Journal of Digital Convergence
Vol. 16. No. 6, pp. 151-161, 2018

ISSN 1738-1916
https://doi.org/10.14400/JDC.2018.16.6.151

Predicting Employment Earning using Deep Convolutional Neural Networks

Adyan Marendra Ramadhani, Na-Rang Kim*, Hyung-Rim Choi
Department of Management Information Systems, Dong-A University

딥 컨볼루션 신경망을 이용한 고용 소득 예측

마렌드라, 김나랑*, 최형림
동아대학교 경영정보학과

Abstract Income is a vital aspect of economic life. Knowing what their income will help people create budgets that allow them to pay for their living expenses. Income data is used by banks, stores, and service companies for marketing purposes and for retaining loyal customers; it is a crucial demographic element used at a wide variety of customer touch points. Therefore, it is essential to be able to make income predictions for existing and potential customers. This paper aims to predict employment earnings or income based on history, and uses machine learning techniques such as SVMs (Support Vector Machines), Gaussian, decision tree and DCNNs (Deep Convolutional Neural Networks) for predicting employment earnings. The results show that the DCNN method provides optimum results with 88% compared to other machine learning techniques used in this paper. Improvement of the data length such PCA has the potential to provide more optimum result.

Key Words : Income prediction, DCNN, SVM, Gaussian, Decision Tree

요 약 소득은 경제생활에서 중요하다. 소득을 예측할 수 있으면, 사람들은 음식, 집세와 같은 생활비를 지불 할 수 있는 예산을 세울 수 있을 뿐 아니라, 다른 재화 또는 비상사태를 위한 돈을 별도로 저축 할 수 있다. 또한 소득수준은 은행, 상점 및 서비스 회사에서 마케팅 목적 및 충성도가 높은 고객을 유치하는 데 활용 된다. 이는 소득이 다양한 고객 접점에서 사용되는 중요한 인구 통계 요소이기 때문이다. 따라서 기존 고객 및 잠재 고객에 대한 수입 예측이 필요하다. 이 연구에서는 소득을 예측하기 위해 SVM (Support Vector Machines), Gaussian, 의사 결정 트리, DCNN (Deep Convolutional Neural Networks)과 같은 기계 학습 기법을 사용하였다. 분석 결과 DCNN 방법이 본 연구에서 사용 된 다른 기계 학습 기법에 비해 최적의 결과(88%)를 제공하는 것으로 나타났다. 향후 PCA 같이 데이터 크기를 향상 시킨다면 더 좋은 연구 결과를 제시할 수 있을 것이다.

주제어 : 소득 예측, DCNN, SVM, Gaussian, Decision Tree

1. Introduction

Income or earning is a sum of money collected from doing work or received through investments [1]. Income also the sum of all the wages, salaries, profits, interests payments, rents, and other forms of earnings received in a given period of time[2]. Income is an essential aspect of daily economic life, and is crucial not

*This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2015S1A3A2046781)
*Corresponding Author : Na-Rang Kim (whitecoral@hanmail.net)
Received May 7, 2018 Revised June 3, 2018
Accepted June 20, 2018 Published June 28, 2018

only for individuals but for businesses, governments and other organizations as well. Knowing income enables people, businesses and governments to create budgets for daily living expenses, and income is a crucial demographic element used at a wide variety of customer touch points. Income data is also used by banks and other businesses for marketing purposes, and for retaining loyal customers. Employment income is different than business income, employment income is the income that gain from the employment services and business income is the income that gain from the business activity. Today, increased competition is leading companies to search for ways to innovate, retain customers and survive by offering the right products at the right time and maintaining good customer relationships. This means they need fast and accurate decision-making processes to retain loyal customers. However, they must also keep per-unit processing costs low, and provide quick turnaround times for customers. Therefore, it is essential for businesses to be able to make accurate income predictions for both existing and potential customers.

