A Pre-Processing Approach Based on Artificial Bee Colony for Classification by Support Vector Machine

Ahmet Babalık, İsmail Babaoğlu, and Ahmet Özkış

***Abstract*—In this study, handling with the success of pre-processing on classification tasks, artificial bee colony (ABC) algorithm is used as a pre-processor in order to improve accuracy of the support vector machine (SVM) classifier. Proposed approach is examined on three different online available dataset by using k-fold cross validation method. The results obtained are compared with the results of the classification of the datasets with pure SVM classifier. The increase of the classification accuracy is observed. By altering parameters of the suggested approach, it is thought the approach would be more successful on the different datasets.**

***Index Terms*—Artificial bee colony, support vector machine, medical data classification.**

I. Introduction

Classification and clustering are some of the basic problems encountered in human life. The main aim of the classification problems is to develop robust, stable, successful and fast classification models utilizing features or attributes of the problem dataset. On the classification process of the medical data various mathematical, statistical and artificial intelligence methods are used [1]. Researchers are suggesting novel techniques and hybrid models in order to improve the success rate. Selection of the classifier influences the success of the classification accuracy [2]. In addition to this, it is known that pre and post processing positively affects the success rate of the classifier. Mathematical, statistical, fuzzy logic, meta-heuristic algorithms and etc. are used on the pre-processing [3].

Dehuri et al. suggested an improved particle swarm optimization technique. Authors used this technique for classification in order to train the functional link artificial neural network [1]. Fan et al. suggested a hybrid model for medical data classification. For this aim, they developed the model by integrating a case-based data clustering method and a fuzzy decision tree [2]. Mok et al. proposed a new clustering analysis method that identifies the desired cluster number and procedures [4]. Chinneck suggested a novel integrated method that simultaneously selects features while placing the separating hyperplane [5]. Alkım et al. used learning vector quantization (LVQ) for classification of medical data. In order to increase the classification ability of LVQ network, they developed a reinforcement mechanism by embedding the algorithm to LVQ [6]. Chen et al. proposed a rough set based support vector machine (SVM) classifier for breast cancer diagnosis. For removing redundant features, rough set reduction algorithm was employed as a feature selection tool [7]. Cedeno et al. developed a novel algorithm by improving neural network training for pattern classification. Authors was inspired this algorithm by the biological metaplasticity property of neurons and Shannon's information theory. They utilized this algorithm for classification of breast cancer dataset [8]. Ramirez et al. experimented the performance of two decision tree procedures and four Bayesian network classifiers as potential decision support systems in the cytodiagnosis of breast cancer. Fan et al. developed a hybrid model by integrating a case-based data clustering method and a fuzzy decision tree for medical data classification [9].

In this study, swarm optimization based ABC algorithm is utilized as a pre-processing method, and its effect on the classification rate is investigated.

Manuscript received July 23, 2012; revised September 20, 2012.

The authors are with the Department of Computer Engineering, Selçuk University, Konya, Turkey (e-mail: ababalik@selcuk.edu.tr).

II. Materials and Methods

A. Dataset Description

Three of the online available datasets achieved from University of California at Irvine (UCI) are used in this study. The description of the datasets [10] could be given as follows;

Breast Cancer Wisconsin (Diagnostic) Dataset: This dataset is related to diagnosis of people with breast cancer. Features of the dataset are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. The characteristics of the cell nuclei present in the image are described. The dataset includes 569 instances having 32 real valued attributes in 2 classes. The classes were named benign and malignant.

Breast Cancer Wisconsin (Original) Dataset: This dataset is related to diagnosis of people with breast cancer. The dataset includes 699 instances having 11 integer valued attributes in 2 classes. Each instance has 9 cytological characteristics differed significantly between benign and malignant samples. The classes were named benign and malignant. This dataset includes missing values in 16 instances. These instances are excluded in classification process of this study.

Vertebral Column Dataset: This dataset is related to diagnosis of people with vertebral column disorders. Features of the dataset are obtained from sagittal panoramic radiographies of the spine. The dataset includes 310 instances having 6 real valued attributes in 3 classes. The classes were named normal, disk hernia and spondylolisthesis. The categories Disk Hernia and Spondylolisthesis could be merged into a single category named as abnormal. Thus, the

classification process is implemented using both 2 (normal - abnormal) and 3 (normal - disk hernia - spondylolisthesis) classes in this study.

B. Support Vector Machine (SVM)

Support Vector Machine is a trained classification algorithm oriented from statistical learning theory. SVM was developed by Vapnik. This method is utilized to classify classification problems having two distinct classes [11]. Recently, SVM is used for many classification problems [7], [12], [13].

In SVM, it's aimed to find the optimum hyperplane that separates two different classes. This optimum hyperplane divides the sample space so that distances of the different classes samples to the hyperplane are the most (shown in Fig. 1). The functions which are used to generate the hyperplane are called as kernel functions. These functions could be linear or non-linear. Linear, polynomial and radial basis function (RBF) kernels are most widely used kernel functions [14].

