IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 14, 2018, accepted January 15, 2019, date of publication January 23, 2019, date of current version February 22, 2019.

*Digital Object Identifier 10.1109/ACCESS.2019.2894366*

# Feature Selection and Its Use in Big Data: Challenges, Methods, and Trends

**MIAO RONG[1], DUNWEI GONG[2], (Member, IEEE), AND XIAOZHI GAO[3]**

[1]School of Information and Electrical Engineering, China University of Mining and Technology, Xuzhou 221006, China
[2]School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China
[3]School of Computing, University of Eastern Finland, 70211 Kuopio, Finland

Corresponding author: Dunwei Gong (dwgong@vip.163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61773384, Grant 61876184, Grant 61876185, Grant 61873105, Grant 61703188, Grant 61573361, Grant 61573362, Grant 61673404, Grant 61763026, and Grant 61473299, and in part by the National Key Research and Development Program of China under Grant 2018YFB1003802-01. The work of X. Gao was supported by the National Natural Science Foundation of China under Grant 51875113.

**ABSTRACT** Feature selection has been an important research area in data mining, which chooses a subset of relevant features for use in the model building. This paper aims to provide an overview of feature selection methods for big data mining. First, it discusses the current challenges and difficulties faced when mining valuable information from big data. A comprehensive review of existing feature selection methods in big data is then presented. Herein, we approach the review from two aspects: methods specific to a particular kind of big data with certain characteristics and applications of methods in classification analysis, which are significantly different to the existing review work. This paper also highlights the current issues of feature selection in big data and suggests the future research directions.

**INDEX TERMS** Feature selection, big data, data mining, applications.

## I. INTRODUCTION

Big data, as one of the IT buzzwords, generally have the following three characteristics: large volume, wide variety, and rapid change [1]–[3]. In terms of volume, Bezdek and Hathaway divided data into a number of categories according to the size of a datum, and pointed out that the size of big data should go up to 10 Terabytes (TB, 1TB=$10^{12}$ bytes) or more [4]–[6]. With respect to variety, statistics have shown that in numerous fields, such as the Internet [7], [8], astronomy [3], biomedicine [5], and geoinformatics [9], there are massive data to be exploited with a great variety of features. In terms of their types and formats, there are not only plenty of structured data, but also massive semi-structured and/or non-structured data, such as text messages and videos, which are widespread, where the former has a fixed format; whereas the data formats of the latter two are often inflexible. Compared with structured data, non-structured data exist more widely in practice. However, they are less useful due to the disordered distribution of these data [3]. Semi-structured data have additional information to facilitate processing them, even though they are not structured, which can be recognized as a particular type of structured data. It should be noted that non-structured data cover many more areas than their structured counterparts. Regarding change, the concept of dataflow is proposed to describe data changing continuously by time, which is required to be utilized from the dynamic environment. Based on these, Manyika *et al.* [10] defined big data as " a dataset whose size exceeds the capability of conventional dataset manage systems in acquiring, storing, processing, and analyzing ". The challenge due to those 3V characters, i.e., volume, variety and velocity, has become the focus of learning methods when dealing with big data. Additionally, redundancy and irrelativeness, which are significant in big data with the goal of avoiding losing effective material, often make the mining process more crucial.

The associate editor coordinating the review of this manuscript and approving it for publication was Ruqiang Yan.

### A. CONTRIBUTIONS OF FEATURE SELECTION

One contributing commitment, feature selection(FS), has already facilitated data mining for its good performance of seeking correlated features and deleting redundant or uncorrelated features from the original dataset [11], [12]. Feature selection is one of the most important data processing techniques, and is frequently exploited to seek correlated features and delete redundant or uncorrelated features from a

feature set [13]. Random or noisy features often disturb a classifier learning correct correlations, and redundant or correlated features increase the complexity of a classifier without adding any useful information to the classifier [14]. A variety of feature selection methods, such as filter, wrapper, and embedded approaches [13] [15], have been developed.

As mentioned above, scalability is a major issue in big data processing systems. The enormous redundancy or irrelevance absolutely accounts for it, not only consuming computing resources, but also affecting processing performance. On this occasion, if this useless information can be removed while valuable clues are retained, the dimension of big data will be greatly lowered, and as a consequence, apart from the computational efficiency, the processing performance of big data will be improved. As a result, studying feature selection approaches for big data so as to obtain a feature subset with superior divisibility is of considerable necessity.

Recently, some researchers have applied these methods to high dimensionality domains, such as DNA microarray analysis [16]–[19], text classification [20]–[23], information retrieval [24]–[26], and web mining [27]–[29]. Online feature selection methods have also been applied to streaming data [30] and valuable information has been extracted from noisy data [31], albeit on a small-scale and with a huge dimension [32], [33].

### B. CHALLENGES OF FEATURE SELECTION

Compared to traditional data, some influential points need to be highlighted on extracting valuable information from big data. Taking the 3V characteristics into consideration, traditional feature selection methods face the following threefold challenges with respect to the case of big data: (1) traditional feature selection methods usually require large amounts of learning time, so it is hard for processing speed to catch up with the change of big data; (2) generally, big data not only include an immense amount of irrelevant and/or redundant features, but also have possible noises of different degrees and different types, which greatly increases the difficulty of selecting features; (3) some data are unreliable/forged, due to different means of acquisition, or even loss, which further enhances the complexity of feature selection.

Due to the properties of big data, existing feature selection methods face demanding challenges in a variety of phases, e.g. the speed of data processing, tracing concept drifts, and dealing with incomplete and/or noisy data. Thus, studying pertinent feature selection methods for big data is of considerable urgency. However, the available methods are extremely specific, and how to extract valuable information from big data based on tackling and analyzing them is still an open issue.

Apart from our review, Bolón-Canedo *et al.* [12] presented a review of feature selection in the context of big data, which mainly describes available feature selection methods classified by practical applications and the next future needs [12]. Unlike their work, we aim to review and compare studies to date regarding the threefold challenges mentioned above, with an analysis of possible challenges and trends in future research. Additionally, we discuss the applications of feature selection methods in several specific kinds of data and classification analysis.

The structure of our paper is explained as follows: Section II looks back to the feature selection methods for traditional data. Next, available feature selection methods and difficulties with processing big data are analyzed in Section IV. Section VI summarizes the paper, and provides several promising directions for further research.

## II. BASIC FEATURE SELECTION FRAMEWORK

Feature selection, also known as variable selection, attribute selection or variable subset selection, is a data mining technique targeting at selecting an optimal subset of features from the whole feature set that renders the best performance in terms of well-defined criteria. Here, a feature refers to an attribute of data, which represents the function of these data in a certain aspect. Since feature selection performs well in simplifying the model, shortening training times, and reducing the variance of the model, researchers can interpret and understand the pattern of the data model more easily by using feature selection. Yu *et al.* [34] pointed out that a good feature selection method should be capable of selecting different features with a high degree of correlation and the optimal classification results.

### A. FEATURE SELECTION FRAMEWORK

A feature selection method can be divided into two parts, i.e., a feature set selection technique that accounts for how to select features from the original entire set, and a feature set evaluation technique that presents how to evaluate the feature subsets [14], [35]. The process of feature selection is shown in Algorithm1 and Fig.1.

**Algorithm 1** The process of feature selection

1: input the original dataset, $X$;
2: **while** the termination condition is not met **do**
3: generate the feature subset, $F$, by searching strategies;
4: evaluate the feature subset, $F$, by evaluation criteria;
5: **end while**
6: **return** $F$;

In Algorithm1, the feature subset can be generated by searching strategies (in **Line3**), such as the random search strategy, the stepwise addition or deletion of features, and heuristic search methods. After a feature subset $F$ has been obtained, its performance of it must be assessed (in **Line4**). Figure1 depicts feature selection as a kind of learning method, which aims to find the appropriate variable subset for users.

#### 1) BENEFITS FROM FEATURE SELECTION

The basic idea of using feature selection is to obtain a new dataset with neither redundant features nor irrelevant features,

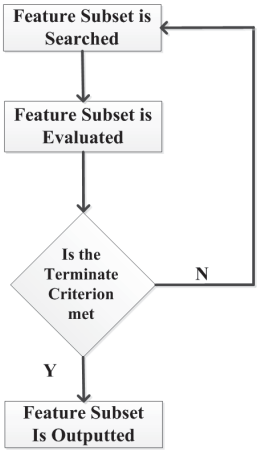



**FIGURE 1.** The flowchart of feature selection.

as well as containing the original data pattern, and not losing any useful information in the original dataset. Nowadays, feature selection methods are widely employed with their capability of dimension reduction, for instance, in the field of written text and DNA microarray analysis, and show their advantages when the number of features is large while the number of samples is small. Compared with feature extraction, feature selection aims to find the features which can describe the original dataset precisely and briefly whereas the latter aims to create new features based on the original dataset. It should be noted that some related features may be redundant since there might be another features which are strongly correlated [36]–[38].

Performing feature selection on a data set has at least the following three advantages: (1) the selected features can be employed to build a brief model for describing original data and thus are beneficial for improving the performance of criteria; (2) the selected features can reflect the core characters of original data and thus are helpful for tracing concept drifts of data expression with good robustness; and (3) the chosen features can help the decision-maker pick valuable information from a large number of noisy data [38], [39].

### *B. TAXONOMIES AND COMPARISONS*

Commonly, feature selection techniques can be classified into filter, wrapper, and embedded according to the means of combining a classifier and a machine learning approach when selecting a feature subset [38], [40]–[42], regardless of supervised or unsupervised methods [37], [43], [44]. The noticeable difference between these two methods lies in whether or not the class labels are available. In the field of feature selection, the former is under condition of available label information employed to evaluate the significance of features and provides rankings of these features. The latter seeks hidden structures in unlabeled data and constructs a feature selector by means of intrinsic properties of data [45]. Due to various machine learning and data mining tasks, there are different evaluation criteria for $F$ Algorithm1 in [46]. For supervised feature selection algorithms, separability criteria, consistency criteria and error rate criteria are favorable for evaluating the performance of methods. In contrast, for those unsupervised feature selection methods, the clustering validity criteria, information theoretical criteria, and feature similarity criteria are now widely seen [47]–[49].