Machine learning and big data are among the current technologies that offer an effective and efficient method for predicting and analyzing data. Machine learning in particular is developing rapidly as a way to optimize interpretive and predictive processing and results. Various machine learning methods and algorithms have been employed over the years to improve the performance of data classification and prediction. A number of researchers have investigated income prediction and classification problems: Azamat [3] developed benchmarking regression algorithms for income prediction modeling, studied the performance of various state-of-the-art regression algorithms (e.g., ordinary least squares regression, beta regression, robust regression, ridge regression, MARS (multivariate adaptive regression splines), ANN (artificial neural network), LSSVM (least squares support vector machine) and CART (classification and regression trees), as well as two-stage models that combine multiple techniques), and applied these to five real-life income datasets for prediction. Lazar [4] used the combination of SVM and PCA (principal component analysis) to test income predictions using CPS income data (income data consisted of text, numbers and a mixture of both, and could be categorized as text). Recently, another method of machine learning has become popular: deep learning. This method adapts the neural system using deep architecture such as DCNN and CNN. Neural network can be used in several sector such as : weather forecasting[17], Drifter Movement prediction[18] early cancer prediction[19], face recognition[20] and many other prediction. Neural network is one of the potential method in machine learning and it can be modified such as using deep learning architecture. Several deep learning studies have looked at DCNN and CNN: Alexis et al. [5] use DCNN for sentence classification and sentiment analysis, while Li et al. [6] use CNN for text classification based on Chinese characters－their research shows those methods optimize sentence classification and sentiment analysis whether the dataset language was English or another language.

Based on recent work and current technology, this research uses a DCNN and other machine learning techniques (SVM, Gaussian naïve Bayes, and decision tree) for comparing the prediction and analysis of employment income.

This paper consists of the following sections: Section I provides an introduction to the research; Section II introduces related works and other research in this field; the study model is presented in Section III; and the results and conclusion are covered in Section IV.

2. Related Works

The aim of this paper is to contribute to and optimize the prediction of employment earnings or income based on historical data.

Some previous research has already examined this

topic: Azamat [3] developed benchmarking regression algorithms for income prediction modeling, studying the performance of various state-of-the-art regression algorithms (e.g., ordinary least squares regression, beta regression, robust regression, ridge regression, MARS (multivariate adaptive regression splines), ANN (artificial neural network), LSSVM (least squares support vector machine) and CART (classification and regression trees), as well as two-stage models that combine multiple techniques) and applying these to five real-life income datasets for prediction. Lazar [4] used the combination of SVM and PCA (principal component analysis) for testing income prediction with CPS income data.

Bjelland, J. et all [7] used deep learning method to predict individual income data based on the mobile phone data.

Based on all research, the current research use the machine learning method to predict the income earning using different dataset and data structure.

While the current research didn't use the deep learning architecture and use the non-sentences data structure, this research use a DCNN, a convolutional neural network based on deep learning architecture and using the sentences data structure.

3. Research Model and Methodology

The proposed model for this research is experimental, using a quantitative approach. The configuration of the proposed method is illustrated in Fig. 1 The proposed experiment will use a DCNN to predict and classify employment income.

3.1 Dataset

The income dataset was extracted from the Census Bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). This dataset was selected and distributed for normal distribution.

Table 1. Dataset Detail

No.	Dataset	Description
1	Train	35000
2	Test	10500

3.2 Deep Convolutional Neural Network

Convolutional neural networks (CNNs) are a particular kind of neural architecture specially designed to handle image data. Since their introduction by LeCun et al. (1989), CNNs have demonstrated excellent

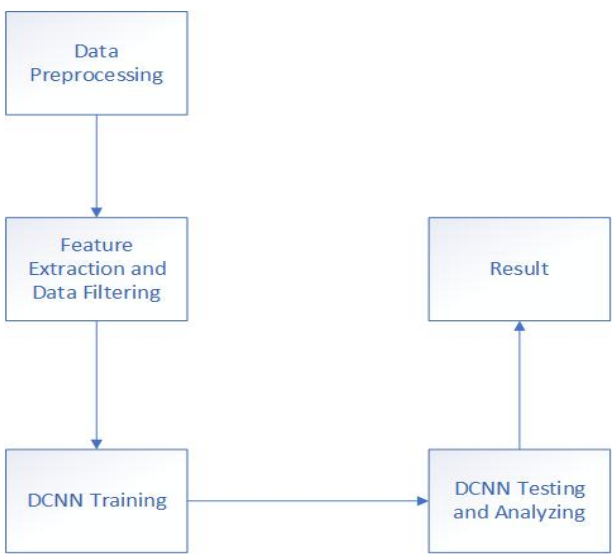

Fig. 1. Research model methodology.