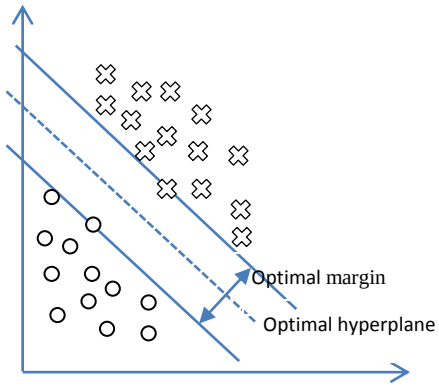

Fig. 1. SVM hyperplane illustration for linear kernel

C. Artificial Bee Colony (ABC)

In ABC algorithm, bees are split up to three namely employed bee, onlooker bee and scout bee. Employed bees are responsible for searching research space according to parameters of foods. Initial position of food sources are chosen by scout bees as randomly in search area. Afterward, scout bees which determine initial positions are changed to employed bees. Onlooker bees look at dance of employed bees. Dance of employed bees consider fitness value of each nectar. Scout bees explore food source randomly in search area. Onlooker and scout bees are considered as unemployed bees. Quality of a food source is tried to progress by employed and onlooker bees. This exploitation lasts till failure limit. If a food source can't get better nectar as failure limit time, another words being unsuccessful exceeds failure limit, employed bee who belong to food source which can't be progressed becomes a scout. The scout bee explores a position in space randomly. In standard ABC, just one employed bee can become scout and make exploration in a cycle. Count of all bees (employed, onlooker and scout) in swarm is equal to two times of food source. In other words number of employed bees is equal to food source [15].

Main structure of standard ABC algorithm could be given below:

1) Initialize stage of population
   Repeat
2) Searching stage of employed bees
3) Placing stage of onlooker bees in accordance with nectar amounts
4) Delivery stage of scout bees
5) Register the best food source obtained so far
   Until

In initialization stage, position of food source is launched randomly by scouts with respect to parameters are chosen.

In employed bee stage, each employed bee looks for new food source which has better nectar within neighborhood of itself. Employed bees make neighborhood choice of food source randomly. After the choice, fitness of new food sources are calculated and greedy selection is done. If new fitness value is better than previous one, food source is moved new position and previous position clear from the memory. Furthermore, because of renewing the food source, failure limit of that source is reset. If new fitness value is worse than previous one, food source stays previous position and increases failure limit. After all this process, employed bees let onlooker bees know of fitness value by dancing in beehive.

In onlooker bee stage, onlooker bees determine a food source according as fitness ratio of food sources. When the nectar quantity of a food source is rises, selection likelihood of that food source also rises. When food source is determined for an onlooker by using likelihood, a neighbor food source is also selected and found a candidate food source. Greedy selection is performed between the candidate food source and previous one.

In scout bee stage, food sources which can't be progressed as experimentation of failure limit are evacuated by employed bees. The employed bees whose food sources are evacuated become scouts and search a new solution in search area. In standard ABC, just one employed can become scout in a cycle [15].

D. Proposed Approach

In the proposed approach, it's aimed to improve the classification accuracy by weighting the datasets using pre-processing. Firstly, optimum c and γ parameters of SVM models are obtained using grid search method [14], [16]. Average success rates of the SVM models are evaluated. Then, utilizing these SVM models, the weight vectors which would be applied to the feature vectors are obtained by using ABC algorithm. Each dataset is weighted employing the individual weight vector, and the weighted datasets are classified by using SVM models with pre-obtained c and γ parameters. For weighted datasets, average success rates of the SVM models are evaluated. The illustration of the proposed approach is given in Fig. 2.

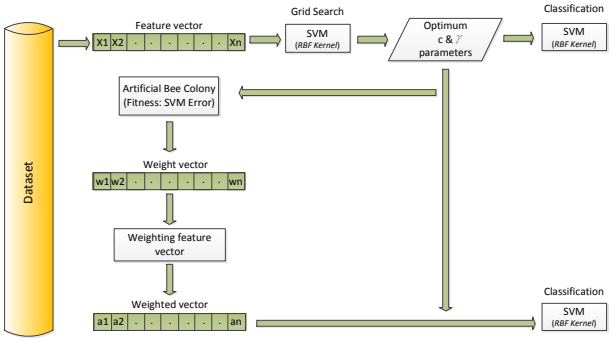

Fig. 2. The proposed approach.

III. Results and Discussion

In order to test the proposed approach *Breast Cancer Wisconsin (Diagnostic) Dataset*, *Breast Cancer Wisconsin (Original) Dataset* and *Vertebral Column Dataset* are used. Proposed approach is implemented in MATLAB platform and SVM models are evaluated using LIBSVM package [17].

