Credit Card Clients Classification Using Hybrid Guided wheel with Particle Swarm Optimized for Voting Ensemble

Khadija Shazly [*,1], Nima Khodadadi [2]

[1] Faculty of Computer and Information, Mansoura university, Egypt
[2]University of Miami, 1251 Memorial Drive, Coral Gables, 33146, FL, USA.
Emails: khadijashazly@students.mans.edu.eg; nxk682@miami.edu

Abstract

Credit card use is rapidly increasing as a result of the widespread availability of these cards, the ease of making electronic transfers, and the ubiquity of online shopping. But credit card debt poses a serious risk to businesses and governments alike, not to mention individual savers and investors. Consequently, the need for efficient, timely, and reliable ways to anticipate credit card risk has grown. In this study, we offer a framework that combines three classifiers, namely, support vector machines, multilayer perceptron and decision trees, to improve the network's accuracy. The proposed strategy is shown to be very competitive with others through simulation.

Keywords: Credit scoring; Credit card; Machine learning; Classification; Metaheuristic optimization; K-Nearest neighbor; Random Forest; Support vector machines

1. Introduction

Banks and customers alike face serious difficulties with cash flow and credit card debt. The banks may give too many credit cards to those who don't deserve them in an effort to expand their customer base, which can lead to excessive card use and mounting debt. As a result, accurate risk assessment is fundamental to maintaining a secure financial infrastructure and mitigating the losses and unpredictability that might result from unexpected events. Predicting the likelihood of a client defaulting on a payment is more beneficial in risk control than classifying them as either "risky" or "non-risky" [1]. In this research, we improve performance by combining an ANN strategy with a metaheuristics-based feature selection technique. Algorithms like the support vector machine (SVM), logistic regression (LGR), and the random forest classifier are used to compare the final outcome to industry standards (RDC).

Due to their ability to investigate the connection between input data and their respective labels, neural networks are well-suited to difficult machine learning challenges. However, other types of machine learning models, such linear regression and Support Vector Machine (SVM), are more suited to solving the smaller issues. As a result, it is important to examine individual issues and determine whether or not a neural network is truly required. In addition, "neural network" is a broad term that encompasses a variety of models. When calculating neurons in the next layer, a feedforward neural network employs all of the neurons in the current layer simultaneously. Neurons are "parallel" in this process, despite the fact that their weights make them distinct from one another. However, sequential datasets benefit greatly from the Recurrent Neural Network paradigm. It enables inputs from the past to shape how inputs in the future are handled. Comparing the accuracy of these two networks is important.

Doi: https://doi.org/10.54216/JAIM.020105
Received: May 11, 2022 Accepted: October 26, 2022

2. Literature Review

Credit scoring has benefited from the widespread use of several data mining tools. Decision-making tools such as decision trees, discriminant analysis, and dynamic programming are discussed in [2]. Complex high-dimensional datasets were successfully analysed using Bayesian techniques and Markov chain in [3]. The authors aimed to improve the model's adaptability and computing efficiency by reducing the need for global specifications through the use of conditional graphs. As an improvement over both logistic regression and classic discriminant analysis, the authors of [4] developed a hybrid method in which neural networks are used in conjunction with discriminant analysis. Analysing texts for similarities is the plan of attack.

To simplify further analysis based on redundancy and relevance, we choose a subset of features in FS and exclude the irrelevant, noisy features. According to these two criteria, Yu and Liu [5] divide the feature subset into four groups: noisy and irrelevant; redundant and weakly relevant; weakly relevant and nonredundant; and very relevant. We say that a feature is irrelevant if it does not contribute to the model's ability to make accurate predictions. Models, search methods, feature quality metrics, and feature assessment are just some of the many methods that may be put into practice with the use of filter and wrapper techniques. When developing hypotheses for predictive models, a comprehensive set of characteristics is essential. There is a one-to-one relationship between the number of features and the size of the hypothesis space; in other words, if you increase the number of features, you also increase the size of the search space. In one scenario, the search space for a dataset includes numerous permutations if it has M features labelled with a binary class. Based on their relationship to the underlying learning model, FS approaches can be classified as either filter, wrapper, or embedding. Selecting statistical characteristics is at the heart of the filter approach. This method reduces processing effort since it is not reliant on the learning algorithm. The significance of the characteristics may be gauged by the application of statistical tools as information gain, chi-square test [6], Fisher score, correlation coefficient, and variance threshold. But the wrapper approach is rather classifier dependent in terms of performance. Based on the classification outcome, the optimal collection of characteristics is chosen. Since the wrapper technique must frequently operate in parallel with the classifier, it incurs significantly higher computational costs than filter approaches. While the filter approach is the most common, these alternatives are more reliable. Feature elimination using recursion [7], sequential feature selection algorithms [8], and evolutionary algorithms [9] are only a few examples of algorithms used as wrappings. The third strategy is an embedded technique that uses hybrid learning and ensemble learning to pick features. Due to the inclusion of a group consensus process, the results of this strategy are superior than those of its predecessor. When compared to wrapper approaches, for instance, random forest requires significantly less processing power. The embedded approach has the issue of being model-dependent.