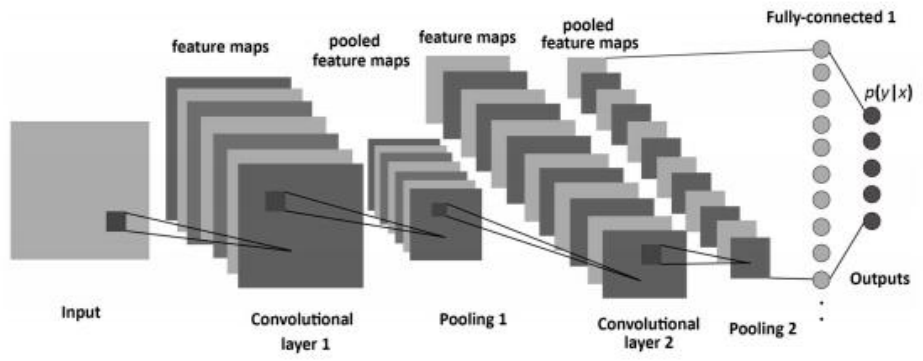

Fig. 2. Structure of a CNN [10]

performance at tasks such as handwritten digit classification and face detection [8]. ConvNets have been successful in identifying faces, objects and traffic signs apart from powering vision in robots and self driving cars[9].The efficacy of convolutional nets (ConvNets or CNNs) in image recognition is one of the main reasons why the world has woken up to the efficacy of deep learning. They are powering major advances in computer vision (CV)[10].

CNN models have a standard structure consisting of alternating convolutional layers and pooling layers (often, each pooling layer is placed after a convolutional layer). The last layers are a small number of fully connected layers, and the final layer is a softmax classifier, as shown in Fig. 2 CNNs are usually trained by backpropagation via stochastic gradient descent (SGD) to find weights and biases that minimize specific loss function to map the arbitrary inputs to the targeted outputs as carefully as possible [11].

The proposed CNN method to be used in this experiment is the DCNN (deep convolutional neural network), inspired by VGG and VDCNN architecture. The proposed method is illustrated in Fig. 3. The input of the network was word vectorization, as the income data was based on text and numbers. It contains 2048 fixed, embedded characters. The convolutional networks in the proposed method include three convolutional layers and one fully connected layer. Using the VGG and VDCNN methods[4], several rules are applied on the network, including the following:

- For the same output temporal resolution, the layers have the same number of feature maps.
- When the temporal resolution is halved, the number of feature maps is doubled.

This helps reduce network usage of memory and memory tracking. For the classification tasks, the temporal resolution of the convolutional network's output is first down-sampled to a fixed dimension using k-max pooling. The network extracts the k most essential features, independently of the position they appear in the sentence. The 512 × k resulting functions are transformed into a single vector, which is the input to a three-layer, fully connected classifier with ReLU hidden units and softmax outputs [12].

3.3 SVM (Support Vector Machine)

The SVM is a state-of-the-art classification method introduced in 1992 by Boser, Guyon, and Vapnik [13]. The support vector machine (SVM) machine learning method is used for resolving binary discrimination and prediction problems. For binary classification, the basic idea of SVM is to find a hyperplane that separates positive and negative training observations and maximizes the margin between these observations and the hyperplane [14].

The proposed SVM to be used in this experiment is the SVM one versus all (OVA) method for the

