Recibido: 22-12-2018 / Revisado: 28-03-2019
Aceptado: 28-03-2019 / Publicado: 28-03-2019

Towards an improved classification model based on Deep Learning and nearest rules strategy

Hacia un modelo de clasificación mejorado basado en el aprendizaje profundo y la estrategia de reglas más cercanas

Mohammed El Fouki[1], Noura Aknin[1], Kamal E. El Kadiri[1]

[1] Abdelmalek Essaadi University, Morocco

melfouki@uae.ac.ma , noura.aknin@uae.ac.ma , kelkadiri@uae.ac.ma

ABSTRACT. In this paper we present a comparison between two improved approaches i.e. hybrid rules-based method with a proposed wrapper and nearest rule strategy, deep principal component analysis. We also perform several experiments with an analysis dataset from a distance learning platform. Several classifiers were developed to compare the performance of the proposed approaches, using accuracy, TP rate, F measure, PRC area, MCC, precision, recall and receiver operating characteristics area (AROC) as metrics. The result con□rms the utility of these algorithms for classification and shows clearly the superiority of our approaches.

RESUMEN. En este artículo presentamos una comparación entre dos enfoques mejorados, es decir, un método híbrido basado en reglas con una envoltura propuesta y una estrategia de reglas más cercana, análisis profundo de componentes principales. También realizamos varios experimentos con un conjunto de datos de análisis desde una plataforma de aprendizaje a distancia. Se desarrollaron varios clasificadores para comparar el rendimiento de los enfoques propuestos, utilizando la precisión, la tasa de TP, la medida de F, el área de PRC, el MCC, la precisión, la recuperación y el área de características operativas del receptor (AROC) como métricas. El resultado confirma la utilidad de estos algoritmos para la clasificación y muestra claramente la superioridad de nuestros enfoques.

KEYWORDS: Educational Data Mining (EDM), Classification, Rules-based, Nearest rules strategy, Wrapper approach, Entropy condition, Deep Learning.

PALABRAS CLAVE: Minería de datos educativos (EDM), Clasificación, Basado en reglas, Estrategia de reglas más cercanas, Enfoque envolvente, Condición de entropía, Aprendizaje profundo.

El Fouki, M.; Aknin, N.; El Kadiri, K. E. (2019). Towards an improved classification model based on Deep Learning and nearest rules strategy. *Campus Virtuales*, 8(1), 111-119.

1. Introduction

Recently, artificial intelligence starts to gain more attention in educational environ-ments, to discover new interesting and useful knowledge about students. EDM is one of the emerging discipline of machine learning that focuses on applying data mining techniques to several educational systems and all its related data (Scheuer & McLaren, 2011). Thus, like described in section II, there is a wide range of topics within EDM. In this paper we will focus exclusively on rules-based and deep learning algorithms, especially, the ways using them to improve student success and their learning pro-cess. EDM receives settings from areas like Learning Analytics, Statics, Text Mining, Computing and Data Mining, Psychometrics, Artificial Intelligence and machine learning, Information Technology, Database Management System, learning systems. However, researchers start to explore various data mining methods to help educators to understand their students' learning processes and to improve the teaching perfor-mance (Romero & Ventura, 2010). We will focus exclusively on ways that data min-ing is used to increase classifier performance, and then to predict students' finals marks, student success, and evaluation of student performances within other e-earning sittings. Educational data mining is now being used in adaptive and intelligent hy-permedia (Brusilovsky & Peylo, 2003), learning management systems (Hidalgo Cajo, 2018; Martínez, 2017; Abdelouarit, Sbihi & Aknin, 2015) and intelligent tutoring systems (Sleeman & Brown, 1982).

The goal of this technique is to estimate a target value of a variable that describes the student. The values usually predicted are knowledge, performances, score or mark. Hens, we can discover two types of values generated by prediction models; the first is a numerical or continuous value, the second is a categorical or discrete value. The first type of values represents a task starts by exploiting datasets in which the target labels are known. For example, a regression model that predicts surgery duration could be developed based on supervised data for many surgeries over a period of time. On the other hand, the categorical value assumes the classification task.

