A Multi-Objective Artificial Bee Colony Approach to Feature Selection Using Fuzzy Mutual Information

Emrah Hancer[*†], Bing Xue[†], Mengjie Zhang[†], Dervis Karaboga[*] and Bahriye Akay[*]
[*]Department of Computer Engineering, Erciyes University, Kayseri 38039, Turkey
{emrahhancer, karaboga, bahriye}@erciyes.edu.tr
[†]School of Engineering and Computer Science, Victoria University of Wellington, Wellington 6140, New Zealand
{Bing.Xue, Mengjie.Zhang}@ecs.vuw.ac.nz

***Abstract*—Feature selection often involves two conflicting objectives of minimizing the feature subset size and the maximizing the classification accuracy. In this paper, a multi-objective artificial bee colony (MOABC) framework is developed for feature selection in classification, and a new fuzzy mutual information based criterion is proposed to evaluate the relevance of feature subsets. Three new multi-objective feature selection approaches are proposed by integrating MOABC with three filter fitness evaluation criteria, which are mutual information, fuzzy mutual information and the proposed fuzzy mutual information. The proposed multi-objective feature selection approaches are examined by comparing them with three single-objective ABC-based feature selection approaches on six commonly used datasets. The results show that the proposed approaches are able to achieve better performance than the original feature set in terms of the classification accuracy and the number of features. By using the same evaluation criterion, the proposed multi-objective algorithms generally perform better than the single-objective methods, especially in terms of reducing the number of features. Furthermore, the proposed fuzzy mutual information criterion outperforms mutual information and the original fuzzy mutual information in both single-objective and multi-objective manners. This work is the first study on multi-objective ABC for filter feature selection in classification, which shows that multi-objective ABC can be effectively used to address feature selection problems.**

I. Introduction

Classification is the task of categorizing the dataset members into the defined or known classes according to the information described by features. However, not all the information provided by features are necessary or useful for the classification, i.e., irrelevant or redundant features may inversely affect the classification performance and computational time. Therefore, it is better to select a subset of relevant features to reach similar or even higher classification performance. By this way, not only higher classification performance and lower computational time, but also the simplified learnt classifier and the reduced dimensionality of data are obtained [1].

It is difficult to determine a feature as redundant or irrelevant in a feature subset due to the complex interaction between features. For instance, a feature individually may not have a significant effect to the target, but when using together with other features, it may improve the classification performance. Also, a relevant feature individually may become redundant when it is interconnected with others. Another challenging issue making feature selection process difficult is the large search space, the size of which increases exponentially proportional to the number of features in a dataset. Most of the existing feature selection approaches are high time consumption and may converge to local minima [1]. Therefore, evolutionary computation (EC) based algorithms attract attention due to their global search ability. Genetic algorithms (GAs) [2], genetic programming (GP) [3], ant colony optimization (ACO) [4] and particle swarm optimization (PSO) [5]–[7] are the most well-known techniques in feature selection. The researchers also have recently concentrated on the artificial bee colony (ABC) algorithm [8] for feature selection problems.

Feature selection can be handled as a multi-objective optimization problem since it aims to maximize the classification performance and minimize the number of selected features simultaneously. By this way, it is expected to meet different requirements in real world applications by obtaining a set of non-dominated feature subsets. However, treating feature selection as a multi-objective problem has just come into consideration in last decade. In other words, most of the existing EC based feature selection algorithms are based on single objective and there exist only a few studies based on multi-objective EC techniques [9], [10]. ABC is one of the most recent swarm-intelligence algorithms and has been successfully applied to various problems [11]. However, the thought of applying ABC for multi-objective feature selection has not been considered yet.

A. Goals

The overall goal of this paper is to propose an ABC based multi-objective filter feature selection approach to receiving the non-dominated solutions, comprising of a smaller number of features and higher classification performance. To satisfy this goal, a new multi-objective ABC (MOABC) framework is proposed, which is inspired by the non-dominated sorting genetic algorithm II (NSGAII) [12]. The other goal is to propose an improved fuzzy mutual information based criterion. This goal is achieved by using fuzzy relevance between every two features and the class labels, and fuzzy redundancy among the selected features. In addition to the proposed criterion, two existing criteria considering basic mutual information between each feature and the class labels, and fuzzy mutual information between each feature and the class labels are also employed. To our knowledge, fuzzy mutual information is used with multi-objective EC techniques in the study for the first time. Specifically, the following cases are investigated:

- whether a single objective ABC approach with three mutual information criteria can select a small number

978-1-4799-7492-4/15/$31.00 © 2015 IEEE

of features and improve the classification performance over using all features;

- whether MOABC-based multi-objective filter feature selection approaches can choose a smaller number of features and obtain better classification performance than single-objective approaches; and
- whether MOABC based multi-objective approach based on the proposed criterion can perform better than the other two MOABC multi-objective approaches based on the existing criteria.

B. Organization

The rest of the paper is organized as follows. Section II gives an outline of the basic ABC algorithm and provides a background on mutual information and existing feature selection approaches. Section III presents the proposed multi-objective algorithms and Section IV defines the experimental design. Section V presents the experimental results and discussions. Finally, Section VI concludes the study and provides an insight into the future trends.

II. Background

A. Artificial Bee Colony (ABC)

ABC is one of the recently proposed swarm intelligence (SI) based algorithms inspired by foraging behaviours of honey bee swarm [13]. In ABC, there exist three kind of bees, including employed bees, onlookers and scouts. In the hive, each solution is associated with one employed bee, i.e., the number of employed bees is equal to the number of solutions. The number of employed bees is also equal to the number of onlookers, but onlookers tend to investigate better solutions in a probabilistic manner. The number of scouts is not predefined in the hive, i.e., it is generated if there exist any abandoned solution determined by the number of trials, known as "limit".