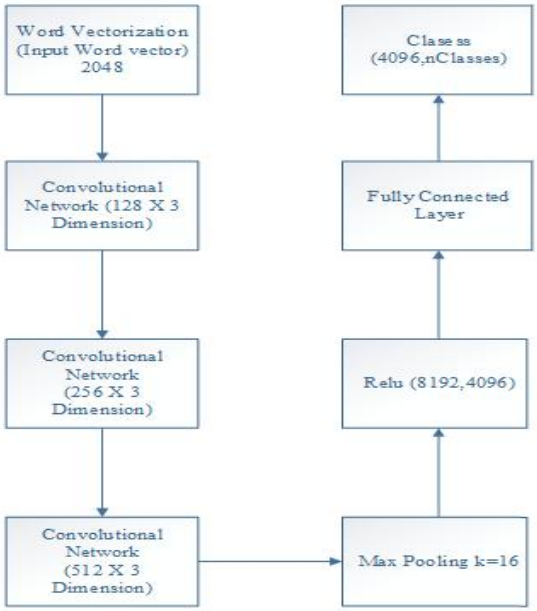

Fig. 3. Proposed DCNN in experiment.

multiclass problem. The binary SVM is applied several times using the technique of one versus all (OVA) for each class.

The SVM decision function is expressed as shown in Equation 1:

$$f(x) = \sum_{i=1}^{n} y_i \alpha_i K(X_{i,x}) + b \quad (1)$$

where $K(X_{i,x}) + b$ is the kernel function and y_i α_i is the label function. The output class is taken from the classifier with the largest positive answer.

3.4 Gaussian Naïve Bayes

Naïve Bayes is a classifier that uses Bayes' Theorem. It predicts membership probabilities for each class, such as the probability that a given record or data point belongs to a particular class. The class with the highest probability is considered as the most likely class. This is also known as the maximum a posteriori (MAP) estimation [15].

The MAP formula used in this experiment is shown in Equation 2:

$$MAP(H) = \max(P(H|E)) = \max((P(E|H)*P(H))/P(E)) = \max(P(E|H)*P(H)) \quad (2)$$

Where:

P(E) = Evidence probability

P(H)= Hypothesis probability

3.5 Decision Tree

The decision tree is a type of supervised learning algorithm (having a pre-defined target variable) that is used mostly in classification problems. It works for both categorical and continuous input and output variables. In this technique, the decision tree splits the population or sample into two or more similar sets (or sub-populations) based on most significant splitter/differentiator in input variables [16]. The proposed decision tree to be used in this experiment is regression tree analysis. Regression tree analysis is useful when the output can be considered as a real number.

the accuracy measurement is based on :

$$\text{Precision (P)} = \frac{TP}{TP + FN}$$

$$\text{Recall (R)} = \frac{TP}{TP + FP}$$

Where :

TP : True Positive

FN : False Negative

FP : False Positive

4. Experimental Results and Analysis

The proposed system used an Intel i5 Core with 16 GB RAM, and was developed using Python and TensorFlow. The performance of the proposed method is evaluated using the classification accuracy obtained from the experiment performed using DCNN and another method for validation and comparison.

The income dataset was extracted from the Census Bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). The dataset consists of the following:

- Age
- Work class
- Fnlwgt
- Education
- Education-num
- Marital-status
- Occupation
- Relationship
- Race
- Sex
- Capital-gain
- Capital-loss
- Hours-per-week
- Native-country

