A Data Mining of Supervised learning Approach based on K-means Clustering

Bilal Sowan[†] and Hazem Qattous[††]

[†]Department of Computer Network Systems, Applied Science Private University, Amman, Jordan
[††]Department of Computer Information Systems, Applied Science Private University, Amman, Jordan

Summary
A diversity of application fields include a massive number of datasets. Each dataset consists of a number of variables (features). One of these variables that is considered as a dependent variable (target variable) and is used for prediction in data mining of the supervised learning task. Data mining is necessary for building an automatic analysis in order to extract knowledge from datasets. Knowledge extraction is useful for recommendation system and decision making which can be accomplished by data mining tasks. Different data types and characteristics of dependent variable play an important role in selecting such a specific data mining task. One of the most challenging issues in the data mining research is selecting the most appropriate and perfect technique for a particular dataset. This paper proposes a supervised learning approach by utilizing k-means clustering in order to convert a regression task into a classification task. The proposed approach is a flexible data mining approach that employs variety techniques. The flexibility means that a dependent variable of a numeric data type in a dataset is not only considered for a regression task. Instead, the approach is also able to apply the same dataset in the classification task by categorizing dependent variable into class labels. The experimental results validate the application of the proposed approach using two datasets. The first dataset is CPU dataset from UCI repository datasets, while the second one is a road traffic dataset from a real-world domain. The results show the effectiveness of the proposed approach that integrates different techniques namely MLP, REPTree, and CART, which are widely used for both classification and regressions tasks. The results also demonstrate that by clustering the dependent variable from numeric values into class labels can produce high accuracy for the used datasets.

Key words:
Data Mining, Regression; Classification, K-means clustering, Clustering, Supervised learning.

1. Introduction

Nowadays, a wide range of fields produce a large number of databases of different varieties. Database implicitly hide an important knowledge, and it is difficult to obtain this knowledge manually. Hence, applying knowledge discovery and data mining techniques as an automatic tool to extract knowledge is a proper solution. Extracted knowledge helps in many aspects. One of the important aspect is to assist a human in making a decision [1, 2]. An extracted knowledge can be in different forms such as rule based and mathematical formulas. Data mining is an emerging research topic for a requirement to develop a significant technique that is to deal with such data variety [3]. Thus, nature of a dataset variables (features or attributes including data types) determines the applied technique.

Data type of dataset variables can be either discrete or continuous values. Discrete variables include categorical, whereas continuous or numeric variables include integers or real values. Categorical can be binary, nominal or ordinal [4].

In general, data mining tasks categorized into two main types. These types are descriptive and predictive tasks. The descriptive tasks concern about finding the correlation between data attributes (features), whereas the predictive tasks are intended to predict (forecast) a future value or unknown value of a dependent variable (target variable). The descriptive task includes many tasks such as: summarization, clustering, and association rules mining. The predictive task includes many tasks such as: classification and regression [2, 3]. Formally, every dataset consists of a number of data objects o_1, o_2, ..., o_i (called records or instances). These objects belong to different variables. Let x_1, x_2, ..., x_n be an input (independent) variables and y_1, y_2, ..., y_m be an output (dependent) variables. Then, the goal of a predictive task is to construct a model that is able to predict a future value of a dependent variable either the value is a numeric or categorical. As a result, if the value of a dependent variable is a numeric, then it is called a regression task, on the other hands, if the value of a dependent variable is categorical (i.e. represented by class labels c_1, c_2, ..., c_j) then it is called a classification task.

Basically, once a data mining model is constructed, if a dependent variable exists in a dataset, then this concept is called a supervised learning. Otherwise, if a dependent variable does not exist in a dataset, then this is called unsupervised learning. Commonly, unsupervised learning employs a descriptive data mining tasks, while a supervised learning apply a predictive data mining tasks [2, 5]. The following data mining tasks are described below:

Manuscript received January 5, 2017
Manuscript revised January 20, 2017

- Association rule is a descriptive data mining task aims at discovering a correlation (a relationship) between data variables. A well-known paradigm of adopting association rule techniques is a market basket analysis for studying customers' behavior [1, 6].
- Clustering is a descriptive data mining task aims at grouping data instances into shared characteristics. In other words, the grouping instances is to maximize the similarity between data instances inside a group and maximize the difference between data instances of different groups [1, 6, 7].
- Regression is a predictive data mining task aims at predicting (forecasting) a future value or unknown value of a dependent variable. A data type of a dependent variable should be a quantitative (numeric) variable [1, 2, 7].
- Classification is a predictive data mining task aims at predicting (forecasting) a future value or unknown value of a dependent variable. A data type of a dependent variable should be a categorical variable and each value belongs to this variable is called a class label [1, 6, 7].