The basic implementation steps of ABC are as follows. Firstly, the population number, maximum cycle and limit parameters are predefined, and then solutions are randomly initialized by Eq. (1). Secondly, a new solution is evolved in the neighborhood of each solution by its associated employed bee through Eq. (2). If the evolved solution V_i is better than the old one, V_i is memorized. After that, a probabilistic value is calculated using roulette wheel for each solution by Eq. (3), and onlookers then select solutions according to these probabilistic values to evolve a new solution by Eq. (2). The quality solutions have more chance to be selected by the onlookers than the other solutions. After onlookers complete their process, if there exist any abandoned solution exceeding limit value, a new solution is produced by a scout bee through Eq. (1) instead of abandoned one.

$$x_{ij} = x_j^{min} + rand(0,1)(x_j^{max} - x_j^{min}) \tag{1}$$

where X_i is the ith solution s.t. $\{x_{i1}, x_{i2}, x_{i3}, ..., x_{ij}, ..., x_{iD}\}$, $i = \{1, 2, ..., SN\}$ and SN is the number of solutions; $j = \{1, 2, ..., D\}$ and D is the number of dimensionality of the search space; x_j^{min} and x_j^{max} are the minimum and maximum predefined values of parameter j.

$$\upsilon_{ij} = x_{ij} + \phi_{ij}(x_{ij} - x_{kj}) \tag{2}$$

where i represents the index of current solution (X_i); k represents the index of neighbor solution (X_k); j is the randomly selected parameter for modification; V_i is the generated solution determined by modifying one parameter of X_i; and ϕ_{ij} is a random number within $[-1, 1]$.

$$p_i = \frac{fitness_i}{\sum\limits_{i=1}^{SN} fitness_i} \tag{3}$$

where $fitness_i$ is the fitness value of source X_i. After the calculation of probability (p_i), a random number in the range of 0 and 1 ($rand(0,1)$) is generated for each solution i.

B. Fuzzy Entropy and Mutual Information

The most well-known entropy criteria are based on the notion of probability and the notion of possibility [14]. For the probabilistic notion, Shannon entropy is one of the most-widely used measure of uncertainty in the literature, defined by:

$$H\left(X\right) = -\sum_{k} p\left(x_k\right) \log_2 p\left(x_k\right) \tag{4}$$

$$H\left(X, Y\right) = -\sum_{k,z} p\left(x_k, y_z\right) \log_2 p\left(x_k, y_z\right) \tag{5}$$

where $X = \{x_1, x_2, ...x_k, ..., x_n\}$ and $Y = \{y_1, y_2, ...y_z, ..., y_m\}$ are discrete random variables. By using entropy, mutual information between X and Y is calculated by:

$$MI(X;Y) = H(X) + H(Y) - H(X,Y) \tag{6}$$

To integrate the fuzziness into the Shannon entropy, $p(x_k)$ is replaced by $\mu(x_k)$ based probability. Suppose that the mean of the data samples belonging to class i is defined as $\overline{x_i}$ and the radius of the data is defined as $r = \max \|\overline{x_i} - x_k\|_\sigma$. The marjinal and joint fuzzy entropy [15] is defined by:

$$FH(X) = -P_{X_{S_i}} \log P_{X_{S_i}} \tag{7}$$

$$FH(X, c_i) = -P(X, c_i) \log P_{X,c_i} \tag{8}$$

where X is a variable, c_i is the ith class, and P_{X,c_i} is the fuzzy equivalent to the joint probability of the training samples that belong to class i. $P_{X_{S_i}}$ and P_{X,c_i} are defined by:

$$P_{X_{S_i}} = \frac{\sum_k \mu(x_{ik})}{NP} \tag{9}$$

$$P\left(X, c_i\right) = \frac{\sum_{k \in A_i} \mu(x_{ik})}{NP} \tag{10}$$

where $\mu(x_{ik})$ the fuzzy membership of the kth vector in the ith class; calculated by:

$$\mu_{ik} = \left(\frac{\|\overline{x_i} - x_k\|_\sigma}{r \pm \epsilon}\right)^{\frac{-2}{m-1}} \tag{11}$$

where m is the fuzzification parameter and σ is the standard deviation involved in the distance computation. And finally fuzzy mutual information for two different features Y and Y is defined by:

$$FMI(X;Y) = FH(X) - FH(X|Y) \tag{12a}$$

$$FMI(X;Y) = FH(Y) - FH(Y|X) \tag{12b}$$

$$FMI(X;Y) = FH(X) + FH(Y) - FH(X,Y) \tag{12c}$$

C. Existing Feature Selection Approaches

Based on evaluation criterion, feature selection approaches can be categorized into wrapper and filter approaches. While wrapper approaches use a learning/classification algorithm as a part of evaluation criterion to evaluate the goodness of the selected features, filter approaches use statistical characteristics of data to measure the interactions between features. In other words, filter approaches do not depend on a classifier as in wrapper approaches. That usually makes filter approaches more general and computationally less expensive than wrappers.

Traditional Filter Feature Selection Approaches. Hall [16] tried to determine the relation between the features and class labels through correlation. Another correlation based algorithm, Relief is proposed by Kira and Rendell [17]. However, correlation based approaches do not properly work in most situations since they do not consider redundant features. FOCUS [18] deeply searches all possible feature subsets, and then selects the smallest one. However, considering all possible feature subsets makes the algorithm computationally inefficient.