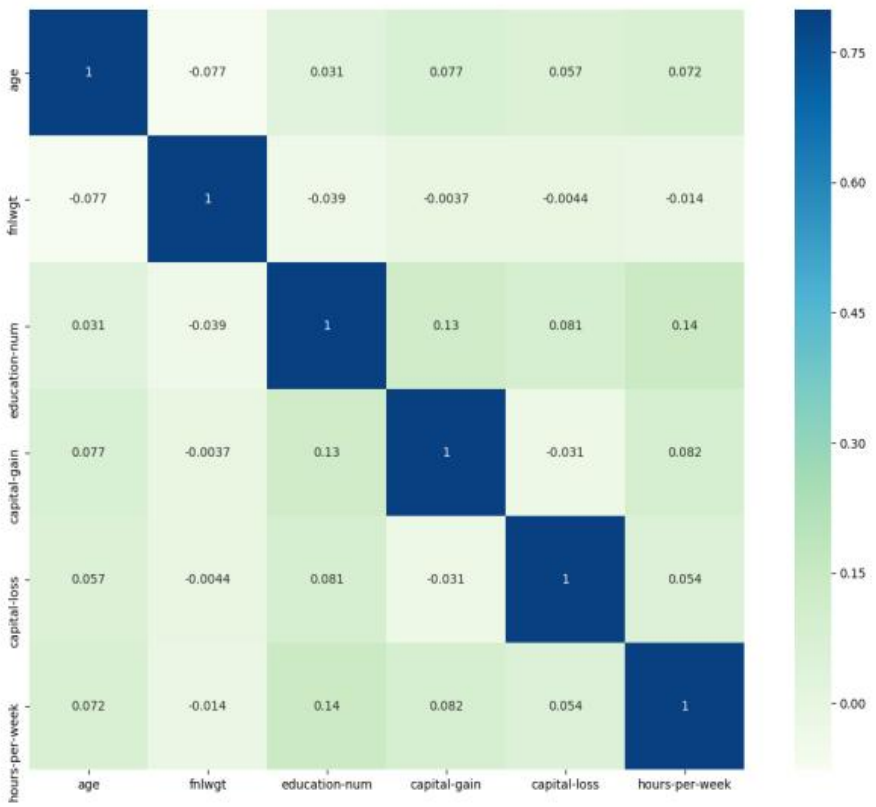

Fig. 4. Data distribution on income dataset.

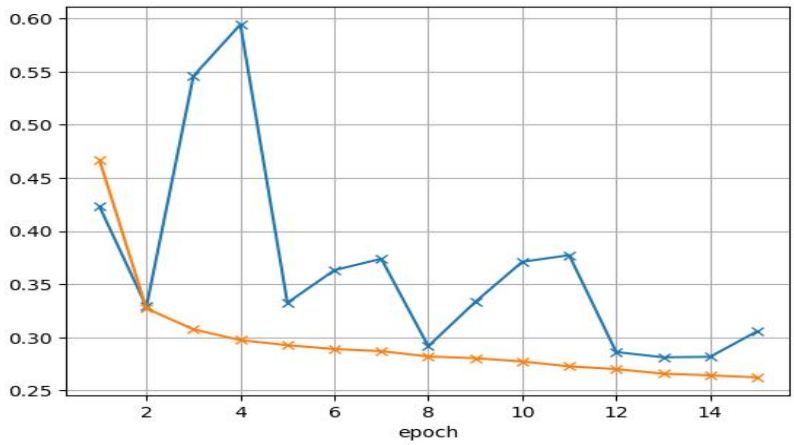

Fig. 5. Loss graph of DCNN (very deep convolutional neural network).

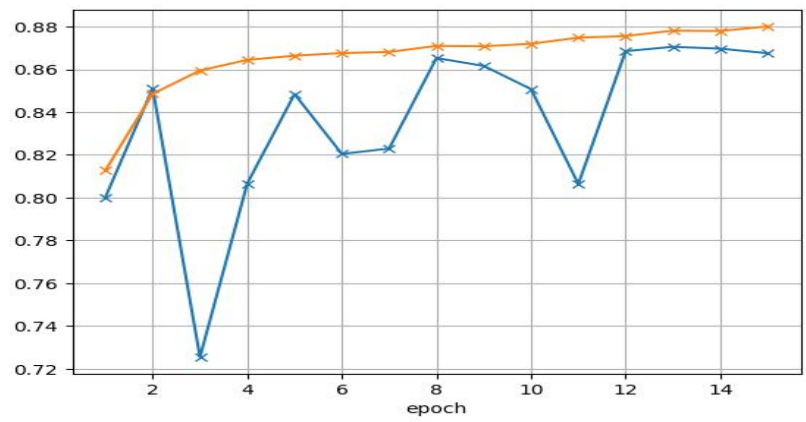

Fig. 6. Accuracy graph of DCNN (very deep convolutional neural network)

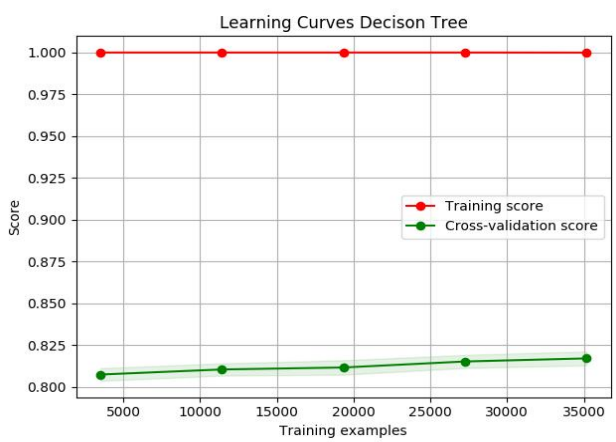

Fig. 7. Decision tree learning curves.