In addition, a plethora of evolutionary metaheuristics-based feature selection methods are offered, with many of these being wrapper types due to the shown superior performance of wrappers [10]. Based on the grey wolf optimizer (GWO) presented by Mirjalili et al. [12], Too et al. [11] suggested a competitive binary grey wolf optimizer (CBGWO) for the feature selection issue in EMG signal categorization. Based on the data collected, CBGWO was shown to be the most effective algorithm for this particular use case. Binary grey wolf optimization (BGWO) [13], binary particle swarm optimization (BPSO) [14], ant colony optimization (ACO) [15], and binary differential evolution (BDE) [16] are just some examples of wrapper-based feature selection algorithms that have been introduced in many works to select a subset of features.

3. Proposed Methodology

This study, which investigates the causes of customer default payments in Taiwan, makes use of data culled from the credit card clients' default dataset. There is a copy of the data set in the UCI repository. The data came from a Taiwanese financial institution that issues credit cards and cash, and the intended recipients were credit cardholders of that institution. There are just two possible answers in a binary classification, and they both include the category variable "default payment" (Yes = 1; No = 0). In this case, the predictor characteristic is "default payment," which is a binary classification issue because it can take on just two possible values. There are 30,000 records in the collection, along with 24 characteristics. All of these characteristics are physiological indicators. There are 23 potential explanatory factors listed in Table 1.

Doi: https://doi.org/10.54216/JAIM.020105
Received: May 11, 2022 Accepted: October 26, 2022

A. The Proposed Ensemble Model

We start by showing you the data we used to boat-train our classifiers after we let go, then we walk you through the processes we propose doing, and lastly we show you the results. Three distinct classification strategies were employed in the uploading of this image: Different kinds of classifiers include decision trees (DT), multilayer perceptrons (MLP), and support vector machines (SVM). We employ stochastic fractal search (SFS) in conjunction with the whale optimization technique to boost the weight of these classifiers' votes inside an ensemble model (WOA).

Table 1 Features and explanation of the dataset

ID	Feature	Explanation
1	X1	Amount of the given credit (NT dollar)
2	X2	Gender (1 = male; 2 = female).
3	X3	Education (1 = graduate school; 2 = university; 3 = high school; 4 = others)
4	X4	Marital status (1 = married; 2 = single; 3 = others)
5	X5	Age (year)
6	X6	The repayment status in Sep 2005
7	X7	The repayment status in Aug 2005
8	X8	The repayment status in Jul 2005
9	X9	The repayment status in June 2005
10	X10	The repayment status in May 2005
11	X11	The repayment status in April 2005
12	X12	The amount of bill statement in Sep 2005
13	X13	The amount of bill statement in Aug 2005
14	X14	The amount of bill statement in Jul 2005
15	X15	The amount of bill statement in June 2005
16	X16	The amount of bill statement in May 2005
17	X17	The amount of bill statement in April 2005
18	X18	The amount paid in Sep 2005
19	X19	The amount paid in Aug 2005
20	X20	The amount paid in Jul 2005
21	X21	The amount paid in June 2005
22	X22	The amount paid in May 2005
23	X23	The amount paid in April 2005

Since the gradient descent works best with normalized (scaled) values, we standardize this dataset before using it for training. If we don't normalize the data, we might end up with numbers on a wide range of scales. It would require additional time for our model to train on this data in order to make the necessary weight adjustments. If we use normalization techniques to get all of our numbers to be on the same scale, our model will train considerably more quickly, and gradient descent will work well.

