Optimize the Combination of Categorical Variable Encoding and Deep Learning Technique for the Problem of Prediction of Vietnamese Student Academic Performance

Do Thi Thu Hien[1], Cu Thi Thu Thuy[2], Tran Kim Anh[3], Dao The Son[4], Cu Nguyen Giap[5]*
Department of Information Technology, Thuongmai University, Hanoi, Vietnam[1]
Department of Econometrics, Academy of Finnance Institute, Hanoi, Vietnam[2]
Department of Economics, Thuongmai University, Hanoi, Vietnam[3, 4]
Department of Informatics, Thuongmai University, Hanoi, Vietnam[5]

***Abstract*—Deep learning techniques have been successfully applied in many technical fields such as computer vision and natural language processing, and recently researchers have paid much attention to the application of this technology in socio-economic problems including the student academic performance prediction (SAPP) problem. In this specialization, this study focusses on both designing an appropriate Deep learning model and handling categorical input variables. In fact, categorical data variables are quite popular in student academic performance prediction problem, and deep learning technique in particular or artificial neural network in general only work well with numerical data variables. Therefore, this study investigates the performance of the combination categorical encoding methods including label encoding, one-hot encoding and "learned" embedding encoding with deep learning techniques including Deep Dense neural network and Long short-term memory neural network for SAPP problem. In experiment, this study compared these proposed models with each other and with some prediction methods based on other machine learning algorithms at the same time. The results showed that the categorical data transformation method using the "learned" embedding encoding improved performance of the deep learning models, and its combination with long short-term memory network gave an outstanding result for the researched problem.**

Keywords*—Deep learning techinique; categorical data type; "learned" embedding encoding; student academic performance prediction

I. Introduction

Modern education system has been changed with a new facet distance education, also called distance learning [1]. Educational technology facilitates distance learning and creates the shift from traditional education model to new model that concerns to virtual community of learners [2]. The recent global disease pandemic also encourages the researchers investigate in design and implement new higher education model [3]. In educational technologies, information communication technology (ICT) is a fundamental infrastructure to supply a massive online courses to learners [1]. However, in this trend, the student need smart systems that support them during learning process. The student academic performance prediction (SAPP) plays the key role in such smart systems. Base on academic performance prediction, the systems give learners suitable advices, early warning and many useful instructions.

Utilization of deep learning techniques on student academic performance prediction becomes a desirable research with many achievements recently [4, 5]. Education management information system (EMIS) is popular and it supplies available data resource for researches on educational data mining using deep learning techniques. However, one challenge when applying deep learning techniques for SAPP problem is that the input data of this problem often contains many categorical variables that the deep learning techniques or artificial neural networks in general do not work well directly [6, 7]. Therefore, research on how to convert categorical data to numerical data for construction and training deep learning models in SAPP problem is necessary.

This study focuses on the analysis of categorical variable transformation methods and its compatibility with deep learning models simultaneously. More specifically, the methods of transforming the categorical data is not studied separately from the classification models. Instead, each categorical data encoding method can be adapted to a classification model using some design of deep learning network. Therefore, this study investigates how categorical data conversion is associated with the corresponding deep learning model. There are several deep learning network architectures that can be used to develop a classification model, however, in the scope of this study the focused models include the long short-term memory recurrent network and the Deep Dense network. These deep learning network models were evaluated and considered as great solutions for SAPP data sets in a number of studies [5, 8]. Besides, the conversion of categorical variable to numerical variable in classification problems take an important position, especially classifiers are built based on artificial neural network or deep learning techniques [9, 10]. The common methods such as label encoding, one-hot encoding and its modifications and new "learned" embedding encoding [11, 12] are interest and they are going to be estimated carefully. Analysis experimental

*Corresponding Author

result of the compounds of categorical data transforms method with deep learning models and compare to result of other machine learning algorithms gives important conclusions for solving SAPP problem. The main analyzed indicator is the accuracy of the prediction. Although this index does not reflect all facets of a predictive classifier, it is still a most popular indicator used in this research area.

