The Optimized Function Selection Using Wolf Algorithm for Classification

Duangjai Jitkongchuen and Worapat Paireekreng*
College of Innovative Technology and Engineering
Dhurakij Pundit University
Bangkok, Thailand
{duangjai.jit,worapat.png}@dpu.ac.th
www.dpu.ac.th

Abstract

Several classification techniques have been widely explored during the past decade. One of the novel approaches in recent years is Nature Based Algorithm. This approach is appropriate to imbalanced dataset. The focus of Nature Based Algorithm mostly is related to selection the optimized functions for self-learning. This is used to solve the NP-hard problems. However, Some Nature Based Algorithms are suitable for general situation; some may be suitable for customized situation. This research proposes the Featured-Wolf (F-Wolf) algorithm to optimize the function selection problem in classification. The proposed algorithm applies the movement of a wolf and characteristics of wolves' leaders which can be more than one leader in a pack. Therefore, the pack can have more than one dominant leader which can help to select the most optimized functions to selection the most relevant features in the dataset. The experiment shows the comparison among other popular Nature Based Algorithms such as Ant Colony Optimization and other classification techniques. The results show that F-Wolf performs better results in terms of accuracy rate.

Keywords: Nature Based Algorithms, Wolf algorithm, Classification

1. Introduction

Data science is now crucial for almost all areas. Data from various sources needs to be analyzed and uses specific analytic tools. One of the most important domain to present the analyzed data is classification. The common techniques used in data mining to solve the classification problem. Some research combined the classification techniques using ensemble technique to improve the classification performance[1].

In addition, the new classification techniques have been investigated to enhance the performance because the characteristics of each dataset have their uniqueness. The adjusted classification model and classifiers needs to be improved. One of the novel approaches in recent years is Nature Based Algorithm (NBA). The approach is related to selection the optimized functions for self-learning. This is the most appropriate technique to solve the NP-hard problems. It also works well with imbalanced dataset. The examples of NBA are such as Grey Wolf Optimizer (GWO)[2] and Ant Colony Optimization [3].

Furthermore, there is an attempt to improve the speed of computational time for classification performance. One of the stage in the pre-processing phase is feature selection. The feature selection has been used for the objective of selecting the most relevant attributes of the data set as input variables. These variables impact on the performance of classification if the relevant variables are chosen to build the classification model. The pre-processing stage eliminates the irrelevant attributes in order to increase the computational time of the model building stage. The example of feature selection can be seen from several data mining, machine learning, and pattern recognition[4]. Moreover, the input variables are the challenge of this area by selecting the appropriate subset of attribute with maintaining of classification performance [5,6].

However, Some Nature Based Algorithms are suitable for general situation; some may be suitable for customized situation. Therefore, some enhancing technique is needed for improving the performance of classification using NBA approach. Feature selection technique can be included in the NBA to perform better in terms of accuracy rate.

This paper proposed the Featured-Wolf (F-Wolf). The main idea of F-Wolf is to imitate the grey wolf behavior including feature selection in the pack. The algorithm uses information exchange within population which leads to generate new candidate individuals for feature selection and perform the better results for classification.

2. Related Works

2.1. *Data Mining and Classification Techniques*

There are many classification techniques that have been used to deal with classification problems and predictions. Examples of these common used classification techniques are Decision Trees (DT), Naïve Bayes (NB), Artificial Neural Networks (ANN) and Support Vector Machines (SVM)[7]. There are different in terms of advantages and disadvantages based on each technique. The factors to decide which technique is suitable for dataset are such as simplicity of the algorithm, up-sampling scale, robustness and outlier handing.

Classification techniques have also been incorporated into the several applications. For example, the mobile services incorporated Artificial Neural Networks (ANN) with feed-forward back-propagation neural network in order to select the different types of particular services[8,9].

The performance of classification model can be assessed by confusion matrix. It is shown in *Table 1*. The confusion matrix presents the results of amount of correct and incorrect instance in each class.

Table 1. Confusion Matrix Terminology.

	Positive Prediction	**Negative Prediction**
Positive class	True Positive (TP)	False Negative (FN)
Negative class	False Positive (FP)	True Negative (TN)

True Positive (TP) and True Negative (TN) are the positive instances and negative instances correctly classified respectively whereas False Positive (FP) and False Negative (FN) are the negative instances and positive instances misclassified respectively.

To measure the performance of the classification, the traditional accuracy rate (1) has been used.

$$ACC = \frac{TP+TN}{TP+FN+FP+TN} \qquad (1)$$

2.2. *Soft Computing and Nature Based Algorithm*

The soft computing has been introduced to solve the uncertain problems with intensive computation. The principle of soft computing can implement probabilistic reasoning approach to solve the problems. The examples of techniques used in soft computing are such as evolutionary algorithm. One of the commonly used technique is Ant Colony Optimization (ACO) which can be seen in the recent research[3].