B. Support Vector Machines (SVM)

Authors in [17-20] mark the debut of the SVM. One of the more recent developments in the realm of supervised machine learning is the support vector machine (SVM). It works well for a small dataset with few outliers. The concept is to identify a hyper segmentation lanes to divide information. The area is segmented along this hyperplane, with each sector holding a certain kind of information (Fig. 1). Multiple hyperplanes can be selected to differentiate between the two types of information. The aircraft with the largest margin is what we're looking for. To calculate the margin, find the two data points closest to the hyperplane that correspond to the two categories. In order to improve the method,

Doi: https://doi.org/10.54216/JAIM.020105
Received: May 11, 2022 Accepted: October 26, 2022

support vector machines (SVMs) look for the super planar with the largest margin value, splitting the data into two equally-sized halves. The closest data points to the hyperplane are referred to as the "support vectors" (Fig. 2). Linear surface that cuts space in half, like a hyperplane does. If you want to divide up your space into two categories, then you need a hyperplane, which is a one-dimensional subspace.

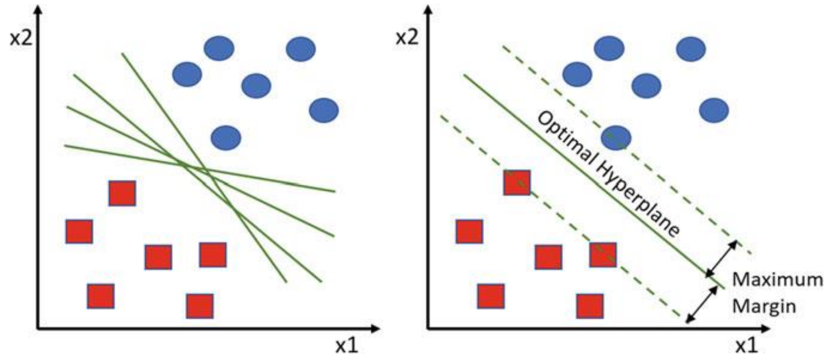

Figure 1: Structure of support vector machines.

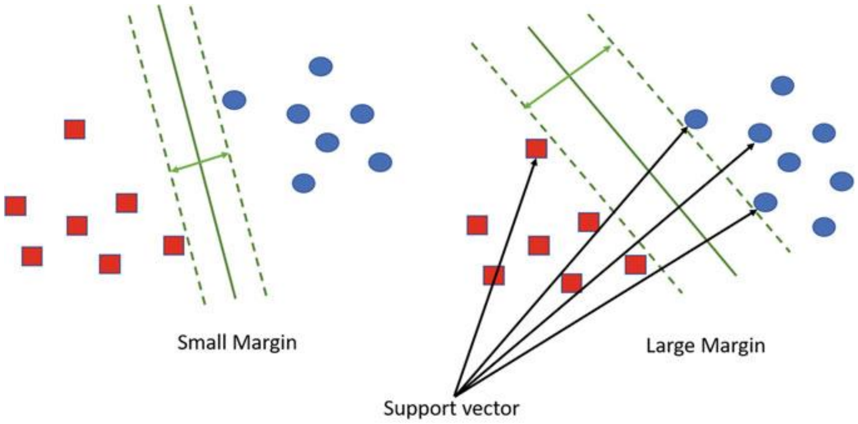

Figure 2: Support vector

C. Multilayer Perceptron (MLP)

It's been pointed out that you can use a multilayer perceptron neural network. For those situations when the variables cannot be separated in a straightforward manner. In order to address this issue, a multilayer perceptron is constructed by adding additional layers to a single-layer perceptron. As can be seen in Figure 3, the MLP network is a kind of feed-forward neural network that may have anywhere from one-to-many hidden layers. n neutrons are input to the network, which also includes n hidden neurons, and n neutrons are output.