Several data mining algorithms are used in the previous studies for building prediction models. These algorithms are Fuzzy Neural Networks (FNN) [8], Support Vector Machine (SVM) [9], Artificial Neural Network (ANN) [9, 10], Classification and Regression Tree (CART) [9, 11], ANN and K-means clustering [12], C4.5 [13] and Reduced Error Pruning Tree (REPTree) [14]. Multi-Layer Perceptron (MLP) is the well-known ANN technique which can be used for both, classification and regression problems. ANN is a non-linear function based on neuron biological inspiration [15]. ANN is used to build a prediction model from a complex (nonlinear) relationship between input and output variables [12].

CART is one of the popular decision tree techniques introduced by Breiman et al. [16]. CART basically constructs a model based on a binary tree. A dataset consists of a number of instances with different variables. These variables have independent (input) variables that are used to predict a dependent (output) variable. A tree starts with a root node of an independent variable, then each node is split recursively to reach a leaf node. A leaf node of the tree represents a dependent variable, which is helped for prediction.

REPTree is a simple and fast decision tree technique based on pruning method. The pruning is used in order to reduce error [17]. REPTree applied an information gain method as a decision criterion for splitting a tree [18]. REPTree is based on C4.5 algorithm [19] which is able to build a decision tree for both, classification and regression tasks [14].

K-means clustering is one of the most commonly used techniques for clustering. It is unsupervised learning technique for partitioning dataset instances into groups (number of *k* clusters). Partitioning is performed randomly by calculating a center for each cluster. In other words, cluster centers are initially selected. Then, iteratively each data instance is assigned to the adjacent center (i.e. each data instance belongs to the closest mean cluster)[20, 21].

ANN, REPTree, and CART are the most effective supervised learning techniques, which applied in data mining research community [13, 17]. Despite of many studies depicted classification and regression tasks, a few concentrated on building an approach that is able to deal with both, classification and regression tasks, on the same dataset. This study addresses very important point in prediction by the capability of transforming a numeric dependent variable of a dataset into class labels by adopting clustering. The aim of this paper is to study the effect of clustering a numeric dependent variable into class labels. In this paper a supervised learning approach is constructed. The approach is able to deal with both, regression and classification tasks, on the same datasets. A number of data mining techniques are used to build the proposed approach namely, MLP, REPTree, and CART. Clustering attempt to convert a numeric dependent variable into class labels using k-means clustering technique. The proposed approach is evaluated using two datasets. First, CPU dataset that is collected from the UCI Machine Learning Repository. Second, Total Fuel Consumption (TFC) dataset that is selected from a real-world road traffic domain. The following are the contributions of this paper:

- Study and examine the effect of k-means clustering technique in determining a data class labels.
- Combine the regression and classification tasks to be applied in the same datasets.
- Build a capable data mining approach that incorporates and supports regression and classification tasks on road traffic dataset of a real-word domain.
- Design experiments and conduct evaluation on two datasets namely CPU and TFC.
- Compare between regression and classification task on the datasets using different performance measures. In addition, study the importance of the integration between regression and classification in one approach.

The remainder of this paper is organized as follows. Section 2, provides a related works. Section 3, explores the proposed approach, describes the datasets and illustrates the performance measures. Section 4, presents experimental results evaluation. Finally, the conclusions are provided in Section 5.

2. Related Works

Many studies have been demonstrated the importance of using clustering techniques to improve prediction performance [21, 22]. Banitaan et al. [21] presented an approach for improving classification accuracy by decomposing and partitioning (clustering) data instances of each class label into subclasses. That means, the result of clustering is considered as a new subclass label for each data instances. Kou et al. [20] suggested an approach based on a Multiple Criteria Decision Making (MCDM) to evaluate clustering techniques quality. The approach is based on the evaluation and selection of clustering techniques among well-known techniques after ranking. The approach has been applied in financial risk analysis field. Ashfaq et al. [23] proposed an approach that is able to identify and calculate unlabeled instances (unclassified instances) based on fuzziness method. The unlabeled instances are resulting from misclassification in supervised learning algorithm tasks (classification task). The proposed approach aimed at enhancing the classification performance on intrusion detection field. The results showed that using the proposed approach abled to enhance classification performance. The comparison has been conducted with classification techniques such as Naïve Bayes, SVM, and Random Forests.