This article is divided into five main sections. Besides the introduction, the remainder includes Section 2 which presents related researches. Section 3 describes the research methodology and design of the proposal deep learning models for the SAPP problem. Section 4 presents the results and the last part is conclusion and some future research recommendations.

II. Related Works

Utilization of deep learning techniques and machine learning techniques for SAPP problem has been concerned in many researches [4, 5, 13, 14]. Using these techniques can improve the prediction quality [5, 13, 14] and opens many applications in reality [4]. In which, the application of deep learning techniques is a research direction that has received great attention in recent times [5]. For example, in the study [15], the authors proved that the Deep Dense neural network improve the accuracy of failure-prone student prediction. The convolutional neural network was investigated for the same researching area in [16]. These deep learning techniques were utilized for predicting student final performance in [8] and for predicting students' future development in [17]. The results showed that proposed deep learning models worked well in many interesting cases of educational data analysis.

More specifically, the deep learning model built on Deep Dense network, a multilayer perceptron network architecture, was proposed for the SAPP problem [4, 15, 18, 19]. The authors announced the good performance of Deep Dense network and also compared the proposed models with other algorithms such as the decision tree algorithm C4.5, random forest, logistic regression and support vector machine. Deep learning models for SAPP problem based on the Long short-term memory network and convolutional neural network were introduced in [5, 20, 21]. The results showed that the proposed deep learning models have superior results compared to other tested algorithms.

When applying neural networks or deep learning techniques to the problem of classification, one of the problems need to be solved is transforming the categorical data type into numerical data type [22]. There are different transformation methods that are commonly applied, in [22] the authors divided these data transformation methods into three groups: predetermined transformation methods, algorithmic methods and automatic transformation methods. These methods have a certain interference, but the first method focuses on the laws of clear change and often has very low complexity. The second method may give predetermined results, but the algorithmic method is geared towards more complex computational and processing methods. A third method uses machine learning techniques and one of them is neural networks to automatically mine data. Choosing the right method of transforming classified data is one of the challenges that need to be addressed when building a model for SAPP problem and classification problem in general. In the study [23], the researchers used the one-hot encoding conversion method. Although [23] did not evaluate the effectiveness of this transformation method, the results of the accuracy of the tested classification methods showed that this transformation method has good applicability. Besides, the "learned" embedding encoding method introduced in [24, 25] can also be applied. Research [26] was an example showing the suitability when applying the "learned" embedding encoding in categorical variable conversion.

Although there are many studies related to the application of deep learning techniques in the SAPP problem and processing of classification data in neural network, but these two groups of studies are quite independent with each other. The researching approach that directly combines two issues in one optimization process for the problem of SAPP will help improve the quality of prediction. Naturally, this study will focus on this approach.

III. Research Methodology

A. Categorical Variable Encoding

When processing input data of categorical data type for an artificial neural network, it is likely to convert a categorical variable to a numerical variable or a vector that its elements are numerical data type. Three common methods used are: 1) Label Encoding; 2) One-hot Encoding and its modification; 3) "Learned" Embedding encoding. In the Label encoding method each label of a categorical data variable is assigned to a most suitable integer number. This method is very simple, however, finding the right assignment for a specific problem is relatively difficult (especially with categorical variables representing unordered data). The second method, One-hot encoding, turns a categorical variable's value into a binary vector where one element takes value 1 presents the appearance and remained elements get value 0 presents the absence. Element takes value 1 in the position equivalent to each classifier label specified in the encryption method. Both the processing methods mentioned above are relatively easy to implement, and the data transformation process is independent of the data processing model, classification model in this researching situation.

In the third processing method, the "learned" embedding encoding method, a classification data will be transformed into the distribution vector, which is associated with the training and optimization of an artificial neural network. A well-trained vector space will provide a projection in which labels in the classification data are represented by naturally close clusters. This encoding method is first proposed for the word processing problem in natural language processing. Later, this method was used in the analysis of classification data, especially when used with traditional artificial neural network models or deep learning models.