Doi: https://doi.org/10.54216/JAIM.020105
Received: May 11, 2022 Accepted: October 26, 2022

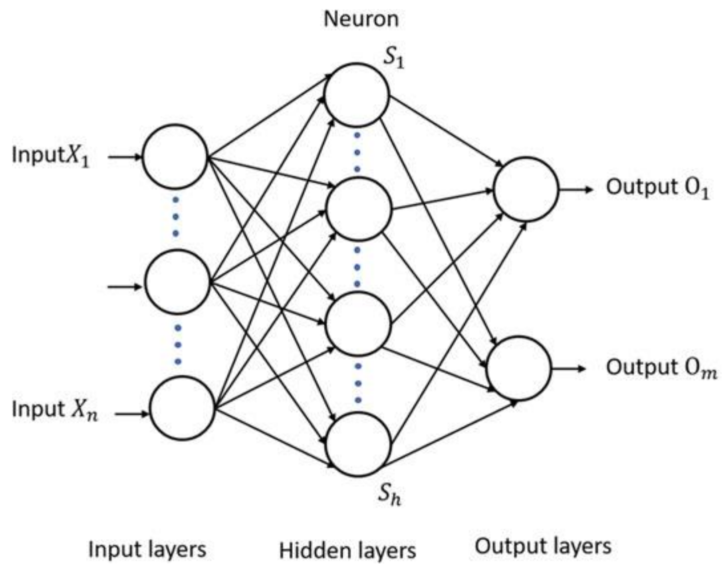

Figure 3: Structure of a multilayer neural network.

Because we'll be using a single hidden layer, our input matrix will have the form (batch size, number of attributes), and our weight matrices will contain one weight each for connections to and from the hidden layer and the output layer. So, the input layer is a matrix with the form of (number of features, number of hidden neurons), the hidden layer is a matrix with the form of (number of neurons, number of classes), and the output layer is a matrix with the form of (number of features, number of hidden neurons, number of classes) (batch size, number of classes).

D. Decision Trees (DT)

DT first appeared in [21, 22]. Each node in the tree represents a property, and each branch represents a possible value for that property. The decision tree's predictive value may be derived from it by tracing the tree's nodes in accordance with their associated property values. To begin, you need a tree algorithm to accurately anticipate the final result. Our data, its characteristics, and the categorical (dummy) values inside it require scrutiny. In addition, we can use the following formulae to deal with entropies and discriminative powers to determine the optimal characteristics for our tree. Presented below is a condensed version of the decision tree in Figure 4.

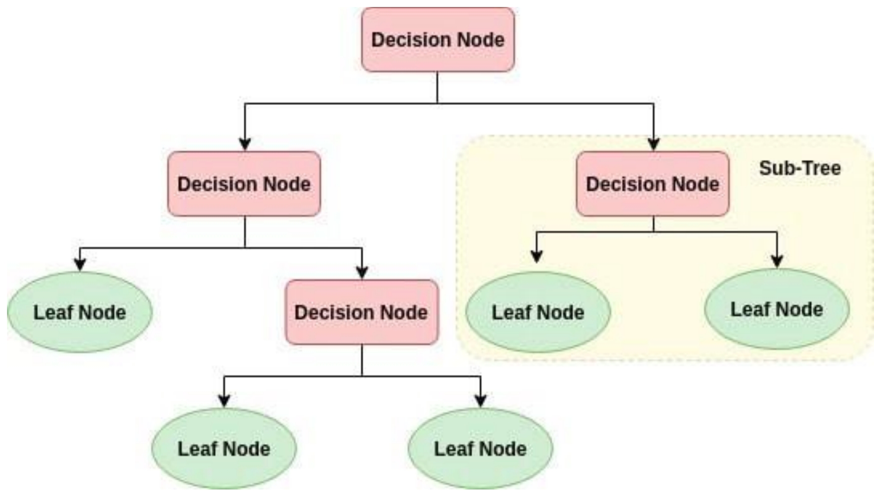

Figure 4: Structure of a decision tree.

For example, we find that the median age is the perfect divider between two groups of people in our data. This method was also applied to additional characteristics. After that, we've reached a point where we've located the most promising opportunities to split, ultimately determining whether certain nodes in our tree would be on the left or the right. To divide our data in half, we use a certain column

Doi: https://doi.org/10.54216/JAIM.020105
Received: May 11, 2022 Accepted: October 26, 2022

and value to identify the ideal splits while we are developing our tree with nodes (left child and right child of a node in a tree). In the recursive section, we apply the same strategies used in the previous sections to each successive level of the tree. There is no need to change the current node to a leaf node unless there is a question to be asked.