However, to handle different type of data such as categorical and continuous data in the real world, Customized techniques is needed to obtain the better performance in terms of accuracy. Some problems domain such as NP-hard problems also needs the specific techniques and methodology to solve the uncertain variables and alternatives. Therefore, it is also the challenge to implement the adapted Nature Based Algorithm for classification performance.

2.3. *Grey Wolf Optimizer*

The grey wolf optimizer algorithm (GWO) was proposed by Mirjalili et al[10]. It is the recent Nature Based Algorithm which simulate the grey wolf behavior to live in a pack. The social dominant hierarchy of the wolf pack can be defined as leaders, subordinate wolves and members which consist of scouts, sentinels, elders, hunters and caretakers. Each member has a different role in a pack. For example, Sentinels protect and guarantee the safety of the pack. Therefore each level and member is defined as variables called alpha, beta and delta respectively. In addition, the lowest level is omega. The omega wolves have to comply with all the other dominant wolves.

The grey wolves show naturally ability to encircle and identify the position of a prey and other wolves help the hunting for a pack. This behavior can be explained in mathematical model which assumes the leader to be alpha (α). The beta (β) and delta (δ) is similar to the second and the third optimal solutions, respectively. Whereas, the rest of the candidate solutions are assumed to be omega (ω). The hunting is guided by alpha, beta and delta. Besides, the omega wolves would consider the best solutions from three different positions and update the information to the pack.

Nevertheless, the traditional grey wolf algorithm basically simulate with only one pack which leads to encircle on local search. Therefore, splitting the pack of the grey wolf is challenge. The proposed method is included algorithm to reduce the optimum ratio and distribute value to in order to obtain the larger search space. This is also including feature selection.

2.4. *Feature Selection for Classification Problem*

Due to the consuming computational time of building the classification model for training dataset, some methodologies are needed to address to problem. The common methodology used is feature selection. The purpose is to select the subset of input variables which is useful and impact on performance of the classifier. This includes elimination of some irrelevant variables with few contributions towards the prediction results. The use of feature selection can be seen from data mining, machine learning and pattern recognition[4, 11]. The objective of feature selection is to select a subset of useful features from the input variables that impact on accuracy of the classifier and eliminate irrelevant features with little contribution towards the prediction results. Moreover, the challenge is to choose the minimum subset of features with little or no loss of classification accuracy[5,6].

Feature selection can be divided into two categories: model-free method and model-based methods. Model-free methods are based on statistical tests, properties of function and available data such as linear regression, whereas, model-based methods such as neural network develop model to find the significant features and minimize the model output error[12,13]. Therefore, Nature Based Algorithm (NBA) can be also incorporated into classification model in order to select the most suitable variables for classification problem with less impact on performance.

3. Research Methodology

3.1. *Experimental Design*

The data source used for the experiment was obtained from University of California Irvine (UCI) machine learning repository. There are 10 datasets used in this research.

3.2. *The Proposed Featured Wolf (F-Wolf) and Procedure*

The overall and details of F-Wolf procedure are shown below in "Fig. 1.".

```
Initialize the grey wolf population $X_i (i = 1, 2, ..., n)$
where each wolf consists of choosing attribute with value yes or no.
Initialize parameters (a, A and C)
Calculate the fitness of each search agent based on accuracy rate
$X_\alpha$ = the best search agent
$X_\beta$ = the second best search agent
$X_\delta$ = the third best search agent
while (t < Max number of iterations)
    Update current wolf's position based on top three wolves
    Crossover top three wolves to create new member using choosing attribute
        Uses voting to find the maximum amount of value chosen by leaders
        The values of the attribute is the best entropy is selected
    Sort all wolves based on suitable value
    Insert new member to wolf with least suitable value
    Update a, A and C
    Calculate the fitness of all search agents
    Update $X_\alpha$, $X_\beta$, $X_\delta$
end while
```

Fig. 1. Procedure of F-Wolf

4. Experimental Results

After the datasets have been prepared, the next step is to build the classification model for prediction. There were 4 different types of algorithms used to compare in the experiment, specifically Featured Wolf (F-Wolf), Ant Colony Optimization (ACO), C4.5 and PART. In this research, the metric to determine the performance of the classification is based on accuracy rate. The preliminary results are shown in *Table 2.*

Table 2. Comparison Results.