However, predicting distance learning students' performance still one of the oldest and most difficult problems of data mining in education, and various models and techniques have been applied such as neural networks (Oladokun, Adebanjo & Charles-Owaba, 2008), Bayesian networks (Conati, Gertner & Vanlehn, 2002), rule-based systems (Golding, Rey & Rosenbloom, 1991), regression (Ibrahim & Rusli, 2007) and correlation analysis (Boulaajoul & Aknin, 2019).

2. Background: rules-based & classification techniques

Data mining and machine learning represent the most area that educational data min-ing methods can be drowning from, by exploiting the hierarchical levels of meaning-ful features in educationally related data. Romero and Ventura (2010) summarize all the EDM research papers into the following categories: clustering, classification, out-lier detection, associations rule mining, sequential pattern mining and text mining.

There are many classifications' applications and predictions' tasks in educationally related environments that have been resolved by educational data mining methods DM. For example:

- Analysis and visualization of data.
- Students modeling.
- Students Profiling
- Predicting student performance.
- Grouping students.
- Recommendations for students.
- Domain modeling and clustering.
- Providing feedback for instructors.
- Trend Analysis.
- Detecting undesirable student behaviors.
- Improved teaching support.

- Social network analysis.
- Planning and scheduling.
- Developing concept maps.
- Students Behavior Modeling.
- Constructing courseware.

The purpose of this section is to define the most EDM methods used recently to promote the educational environments, by regrouping them into four categories: clas-sification, neural networks, associations rules mining and dimensionality reduction technique.

2.1. Classification

The classification aims to exploit the input instances and to develop a model (accu-rate description for each class) using the knowledge present in the dataset. Basically, the classification algorithm attempts to determine the pattern between the selected attributes from datasets, which would make it possible to develop the classification model and to predict the outcome class. Hence, the key objective of the classification is to develop a model based on some instances' features to describe the class, or one feature to describe a group of classes. Then, the developed model would be used to predict the group of class (output attributes) of new instances from the datasets based on previous instances values.

For example, T. Larrain (Larrain, Bernhard, Mery & Bowyer, 2017) addresses the prob-lem of unconstrained face recognition by proposing a new model called sparse finger-print classification algorithm (SFCA).

They implemented two main phases, a training step to construct representative dic-tionaries by is extracting from each subject's face images a grid of patches. Testing phase to create the fingerprint of the face by extracting the grid from the query image, transforming every path into a binary spare representation. Another work (Pham, Nguyen, Dinh, Nguyen & Ha, 2017), proposes a semi-supervised MLC algorithm to process just the unlabeled data for enhancing the model performance. In the training phase, the proposed algorithm uses specific attributes per prominent group label se-lected by a greedy model implemented in the LIFT algorithm, and unlabeled data consumption mechanism from TESC.

2.2. Neural networks

Neural Network (NN) is an information-processing pattern developed based on the properties of a large number of highly organized neurons working in one union to solve a variety of classification problems in pattern recognition, prediction, optimiza-tion, associative memory, and control. However, in this study we will introduce an advanced version of NN named Deep Learning. Basically, Deep Neural Network (DNN), deep learning or deep machine learning, or deep structured learning is a new big trend in Machine Learning research, which has been introduced based on artificial neural network with the objective of relating machine learning to artificial intelli-gence. First, why deep learning is better from the other data mining techniques? The simple definition is that deep learning refers to neural nets with more than one hidden layer (El Fouki, Aknin & El Kadiri, 2017). The depth of the neural net uses a schema of several layers of nonlinear processing neurons for feature selection/extraction or even transformation. The output of each successive layer is received from the previous layer as input. This feature hierarchy and the learning process in the data are started automatically when deep neural network learn to reconstruct the data.

Also, deep learning is a class of unsupervised machine learning techniques that are based on a learning process of multiple levels of data representations features, which constitutes the majority of the word datasets. Deep learning becomes one of the best machine learning algorithms, because, using just the traditional machine learning techniques we are unable to handle unsupervised data. Cireşan and Meier (Cireşan, Meier, Masci & Schmidhuber, 2012) describe the approach that won the final phase of the German traffic sign recognition benchmark.