#### 1) FILTER TECHNIQUES

The filter feature selection method is the algorithm that selects the features without evaluating the performance metric of the classifier's model and the selected features [50]. It assumes the data are completely independent of classification algorithms and forms the feature subset according to the importance of a feature measured by its contribution to the class attributes. The performance metric on the output of the classification algorithm is not employed to assess a feature subset, while the measurement works only by data distribution [39], [51], [52].

There are a variety of filter measures which are classified according to the way of combining the features and the class attributes, such as distance-based measures, probability-based measures, mutual information-based measures, consistency measures, and neighborhood graph-based measures [53]. Therefore, the key to a filter approach lies in defining and exploring the relevance between each feature and the class attributes. Accordingly, the measurement of strongly relevant, weakly relevant and irrelevant features is presented in different aspects by researchers [30], [52], [54]. For example, Xindong *et al.* [30] defined relevance based on the exclusion of the conditional independence, whereas Kira and Rendell [55] described the RELIEF algorithm to estimate the weights of the features. Representative filter methods are RELIEF [55], FOCUS [56], and MIFS [54].

The benefits of filter methods are that they are independent of a learning process, have good robustness for the concept drift of data expression [53], [57], [58], and are time-effective because there is less computational complexity. However, they have the following drawbacks: greatly relying on the stopping criteria (a threshold for determining when to stop these methods) and the mechanisms for calculating the importance of a feature [59]. Besides, the strategy of seeking features is an influential factor on filter-based feature subset evaluation methods.

Although the selection process of filters relies little on the classification algorithms, the best filter measure is likely to be classifier specific, since different classifiers perform differently when combined with the same filter [35]. Recently, Freeman *et al.* [53] compared 16 commonly used filters and combined them with two classifiers, K-Nearest Neighbor(KNN) and Support Vector Machine(SVM)for 40 datasets. Their empirical results in terms of classification accuracy give an indication of which filter measures may be appropriate for use with different classifiers.

### 2) WRAPPER TECHNIQUES

Wrapper is a kind of black-box procedure, as knowledge is required in advance [35], [40]. It employs a performance metric based on the classification algorithm to evaluate a candidate feature subset, and conducts a search for the optimal subset based on the evaluation results [52], [60]. It is a kind of subset evaluation techniques with the exception of learning algorithms measuring classification performances. Wrapper methods first divide samples into the training subset and the testing subset. Next, the training subset is used to train the classifier, and the testing subset is employed to verify the classification performances. The advantages of wrapper methods are as follows: they model features' dependencies [61], interact with the classifier, and often yield good results [62]. However, they consume a huge amount of computational time, are highly reliant on a learning process and have a high risk of overfitting for small data.

The process of wrapper is depicted in Figure 2.

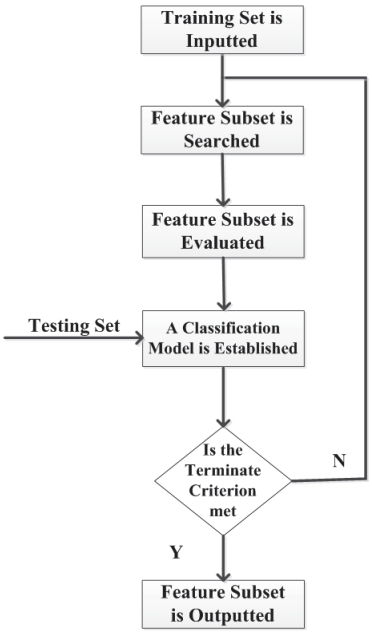



**FIGURE 2.** The wrapper procedure.

From Figure2, an initial state, a termination condition, and a search engine are required during the process of a wrapper method. In addition, as wrappers are associated with the learning algorithm, the combination of features, the criterion for evaluating the performance and the type of a classifier are crucial factors that influence the classification results [63]. Similar to filter, the type of classifiers has a contributing impact on the performance of wrapper-based feature selection methods according to the research conducted by Shanab et al. [62].

### 3) EMBEDDED TECHNIQUES

Embedded methods incorporate the learning progress of a classifier into feature selection [64] and search an optimal feature subset by optimizing a function designed in advance [65]–[67]. In the learning process, the classifier deletes features that have a minor influence on the classification result, and retains good features into a feature subset. Like wrapper methods, embedded methods are specific to classifiers. The benefits of embedded techniques are that they communicate with the classifier, and have a smaller computation complexity than wrappers. The computer-load necessity can be avoided in embedded measures for not reevaluating the performance of a classifier by dividing the whole feature set into the above two parts. However, they are apt to be influenced by the function for optimizing the performance of a feature subset [67], the type of a classifier, and the settings of its related parameters. These factors have a significant impact on the speed and accuracy of an embedded method [64]. In addition, embedded methods cost more in terms of computation than filter ones. Popular embedded methods are Recursive Feature Elimination for Support Vector Machines (SVM-RFE) [68]–[70] and Feature Selection-Perceptron (FS-P) [71]–[73].

Figure3 indicates the characteristics of the embedded-based feature selection techniques.

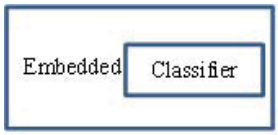



**FIGURE 3.** The basic idea of embedded techniques from [13].

### 4) COMPARISONS

The above studies have resulted in many feature selection methods, most of which, however, aim only at specific backgrounds. In addition, comparing the performances of these methods is not easy. Herein, we provide a brief comparison in Table1 of filter, wrapper, and embedded methods.

Bolón et al. [13] compared some frequently used filter methods mentioned above in terms of whether they are univariate or multivariate and their computing cost. According to them, univariate methods are fast and scalable, but ignore feature dependencies, while multivariate filters model feature dependencies at the cost of being slower and less scalable than univariate techniques [13].

Table2 presents the comparison description of these methods (where $n$ is the number of samples and $m$ is the number of features).

## III. VARIANTS AND EXTENSIONS OF FEATURE SELECTION

### A. HYBRID AND ENSEMBLE METHODS

By combining the advantages of the above methods, various hybrid feature selection methods have been developed [16], [74]–[78], including the combination of two filter methods, and that of one filter strategy with one wrapper strategy. These hybrid methods can take advantages

**TABLE 1. Comparison of Commonly Used Feature Selection Methods.**

| Accuracy | Filter | Wrapper | Embedded |
|---|---|---|---|
| Interact with classifiers | No | Yes | Yes |
| Computational cost | Comparatively low | Comparatively high | Depends |
| Accuracy | Comparatively low | Comparatively high | Comparatively high |
| Model feature dependence | Depends | Yes | Yes |
| Robustness | Yes | Yes | Yes |
| Risk of overfitting | No | Yes | Yes |

**TABLE 2. Comparison of commonly used feature selection methods according to Bolón et al. [13].**

| | Type | Uni/multivariate | complexity |
|---|---|---|---|
| RELIEF-F | Filter | Multivariate | $n^2m$ |
| mRMR | Filter | Multivariate | $nm^2$ |
| INTERACT | Filter | Multivariate | $nm^2$ |
| FCBF | Filter | Multivariate | nmlogm |
| Fisher score | Wrapper | Univariate | nm |
| SVM-RFE | Embedded | Multivariate | $nm^2$ |

of multiple feature selection methods and outperform a single method.

Zhang *et al.* [76] proposed a hybrid feature selection method combining Relief-F and mRMR for gene expression data. Relief-F is first used to look for a candidate gene set, and then the mRMR method is used to directly reduce redundancy for selecting a compact yet effective gene subset from the candidate set. Luo *et al.* [79] have proposed a two-step algorithm to combine the feature selectors for textual information in advertisements on the web. The algorithm first intersects two global feature selection results and then performs a local feature selection. Their experimental results indicate that their combination methods are efficient for a specific background. However, they cannot guarantee an optimal feature subset [80].

Due to the weaknesses of hybrid methods, the ensemble method has been studied to overcome the above drawback [50], [81], [82]. In this method, several strategies are simultaneously employed to find a number of feature subsets, and the final feature subset is produced by integrating these subsets [19], [83], [84]. The benefits of this kind of method are that they obtain the optimal subsets by combining base classifiers built with different feature subsets and show good capability in tackling dynamic data. However, for a specific background, the ensemble measures can present various drawbacks [19]- one main problem which needs to be considered when building an ensemble model is diversity [85]. Diversity can be achieved through various data sets, feature subsets, or classifiers.

Xia *et al.* [86] proposed a feature ensemble plus sample selection method for domain adaptation in sentiment classification. This approach can yield significant improvements compared to individual feature ensemble or sample selection methods to take full account of two attributes, i.e. labeling adaptation and instance adaptation. In addition, some effective methods for feature selection problems have been proposed, such as improved Fisher score algorithm [87] and enhanced bare-bones particle swarm optimization (BPSO) [88]. Moreover, for some specific problems, such as unreliable data [89], [90], incomplete data [91]–[94], text data [95], and costly data [96], researchers have also proposed the corresponding feature selection methods. This kind of methods usually concentrates on improving the performance of search strategy of the optimal subset.

### B. FEATURE SELECTION BASED ON HEURISTIC ALGORITHMS

Due to the advantages of heuristic algorithms, e.g. a small number of parameters to be tuned that are easy to implement and independent of the gradient of an optimization objective, more and more studies have been focused on utilizing these heuristic algorithms to deal with feature selection problems. Representative heuristic algorithms include genetic algorithms [68], [81], [97] [98], differential evolutional algorithms [99], [100], simulated annealing [14], particle swarm optimization [101]–[103], tabu search [104]–[106],and Fisher score algorithms [87], etc. These methods can generally achieve a good feature subset with a fast speed, making the study of feature selection incorporated with search strategies a new trend.

Oh *et al.* [107] first attacked the problem of feature selection with genetic algorithms, followed by a review of the popular feature selection methods based on genetic algorithms conducted by Abd-Alsabour [15], where they pointed out that less knowledge required related to the domain of a problem makes genetic algorithms more suitable for feature selection than the traditional search strategies.

Particle swarm optimization(PSO) is a relatively new heuristic technique inspired by the behavior of bird flocks. Due to its advantages, such as simplicity, fast convergence, and population-based search, researchers have employed the PSO to select the feature subset. Wang*et al.* [108] proposed a feature selection method based on the rough set and PSO. To address the shortcomings of the standard PSO [109], various variants of PSO have been studied and applied to the problem of feature selection, including geometric PSO [110], chaotic binary PSO [111], discrete PSO with adaptive feature selection [112], Taguchi chaotic binary PSO [113], and bare-bones PSO [89], which we have done in our previous work.

