VOL 1 (2017) NO 4 - 2

e-ISSN : 2549-9904
ISSN : 2549-9610

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

Classification of Alcohol Consumption among Secondary School Students

Shamala Palaniappan[#], Norhamreeza A Hameed[#], Aida Mustapha[#], Noor Azah Samsudin[#]

[#] *Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Malaysia*
E-mail: shamala@uthm.edu.my, shamala@uthm.edu.my, aida@uthm.edu.my, azah@uthm.edu.my

***Abstract*— In 2016, the National Institute of Health reported that 26% of 8th graders, 47% of 10th graders, and 64% of 12th graders have all had experience in consuming alcoholic drinks. This finding indicates an accelerating trend in alcohol use among school students, hence a growing concerns among the public. To address this issue, this paper is set to model the alcohol consumption data among the secondary school students and attempt to predict the alcohol consumption behaviors among them. A set of classification experiments are carried out and the classification accuracies are compared between two variations of neural network algorithms; a self-tuning multilayer perceptron classifier (AutoMLP) against the standard MLP using the student alcohol consumption dataset. It is found that AutoMLP produced better accuracy of 64.54% than neural network with 61.78%.**

***Keywords*— Neural network, classification, data mining, alcohol consumption.**

I. Introduction

Alcoholic drink is a drink that contains certain amount of ethanol or alcohol that is able to cause negative effects to the consumers [4]. Short-term effects of alcohol usage include slurring in speech, drowsiness, emotional changes, sleep disruption and lower body temperature. In the long run, however, continuous consumption will cause death of brain cells which can lead to brain disorder and weaken the immune system, making the body a much easier target for disease. Taking alcohol especially during teen age can cause long term bad impacts because teenagers are not aware with the danger of alcoholic addiction. Addiction at younger age will affect their brain development and induce such as vandalism and bullying. Besides, they are exposed to variety of health problems such as sleep disorder, headaches or even worse mental health problems such as depression and suicidal.

A recent survey by the National Institute of Health (NIH) revealed that despite the law prohibiting purchase of alcoholic beverages among secondary schools and most college students, most of them have had a substantial amount of experience consuming alcohol as early as in 8th grade [9]. To study this issue, this research is set to predict alcohol addiction among the secondary school students given a set of parameters or factors leading to alcohol addiction. A number of alcohol consumption models have been proposed in the literature such as by [3, 6, 7].

To achieve this objective, a classification task from a data mining methodology is proposed [6, 9, 11, 12]. Classification involves determining an object into predefined groups of classes [1, 5, 14]. Classification is proficient when processing a large amount of data. It can be employed to predict label of categorical class and classifies data based on the training set and labels of class. The remainder of this paper proceeds as follows. Section II presents the materials and methods used to achieve the objective of this research, Section III presents the experimental results, and finally Section IV concludes with some indication for future work.

II. Methods

Figure 5 shows the methodology. The classification of alcohol consumption among secondary school students requires input, pre-processing, classification and output.

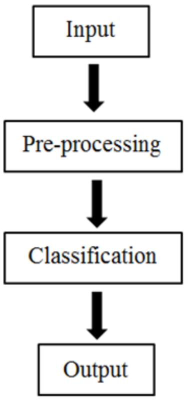

Fig. 1 Classification methodology.

Input is the collection of data that consists of instances and attributes. A pre-processing step is required to handle incomplete, noisy and uncertain data. The classification experiment will compare between two classification algorithms, which are multilayer perceptron (MLP) neural network and a self-tuning multilayer perceptron classifier (AutoMLP). Finally, the output of the classification experiment is measured in the form of classification accuracy.

Dataset

The Student Alcohol Consumption dataset was sourced from [13], originating from secondary school student performance dataset by [2]. Their goal was predicting alcohol consumption by secondary school student by studying the correlation between alcohol usage and the social, gender and study time attributes for each student. The dataset consists of 1,024 student information with 33 attributes. The previous research in [13] has identified a common pattern that forms 14 attributes as shown in Figure 2.

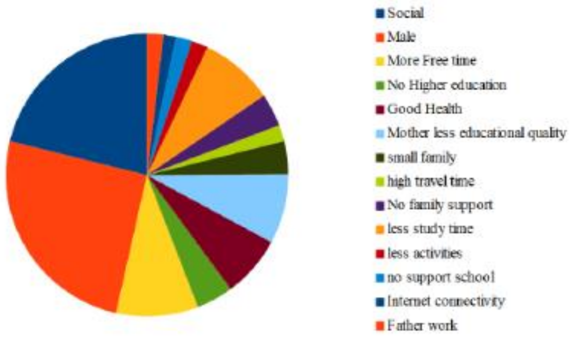

Fig. 2 Attributes proposed by [13].