Mutual Information Based Feature Selection Approaches. It is known that maximizing the relevance between each feature and the class labels, and minimizing the redundancy between features can improve the classification performance. One of the most well-known ways to measure the mutual dependence between two features is the mutual information. Mutual information has been used by a number of researchers for feature selection. The most well-known examples are the mutual information based feature selection (MIFS) [19], minimum redundancy and maximum relevance (mRmR) [20], joint mutual information (JMI) [21], and MIFS-U [22]. In detailed, MIFS and mRmR are equivalent except for the usage of a parameter, in which features are added to feature subset according to the max-relevance and min-redundancy criterion. On account of the difficulties on calculating probabilities of continues variables via mutual information, researchers have concentrated on fuzzy mutual information. Khushaba et al. [15] proposed a fuzzy mutual information based wavelet packet transform (FMIWPT) feature selection approach for classifying the driver drowsiness state into one of the predefined drowsiness levels. Yu et al. [23] proposed a fuzzy version of mRmR criterion to classify datasets. The obtained results showed that the proposed approach outperformed the mRmR, Relief and correlation based feature selection approaches.

EC Based Filter Feature Selection Approaches. Based on fuzzzy sets, Chakraborty [2] improved GA and PSO based filter approaches. Wang et al. [5] proposed an approach based on improved binary PSO (BPSO) and rough sets theory. In this BPSO model, velocity is defined as the number of elements. The obtained results showed that BPSO outperformed GA. Cervante et al. [24] used mRmR and information gain (IG) based criteria with BPSO. Unfortunately, the balance weights between relevance and redundancy needs to be determined. To cover this issue, Xue et al. [9], [10], [25], [26] proposed various multi-objective optimization filter feature selection algorithms using NSGAII, strength Pareto evolutionary algorithm 2 (SPEA2), non-dominated sorting based PSO (NSPSO) [27], and crowding distance based PSO (CMBPSO) [28] with mutual information, entropy, basic rough-set and probabilistic rough-set theories. The results showed that multi-objective feature selection approaches can get a smaller number of features and better classification performance than single-objective approaches.

Most of the existing filter based feature selection approaches are single-objective and the thought of handling feature selection as a multi-objective problem has just been investigating for the last 5-6 years. Moreover, there exist only one single-objective ABC based filter feature selection approach [29] in the literature. In addition, feature selection has not been considered in multi-objective ABC concept, and fuzzy mutual information has not been used with multi-objective EC techniques yet. Therefore, the development of multi-objective ABC based feature selection approach, and applying fuzzy mutual information with multi-objective EC techniques is still an open issue.

III. Proposed Feature Selection Approaches

In this section, two exiting mutual information criteria are first described, and then an improved fuzzy mutual information is proposed. Based on the criteria, three single-objective ABC based feature selection approaches (ABC-MI, ABC-FMI and ABC-IFMI) are developed. Then, a new multi-objective ABC framework (MOABC) is designed, and by applying this framework, three new multi-objective feature selection approaches (MOABC-MI, MOABC-FMI and MOABC-IFMI) are also proposed.

A. Single-Objective Approaches Based on Existing Criteria

ABC-MI: Using the components of mRmR to evaluate the relevance between a feature and the class labels, and the redundancy beween two features, Cervante et al. [24] improved a BPSO based filter feature selection approach (PSOfsMI). Inspired by this approach, the objective function based on mRmR is applied to ABC in this study, shown by Eq. (13):

$$F_{mi} = Rel - Red \tag{13a}$$

$$Rel = \sum_{X \in S} MI(X;c) \tag{13b}$$

$$Red = \sum_{X,Y \in S} MI(X;Y) \tag{13c}$$

where X and Y are individual features in the selected feature subset S, and c is the class label; Rel represents the sum of relevance between each feature and the class labels; and Red represents the redundancy among features in the selected feature subset.

ABC-FMI: The fuzzy version of MRMR [23] based on fuzzy relevance between a feature and the class labels, and fuzzy redundancy between two features was first proposed to use the mRmR in continuous feature selection problems. To our knowledge, fuzzy mRmR has not been applied to ABC yet. Thus, the objective function based on fuzzy mutual information is integrated to ABC by Eq. (14):

$$F_{fmi} = FRel - FRed \tag{14a}$$

$$FRel = \sum_{X \in S} FMI(X;c) \tag{14b}$$

$$FRed = \sum_{X,Y \in S} FMI(X;Y) \tag{14c}$$

where $FRel$ represents the sum of fuzzy relevance between each feature and the class labels; and $FRed$ represents the fuzzy redundancy among features.

B. Single-Objective Approach Based on Improved Criterion

The existing feature selection criteria just concern with the relation between one feature and the class labels. In other words, they only consider two-way interactions between features. However, it is known that feature interaction may be more complex than two-ways. To cover this issue, an improved fuzzy mutual information criterion (IFMI) based on relevance between two features and the class labels is proposed by Eq. (15), and it is implemented with ABC, defined as ABC-IFMI.

$$F_{ifmi} = IFRel - FRed \tag{15a}$$

$$IFRel = \sum_{X,Y \in S} FMI(C;X;Y) \tag{15b}$$

$$FMI(C;X;Y) = FH(C) - FH(C|X,Y) \tag{15c}$$

$$\begin{aligned} FH(C|X,Y) &= FH(C,X,Y) - FH(X,Y) \\ &= -\sum_{c,x,y} p(c,x,y) \log p(c,x,y) \\ &\quad + \sum_{x,y} \left(\sum_{x} p(c,x,y) \right) \log p(c,x,y) \end{aligned} \tag{15d}$$

where $IFRel$ representing the sum of fuzzy relevance between every two features and the class labels can be accepted as an improved version of $FRel$. Specifically, F_{ifmi} is diverged from F_{fmi} by applying two features to measure the relevance of them with the class labels. By this way, further probabilities are evaluated to investigate the information between features and the class labels.