Technique	**Accuracy Rate (%)**			
Dataset	**F-Wolf**	**ACO**	**C4.5**	**PART**
Credit-G	***87.96***	79.29	71.70	71.90
Haberman	76.30	79.63	70.59	70.59
Heart-C	***78.00***	69.38	49.17	55.12
Heart-H	70.00	73.47	64.29	62.59
Heart-statlog	***86.67***	77.24	62.22	64.07
Horse	82.50	88.00	82.00	82.33
Iris	***99.09***	98.22	94.67	95.33
Pima diabets	***87.00***	78.89	73.31	73.44
Shuttle	***96.30***	75.60	53.33	53.33
Sonar	***97.59***	87.07	76.92	73.56
Average	86.14	80.68	69.82	70.23

It appeared that the proposed F-Wolf classification model can perform better in terms of accuracy rate compared to other classification techniques. In addition, it performed better than C4.5 and PART. To compare with ACO, the Nature Based Algorithm, the results show that the proposed technique presented the better results with 7 out of 10 datasets. Furthermore, the average accuracy rate for F-Wolf is higher significantly compared to other techniques.

5. Discussion and Conclusions

Classification is an important problem domain for data related area. Several techniques have been proposed to deal with classification. One of the novel approaches in recent years is Nature Based Algorithm (NBA) which appropriate to imbalanced dataset including NP-hard problems. Grey Wolf Optimizer (GWO) is one of the NBA to address the classification problems. However, it may be suitable for one pack situation, therefore, Featured Wolf Algorithm (F-Wolf) is proposed using information exchange within population. This which leads to generate new candidate individuals for feature selection in order to enhance the classification performance.

The results from the experiments based on UCI datasets have shown that the proposed F-Wolf provided the better classification results. The average of accuracy rate for all datasets in the experiment was also higher than other classification techniques. The F-Wolf used crossover to keep the best value in each round but the traditional did not cover.

In the future works, it is important to consider the imbalanced data sets with other Nature Based Algorithm. This includes the investigating the improved performance on prediction results of the classification model. The hybrid Nature Based Algorithm is therefore needed to address the problem. Moreover, the other soft computing techniques such as fuzzy-based techniques or machine learning techniques can be incorporated in the feature selection stage. This is to build the improved classification model.

References

1. W. Paireekreng and T. Prexawanprasut, An Integrated Model for Learning Style Classification in University Students Using Data Mining Techniques, in Proc. *12th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, (Hua Hin, Thailand, 2015).
2. D. Jitkongchuen, P. Phaidang and P. Pongtawevirat, Grey Wolf Optimization Algorithm with Invasion-based Migration Operation, in Proc. *15th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2016)*, (Okayama, Japan,2016).
3. W. Paireekreng, D. Jitkongchuen, W. Sukpongthai and R. Suwannakoot, Improving Soft Computing Performance with Ant Colony Optimization for Multiclass Classification: The Application for Learning Style Classification, in *Proc. 7th International Conference on Intelligent Systems, Modelling and Simulation (ISMS2016)*, (Bangkok, Thailand, 2016) pp.101-105.
4. P. Mitra, C. A. Murthy and S. K. Pal, Unsupervised feature selection using feature similarity, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3) (2002) 301–312.
5. H. Almuallim and T. G. Dietterich, Learning boolean concepts in the presence of many irrelevant features, *Artificial Intelligence*, 69 (1-2) (1994) 279–305.
6. D. Koller and M. Sahami, Toward optimal feature selection, in *Proc. 13th International Conference on Machine Learning*, (1996).
7. X. Wu, et al., Top 10 Algorithms in Data Mining, *Knowledge and Information Systems*, 14 (2008) 1-37.
8. Q. H. Mahmoud, et al., Design and implementation of a smart system for personalization and accurate selection of mobile services, *Requirement Engineering*, 12 (2007) 221-230.
9. A. Cufoglu, et al., A Comparative Study of Selected Classifiers with Classification Accuracy in User Profiling, in *Proc. World Congress on Computer Science and Information Engineering*, (Los Angeles, California, USA, 2009).
10. S. Mirjalili, S. M. Mirjalili and A. Lewis, Grey wolf optimizer, *Advances in Engineering Software*, 69 (2014), 46-61.
11. Miller, Subset Selection in Regression, 2nd edn. (Chapman & Hall/CRC, 2002).
12. S. M. Vieira, J. M. C. Sousa and T. A. Runkler, Two cooperative ant colonies for feature selection using fuzzy models, *Expert Systems with Applications*, 37(4) (2010) 2714-2723.
13. A. E. Isabelle Guyon, An Introduction to Variable and Feature Selection, *Journal of Machine Learning Research*, 3 (2003) 1154-1182.