The "learned" embedding encoding method is a learning method, therefore it requires a suitable training data. Besides, this method is an algorithm with several parameters, so the study of parameter optimization is also a problem to solve. The most important of these parameters is the size of the output-

dimension vector. The studies related to parameter optimization for "learned" embedding encoding are mainly done with word processing problem. In this study, the output-dimension vector is calculated by a recommended formula in many studies, as follows:

$$Emb\_size = \min(50, (n\_cat/2)+1) \quad (1)$$

Emb_size: is size of output dimension of an embedding layer.

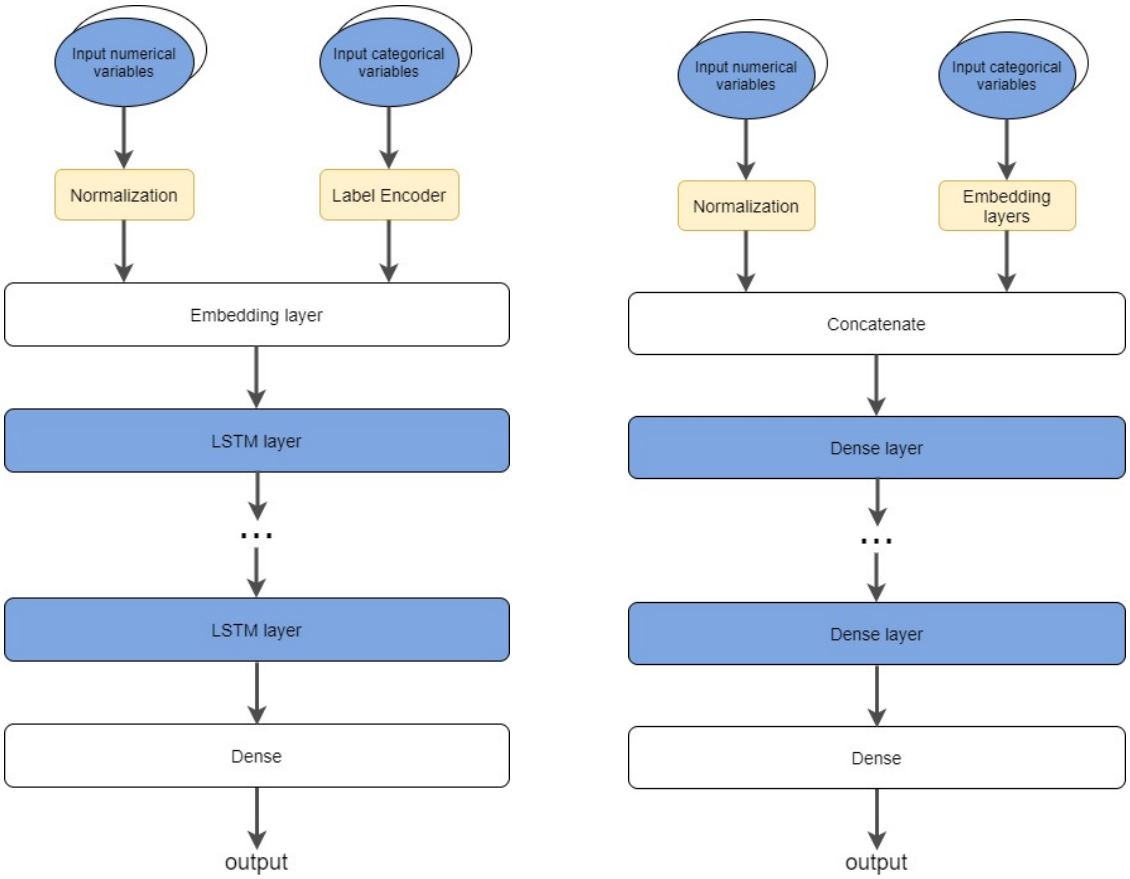
n_cat: is the force of the categorical variable.