E. Metaheuristic Optimization

In computer science and mathematical optimization, metaheuristics are used to locate, build, or choose a heuristic (partial search algorithm) that may provide a viable solution to an optimization issue despite a lack of complete knowledge or computing resources. It would be difficult to collect an accurate cross-section of the space of feasible answers without the help of metaheuristics. Metaheuristics are often more flexible than other optimization methods since they may be used in a variety of settings with little in the way of preparation or in-depth understanding of the optimization issue at hand. In place of traditional optimization methods and iterative procedures, metaheuristics are increasingly being adopted; nevertheless, using a metaheuristic does not guarantee that the best possible solution will be discovered for a given class of problems. The final solution obtained by several prominent metaheuristics, stochastic optimization included, might vary depending on the values of the random variables used. In combinatorial optimization, hypotheses regarding the ideal solution are generated using metaheuristics, which can be more efficient than optimization algorithms, iterative techniques, and even simple heuristics. For this reason, they may stand for alternative ways of addressing optimization issues. There are a lot of articles on this. The bulk of published metaheuristics publications are inherently experimental since they relate the author's own experiences implementing the algorithm in the real world. However, there are also some formal theoretical results that provide light on issues like convergence and the possibility of achieving a global optimum. Multiple metaheuristic methodologies have been presented in recent research, and they may all have useful implications for the area as a whole. Imprecise terminology, a failure to effectively address critical topics, poor methodology, and a lack of acceptable citations have all hindered previous research on this topic.

4. Results

The study creates three different classifiers—the DT, the KNN, and the SVM. The dataset is split 70:30, with the former utilized for model training and the latter for checking the models' performance in a simulated environment. Furthermore, we use two distinct datasets—outlier data and removal outlier data—in our investigation. Table 2 displays the comparison of the classifier methods based on the accuracy, recall, precision, and F1 score obtained from the outlier data. The suggested ensemble model achieves the best accuracy in the findings.

Table 2: Classification results using the proposed method compared to other methods

	Accuracy	Sensitivity	Specificity	Pvalue	Nvalue	F-score
NN	0.8945	0.9967	0.6667	0.8696	0.9890	0.9288
SVM	0.8844	0.9967	0.6429	0.8571	0.9890	0.9217
DT	0.8369	0.9967	0.5455	0.8000	0.9890	0.8876
Guided WOA +PSO	0.9420	0.9967	0.7965	0.9288	0.9890	0.9615

Statistical evidence supporting the voting ensemble classifier's superiority is presented in Table 3. With the optimum voting ensemble that was presented, we get markedly better results.

Table 3: Statistical analysis of the results recorded by the proposed method

	NN	SVM	DT	Guided WOA +PSO
Number of values	10	10	10	10
Minimum	0.8745	0.8844	0.8369	0.942
25% Percentile	0.892	0.8844	0.8369	0.942
Median	0.8945	0.8844	0.8369	0.942
75% Percentile	0.8945	0.8844	0.8394	0.942

Doi: https://doi.org/10.54216/JAIM.020105
Received: May 11, 2022 Accepted: October 26, 2022

Maximum	0.8945	0.8944	0.8569	0.942
Range	0.02	0.01	0.02	0
10% Percentile	0.8755	0.8844	0.8369	0.942
90% Percentile	0.8945	0.8934	0.8559	0.942
95% CI of median				
Actual confidence level	97.85%	97.85%	97.85%	97.85%
Lower confidence limit	0.8845	0.8844	0.8369	0.942
Upper confidence limit	0.8945	0.8844	0.8469	0.942
Mean	0.8915	0.8854	0.8399	0.942
Std. Deviation	0.006749	0.003162	0.006749	0
Std. Error of Mean	0.002134	0.001	0.002134	0
Lower 95% CI of mean	0.8867	0.8831	0.8351	0.942
Upper 95% CI of mean	0.8963	0.8876	0.8447	0.942
Coefficient of variation	0.7571%	0.3572%	0.8036%	0.000%
Geometric mean	0.8915	0.8853	0.8399	0.942
Geometric SD factor	1.008	1.004	1.008	1
Lower 95% CI of geo. mean	0.8866	0.8831	0.8351	0.942
Upper 95% CI of geo. mean	0.8963	0.8876	0.8447	0.942
Harmonic mean	0.8914	0.8853	0.8399	0.942
Lower 95% CI of harm. mean	0.8866	0.8831	0.8351	0.942
Upper 95% CI of harm. mean	0.8964	0.8876	0.8446	0.942
Quadratic mean	0.8915	0.8854	0.8399	0.942
Lower 95% CI of quad. mean	0.8867	0.8831	0.8351	0.942
Upper 95% CI of quad. mean	0.8963	0.8876	0.8448	0.942
Skewness	-2.277	3.162	2.277	
Kurtosis	4.765	10	4.765	
Sum	8.915	8.854	8.399	9.42