Similarly, differential evolution (DE) is an optimization method based on population with super global search capability. In recent years, it has been used for feature selection with interesting classification results [38], [114], [115]. Additionally, two of our previous works on multi-objective feature selection are based on DE [96], [116]. As these two techniques are focused on different backgrounds, it is hard to compare and discuss which of the two is better. This is also a problem when comparing most of the available feature selection measures.

Huang [117] proposed a classification method using ant colony optimization (ACO), which optimizes both the parameters of a feature subset and SVM. Su and Lin [57] incorporated an electromagnetism-like mechanism into a wrapper method. Lin *et al.* [118] integrated the simulated annealing with SVM for feature selection.

### C. OPTIMALITY MODELS

Note that before selecting features, researchers should formulate the problem of feature selection as an optimization model. Some researchers have considered the problem as a single-objective model, e.g. maximizing the accuracy in classification [114], [115]. A feature selection problem generally includes several conflicting objectives, e.g. the number of features, the performance in classification, and/or the reliability of features [89], [96]. Formulating feature selection as a multi-objective problem is the premise of obtaining a series of non-dominated feature subsets, which is beneficial to meet various requirements in real-world applications. Herein, we introduce two examples. One is the previous results obtained by our experiments on the dataset, 'Sonar', on this two-objective feature selection problem with the number of selected features and the classification accuracy, shown separately in Fig.4.

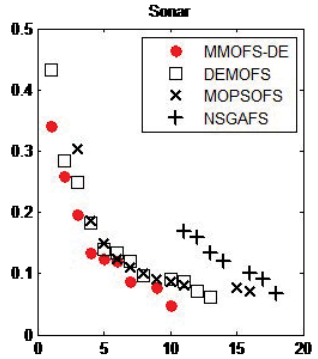



**FIGURE 4.** An example of a two-objective feature selection problem.

Fig.4 has a trend that if more features are selected, a higher accuracy can be obtained. However, the maximal size of the variable subset is still far below the variety of original dataset. This means that the redundant and irrelevant features are removed from the original dataset, and our feature selection approach is able to work well. However, a set of results can be viewed simply as a reference; for a multi-objective feature selection problem, balancing each objective and looking for the most suitable solution are desirable.

The other example is the mathematical model of a wrapper method for unreliable data [89]. Since sample data are unreliable, the reliability degree (RD) in Equal.1, not merely the classification accuracy (CA) in Equal.2, is taken into account for evaluating a feature subset. Herein, if a feature $i$ is selected, then $x_i$ is set as 1; otherwise, $x_i$ is set as 0. $e_i$ is the value within [0,1] to represent the reliability of a feature without loss of generality. The bigger the value, the higher the RD value of this feature. For measuring CA, we adopt the one nearest neighbor classifier, where the testing dataset only contains one attribute each time to evaluate the classification accuracy of a solution (feature subset). If the constructed classifier can predict the class of the testing data clue $x$, then $S_i$ is set as 1; otherwise, $S_i$ is set as 1.

$$f_1(x) = \frac{\sum_{i=1}^{N} x_i \cdot e_i}{\sum_{i=1}^{N} x_i} \tag{1}$$

$$f_2(x) = \frac{1}{K} \sum_{i=1}^{K} S_i(x) \tag{2}$$

The above studies have resulted in many feature selection methods, most of which, however, aim only at specific backgrounds. In addition, comparing the performance of these methods is not easy.

### D. LIMITATIONS OF TRADITIONAL FEATURE SELECTION METHODS FOR BIG DATA

Since practical data may include noises with different degrees, types and formats, or values that are approximate zero, which makes the determination of the correlation degree between features difficult, feature selection is still an unsolved problem. In the case of big data, as vast ever growing data emerge while the existing measurements are inadequate, there is a growing need for efficient feature selection methods for big data.

Due to the three characteristics for big data mentioned above, different analytical modes must be considered for different application requirements [8].

Feature selection methods for traditional data are like a kind of offline method. However, as there is an immense amount of irrelevant and/or redundant features as well as the large volume, how to decrease the computational cost without the classification accuracy deteriorating is an urgent issue. Additionally, with respect to the wide variety of big data, efficient feature selection methods are required to extract valuable information from data with a small size and various formats or types. Moreover, for dynamic data, traditional feature selection methods have difficulty in tracking the changes of the data and since no complete knowledge is known in advance, constructing the classifier's model difficult using

traditional methods. Finally, on account of the precision of equipment or environmental disturbance, dealing with the severe lack or unreliability of some attributes's values from big data needs more attempts.

## IV. FEATURE SELECTION METHODS FOR BIG DATA

Complex characteristics of data bring about a difficulty in obtaining a common feature selection method for big data. A method specific to a background is feasible. Accordingly, in this section, we will review the available feature selection methods for big data according to the particular types of data they use to handle and the applications in analysis. The first part includes static big data, dynamic data, missing data, heterogeneous data, unreliable data and imbalanced data, while the latter part consists of applications in text analysis. In addition, after looking back to available feature selection methods for a kind of data, we also describe what we can do in terms of further research.

### A. SPECIFIC TO SEVERAL PARTICULAR KINDS OF DATA

#### 1) STATIC BIG DATA

The progress of science and technology contributes to a world full of information, and data is the clue to information. Some common characteristic or even rational effect from historical data may facilitate policy-making. For example, taking rainfall data or other meteorological information for an area during the past few decades, the month this year in which the heavy rain is most likely to occur can be inferred. Therefore, we can work on an outside activity more reasonably or even make some protection to avoid flooding. Clinical data have to be well-preserved due to long-time research for pathology. Moreover, the symptoms determine the diagnosis from the doctor and the next treatment. As a consequence, the relevance between symptoms and the diagnosis has to be learnt and unnecessary physical examinations can be avoided.

For static big data, due to large scale or high dimension, the aim is to look for the inner pattern or construction of data, followed by extracting useful information which will be subject to further use, for example for prediction. Consequently, feature selection methods work like a pre-processor for finding valuable information from big data. Herein, we discuss methods from two aspects, large-scale data with a high dimension and data with a small sample but a high dimension.

##### *a: LARGE-SCALE DATA WITH A HIGH DIMENSION*

As we have discussed in SectionII-B.1, a series of measurements can be used to estimate the relevance between the features and class attributes. For large-scale data, mRMR(max-relevancy and min-redundancy) is an efficient tool that can search a set of features where the relevance between the feature and the class is maximized (max-relevancy) while the pairwise information between the features in the set is minimized (min-redundancy) [119]–[122]. This is one of the mutual information-based measures, which are developed to cope with computational complexity, since pairwise comparisons for calculating the correlations between features are conducted [34], [45], [93], [123], [124].

To further improve the performance of mRMR, Wang *et al.* [45] proposed an unsupervised feature selection method for dimensionality reduction.

*MPMR:* This measure provides a new criterion for unsupervised feature selection. The new criterion is called the maximal projection and minimal redundancy, which is formulated with the use of a projection matrix.

*$mr^2PSO$:* Unler *et al.* [123] presented a relevance and redundancy criterion based on mutual information. The basic idea of the proposed relevance and redundancy criterion is to maximize the prediction accuracy of the selected feature subset, which varies from the mRMR to determine the information property of a feature subset. That means the relevance and redundancy mutual information acts only as an intermediate measure in the PSO algorithm to improve the speed and performance of the search.

##### *b: SMALL-SAMPLE DATA WITH A HIGH DIMENSION*

For big data with a number of dimensions much bigger than that of data, the dimension becomes a major barrier to developing a predictive model with a high precision or improving the efficiency of a feature selection method. Viewing this, many scholars have attempted to design methods targeting at data with high dimension. He *et al.* [124] proposed a feature selection method based on mRMR, called MINT, which performs feature selection using both the training data and the unlabeled test data.

Apart from the mutual information-based measure, the distance-based measure also facilitates the process of feature selection with small-scale and high dimensional data. For example, Vijay *et al.* [125] presented an embedded technique by incorporating sparsity into a classifier, and Fang *et al.* [126] proposed an unsupervised feature selection method based on localities and similarities.

The aforementioned methods can effectively handle the problem of feature selection for high dimensional data. However, on one hand, they require information on all the features of all the data before selecting features, which is unpractical for big data. On the other hand, filter and embedded techniques take advantage of small computing consumption at the expense of a high degree of accuracy. How to improve accuracy with a small computing cost needs more attention in the future.

#### 2) DYNAMIC DATA

Online feature selection methods belong to the stream mode, which reevaluate the existing features based on the newly received datum. The key challenge of online feature selection is how to make accurate predictions for an instance using a small number of active features [127]. A number of research has been proposed to deal with dynamic data by means of feature selection [128]–[130].

*SFS:* Zhou [131] proposed the streamwise feature selection which only evaluates each feature once when it is generated. The benefits of streamwise methods are that features are generated dynamically and overfitting can be controlled by dynamically adjusting the threshold for adding features to the model. In their work, the candidate feature set is regarded as a dynamically generated stream, while knowledge on the structure of the feature space is required prior to heuristically controlling the choice of candidate features, which is often infeasible for real-world applications.

*OFS:* Wang *et al.* [127] assume that the dimension of data is fixed and the pattern of a datum can be achieved with the datum over time, where an online learner is employed to maintain a classifier involving only a small and fixed number of features.

*OSFS:* Conversely, with the number of training examples dynamically changing and more attention being paid to streaming features, Xindong *et al.* [30] described streaming features as features that flow in one by one over time. They studied a framework with small complexity cost, Fast-OSFS, to estimate the relevance of features and class attributes by calculating some probability values. An interesting point is that Fast-OSFS has its memory for redundant features due to the definition of the relationship based on conditional probability. This guarantees that even though a redundant feature has been removed earlier, a new feature with the same kind of redundant information as the former feature can also be discovered and eliminated.

*SAOLA:*In contrast to the estimation of mutual information by conditional probability, Yu *et al.* [34] defined the redundancy between features and the relevance between features and class attributes using entropy models. Additionally, once a feature is deleted, it will not be investigated any more by the greedy algorithm, which only adds new features but never deletes them. Therefore, their method can have a more rapid speed than Fast-OSFS.

For streaming data with a high dimension, great computing consumption may give rise to an incremental searching space, especially in an exponential way. In view of this, Fong *et al.* [132] have proposed a light-weight feature selection method based on heuristic algorithms.