B. Proposal Deep Learning Models

Combining methods of processing variables of classification data and deep learning techniques, two deep learning models are proposed for the SAPP problem. In which, a model uses Deep Dense network architecture (Fig. 1a), and one uses a Long short-term memory recurrent network architecture (Fig. 1b).

In particular, with LSTM network model, input categorical variables are converted by label encoding first. This data is combined with input numerical variables that are standardized to make input data for the LSTM network. The requirement of input data of a LSTM network is a 3-dimensional data type, so the input data is processed by an embedding layer first. The number of hidden LSTM layers added follow embedding layer and the number of hidden nodes of each layer can vary by designer. Each LSTM layer followed by a dropout layer to improve network performance based on reducing effect of the overfit problem. The original output variable is a categorical data type, and it is converted to a set of binary variables by One-hot encoding. Therefore, the output layer is a dense layer with the number of nodes equals to the number of output variables.

In the second model, each categorical variable is handled by a separate embedding layer. Then, all output of these embedding classes is aggregated with the normalized numerical variables by a concatenate layer. The output of the concatenate layer is the input to the dense network with the number of hidden layers depending on the design. Each Dense layer followed by a dropout layer too. The output of the problem is a set of binary variables, so the second model uses a dense layer as an output layer similar to the first model. Besides embedding encoding approach, two other encoding methods including label encoding and one-hot encoding are also experimented with Deep Dense network in the same structure.



(a) Long Short-Term Memory Network. (b) Deep Dense Network.

Fig. 1. Proposal Deep Learning Models.

IV. Experimental Result

A. Data Collection

Data of student academic performance was collected from Vietnamese universities, focused on students who followed disciplines related to economics. These data are extracted under the admission of grant number B2019-TMA-2 supported by Vietnamese Ministry of Education and Training. The data included information about student's learning capability at the entry time and two first years at university and final grade. Homology related information was collected but it was removed due to privacy concerns.

The experimental data includes 41 input attributes and one classification output, these attributes are described in Table I. There are 36 categorical input attributes, it takes 87.80% of input data. It means that categorical variable encoding is an important task for this problem. The remaining input attributes are 5 numerical variables. The data includes 524 observations, in which the ratio between normal classes such as pass and good grades and rare classes such as fail and distinction grades is quite large. This data is imbalanced data and it makes predictive classification problem become harder.

B. Experimental Result

The deep learning models proposed in Section III was implemented by Python programming language and this program used the Tensor Flow and Keras libraries. The experiment was run on a computer has following configuration: Core i7 2.0GHz, 16GB RAM, and 2GB GPU. Other machine learning algorithms its results were used to compare with the proposed models were implemented in Weka, an instance machine learning tool.

The method of evaluating the forecasting models is based on accuracy indicator. According to the accuracy measurement, is calculated by the formula: number of correct classified observations/ total number of observations, and, average accuracy is assessed by 10 times running on 10 testing data sets randomly spliced from collected data.

In the scope of this study, there are several parameters of deep learning models fixed. That are the number of hidden layers of the neural network is ignited at 2. The training process uses cell training with batch size = 100, Adam optimization method with learning rate parameter= 0.001. The fraction of the input units to drop is set at 0.15 at each dropout layer. Training accuracy of proposed models with number of interactions epochs=1000, is shown in the following figure.

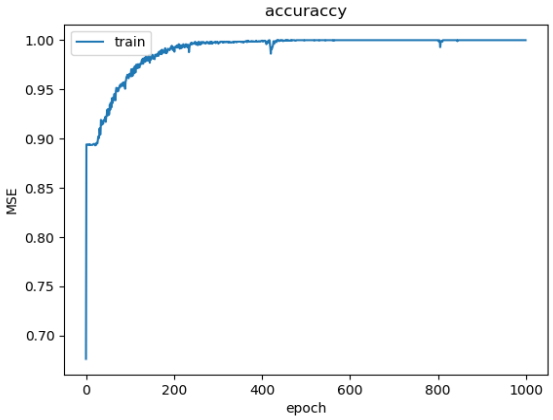
In Fig. 2, it can be seen that the training processes of all proposed deep learning models have a stable convergence and its reach expected error within 1000 epochs. For example, in Fig. 1, the Deep Dense networks converged to highest training accuracy at 100%. Besides, the Deep Dense network combines with "learned" embedding encoding for categorical variables converged faster than other testing deep learning models.