By using a fully parameterizable GPU model based on DNN, they could achieve a better-than-human recognition rate of 99.46%. Another paper (Saxe & Berlin, 2015) presents a deep neural network model for malware classification, based on more than 400,000 software binaries.

2.3. Principal component analysis

Principal component analysis (PCA) is a dimensionality reduction technique aims to reduce the dimensionality number of variables in the dataset without sacrificing the classifier accuracy. However, PCA is a linear technique. Therefore, nonlinear associa-tions between the variables of the original datasets may be lost in the phase of prepro-cessing, especially, if the main objective is to further use nonlinear data analysis models on the reduced datasets. When a large number of quantitative variables are studied simultaneously, it is difficult to represent a global graph for all features. The difficulty arises from the fact that the studied individuals are no longer represented in a space of two dimensions, but in a larger dimension space. The PCA method has been extensively applied in several published works including the use of neural net-works as a means of reducing the dimensionality of input variable. Lifang Shang (Shang, Cheng Lv & Yi, 2006) proposes an automatic method to register computed tomography (CT) and magnetic resonance (MR) brain images by using first principal directions of feature images. Another work (Akinina, Akinin, Taganov, Sokolova & Nikiforov, 2016) simplify the classification of pixels of images obtained with the help of unmanned aerial vehicles for the organization received information processing in real time. This experiment proved the effectiveness of such use of preprocessing mode based on principal component analysis. In the same area, another paper (Li, Hu, Liu & Xue, 2015) describe an optimized ANN model for hourly prediction of building electricity consumption. They used in this investigation datasets collected from the Energy Prediction Shootout Contest.

2.4. Association rule mining

Proposed by Agrawal et al. (Agrawal & Srikant, 1994) in 1993. Association rule mining, is one of the important data mining models for discovering interesting rela-tions between attributes in large databases. This may take the form of attempting to recognize which attributes are most powerfully associated with a single attribute of particular interest, or may take the form of trying to find out which associations be-tween any two attributes are strongest. In education association rule mining aims to recognize specific relationships between data. This technique is effective to identify students' failure patterns (Kumar & Chadha, 2012), Planning and scheduling, parame-ters related to the admission process, migration, student modeling, contributions of alumni, pattern between different groups of students (Jindal & Borah, 2013).

When classifying a new object, there may be no decision rules that satisfy the charac-teristics of the classified object. In this case, the default rule will be applied to the classification model, especially in the case of minimum rule systems, which generally contain only a limited number of optimal rules (Ming, Wenying & Xu, 2009). The strategy described in this work ignores the learning instances that have already been processed to generate classifying rules. This technique aims to optimize the learning process and also to keep just the interesting rules. This problem occurs especially when different prediction models are shared by a large proportion of common instances.

3. Experiments and discussion

3.1. Data preparation

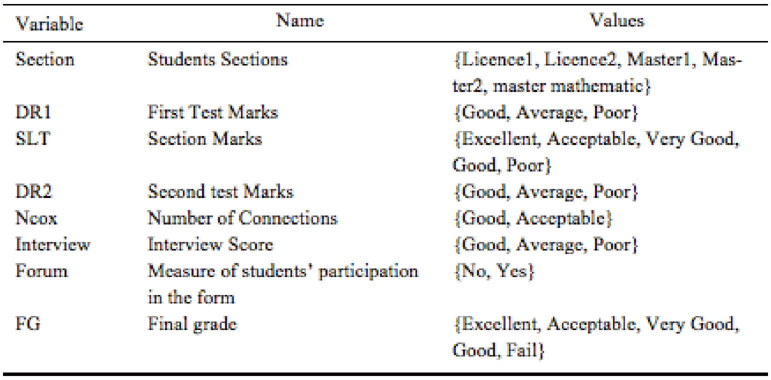
The student's dataset used in this study was obtained from the database of an e-learning platform operated jointly by the University Abdelmalek Essaadi and the University of Picardie Jules Verne for a public of distance learning (Master of Com-puter Applications) from session 2011 to 2015. Then, the finals grade of students in the dataset are classifieds into excellent, acceptable, very good, good or fail based on:

