Hindawi
Journal of Sensors
Volume 2018, Article ID 3419213, 12 pages
https://doi.org/10.1155/2018/3419213

Research Article

Evolutionary Multilabel Feature Selection Using Promising Feature Subset Generation

Jaesung Lee, Wangduk Seo, Ho Han, 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 8 June 2018; Accepted 7 August 2018; Published 18 September 2018

Academic Editor: Grigore Stamatescu

Recent progress in the development of sensor devices improves information harvesting and allows complex but intelligent applications based on learning hidden relations between collected sensor data and objectives. In this scenario, multilabel feature selection can play an important role in achieving better learning accuracy when constrained with limited resources. However, existing multilabel feature selection methods are search-ineffective because generated feature subsets frequently include unimportant features. In addition, only a few feature subsets compared to the search space are considered, yielding feature subsets with low multilabel learning accuracy. In this study, we propose an effective multilabel feature selection method based on a novel feature subset generation procedure. Experimental results demonstrate that the proposed method can identify better feature subsets than conventional methods.

1. Introduction

Recent progress in the development of sensor networks improves the precision of continuous data sensing [1], which increases the coverage of ambient applications such as activity monitoring in daily routines that may involve the concurrent prediction of the activity level and caloric expenditure [2, 3]. Owing to limitations in computational and storage capability [4, 5] and redundant data sensing for denoising [6, 7], composing a strategy that would produce the best accuracy under given data collection conditions is considered one of the most important issues in this field [8]. Consequently, multilabel learning is considered to be a promising approach because it allows for improvements in accuracy by exploiting the dependency among labels [9, 10].

Let $W \subset \mathbb{R}^{|F|}$ denote the set of patterns described by a set of features $F = \{f_1, \cdots, f_d\}$. Then, each pattern $w_i \in W$, where $1 \le i \le |W|$, is assigned to a certain label subset $\lambda_i \subseteq L$ in which $L = \{l_1, l_2, l_3, \cdots, l_{|L|}\}$ and represents a finite set of labels. To attain additional improvements in accuracy, the algorithm has to exploit useful dependencies among labels based on input feature values [11]. For this purpose, the multilabel feature selection that identifies a subset $S \subset F$ with maximum $n << d$ features that provide the largest dependency on L can be used as a promising preprocessing step because it remedies the complicated relation among features and labels by selecting important features and discarding unnecessary ones [12, 13].

Basically, multilabel feature selection is a search problem [14]; it can be achieved by identifying the optimal feature subset that gives the best prediction accuracy from

$$\sum_{k=1}^{n} \binom{d}{k} \tag{1}$$

candidate feature subsets [15]. Because the examination of all feature subsets is impractical, conventional methods employ a heuristic search method that identifies a feasible solution within limited computational costs by sacrificing optimality [16]. Of the many search methods, the evolutionary search

method is considered a promising approach because it effectively narrows down the search space by examining neighbor solutions or feature subsets of the best solutions created from past generations [17, 18].

In the evolutionary search method, the best solution is replaced if a newly created neighbor solution yields a better fitness value. Therefore, generating promising solutions determines the effectiveness of the search. Owing to the extensively wide search space and limited computational cost, a conventional strategy tackling this difficulty is to employ a cheap evaluation method that measures the potential of possible solutions, filtering out unpromising solutions and then validating the exact fitness value of the remaining solutions [19]. However, to the best of our knowledge, there is no serious investigation on this direction from the literatures related to intelligent sensor applications and multilabel feature selection.

In this study, we propose a novel effective evolutionary search method for multilabel datasets. Previous studies considering the intelligent sensor applications incurring multilabel feature selection did not tackle the issue related to the generation of promising feature subsets, resulting in a degeneration of search effectiveness. Our contribution can be summarized as follows:

(i) The proposed method improves search effectiveness by producing a large number of feature subsets with important features and then filters out unpromising feature subsets using a cheap evaluation method.

(ii) A cheap feature subset evaluation method is employed to filter out unpromising feature subsets without checking the fitness value which demands expensive computational cost.

(iii) We compared the performance of conventional multilabel feature wrapper methods and the proposed method on 14 multilabel datasets and conducted 53 standard statistical tests to validate the superiority of the proposed method

2. Related Work

Because multilabel feature selection can improve the learning accuracy as well as the efficiency of a later algorithm by highlighting important features such as multilabel classifier for the concurrent prediction, it gained significant attention from diverse fields [20, 21].

Feature selection methods come in two categories: filters and wrappers. Filter methods rank features based on their own criterion by evaluating the importance of each feature. For multilabel feature selection on multilabel datasets, a simple strategy that changes the label sets to a single label set was often considered, such as a label powerset [22]. This method is advantageous because it enables conventional feature selection methods for single-label datasets. Several conventional filter methods have been reported [23]; however, filter methods commonly suffer from low multilabel classification accuracy, owing to noninteraction with multilabel classifiers or subsequent problems such as imbalance in transformed single-label data. By contrast, wrapper methods evaluate created feature subsets and improve them. In detail, they locate promising feature subsets using a search method employed and then evaluate them using a later learning algorithm [17]. Although the learning algorithm can be different according to the application, recent review indicated that the most frequent choice for the search method is the evolutionary search [24] because it is effective at searching for feasible solution in global perspective. Zhang et al. [14] proposed a multilabel feature selection method based on genetic algorithms. However, a major drawback of the genetic algorithm is their premature convergence to unrefined solutions [17]. On the other hand, a genetic algorithm-based nondominated sorting genetic algorithm-II [25] and multiobjective particle swarm optimization [26] have been used for multilabel feature selection.