Base on convergence of the accuracy during training process, the proposed Deep learning models are continuously experimented with the same setting above, but the number of epochs is settled at 500. This value ensures that the experiment can reach expected correctness and it saves time also. The experimental result is calculated on the test data, which were generated by randomly selecting 25% of the observations from the original data set.

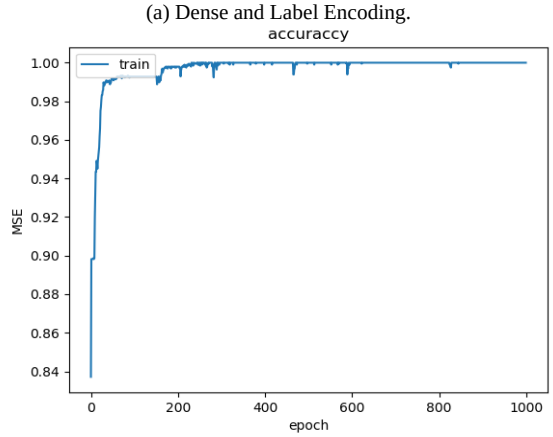
Optimization of number of hidden layers and number of hidden layers nodes are necessary for deep learning model design. However, there are not a thumb rule for setting these values in all situations. The popular approach lays down in experimentation. This study used this approach, specifically, the number of candidates for good setting was retrieved by literature review, and then these candidates were investigated by experiment to seek the optimum solution among candidates.

TABLE I. Description of Input Variables

Variable Name	Range	Description
gap_year	[0-3]	Elapsed time between high school graduation and university entrance.
Mark1, Mark2, Mark3	[0-10]	Individual marks of entrance exam
Total_mark	[0-30]	Total mark of admission
Group_Sub	1-3	Student's group
Gen_Sub_i, i=[1-22]	A, B, C, D, F	Scores of general subjects
$Spec\_Ground\_Sub_i$, i= [1-7]	A, B, C, D, F	Scores of specialized round subjects/modules
$Spec\_Sub_i$, i= [1-6]	A, B, C, D, F	Scores of other specialized subjects/modules



(a) Dense and Label Encoding.



(b) Dense and "Learned" Embedding Encoding.

Fig. 2. Training Convergence.

Based on the literature review and experiment the optimum number of hidden layers in all proposed deep learning models is 2. The average accuracy was not improved when the number of hidden layers was increased to 3 and 4, but the training time increased. Therefore, all deep learning models had 2 hidden layers in on going test.

Fig. 3 presents the performance of Deep Dense network combines with label encoding method or categorical data type. It can be seen that two values of number of hidden nodes give better performances than others. In this case, the best number of hidden nodes is chosen at 100, even though it give average accuracy 83.359% that is very close to result of the selection of 300 hidden nodes. Because the smaller number of hidden nodes leads to faster training process.

According to deep experimentation, the optimum number of hidden layer nodes in Deep Dense network with both Label encoding and One-hot encoding methods is 100, while the optimum setting for Deep Dense network with "learned" embedding encoding is 150, while the optimum number of hidden layer nodes in LSTM Network with "learned" embedding encoding is 100.

The performances of deep learning models with the optimum parameters are depicted in Fig. 4. The accuracy is measured by the average results of 15 run times with the experimental appropriate set of parameters mentioned above. It can be seen in Fig. 4, the box plot chart showed the LSTM-Network with "learned" embedding encoding method had the best performance. This network has highest average accuracy that was 86.26%, and this network was also more stable presented by median, lower quartile and upper quartile lines. Moreover, the box plot of LSTM model showed that most of running test often gave testing accuracy closes to maximum value.