A comparison of the suggested technique to the alternatives is made using the Wilcoxon signed-rank test. Table 4 displays the data gathered throughout this study. The p-value listed in the table is the supporting evidence.

Table 4: Wilcoxon signed rank test of the recorded results of the proposed method

	NN	SVM	DT	Guided WOA +PSO
Theoretical median	0	0	0	0
Actual median	0.8945	0.8844	0.8369	0.942
Number of values	10	10	10	10
Wilcoxon Signed Rank Test				
Sum of signed ranks (W)	55	55	55	55
Sum of positive ranks	55	55	55	55
Sum of negative ranks	0	0	0	0
P value (two tailed)	0.002	0.002	0.002	0.002
Exact or estimate?	Exact	Exact	Exact	Exact
P value summary	**	**	**	**
Significant (alpha=0.05)?	Yes	Yes	Yes	Yes

Doi: https://doi.org/10.54216/JAIM.020105
Received: May 11, 2022 Accepted: October 26, 2022

How big is the discrepancy?				
Discrepancy	0.8945	0.8844	0.8369	0.942

The Figure 7 graph shows how the optimal voting ensemble classifier performs in comparison to the reference models. An illustration of the improved efficiency of the suggested approach is shown below.

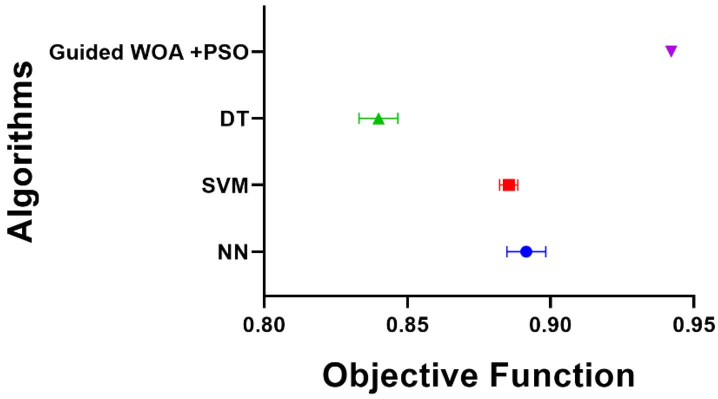

Figure 7: The accuracy of the proposed method compared to other methods.

5. Conclusion

A feature selection approach is provided with a basic neural network architecture in this research. Only the top 17 features from the original set were chosen. Results from a sensitivity analysis that varied the number of neurons and training epochs showed that the MLP consistently produced accurate predictions. The optimal setup, which beat the other algorithms considered here, included 45 neurons in a single layer and 100 epochs. The result was an accuracy of 85.24 percent. We want to further improve the network architecture and the feature selection method in our future research. In place of the standard stochastic gradient descent, hybrid metaheuristics that retain the best features of their parent algorithms might be implemented to expedite the training process. In [23], an adaptive particle swarm—grey wolf optimization technique is presented as a viable solution.

Funding: "This research received no external funding"

Conflicts of Interest: "The authors declare no conflict of interest."

References

[1] The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients | Semantic Scholar. https://www.semanticscholar.org/paper/Thecomparisons-of-data-mining-techniques-for-the-Yeh-Lien/1cacac4f0ea9fdff3cd88c151c94115a9fddcf33. Accessed 6 Sep 2020

[2] Quantitative methods in credit management: a survey. Oper Res. https://pubsonline.informs.org/doi/abs/10.1287/opre.42.4.589. Accessed 6 Sep 2020

[3] Bayesian data mining, with application to benchmarking and credit scoring—Giudici—2001.Applied stochastic models in business and industry. WileyOnline Library. https://onlinelibrary.wiley.com/doi/abs/10.1002/asmb.425. Accessed 6 Sep 2020