Although most studies consider single-label sensory datasets, there are several studies on feature selection methods because of the promising potential. To apply automatic view generation, a semisupervised feature selection method for features extracted from very high-resolution remote sensing images was proposed [27]. Specifically, features are categorized into a series of disjoint groups, and then important features in each group are selected by solving the $l_{11,2}$-norm-based minimization problem. Similarly, a refined feature subset from discrete wavelet transform coefficient features, extracted from artificial tongue sensor signals, was selected by using a dispersion ratio computation [12]. Activity recognition using accelerometers was also shown to be improved by feature selection [28]. There are several studies related to the identification of a set of important features based on the fitness or classification accuracy derived from the learning algorithm. For example, a feature subset can be obtained by iteratively including the best feature at each step, which is referred to as the sequential forward selection algorithm [29]. This technique is applied to the application of chiller fault detection [30], which is an instantiation of an automatic fault detection problem in a smart factory [31]. The genetic algorithm which is one of the most famous evolutionary search methods in the machine learning community was also considered for selecting discriminative features for online bearing fault diagnosis [32]. In addition, the particle swarm optimization technique, which is another popular evolutionary search method, was also used to find the optimal feature subset for intrusion detection [13]. Support vector machine recursive feature elimination has been used for the analysis of correlated gas sensor data [7]. Energy consumption was minimized and the classification accuracy was improved by feature selection from sensor data [5].

3. Proposed Method

3.1. Preliminary. Of the various evolutionary search methods, estimation of distribution algorithm (EDA) has proven effective for solving various problems [24, 33]. Unlike typical evolutionary search methods, to generate new feature subsets, EDAs do not use genetic operators [19]. Instead, conventional EDAs generate new solutions or candidates

using a probability model and update the probability model based on a statistical distribution estimated from the representation of solutions. Thus, it provides an opportunity to generate promising feature subsets by manipulating the probability model. The probability model can be implemented as follows [33, 34]:

$$P^{t+1}(i) = P^t \times (1 - \mathrm{LR}) + F^t(i) \times \mathrm{LR}, \quad (2)$$

where $P^t(i)$ is the selection probability of the i-th feature in the t-th generation, $F^t(i)$ is the probability associated with the i-th feature in the top 50% feature subsets in the t-th generation that are ranked in terms of their fitness values, and LR is the learning rate, which is a user-defined parameter that controls the influence of $F(i)$ to the probability model in the next generation. Through (2), the probability of selecting a feature in the $(t+1)$-th generation, P^{t+1}, is calculated, and in the $(t+1)$-th generation, feature subsets are built. This process is repeated until the maximum allowed computational cost is exhausted. Although there are many stopping criteria, we set the number of spent fitness function calls (FFCs) as the termination condition for all evolutionary search methods employed in this study for a fair comparison against diversified settings and implementations [35].

In the feature selection problem, the algorithm should be capable of searching a huge parametric space; thus, significant computational cost is associated with finding a promising solution. Although simple probability models are easy to implement, it can be insufficient for solving complicated problems, such as pinpointing promising feature subsets in a large search space [36]. For example, in the conventional EDA-based feature selection method, all features are initially assigned the selection probability of 0.5. This means that nonpromising features can be also present in feature subsets. To overcome this drawback, we devise a process for generation of a promising feature subset. Specifically, when creating a feature subset, the algorithm will consider important features more frequently by setting the priority to such features given by an individual feature filter.

After creating the feature subsets, the next step amounts to selecting promising feature subsets. Although good feature subsets can be created using filter methods, there can be nonpromising feature subsets because the creation process is probabilistic and there can be efficient interaction among features. Nonpromising feature subsets consume FFCs and negatively affect the search efficiency. To overcome this problem, we propose a feature subset evaluation method consuming a cheap computational cost. Using the methods of information theory, the proposed method calculates, for each subset, the relevance and redundancy of the subset features. Then, the proposed method selects feature subsets with maximal relevance and minimal redundancy. Because there is a possibility that the proposed solution will be only locally promising, the proposed method uses roulette wheel selection as the selection algorithm [37]. Thus, nonpromising feature subsets are filtered out from the neighbor set, without exact evaluation.

In the proposed method, there are two key functions for the feature subset generation. *create* function makes candidate feature subsets that is composed of relevant features. *select* function selects promising feature subsets among created ones by using roulette wheel selection based on their potential given by a feature subset evaluation method. Figure 1 schematically shows the proposed method. In the first stage, the probability model is initialized, indicating feature subsets containing randomly chosen features will be created frequently. The probability model is represented as a vector where each element encodes the presence of each feature. In the next step, feature subsets are created using *create* function. All feature subsets are assigned random integers, ranging from one to n. If one feature subset is determined to choose two features, the proposed method ranks the features in terms of their importance, using a filter method. In the first iteration, the most important feature is f_4. Then, the proposed method chooses a random number r between 0 and 1 and compares r to the selection probability of f_4 in the probability model, p_4. Since p_4 is greater than r in the example, f_4 is selected and added to the feature subset. In the second iteration, features are again ranked using the filter method. In this case, the features' importance is measured again in terms of relevance and redundancy under the selection of f_4. Thus, the features' ranks can change. In this example, f_2 is the most important feature. Then, another random number r is drawn and compared to the selection probability of f_2, p_2. However, p_2 is lower than r; thus, f_2 is not selected. Then, the second most important feature can be selected. In this example, f_5 is added to the subset of features, and iteration is terminated. Through this process, the proposed method creates a series of new feature subsets including important features. The next step amounts to selecting promising feature subsets among created feature subsets using *select* function. The proposed method chooses m promising feature subsets by roulette wheel selection biased by the proposed feature subset evaluation. Finally, the probability model is updated using m promising feature subsets and (2), to reflect the presence of features in the best half of new feature subsets ranked by fitness value.