In the same point of view, the embedding encoding method is also good to combine with Deep Dense network, it helps this network architecture works better than other encoding method. In tested SAPP problem, Deep Dense network with Label encoding method was better than using this network architecture with One-hot encoding method. The reason seems to be that most categorical variables in input data are ordinal data type.

In general, the performance of all testing algorithms is showed in Fig. 5. Experimentation proves that for student performance prediction the deep learning models combined with the embedding encoding method called "learned embedding" encoding for categorical variables gave good results. Deep Dense models with label encoding and one-hot encoding methods have competitive results with the best testing machine learning algorithms (SMO and random forest). However, Deep Dense network and LSTM network with "learned" embedding encoding have out performance.

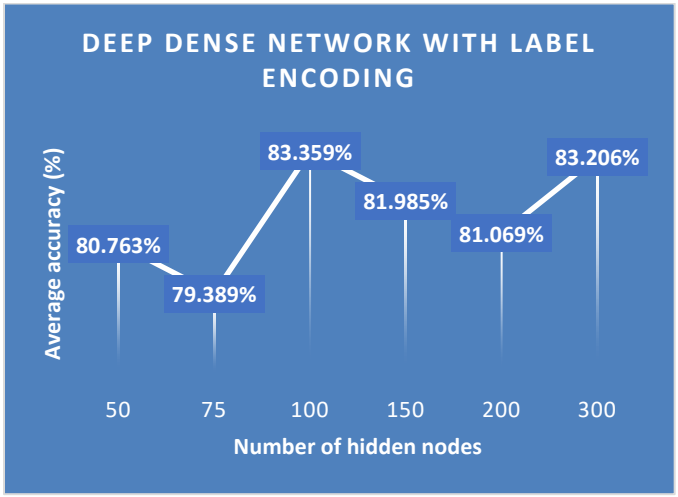


Fig. 3. The Performance of Deep Dense Network with Label Encoding.

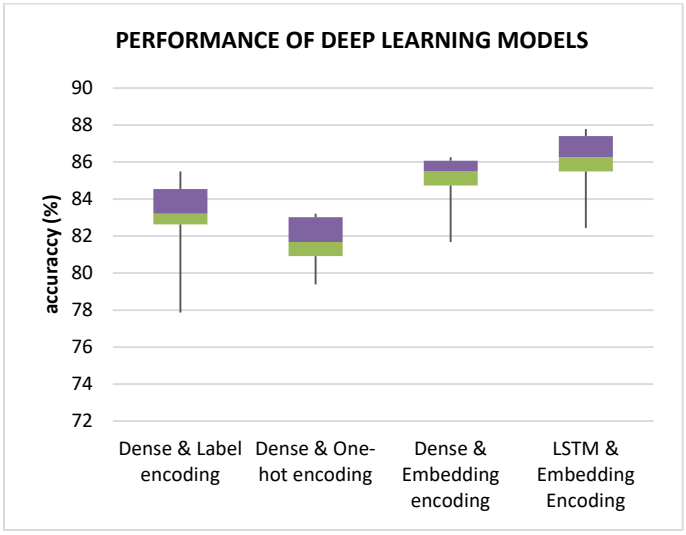


Fig. 4. Performances of Proposed Deep Learning Models.