[4] Lee T-S, Chiu C-C, Lu C-J, Chen I-F (2002) Credit scoring using the hybrid neural discriminant technique. Expert Syst Appl 23(3):245–254. https://doi.org/10.1016/S0957-4174(02)00044-1

[5] Yu L, Liu H (2004) Efficient feature selection via analysis of relevance and redundancy. JMach Learn Res 5:1205–1224

[6] Yang Y, Pedersen J (1997) A comparative study on feature selection in text categorization

Doi: https://doi.org/10.54216/JAIM.020105
Received: May 11, 2022 Accepted: October 26, 2022

[7] Yan K, Zhang D (2015) Feature selection and analysis on correlated gas sensor data with recursive feature elimination. Sens Actuators B Chem 212:353–363. https://doi.org/10.1016/j.snb.2015.02.025

[8] Jain A, Zongker D (1997) Feature selection: evaluation, application, and small sample performance. IEEE Trans Pattern Anal Mach Intell 19(2):153–158. https://doi.org/10.1109/34.574797

[9] De Stefanoc, Fontanella F, Marrocco C, Scotto di FrecaA(2014)AGA-based feature selection approach with an application to handwritten character recognition. Pattern Recogn Lett 35:130–141. https://doi.org/10.1016/j.patrec.2013.01.026

[10] Zorarpacı E, Özel SA (2016) A hybrid approach of differential evolution and artificial bee colony for feature selection. Expert Syst Appl 62:91–103. https://doi.org/10.1016/j.eswa.2016.06.004

[11] Too J, Abdullah AR, Mohd Saad N, Mohd Ali N, TeeW(2018) A new competitive binary grey wolf optimizer to solve the feature selection problem in EMG signals classification. Computers 7(4), Art. no. 4. https://doi.org/10.3390/computers7040058

[12] Mirjalili S, Mirjalili SM, Lewis A (2014) Grey Wolf Optimizer. Adv Eng Softw 69:46–61.https://doi.org/10.1016/j.advengsoft.2013.12.007

[13] Binary grey wolf optimization approaches for feature selection. Request PDF. ResearchGate.https://www.researchgate.net/publication/280836997_Binary_Grey_Wolf_Optimization_Approaches_for_Feature_Selection. Accessed 6 Sep 2020

[14] Chuang L-Y, Chang H-W, Tu C-J, Yang C-H (2008) Improved binary PSO for feature selection using gene expression data. Comput Biol Chem 32(1):29–38. https://doi.org/10.1016/j.compbiolchem.2007.09.005

[15] Text feature selection using ant colony optimization | Request PDF. ResearchGate. https://www.researchgate.net/publication/222669067_Text_feature_selection_using_ant_colony_optimization. Accessed 6 Sep 2020

[16] Feature selection with discrete binary differential evolution. IEEE conference publication. https://ieeexplore.ieee.org/document/5376334. Accessed 6 Sep 2020

[17] Wright RE (1995) "Logistic regression", in reading and understanding multivariate statistics. American Psychological Association, Washington, pp 217–244

[18] An introduction to kernel and nearest-neighbor nonparametric regression: The American statistician: Vol 46, No 3. https://www.tandfonline.com/doi/abs/10.1080/00031305.1992.10475879.Accessed 6 Sep 2020

[19] Statistical Learning Theory. Wiley. https://www.wiley.com/en-us/Statistical+Learning+Theory-p-9780471030034. Accessed 6 Sep 2020

[20] Ting KM, Zheng Z (1999) Improving the performance of boosting for Naive Bayesian classification. In: Methodologies for knowledge discovery and data mining, Berlin, Heidelberg, pp 296–305, https://doi.org/10.1007/3-540-48912-6_41

[21] Quinlan JR (1986) Induction of decision trees. Mach Learn 1(1):81–106. https://doi.org/10.1007/BF00116251

[22] Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324

[23] Dao S, Pham T (in press) Capacitated vehicle routing problem—a new clustering approach based on hybridization of adaptive particle swarm optimization and grey wolf optimization. In: Evolutionary data clustering: algorithms, and applications. Springer

Doi: https://doi.org/10.54216/JAIM.020105
Received: May 11, 2022 Accepted: October 26, 2022