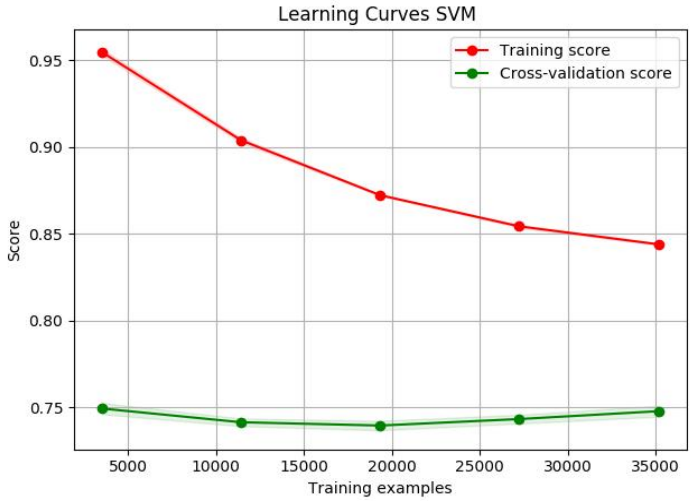

Fig. 8. SVM learning curve

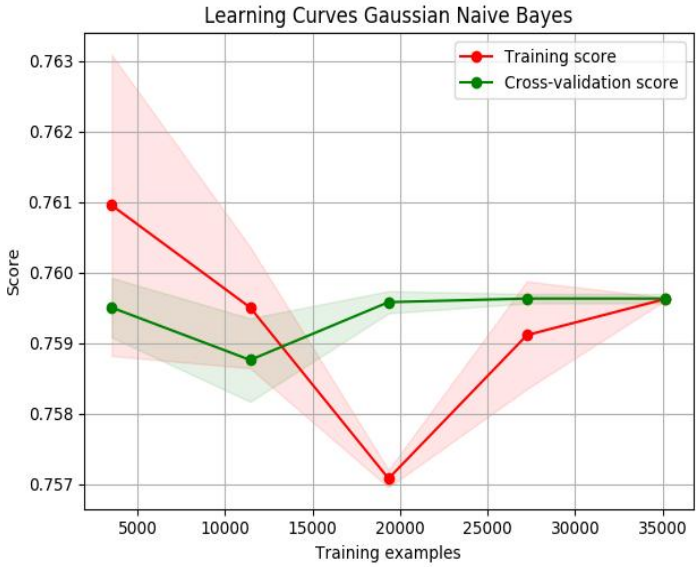

Fig. 9. Gaussian naïve Bayes learning curve

In this experiment, the data were chosen randomly for normal distribution. If the data do not distribute normally, the program will create the extra data or change data. Fig. 4 depicts the data distribution in the income dataset; based on the graph shown, the dataset was normally distributed.

Fig. 3 and 4 depict the loss and accuracy graph and the learning rate in DCNN. In Fig. 3 and 4, the orange line represents the training score, while the blue line represents the validation or test score; in the loss graph, the validation score and the test score decrease by epoch iteration although the test score fluctuated as the graph went down to create an ROC-shaped curve. In the accuracy graph, the test score fluctuated, but

increased for higher accuracy; so too did the orange line in the accuracy graph. Based on these graphs, the DCNN trains smoothly.

Fig. 5, 6 and 7 represent the learning rates for the decision tree, SVM and Gaussian naïve Bayes methods. Fig. 5 depicts the learning curve on the decision tree; the red line portrays the training score, while the green line portrays the cross-validation score. Based on Fig. 5, the decision tree did not train smoothly; the red and green lines in the decision tree learning graph show underfitting in the training process, as does Fig. 6 for the SVM training process, while the Gaussian naïve Bayes method trains well but not smoothly. Based on Fig. 7, the Gaussian naïve Bayes method shows training stuck between scores of 0.76 and 0.75, indicating an overfitting problem.