- Students Sections
- First Test Marks

- Selection Marks
- Second Test Marks
- Number of Connections
- Interview Score
- Students' participations in the Forum

3.2. Data selection and transformation

At this stage, only few variables were selected which were required for data mining. Thus, only the principal component variables have been chosen. While some of the information for the attributes was extracted from the database, the preliminary analy-sis revealed inconvenient coding schemes and coding errors. All the predictor input and response classes which were extracted from the database are given in Table 1.



Variable	Name	Values
Section	Students Sections	{Licence1, Licence2, Master1, Master2, master mathematic}
DR1	First Test Marks	{Good, Average, Poor}
SLT	Section Marks	{Excellent, Acceptable, Very Good, Good, Poor}
DR2	Second test Marks	{Good, Average, Poor}
Ncox	Number of Connections	{Good, Acceptable}
Interview	Interview Score	{Good, Average, Poor}
Forum	Measure of students' participation in the form	{No, Yes}
FG	Final grade	{Excellent, Acceptable, Very Good, Good, Fail}

Table 1. Student related variables. Source: Own elaboration.

3.3. Nearest rules and validation process

The nearest-neighbor rule is used here as a decision rule in our pattern classification problem. It entirely relies on the given features and the distance measurement defined to the algorithm. So, it is based on a comparison with everything that has been stored, and without any special prior assumptions about the structures from which the training instances are drawn, the Nearest Neighbor Rules could achieve consistently high performance. After selecting just, the rules we needed, the entropy model starts determining probabilities based on the theorem of decreasing the number of assumptions as possible, rather than the constraints imposed. The constraints varia-bles are derived from the training and validation process which defines the relation-ships between the input features and the outcome. For, example, in text classifica-tion, the highest entropy is the model which selects a class (C) of each and every word (W) for a document (D) in the training and validation dataset.

3.4. Data preprocessing

Preprocessing is employed here as a means to convert the raw measurement values into more compact and more abstract information that the models can more easily handle, and to speed up the convergence process. Therefore, dimensionality reduction techniques and feature selection, are implemented in this study, because they compact the variables / attributes in the datasets while attempting to conserve the greatest amount of information and to preserve the classifier accuracy. For Example, when preprocessing is used because of the related qualities with pixels the learning rate increments significantly. Thus, we examine the extent to which PCA and wrapper approach aid the diagnostic capabilities of the DNN and nearest rules models.

3.5. Network training and validation process

The next step in modeling the neural network model is the resolution of the number of processing items and hidden layers in the network. Choosing the number of pro-cessing items and hidden layers isn't easy, because having a limited number of hid-den layers in a neural network lowers the processing capability of the network. Simi-larly, a large number of processing items will progressively slow down the training process (El Fouki, Aknin & El Kadiri, 2019).

In determining the number of hidden layers to be used, we could cite two approaches: growing approach which begins with a small network and then increase its size, the other method is pruning approach which starts with a large network and then reduce its size by removing useless items. The Growing Method was applied in the model-ing of the neural network model. Therefore, the investigation requires initializing with no hidden layers and then gradually increasing them.

3.6. Results and test process

In order to have meaningful results it was imperative to test our hypothesis on a real-world problem, so, the prediction model has been developed based on certain input variables selected from a database of an e-learning platform operated jointly by the University Abdelmalek Essaadi and the University of Picardie Jules Verne for a pub-lic of distance learning. In this classification strategy, we automatically select features in our data that are most useful and most relevant for the problem we are working on, then, a set of rules is used to classify new items (instances not present in the learning process) by matching these instances to condition parts of the rules. This is a process called feature selection it should not overfeed the training data and lose the ability to generalize the unseen data. In such process, the sets of rules can be either some orga-nized rules (priority list) or unordered. Therefore, the matching starts as long as we get the first rule which is used to classify a new instance and the remaining rules are ignored. The default rule is always the last one, and it is used if no other rule has been associated. The overall results of classifier performance on our dataset are shown in Figure 1. Nearest rules model has the best performance with 93.14% accuracy.