Al Snousy et al. [14] constructed models that compared different classification decision tree algorithms. The models are applied on cancer gene expression datasets. The outcomes of this study have been confirmed that decision tree algorithms produced a satisfactory results. The importance of decision tree algorithms due to their classification accuracy and interpretability. Wang et al. [24] proposed a prediction model in railway crossover systems domain. The model is based on an integration of a modified Bayesian network algorithm and Monte Carlo simulation. The prediction model has been carried out to find a relation between the weather and inability of rail switching (railway turnouts). The proposed model has been tested on real dataset and compared with other prediction algorithms such as SVM, ANN, and AdaBoost.RT.

3. Proposed Approach

This paper proposes a supervised learning approach to deal with both, regression and classification tasks. The approach is based on clustering technique. A variety of data mining techniques are employed in the proposed approach. These techniques are: MLP, REPTree and CART. The purpose behind using these techniques is that they can be adopted in both, classification and regression prediction tasks.

The aim of the proposed approach is to accept several dataset regardless of its data type dependent variable characteristics. The data type dependent variable either numeric or categorical is accepted. This section provides a details of the proposed approach steps, dataset description, and performance measures. The following steps illustrates the proposed approach as shown in Fig. 1.

Step 1: Select a dataset in a particular domain.

Step 2: Clean a dataset by determining outliers and handling missing values.

Step 3: Determine a data mining task either regression or classification.

Step 4: Examine a data type of a dependent variable for each dataset. The following cases can occur:

- If a data mining task is regression and a data type of a dependent variable is numeric. Then a prediction is applied.
- If a data mining task is classification and a data type of a dependent variable is categorical. Then a prediction is applied.
- If a data mining task is classification and a data type of a dependent variable is numeric. Then k-means clustering technique is adopted to transform the dependent variable data type into a categorical values as shown in Fig. 2. In clustering, each numeric value is replaced with a class label that belongs to a specific cluster. As a result, a prediction is then applied.

Step 5: Assess the prediction quality of the proposed approach using different performance measures.

3.1 Dataset

This subsection describes two datasets. The first dataset, called CPU, was collected from UCI Machine Learning Repository datasets [25]. The second one was generated using a simulation model for road traffic (METANET model) [26]. Table 1 shows summary of the datasets structures. The details description of the datasets used in this study are explained as follows:

- The first dataset, called CPU, consists of 209 instances and 9 integer variables.
- The second dataset in road traffic field, consists of 5000 instances and 5 real-values variables. The road traffic dataset variables represent the following (shown in Fig. 3):
  - Traffic flow (the number of vehicles per hour) represented by the main road and the service road. Traffic flow consists of traffic density 1 and traffic density 2.
  - Traffic demand (the number of vehicles that is required to use the traffic network) for the main road and service road. Traffic demand

consists of traffic demand 1 and traffic demand 2.

- The Predicted Total Fuel Consumption (TFC) (liters required for a vehicle to cross the traffic network).