After the training process, accuracy was measured. The accuracy is obtained from precision and recall calculation. The Table 2 shows the prediction accuracy being produced by the machine learning method, with an average prediction accuracy value of 80%. Based on the graph and accuracy results, the DCNN method shows better and optimum performance for predicting income, while other machine learning methods (SVM, decision tree, and naïve Bayes) suffer from underfitting and overfitting.

Table 2. Comparison of income prediction results

No.	Methods	Accuracy Result	Description
1	DCNN	88%	
2	Naïve Bayes	75%	
3	SVM	85%	
4	Decision Tree	81%	

5. Conclusion

Based on the results of the deep convolutional neural network, this method shows an average optimum result for income prediction. The DCNN method performs better than the SVM, Gaussian naïve Bayes and decision tree methods, showing an average accuracy prediction of 88%. By contrast, the SVM, Gaussian naïve Bayes and decision tree methods struggle with overfitting or underfitting data when handling income prediction. This result uses the sentences data structure model, because the sentences data structure model fit with deep convolutional neural network. Based on all the result analysis and experiment DCNN/CNN could be used for predicting the income of the employer and be used for business purpose.

Even though the DCNN/CNN method was originally used for computer vision and images, it can be used in text problems as well. Images and small amounts of text have similar properties. Texts are also compositional in many languages; characters combine to form n-grams, stems, words, phrase, sentences, etc. These similar properties make it productive to compare computer vision and natural language processing.

This research aims to contribute to current understanding by seeking an optimum result for income and earning prediction, and to identify an efficient and effective method for doing so when dealing with large volumes of data and lengthy datasets. However, due to hardware and data length limitations, this experiment was able to handle an income dataset of no more than 35,000.

Based on current limitations and conditions, this research suggests several directions for future research. First, regarding the hardware limitation, future research is recommended using high-specification hardware to handle the training and preprocessing phase. Second, regarding the depth of the CNN, future research should adapt the depth of the CNN to current and future dataset conditions. Finally, regarding data structure, future research should shorten data structure and data length for better and optimum performance; Gensim, PCA and other optimization methods can be applied to achieve these conditions.

REFERENCES

[1] Cambridge University Press. (2008). *Cambridge online*

dictionary, Cambridge Dictionary online. http://www.temoa.info/node/324

[2] Case, K. & Fair, R. (2007). *Principles of Economics*. Upper Saddle River, NJ: Pearson Education. 54.

[3] A. Lazar. (2004). *Income prediction via support vector machine. 2004 International Conference on Machine Learning and Applications,* Proceedings, 143-149.

[4] Conneau, A., Schwenk, H., Barrault, L. & Lecun, Y. (2017). *Very Deep Convolutional Networks for Text Classification. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. 1. Long Papers. DOI:10.18653/v1/e17-1104

[5] A. Kibekbaev & E. Duman. (2015). *Benchmarking Regression Algorithms for Income Prediction Modeling. 2015 International Conference on Computational Science and Computational Intelligence (CSCI),* Las Vegas, 180-185. DOI: 10.1109/CSCI.2015.162

[6] K. Chen, L. Tian, H. Ding. M. Cai, L. Sun, S. Liang & Q. Huo (2017). *A Compact CNN-DBLSTM Based Character Model for Online Handwritten Chinese Text Recognition, 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Kyoto, 1068-1073.

[7] Bjelland, J., Reme B.A., Iqbal A. & Jahani, E. (2016), *Deep learning applied to mobile phone data for Individual income classification. International conference on Artificial Intelligence: Technologies and Applications (ICAITA)*, Atlantic Press, 96-99.

[8] The Data Science Blog. *(2018), An Intuitive Explanation of Convolutional Neural Networks. https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/*

[9] DL4J (2018). *A Beginner's Guide to Deep Convolutional Neural Networks (CNNs) - Deeplearning4j: Open-source, Distributed Deep Learning for the JVM. https://deeplearning4j.org/convolutionalnetwork*

[10] Besbes, A. (2018). *Understanding Deep Convolutional Neural Networks with a practical use-case in Tensorflow and Keras.* https://www.kdnuggets.com/2017/11/understanding-deep-convolutional-neural-networks-tensorflow-keras.html

[11] Namatēvs, I. (2017). *Deep Convolutional Neural Networks: Structure, Feature Extraction and Training. Information Technology and Management Science, 20(1),* 40 - 47 .