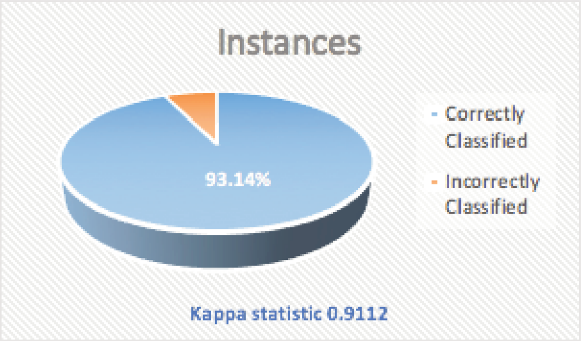


Figure 1. Instances correctly classified. Source: Own elaboration.

The second algorithm described here starts by preprocessing the ENS data, so it is preceded by preprocessing using Principal Component Analysis, followed by trans-forming the highly correlated input variables to a small number of orthogonal varia-bles and improve the training speeds achieved due to the extensive reduction in the number of parameters. Input layer neurons are being connected to the hidden layer neurons through weighted (random generated weight) connections. The data were then split into three datasets; training, validation and tests. The training set was used to train the DNN, the training process and the test set was used to evaluate the models' performance after achievement of the training process.

The goods of applying the PCA method to a DNN algorithm were examined to pre-dicting the student's performance based on ENS platform data. For ease of evaluation, the final grade has been normalized within excellent, acceptable, very good, good and fail. For the investigation purpose, two types of transformations were used, either the Fourier transforms coefficients or principal component scores. The overall results of classifier performance on our dataset are shown in Figure 2. DNN has the best per-formance with 91.1% accuracy.

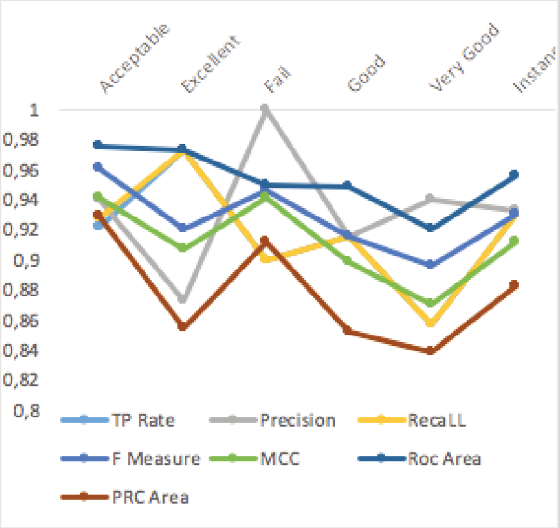


Figure 2. Results obtained for each class. Source: Own elaboration.

As for the Receiver Operating Characteristics area (AROC) measurement is one of the most important measures for evaluating a classier. An "excellent" classifier will have ROC area measures approaching 1 with 0.5 being related to "random guessing," the same as a classifier with Kappa statistics of 0.

The Margin curve defines points illustrating the prediction margin. The margin is described as the variation between the output probability predicted for the selected class and the highest output probability predicted for the other classes (Yang, Fong, Sun & Wong, 2012). Figure 3 describes, the most of the instances are correctly clas-sified by deep neural network model, since they are organized in the area of probabil-ity one (the right part of the graph).

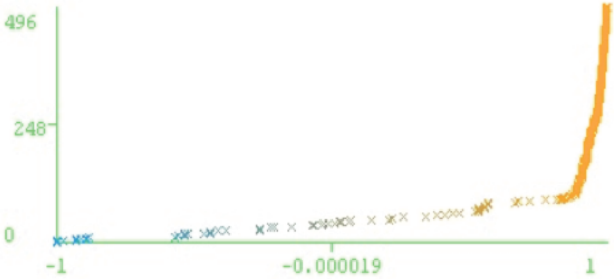


Figure 3. Margin Curve of DNN model. Source: Own elaboration.