*APSOFS:* APSOFS [132] finds a preferred combination of classification algorithms and the light-weight feature selection algorithms. An interesting aspect of their work is the discussion on how the new functions of data stream mining algorithms can help overcome the incremental computation.

An interesting issue is streaming labels, namely, the number of class attributes being unknown and the size of feature subset being constant.

*MLFS:* Like streaming features, Lin *et al.* [133] made an assumption that labels arrive one at a time. Under this scenario, they first obtain individual feature rank list weights based on mRMR for each newly arrived label. Afterwards, on the basis of the fixed weight values, the distance between the final feature rank list and each individual feature rank list is calculated, and the final feature rank list which makes the distance minimal is what is needed. This is a kind of embedded methods with a filter to rank and a learning method seeking the optimal feature subset. It provides a new idea for streaming-label feature selection problems, which attract many domains like image retrieval and medical diagnosis.

In summary, Table3 compares the methods discussed above briefly.

Similar to feature selection methods for large-scale data, in terms of online feature selection problems, filter and embedded techniques show a great potential due to their small computational cost, while wrapper methods are seldom employed. However, the small complexity of filter methods is often accompanied by a low degree of accuracy. Therefore, embedded methods as well as combined methods will become new trends for further research.

### 3) MISSING DATA

Missing data is very common in big data on account of software disasters or low resolution of hardware [134]–[136]. Bu *et al.* [136] has pointed out that the existence of missing data greatly increases greatly increases the difficulty in processing data. For traditional data, a simple approach to processing a dataset with missing data is to directly delete these missing data. It is clear that this approach can reduce the number of data, but some valuable information may also be ignored [93]. As mentioned above, even though some information is redundant or irrelevant, it is still retained in the original big data-set in its entirely for further analysis. Therefore, this approach goes against the intention of big data. Another simple way is to seek features with a high degree of relevance to those of the missing data, and to take the average value of the related features as the value of the missing data [137]. This kind of method assumes similarity measured by distance. Expectation maximization for missing data is established using a probabilistic model where iteration cannot be ignored when looking for a promising estimation. Similarly, low-rank approximation for missing data also has the demerit of repeated iterations [138]. Commonly, a parameter interpreted as probability is required, which is more difficult to determine for big data. In summary, repeated iterations are not suitable for big data mining, which may cause an incremental computation.

What can we do to deal with incomplete big data? Bu *et al.* [139] have attempted to employ feature selection for clustering incomplete big data. In their method, feature selection aims to filter undesirable features in a set of complete data measured by some entropy-based definitions, followed by a clustering model based on the selected feature subset. Yuan *et al.* [140] employed a feature selection method to multi-source learning of incomplete neural imaging data, which is a kind of large-scale data. The method utilizes

**TABLE 3. Comparison of commonly used feature selection methods for dynamic data.**

| Method | Data/Feature Sequence | Techniques | Search Strategy | Construction Style of the Feature Subset | Un/supervised | Limitation | Reference |
|---|---|---|---|---|---|---|---|
| SFS | Data | Filter | Alpha-investing | Adding relevant features | Supervised | Knowledge is required in advance and redundancy is never evaluated | [132] |
| OFS | Data | Wrapper | k-greedy search | Adding relevant features | Supervised | The dimension of data is fixed and redundancy is never evaluated | [128] |
| OSFS | Feature | Filter | Correlation with class attributes | Adding relevant features | Supervised | The size of the maximum conditioning subset has an influence on the performance | [30] |
| SAOLA | Feature | Filter | Correlation with the current feature subset and class attributes & k-greedy search | Adding relevant features & Removing redundant features | Supervised | The relevance threshold has an influence on the performance | [34] |
| APSOFS | Data | Wrapper | Accelerated particle swarm optimization | Regarding the best candidate subset as the final feature subset | Supervised | Large scale of dataset leads to an unsatisfied computational cost | [133] |
| MLFS | Data | Wrapper | Correlation between original features and the newly-arriving label | Regarding the best candidate subset as the final feature subset | Supervised | The size of features are fixed in advance | [134] |

missing blocks to partition a dataset into several independent learning tasks, with each having a classification model based on a feature selection method. These methods regard feature selection as a tool for reconstruction of the original data. If a feature has a redundancy with the others, or it is not relevant to the class attributes, tackling it at the expense of computing resources is unnecessary. When a feature without a recognized value is regarded as a redundant or irrelevant feature, it will then be eliminated.

Moreover, since feature selection is clearly desirable due to the abundance of missing features in many real-world applications, researchers have attempted to select a subset of features by rough sets although some features are missing, and to preserve the meaning of features contained in the data set to avoid information loss. Qian and Shu [141] proposed a feature selection method based on mutual information measured by rough sets for incomplete data, which takes into account a greedy forward search strategy from a whole set to accelerate the selection speed. The challenge of this kind of method is the use of the mutual information based on rough set theory for constructing data models.

Incomplete data is an interesting but formidable issue for data mining. Due to the efficiency of feature selection in seeking the relationship between features and features with class attributes, feature selection facilitates the reconstruction of a data model in line with the original data set with much missing information. Further data mining techniques can then be processed. However, the large scale or the high dimension of big data makes feature selection difficult, let alone big data in the dynamic environment. One challenge for available methods is the improvement of the consumption cost, which performs well in reconstructing a data model without repeated iterations. Moreover, how to apply efficient methods of big data in a static environment to dealing with big data in dynamic environment is still an open issue.

#### 4) HETEROGENEOUS DATA

In most practical problems, data are often collected from different sources. Their features are often heterogeneous and consist of numerical and non-numerical features with different properties [142]–[145]. For example, in clinical research [146] medical data are collected from different sources, such as demographics, disease history, medication, allergies, biomarkers, medical images, or genetic markers, each of which offers a different partial view of a patientąŕs condition [147]. As a result, it is difficult to evaluate heterogeneous features concurrently. As discussed previously, feature selection methods assign each feature a value of importance, and accordingly retain or get rid of a feature according to their inner measurement. Therefore, for heterogeneous data, data format differences contribute to the major obstacles for data mining, in particular in the field of big data.

The available feature selection methods for heterogeneous data can be roughly divided into numerical [148], [149] and non-numerical [150], [151] feature selection algorithms. Rough set [152], [153] and mutual information [145] are

two efficient tools for dealing with heterogeneous feature selection, while the convincing difference lies in methods based on the former being computationally extensive.

Under the circumstance of ever increasingly heterogeneous data, effective methods are in great demand in terms of size or formats. However, there are a very limited number of methods for heterogeneous data in the context of big data.

#### 5) UNRELIABLE DATA AND IMBALANCED DATA

Unreliable data are collected from equipment with deviations or with effects from the outside environment. Each feature has a reliability value resulting from the sensor precision, faulty equipment, environmental temperature changes, incorrect operation, etc., [90], [154], [155]. Gong *et al.* [89] propose a feature selection method for unreliable data where the reliability of a feature is represented by a value between 0 and 1, and the mathematic model is constructed using the reliability value. Commonly, when dealing with unreliable data, fuzzy methods have the capability of describing the degree of uncertainty for each variable. For example, Chen [156] employs a cost-based fuzzy decision model to deal with unreliable systems. Xie *et al.* [157] incorporate the designed fuzzy weighting function into their fuzzy control model under communication links. Since feature selection is a tool for mining data, if a datum is not totally reliant, or if there is a a breakdown or faulty operation when collecting the data, how can we make full use of it? In an industrial and mining context, there may be hundreds of sensors with their own individual accuracy in the underground production line, and the supervised data from these sensors is delivered in real-time to the upper detecting chamber. The transmitted data will surely have a significant influence on the judgment of the upper control. Consequently, designing useful feature selection methods for unreliable big data, in particular when data are transmitted in real-time for further processing, is of great importance.

In machine learning and data mining, when the number of observations in one class is significantly rarer than in other classes [158], the processing for the minor samples is difficult, which may contribute to misclassifying or even ignoring these minor samples. These minor samples, however, often contain more valuable information. Imbalanced data arise from the accumulated amount of data, in particular in the field of big data [159]. For instance in a detecting chamber of a pit, data collected from sensors consist of the major samples under no fault environment. When a breakdown occurs, for instance a leash deviates from its set track, an abnormal value from the sensor is then transmitted into the detecting chamber. These abnormal values consist of minor samples.

Since one or more classes are underrepresented in the data set [160], some researchers treat all imbalanced data consistently with a versatile algorithm [159], [161], [162] while others deal with imbalanced data with various dimensions, ratios and the number of classes [158].

For imbalanced data, several feature selection models have been attempted [158], [161], [163]–[168]. Embedded methods has been proven to be the most efficient tool for dealing with imbalanced data [169], [170], while before that, data level approaches with the aim of balancing all class attributes by re-sampling the training dataset are necessary. For example, over-sampling and under-sampling, where the former creates new samples in minor class, such as SMOTE [171], while the latter reduces the number of majority class samples, such as ACOS [172].

One obstacle to selecting feature subsets for imbalanced data is avoiding losing potentially useful information and altering the original data distribution since feature selection has to achieve a trade-off between eliminating irrelevance and redundancy and retaining valuable features. Under the environment of imbalanced big data, the inner pattern is hard to recognize and the computing cost should be lowered. Besides, it will have an extreme effect on the classifier's accuracy if potential features are removed. Therefore, the improvement of the performance in feature selection methods for imbalanced data is a major challenge for imbalanced mining.

### B. APPLICATIONS IN CLASSIFICATION ANALYSIS

#### 1) TEXT CLASSIFICATION

The basic motivation for studying big data is to aggregate and process a huge amount of data as rapidly as possible to mine valuable information [173]–[177]. Based on this desirable information, researchers can avoid risks, confirm reasons, and predict events [178]. One of the most important tasks of processing big data is to classify texts.

Today, the Internet is an essential part of people's lives, offering a great deal of convenience and reference sources. The open review platform on the web provides users with plenty of available contents. Many chaotic reviews, however, consume valuable decision-making time for users. Under this circumstance, feature selection is a promising way to provide a filter and reference. Wang *et al.* [179] proposed an effective feature selection method for classifying sentiment text in product reviews, where the experiment includes 1006 car reviews documents of cars. Although the method performs well in text classification, the attempts for big data where perhaps hundreds of or thousands of clues need to be analyzed. Assuming that there is a high degree of redundancy and irrelevance in these clues, one difficult task for decision makers is to choose goods with the preferred performance and a low expense.