The patterns shown in Figure 2 cover most of family information such as the social status, free times, education level, parent's educational level, size of family, financial support, Internet connection, schooling and past times activities. The final attributes used in the classification experiment is shown in Table 1.

TABLE 1
DATASET ATTRIBUTES AND DESCRIPTIONS.

Attributes	Descriptions
age	student's age (numeric: from 15 to 22)
medu	mother's education
fedu	father's education
traveltime	home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
studytime	weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
failures	number of past class failures (numeric: n if 1<=n<3, else 4)
famrel	quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
freetime	free time after school (numeric: from 1 - very low to 5 - very high)
goout	going out with friends (numeric: from 1 - very low to 5 - very high)
dalc	workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
health	current health status (numeric: from 1 - very bad to 5 - very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20, output target)

Pre-Processing

Pre-processing is necessary because the dataset contains noisy, inconsistent, missing and outdated values. In this research, data pre-processing selected 15 attributes as shown in Table 1. The first step involved is data cleaning. The aim of data cleaning is to fill in missing value, smooth the data, identify or remove the outliers and noisy data and finally to resolve data inconsistencies. The dataset was also normalized using Z-transformation.

Classification

Classification is a two-step process; training and testing. The training or learning step is when the classification model is constructed based on the input data. The classifier is built describing a predetermined set of data classes or concepts. During testing, the classifier model constructed is then used to predict class labels for the given data in testing set. The classification experiments were carried out using the Rapid Miner software (https://rapidminer.com/). The dataset was divided into training and testing set and then passed as input to a standard multilayer perceptron (MLP) and an AutoMLP algorithm in Rapid Miner. A Multi-Layer Perceptron (MLP) is a type of feed-forward neural network for mapping sets of input data onto a set of appropriate outputs. MLP is usually characterized by three layers of neurons; input layer, hidden layer and output layer with nonlinear activation functions at the hidden layer units.

AutoMLP is a simple algorithm that is able to adjust both of its learning rate and size of the neural network during training itself. An AutoMLP network maintains a small ensemble of networks trained in parallel with different rates and different numbers of hidden units. After a certain epoch, the error rate is calculated based on the testing dataset and the highest error rate will be replaced with the network of smallest error rate. Finally, the number of hidden units and the learning rates are drawn based in probability distribution derived from successful rates and sizes. The standard MLP and AutoMLP models were constructed based on the training set, which were then used to classify the new data in the training set. The classification experiment applied 10-fold cross-validation for a more consistent result.

III. RESULTS AND DISCUSSION

Based on the two classification algorithms, a standard MLP and the self-tuning MLP called the AutoMLP, the results of performance measure for the classification experiment in student alcohol consumption in terms of the classification accuracy, squared error and root mean squared error are reported in Table 2. The estimate for accuracy is the overall number of correct classifications from the epochs, divided by total number of tuples in the initial data.

TABLE 2
COMPARATIVE PERFORMANCE MEASURES.

Performance Measures	MLP (%)	AutoMLP (%)
Accuracy	61.78	64.54
Squared error	0.360	0.322
Root mean squared error	0.599	0.566

From the results, AutoMLP produced better result as compared to the standard MLP with accuracy of 64.54%, squared error of 0.322 and root mean squared error of 0.566. Figure 3 shows the overview of the neural network diagram for the dataset.

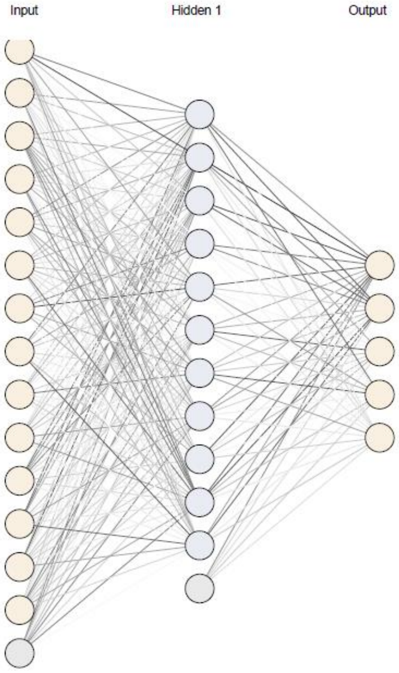

Fig. 3 Overview of the neural networks.