C. Proposed Multi-Objective ABC Framework

Gaining from Section II.C, handling feature selection in a multi-objective concept via EC techniques is able to solve feature selection problems more effectively than single-objective design. However, the use of ABC for multi-objective feature selection has not been investigated yet. This motivates us to develop a multi-objective ABC (MOABC) framework for feature selection problems. The proposed MOABC framework is inspired by the idea of NSGAII. Not only non-dominated sorting, but also genetic crossover and mutation mechanisms of NSGAII are integrated to the ABC algorithm. As seen in Algorithm 1, for each current solution, a neighbor solution is selected from set, and then simulated-binary crossover (SBX) [30] is applied between them. By this way, two solutions are generated. After the process of crossover, the polynomial mutation [31] operator is applied to the current and its neighbor solutions. Therefore, four solutions are produced for each original solution. These solutions are then added to the set. The solution selection mechanism for evolution in employed bee phase, the same as standard ABC, is applied through randomness. However, the selection mechanism in onlooker bee phase cannot be applied as in standard ABC since probabilistic values of solutions comprising of more than one objective function values cannot be calculated by Eq. 3. As for calculating the probabilistic values based on fitness values to select a solution for evolution in onlooker bee phase, the fitness assignment mechanism [32], [33] is applied by Eq. (16). At the end of employed bee and onlooker bee phase, non-dominated sorting is applied to the pool of parent solutions and their mutants.

```
begin
    Divide dataset into training and test set;
    Initialize solution set $X = X_1, X_2, ..., X_n$ by Eq. (1);
    Evaluate two objective of solutions;
    // number of features and relevance (Rel in
       MOABC-MI, FRel in MOABC-FMI and IFRel in
       MOABC-IFMI
     Apply non-dominated sorting to solutions;
    for $cycle \leftarrow 1$ to $MCN$ do
        foreach employed bee $i$ do
            Randomly choose a solution $X_k$ in the neighborhood of $X_i$;
            Apply crossover between $X_i$ and $X_k$;
            Apply mutation to $X_i$ and $X_k$;
            Evaluate two objectives of evolved solutions;
            Add evolved solutions to $X$;
        end
        Apply non-dominated sorting on $X$;
        Select best $SN$ solutions based on rank and crowding distance to
        renew population;
        foreach onlooker bee $i$ do
            Select a food source $X_i$ depending on probability $p_i$;
            Randomly choose a solution $X_k$ in the neighborhood of $X_i$;
            Apply crossover between $X_i$ and $X_k$;
            Apply mutation to $X_i$ and $X_k$;
            Evaluate two objectives of evolved solutions;
            Add evolved solutions to $X$;
        end
        Apply non-dominated sorting on $X$;
        Select best $SN$ solutions based on rank and crowding distance to
        renew population;
        if there exits an abondoned solution then
            Scout bee determines a new solution by Eq. (1);
        end
    end
    Calculate the classification accuracy of the feature subsets (solutions) in
    the Front 1 on the test set;
    Return the solutions and their classification accuracy rates;
end
```

Algorithm 1: Pseudo-code of MOABC-MI, MOABC-FMI and MOABC-IFMI

$$fitness_i = \frac{1}{R(i) - TS(i) - d(i)} \tag{16}$$

where $R(i)$ is the Pareto font rank value of the solution i; $T > 0$ is a temperature; $d(i)$ is the crowding distance; and $S(i)$ is calculated by:

$$S(i) = -p_T(i) \log_{p_T}(i) \tag{17a}$$

$$p_T(i) = (1/Z) \exp(-R(i)/T) \tag{17b}$$

$$Z = \sum_{1}^{N} \exp(-R(i)/T) \tag{17c}$$

where $p_T(i)$ is the Gibbs distribution [34] and N is the population size.

Based on MOABC, three filter multi-objective approaches are proposed: 1) MOABC-MI based on the number of features and Rel (Eq. (13b)), 2) MOABC-FMI based on the number of features and $FRel$ (Eq. (14b)), and 3) MOABC-IFMI based on the number of features and $IFRel$ (Eq. (15b)). The solutions

TABLE I: Datasets

Data set	#Samples	#Features	#Classes
Lymph	148	18	4
Mushroom	8124	22	2
Dermatology	366	34	6
Soybean large	307	35	19
Chess	3196	36	2
Connect4	44473	42	3

acting as probable feature subsets are represented via continuous variables in the range between 0 and 1 in both ABC and MOABC approaches. Each dimension of a solution indicates the condition (selected/unselected) of a related feature among all available features for the feature subset. If the value of a dimension is equal to or more than 0.5, regarding feature is chosen for the feature subset; otherwise, it is not chosen. Note that, the objectives of a solution are calculated after the determination of feature subset.

IV. Experimental Design

The performance anaylsis of the feature selection approaches is established on six benchmark datasets (listed in Table 1), which were chosen from the UCI machine learning repository [35]. The datasets comprise of various number of samples, features and classes, satisfying a comprehensive analysis of the feature selection approaches. For each dataset, samples are randomly separated into two sets: 70% as the training set and 30% as the test set [36].

The feature selection algorithms first run on the training set to get an optimal feature subset(s). The performance of the obtained feature subset(s) is then evaluated by a classification algorithm on the test set. For both standard and multi-objective ABC, the population size, limit value and maximum number of evaluations are set to 50, 100 and 10000, respectively. Each feature selection approach has been implemented for 30 independent runs on each dataset. For the evaluation of the classification performance on the test set, one of the most widely used classifier, KNN, is employed. In KNN, the number of nearest neighbours, K is chosen as 5. The classification performance is calculated by:

$$Accuracy = \frac{TP + TN}{FP + FN + TP + TN} \tag{18}$$

where TP and TN are true positives and negatives, and FP and FN are false positives and negatives.

While single-objective approaches get a single result in each of 30 runs for each dataset, multi-objective approaches get a non-dominated solution set in each run. Over 30 runs, single-objective approaches get 30 solutions and multi-objective approaches get 30 sets. The sets achieved by multi-objective approaches are then collected into one union set. The results in the union set can be presented in two ways. The first way is to present the mean values of the classification accuracies having the same feature subset size, referred as "average Pareto front". The other way is to present the non-dominated solutions in the union set.