If a dynamic environment exists in text classification, what form could a feature selection method take? Nakanishi Takafumi introduces a vector space model, which is the most popular and basic method for the comparison of concepts, events, and phenomena [180] to measure similarity or correlation between queries and information resources that are relevant to users' information. It is worth mentioning that the work of Takafumi can change space dynamically for semantic computations and analyses with the update of data. This is a breakthrough in text classification. In the big data environment, dynamic feature selection needs more attention.

#### 2) MICROARRAY ANALYSIS

Microarray data are generated from microarray experiments, which generally have high dimensions and a small number of samples. A key issue in microarray experiments is the large number of irrelevant and redundant genes. Their elimination should make the process of obtaining the classifier easier [181]–[184]. As mentioned above, feature selection methods for big data with high dimension and small samples has been discussed, and these are not suitable for a large amount of repeated iterations.

Under this circumstance, Apolloni *et al.* [184] developed an efficient hybrid feature selection method combining a wrapper based on binary differential evolution with a rank-based filter, where the initial population consists of solutions from the most relevant features obtained by the filter and solutions randomly generated to promote the diversity of solutions. Similarly, Tabakhi *et al.* [185] proposed an embedded feature selection method for genes, which is an unsupervised method that is different from the former method.

In view of the particularity in some sensitive cases, e.g. medical diagnosis, their research data are commonly very large on account of constant preservation [6]. Filter taxonomy is often taken into consideration in terms of the advantage of saving computation cost. However, in order to increase the desired classification accuracy, researchers tend to incorporate filters into other taxonomies or heuristic search strategies. Therefore, the accuracy of feature selection methods needs to be improved without increasing the complexity of these methods. Besides, the small number of samples creates a stepping stone for the improvement of the classification accuracy.

#### 3) IMAGE CLASSIFICATION AND BACKGROUND SUPPRESSION

In the domain of image classification, diverse information is required for images to be classified. Current attempts in terms of this are like [50] and [186]–[190]. On account of storage and computing costs, a large number of features does not represent better classification. Therefore, feature selection is taken into consideration for image classification. Shang and Barnes [191] proposed a fuzzy-rough feature selection method which is then incorporated into machine learning for Mars image classification. Hierarchical image content analysis is provided by Vavilin and Jo [192] for dealing with image classification and retrieval for natural and urban scenes. Moreover, Chang *et al.* [193] employed k-fold feature selection, based on the concept similar to k-fold cross-validation for image classification.

Although these studies are able to obtain a good classification result, the human factor has been ignored. In view of this, Zhou *et al.* [194] proposed an eye-guided tracking feature selection method for this field, which explores the mechanisms of the human eye for processing visual information based on mRMR and SVM. Their method takes a new look at image classification even though they do not consider dynamic images. In the context of big data, certain types of images tend to have an influence on eye tracking data. Diverse properties for images, such as the color, edge distribution, illumination, weather, season, daytime, saturation, buildings, cars, trees, the sky, and roads [192] make the image classification method specific to different scenarios. From this big domain of images, the way that we can extract valuable information and make use of it attract attention due to the widespread use of digital equipment.

Background suppression targets detecting and analyzing text from video frames, where a media sequence is assumed in an unknown video length [195], [196]. Feature selection methods are able to seek information, and Nguyen et al. employ feature selection to deal with the background suppression problem [197]. Similar to unreliable data, features for this kind of problem often have different importance, leading to more difficulty for feature selection. Moreover, their streamwise styles of data undoubtedly produce a greater challenge.

## V. FUTURE RESEARCH ISSUES

This paper has reviewed available feature selection methods for dealing with big data. Some possible future directions will be discussed in this section.

*Data model:* As mentioned above, we are now in the era of big data with extremely large sizes and rapid changes. A changing environment under the conditions of large-scale data should also be taken into consideration for feature selection. The dynamic feature selection method is an open issue for researchers, since not only is the cost of accessing the features or data high, but also we usually cannot obtain all the features or data in advance in real-world applications. With the growth of streaming data and the development of the dynamic feature selection methods, how to combine the two aspects for an efficient and fast classification requires a lot of work.

*High-dimension:* Although some available feature selection strategies based on mutual information for high-dimension data have the capability in computational consuming, the definition of the relationships between features or class attributes, which has an influence on the final selection results, is still a great challenge.

*Large-scale:* With respect to various formats of data, combined with a large scale, cloud computing and cooperative computing are new topics, while there is a handful of research incorporating these parallel processing methods into feature selection for big data.

*Data structure:* In terms of semi-structured data and non-structured data, the importance of normalization should be stressed. If feature selection methods that facilitate seeking the internal patterns of these kinds of big data are designed, then it will be easier for our interpreters to process and utilize them. However, most of the current research aims to find the feature subset

of structured big data, where the semi-structured and non-structured big data are absent.

*Dynamic environment:* Concerning the data whose format, scale or other characteristics are changing over time, namely, under the dynamic occasion, there have been a limited amount of research, even though several streamwise models are available. The processing speed is a main obstacle of these kinds of data, since in some cases, it is more necessary for our data users to obtain a brief and simple model, which can describe the main characteristics of the original model rather than a precise one. Using the obtained data model, future data can be processed roughly online, followed by other processing techniques offline.

*Combined with parallel methods:* For the characteristics of big data, parallel processing methods have been applied to improve the efficiency of mining big data [198]. Since feature selection methods can lighten the processing load in inducing a data mining model, the combination of the merits of both feature selection and parallel processing is worth investigating.

*a. Combined with CoEA:* The cooperative evolutionary algorithm (CoEA), a parallel evolutionary algorithm (EA), has generally a rapid processing speed. If the divide-and-conquer idea of CoEA to feature selection is applied to big data, the problem of feature selection with high dimensions is split into several subproblems of feature selection with low dimensions. Therefore, these subproblems with low dimensions can be easily handled in parallel by EAs, which provides a feasible approach for improving the efficiency of feature selection.

*b. Combined with cloud computing:* In recent years, cloud computing has been intensively studied. This lays the foundation for remote storage and distributed processing for big data. Heterogeneous parallel processing based on a cloud environment, however, can lead to many problems, such as the division of processing tasks and cooperation among cloud resources. Moreover, in the light of the remarkable performance of GPUs in float point arithmetic and large scale data processing, implementing feature selection methods on GPUs is also an effective way to improve the efficiency of big data mining.

*Energy saving:* Filter methods offer a lower computational cost than both wrapper and embedded methods, despite the fact that they perform at the expense of high accuracy. Therefore, hybrid methods, efficient embedded methods and parallel techniques are desirable to save computing cost while not lowering the accuracy when dealing with big data.

*Performance in real-time processing:* Since the usefulness of a datum will degrade over time, the time consumption for processing big data must be taken into account when designing a processing technique. Due to the merits of the dimensional reduction brought from feature selection, some researchers have attempted to apply feature selection methods to processing data in real time [199], [200]. For example, Zhang *et al.* [160] proposed a feature selection method based on the Fisher filter and wrapper to reduce the number of features so as to reduce the processing time. For dealing with problems with the requirement of real-time processing, apart from the efficiency of feature selection, just as much attention must be paid to the accuracy. The well-established data model will help further processing whereas an incorrect one contributes to a great impact on subsequent processing.

*Practical problems:* Since feature selection targets recognizing the inner pattern of a dataset and eliminating the irrelevance or the redundancy of the dataset, applying feature selection as a data mining tool to deal with practical problems is desirable in today's world of big data. However, there is a lack of research into practical cases on the basis of feature selection in the context of big data. It will be appreciated if the unreliability, the unbalance, and the allotropy of data are taken into consideration.

## VI. CONCLUSION

Mining valuable information from big data is indeed difficult and challenging. As an important data preprocessing technique, feature selection can greatly improve the efficiency of utilizing data. This paper first reviews feature selection methods for traditional data and then comments in detail on available feature selection methods for big data. On the one hand, although researchers have developed a large variety of available feature selection methods for traditional data, they still have difficulties tackling the problem of feature selection for big data. On the other hand, the existing methods of big data feature selection have severe limitations in achieving an appropriate tradeoff between the accuracy of solutions and computational complexity. Moreover, for practical problems, even though more work is essential, we have some strategies and techniques specific to a background, which are reviewed in this paper. Besides, more attention is paid to the applications of feature selection methods specific to several particular kinds of data and classification analysis. It will be appreciated if our review work provides a reference for those who would like to explore big data mining via feature selection.

## REFERENCES

[1] S. El-Sappagh, F. Ali, S. El-Masri, K. Kim, A. Ali, and K.-S. Kwak, "Mobile health technologies for diabetes mellitus: Current state and future challenges," *IEEE Access*, to be published.

[2] R. Elshawi, S. Sakr, D. Talia, and P. Trunfio, "Big data systems meet machine learning challenges: Towards big data science as a service," *Big Data Res.*, vol. 14, pp. 1–11, Dec. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2214579617303957

[3] V. Mayer-Schönberger and K. Cukier, *Big Data: A Revolution That Will Transform How We Live, Work and Think*. Boston, MA, USA: Houghton Mifflin, 2013.

[4] R. J. Hathaway and J. C. Bezdek, "Extending fuzzy and probabilistic clustering to very large data sets," *Comput. Statist. Data Anal.*, vol. 51, no. 1, pp. 215–234, 2006.

[5] C. Lynch, "Big data: How do your data grow?" *Nature*, vol. 455, no. 7209, pp. 28–29, 2008.

[6] K. Davis, *Ethics of Big Data: Balancing Risk and Innovation*. Newton, MA, USA: O'Reilly and Associates Inc, 2012.

[7] O. B. Sezer, E. Dogdu, and A. M. Ozbayoglu, "Context-aware computing, learning, and big data in Internet of Things: A survey," *IEEE Internet Things J.*, vol. 5, no. 1, pp. 1–27, Feb. 2018.

[8] C. K. Emani, N. Cullot, and C. Nicolle, "Understandable big data: A survey," *Comput. Sci. Rev.*, vol. 17, pp. 70–81, Aug. 2015.

[9] A. Nara, "Big data: Techniques and technologies in geoinformatics," *Int. J. Geograph. Inf. Sci.*, vol. 29, no. 4, pp. 694–696, Apr. 2015.