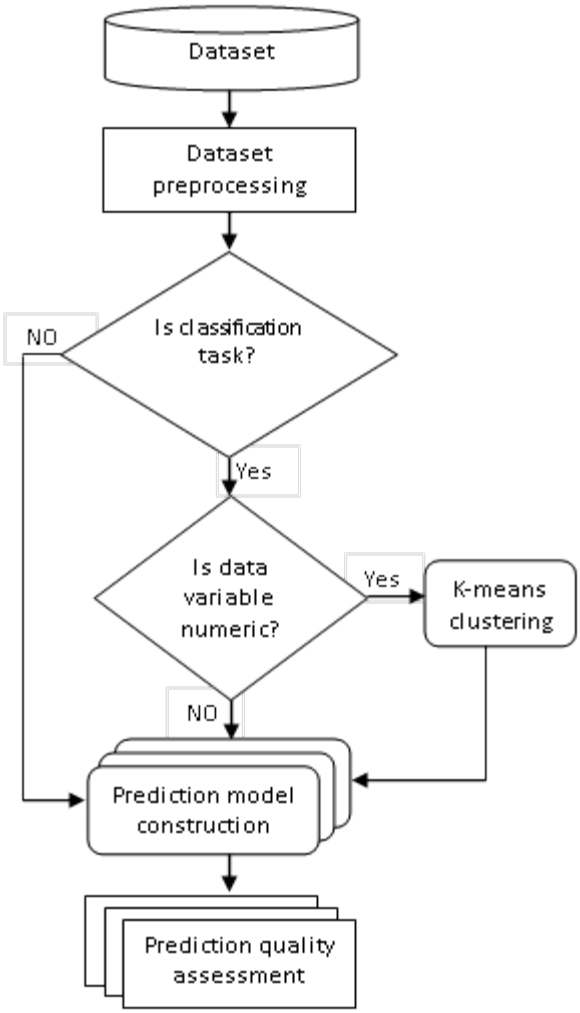

Fig. 1 The proposed approach.

Table 1: Summary of datasets structures.

Dataset	Number of instances	Number of variables	Data type
CPU	209	9	Integer
TFC	5000	5	Real

Input: Dataset $D_i = (D_{ix_1}, \dots, D_{ix_n}, D_{iy_1}, \dots, D_{iy_m})$ where x is independent variable, y is dependent variable, k is number of clusters.
Output: Categorize y
Method:
for each $D_{iy_1}, \dots, D_{iy_m}$ of each D_i
if D_{iy_j} is numeric, where $j = 1,2, \dots, m$
Categorize each D_{iy_j} using k-means clustering based on k
endif
endfor

Fig. 2 Categorization algorithm.

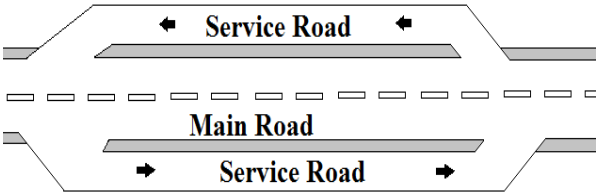

Fig. 3 The sub-network of road traffic.

3.2 Performance Measures

Several performance measures are applied in previous studies. This study selected Relative Absolute Error (RAE) measure because it is one of the rare measures that compares regression and classification techniques. This helps in performing a comprehensive validation of the proposed approach. The other selected measure is the correlation coefficient which measures the model fitting. These measures are described in Eq. (1) [27] and Eq. (2) [28]:

$$\text{Relative Absolute Error (RAE)} = \frac{\sum_{i=1}^{n}|\hat{y}_i - y_i|}{\sum_{i=1}^{n}|\bar{y} - y_i|} * 100\% \quad (1)$$

$$\text{Correlation coefficient} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}} \quad (2)$$

where, y: is the actual value of a dependent variable, $\hat{y}$: is the predicted value of a dependent variable, $\bar{y}$: is the mean of actual value, $\bar{\hat{y}}$: is the mean of predicted value, n: is the total number of data instances.

The correlation coefficient is used to find the relationship between predicted and actual dependent variable in regression tasks. Its value ranges from -1 to 1. If its value is 0, this indicates that there is no correlation at all between the actual and predicted values. If its value is -1, this

indicates the negative correlation between the actual and predicted values. On the contrast, the actual and predicted values are the same, if the correlation value is 1.

4. Experimental Results

The empirical study applies three data mining techniques, and select two datasets in different fields to validate and evaluate the proposed approach. The prediction performance for both, regression and classification tasks, is performed using 10 folds cross-validation method. Each dataset in cross-validation method is divided into training and testing sets for better evaluation results. All experiments conducted in this study were performed using WEKA data mining tool [29] and MATLAB platform.

The proposed approach has employed a widely used techniques in order to justify the prediction performance. The techniques are MLP, REPTree and CART. The key behind using these techniques in the proposed approach is due to their ability to work in both, regression and classification data mining tasks. MLP has a powerful ability in learning process. REPTree is a promising and effective technique for prediction with good accuracy results [17, 30]. CART is based on building a decision tree that produces rules. CART also can be operated very well in prediction for both small and large datasets [11]. Furthermore, REPTree and CART are based on a decision tree which produce rules. These rules are used for better interpretability and understandability a domain field.

Fig. 4 shows the correlation coefficient for CART, MLP, and REPTree techniques that were applied on CPU dataset of numeric dependent variable. It can be noticed that all techniques produced high values. The minimum value produced by REPTree was (0.82) while the maximum value produced by CART was (0.93).