IV. SUMMARY

This paper was set to predict alcohol consumption among secondary school students using a Portuguese secondary school dataset composed by [2]. The experiment was motivated by the shocking report by the National Institute of Health that revealed a high percentage of teenagers have had a substantial amount of experience with alcohol during the school days. Given the dataset, a series of classification was designed and carried out. In order to test the model performance, the experiment then compared the classification accuracy between two variants of neural network algorithms, which is a self-tuning multilayer perceptron classifier (AutoMLP) and a standard MLP classifier. It is found that AutoMLP produced better accuracy of 64.54% than neural network with 61.78%. In the future, this dataset must be further enrich with other indicative variables that relates to alcohol consumption because this dataset is originally design for predicting school student performance where alcohol consumption is only part of the parameters they observed.

ACKNOWLEDGEMENT

This research is sponsored by Universiti Tun Hussein Onn Malaysia.

REFERENCES

[1] Aggarwal, C.C. (2015). Data mining. Springer.

[2] Cortez, P. & Silva, A.M.G. (2008). Using data mining to predict secondary school student performance.

[3] Crutzen, R., Giabbanelli, P.J., Jander, A., Mercken, L., & Vries, H.d. (2015). Identifying binge drinkers based on parenting dimensions and alcohol-specific parenting practices: building classifiers on adolescent-parent paired data. BMC Public Health, 15(1):747.

[4] Gunzerath, L., Faden, V., Zakhari, S., & Warren, K. (2004). National Institute on Alcohol Abuse and Alcoholism report on moderate drinking. Alcoholism: Clinical and Experimental Research, 28, 829-847.

[5] Gupta, G. (2014). Introduction to data mining with case studies: PHI Learning Pvt. Ltd.

[6] Hamid, N.A., Nawi, N.M., & Ghazali, R. (2011). The Effect of Adaptive Gain and Adaptive Momentum in Improving Training Time of Gradient Descent Back Propagation Algorithm on Classification Problems. Proceedings of the International Conference on Advanced Science, Engineering and Information Technology, 178-184.

[7] Hariharan, B., Krithivasan, R., & Angel, D. (2016). Prediction of Secondary School Students' Alcohol Addiction using Random Forest. International Journal of Computer Applications, 149(6).

[8] Jackson, N., Denny, S., Sheridan, J., Fleming, T., Clark, T., Teevale, T., & Ameratunga, S. (2014). Predictors of drinking patterns in adolescence: A latent class analysis. Drug and Alcohol Dependence, 135: 133-139.

[9] Lashari, S.A., Ibrahim, R., Senan, N., Yanto, I.T., & Herawan, T. (2016). Application of Wavelet De-noising Filters in Mammogram Images Classification Using Fuzzy Soft Set. Proceedings of the International Conference on Soft Computing and Data Mining, p. 529-537.

[10] Miech, Richard A., et al. (2016). Monitoring the Future national survey results on drug use, 1975-2015: Volume I, Secondary school students.

[11] Nawi, N.M., Hamid, N.A., Harsad, H., & Ramli, A.A. (2016). Second Order Back Propagation Neural Network (SOBPNN) Algorithm for Medical Data Classification. Proceedings of the Computational Intelligence in Information Systems: Proceedings of the Fourth INNS Symposia Series on Computational Intelligence in Information Systems (INNS-CIIS 2014). In S. Phon-Amnuaisuk and T. W. Au (Eds.), Springer International Publishing, 73-83.

[12] Nawi, N.M., Ransing, R.S., Salleh, M.N.N., Ghazali, R., & Hamid, N.A. (2010). An Improved Back Propagation Neural Network Algorithm on Classification Problems. Database Theory and Application, Bio-Science and Bio-Technology, In Y. Zhang, A. Cuzzocrea, J. Ma, K.-i. Chung, T. Arslan, and X. Song (Eds.), 108, 177-188.

[13] Pagnotta, F. & Amran, M.H. (2016). Using data mining to predict secondary school student alcohol consumption. Department of Computer Science, University of Camerino.

[14] Witten, I.H., Frank, E., Hall, M.A., & Pal, C.J. (2016). Data Mining: Practical machine learning tools and techniques: Morgan Kaufmann.