V. Experimental Results

The results of MOABC-MI, MOABC-FMI, MOABC-IFMI based on the three criteria, i.e. mutual inforation, fuzzy mutual information and the improved fuzzy mutual information, are shown in Figs 1, 2, and 3, respectively. Fig 4 further shows the non-dominated solutions achieved by MOABC-MI, MOABC-FMI and MOABC-IFMI over the 30 independent runs. On the top of each chart, the number of available features and the classification accuracy using all features are presented in the brackets. In the charts, "-A" represents the "average Pareto front" and "-B" represents the non-dominated solutions over the 30 independent runs, and the points represent the 30 solutions achieved by the single-objective algorithms in the 30 independent runs. In some charts of Fig. 1 and Fig. 2, to clearly show the difference between the single-objective and multi-objective approaches, the classification accuracies for 1 or 2 features are not shown, but they are presented in Fig. 4. Moreover, there may be fewer than 30 distinct points for single-objective approaches in a chart since single-objective approaches may evolve the same feature subset(s) in different runs, which are shown in the same point in the charts.

A. *MOABC-MI vs. ABC-MI*

From Fig. 1, it can be observed that ABC using mutual information as the evaluation criterion can reduce around 75% of the available features in most cases, and the classification performance is often similar or slightly worse than using all features. For instance, the feature subset size obtained by ABC-MI is between 8 and 11 on the Soybean datasets with 35 available features. Fig. 1 also shows that in almost all cases, MOABC-MI can obtain smaller feature subsets and similar or higher classification accuracy than using all features. For example, on the Connect4 dataset, MOABC-MI reduced the dimensionality (i.e. number of features) from 42 to 9, but increases the classification accuracy from 71.69% to 92.52%. The results show that MOABC-MI is able to or has the potential to significantly reduce the dimensionality of the data and maintain or increase the classification performance, which suggests that the proposal of using MOABC and mutual information for feature selection is successful.

By comparing ABC-MI with MOABC-MI, it can be seen that most of the ABC-MI points are appeared under the lines of both MOABC-A and MOABC-B, which shows that MOABC-MI outperforms ABC-MI in terms of the number of features and the classification performance except for Dermatology. The comparisons suggest that MOABC employs the multi-objective search mechanism can explore the search space more effectively than single-objective ABC to find better feature subsets.

B. *MOABC-FMI vs. ABC-FMI*

When considering the FMI based approaches, according to Fig. 2, ABC-FMI performs better than using all features in terms of feature subset size, and the classification performance is often similar or slightly worse than using all features like as ABC-MI. Fig. 2 also shows that MOABC-FMI generally can achieve higher classification accuracy and a smaller number of features than using all features. For instance, MOABC-MI can reach 100% accuracy and the number of selected features can be reduced to 10 in Mushroom with 22 features.

By comparing ABC-FMI with MOABC-FMI, it can be observed that MOABC-FMI is superior to ABC-FMI (except for the Lymph and Dermatology datasets) since most of the