The performance measures precision, TP Rate, recall, and F-measure in figure 4 de-termine how our hybrid algorithm, the rule-based, and the DPCA methods perform in identifying instance classes. The performance measures "accuracy" determines the number of well-classified instances, as an absolute value, or a percentage of the total examples number. The mean absolute error is the difference between the probability (calculated by the classifier) of an instance class, and its initial probability class, which has been fixed in the dataset. The sum of these errors is then divided by the number of instances. It is clear that our algorithm had the best accuracy, by compar-ing all the five measures for the three classifiers we distinct advantage of the hybrid developed models.

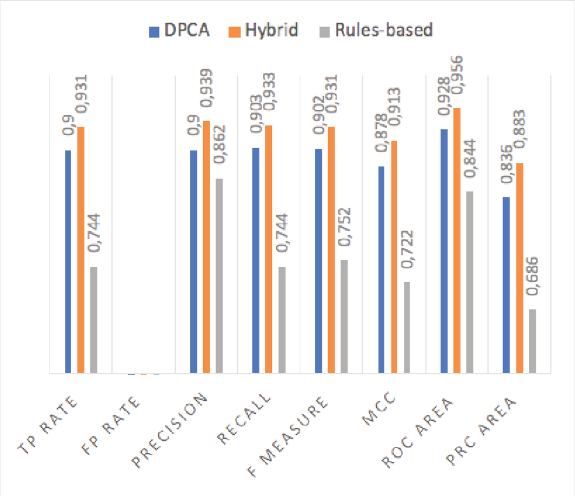


Figure 4. Comparison of the results obtained by the three classifiers over INES Data. Source: Own elaboration.

4. Conclusion

This study shows that the two proposed algorithms achieve better results than the usually used techniques in the literature. Therefore, the deep principal component analysis and wrapper approach could improve the classification models and even in-crease the prediction performance of a deep neural network algorithm by reducing the dimensionality number of variables of datasets and by optimizing the associated rules without sacrificing the classifier accuracy and reliability. Therefore, preprocessing the data before the deep learning process helps in reducing its dimension which will de-crease the correlated/related inctances caused by overlapping input instances, which will reduce the network training time and improve the performance of the classifica-tion systems, particularly in solving complex problems involving a large number of input data.

The finals grade of students in datasets are classifieds into excellent, very good, good, acceptable and fail classes on the basis of their section, first test marks, selection marks, second test marks, number of connections to the platform, interview score and measure of their participation in the forum. The result of classification can be used to take remedial actions on students' learning process.

Cómo citar este artículo / How to cite this paper

El Fouki, M.; Aknin, N.; El Kadiri, K. E. (2019). Towards an improved classification model based on Deep Learning and nearest rules strategy. *Campus Virtuales,* 8(1), 111-119. (www.revistacampusvirtuales.es)

References

Abdelouarit, K.; Sbihi, B.; Aknin, N. (2015). Big Data at the Service of Teaching and Scientific Research within the UAE. J. Educ. Vocat. Res., 6(4), 72-75.

Agrawal, R.; Srikant, R. (1994). Fast algorithms for mining association rules. 94 Proc. 20th Int. Conf. Very Large Data Bases, 1215, 487-499.

Akinina, N. V.; Akinin, M. V.; Taganov, A. I.; Sokolova, A. V.; Nikiforov, M. B. (2016). Neural network implementation of a principal component analysis tasks on board the unmanned aerial vehicle information processing in real time. In 5th Mediterranean Conference on Embedded Computing (MECO) (pp. 326-330).

Boulaajoul, M.; Aknin, N. (2019). The Role of the Clusters Analysis Techniques to Determine the Quality of the Content Wiki. Int. J. Emerg. Technol. Learn., 14(01), 150.