[12] B. E. Boser, I. M. Guyon, and V. N. Vapnik.(1992), *A training algorithm for optimal margin classifiers, COLT '92 Proceedings of the fifth annual workshop on Computational learning theory,* Pittsburgh, PA, ACM Press, 144 - 152.

[13] K. Nurhanim, I. Elamvazuthi, L. I. Izhar and T. Ganesan. (2017). *Classification of human activity based on smartphone inertial sensor using support vector machine, 2017 IEEE 3rd International Symposium in Robotics and Manufacturing Automation (ROMA)*, Kuala Lumpur, Malaysia, 1-5.

[14] Dataaspirant. (2017). *How the Naive Bayes Classifier works in Machine Learning.* http://dataaspirant.com/2017/02/06/naive-bayes-classifier-machine-learning/

[15] Analytics Vidhya Content Team. (2016). *A Complete Tutorial on Tree Based Modeling from Scratch (in R & Python).* https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/

[16] Albelwi, S. and Mahmood, A. (2017). *A Framework for Designing the Architectures of Deep Convolutional Neural Networks*. Entropy, *19(6)*, 242.

[17] G. D. Kim, Y. H. Kim.(2017). A Survey on Oil Spill and Weather Forecast Using Machine Learning Based on Neural Networks and Statistical Methods, *Journal of the Korea Convergence Society, 8(10)*, 1-8.

[18] C. J. Lee,. G. D. Kim, Y. H. Kim, (2017). Performance Comparison of Machine Learning Based on Neural Networks and Statistical Methods for Prediction of Drifter Movement, *.Journal of the Korea Convergence Society. 8(10),* 45-52.

[19] H. J. Yoon. (2017). Classification for early diagnosis for breast cancer base on Neural Network, *Journal of the Korea Convergence Society, 8(12)*, 49-53.

[20] K. T. Kim, J. Y. Choi. (2018). Facial Local Region Based Deep Convolutional Neural Networks for Automated Face Recognition, *Journal of the Korea Convergence Society, 9(4),* 47-55.

마렌드라(Ramadhani, Adyan Marendra) [학생회원]

- 2014년 1월 : Indonesia Universty of Education 자연과학대학 컴퓨터 과학 교육학과 학사
- 2016년 2월 : Bandung Institute of Technology 전기 공학 및 정보학대학 전기공학 석사(컴퓨터 공학)
- 2016년 2월 : 부경대학교 IT융합응용공학과 대학원정보시스템협동과정 석사
- 2016년 2월 ~ 현재 : 동아대학교 경영대학 경영정보학과 박사 과정
- 관심분야 : 인공지능, 정보시스템, 경영정보시스템, 블록체인
- E-Mail : adyan.rendra@gmail.com/ adyan.rendra@donga.ac.kr

김 나 랑(Na Rang Kim) [정회원]

- 1999년 2월 : 부산대학교 문헌정보학과(문헌정보학사)
- 2002년 2월 : 동아대학교 경영정보학과(경영학석사)
- 2007년 8월 : 동아대학교 경영정보학과(경영학박사)
- 2018년 4월 ~ 현재 : 동아대학교 경영대학 경영정보학과 조교수
- 관심분야 : Co-creation, 오픈이노베이션, 플랫폼, 빅데이터
- E-Mail : whitecoral@hanmail.net

최 형 림(Hyung Rim Choi) [정회원]

- 1979년 2월 : 서울대학교 경영학과(경영학학사)
- 1986년 2월 : 한국과학기술원 경영과학과(경영학석사)
- 1993년 8월 : 한국과학기술원 경영과학과(경영학박사)
- 1998년 10월 ~ 현재 : 동아대학교 경영대학 경영정보학과 교수
- 관심분야 : 항만물류, 유비쿼터스, RFID
- E-Mail : hrchoi@dau.ac.kr