3.2. Proposed Search Procedure. The proposed algorithm creates feature subsets to large searching spaces and filters nonpromising subsets using the proposed subset evaluation method that does not incur exact evaluation. Algorithm 1 shows the proposed method. For the population size m and maximal number of FFCs v, the method initializes feature subsets $O(t)$ and the probability model P (line 3). The method generates a set of m feature subsets $O(t)$ through a random assignment of maximum $|F|$ binary bits. The probability model P is an $|F|$ length vector, and each entry in the vector refers to the probability of choosing coordinated features. Each entry is initialized for the distribution of the features of $O(t)$. Then, the created set $O(t)$ is evaluated (line 4). The method set consumed FFCs u to 0

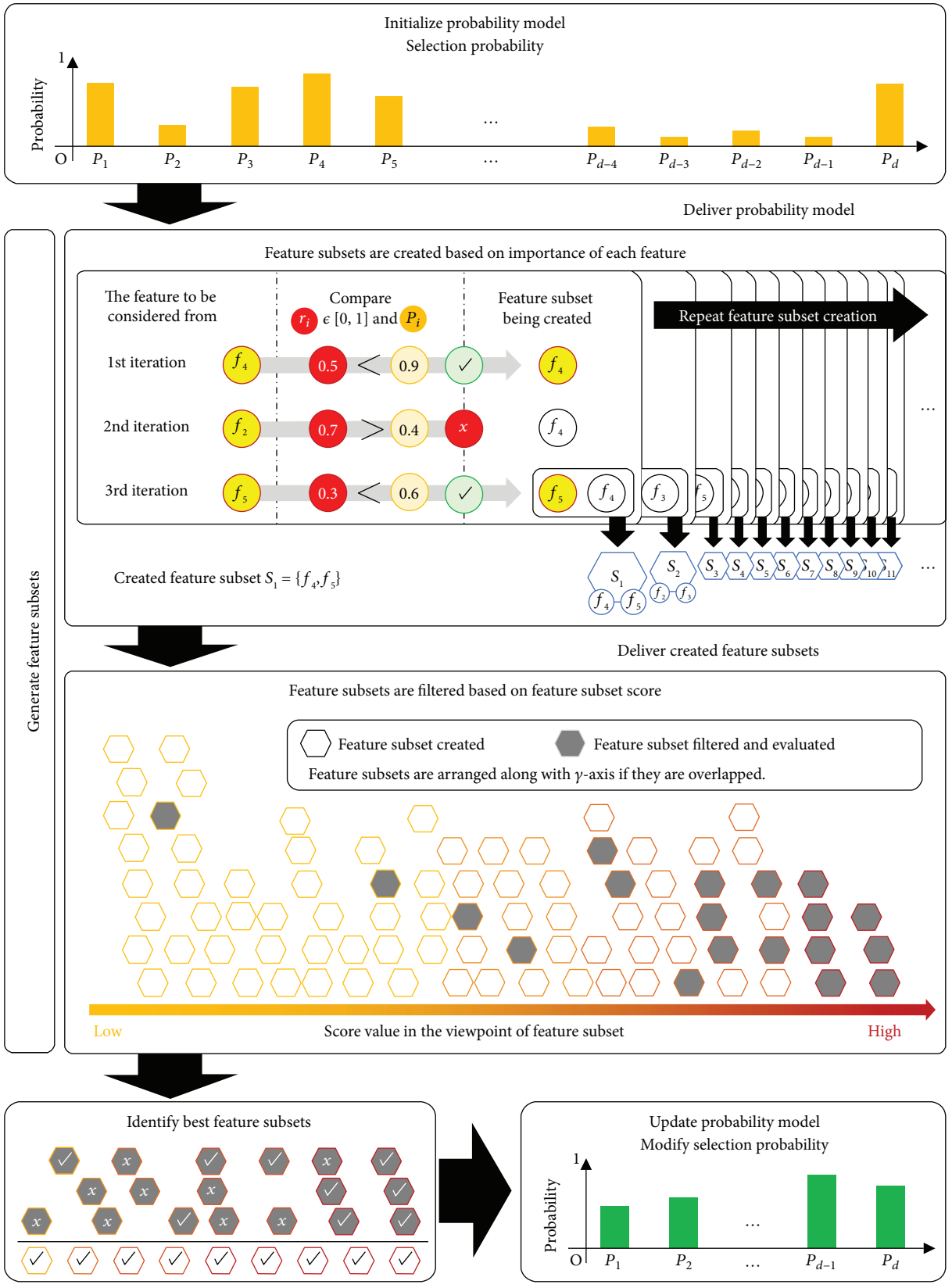

Figure 1: Schematic overview of the proposed method.

1: **Input:** population size m, max FFCs v
2: **Output:** best feature subset S_g
3: initializing $O(t)$ and probability model p
4: evaluating $O(t)$
5: $u \leftarrow 0$ ▷ Set consumed FFCs to 0
6: store the global best feature subset to S_g
7: **while** $u \leq v$ **do**
8: update P by Eq. (2)
9: $E(t) \leftarrow$ *create* neighbor set
10: $O(t+1) \leftarrow$ *select* feature subsets in $E(t)$
11: evaluate $O(t+1)$
12: $u \leftarrow u+m$ ▷ update consumed FFCs
13: update global best feature subset S_g
14: **end while**

Algorithm 1: Proposed method.

1: **Input:** neighbor population size e, probability model P
2: **Output:** neighbor set $E(t)$
3: $E(t) \leftarrow \varnothing$
4: **for** $k = 1$ to e **do**
5: $n \leftarrow$ *random integer value in* $[0, |F|]$ ▷ set a feature subset size randomly
6: $S_k \leftarrow \{\varnothing\}$
7: **for** $i = 1$ to n **do**
8: $R \leftarrow \{f_1, f_2, f_3, \cdots\}$ ▷ ranked features by Eq. (6)
9: **for** $j = 1$ to $|F|$ **do**
10: **if** $P^t(f_j) >$ *random value in* [0,1] **then**
11: $S_k \leftarrow S_k \cup j$
12: **break**
13: **end if**
14: **end for**
15: **end for**
16: $E(t) \leftarrow E(t) \cup S_k$
17: **end for**

Algorithm 2: *create* function.