[10] J. Manyika *et al.*, *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. McKinsey Global Institute, 2011.

[11] Q. Tuo, H. Zhao, and Q. Hu, "Hierarchical feature selection with subtree based graph regularization," *Knowl.-Based Syst.*, vol. 163, pp. 996–1008, Jan. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950705118305094

[12] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "Recent advances and emerging challenges of feature selection in the context of big data," *Knowl.-Based Syst.*, vol. 86, pp. 33–45, Sep. 2015.

[13] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "A review of feature selection methods on synthetic data," *Knowl. Inf. Syst.*, vol. 34, no. 3, pp. 483–519, 2013.

[14] J. Wu and Z. Lu, "A novel hybrid genetic algorithm and simulated annealing for feature selection and kernel optimization in support vector regression," in *Proc. IEEE 5th Int. Conf. Adv. Comput. Intell. (ICACI)*, Oct. 2012, pp. 999–1003.

[15] N. Abd-Alsabour, "A review on evolutionary feature selection," in *Proc. Eur. Modelling Symp.*, 2014, pp. 20–26.

[16] A. A. Raweh, M. Nassef, and A. Badr, "A hybridized feature selection and extraction approach for enhancing cancer prediction based on DNA methylation," *IEEE Access*, vol. 6, pp. 15212–15223, 2018.

[17] A. Brankovic, M. Hosseini, and L. Piroddi, "A distributed feature selection algorithm based on distance correlation with an application to microarrays," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published.

[18] E. Bonilla-Huerta, A. Hernández-Montiel, R. Morales-Caporal, and M. Arjona-López, "Hybrid framework using multiple-filters and an embedded approach for an efficient selection and classification of microarray data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 13, no. 1, pp. 12–26, Jan./Feb. 2016. [Online]. Available: https://doi.org/10.1109/TCBB.2015.2474384

[19] M. Reboiro-Jato, F. Díaz, D. Glez-Peña, and F. Fdez-Riverola, "A novel ensemble of classifiers that use biological relevant gene sets for microarray classification," *Appl. Soft Comput.*, vol. 17, pp. 117–126, Apr. 2014.

[20] F. P. Shah and V. Patel, "A review on feature selection and feature extraction for text classification," in *Proc. Int. Conf. Wireless Commun., Signal Process. Netw. (WiSPNET)*, Mar. 2016, pp. 2264–2268.

[21] H. Wang, F. Dong, and L. Song, "Bubble-forming regime identification based on image textural features and the MCWA feature selection method," *IEEE Access*, vol. 5, pp. 15820–15830, 2017.

[22] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *J. Mach. Learn. Res.*, vol. 3, pp. 1289–1305, Mar. 2003.

[23] J. C. Gomez, E. Boiy, and M.-F. Moens, "Highly discriminative statistical features for email classification," *Knowl. Inf. Syst.*, vol. 31, no. 1, pp. 23–53, 2012.

[24] Z. Zhao, X. He, D. Cai, L. Zhang, W. Ng, and Y. Zhuang, "Graph regularized feature selection with data reconstruction," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 3, pp. 689–700, Mar. 2016.

[25] J. G. Dy, C. E. Brodley, A. Kak, L. S. Broderick, and A. M. Aisen, "Unsupervised feature selection applied to content-based retrieval of lung images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 3, pp. 373–378, Mar. 2003.

[26] P. Saari, T. Eerola, and O. Lartillot, "Generalizability and simplicity as criteria in feature selection: Application to mood classification in music," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 6, pp. 1802–1812, Aug. 2011.

[27] M. H. Kamarudin, C. Maple, T. Watson, and N. S. Safa, "A LogitBoost-based algorithm for detecting known and unknown Web attacks," *IEEE Access*, vol. 5, pp. 26190–26200, 2017.

[28] L. Zhao and X. Dong, "An industrial Internet of Things feature selection method based on potential entropy evaluation criteria," *IEEE Access*, vol. 6, pp. 4608–4617, 2018.

[29] R. Wald, T. M. Khoshgoftaar, A. Napolitano, and C. Sumner, "Using Twitter content to predict psychopathy," in *Proc. IEEE 11th Int. Conf. Mach. Learn. Appl. (ICMLA)*, vol. 2, Dec. 2012, pp. 394–401.

[30] X. Wu, K. Yu, W. Ding, H. Wang, and X. Zhu, "Online feature selection with streaming features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1178–1192, May 2013.

[31] W. Shu, W. Qian, and Y. Xie, "Incremental approaches for feature selection from dynamic data with the variation of multiple objects," *Knowl.-Based Syst.*, vol. 163, pp. 320–331, Jan. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950705118304246

[32] P. Bugata and P. Drotár, "Weighted nearest neighbors feature selection," *Knowl.-Based Syst.*, vol. 163, pp. 749–761, Jan. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950705118304908

[33] A. Lagrange, M. Fauvel, and M. Grizonnet, "Large-scale feature selection with Gaussian mixture models for the classification of high dimensional remote sensing images," *IEEE Trans. Comput. Imag.*, vol. 3, no. 2, pp. 230–242, Jun. 2017.

[34] K. Yu, X. Wu, W. Ding, and J. Pei, "Towards scalable and accurate online feature selection for big data," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Dec. 2014, pp. 660–669.

[35] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, nos. 1–2, pp. 273–324, 1997.

[36] H. Zhou, S. Han, and Y. Liu, "A novel feature selection approach based on document frequency of segmented term frequency," *IEEE Access*, vol. 6, pp. 53811–53821, 2018.

[37] W. Zhou, C. Wu, Y. Yi, and G. Luo, "Structure preserving non-negative feature self-representation for unsupervised feature selection," *IEEE Access*, vol. 5, pp. 8792–8803, 2017.

[38] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, no. 6, pp. 1157–1182, Jan. 2003.

[39] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Neural Comput. Appl.*, vol. 24, no. 1, pp. 175–186, 2014.

[40] X.-Y. Liu, Y. Liang, S. Wang, Z.-Y. Yang, and H.-S. Ye, "A hybrid genetic algorithm with wrapper-embedded approaches for feature selection," *IEEE Access*, vol. 6, pp. 22863–22874, 2018.

[41] F. Bagherzadeh-Khiabani, A. Ramezankhani, F. Azizi, F. Hadaegh, E. W. Steyerberg, and D. Khalili, "A tutorial on variable selection for clinical prediction models: Feature selection methods in data mining could improve the results," *J. Clin. Epidemiol.*, vol. 71, pp. 76–85, Mar. 2016.

[42] M. Dash and H. Liu, "Feature selection for classification," *Intell. Data Anal.*, vol. 1, nos. 1–4, pp. 131–156, 1997.

[43] R. Chen, N. Sun, X. Chen, M. Yang, and Q. Wu, "Supervised feature selection with a stratified feature weighting method," *IEEE Access*, vol. 6, pp. 15087–15098, 2018.

[44] S. Wang and W. Zhu, "Sparse graph embedding unsupervised feature selection," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 48, no. 3, pp. 329–341, Mar. 2018.

[45] S. Wang, W. Pedrycz, Q. Zhu, and W. Zhu, "Unsupervised feature selection via maximum projection and minimum redundancy," *Knowl.-Based Syst.*, vol. 75, pp. 19–29, Feb. 2015.

[46] 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, Mar. 2015.

[47] J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed, "Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 13, no. 5, pp. 971–989, Sep. 2016.

[48] J. Li and H. Liu, "Challenges of feature selection for big data analytics," *IEEE Intell. Syst.*, vol. 32, no. 2, pp. 9–15, Mar./Apr. 2017.

[49] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 4, pp. 491–502, Apr. 2005.

[50] A. Moghimi, C. Yang, and P. M. Marchetto, "Ensemble feature selection for plant phenotyping: A journey from hyperspectral to multispectral imaging," *IEEE Access*, vol. 6, pp. 56870–56884, 2018.

[51] C. Yao, Y.-F. Liu, B. Jiang, J. Han, and J. Han, "LLE score: A new filter-based unsupervised feature selection method based on nonlinear manifold embedding and its application to image recognition," *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5257–5269, Nov. 2017.

[52] I. Tsamardinos and C. F. Aliferis, "Towards principled feature selection: Relevancy, filters and wrappers," in *Proc. 9th Int. Workshop Artif. Intell. Statist.* San Mateo, CA, USA: Morgan Kaufmann, 2003.

[53] C. Freeman, D. Kulić, and O. Basir, "An evaluation of classifier-specific filter measure performance for feature selection," *Pattern Recognit.*, vol. 48, no. 5, pp. 1812–1826, 2015.

[54] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. Neural Netw.*, vol. 5, no. 4, pp. 537–550, Jul. 1994.

[55] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Proc. Int. Workshop Mach. Learn.*, 1992, pp. 249–256.

[56] H. Almuallim and T. G. Dietterich, "Learning Boolean concepts in the presence of many irrelevant features," *Artif. Intell.*, vol. 69, nos. 1–2, pp. 279–305, 1994.

[57] C.-T. Su and H.-C. Lin, "Applying electromagnetism-like mechanism for feature selection," *Inf. Sci.*, vol. 181, no. 5, pp. 972–986, 2011.

[58] P. Bermejo, L. de la Ossa, J. A. Gámez, and J. M. Puerta, "Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking," *Knowl.-Based Syst.*, vol. 25, no. 1, pp. 35–44, 2012.

[59] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artif. Intell.*, vol. 97, no. 1, pp. 245–271, Dec. 1997.

[60] L. Ma, M. Li, Y. Gao, T. Chen, X. Ma, and L. Qu, "A novel wrapper approach for feature selection in object-based image classification using polygon-based cross-validation," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 3, pp. 409–413, Mar. 2017.

[61] T. M. Khoshgoftaar, A. Fazelpour, H. Wang, and R. Wald, "A survey of stability analysis of feature subset selection techniques," in *Proc. IEEE 14th Int. Conf. Inf. Reuse Integr. (IRI)*, Aug. 2013, pp. 424–431.

[62] A. A. Shanab, T. M. Khoshgoftaar, and R. Wald, "Evaluation of wrapper-based feature selection using hard, moderate, and easy bioinformatics data," in *Proc. IEEE Int. Conf. Bioinform. Bioeng. (BIBE)*, Nov. 2014, pp. 149–155.