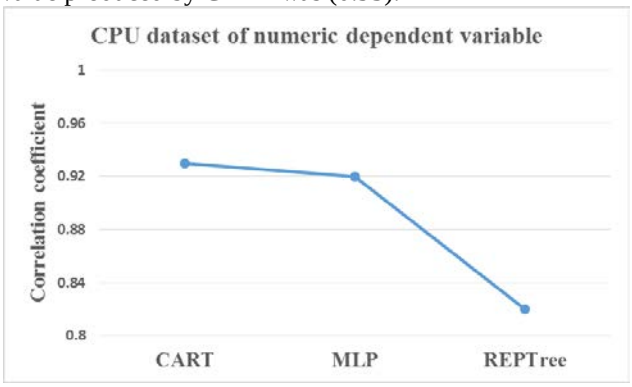

Fig. 4 Correlation coefficient of all techniques on CPU dataset.

Fig. 5 shows the correlation coefficient for CART, MLP, and REPTree techniques that were applied on TFC dataset of numeric dependent variable. It can be seen that all techniques also produced very high values. MLP produced the highest value (0.99), as well as CART and REPTree produced very high value (0.98) and (0.97) respectively. In other words, the techniques used in this study are highly predictive and perfect models on CPU and TFC datasets.

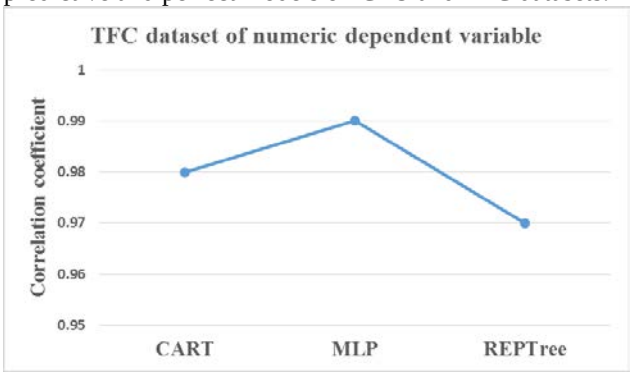

Fig. 5 Correlation coefficient of all techniques on TFC dataset.

Fig. 6 and Fig. 7 depict the results after clustering the dependent variable in each dataset. The clustering means that converting a dependent variable data type from numeric to categorical values (class labels). In the clustering step, many experiments are constructed to select most appropriate number of clusters that produces a satisfactory results. The number of clusters *k* is varying from 2 to 5 clusters.

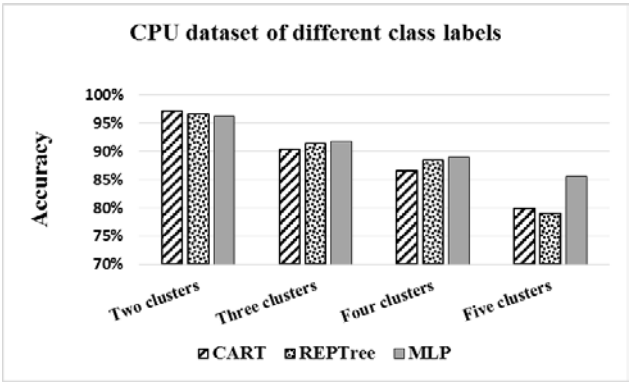

Fig. 6 Classification accuracy of all techniques on CPU dataset.

It can be observed from Fig. 6 that the case of "two clusters" (*k*=2) obtained the highest accuracy. All three techniques show almost the same classification accuracy when (*k*=2) for CPU dataset. MLP, CART and REPTree produce (96%), (97%), and (97%) respectively. Upon close observation, classification accuracy is decreased gradually when using a number of clusters more than two clusters (three, four, and five clusters). This can be explained as the clustering of a numeric values into a categorical values is accepted up to a valid number of clusters. Furthermore, increasing the number of clusters will be affected on a dependent variable homogeneity. It

can be noticed in Fig. 6 that the difference in classification accuracy between a number of clusters (k=2) (highest classification accuracy) and a number of clusters (k=5) (lowest classification accuracy) is approximately (10%) for MLP, and is approximately (17%) for CART and REPTree on CPU dataset.

It can be noticed from Fig. 7 that two clusters (k=2) also obtained the highest accuracy. MLP technique produced the highest classification accuracy (99%) for TFC dataset. Furthermore, CART and REPTree produced (93%) and (92%) respectively. Similarly, the classification accuracy is decreased gradually when using a number of clusters more than two clusters (three, four, and five clusters). It is also observed that MLP produced the highest classification accuracy compared to other techniques when a number of clusters are three, four, and five clusters.