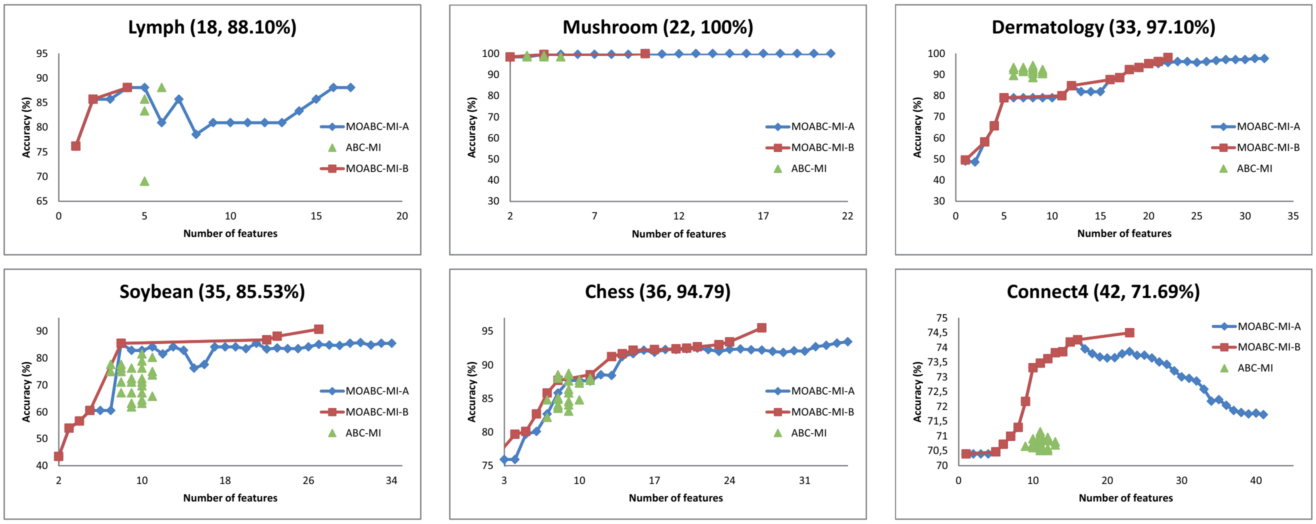

Fig. 1: Experimental results of ABC-MI and MOABC-MI

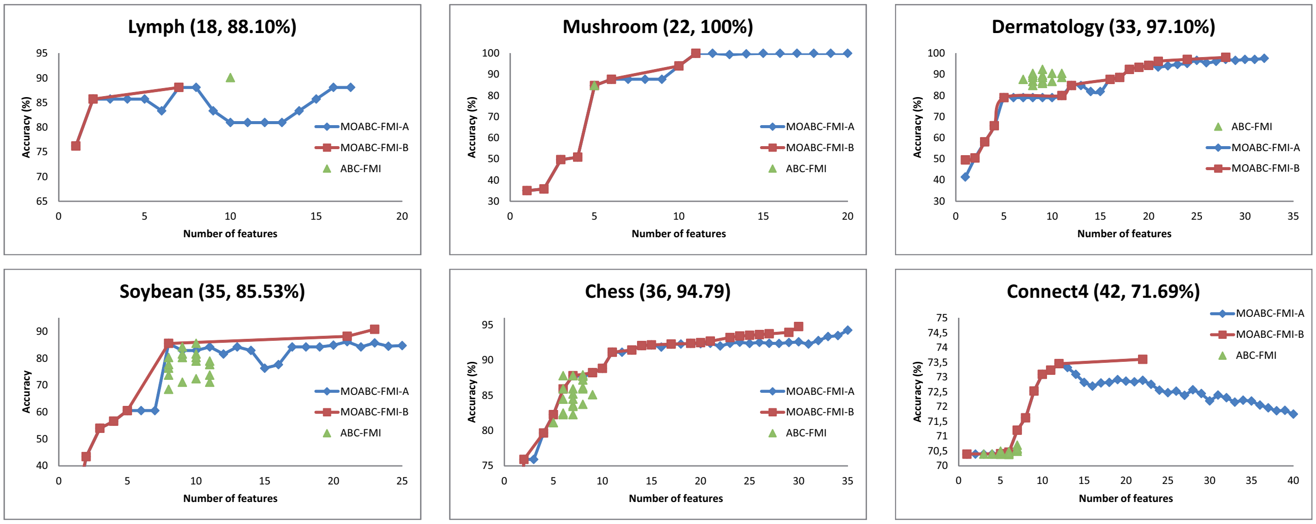

Fig. 2: Experimental results of ABC-FMI and MOABC-FMI

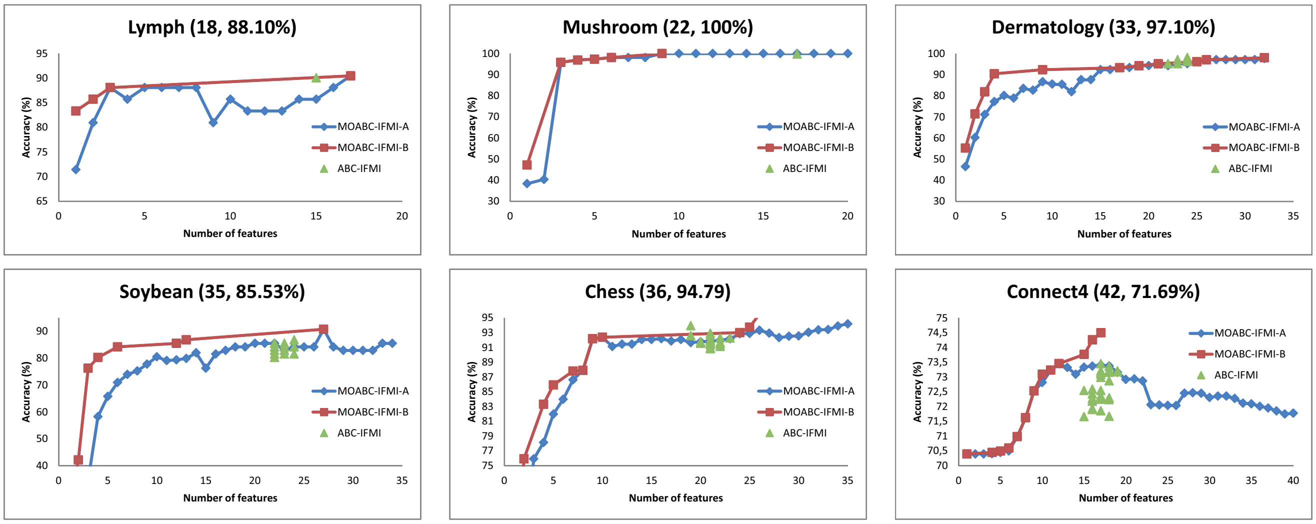

Fig. 3: Experimental results of ABC-IFMI and MOABC-IFMI

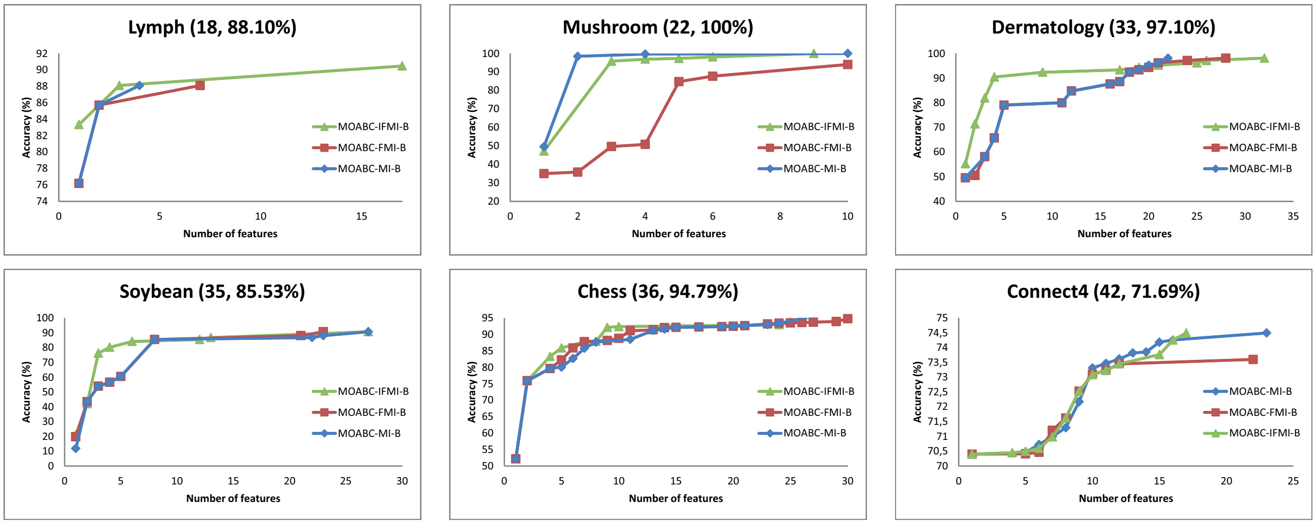

Fig. 4: Experimental results of MOABC-MI-B, MOABC-FMI-B and MOABC-IFMI-B

points obtained from ABC-FMI are dominated by the results shown by MOABC-FMI-B and MOABC-FMI-A. The Soybean dataset can be easily given as a typical example. This further confirms that using a multi-objective search mechanism in ABC could find better solutions than single-objective ABC regardless of the filter measure as mutual information or fuzzy mutual information.

C. MOABC-IFMI vs. ABC-IFMI

For the approaches using the proposed IFMI criterion, Fig. 3 shows that ABC-IFMI also brings advantages versus using all features in terms of the classification performance and the number of features needed for classification. The results in Fig. 3 also show that MOABC-IFMI with the proposed IFMI as the evaluation criterion can be successfully used for multi-objective feature selection to obtain a set of non-dominated solutions, which reduce the number of features and maintain or increase the classification performance.

Fig. 3 also indicates that MOABC-IFMI can eliminate irrelevant or redundant features more effectively than ABC-IFMI to select much smaller feature subsets, although they have similar classification accuracies in most cases. This further shows the advantage of using multi-objective ABC in feature selection, i.e. optimizing two objectives separately can at least maintain the classification performance obtained by single-objective ABC, but further optimize (i.e. reduce) the number of features.

D. Comparisons between Different Criteria

To compare the three different evaluation criteria, Figs. 1, 2 and 3 show that ABC-IFMI is able to obtain higher classification accuracy than ABC-MI and ABC-FMI. This suggests that the proposed IFMI measure can better reflect the relevance between a subset of features and the class labels to improve the classification accuracy. However, the number of features selected by ABC-IFMI is often larger than that of ABC-MI and ABC-FMI. This is not unexpected since there is often a trade-off between the number of features and the classification performance, which is also the motivation of investigating multi-objective ABC for feature selection.

To further compare the three criteria in multi-objective approaches, Fig. 4 shows the non-dominated solutions obtained by MOABC-MI, MOABC-FMI, and MOABC-IFMI in the 30 independent runs, i.e. shown by “-B”. The results indicate that the proposed IFMI outperforms or achieves similar classification performance to MI and FMI on five out of the six datasets. Particularly, when the number of selected features is small, for example on the Soybean and Dermatology datasets, the classification performance of IFMI is much higher than that of MI and FMI. This further shows that the proposed IFMI criterion is a better filter measure than MI and the standard FMI.

VI. Conclusions

The overall goal of this paper was to propose an ABC based multi-objective approach to feature selection in classification. This goal was achieved by developing a multi-objective ABC framework (MOABC) inspired by the NSGAII algorithm. The other goal was to propose a filter criterion to measure the relevance between a subset of features and the class labels, which was achieved by developing an improved fuzzy mutual information measure (IFMI) by integrating the relevance between two features and the class labels into the original fuzzy mutual information measure. Then three multi-objective algorithms (MOABC-MI, MOABC-FMI and MOABC-IFMI) were proposed for feature selection, which were based on mutual information, fuzzy mutual information, and the proposed improved fuzzy mutual information, respectively. The performances of the proposed algorithms were demonstrated by comparing them with three single-objective algorithms (ABC-MI, ABC-FMI and ABC-IFMI) in terms of the classification accuracy and the number of features on six benchmark classification tasks.

The experimental results show that the proposed single-objective algorithms can remove irrelevant and/or redundant features effectively and maintain the classification performance in most cases. The proposed multi-objective algorithms can evolve a set of non-dominated feature subsets that include smaller feature subsets and higher classification accuracy than using all features, and they can further remove irrelevant or redundant features over single-objective algorithms without decreasing or even increasing the classification performance. By comparing the methods using different evaluation criteria, the proposed IFMI measure is superior to the mutual information measure and the original fuzzy mutual information measure, especially in terms of the classification performance.

This paper represents the first ABC-based multi-objective approach for feature selection and is also the first study applying fuzzy mutual information with multi-objective EC techniques for feature selection. In future, we will further examine the generality of the proposed algorithms and compare them with other EC-based multi-objective approaches. We also intend to develop novel multi-objective ABC based approaches to better search the Pareto front of non-dominated solutions in large-scale feature selection problems.

REFERENCES

[1] M. Dash and H. Liu, "Feature selection for classification," *Intelligent Data Analysis*, vol. 1, no. 14, pp. 131–156, 1997.

[2] B. Chakraborty, "Genetic algorithm with fuzzy fitness function for feature selection," in *Proceedings of the IEEE International Symposium on Industrial Electronics (ISIE'2002)*, vol. 1, 2002, pp. 315–319 vol.1.

[3] K. Neshatian and M. Zhang, "Pareto front feature selection: Using genetic programming to explore feature space," in *11th Annual Conference on Genetic and Evolutionary Computation*, 2009, pp. 1027–1034.

[4] R. Jensen, "Performing feature selection with ACO," in *Swarm Intelligence in Data Mining*, ser. Studies in Computational Intelligence. Springer Berlin Heidelberg, 2006, vol. 34, pp. 45–73.

[5] X. Wang, J. Yang, X. Teng, W. Xia, and R. Jensen, "Feature selection based on rough sets and particle swarm optimization," *Pattern Recognition Letters*, vol. 28, no. 4, pp. 459 – 471, 2007.

[6] B. Xue, M. Zhang, and W. Browne, "Novel initialisation and updating mechanisms in PSO for feature selection in classification," in *Applications of Evolutionary Computation*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, vol. 7835, pp. 428–438.

[7] L. Cervante, B. Xue, L. Shang, and M. Zhang, "A dimension reduction approach to classification based on particle swarm optimisation and rough set theory," in *AI 2012: Advances in Artificial Intelligence*. Springer, 2012, pp. 313–325.

[8] M. Schiezaro and H. Pedrini, "Data feature selection based on artificial bee colony algorithm," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, pp. 1–8, 2013.

[9] B. Xue, L. Cervante, L. Shang, and M. Zhang, "A particle swarm optimisation based multi-objective filter approach to feature selection for classification," in *PRICAI*, ser. Lecture Notes in Computer Science, 2012, vol. 7458, pp. 673–685.

[10] B. Xue, L. Cervante, L. Shang, W. Browne, and M. Zhang, "A multi-objective PSO for filter-based feature selection in classification problems," *Connection Science*, vol. 24, no. 2-3, pp. 91–116, 2012.

[11] D. Karaboga, B. Gorkemli, C. Ozturk, and N. Karaboga, "A comprehensive survey: artificial bee colony ABC algorithm and applications," *Artificial Intelligence Review*, vol. 42, no. 1, pp. 21–57, 2014.

[12] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.

[13] D. Karaboga, "Artificial bee colony algorithm," *Scholarpedia*, vol. 5, no. 3, p. 6915, 2010.

[14] R. Khushaba, A. Al-Jumaily, and A. Al-Ani, "Novel feature extraction method based on fuzzy entropy and wavelet packet transform for myoelectric control," in *International Symposium on Communications and Information Technologies (ISCIT'07)*, 2007, pp. 352–357.

[15] R. Khushaba, S. Kodagoda, S. Lal, and G. Dissanayake, "Driver drowsiness classification using fuzzy wavelet-packet-based feature-extraction algorithm," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 121–131, 2011.

[16] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999.

[17] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Proceedings of the Ninth International Workshop on Machine Learning*, ser. ML92. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1992, pp. 249–256.

[18] H. Almuallim and T. G. Dietterich, "Learning boolean concepts in the presence of many irrelevant features," *Artificial Intelligence*, vol. 69, pp. 279–305, 1994.

[19] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on Neural Networks*, vol. 5, no. 4, pp. 537–550, 1994.

[20] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[21] H. H. Yang and J. Moody, "Data visualization and feature selection: New algorithms for nongaussian data," in *in Advances in Neural Information Processing Systems*. MIT Press, 1999, pp. 687–693.

[22] N. Kwak and C.-H. Choi, "Input feature selection for classification problems," *IEEE Transactions on Neural Networks*, vol. 13, no. 1, pp. 143–159, 2002.

[23] D. Yu, S. An, and Q. Hu, "Fuzzy mutual information based min-redundancy and max-relevance heterogeneous feature selection," *International Journal of Computational Intelligence Systems*, vol. 4, no. 4, pp. 619–633, 2011.

[24] L. Cervante, B. Xue, M. Zhang, and L. Shang, "Binary particle swarm optimisation for feature selection: A filter based approach," in *IEEE Congress on Evolutionary Computation (CEC'2012)*, 2012, pp. 1–8.

[25] B. Xue, L. Cervante, L. Shang, W. N. Browne, and M. Zhang, "Multi-objective evolutionary algorithms for filter based feature selection in classification," *International Journal on Artificial Intelligence Tools*, vol. 22, no. 04, p. 1350024, 2013.

[26] ——, "Binary PSO and rough set theory for feature selection: A multi-objective filter based approach," *International Journal of Computational Intelligence and Applications*, vol. 13, no. 02, p. 1450009, 2014.

[27] X. Li, "A non-dominated sorting particle swarm optimizer for multiobjective optimization," in *Genetic and Evolutionary Computation Conference*, 2003, vol. 2723, pp. 37–48.

[28] M. R. Sierra and C. A. C. Coello, "Improving pso-based multi-objective optimization using crowding, mutation and epsilon-dominance," in *Evolutionary Multi-Criterion Optimization*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2005, vol. 3410, pp. 505–519.

[29] N. Suguna and K. G. Thanushkodi, "An independent rough set approach hybrid with artificial bee colony algorithm for dimensionality reduction," *American Journal of Applied Sciences*, vol. 8, no. 3, pp. 261–266, 2011.

[30] R. B. Agrawal, K. Deb, K. Deb, and R. B. Agrawal, "Simulated binary crossover for continuous search space," Tech. Rep., 1994.

[31] K. Liagkouras and K. Metaxiotis, "An elitist polynomial mutation operator for improved performance of moeas in computer networks," in *22nd International Conference on Computer Communications and Networks (ICCCN'2013)*, 2013, pp. 1–5.

[32] X. Zou, M. Liu, L. Kang, and J. He, "A high performance multi-objective evolutionary algorithm based on the principles of thermodynamics," in *Parallel Problem Solving from Nature - PPSN VIII*, ser. Lecture Notes in Computer Science, 2004, vol. 3242, pp. 922–931.

[33] B. Akay, "Synchronous and asynchronous pareto-based multi-objective artificial bee colony algorithms," *Journal of Global Optimization*, vol. 57, no. 2, pp. 415–445, 2013.

[34] L. D. Landau and E. M. Lifshitz, *Statistical Physics. Course of Theoretical Physics 5*, 3rd ed. Oxford: Pergamon Press, 1980.

[35] K. Bache and M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[36] B. Xue, Z. Mengjie, and W. N. Browne, "Particle swarm optimization for feature selection in classification: A multi-objective approach," *IEEE Transactions on Cybernetics*, vol. 43, no. 6, pp. 1656–1671, 2013.