(line 5) and stores the global best feature subset to S_g (line 6). P is updated by (2) (line 8). The method creates a set of neighbor feature subsets E, which is based on a filter method by *create* function (line 9). Then, m feature subsets are selected by roulette wheel selection weighted by *select* function in the set $E(t)$ and yield the new generation $O(t+1)$ (line 6). The feature subset $O(t+1)$ is evaluated (line 11), and sets consumed FFCs u (line 12). The feature subset S_g, which offers globally optimal performance, is stored and replaced in the procedure (line 13). After all allowed FFCs are consumed, the algorithm returns the feature subset S_g.

Algorithm 2 is a *create* function that shows the process of creating feature subsets. Each feature subset selects random size n features (line 5). To introduce important features more frequently, first, each feature should be ranked by their importance value. To achieve this, we evaluate the importance of each feature using the relevance criterion [20]:

$$I(f_i) = \mathrm{Rel}(f_i) - \mathrm{Red}(f_i), \tag{3}$$

where $\mathrm{Rel}(f_i)$ and $\mathrm{Red}(f_i)$ denote the relevance and redundancy of the i-th feature and $I(f_i)$ denotes the importance of the i-th feature. Although both functions can be implemented differently according to the subject of each study, we use a recent filter method for measuring the importance of features. In the work of [15], we proposed a filter method for multilabel dataset, and was shown to outperform conventional filter methods. Because of this reason, we use this method for measuring the importance of features. Accordingly, $\mathrm{Rel}(f_i)$ can be implemented as

$$\mathrm{Rel}(f_i) = \sum_{l \in L} M(f_i; l), \tag{4}$$

where $M(x;y) = H(x) - H(x,y) + H(y)$ indicates the mutual information between variables x and y and $H(x) = -\sum P(x) \log P(x)$ is the joint entropy obtained from the probability $P(x)$, $P(y)$, and $P(x,y)$. Next, $\mathrm{Red}(f_i)$ can be implemented as

$$\mathrm{Red}(f_i) = \sum_{f \in S} \sum_{l \in L} \frac{M(f_i;l)}{H(f_i)} M(f_i;f). \quad (5)$$

Thus, the feature f_i's importance is measured by

$$I(f_i) = \sum_{l \in L} M(f_i;l) - \sum_{f \in S} \sum_{l \in L} \frac{M(f_i;l)}{H(f_i)} M(f_i;f). \quad (6)$$

Then the rank of each feature can be determined by using (6) and remembered (line 8). After then, the function decides whether to choose a feature from the most important subsets by P (lines 9 to 13). If a feature is chosen, it is added to subset S_k (line 11). In addition, after a subset is created, it is added to the set of neighbor feature subsets E (line 16).

It is well-known fact from the feature selection community that a set of individually good features is not necessarily a good feature subset due to the interaction among features. This means that the created feature subset can be unpromising even though (6) only included important features. To achieve this, *select* function described in Algorithm 3 that shows the process for selecting promising feature subsets in a neighbor set is necessary. In *select* function, a new feature subset filter method is employed [38]. Specifically, it evaluates the fitness of the feature subset as

$$E(S) = \sum_{f_i \in S} \sum_{l \in L} M(f_i;l) - \sum_{f_i \in S} \sum_{f_j \in S} M\left(f_i;f_j\right). \quad (7)$$

By using (7), *select* function ranks feature subsets in the neighbor set E (line 3). Next, the algorithm selects m feature subsets $G(t)$ using roulette wheel selection [37], which is a biased selection weight by (7) (line 4).

In summary, in the generation of a feature subset, the algorithm ranks the importance of features using the filter method and selects the most important feature i based on the probability $P^t(i)$ considering subset S selected at this point. If the i-th feature is not chosen, the next most important feature j can be selected with the probability $P^t(j)$, and the process repeats until a feature is selected. Then, for each neighbor feature subset, (7) ranks the importance of feature subsets, and feature subsets with highest $E(\cdot)$ values are likely to selected.

1: **Input:** neighbor set E, population size m
2: **Output:** filtered set $G(t)$
3: rank feature subsets in set E by Eq. (7)
4: select m feature subsets by roulette wheel selection
5: $G(t) \leftarrow$ selected feature subsets

Algorithm 3: *select* function.

4. Experimental Results

We conducted experiments on 14 datasets from various domains. The Birds dataset is audio data containing samples of multiple bird calls. The Enron and Language Log (Llog) datasets are generated from text mining applications, where each feature corresponds to the presence of a word and each label represents the relevance of each text pattern to a specific subject. The Mediamill dataset contains video data from an automatic detection system. The Medical dataset is sampled from a large corpus of suicide letters obtained from the natural language processing of clinical free texts. The TMC2007 dataset contains safety reports of a complex space system. The remaining eight datasets came from the Yahoo dataset collection. We performed unsupervised dimensionality reduction on datasets, including the TMC2007 and Yahoo collections, consisting of more than 10,000 features. Because our algorithm uses information theory, for numeric features, we performed discretization using the supervised discretization method [39]. Table 1 shows the standard characteristics of the multilabel datasets used in our experiments, including the number of patterns in the datasets $|W|$, number of features $|F|$, type of features, and number of labels $|L|$. The label cardinality measure C ard represents the average number of labels for each instance. The label density measure 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 applications associated with the extracted datasets.

We compared the proposed method with conventional methods, including the genetic algorithm (GA) [14], nondominated sorting genetic algorithm-II (NSGA-II) [25], and multiobjective particle swarm optimization feature selection (MPSOFS) [26]. We considered a conventional multilabel classifier, namely, the multilabel naïve Bayes (MLNB) classifier [14]. We used conventional hold-out cross-validation for each dataset. Of the patterns, 80% were randomly chosen as a training set and the remaining 20% were chosen as a test set. We set the size of the population to 20, and the maximal number of FFCs was limited to 100. In our proposed method, we created 500 feature subsets using the probability model and set the learning rate (LR) to 0.4. The GA and NSGA-II created two offspring feature subsets and one feature subset from mutation operators in each generation. The MPSOFS preserved the global best particle solutions and each particle's best solutions. Thereafter, the MPSOFS updated the velocity values. All experiments were repeated 10 times, and the average measured values were used to compare the performances of the methods.