It can be found in Fig. 7 that the difference in classification accuracy between a number of clusters (k=2) (highest classification accuracy) and a number of clusters (k=5) (lowest classification accuracy) is approximately (10%) for MLP, and is approximately (12%) for CART and REPTree on TFC dataset. However, the difference for both CART and REPTree is approximately (5%) between two datasets, which is a bit higher than MLP. This variation in the difference is justified by the difference in data type characteristics and datasets sizes between two datasets used in this study. It can be concluded that MLP has generated a consistent results when selecting a different number of clusters on different datasets. It also can be concluded that the best result is obtained when a number of clusters is set to two clusters (k=2) for all techniques on both datasets.

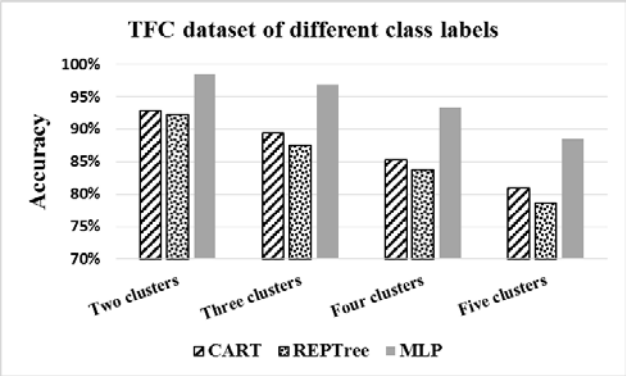

Fig. 7 Classification accuracy of all techniques on TFC dataset.

Table 2 indicates the result of RAE measure for all techniques on TFC dataset. It is clear that, the classification task provides a lower value of RAE for all techniques compared to the regression task. This indicates that the classification task produces a better results than the regression task. This can be explained by two reasons; firstly, a difference in data type characteristics of dependent variable. Secondly, the effectiveness of the proposed approach using clustering of a dependent variable, which provides a better results.

Table 2: The result of RAE measure for all techniques on TFC dataset.

Performance measures/techniques	RAE in regression	RAE in classification
CART	20.16%	18.73%
MLP	12.93%	5.05%
REPTree	22.66%	22.47%

5. Conclusion and Future Work

This paper proposes a supervised learning approach that utilizes k-means clustering technique by categorizing the dependent variable of numeric values in a dataset. After employing clustering, each numeric value is clustered (grouped) into a specific cluster that is used to replace its value with that cluster as labeling value (class label). The approach was constructed based on MLP, REPTree and CART that are used for both, classification and regression tasks. The performance of the proposed approach was assessed using a collection of performance measures. The proposed approach was validated using experiments on two datasets. First dataset, CPU dataset, was collected from UCI repository. Second dataset is a real dataset in a road traffic domain. Results show that the proposed approach achieved a satisfactory high accuracy. The use approach also calculated a lower RAE value in classification task compared to RAE value in regression task on both datasets. In addition, the best results were achieved in classification when the number of clusters k is set to (k=2) for both datasets. Future research direction will include further investigation of other classification, regression and clustering techniques. Moreover, clustering validity to determine the suitable number of clusters will be investigated. Considering independent variables in clustering step to determine a class label can also be one of the future work.

Acknowledgments

The authors are grateful to the Applied Science Private University, Amman, Jordan, for the full financial support granted to this research.

References

[1] Goebel, M. and L. Gruenwald, A survey of data mining and knowledge discovery software tools. ACM SIGKDD explorations newsletter, 1999. 1(1): p. 20-33.

[2] Peña-Ayala, A., Educational data mining: A survey and a data mining-based analysis of recent works. Expert systems with applications, 2014. 41(4): p. 1432-1462.