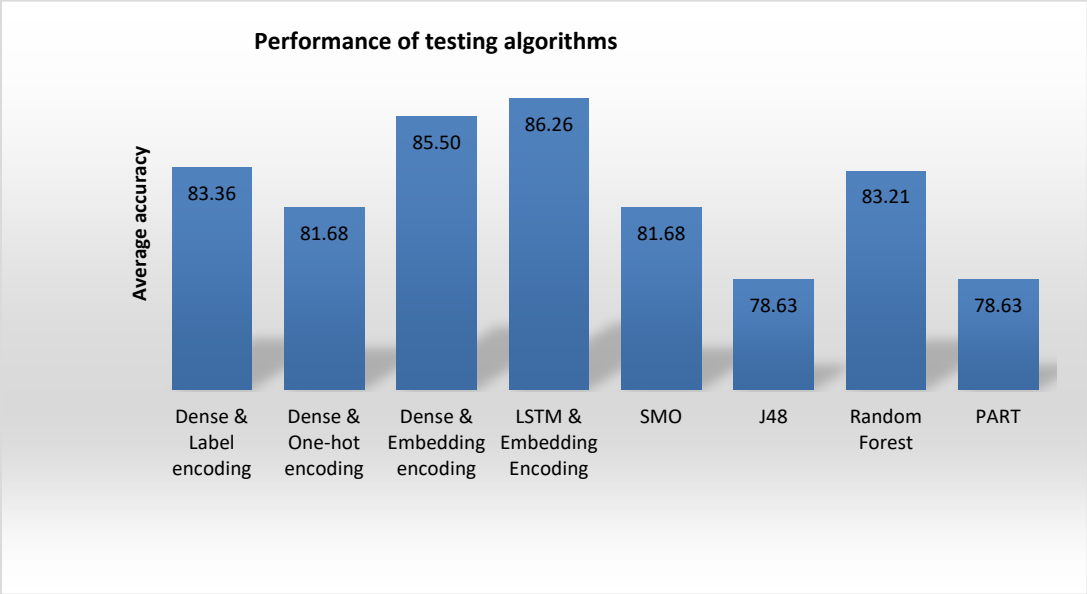


Fig. 5. Performance of All Testing Algorithms.

V. Conclusion

This study is an empirical study to solve the problem of student performance prediction with deep learning techniques, in which the construction of a deep learning model is studied coincidentally with categorical variables processing. Three methods of converting categorical variables to numeric variables including label encoding, One-hot encoding, and the embedding encoding methods were studied. Meanwhile, Deep Dense network and Long-short term memory network architectures designed in accordance with these encoding methods.

The experimental results give good insights and demonstrated the effectiveness of the "learned" embedding encoding method, because this encoding method has ability to learning simultaneously with the neural network in training process. In which using this data transform method with LSTM deep learning architecture gave the best result between other testing methods. This model has average accuracy at 86.26%. Embedding encoding method also improves performance of Deep Dense network also, and it also point out that to solve the prediction problems, which have categorical input variables, a deep learning model need to be designed with categorical variable encoding simultaneously.

The study also has a limitation that it takes time for experiment to find the optimal set of parameters in deep learning model design. Therefore, in this study, some parameters are selected according to recommendations. For example, the commonly used activate function is the "sigmoid" function, the recommended number of hidden layers is two. In the future, this study is going to be extended to solve these limitation.

Acknowledgment

This study was supported by Vietnamese Ministry of Education and Training and Thuongmai University under grant number B2019-TMA-02.

References

[1] Kaplan, A. M., & Haenlein, M. (2016). Higher education and the digital revolution: About MOOCs, SPOCs, social media, and the Cookie Monster. Business Horizons, 59(4), 441-450.

[2] Garrison, D. R. (2011, 20 May). E-learning in the 21stcentury: A framework for research and practice. New York: Taylor & Francis. ISBN:0-203-83876-9.

[3] Ilmiyah, S., & Setiawan, A. R. (2020). Students' Worksheet for Distance Learning Based on Scientific Literacy in the Topic Coronavirus Disease 2019 (COVID-19). EdArXiv, 7 Apr. 2020. Web

[4] Muniasamy, A., & Alasiry, A. (2020). Deep Learning: The Impact on Future eLearning. International Journal of Emerging Technologies in Learning (iJET), 15(01), 188-199.

[5] Doleck, T., Lemay, D. J., Basnet, R. B., & Bazelais, P. (2020). Predictive analytics in education: A comparison of deep learning frameworks. Education and Information Technologies, 25(3), 1951-1963.

[6] Hancock, J. T., & Khoshgoftaar, T. M. (2020). Survey on categorical data for neural networks. Journal of Big Data, 7, 1-41.