To measure the methods' performances, we employed the following four evaluation metrics: multilabel accuracy,

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
Enron	1702	1001	Nominal	53	3.378	0.064	753	Text
Llog	1460	1004	Nominal	75	1.180	0.016	304	Text
Mediamill	43,907	120	Numeric	101	4.376	0.043	6555	Video
Medical	978	1449	Nominal	45	1.245	0.028	94	Text
TMC2007	28,596	49,060	Numeric	22	2.158	0.098	1341	Text
Business	11,214	1096	Numeric	30	1.599	0.053	233	Text
Education	12,030	1377	Numeric	33	1.463	0.044	511	Text
Entertainment	12,730	1600	Numeric	21	1.414	0.067	337	Text
Health	9205	1530	Numeric	32	1.644	0.051	335	Text
Reference	8027	1984	Numeric	33	1.174	0.036	275	Text
Science	6428	1859	Numeric	40	1.450	0.036	457	Text
Social	12,111	2618	Numeric	39	1.279	0.033	361	Text
Society	14,512	1590	Numeric	27	1.670	0.062	1054	Text

hamming loss, ranking loss, and normalized coverage. Multilabel accuracy is defined as

$$\text{mlacc}(T) = \frac{1}{|T|}\sum_{i=1}^{|T|}\frac{|\lambda_i \cap Y_i|}{|\lambda_i \cup Y_i|}, \tag{8}$$

where T is a given test set. Hamming loss is defined by

$$\text{hloss}(T) = \frac{1}{|T|}\sum_{i=1}^{|T|}\frac{1}{|L|}|\lambda_i \Delta Y_i|, \tag{9}$$

where λ denotes the correct label subset and Δ denotes the symmetric difference between the two sets. Ranking loss is defined by

$$\text{rloss}(T) = \frac{1}{|T|}\sum_{i=1}^{|T|}\frac{\left|\left\{(a,b)\middle| a \in \overline{\lambda}_i, \psi_{i,a} \le \psi_{i,b}\right\}\right|}{|\lambda_i|\left|\overline{\lambda}_i\right|}, \tag{10}$$

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

$$\text{ncov}(T) = \frac{1}{|L|}\left(\frac{1}{|T|}\sum_{i=1}^{|T|}\max_{l\in\lambda_i}\text{rank}\,(l) - 1\right), \tag{11}$$

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.

Tables 2, 3, 4, and 5 list the experimental results for the different performance measures as averages over the experiments on the employed datasets. The best performance of each dataset is indicated by a bold font. In each table, the last column shows the average rank (Avg. rank) of each comparison method over all the multilabel datasets. In terms of the multilabel accuracy and ranking loss measures, the proposed method outperformed the GA, NSGA-II, and MPSOFS, on all datasets. In terms of the hamming loss, the proposed method outperformed conventional methods on all datasets except TMC2007. In terms of the normalized coverage, the proposed method outperformed the conventional methods on all datasets except Llog.

After measuring the performance of the methods on all datasets, we analyzed the performance using statistical tools. We employed the Friedman test, a widely used statistical test, for comparing multiple methods over a number of datasets [40]. Supposing there are k methods and N datasets, and let R_j denote the average rank for the j-th method under the null hypothesis (i.e., when all of the methods perform equally well). Then, the following Friedman statistic F_F is distributed according to the F-distribution with $k-1$ numerator degrees of freedom and $(k-1)(N-1)$ denominator degrees of freedom as parameters:

$$F_F\frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2}, \tag{12}$$

where χ_F^2 is defined as

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_{j=1}^{k}R_j^2 - \frac{k(k+1)^2}{4}\right]. \tag{13}$$

If F_F is larger than the critical value at a significance level α, the null hypothesis is rejected, implying that the compared methods have different performances. After the

Table 2: Comparison results in terms of multilabel accuracy.

Method	Birds	Enron	Llog	Mediamill	Medical
Proposed	**0.527 ± 0.044**	**0.383 ± 0.012**	**0.249 ± 0.014**	**0.362 ± 0.004**	**0.427 ± 0.030**
GA	0.491 ± 0.055	0.284 ± 0.022	0.209 ± 0.018	0.336 ± 0.031	0.303 ± 0.058
NSGA	0.480 ± 0.040	0.282 ± 0.022	0.208 ± 0.014	0.347 ± 0.013	0.297 ± 0.018
MPSOFS	0.453 ± 0.031	0.206 ± 0.012	0.042 ± 0.002	0.163 ± 0.007	0.286 ± 0.033
Method	TMC2007	Business	Education	Entertainment	Health
Proposed	**0.441 ± 0.004**	**0.672 ± 0.011**	**0.318 ± 0.009**	**0.396 ± 0.004**	**0.537 ± 0.011**
GA	0.435 ± 0.010	0.657 ± 0.021	0.316 ± 0.010	0.361 ± 0.011	0.499 ± 0.011
NSGA	0.434 ± 0.005	0.662 ± 0.013	0.318 ± 0.010	0.362 ± 0.010	0.495 ± 0.009
MPSOFS	0.420 ± 0.005	0.634 ± 0.008	0.283 ± 0.005	0.365 ± 0.010	0.496 ± 0.013
Method	Reference	Science	Social	Society	Avg. rank
Proposed	**0.436 ± 0.009**	**0.288 ± 0.010**	**0.546 ± 0.008**	**0.371 ± 0.008**	**1.00**
GA	0.422 ± 0.012	0.231 ± 0.009	0.517 ± 0.012	0.258 ± 0.011	2.79
NSGA	0.429 ± 0.009	0.237 ± 0.010	0.526 ± 0.009	0.267 ± 0.010	2.64
MPSOFS	0.414 ± 0.017	0.234 ± 0.013	0.527 ± 0.012	0.239 ± 0.007	3.57

Table 3: Comparison results in terms of hamming loss.

Method	Birds	Enron	Llog	Mediamill	Medical
Proposed	**0.061 ± 0.008**	**0.060 ± 0.003**	**0.016 ± 0.001**	**0.034 ± 0.000**	**0.020 ± 0.001**
GA	0.072 ± 0.015	0.100 ± 0.033	0.075 ± 0.073	0.048 ± 0.022	0.023 ± 0.001
NSGA	0.064 ± 0.008	0.104 ± 0.026	0.072 ± 0.071	0.054 ± 0.032	0.022 ± 0.001
MPSOFS	0.135 ± 0.018	0.198 ± 0.007	0.292 ± 0.008	0.174 ± 0.005	0.023 ± 0.001
Method	TMC2007	Business	Education	Entertainment	Health
Proposed	**0.088 ± 0.002**	**0.029 ± 0.001**	**0.042 ± 0.001**	**0.055 ± 0.001**	**0.039 ± 0.001**
GA	0.088 ± 0.004	0.035 ± 0.005	0.046 ± 0.003	0.070 ± 0.007	0.051 ± 0.004
NSGA	0.086 ± 0.004	0.037 ± 0.010	0.048 ± 0.004	0.068 ± 0.004	0.054 ± 0.005
MPSOFS	0.117 ± 0.002	0.079 ± 0.003	0.061 ± 0.002	0.105 ± 0.003	0.067 ± 0.002
Method	Reference	Science	Social	Society	Avg. rank
Proposed	**0.034 ± 0.006**	**0.035 ± 0.003**	**0.025 ± 0.003**	**0.054 ± 0.002**	**1.07**
GA	0.055 ± 0.013	0.051 ± 0.012	0.042 ± 0.007	0.062 ± 0.005	2.71
NSGA	0.047 ± 0.008	0.045 ± 0.008	0.040 ± 0.010	0.060 ± 0.002	2.29
MPSOFS	0.086 ± 0.005	0.110 ± 0.005	0.070 ± 0.002	0.144 ± 0.007	3.93

null hypothesis is rejected, we perform a post hoc test to analyze whether the proposed method performs significantly better than other methods. The Bonferroni–Dunn test is employed [41]. Critical difference (CD) is used to compare the proposed method and one comparison method. CD is defined as

$$\mathrm{CD} = q_\alpha \sqrt{\frac{k(k+1)}{6N}}, \tag{14}$$

where the critical value q_α is constant and is determined by the number of methods and the significance level. If the difference between the two compared methods' average ranks is greater than CD, the better-ranking method is concluded to perform significantly better than the other method. Because our experiment used four methods, including the proposed method, and 14 datasets, we set $k = 4$ and $N = 14$. We employed the Friedman test when the significance level α was 0.05. Table 6 shows the summary of the employed Friedman test. The critical value for 3 and 39 degrees of freedom was 2.845. The Friedman statistic F_F for all performance measures was above the critical value. Thus, the null hypothesis that the compared methods perform equally well was rejected.

To employ the Bonferroni–Dunn test, the calculated CD with $\alpha = 0.05$ was 1.168 since $q_\alpha = 2.394$ at the significance level α of 0.05. Figure 2 shows the CD diagrams for all evaluation measures, where the average rank of each method is on the top of each figure. Our proposed method significantly outperforms other, conventional, methods on all evaluation measures.

Table 4: Comparison results in terms of ranking loss.

Method	Birds	Enron	Llog	Mediamill	Medical
Proposed	**0.115 ± 0.017**	**0.100 ± 0.008**	**0.155 ± 0.023**	**0.060 ± 0.001**	**0.115 ± 0.026**
GA	0.129 ± 0.015	0.133 ± 0.024	0.164 ± 0.023	0.066 ± 0.007	0.145 ± 0.026
NSGA	0.125 ± 0.017	0.149 ± 0.031	0.163 ± 0.023	0.067 ± 0.011	0.139 ± 0.026
MPSOFS	0.132 ± 0.017	0.194 ± 0.011	0.180 ± 0.021	0.159 ± 0.004	0.140 ± 0.024
Method	TMC2007	Business	Education	Entertainment	Health
Proposed	**0.073 ± 0.001**	**0.062 ± 0.027**	**0.089 ± 0.003**	**0.111 ± 0.002**	**0.085 ± 0.028**
GA	0.075 ± 0.002	0.070 ± 0.029	0.100 ± 0.004	0.140 ± 0.007	0.098 ± 0.027
NSGA	0.076 ± 0.002	0.067 ± 0.027	0.100 ± 0.003	0.137 ± 0.009	0.097 ± 0.028
MPSOFS	0.078 ± 0.002	0.096 ± 0.027	0.101 ± 0.003	0.153 ± 0.005	0.098 ± 0.028
Method	Reference	Science	Social	Society	Avg. rank
Proposed	**0.111 ± 0.023**	**0.118 ± 0.003**	**0.075 ± 0.010**	**0.135 ± 0.003**	**1.00**
GA	0.130 ± 0.021	0.154 ± 0.007	0.088 ± 0.012	0.151 ± 0.008	2.86
NSGA	0.128 ± 0.025	0.150 ± 0.006	0.085 ± 0.011	0.153 ± 0.011	2.29
MPSOFS	0.140 ± 0.023	0.157 ± 0.004	0.097 ± 0.012	0.212 ± 0.005	3.86

Table 5: Comparison results in terms of normalized coverage.

Method	Birds	Enron	Llog	Mediamill	Medical
Proposed	**0.194 ± 0.015**	**0.277 ± 0.015**	**0.202 ± 0.025**	**0.196 ± 0.003**	**0.155 ± 0.029**
GA	0.223 ± 0.025	0.337 ± 0.038	0.201 ± 0.021	0.205 ± 0.005	0.180 ± 0.027
NSGA	0.212 ± 0.020	0.336 ± 0.038	0.199 ± 0.025	0.208 ± 0.019	0.178 ± 0.029
MPSOFS	0.222 ± 0.023	0.413 ± 0.014	0.201 ± 0.023	0.330 ± 0.006	0.179 ± 0.027
Method	TMC2007	Business	Education	Entertainment	Health
Proposed	**0.203 ± 0.002**	**0.132 ± 0.023**	**0.148 ± 0.004**	**0.197 ± 0.003**	**0.157 ± 0.023**
GA	0.208 ± 0.004	0.139 ± 0.024	0.159 ± 0.003	0.221 ± 0.009	0.167 ± 0.026
NSGA	0.209 ± 0.004	0.141 ± 0.021	0.159 ± 0.004	0.219 ± 0.009	0.167 ± 0.026
MPSOFS	0.211 ± 0.002	0.168 ± 0.023	0.159 ± 0.004	0.233 ± 0.003	0.166 ± 0.025
Method	Reference	Science	Social	Society	Avg. rank
Proposed	**0.157 ± 0.023**	**0.182 ± 0.004**	**0.127 ± 0.010**	**0.242 ± 0.007**	**1.21**
GA	0.177 ± 0.022	0.215 ± 0.005	0.140 ± 0.011	0.257 ± 0.006	2.86
NSGA	0.171 ± 0.021	0.213 ± 0.006	0.136 ± 0.014	0.261 ± 0.008	2.43
MPSOFS	0.179 ± 0.022	0.215 ± 0.006	0.145 ± 0.011	0.317 ± 0.007	3.50

Table 6: Summary of the Friedman statistics F_F ($k = 4$, $N = 14$) and critical value in terms of each evaluation measure.

Evaluation measure	F_F	Critical value ($\alpha = 0.05$)
Multilabel accuracy	30.333	2.845
Hamming loss	65.642	
Ranking loss	75.472	
Normalized coverage	16.355	

5. Conclusion

To handle multilabel sensor datasets, we proposed an effective search based on a promising feature subset generation method for multilabel feature selection problem. The main contribution of this work is to propose and validate a new feature subset generation method. Specifically, the proposed method generates candidate feature subsets using important features and chooses promising subsets of features without consuming significant computational cost. Experimental results show that our method converges faster than other conventional methods. In the future, we would like to investigate a new feature subset generation that is more effective because the proposed feature subset generation is strongly dependent on the employed filter method, and it may result redundant feature subsets during the search process. In addition, we would like to apply the proposed method to various sensor datasets and compare the performance with conventional feature selection methods considered from sensory data analysis.

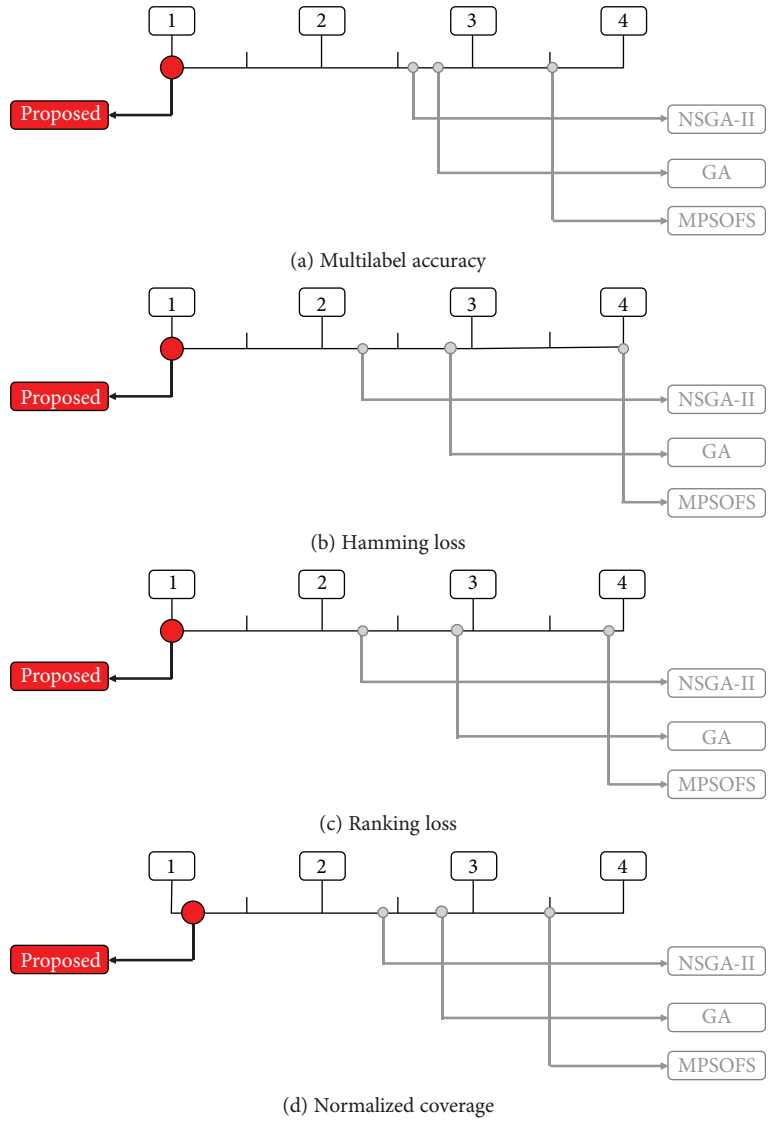

Figure 2: Bonferroni-Dunn test results of four comparison methods with four evaluation measures (significance level $\alpha = 0.05$).

Data Availability

All the employed data used to support the findings of this study have been deposited in the MULAN repository (http://mulan.sourceforge.net/datasets-mlc.html).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by 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] S. Qadri, D. M. Khan, S. F. Qadri et al., "Multisource data fusion framework for land use/land cover classification using machine vision," *Journal of Sensors*, vol. 2017, Article ID 3515418, 8 pages, 2017.