[63] L.-Y. Qiao, X.-Y. Peng, and Y. Peng, "BPSO-SVM wrapper for feature subset selection," *Dianzi Xuebao (Acta Electronica Sinica)*, vol. 34, no. 3, pp. 496–498, 2006.

[64] H.-H. Hsu, C.-W. Hsieh, and M.-D. Lu, "Hybrid feature selection by combining filters and wrappers," *Expert Syst. Appl.*, vol. 38, no. 7, pp. 8144–8150, 2011.

[65] C. Hou, F. Nie, X. Li, D. Yi, and Y. Wu, "Joint embedding learning and sparse regression: A framework for unsupervised feature selection," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 793–804, Jun. 2014.

[66] K. Mistry, L. Zhang, S. C. Neoh, C. P. Lim, and B. Fielding, "A micro-GA embedded PSO feature selection approach to intelligent facial emotion recognition," *IEEE Trans. Cybern.*, vol. 47, no. 6, pp. 1496–1509, Jun. 2017.

[67] J. Xie and W.-X. Xie, "Several feature selection algorithms based on the discernibility of a feature subset and support vector machines," *Chin. J. Comput.*, vol. 37, pp. 1704–1718, Aug. 2014.

[68] T. Kari *et al.*, "Hybrid feature selection approach for power transformer fault diagnosis based on support vector machine and genetic algorithm," *IET Gener., Transmiss. Distrib.*, vol. 12, no. 21, pp. 5672–5680, 2018.

[69] J. Spilka, J. Frecon, R. Leonarduzzi, N. Pustelnik, P. Abry, and M. Doret, "Sparse support vector machine for intrapartum fetal heart rate classification," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 3, pp. 664–671, May 2017.

[70] A. Rakotomamonjy, "Variable selection using SVM based criteria," *J. Mach. Learn. Res.*, vol. 3, pp. 1357–1370, Mar. 2003.

[71] R. Chakraborty and N. R. Pal, "Feature selection using a neural framework with controlled redundancy," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 1, pp. 35–50, Jan. 2015.

[72] E. Romero and J. M. Sopena, "Performing feature selection with multilayer perceptrons," *IEEE Trans. Neural Netw.*, vol. 19, no. 3, pp. 431–441, Mar. 2008.

[73] B. Lerner, M. Levinstein, B. Rosenberg, H. Guterman, I. Dinstein, and Y. Romem, "Feature selection and chromosome classification using a multilayer perceptron neural network," in *Proc. IEEE Int. Conf. Neural Netw., IEEE World Congr. Comput. Intell.*, vol. 6, Jun. 1994, pp. 3540–3545.

[74] R. Alzubi, N. Ramzan, H. Alzoubi, and A. Amira, "A hybrid feature selection method for complex diseases SNPs," *IEEE Access*, vol. 6, pp. 1292–1301, 2018.

[75] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, "Feature selection for SVMS," in *Proc. NIPS*, vol. 12, 2000, pp. 668–674.

[76] Y. Zhang, C. Ding, and T. Li, "Gene selection algorithm by combining reliefF and mRMR," *BMC Genomics*, vol. 9, no. Suppl 2, p. S27, 2008.

[77] Y. Peng, Z. Wu, and J. Jiang, "A novel feature selection approach for biomedical data classification," *J. Biomed. Inform.*, vol. 43, no. 1, pp. 15–23, 2010.

[78] A. E. Akadi, A. Amine, A. E. Ouardighi, and D. Aboutajdine, "A two-stage gene selection scheme utilizing MRMR filter and GA wrapper," *Knowl. Inf. Syst.*, vol. 26, no. 3, pp. 487–500, 2011.

[79] Y. Luo, Y. Li, C. Zhou, and C. Xu, "Combining feature selectors in a product advertisement classification system," in *Proc. IEEE 1st Asian Conf. Pattern Recognit. (ACPR)*, Nov. 2011, pp. 184–188.

[80] L. Hedjazi, J. Aguilar-Martin, M.-V. Le Lann, and T. Kempowsky-Hamon, "Membership-margin based feature selection for mixed type and high-dimensional data: Theory and applications," *Inf. Sci.*, vol. 322, pp. 174–196, Nov. 2015.

[81] K. Nag and N. R. Pal, "A multiobjective genetic programming-based ensemble for simultaneous feature selection and classification," *IEEE Trans. Cybern.*, vol. 46, no. 2, pp. 499–510, Feb. 2016.

[82] F. Alzami *et al.*, "Adaptive hybrid feature selection-based classifier ensemble for epileptic seizure classification," *IEEE Access*, vol. 6, pp. 29132–29145, 2018.

[83] S. Nagi and D. K. Bhattacharyya, "Classification of microarray cancer data using ensemble approach," *Netw. Model. Anal. Health Inform. Bioinform.*, vol. 2, no. 3, pp. 159–173, 2013.

[84] O. P. Günther *et al.*, "A computational pipeline for the development of multi-marker bio-signature panels and ensemble classifiers," *BMC Bioinform.*, vol. 13, no. 1, p. 326, 2012.

[85] H. Wang, T. M. Khoshgoftaar, and A. Napolitano, "A comparative study of ensemble feature selection techniques for software defect prediction," in *Proc. IEEE 9th Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2010, pp. 135–140.

[86] R. Xia, C. Zong, X. Hu, and E. Cambria, "Feature ensemble plus sample selection: Domain adaptation for sentiment classification," *IEEE Intell. Syst.*, vol. 28, no. 3, pp. 10–18, May 2013.

[87] N. Gu, M. Fan, L. Du, and D. Ren, "Efficient sequential feature selection based on adaptive eigenspace model," *Neurocomputing*, vol. 161, pp. 199–209, Aug. 2015.

[88] 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, Jan. 2015.

[89] D. Gong, Y. Hu, and Y. Zhang, "Feature selection method for unreliable data based on particle swarm optimization," *Acta Electronica Sinica*, vol. 7, no. 7, pp. 1320–1326, 2014.

[90] Z. Yong, G. Dun-Wei, and Z. Wan-Qiu, "Feature selection of unreliable data using an improved multi-objective PSO algorithm," *Neurocomputing*, vol. 171, pp. 1281–1290, Jan. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231215010632

[91] J. DePasquale and R. Polikar, "Random feature subset selection for analysis of data with missing features," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IJCNN)*, Aug. 2007, pp. 2379–2384.

[92] A. Aussem and S. R. de Morais, "A conservative feature subset selection algorithm with missing data," *Neurocomputing*, vol. 73, nos. 4–6, pp. 585–590, 2010.

[93] G. Doquire and M. Verleysen, "Feature selection with missing data using mutual information estimators," *Neurocomputing*, vol. 90, pp. 3–11, Aug. 2012.

[94] M. Ramoni and P. Sebastiani, "Robust learning with missing data," *Mach. Learn.*, vol. 45, no. 2, pp. 147–170, 2001.

[95] W. Zong, F. Wu, L.-K. Chu, and D. Sculli, "A discriminative and semantic feature selection method for text categorization," *Int. J. Prod. Econ.*, vol. 165, pp. 215–222, Jul. 2015.

[96] Y. Zhang, D.-W. Gong, and M. Rong, "Multi-objective differential evolution algorithm for multi-label feature selection in classification," in *Advances in Swarm and Computational Intelligence*. Beijing, China: Springer, 2015, pp. 339–345.

[97] S. Peng, Q. Xu, X. B. Ling, X. Peng, W. Du, and L. Chen, "Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines," *FEBS Lett.*, vol. 555, no. 2, pp. 358–362, 2003.

[98] C. H. Ooi and P. Tan, "Genetic algorithms applied to multi-class prediction for the analysis of gene expression data," *Bioinformatics*, vol. 19, no. 1, pp. 37–44, 2003.

[99] L. Yu, L. Hu, and L. Tang, "Stock selection with a novel sigmoid-based mixed discrete-continuous differential evolution algorithm," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 7, pp. 1891–1904, Jul. 2016.

[100] A. Gosh, S. Das, R. Mallipeddi, A. K. Das, and S. S. Dash, "A modified differential evolution with distance-based selection for continuous optimization in presence of noise," *IEEE Access*, vol. 5, pp. 26944–26964, 2017.

[101] Y. Zhang, D.-W. Gong, and J. Cheng, "Multi-objective particle swarm optimization approach for cost-based feature selection in classification," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 1, pp. 64–75, Jan./Feb. 2017. [Online]. Available: https://doi.org/10.1109/TCBB.2015.2476796

[102] A. A. Naeini, M. Babadi, S. M. J. Mirzadeh, and S. Amini, "Particle swarm optimization for object-based feature selection of VHSR satellite images," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 3, pp. 379–383, Mar. 2018.

[103] B. Tran, B. Xue, and M. Zhang, "A new representation in PSO for discretization-based feature selection," *IEEE Trans. Cybern.*, vol. 48, no. 6, pp. 1733–1746, Jun. 2018.

[104] C. Yan, J. Ma, H. Luo, and J. Wang, "A hybrid algorithm based on binary chemical reaction optimization and tabu search for feature selection of high-dimensional biomedical data," *Tsinghua Sci. Technol.*, vol. 23, no. 6, pp. 733–743, Dec. 2018.

[105] M. A. Tahir and A. Bouridane, "Novel Round-Robin Tabu search algorithm for prostate cancer classification and diagnosis using multispectral imagery," *IEEE Trans. Inf. Technol. Biomed.*, vol. 10, no. 4, pp. 782–793, Oct. 2006.

[106] S. A. Toussi, H. S. Yazdi, E. Hajinezhad, and S. Effati, "Eigenvector selection in spectral clustering using Tabu search," in *Proc. IEEE 1st Int. eConf. Comput. Knowl. Eng. (ICCKE)*, Oct. 2011, pp. 75–80.

[107] I.-S. Oh, J.-S. Lee, and B.-R. Moon, "Hybrid genetic algorithms for feature selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 11, pp. 1424–1437, Nov. 2004.

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

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

[110] E.-G. Talbi, L. Jourdan, J. Garcia-Nieto, and E. Alba, "Comparison of population based metaheuristics for feature selection: Application to microarray data classification," in *Proc. ACS/IEEE Int. Conf. Comput. Syst. Appl.*, Mar./Apr. 2008, pp. 45–52.