[3] Mitra, S., S.K. Pal, and P. Mitra, Data mining in soft computing framework: a survey. IEEE transactions on neural networks, 2002. 13(1): p. 3-14.
[4] PHUA, C., et al., A Comprehensive Survey of Data Mining-based Fraud Detection Research. Artificial Intelligence Review, 2005: p. 1-14.
[5] Chakrabarti, S., Data mining for hypertext: A tutorial survey. ACM SIGKDD Explorations Newsletter, 2000. 1(2): p. 1-11.
[6] Tsai, C.-W., et al., Data mining for internet of things: a survey. IEEE Communications Surveys & Tutorials, 2014. 16(1): p. 77-97.
[7] Ngai, E., et al., The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision Support Systems, 2011. 50(3): p. 559-569.
[8] Dahal, K., et al., GA-based learning for rule identification in fuzzy neural networks. Applied Soft Computing, 2015. 35: p. 605-617.
[9] Huang, C.-L., M.-C. Chen, and C.-J. Wang, Credit scoring with a data mining approach based on support vector machines. Expert systems with applications, 2007. 33(4): p. 847-856.
[10] Islam, S., et al., Empirical prediction models for adaptive resource provisioning in the cloud. Future Generation Computer Systems, 2012. 28(1): p. 155-162.
[11] Razi, M.A. and K. Athappilly, A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. Expert Systems with Applications, 2005. 29(1): p. 65-74.
[12] Capozzoli, A., F. Lauro, and I. Khan, Fault detection analysis using data mining techniques for a cluster of smart office buildings. Expert Systems with Applications, 2015. 42(9): p. 4324-4338.
[13] Wu, X., et al., Top 10 algorithms in data mining. Knowledge and information systems, 2008. 14(1): p. 1-37.
[14] Al Snousy, M.B., et al., Suite of decision tree-based classification algorithms on cancer gene expression data. Egyptian Informatics Journal, 2011. 12(2): p. 73-82.
[15] Delen, D., G. Walker, and A. Kadam, Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine, 2005. 34(2): p. 113-127.
[16] Breiman, L., et al., Classification and regression trees. 1984: Wadsworth International Group, Belmont, California.
[17] Nisanci, M., et al., The prediction of the electric field level in the reverberation chamber depending on position of stirrer. Expert Systems with Applications, 2011. 38(3): p. 1689-1696.
[18] Zhao, Y. and Y. Zhang, Comparison of decision tree methods for finding active objects. Advances in Space Research, 2008. 41(12): p. 1955-1959.
[19] Quinlan, J.R., C4. 5: Programming for machine learning. Morgan Kauffmann San Mateo, CA, USA 1993.
[20] Kou, G., Y. Peng, and G. Wang, Evaluation of clustering algorithms for financial risk analysis using MCDM methods. Information Sciences, 2014. 275: p. 1-12.
[21] Banitaan, S., A.B. Nassif, and M. Azzeh. Class Decomposition Using K-Means and Hierarchical Clustering. in 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA). 2015. IEEE.
[22] Japkowicz, N. Supervised learning with unsupervised output separation. in International Conference on Artificial Intelligence and Soft Computing. 2002.
[23] Ashfaq, R.A.R., et al., Fuzziness based semi-supervised learning approach for intrusion detection system. Information Sciences, 2016. 378: p. 484–497.
[24] Wang, G., et al., A Bayesian network model for prediction of weather-related failures in railway turnout systems. Expert Systems with Applications, 2017. 69: p. 247-256.
[25] Lichman, M. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. 2013.
[26] Messner, A. and M. Papageorgiou, METANET: A macroscopic simulation program for motorway networks. Traffic Engineering & Control, 1990. 31(8-9): p. 466-470.
[27] Öztürk, A., S. Kayalıgil, and N.E. Özdemirel, Manufacturing lead time estimation using data mining. European Journal of Operational Research, 2006. 173(2): p. 683-700.
[28] Quek, C., M. Pasquier, and B.B.S. Lim, POP-TRAFFIC: a novel fuzzy neural approach to road traffic analysis and prediction. IEEE Transactions on Intelligent Transportation Systems, 2006. 7(2): p. 133-146.
[29] Hall, M., et al., The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 2009. 11(1): p. 10-18.
[30] Natek, S. and M. Zwilling, Student data mining solution–knowledge management system related to higher education institutions. Expert systems with applications, 2014. 41(14): p. 6400-6407.

Dr. Bilal Sowan is currently an assistant professor at the Faculty of Information Technology, Applied Science Private University, Amman, Jordan. Dr. Sowan holds a Ph.D. degree in Computing from University of Bradford, UK. His research interests are in data mining and human computer interaction.

Dr. Hazem Qattous is currently an assistant professor at Applied Science Private University in Amman, Jordan. Dr. Qattous holds a Ph.D. degree in Computing from Glasgow University, UK. His research interests are in Human-Computer Interaction.