[2] A. Alhamoud, V. Muradi, D. Bohnstedt, and R. Steinmetz, "Activity recognition in multi-user environments using techniques of multi-label classification," in *Proceedings of the 6th International Conference on the Internet of Things - IoT'16*, pp. 15–23, Stuttgart, Germany, 2016.

[3] R. Gravina, P. Alinia, H. Ghasemzadeh, and G. Fortino, "Multi-sensor fusion in body sensor networks: state-of-the-art and research challenges," *Information Fusion*, vol. 35, pp. 68–80, 2017.

[4] B. Du, Z. Wang, L. Zhang, L. Zhang, and D. Tao, "Robust and discriminative labeling for multi-label active learning based on maximum correntropy criterion," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1694–1707, 2017.

[5] 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.

[6] D. Li, Y. Zhou, G. Hu, and C. J. Spanos, "Optimal sensor configuration and feature selection for AHU fault detection and diagnosis," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 3, pp. 1369–1380, 2017.

[7] K. Yan and D. Zhang, "Feature selection and analysis on correlated gas sensor data with recursive feature elimination," *Sensors and Actuators B: Chemical*, vol. 212, pp. 353–363, 2015.

[8] S. Cheng, Z. Cai, J. Li, and H. Gao, "Extracting kernel dataset from big sensory data in wireless sensor networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 4, pp. 813–827, 2017.

[9] R. Kumar, I. Qamar, J. S. Virdi, and N. C. Krishnan, "Multi-label learning for activity recognition," in *2015 International Conference on Intelligent Environments*, pp. 152–155, Prague, Czech Republic, 2015.

[10] J. Read, L. Martino, P. M. Olmos, and D. Luengo, "Scalable multi-output label prediction: from classifier chains to classifier trellises," *Pattern Recognition*, vol. 48, no. 6, pp. 2096–2109, 2015.

[11] 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.

[12] T. Liu, Y. Chen, D. Li, and M. Wu, "An active feature selection strategy for DWT in artificial taste," *Journal of Sensors*, vol. 2018, Article ID 9709505, 11 pages, 2018.

[13] X. Teng, H. Dong, and X. Zhou, "Adaptive feature selection using v-shaped binary particle swarm optimization," *PLoS One*, vol. 12, no. 3, article e0173907, 2017.

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

[15] 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.

[16] Z. Michalewicz and D. B. Fogel, *How to Solve It: Modern Heuristics*, Springer Science & Business Media, 2013.

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

[18] J. Lee, W. Seo, and D.-W. Kim, "Effective evolutionary multi-label feature selection under a budget constraint," *Complexity*, vol. 2018, Article ID 3241489, 14 pages, 2018.

[19] A. Zhou, J. Sun, and Q. Zhang, "An estimation of distribution algorithm with cheap and expensive local search methods," *IEEE Transactions on Evolutionary Computation*, vol. 19, no. 6, pp. 807–822, 2015.

[20] 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.

[21] 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.

[22] J. Read, "A pruned problem transformation method for multi-label classification," *Proceedings of New Zealand Computer Science Research Student Conference*, 2008, pp. 143–150, Christchurch, New Zealand, 2008.

[23] N. Spolaor, 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.

[24] 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.

[25] J. Yin, T. Tao, and J. Xu, "A multi-label feature selection algorithm based on multi-objective optimization," in *2015 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, Killarney, Ireland, 2015.

[26] 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, p. 376, 2017.

[27] X. Chen, W. Liu, F. Su, and G. Zhou, "Semisupervised multi-view feature selection for VHR remote sensing images with label learning and automatic view generation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 6, pp. 2876–2888, 2017.

[28] P. Gupta and T. Dallas, "Feature selection and activity recognition system using a single triaxial accelerometer," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 6, pp. 1780–1786, 2014.

[29] A. Jain and D. Zongker, "Feature selection: evaluation, application, and small sample performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, pp. 153–158, 1997.

[30] K. Yan, L. Ma, Y. Dai, W. Shen, Z. Ji, and D. Xie, "Cost-sensitive and sequential feature selection for chiller fault detection and diagnosis," *International Journal of Refrigeration*, vol. 86, pp. 401–409, 2018.

[31] Y. Luo, Y. Duan, W. Li, P. Pace, and G. Fortino, "A novel mobile and hierarchical data transmission architecture for smart factories," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 8, pp. 3534–3546, 2018.

[32] R. Islam, S. A. Khan, and J.-m. Kim, "Discriminant feature distribution analysis-based hybrid feature selection for online bearing fault diagnosis in induction motors," *Journal of Sensors*, vol. 2016, Article ID 7145715, 16 pages, 2016.

[33] M. Perez, D. M. Rubin, T. Marwala, L. E. Scott, and W. Stevens, "A population-based incremental learning approach to microarray gene expression feature selection," in *2010 IEEE 26-th Convention of Electrical and Electronics Engineers in Israel*, pp. 10–14, Eilat, Israel, 2010.

[34] S. Baluja, "Population-based incremental learning: a method for integrating genetic search based function optimization and competitive learning," Technical Report Carnegie-Mellon University Pittsburgh PA Department of Computer Science, 1994.

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

[36] M. Pelikan, D. E. Goldberg, and F. G. Lobo, "A survey of optimization by building and using probabilistic models," *Computational Optimization and Applications*, vol. 21, no. 1, pp. 5–20, 2002.

[37] A. Lipowski and D. Lipowska, "Roulette-wheel selection via stochastic acceptance," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 6, pp. 2193–2196, 2012.

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

[39] 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.

[40] J. Demsar, "Statistical comparisons of classifier over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[41] O. J. Dunn, "Multiple comparisons among means," *Journal of the American Statistical Association*, vol. 56, no. 293, pp. 52–64, 1961.