Brusilovsky, P.; Peylo, C. (2003). Adaptive and Intelligent Web-based Educational Systems adaptative and intelligent technologies for web-

based educational systems. Int. J. Artif. Intell. Educ., 13, 156-169.
Cireşan, D.; Meier, U.; Masci, J.; Schmidhuber, J. (2012). Multi-column deep neural network for traffic sign classification. Neural Networks, 32, 333-338.
Conati, C.; Gertner, A.; Vanlehn, K. (2002). Using Bayesian Networks to Manage Uncertainty in Student Modeling. User Model. User-adapt. Interact., 12(4), 371-417.
El Fouki, M.; Aknin, N.; El Kadiri, K. E. (2017). Intelligent Adapted e-Learning System based on Deep Reinforcement Learning. In Proceedings of the 2nd International Conference on Computing and Wireless Communication Systems - ICCWCS'17 (pp. 1-6).
El Fouki, M.; Aknin, N.; El Kadiri, K. E. (2019). Multidimensional Approach Based on Deep Learning to Improve the Prediction Performance of DNN Models. Int. J. Emerg. Technol. Learn., 14(02), 30.
Golding, A. R.; Rey, M.; Rosenbloom, P. S. (1991). Improving Rule-Based Systems through Case-Based Reasoning. Knowl. Creat. Diffus. Util.
Hidalgo Cajo, B. G. (2018). Minería de datos en los Sistemas de Gestión de Aprendizaje en la Educación Universitaria. Campus Virtuales, 7(2), 115-128.
Ibrahim, Z.; Rusli, D. (2007). Predicting Students' Academic Performance: Comparing Artificial Neural Network, Decision tree And Linear Regression. In Proc. 21st Annu. SAS Malaysia Forum (pp. 1-6).
Jindal, R.; Borah, M. D. (2013). A survey on educational data mining and research trends. Int. J. Database Manag. Syst., 5(3), 53.
Kumar, V.; Chadha, A. (2012). Mining association rules in student's assessment data. Int. J. Comput. Sci. Issues, 9(5), 211-216.
Larrain, T.; Bernhard, J. S.; Mery, D.; Bowyer, K. W. (2017). Face Recognition Using Sparse Fingerprint Classification Algorithm. IEEE Trans. Inf. Forensics Secur., 12(7), 1646-1657.
Li, K.; Hu, C.; Liu, G.; Xue, W. (2015). Building's electricity consumption prediction using optimized artificial neural networks and principal component analysis. Energy Build., 108, 106-113.
Martínez, D. D. (2017). Profesorado en formación y ambientes educativos virtuales. Campus Virtuales, 6(2), 69-78.
Ming, H.; Wenying, N.; Xu, L. (2009). An improved Decision Tree classification algorithm based on ID3 and the application in score analysis. In Chinese Control and Decision Conference (pp. 1876–1879).
Oladokun, V. O.; Adebanjo, A. T.; Charles-Owaba, O. E. (2008). Predicting students' academic performance using artificial neural network: A case study of an engineering course. Pacific J. Sci. Technol., 9(1), 72-79.
Pham, T. N.; Nguyen, V. Q.; Dinh, D. T.; Nguyen, T. T.; Ha, Q. T. (2017). MASS: A Semi-supervised Multi-label Classification Algorithm with Specific Features.
Romero C.; Ventura, S. (2010). Educational data mining: A review of the state of the art. IEEE Trans. Syst. Man, Cybern. C Appl. Rev., 40(X), 601-618.
Saxe, J.; Berlin, K. (2015). Deep Neural Network Based Malware Detection Using Two Dimensional Binary Program Features.
Scheuer, O.; McLaren, B. M. (2011). Educational Data Mining. Encycl. Sci. Learn.
Shang, L.; Cheng Lv, J.; Yi, Z. (2006). Rigid medical image registration using PCA neural network. Neurocomputing, 69(13-15), 1717-1722.
Sleeman, D.; Brown, J. S. (1982). Intelligent tutoring systems.
Yang, H.; Fong, S.; Sun, G.; Wong, R. (2012). A Very Fast Decision Tree Algorithm for Real-Time Data Mining of Imperfect Data Streams in a Distributed Wireless Sensor Network. Int. J. Distrib. Sens. Networks, 8(12), 863545.