[7] Zhang W., Du T., Wang J. (2016) Deep Learning over Multi-field Categorical Data. In: Ferro N. et al. (eds) Advances in Information Retrieval. ECIR 2016. Lecture Notes in Computer Science, vol 9626. Springer, Cham.

[8] Aydoğdu, Ş. (2020). Predicting student final performance using artificial neural networks in online learning environments. Education and Information Technologies, 25(3), 1913-1927.

[9] Hancock, J. T., & Khoshgoftaar, T. M. (2020). Survey on categorical data for neural networks. Journal of Big Data, 7, 1-41.

[10] Zhang W., Du T., Wang J. (2016) Deep Learning over Multi-field Categorical Data. In: Ferro N. et al. (eds) Advances in Information Retrieval. ECIR 2016. Lecture Notes in Computer Science, vol 9626. Springer, Cham.

[11] Chen T, Tang L-A, Sun Y, Chen Z, Zhang K. Entity embedding-based anomaly detection for heterogeneous categorical events. 2016. arXiv:1608.07502.

[12] Goyal P, Ferrara E. Graph embedding techniques, applications, and performance: a survey. Knowl Based Syst. 2018;151:78–94.

[13] Manjarres, A. V., Sandoval, L. G. M., & Suárez, M. S. (2018). Data mining techniques applied in educational environments: Literature review. Digital Education Review, (33), 235-266.

[14] Tran, T. O., Dang, H. T., Dinh, V. T., & Phan, X. H. (2017). Performance prediction for students: a multi-strategy approach. Cybernetics and Information Technologies, 17(2), 164-182.

[15] Kostopoulos, G., Tsiakmaki, M., Kotsiantis, S., & Ragos, O. (2020). Deep Dense Neural Network for Early Prediction of Failure-Prone Students. In Machine Learning Paradigms (pp. 291-306). Springer, Cham.

[16] Akour, M., Al, S. H., & Al Qasem, O. (2020). The effectiveness of using deep learning algorithms in predicting students achievements. Indonesian J. Elect. Eng. Comput. Sci., 19(1), 387-393.

[17] Fok, W. W., He, Y. S., Yeung, H. A., Law, K. Y., Cheung, K. H., Ai, Y. Y., & Ho, P. (2018, May). Prediction model for students' future development by deep learning and tensorflow artificial intelligence engine. In 2018 4th international conference on information management (ICIM) (pp. 103-106). IEEE.

[18] Ha, D. T., Loan, P. T. T., Giap, C. N., & Huong, N. T. L. (2020). An Empirical Study for Student Academic Performance Prediction Using Machine Learning Techniques. International Journal of Computer Science and Information Security (IJCSIS), 18(3).

[19] Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. Computers in Human Behavior, 104, 106189.

[20] Kim, B. H., Vizitei, E., & Ganapathi, V. (2018). GritNet: Student performance prediction with deep learning. arXiv preprint arXiv:1804.07405.

[21] Akour, M., Al, S. H., & Al Qasem, O. (2020). The effectiveness of using deep learning algorithms in predicting students achievements. Indonesian J. Elect. Eng. Comput. Sci., 19(1), 387-393.

[22] Hancock, J. T., & Khoshgoftaar, T. M. (2020). Survey on categorical data for neural networks. Journal of Big Data, 7, 1-41.

[23] Wen, H., & Huang, F. (2020, May). Personal Loan Fraud Detection Based on Hybrid Supervised and Unsupervised Learning. In 2020 5th IEEE International Conference on Big Data Analytics (ICBDA) (pp. 339-343). IEEE.

[24] Cheng G, Berkhahn F. Entity embeddings of categorical variables. CoRR. 2016. arXiv:1604.06737.

[25] Goyal P, Ferrara E. Graph embedding techniques, applications, and performance: a survey. Knowl Based Syst. 2018;151:78–94.

[26] Chen T, Tang L-A, Sun Y, Chen Z, Zhang K. Entity embedding-based anomaly detection for heterogeneous categori-cal events. 2016. arXiv :1608.07502