In the classification process of both datasets and weighted datasets, SVM is used as the classifier. RBF kernel is used as the kernel function in SVM. k-fold cross validation algorithm is used in order to improve the reliability of the study, and k is used being equal to 5. 4 fold of the datasets are used in training process and 1 fold is used in test process. c and γ parameters of SVM are obtained using grid search algorithm [16]. Various pairs of c and γ parameters are tried by exponentially growing sequences. The grid size is utilized as $[2^{-20}, 2^{20}]$ and $[2^{-20}, 2^{20}]$ for c and γ parameters respectively. After obtaining optimum c and γ parameters the classification accuracy for each dataset is evaluated.

In weighting process, ABC algorithm is used as a pre-processor. The weight vectors are obtained for each dataset using associated optimum c and γ parameters with related SVM model. The values of weight vector are achieved in the range [0, 1]. The number of the colony size of the ABC algorithm is used as 20, and max iteration is used as 1000. The error value obtained from optimum c and γ parameters and related SVM model is used as the fitness function in ABC algorithm.

The classification results are given in Table I. According to the classification results given in Table I, it is observed that classification results are slightly improved by the proposed approach due to the high success rate of classification of the un-weighted data with SVM. Besides, the classification results are explicitly improved by the proposed approach due to lower success rate of classification of the un-weighted data with SVM. The proposed approach can be experimented on different dataset or problems as a future work.

TABLE I: The Classification Results (%)

Dataset	**SVM** *(Mean of 5 Fold)*	**Weighted SVM with ABC** *(Mean of 5 Fold)*
Breast Cancer Wisconsin (Diagnostic) Dataset	97.36	97.72
Breast Cancer Wisconsin (Original) Dataset	97.51	97.95
Vertebral Column Dataset (2 class)	89.03	91.61
Vertebral Column Dataset (3 class)	89.35	90.97

Acknowledgements

This study has been supported by Scientific Research Project of Selçuk University.

References

[1] S. Dehuri, R. Roy, S. Cho, and A. Ghosh, "An improved swarm optimized functional link artificial neural network (ISO-FLANN) for classification," *The Journal of Systems and Software*, vol. 85, pp. 1333 – 1345, 2012.

[2] C.-Y. Fan, P.-C. Chan, J.-J. Lin, and J. C. Hsieh, "A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification," *Applied Soft Computing*, vol. 11, pp. 632-644, 2011.

[3] L. Xiang-wei and Q. Yian-fangi, "A data preprocessing algorithm for classification model based on Rough sets," *Physics Procedia*, vol. 25, pp. 2025-2029, 2012.

[4] P. Y. Mok, H. Q. Huang, Y. L. Kwok, and J. S. Au, "A robust adaptive clustering analysis method for automatic identification of clusters," *Pattern Recognition*, vol. 45, pp. 3017-3033, 2011.

[5] J. W. Chinneck, "Integrated classifier hyperplane placement and feature selection," *Expert Systems with Applications*, vol. 39, pp. 8193-8203, 2012.

[6] E. Alkım, E. Gürbüz, and E. Kılıç, "Afast and adaptive automated disease diagnosis method with and innovative neural network model," *Neural Networks*, Article in press.

[7] H.-L. Chen, B. Yang, J. Liu, and D.-Y. Liu, "A support vector machine classifier with rough set-based feture selection for breast cancer diagnosis," *Expert Systems with Applications*, vol. 38, pp. 9014-9022, 2011.

[8] A. M. Cedeno, J. Q. Dominguez, and A. Andina, "WDBC breast cancer database classification applying artificial metaplasticity neural network," *Expert Sysems with Applications*, vol. 38, pp. 9573-9579, 2011.

[9] N. C. Ramirez, H. G. A. Mesa, H. C. Calvet, and R. E. B. Martinez, "Discovering interobserver variability in the cytodiagnosis of breast cancer using decision trees and Bayesian networks," *Applied Soft Computing*, vol. 9, pp. 1331-1342, 2009.

[10] UCI Repository of Machine Learning databases. [Online]. Available: http://archive.ics.uci.edu/ml

[11] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273:297, 1995.

[12] I. Maglogiannis, E. Zafiropoulos, *et al.*, "An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers," *Applied Intelligence*, vol. 30, no. 1, pp. 24-36, 2009.

[13] M. A. H. Farquad and I. Bose, "Prepocessing unbalanced data using support vector machine," *Decision Support Systems*, vol. 53, pp. 226-233, 2012.

[14] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," *Technical Report Department of Computer Science, National Taiwan University, Taipei, Taiwan.* 2008.

[15] D. Karaboğa, B. Görkemli, C. Öztürk, and N. Karaboğa, "A compherensive survey: artificial bee colony (ABC) algorithm and applications," *Artificial Intelligence Rev.* Article in press.

[16] T. H. Cormen, C. E. Leiserson, R. L. Rivest. And C. Stein. *Introduction to Algorithms, 2nd edition*. MIT Press, 2001.

[17] C.-C. Chang and C.-J. Lin, LIBSVM: a library for support vector machines, 2001. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm.