Hindawi
Complexity
Volume 2018, Article ID 3241489, 14 pages
https://doi.org/10.1155/2018/3241489

Research Article

Effective Evolutionary Multilabel Feature Selection under a Budget Constraint

Jaesung Lee, Wangduk Seo, and Dae-Won Kim

School of Computer Science and Engineering, Chung-Ang University, 221 Heukseok-dong, Dongjak-gu, Seoul 06974, Republic of Korea

Correspondence should be addressed to Dae-Won Kim; dwkim@cau.ac.kr

Received 29 September 2017; Accepted 21 January 2018; Published 7 March 2018

Academic Editor: Kevin Wong

Multilabel feature selection involves the selection of relevant features from multilabeled datasets, resulting in improved multilabel learning accuracy. Evolutionary search-based multilabel feature selection methods have proved useful for identifying a compact feature subset by successfully improving the accuracy of multilabel classification. However, conventional methods frequently violate budget constraints or result in inefficient searches due to ineffective exploration of important features. In this paper, we present an effective evolutionary search-based feature selection method for multilabel classification with a budget constraint. The proposed method employs a novel exploration operation to enhance the search capabilities of a traditional genetic search, resulting in improved multilabel classification. Empirical studies using 20 real-world datasets demonstrate that the proposed method outperforms conventional multilabel feature selection methods.

1. Introduction

Multilabel classification has emerged as a promising technique for various applications, including lifelong structure monitoring [1], functional proteomics [2], and sentiment analysis [3]. These applications produce a series of labels for describing complicated concepts, which are compounded when high-level concepts are composed of multiple subconcepts, such as the environmental and operational conditions of structures [1, 4, 5]. Let $W \subset \mathbb{R}^d$ denote a set of patterns constructed from a set of features F. Then, each pattern $w_i \in W$, where $1 \le i \le |W|$, is assigned to a certain label subset $\lambda_i \subseteq L$, where $L = \{l_1, \dots, l_{|L|}\}$ and is a finite set of labels. Therefore, the task of multilabel classification is to identify a function that maps given instances into one of $2^{|L|}$ label subsets based on input feature values.

In practice, there can be a maximum number of features allowed because of the limits on data acquisition rates or energy consumption [6–8]. In reality, for example, this problem can emerge from the music applications on lightweight mobile devices. Applications for mobile devices typically have a limitation in computational capacity and there is a maximum number of allowed features to be extracted [9, 10]. This is because an overly excessive number of extracted features on mobile devices causes consumers to suffer low quality user experience due to unacceptable waiting or battery consumption.

Given input data with an original feature set F and label set L, the goal of our multilabel feature selection problem is to identify a feature subset $S \subset F$ with the maximum number of features n that yields the best multilabel classification accuracy [11, 12]. This problem is known as budgeted feature selection [13] or feature selection with test cost constraints [8, 14, 15]. However, most studies have been conducted from the perspective of traditional single-label learning. It should be noted, especially when a given constraint n is small, that our multilabel feature selection problem becomes more challenging in terms of classification accuracy due to the fact that a small number of features must support multiple labels simultaneously [16–19].

Multilabel feature selection methods can be categorized according to how they assess the importance of candidate feature subsets [16, 20–22]. Filter-based multilabel feature selection methods identify a final feature subset by focusing on the intrinsic discriminative power of features [21, 23–25]. Some multilabel learning algorithms have a feature selection process embedded in their learning process [26, 27]. In contrast, wrapper-based multilabel feature selection methods assess the importance of feature subsets through a search process by using a multilabel classifier directly. This typically results in better classification accuracy [11, 12]. For this reason, we focus on a multilabel feature wrapper based on an evolutionary search process [28].

During the search process, each chromosome represents a feature subset and selects a number of features less than or equal to n. As a result, most features remain unselected by any chromosome in the population. This can lead to an ineffective search because important features can be continuously neglected. Without negatively affecting the strength of the evolutionary search, this problem can be solved by adding additional chromosomes that convey promising unselected features to the population. In this study, we propose an effective multilabel feature wrapper while considering the constraint of feature subset size. Experimental results demonstrate that the proposed method is able to identify an effective feature subset for multilabel classification with the aid of an enhanced evolutionary search process.

2. Related Work

In traditional single-label feature selection, the budgeted feature selection problem is treated as a special case of the feature selection problem where the algorithm should consider the effectiveness of the feature subset and the acquisition cost for gathering each feature simultaneously. To solve this problem, Zhang et al. [29] proposed a feature selection algorithm based on the bare bones particle swarm optimization, which considers the complexity of an algorithm due to additional parameters. Because the acquisition cost for each feature can be unequal, multiobjective particle swarm optimization approach for cost-based feature selection and return-cost-based binary firefly algorithm for feature selection are also studied [30, 31] which have another objective function of minimizing the cost sum of features.

In multilabel feature selection studies, one of the major trends is the application of a feature selection method for single-label problems by transforming multilabel datasets into single-label datasets [32, 33]. Although this strategy facilitates the use of conventional methods, which has advantages in terms of ease of use [34], algorithm adaptation strategies that directly manage multilabel problems have also been considered [35]. In these approaches, which are largely filter-based, a feature subset is obtained by optimizing a specific criterion, such as a joint learning criterion that involves simultaneous feature selection and multilabel learning [27, 36], $l_{2,1}$-norm function optimization [37], label ranking error [26], Hilbert-Schmidt independence criterion [23], F-statistics [21], or mutual information [16, 24, 38]. However, these methods commonly suffer from low multilabel classification accuracy because of a lack of interaction with multilabel classifiers.

As a notable multilabel feature wrapper study, Zhang et al. [12] proposed a multilabel feature selection method based on a genetic algorithm (GA), which is the most common choice in evolutionary feature wrapper studies [28]. Specifically, their method combined instance- and label-based evaluation metrics [39] as a fitness function to determine label dependency. However, in the original proposal, a maximum number of features to be selected were not considered during the genetic search process. The multilabel classification performance when considering the number of features to be selected was later demonstrated for comparison purposes [11]. During initialization, this method creates chromosomes by selecting a number of features less than n. During the genetic search process, this constraint is continuously satisfied by employing restrictive crossover and mutation operators [40] that immediately discard features randomly if the number of selected features exceeds n. Although this method satisfies the constraint, important features may be discarded, resulting in an ineffective feature subset.

Recent multilabel feature wrapper methods have treated the number of features to be selected as a secondary objective to be achieved by the evolutionary search process (i.e., multiobjective optimization [28]). This is achieved through a specifically designed ranking method for multiobjective optimization problems, known as nondominated sort [41], where the rank of each chromosome is based on the number of times it dominates other chromosomes in terms of two fitness values: multilabel classification accuracy and the number of selected features. Because the ranking of the chromosomes can be determined, it can be directly used in the natural selection process of a GA. Although the most common approach using a nondominated sorting method is NSGA-II [42], nondominated sorting has also been employed in other evolutionary search methods, including particle swarm optimization (PSO) [43]. A common drawback in these methods is that no solution may satisfy the feature number constraint if such a solution is not included in the final Pareto front. Additionally, they may suffer from unnecessary searches of infeasible solutions conveying unacceptable number of features.

Our review indicates that conventional multilabel feature wrappers can fail to identify a final solution that satisfies a given constraint. To remedy this limitation, in addition to the evolutionary process, it is necessary to devise a new process, namely, exploration operation, to find important features in a large set of novel features with the aid of an effective filter and supply them to the population to enhance the evolutionary search process. We summarize subsequent issues and corresponding reasons to our approach as follows.

(i) The exploration operation must be able to identify promising features in a large unselected feature set size of $O(|F| - n) = O(|F|)$. To achieve this, we employ a criterion that measures the relevance score of features.

(ii) The exploration operation must be computationally efficient to circumvent performance degradation of the entire search process. To achieve this, we employ a multilabel feature filter that is confirmed to be efficient because it only requires the dependency between two variables [16].

(iii) Our exploration operation is designed to incur no additional parameter that may cause complicated

```
(1) procedure Proposed algorithm($v, m$)                    ▷ allowed FFC $v$
(2)   $t \leftarrow 0, u \leftarrow 0$                                       ▷ $t$-th generation
(3)   initializing $P(t)$                        ▷ population $P$ of $t$-th generation
(4)   evaluating $P(t)$
(5)   While $u \leq v$ do                          ▷ if spent FFC $u$ is less than $v$
(6)       create $G(t)$ using genetic operators
(7)       create $E(t)$ using exploration operator based on $G(t)$
(8)       $N(t) \leftarrow \{G(t) \cup E(t)\}$                        ▷ offspring set $N(t)$
(9)       evaluate $N(t)$ using a multi-label classifier
(10)      add $N(t)$ to $P(t)$
(11)      $t \leftarrow t + 1$
(12)      select $P(t)$ from $P(t-1)$                           ▷ natural selection
(13)      $u \leftarrow m + 2 \cdot |G(t)| \cdot t$              ▷ update $u$ based on spent FFC
(14)  end while
(15) end procedure
```

Algorithm 1: Procedures of proposed multilabel feature wrapper.

parameter control issues and increase the overall complexity of the algorithm [11, 44]. Based on the number of features given by the evolutionary search, it automatically identifies an effective feature subset that is composed only of novel features.

3. Proposed Method

3.1. Motivation and Approach. In this study, we enhance the performance of a population-based search, such as a GA, for multilabel feature selection with a budget constraint by introducing novel chromosomes that inject promising unselected features into the population. Figure 1 reveals several key issues that should be considered when introducing novel features into the evolutionary search-based multilabel feature selection process with a budget constraint. In the original feature set F, there may be a subset of important features that are strongly dependent on multiple labels, leading to excellent discriminative power in the multilabel classifier if they are included in the final feature subset. After a random initialization process is completed, important features, such as f_1, may be unselected by any chromosome (feature subset) because each chromosome only covers a small number of features under the budget constraint n. It should be noted that $\lceil |F|/n \rceil$ chromosomes should be evaluated to consider all the features at least once, even though all chromosomes are forced to select disjoint feature subsets, which incurs an expensive computational cost. Instead, the proposed method identifies promising features with the help of the employed filter without explicit evaluation of candidate feature subsets.

Next, genetic operators, such as crossovers and mutations, are applied to the population to create new chromosomes. However, unselected important features may not be considered because new chromosomes are created by exchanging the alleles of their ancestors. This means that if ancestors commonly unselect a feature, then their offspring will also unselect that feature. The only chance to add neglected features into the offspring creation process is through the use of a mutation operation. However, this is computationally inefficient because the mutation operation is done by selecting features randomly and, additionally, the mutation rate is set to a small value in order to achieve the convergence. Thus, a large number of iterations or generations should be spent to introduce important features into the population randomly.

In the proposed method, the exploration operator is applied to each of the new offspring to create novel chromosomes that contain promising features that were not considered by the original offspring. During each exploration operation, we calculate the dependency of unselected features on multiple labels $(l_1, l_2, \ldots, l_8)$. After the ranking of each feature is computed (e.g., $f_1 \rightarrow f_{44} \rightarrow f_{32} \rightarrow f_3 \rightarrow \cdots$), a new chromosome that selects the most promising features is created. Finally, exploration and genetic operation-based chromosomes are then merged into a single population.

This paper presents an effective evolutionary search method that remedies the aforementioned issues. In Section 3.2, we discuss the procedural steps of the proposed method and how to handle the issues associated with the exploration operation and the creation of new chromosomes. Section 3.3 presents a mutual-information-based search method for efficiently capturing the relationships between features and labels.

3.2. Algorithm. Algorithm 1 outlines the pseudocode for the procedures used in the proposed method. The terms used for describing the algorithm are summarized in "Terms Used in This Study and Meanings" section. The feature selection vector in a chromosome is a binary string where each bit represents an individual feature, with values of one and zero representing selected and unselected features, respectively. In the initialization step (line (3)), the algorithm generates m chromosomes via random assignment of maximum n binary bits. The selected feature subset (S_c) encoded in $c \in P(t)$ is then evaluated using a fitness function. We use multilabel classification error as the fitness function for the selected

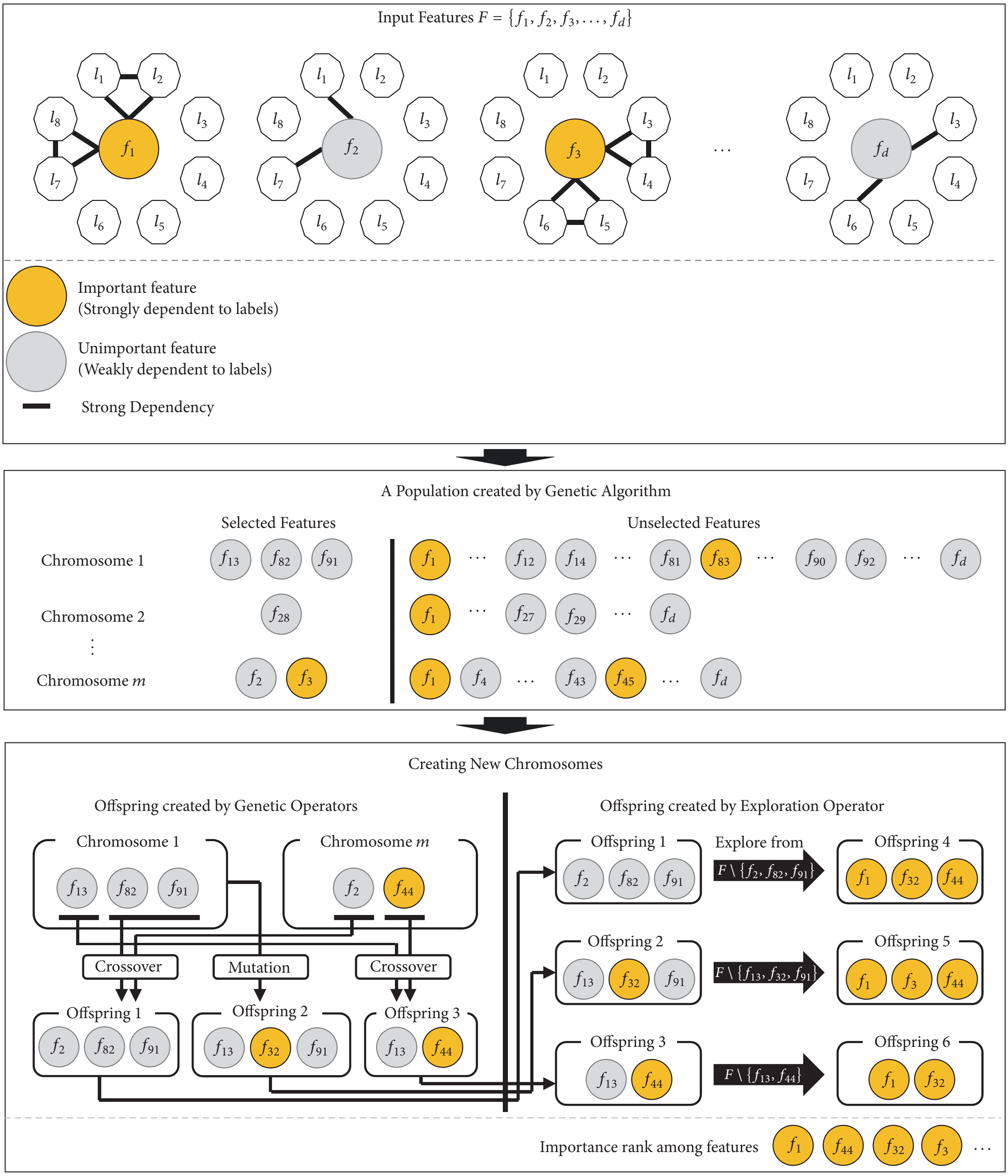

Input Features $F = \{f_1, f_2, f_3, \dots, f_d\}$
l_1 l_2 l_3 l_4 l_5 l_6 l_7 l_8
f_1
f_2
f_3
$\cdots$
f_d
Important feature
(Strongly dependent to labels)
Unimportant feature
(Weakly dependent to labels)
Strong Dependency
A Population created by Genetic Algorithm
Selected Features
Unselected Features
Chromosome 1
f_{13} f_{82} f_{91}
f_1 $\cdots$ f_{12} f_{14} $\cdots$ f_{81} f_{83} $\cdots$ f_{90} f_{92} $\cdots$ f_d
Chromosome 2
f_{28}
f_1 $\cdots$ f_{27} f_{29} $\cdots$ f_d
$\vdots$
Chromosome m
f_2 f_3
f_1 f_4 $\cdots$ f_{43} f_{45} $\cdots$ f_d
Creating New Chromosomes
Offspring created by Genetic Operators
Chromosome 1
f_{13} f_{82} f_{91}
Chromosome m
f_2 f_{44}
Crossover
Mutation
Crossover
Offspring 1
f_2 f_{82} f_{91}
Offspring 2
f_{13} f_{32} f_{91}
Offspring 3
f_{13} f_{44}
Offspring created by Exploration Operator
Offspring 1
f_2 f_{82} f_{91}
Explore from
$F \setminus \{f_2, f_{82}, f_{91}\}$
Offspring 4
f_1 f_{32} f_{44}
Offspring 2
f_{13} f_{32} f_{91}
$F \setminus \{f_{13}, f_{32}, f_{91}\}$
Offspring 5
f_1 f_3 f_{44}
Offspring 3
f_{13} f_{44}
$F \setminus \{f_{13}, f_{44}\}$
Offspring 6
f_1 f_{32}
Importance rank among features
f_1 f_{44} f_{32} f_3 $\cdots$

Figure 1: The cooperation process between genetic and exploration operation.

feature subset. Because m chromosomes must be evaluated in order to obtain their fitness values, m fitness function calls (FFCs) are used in line (4).

After performing the initialization process, the proposed method performs a reproduction process that can be divided into two parts: reproduction via genetic operators and reproduction via the exploration operator. First, the proposed method creates an offspring set $G(t)$ (line (6)) using restrictive crossover and mutation operators to control the number of selected features [40]. Next, the exploration

```
procedure Explore($G(t)$)
    $E(t) \leftarrow \{\emptyset\}$
    for each $c \in G(t)$ do
        $Z \leftarrow \{\emptyset\}$                      ▷ initialize novel feature subset $Z$
        for $i = 1$ to $|S_c|$ do                ▷ feature subset selected by $c$, $S_c$
            find the best feature $f^+ = \arg\max_{f^+ \in \{F \backslash \{S_c \cup Z\}\}} Q(f^+, L)$
            add $f^+$ to $Z$
        end for
        add $Z$ to $E(t)$ as a chromosome
    end for
end procedure
```

Algorithm 2: Procedures of exploration operator.

operator identifies unselected promising features from the perspective of each chromosome in $G(t)$ and encodes them into a new chromosome in $E(t)$ (line (7)). For balance between the genetic and exploration operations, we set the size of $E(t)$ to the same value as that of $G(t)$ because $E(t)$ must be evaluated in order to determine its fitness. These two sets of chromosomes are then combined to form the offspring set $N(t)$ of the tth population (line (8)). To evaluate the fitness of the offspring set, the proposed method uses a certain number of FFCs (line (9)). Specifically, the proposed method uses $2 \cdot |G(t)|$ FFCs in one generation. Next, $N(t)$ is added to $P(t)$ and m chromosomes with higher fitness values are selected (line (11)). This procedure is repeated until the algorithm uses all of its allowed FFCs. This limit is denoted v and is chosen by the user. The output of Algorithm 1 is the best feature subset obtained during evolution.

3.3. Exploration Operator. Because a feature subset selects a small number of features within n and most features will remain unselected, the exploration operator is needed in order to explore a large set of unselected features. Algorithm 2 outlines the pseudocode for the proposed exploration operator. For each offspring generated by the genetic operators, we iteratively select relevant features that maximize the objective function and that were not selected by the offspring c until the subset size becomes $|S_c|$, where $|S_c|$ is the subset size of c. Thus, proposed exploration operation does not incur additional parameter for determining the number of features to be selected.

To implement our exploration operation, we employ an effective filter method called the scalable criterion for large label sets (SCLS) [16] as an objective function $Q(f^+, L)$, where L is the label set. The selection of the ith feature from the set $\{F \backslash \{S_c \cup Z\}\}$, where Z is a feature subset with $i-1$ features when selecting ith feature, is performed by identifying f_i that maximizes the value of the following relevance evaluation [17]:

$$\max_{f_i \in \{F \backslash \{S_c \cup Z\}\}} \left[D(f_i) - R(f_i) \right], \tag{1}$$

where $D(f_i)$ and $R(f_i)$ denote the dependency of f_i on L and the dependency of f_i on the selected features of Z, respectively. From [17], (1) can be reformulated as follows:

$$\max_{f_i \in \{F \backslash \{S_c \cup Z\}\}} \left[\sum_{l \in L} M(f_i; l) - \sum_{f \in S_{i-1}} M(f_i; f) \right], \tag{2}$$

where $M(x; y) = H(x) - H(x, y) + H(y)$ is the mutual information between variables x and y and $H(x) = -\sum P(x) \log P(x)$ is the joint entropy of the probability functions $P(x)$, $P(y)$, and $P(x, y)$. Following from (2), $D(f_2)$ can be calculated as follows:

$$D(f_2) = \sum_{l \in L} M(f_2; l). \tag{3}$$

As (2), $R(f_2)$ can be calculated as

$$R(f_2) = \sum_{l \in L} M(f_2; l). \tag{4}$$

In order to calculate $R(f_2)$ while considering adaptability against the scaling of $D(f_2)$ and avoiding repetitive calculations by $f \in S$ and $l \in L$, let $\mathrm{Red}(f_2)$ be represented as follows:

$$R(f_2) = \alpha \cdot D(f_2) = \alpha \sum_{l \in L} M(f_2; l), \tag{5}$$

where $0 \leq \alpha \leq 1$, which must be estimated, determines the reduction with relevance to f_2 based on $D(f_2)$, while circumventing the repetitive calculations for reduction against each label. According to [16], α can be approximated as follows:

$$\alpha \approx \frac{M(f_2; f_1)}{H(f_2)}. \tag{6}$$

As a result, the relevance evaluation for f_2 is performed as follows:

$$J = \sum_{l \in L} M(f_2; l) - \frac{M(f_2; f_1)}{H(f_2)} \sum_{l \in L} M(f_2; l). \tag{7}$$

Equation (7) represents how the relevance evaluation can be performed when $i = 2$. By considering the previously

Table 1: Standard characteristics of employed datasets.

Dataset	$\|W\|$	$\|F\|$	Type	$\|L\|$	Card.	Den.	Distinct.	Domain
Birds	645	260	Mixed	19	1.014	0.053	133	Audio
Emotions	593	72	Numeric	6	1.869	0.311	27	Music
Enron	1,702	1,001	Nominal	53	3.378	0.064	753	Text
Genbase	662	1,185	Nominal	27	1.252	0.046	32	Biology
LLog	1,460	1,004	Nominal	75	1.180	0.016	304	Text
Mediamill	43,907	120	Numeric	45	1.245	0.028	94	Video
Medical	978	1,494	Nominal	45	1.245	0.028	94	Text
Scene	2,407	294	Numeric	6	1.074	0.179	15	Images
Slashdot	3,782	1,079	Nominal	22	1.181	0.054	156	Text
TMC2007	28,596	981	Numeric	22	2.158	0.098	1,341	Text
Yeast	2,417	103	Numeric	14	4.237	0.303	198	Biology
Arts	7,484	1,157	Numeric	26	1.654	0.064	599	Text
Business	11,214	1,096	Numeric	30	1.599	0.053	233	Text
Computers	12,444	1,705	Numeric	33	1.507	0.046	428	Text
Education	12,030	1,377	Numeric	33	1.463	0.044	511	Text
Entertain	12,730	1,600	Numeric	21	1.414	0.067	337	Text
Health	9,205	1,530	Numeric	32	1.644	0.051	335	Text
Reference	8,027	1,984	Numeric	33	1.174	0.036	275	Text
Science	6,428	1,859	Numeric	40	1.450	0.036	457	Text
Social	12,111	2,618	Numeric	39	1.279	0.033	361	Text
Society	14,512	1,590	Numeric	27	1.670	0.062	1,054	Text

selected features in Z, the final relevance evaluation can be represented as follows:

$$\max_{f_i \in \{F \setminus \{S_c \cup Z\}\}} \left[\sum_{l \in L} M(f_i; l) - \sum_{f \in S_{i-1}} \sum_{l \in L} \frac{M(f_i; l)}{H(f_i)} M(f_i; f) \right]. \quad (8)$$

Equation (8) is the objective function for selecting relevant features from the unselected feature subset used by our exploration operation.

3.4. Experimental Settings. We experimented on 20 different datasets from various domains. The Birds dataset is audio data containing examples of multiple bird calls. The Emotions dataset is music data classified into six emotional clusters. The Enron, Language Log (LLog), and Slashdot datasets were generated from text mining applications, where each feature corresponds to the occurrence of a word and each label represents the relevancy of each text pattern to a specific subject. The Genbase and Yeast datasets come from the biological domain and include information about the functions of genes and proteins. The Mediamill dataset is video data from an automatic detection system. The Medical dataset was sampled from a large corpus of suicide letters obtained from the natural language processing of clinical free text. The Scene dataset is related to the semantic indexing of still scenes, where each scene may contain multiple objects. The TMC2007 dataset contains safety reports of complex space system. The remaining nine datasets come from the Yahoo dataset collection. We performed unsupervised dimensionality reduction on text datasets, including the TMC2007 and Yahoo collections, which were composed of more than 10,000 features. Specifically, the top 2% and 5% of features with the highest document frequency were retained for TMC2007 and the Yahoo datasets, respectively [45]. In the text mining domain, existing studies report that classification performance will not suffer significantly from the retention of 1% of features based on document frequency [46].

Table 1 contains the standard statistics for the multilabel datasets employed in our experiments, including the number of patterns in the dataset $|W|$, number of features $|F|$, type of features, and number of labels $|L|$. When the feature type is numeric, we discretize the features by using the supervised discretization method [47] for multilabel naïve Bayes classifier (MLNB) [12]. Specifically, each observed numeric value is assigned to one of several bins that are automatically determined by using the discretization method. The label cardinality *Card* represents the average number of labels for each instance. The label density *Den* is the label cardinality over the total number of labels. The number of distinct label sets *Distinct* indicates the number of unique label subsets in L. *Domain* represents the application that each dataset was extracted from.

We measured the mean size of the selected feature subsets for both the proposed method and the conventional multilabel feature selection methods (GA with restrictive genetic operators [40] (RGA), NSGA-II [43], and MPSOFS [43]) to determine which methods achieved to select less than 10 features. Specifically, we provide detailed parameter setting to support good reproducibility as follows:

(i) RGA creates $m = 20$ initial solutions by selecting less than $n = 10$ features randomly in accordance with each chromosome. Each solution in the initial population $P(t)$, where $t = 0$, is evaluated using an employed multilabel classifier. Next, the RGA creates an offspring set $N(t)$ by using genetic operators. To apply the crossover operator, two solutions in $P(t)$ are randomly selected and mated; thereafter, one solution in $P(t)$ is randomly selected and mutated. In this study, we employed restrictive crossover and restrictive mutation operators with both crossover rate and mutation rate set to 1.0. Therefore, for each iteration, the GA creates three new solutions to compose $N(t)$. Each newly created solution is evaluated using the multilabel classifier. To create $P(t+1)$, $N(t)$ is added to $P(t)$, and 20 solutions with higher fitness values are selected. This procedure is repeated until the RGA spends 100 FFCs.

(ii) NSGA-II creates $m = 20$ initial solutions randomly, the same number RGA creates. The maximum number of allowed feature is set to $|F|$ because the NSGA-II naturally minimizes the number of selected features. Each solution in $P(t)$ is evaluated using an employed multilabel classifier and the number of features. The NSGA-II then creates $N(t)$ where $|N(t)| = 3$ which is the same setting of RGA. To create $P(t+1)$, $N(t)$ is added to $P(t)$, and the superiority of each solution is determined by the nondominated sort method. After the superiority among solutions in $\{P(t) \cup N(t)\}$ is determined, the top 20 solutions are selected to form $P(t+1)$. This procedure is repeated until the NSGA-II spends 100 FFCs.

(iii) MPSOFS creates 20 initial solutions randomly, the same number RGA creates. Each solution in $P(t)$ is evaluated using an employed multilabel classifier and the number of features and ranked using the nondominated sort method. The MPSOFS then preserves the best solution of $P(t)$ called the global best solution. In addition, the best solution which each chromosome experienced is also preserved; this is called the individual best solution, and therefore there are 20 individual best solutions. Thereafter, the MPSOFS updates the representation of each chromosome based on the global best solution and its own individual best solution using a velocity with inertia weight of 0.7298 and two acceleration coefficients of 1.4962 suggested from the study of [48]. After all chromosomes in $P(t)$ are modified, they are evaluated and regarded as $P(t+1)$. This procedure is repeated until the MPSOFS spends 100 FFCs.

Although different parameter setting may result in better performance, we fixed the size of the population m to 20 and the number of spent FFCs v to 100 for all the methods to ensure a fair comparison. To evaluate the quality of the feature subsets obtained by each method, we used MLNB classifier because it outputs a predicted label subset based on the intrinsic characteristics of a given dataset without requiring any complicated parameter-tuning process that might influence the final multilabel classification performance [39]. For the sake of fairness, we used the hold-out cross-validation method for each experiment [11, 49]. 80% of the samples in a given dataset were randomly chosen as the training set for multilabel feature selection and classifier training, while the remaining 20% of the samples were used as the test set to obtain the multilabel classification performance. For both the RGA and the proposed method, we set the population size to 20 and the maximum number of allowed FFCs to 100. Each experiment was repeated 10 times and the average value was used to represent the classification performance of each feature selection method.

We employed four evaluation metrics: Hamming loss, multilabel accuracy, ranking loss, and normalized coverage. Let $T = \{(T_i, \lambda_i) \mid 1 \leq i \leq |T|\}$ be a given test set where $\lambda_i \subseteq L$ is a correct label subset. For a given test sample T_i, a classifier, such as MLNB, should output a set of confidence values $0 \leq \psi_{i,l} \leq 1$ for each label $l \in L$. If a confidence value $\psi_{i,l}$ is larger than a predefined threshold value, such as 0.5, the corresponding label l will be included in the predicted label subset Y_i. Based on the ground truth λ_i, confidence values $\psi_{i,l}$, and predicted label subset Y_i, multilabel classification performance can be measured with each evaluation metric [33, 45, 50].

Multilabel accuracy is defined as follows:

$$\operatorname{mlacc}(T) = \frac{1}{|T|}\sum_{i=1}^{|T|}\frac{|\lambda_i \cap Y_i|}{|\lambda_i \cup Y_i|}. \quad (9)$$

Hamming loss is defined as follows:

$$\operatorname{hloss}(T) = \frac{1}{|T|}\sum_{i=1}^{|T|}\frac{1}{|L|}|\lambda_i \triangle Y_i|, \quad (10)$$

where $\triangle$ denotes the symmetric difference between two sets. Ranking loss is defined as follows:

$$\operatorname{rloss}(T) = \frac{1}{|T|}\sum_{i=1}^{|T|}\frac{\left|\left\{(a,b) \mid a \in \lambda_i, b \in \overline{\lambda_i}, \psi_{i,a} \leq \psi_{i,b}\right\}\right|}{|\lambda_i|\left|\overline{\lambda_i}\right|}, \quad (11)$$

where $\overline{\lambda_i}$ is a complementary set of λ_i. Therefore, ranking loss measures the average fraction of (a,b) pairs with $\psi_{i,a} \leq \psi_{i,b}$ over all possible relevant and irrelevant label pairs. Finally, normalized coverage is defined as follows:

$$\operatorname{ncov}(T) = \frac{1}{|L|}\left(\frac{1}{|T|}\sum_{i=1}^{|T|}\max_{l \in \lambda_i}\operatorname{rank}(l) - 1\right), \quad (12)$$

where rank$(\cdot)$ returns the rank of the corresponding relevant label $l \in \lambda_i$ according to $\psi_{i,l}$ in nonincreasing order. Therefore, normalized coverage measures how many labels must be marked as positive for all relevant labels to be positive. Higher values of multilabel accuracy and lower values of Hamming loss, ranking loss, and normalized coverage indicate good classification performance.

Additionally, because we are interested in the superiority of the proposed method over conventional multilabel feature selection methods, we perform the Wilcoxon signed-rank test [51] to validate the performance of the proposed method. Let d_i be the difference between the performance of the two methods for the ith dataset. The differences are ranked

Table 2: Comparison results for multilabel feature selection methods in terms of selected feature subset size (mean ± std. deviation). The ✘ symbol is used to indicate that the corresponding method failed to select less than 10 features for the dataset.

Dataset	Proposed	RGA	NSGA-II	MPSOFS
Birds	7 ± 2	5 ± 2	99 ± 51✘	138 ± 4✘
Emotions	8 ± 1	9 ± 1	50 ± 6✘	37 ± 4✘
Enron	8 ± 1	9 ± 1	74 ± 51✘	527 ± 23✘
Genbase	9 ± 0	7 ± 1	974 ± 139✘	637 ± 24✘
LLog	8 ± 1	7 ± 2	205 ± 108✘	522 ± 17✘
Mediamill	5 ± 1	4 ± 0	7 ± 2	52 ± 3✘
Medical	8 ± 1	8 ± 1	664 ± 138✘	762 ± 30✘
Scene	9 ± 1	9 ± 0	137 ± 31✘	147 ± 4✘
Slashdot	9 ± 0	8 ± 2	970 ± 85✘	569 ± 22✘
TMC2007	8 ± 1	8 ± 1	495 ± 93✘	506 ± 20✘
Yeast	9 ± 1	9 ± 1	34 ± 10✘	52 ± 3✘
Arts	9 ± 1	7 ± 1	1,019 ± 85✘	613 ± 25✘
Business	4 ± 2	4 ± 2	130 ± 140✘	578 ± 20✘
Education	8 ± 1	8 ± 2	1263 ± 51✘	742 ± 24✘
Entertainment	9 ± 0	8 ± 2	985 ± 227✘	840 ± 28✘
Health	7 ± 1	4 ± 1	842 ± 184✘	814 ± 33✘
Reference	7 ± 2	8 ± 1	1052 ± 334✘	1058 ± 42✘
Science	9 ± 0	7 ± 1	969 ± 265✘	993 ± 65✘
Social	6 ± 1	8 ± 1	1,699 ± 401✘	1,353 ± 71✘
Society	7 ± 2	4 ± 3	498 ± 95✘	826 ± 28✘

based on their absolute values and the smallest d_i is assigned to the first rank. If ties occur, average ranks are assigned. Let R^+ be the sum of the ranks for the datasets on which the compared method outperforms the proposed method, defined as follows:

$$R^+ = \sum_{d_i>0} \operatorname{rank}(d_i) + \frac{1}{2}\sum_{d_i=0} \operatorname{rank}(d_i). \qquad (13)$$

Let R^- be the sum of the ranks for the datasets on which the proposed method outperforms the compared method. Then, based on the critical values from the Wilcoxon test, for a confidence level of $\alpha = 0.05$ and $N = 20$, the difference between the compared methods is significant if $\min(R^+, R^-)$ is less than or equal to 8. In this case, the null hypothesis of equal performance is rejected.

4. Experimental Results

4.1. Comparison Results. Table 2 contains the results for the mean size and standard deviation of the selected feature subsets of the proposed method and conventional multilabel feature selection methods when the evaluation metric is multilabel accuracy. The ✘ symbol indicates methods that failed to satisfy given constraint for the corresponding dataset. The proposed method and RGA both selected less than 10 features for all datasets. The NSGA-II and MPSOFS methods failed to select less than 10 features for all datasets other than the Mediamill dataset for NSGA-II, despite having objective functions to minimize feature subset sizes. Because the NSGA-II and MPSOFS failed to select less than 10 features for most datasets, we compared the performance of the proposed method with the performance of the RGA from subsequent experiments. It should be noted that n can be set to a larger value than 10, such as 30 or 50. The experimental results in Table 2 show that the NSGA-II or MPSOFS will fail to satisfy the given constraints because they output the final feature subset, which is composed of tens or hundreds of features for most experiments.

Tables 3 and 4 contain the experimental results for the proposed method and RGA on 20 multilabel datasets, presented as the average performances for hold-out cross-validation with corresponding standard deviations. Table 3 contains the performance results for multilabel accuracy and Hamming loss, and Table 4 contains the performance results for ranking loss and normalized coverage. The best performance between the two methods is indicated by bold font and a √ symbol. Finally, Table 5 contains the results of the Wilcoxon signed-rank test for the proposed method against RGA for Genbase dataset with a significance threshold of $\alpha = 0.05$. For each evaluation metric, the winner of each comparison is indicated with bold font and the corresponding sum of the outperformed rank R^+ over the total rank and p values are presented in the parenthesis. We observed a similar tendency from the same experiments on the other multilabel datasets.

As shown in Tables 3 and 4, the proposed method outperformed RGA for most multilabel datasets. Specifically, the proposed method achieved the best performance for 90% of the datasets in terms of multilabel accuracy, 95% of the datasets in terms of Hamming loss, 95% of the datasets in terms of ranking loss, and 100% of the datasets in terms of normalized coverage. Thus, the proposed method

Table 3: Comparison results for multilabel feature selection methods in terms of multilabel accuracy and Hamming loss (mean ± std. deviation). The √ symbol indicates the method that achieves the best performance for each dataset.

	Evaluation measure			
Methods	Multi-label accuracy		Hamming loss	
	Proposed	RGA	Proposed	RGA
Birds	**0.497 ± 0.048**√	0.459 ± 0.048	**0.055 ± 0.005**√	0.056 ± 0.004
Emotions	**0.460 ± 0.020**√	0.447 ± 0.029	**0.243 ± 0.022**√	0.252 ± 0.016
Enron	**0.360 ± 0.021**√	0.271 ± 0.042	**0.056 ± 0.002**√	0.060 ± 0.001
Genbase	**0.886 ± 0.041**√	0.155 ± 0.097	**0.011 ± 0.004**√	0.042 ± 0.003
LLog	**0.213 ± 0.027**√	0.166 ± 0.026	0.016 ± 0.001	**0.016 ± 0.001**√
Mediamill	**0.366 ± 0.002**√	0.359 ± 0.005	0.034 ± 0.000	**0.034 ± 0.000**√
Medical	**0.517 ± 0.048**√	0.097 ± 0.046	**0.018 ± 0.002**√	0.026 ± 0.002
Scene	**0.408 ± 0.019**√	0.352 ± 0.030	0.157 ± 0.002	**0.154 ± 0.005**√
Slashdot	**0.144 ± 0.017**√	0.031 ± 0.011	**0.048 ± 0.001**√	0.053 ± 0.000
TMC2007	**0.372 ± 0.005**√	0.318 ± 0.020	**0.084 ± 0.001**√	0.088 ± 0.001
Yeast	**0.465 ± 0.013**√	0.442 ± 0.019	**0.224 ± 0.006**√	0.225 ± 0.010
Arts	**0.140 ± 0.012**√	0.049 ± 0.015	**0.060 ± 0.001**√	0.063 ± 0.001
Business	0.678 ± 0.011	**0.678 ± 0.008**√	0.029 ± 0.001	**0.029 ± 0.001**√
Education	**0.109 ± 0.019**√	0.033 ± 0.011	**0.042 ± 0.001**√	0.044 ± 0.001
Entertain	**0.233 ± 0.016**√	0.128 ± 0.042	**0.058 ± 0.000**√	0.065 ± 0.002
Health	**0.510 ± 0.018**√	0.402 ± 0.016	**0.040 ± 0.001**√	0.049 ± 0.001
Reference	0.382 ± 0.044	**0.393 ± 0.011**√	**0.030 ± 0.001**√	0.034 ± 0.001
Science	**0.120 ± 0.011**√	0.042 ± 0.015	**0.034 ± 0.001**√	0.036 ± 0.001
Social	**0.546 ± 0.018**√	0.134 ± 0.060	**0.024 ± 0.001**√	0.030 ± 0.001
Society	**0.304 ± 0.135**√	0.280 ± 0.146	**0.055 ± 0.001**√	0.059 ± 0.001

Table 4: Comparison results for multilabel feature selection methods in terms of ranking loss and normalized coverage (mean ± std. deviation). The √ symbol indicates the method that achieves the best performance for each dataset.

	Evaluation measure			
Methods	Ranking loss		Normalized coverage	
	Proposed	RGA	Proposed	RGA
Birds	**0.143 ± 0.015**√	0.166 ± 0.019	**0.227 ± 0.019**√	0.248 ± 0.028
Emotions	0.218 ± 0.025	**0.217 ± 0.029**√	**0.499 ± 0.028**√	0.524 ± 0.030
Enron	**0.098 ± 0.008**√	0.115 ± 0.008	**0.277 ± 0.001**√	0.296 ± 0.010
Genbase	**0.035 ± 0.026**√	0.152 ± 0.037	**0.084 ± 0.026**√	0.212 ± 0.029
LLog	**0.170 ± 0.019**√	0.179 ± 0.021	**0.215 ± 0.024**√	0.223 ± 0.022
Mediamill	**0.057 ± 0.001**√	0.058 ± 0.001	**0.194 ± 0.003**√	0.197 ± 0.002
Medical	**0.093 ± 0.026**√	0.173 ± 0.260	**0.132 ± 0.027**√	0.199 ± 0.023
Scene	**0.159 ± 0.012**√	0.188 ± 0.015	**0.311 ± 0.007**√	0.326 ± 0.012
Slashdot	**0.247 ± 0.004**√	0.297 ± 0.010	**0.301 ± 0.004**√	0.353 ± 0.010
TMC2007	**0.113 ± 0.004**√	0.154 ± 0.006	**0.254 ± 0.004**√	0.316 ± 0.008
Yeast	**0.199 ± 0.007**√	0.200 ± 0.008	**0.550 ± 0.010**√	0.553 ± 0.013
Arts	**0.161 ± 0.017**√	0.180 ± 0.019	**0.260 ± 0.016**√	0.275 ± 0.019
Business	**0.059 ± 0.025**√	0.062 ± 0.025	**0.129 ± 0.024**√	0.132 ± 0.023
Education	**0.095 ± 0.004**√	0.109 ± 0.003	**0.152 ± 0.004**√	0.168 ± 0.004
Entertain	**0.130 ± 0.005**√	0.137 ± 0.005	**0.215 ± 0.005**√	0.222 ± 0.009
Health	**0.089 ± 0.028**√	0.107 ± 0.027	**0.161 ± 0.025**√	0.179 ± 0.026
Reference	**0.110 ± 0.022**√	0.119 ± 0.023	**0.155 ± 0.023**√	0.164 ± 0.022
Science	**0.138 ± 0.005**√	0.152 ± 0.003	**0.201 ± 0.007**√	0.213 ± 0.005
Social	**0.073 ± 0.010**√	0.107 ± 0.027	**0.124 ± 0.010**√	0.132 ± 0.011
Society	0.143 ± 0.005	**0.137 ± 0.005**√	**0.249 ± 0.006**√	0.261 ± 0.005

Table 5: Wilcoxon signed-rank test results for the proposed method against RGA for Genbase dataset with a significance threshold of $\alpha = 0.05$, sum of outperformed rank R^+ over the total rank and p values.

Evaluation measures	Proposed versus RGA		
	Result	Stats	p value
Hamming loss	*Win*	55/55	$2.0e-3$
Multilabel accuracy	*Win*	55/55	$2.0e-3$
Ranking loss	*Win*	55/55	$2.0e-3$
Normalized coverage	*Win*	55/55	$2.0e-3$

significantly outperforms RGA for all evaluation metrics. This is evident from the experimental results shown in Table 5, which clearly demonstrate that the proposed method is statistically superior to RGA.

4.2. Analysis. Figure 2 shows the convergence behaviors of the GA and proposed method according to the number of spent FFCs (u) in terms of the multilabel accuracy; the horizontal axis represents u, and the vertical axis indicates the multilabel accuracy performance. Because the convergence behaviors may differ according to each experiment owing to the stochastic nature of the population-based search methods, we set the same initialized population in both algorithms and averaged the multilabel accuracy performance of the top elitist in the population after conducting the experiment 10 times. Figure 2 shows that the multilabel accuracy performance monotonically improves with u. Because the initialization steps consume 20 FFCs and the two methods have the same initialized population that is randomly created, both methods gradually improve the multilabel accuracy initially. However, the experimental results indicate that the multilabel accuracy value of the proposed method is dramatically improved when $u \geq 20$ because the exploration operator is applied to the population after the initialization. Thus, Figure 2 indicates that the proposed method can efficiently locate a good feature subset from unselected features.

The goal of our exploration operation introduces novel promising features that would effectively improve the multilabel classification performance. To validate the effectiveness of our exploration operation, we conduct an additional experiment by comparing the fitness values of the offspring set created by the proposed exploration operation and the random operation, respectively. Specifically, 50 chromosomes, namely, G, that select 10 or lesser number of features as the same initialization procedure of RGA were used and 50 new chromosomes are then created by applying the proposed exploration operation to each chromosome in G to form the first offspring set. Thereafter, for the sake of comparison, novel features with regard to each chromosome in G are selected randomly and introduced to create the second offspring set. Finally, the fitness values of the first and second offspring sets in terms of the four performance measures are measured. Figure 3 shows the box plots of fitness values given by the two offspring sets of the Genbase dataset. The experimental results indicates that the fitness values of the first offspring set (Proposed) is much better than that of the second offspring set (Random) from the viewpoint of all measures, indicating that the proposed exploration operation has a much better search capability than the random search.

5. Conclusion

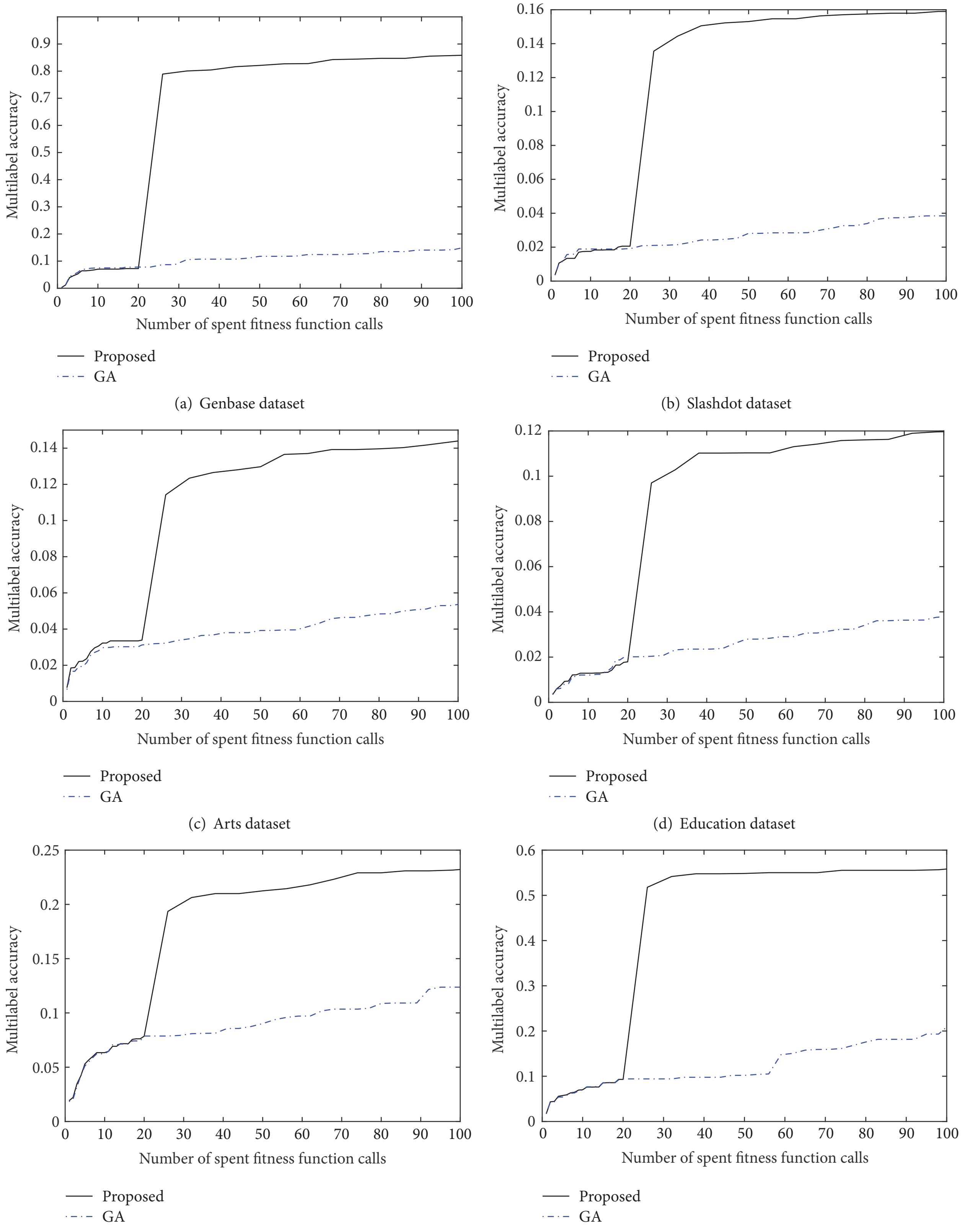
We proposed an effective evolutionary search-based feature selection method with a budget constraint for multilabel classification. As a feature subset selects a small number of features within the maximum allowed number of features and most features are unselected in the budget constraint problem, we employ a novel exploration operation to find relevant features in the large unselected feature subset. Our experiments on 20 real-world datasets demonstrated that proposed exploration operator successfully enhances the search capability of genetic search, resulting in an improvement in multilabel classification. The results also showed that the proposed method can search a feature subset successfully, which does not violate the budget constraint. Statistical tests showed that our method outperformed conventional methods in four performance measures. Although the proposed exploration operation improves the effectiveness of evolutionary search without incurring additional parameters, it cannot be applied directly to certain types of evolutionary search algorithms, such as particle swarm optimization, which do not depend on offspring sets. Thus, an additional consideration should be made to design a new exploration operation for such cases.

A future research direction will be a study on an evolutionary algorithm. The proposed method is a genetic algorithm based feature selection; however, it can be applied to other evolutionary algorithms such as the Estimation of Distribution Algorithm. We would like to study this issue further.

Terms Used in This Study and Meanings

Constants

t: Number of generations
m: The size of the population, $|P(t)| = m$
n: Maximum number of allowed features selected by S_c
c: A chromosome in $P(t)$
S_c: A selected feature subset represented by c
v: Maximum number of allowed fitness function calls (FFCs)
u: Number of spent FFCs, $u = m + 2 \cdot |G(t)| \cdot t$.



(a) Genbase dataset

(b) Slashdot dataset

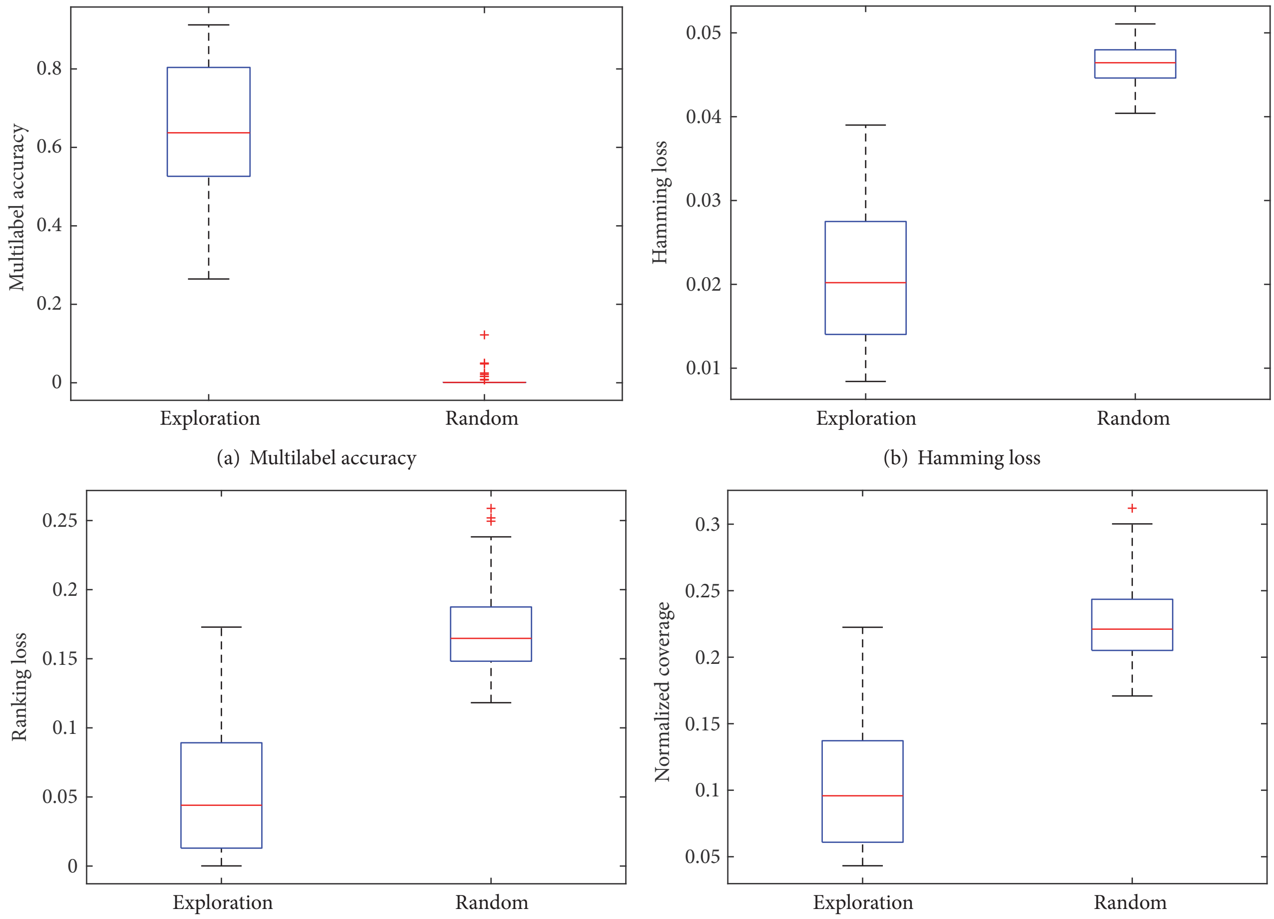
(c) Arts dataset

(d) Education dataset

(e) Entertainment dataset

(f) Social dataset

FIGURE 2: Comparison results of the convergence between RGA and the proposed method in terms of multilabel accuracy (a higher value indicates a good classification performance).



(a) Multilabel accuracy

(b) Hamming loss

(c) Ranking loss

(d) Normalized coverage

Figure 3: Comparison results showing the effectiveness of the proposed exploration operator and random search in terms of the four performance measures on the Genbase dataset.

Sets

$P(t)$: The population at the tth generation
$G(t)$: A set of newly created solutions from genetic operator
$E(t)$: A set of newly created solutions from exploration operator
$N(t)$: A set of newly created solutions from $P(t)$, $G(t) \cup E(t)$.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by the Chung-Ang University Research Grants in 2017 and by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (NRF-2016R1C1B1014774).

References

[1] A. Iliopoulos, R. Shirzadeh, W. Weijtjens, P. Guillaume, D. V. Hemelrijck, and C. Devriendt, "A modal decomposition and expansion approach for prediction of dynamic responses on a monopile offshore wind turbine using a limited number of vibration sensors," *Mechanical Systems and Signal Processing*, vol. 68-69, pp. 84–104, 2016.

[2] S. Diplaris, G. Tsoumakas, P. A. Mitkas, and I. Vlahavas, "Protein classification with multiple algorithms," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface*, vol. 3746, pp. 448–456, 2005.

[3] Y. Rao, "Contextual Sentiment Topic Model for Adaptive Social Emotion Classification," *IEEE Intelligent Systems*, vol. 31, no. 1, pp. 41–47, 2016.

[4] R. Agrawal, A. Gupta, Y. Prabhu, and M. Varma, "Multi-label learning with millions of labels," in *Proceedings of the the 22nd international conference*, pp. 13–24, Rio de Janeiro, Brazil, May 2013.

[5] P. Duygulu, K. Barnard, J. F. de Freitas, and D. A. Forsyth, "Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary," in *Computer Vision — ECCV*

2002, vol. 2353 of *Lecture Notes in Computer Science*, pp. 97–112, Springer Berlin Heidelberg, Berlin, Heidelberg, 2002.

[6] H. Ghasemzadeh, N. Amini, R. Saeedi, and M. Sarrafzadeh, "Power-aware computing in wearable sensor networks: An optimal feature selection," *IEEE Transactions on Mobile Computing*, vol. 14, no. 4, pp. 800–812, 2015.

[7] B. Nushi, A. Singla, A. Krause, and D. Kossmann, "Learning and feature selection under budget constraints in crowdsourcing," in *Proceedings of the in 4th AAAI Conf. Human Computation and Crowdsourcing*, pp. 159–168, Austin, USA, October 2016.

[8] H. Yang, R. Fujimaki, Y. Kusumura, and J. Liu, "Online feature selection: A limited-memory substitution algorithm and its asynchronous parallel variation," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2016*, pp. 1945–1954, USA, August 2016.

[9] H. Blume, B. Bischl, M. Botteck et al., "Huge music archives on mobile devices," *IEEE Signal Processing Magazine*, vol. 28, no. 4, pp. 24–39, 2011.

[10] P. Naula, A. Airola, T. Salakoski, and T. Pahikkala, "Multi-label learning under feature extraction budgets," *Pattern Recognition Letters*, vol. 40, no. 1, pp. 56–65, 2014.

[11] J. Lee and D. W. Kim, "Memetic feature selection algorithm for multi-label classification," *Information Sciences*, vol. 293, pp. 80–96, 2015.

[12] M.-L. Zhang, J. M. Peña, and V. Robles, "Feature selection for multi-label naive Bayes classification," *Information Sciences*, vol. 179, no. 19, pp. 3218–3229, 2009.

[13] H. Yang, Z. Xu, M. R. Lyu, and I. King, "Budget constrained non-monotonic feature selection," *Neural Networks*, vol. 71, pp. 214–224, 2015.

[14] F. Min and J. Xu, "Semi-greedy heuristics for feature selection with test cost constraints," *Granular Computing*, vol. 1, no. 3, pp. 199–211, 2016.

[15] J. Wang, P. Zhao, S. C. H. Hoi, and R. Jin, "Online feature selection and its applications," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 3, pp. 698–710, 2014.

[16] J. Lee and D.-W. Kim, "SCLS: Multi-label feature selection based on scalable criterion for large label set," *Pattern Recognition*, vol. 66, pp. 342–352, 2017.

[17] J. Lee, H. Lim, and D.-W. Kim, "Approximating mutual information for multi-label feature selection," *IEEE Electronics Letters*, vol. 48, no. 15, pp. 929-930, 2012.

[18] H. Lim, J. Lee, and D.-W. Kim, "Multi-label learning using mathematical programming," *IEICE Transaction on Information and Systems*, vol. E98D, no. 1, pp. 197–200, 2015.

[19] Y. Lin, Q. Hu, J. Liu, and J. Duan, "Multi-label feature selection based on max-dependency and min-redundancy," *Neurocomputing*, vol. 168, pp. 92–103, 2015.

[20] G. Doquire and M. Verleysen, "Mutual information-based feature selection for multilabel classification," *Neurocomputing*, vol. 122, pp. 148–155, 2013.

[21] D. Kong, C. Ding, H. Huang, and H. Zhao, "Multi-label ReliefF and F-statistic feature selections for image annotation," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2012*, pp. 2352–2359, usa, June 2012.

[22] J. Lee and D.-W. Kim, "Fast multi-label feature selection based on information-theoretic feature ranking," *Pattern Recognition*, vol. 48, no. 9, pp. 2761–2771, 2015.

[23] X. Kong and P. S. Yu, "GMLC: A multi-label feature selection framework for graph classification," *Knowledge and Information Systems*, vol. 31, no. 2, pp. 281–305, 2012.

[24] J. Lee and D.-W. Kim, "Feature selection for multi-label classification using multivariate mutual information," *Pattern Recognition Letters*, vol. 34, no. 3, pp. 349–357, 2013.

[25] N. Spolaôr, E. A. Cherman, M. C. Monard, and H. D. Lee, "A comparison of multi-label feature selection methods using the problem transformation approach," *Electronic Notes in Theoretical Computer Science*, vol. 292, pp. 135–151, 2013.

[26] Q. Gu, Z. Li, and J. Han, "Correlated multi-label feature selection," in *Proceedings of the the 20th ACM international conference*, p. 1087, Glasgow, Scotland, UK, October 2011.

[27] B. Qian and I. Davidson, "Semi-supervised dimension reduction for multi-label classification," in *Proceedings of the Proc. 24th AAAI Conf. Artificial Intelligence*, pp. 569–574, Atlanta, USA, Jul 2010.

[28] B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A Survey on Evolutionary Computation Approaches to Feature Selection," *IEEE Transactions on Evolutionary Computation*, vol. 20, no. 4, pp. 606–626, 2016.

[29] Y. Zhang, D. Gong, Y. Hu, and W. Zhang, "Feature selection algorithm based on bare bones particle swarm optimization," *Neurocomputing*, vol. 148, pp. 150–157, 2015.

[30] Y. Zhang, D.-W. Gong, and J. Cheng, "Multi-objective particle swarm optimization approach for cost-based feature selection in classification," *IEEE Transactions on Computational Biology and Bioinformatics*, vol. 14, no. 1, pp. 64–75, 2017.

[31] Y. Zhang, X.-F. Song, and D.-W. Gong, "A return-cost-based binary firefly algorithm for feature selection," *Information Sciences*, vol. 418-419, pp. 561–574, 2017.

[32] J. Read, B. Pfahringer, and G. Holmes, "Multi-label Classification Using Ensembles of Pruned Sets," in *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining (ICDM)*, pp. 995–1000, Pisa, Italy, December 2008.

[33] N. Spolaôr, M. C. Monard, G. Tsoumakas, and H. D. Lee, "A systematic review of multi-label feature selection and a new method based on label construction," *Neurocomputing*, vol. 180, pp. 3–15, 2016.

[34] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of imbalanced data: a review," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 4, pp. 687–719, 2009.

[35] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Random k-labelsets for multilabel classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 7, pp. 1079–1089, 2011.

[36] S. Ji and J. Ye, "Linear dimensionality reduction for multi-label classification," in *Proceedings of the Proc. 21th Int. Joint Conf. Artificial Intelligence*, pp. 1077–1082, Pasadena, USA, Jul 2009.

[37] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint l2,1-norms minimization," in *Advances in Neural Information Processing System*, pp. 1813–1821, MIT Press, 2010.

[38] J. Lee and D.-W. Kim, "Mutual Information-based multi-label feature selection using interaction information," *Expert Systems with Applications*, vol. 42, no. 4, pp. 2013–2025, 2015.

[39] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, 2014.

[40] Z. Zhu, S. Jia, and Z. Ji, "Towards a memetic feature selection paradigm," *IEEE Computational Intelligence Magazine*, vol. 5, no. 2, pp. 41–53, 2010.

[41] 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.

[42] J. Yin, T. Tao, and J. Xu, "A Multi-label feature selection algorithm based on multi-objective optimization," in *Proceedings of the International Joint Conference on Neural Networks, IJCNN 2015*, Ireland, July 2015.

[43] Y. Zhang, D.-W. Gong, X.-Y. Sun, and Y.-N. Guo, "A PSO-based multi-objective multi-label feature selection method in classification," *Scientific Reports*, vol. 7, no. 1, article no. 376, 2017.

[44] G. Karafotias, M. Hoogendoorn, and A. E. Eiben, "Parameter Control in Evolutionary Algorithms: Trends and Challenges," *IEEE Transactions on Evolutionary Computation*, vol. 19, no. 2, pp. 167–187, 2015.

[45] M.-L. Zhang and L. Wu, "LIFT: multi-label learning with label-specific features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 1, pp. 107–120, 2015.

[46] Y. Yang and S. Gopal, "Multilabel classification with meta-level features in a learning-to-rank framework," *Machine Learning*, vol. 88, no. 1-2, pp. 47–68, 2012.

[47] A. Cano, J. M. Luna, E. L. Gibaja, and S. Ventura, "LAIM discretization for multi-label data," *Information Sciences*, vol. 330, pp. 370–384, 2016.

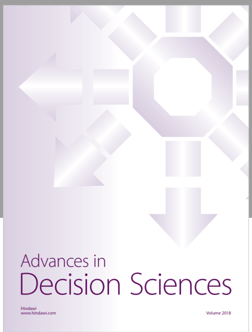
[48] F. van den Bergh and A. P. Engelbrecht, "A study of particle swarm optimization particle trajectories," *Information Sciences*, vol. 176, no. 8, pp. 937–971, 2006.

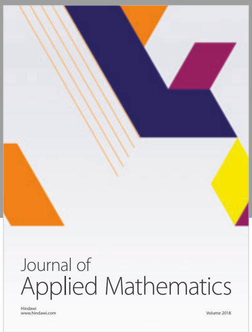
[49] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics Surveys*, vol. 4, pp. 40–79, 2010.

[50] J. Lee, H. Kim, N.-R. Kim, and J.-H. Lee, "An approach for multi-label classification by directed acyclic graph with label correlation maximization," *Information Sciences*, vol. 351, pp. 101–114, 2016.

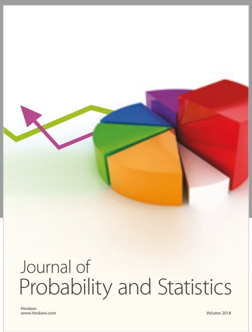
[51] F. Wilcoxon, "Probability tables for individual comparisons by ranking methods," *Biometrics - A Journal of the International Biometric Society*, vol. 3, pp. 119–122, 1947.

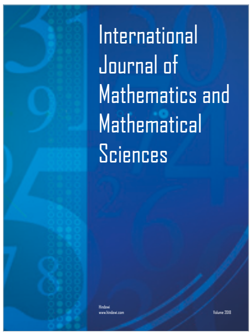
Advances in
Operations Research


Advances in
Decision Sciences

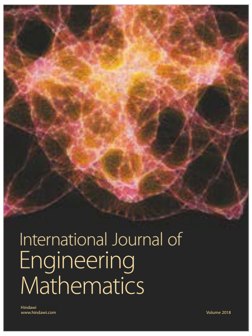

Journal of
Applied Mathematics

The Scientific
World Journal

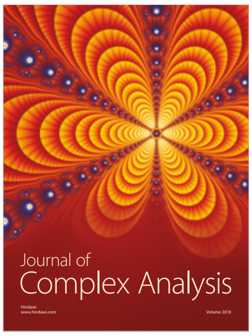

Journal of
Probability and Statistics

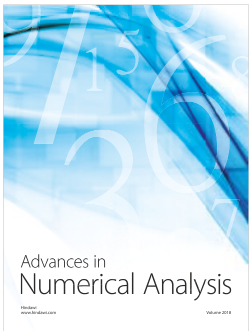

International
Journal of
Mathematics and
Mathematical
Sciences

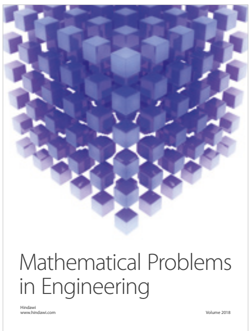
Journal of
Optimization

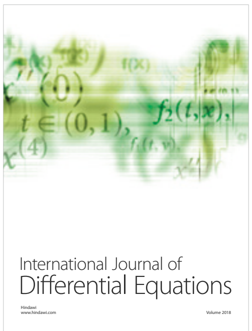

International Journal of
Engineering
Mathematics

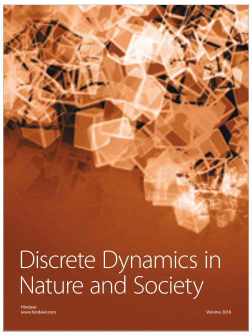
International Journal of
Analysis

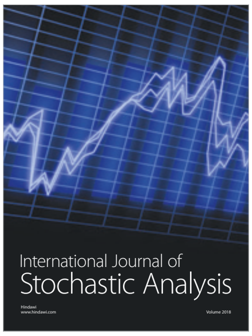

Journal of
Complex Analysis


Advances in
Numerical Analysis


Mathematical Problems
in Engineering

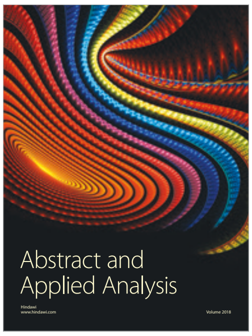

International Journal of
Differential Equations

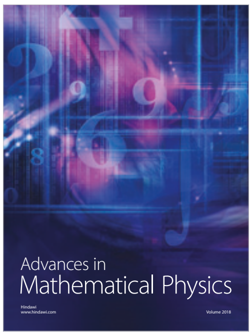

Discrete Dynamics in
Nature and Society


International Journal of
Stochastic Analysis

Journal of
Mathematics

Journal of
Function Spaces


Abstract and
Applied Analysis


Advances in
Mathematical Physics