classifier, and the 'SVM' is the Support Vector Machine.

Table 2. Result of the training methods

	Very Small Data	Small Data	Big Data
Tree	0.41	0.62	0.84
Kneighbor	0.57	0.83	0.95
Euc	0.68	0.86	0.94
SVM	0.72	0.88	0.97

The training result in graph is shown as Fig. 3 below. The graph is programmed in Python, and the x-axis shows the list of the training methods.

The quantity of the input data to train are as follows. The 'Very Small Data' is selected for 100 samples, the 'Small Data' used for 1000, finally, the 'Big Data' used 40,000 numbers of the samples.

For the y-axis shows the performance of the training result with testing data. The testing data is not used for the training data. The value 1.0

means the prediction ratio of 100%.

As I have expected that the result shows the bigger the training data, the better the performance. However, the performance varies among each training methods with the quantity of the training data. The Tree Classifier showed the worst performance among all, but other methods seemed to show the comparatively good performance.

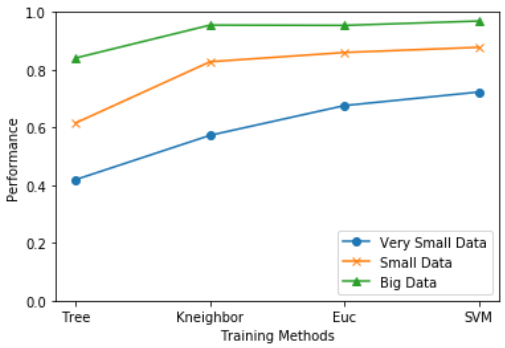

Fig. 3 Comparing training methods with different quantity of input data

The Kneighbor method shows little bit of low performance with small train data. However, if there are enough input data to be trained, the performance of the Kneighbor method slightly exceeds the Euc method.

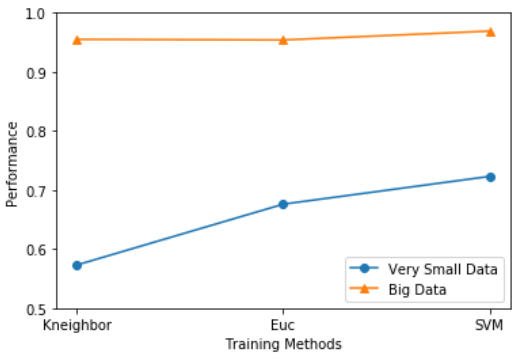

Fig. 4 Comparing kneighbor with other methods

On the contrary, the SVM methods shows the best performance whether the quantity of input data is big or small for the hand written dataset.

V. Conclusion

The various supervised machine learning methods supported by the Tensorflow and Scikit-learn, are used for the training with the hand written digit dataset to predict the test data. Then comparing the performance of 4 types of training models with different quantity of the input data. The Tree classifier showed the worst performance, and the other three models showed better. Especially, the SVM model showed the best performance both on the small and big quantity of the train data. And the Kneighbor model slightly exceeds the Euc model with the big train data supported.

The hand written digit numbers used in this experiment is perfectly set by the user, however, the experiment with other natural pictures taken outside could be different from the work.

Acknowledgement

This Research was supported by the Tongmyong University Research Grants 2019 (2019F001)

Reference

[1] Y. Jung and Y. Bae, "Analysis of Fault Diagnosis for Current and Vibration Signals in Pumps and Motors using a Reconstructed Phase Portrait," *Int. J. of Fuzzy Logic and Intelligent Systems*, vol. 15, no. 3, 2015, pp. 166-171.

[2] R. Sathya, and A. Annamma, "Comparison of Supervised and Unsupervised Learning

Algorithms for Pattern Classification," *International Journal of Advanced Research in Artificial Intelligence,* 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] C. Neocleous, and C. Schizas, "*Artificial Neural Network Learning: A Comparative, Methods and Applications of Artificial Intelligence.*," Hellenic Conference on Artificial Intelligence SETN, Springer, 2002.

[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.

[11] S. Shamim, M. Miah, A. Sarker, and M. Rana, "Handwritten Digit Recognition Machine Algorithms," *Global Journal of Computer Science and Technology*, vol. 18, 2018, pp. 17-23.

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