[111] L.-Y. Chuang, C.-H. Yang, and J.-C. Li, "Chaotic maps based on binary particle swarm optimization for feature selection," *Appl. Soft Comput.*, vol. 11, no. 1, pp. 239–248, 2011.

[112] A. Unler and A. Murat, "A discrete particle swarm optimization method for feature selection in binary classification problems," *Eur. J. Oper. Res.*, vol. 206, no. 3, pp. 528–539, 2010.

[113] L.-Y. Chuang, C.-S. Yang, K.-C. Wu, and C.-H. Yang, "Gene selection and classification using Taguchi chaotic binary particle swarm optimization," *Expert Syst. Appl.*, vol. 38, no. 10, pp. 13367–13377, 2011.

[114] U. K. Sikdar, A. Ekbal, and S. Saha, "Differential evolution based mention detection for anaphora resolution," in *Proc. Annu. IEEE India Conf. (INDICON)*, Dec. 2013, pp. 1–6.

[115] A. Al-Ani, A. Alsukker, and R. N. Khushaba, "Feature subset selection using differential evolution and a wheel based search strategy," *Swarm Evol. Comput.*, vol. 9, pp. 15–26, Apr. 2013.

[116] Y. Zhang, M. Rong, and D. Gong, "A multi-objective feature selection based on differential evolution," in *Proc. IEEE Int. Conf. Control, Autom. Inf. Sci. (ICCAIS)*, Oct. 2015, pp. 302–306.

[117] C.-L. Huang, "ACO-based hybrid classification system with feature subset selection and model parameters optimization," *Neurocomputing*, vol. 73, nos. 1–3, pp. 438–448, 2009.

[118] S.-W. Lin, Z.-J. Lee, S.-C. Chen, and T.-Y. Tseng, "Parameter determination of support vector machine and feature selection using simulated annealing approach," *Appl. Soft Comput.*, vol. 8, no. 4, pp. 1505–1512, 2008.

[119] Y. Zhang, J. Wu, and J. Cai, "Compact representation of high-dimensional feature vectors for large-scale image recognition and retrieval," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2407–2419, May 2016.

[120] F. Gao, X. Zhang, Y. Huang, Y. Luo, X. Li, and L.-Y. Duan, "Data-driven lightweight interest point selection for large-scale visual search," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2774–2787, Oct. 2018.

[121] J. M. N. Abad and A. Soleimani, "Novel feature selection algorithm for thermal prediction model," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 26, no. 10, pp. 1831–1844, Oct. 2018.

[122] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[123] A. Unler, A. Murat, and R. B. Chinnam, "$mr^2PSO$: A maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification," *Inf. Sci.*, vol. 181, no. 20, pp. 4625–4641, 2011.

[124] D. He, I. Rish, D. Haws, S. Teyssedre, Z. Karaman, and L. Parida. (2013). "MINT: Mutual information based transductive feature selection for genetic trait prediction." [Online]. Available: https://arxiv.org/abs/1310.1659

[125] V. Pappu, O. P. Panagopoulos, P. Xanthopoulos, and P. M. Pardalos, "Sparse proximal support vector machines for feature selection in high dimensional datasets," *Expert Syst. Appl.*, vol. 42, pp. 9183–9191, 2015.

[126] X. Fang, Y. Xu, X. Li, Z. Fan, H. Liu, and Y. Chen, "Locality and similarity preserving embedding for feature selection," *Neurocomputing*, vol. 128, no. 5, pp. 304–315, Mar. 2014.

[127] J. Wang, P. Zhao, S. C. H. Hoi, and R. Jin, "Online feature selection and its applications," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 3, pp. 698–710, Mar. 2014.

[128] S. S. Naqvi, W. N. Browne, and C. Hollitt, "Feature quality-based dynamic feature selection for improving salient object detection," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 4298–4313, Sep. 2016.

[129] C. Tong and X. Shi, "Decentralized monitoring of dynamic processes based on dynamic feature selection and informative fault pattern dissimilarity," *IEEE Trans. Ind. Electron.*, vol. 63, no. 6, pp. 3804–3814, Jun. 2016.

[130] M. Pratama, W. Pedrycz, and E. Lughofer, "Evolving ensemble fuzzy classifier," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 5, pp. 2552–2567, Oct. 2018.

[131] J. Zhou, D. P. Foster, R. A. Stine, and L. H. Ungar, "Streamwise feature selection," *J. Mach. Learn. Res.*, vol. 7, no. 1, pp. 1861–1885, Dec. 2006.

[132] S. Fong, R. Wong, and A. V. Vasilakos, "Accelerated PSO swarm search feature selection for data stream mining big data," *IEEE Trans. Serv. Comput.*, vol. 9, no. 1, pp. 33–45, Jan. 2016.

[133] Y. Lin, Q. Hu, J. Zhang, and X. Wu, "Multi-label feature selection with streaming labels," *Inf. Sci.*, vol. 372, pp. 256–275, Dec. 2016.

[134] S. Jin, F. Ye, Z. Zhang, K. Chakrabarty, and X. Gu, "Efficient board-level functional fault diagnosis with missing syndromes," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 35, no. 6, pp. 985–998, Jun. 2016.

[135] W. Xu, Z. He, E. Lo, and C.-Y. Chow, "Explaining missing answers to top-k SQL queries," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 8, pp. 2071–2085, Aug. 2016.

[136] F. Bu, Z. Chen, Q. Zhang, and X. Wang, "Incomplete big data clustering algorithm using feature selection and partial distance," in *Proc. 5th Int. Conf. Digit. Home (ICDH)*, 2014, pp. 263–266.

[137] T. Hastie, R. Tibshirani, G. Sherlock, M. Eisen, P. Brown, and D. Botstein, "Imputing missing data for gene expression arrays," Division Biostatist., Stanford Univ., Stanford, CA, USA, Tech. Rep., 1999.

[138] P. G. Clark, J. W. Grzymala-Busse, and W. Rzasa, "Mining incomplete data with singleton, subset and concept probabilistic approximations," *Inf. Sci.*, vol. 280, pp. 368–384, Oct. 2014.

[139] F. Bu, Z. Chen, Q. Zhang, and L. T. Yang, "Incomplete high-dimensional data imputation algorithm using feature selection and clustering analysis on cloud," *J. Supercomput.*, vol. 72, no. 8, pp. 2977–2990, 2015.

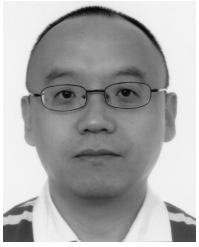
[140] L. Yuan *et al.*, "Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data," *NeuroImage*, vol. 61, no. 3, pp. 622–632, 2012.

[141] W. Qian and W. Shu, "Mutual information criterion for feature selection from incomplete data," *Neurocomputing*, vol. 168, pp. 210–220, Nov. 2015.
[142] D. Song, Y. Luo, and J. Heflin, "Linking heterogeneous data in the semantic Web using scalable and domain-independent candidate selection," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 1, pp. 143–156, Jan. 2017.
[143] S. Huda, J. Yearwood, H. F. Jelinek, M. M. Hassan, G. Fortino, and M. Buckland, "A hybrid feature selection with ensemble classification for imbalanced healthcare data: A case study for brain tumor diagnosis," *IEEE Access*, vol. 4, pp. 9145–9154, 2016.
[144] M. A. Hossain, X. Jia, and J. A. Benediktsson, "One-class oriented feature selection and classification of heterogeneous remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Ses.*, vol. 9, pp. 1606–1612, Apr. 2016.
[145] M. Wei, T. W. S. Chow, and R. H. M. Chan, "Heterogeneous feature subset selection using mutual information-based feature transformation," *Neurocomputing*, vol. 168, pp. 706–718, Nov. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S092523121500733X
[146] Y. Motai, N. A. Siddique, and H. Yoshida, "Heterogeneous data analysis: Online learning for medical-image-based diagnosis," *Pattern Recognit.*, vol. 63, pp. 612–624, Mar. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320316302977
[147] S. Pölsterl, S. Conjeti, N. Navab, and A. Katouzian "Survival analysis for high-dimensional, heterogeneous medical data: Exploring feature extraction as an alternative to feature selection," *Artif. Intell. Med.*, vol. 72, pp. 1–11, Sep. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0933365716300653
[148] T. W. S. Chow and D. Huang, "Estimating optimal feature subsets using efficient estimation of high-dimensional mutual information," *IEEE Trans. Neural Netw.*, vol. 16, no. 1, pp. 213–224, Jan. 2005.
[149] J. Neumann, C. Schnörr, and G. Steidl, "Combined SVM-based feature selection and classification," *Mach. Learn.*, vol. 61, nos. 1–3, pp. 129–150, 2005.
[150] T. W. S. Chow, P. Wang, and E. W. M. Ma, "A new feature selection scheme using a data distribution factor for unsupervised nominal data," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 38, no. 2, pp. 499–509, Apr. 2008.
[151] N. Zhong, J. Dong, and S. Ohsuga, "Using rough sets with heuristics for feature selection," *J. Intell. Inf. Syst.*, vol. 16, no. 3, pp. 199–214, 2001.
[152] Z. Pawlak and A. Skowron, "Rough sets: Some extensions," *Inf. Sci.*, vol. 177, no. 1, pp. 28–40, 2007.
[153] R. Slowinski and D. Vanderpooten, "A generalized definition of rough approximations based on similarity," *IEEE Trans. Knowl. Data Eng.*, vol. 12, no. 2, pp. 331–336, Mar. 2000.
[154] M.-G. Park and K.-J. Yoon, "Optimal key-frame selection for video-based structure-from-motion," *Electron. Lett.*, vol. 47, no. 25, pp. 1367–1369, Dec. 2011.
[155] R. M. Balabin and S. V. Smirnov, "Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data," *Analytica Chim. Acta*, vol. 692, nos. 1–2, pp. 63–72, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0003267011003539
[156] S.-P. Chen, "Time value of delays in unreliable production systems with mixed uncertainties of fuzziness and randomness," *Eur. J. Oper. Res.*, vol. 255, no. 3, pp. 834–844, 2016.
[157] X.-P. Xie, D. Yue, and S.-L. Hu, "Fuzzy control design of nonlinear systems under unreliable communication links: A systematic homogenous polynomial approach," *Inf. Sci.*, vols. 370–371, pp. 763–771, Nov. 2016.
[158] L. Yijing, G. Haixiang, L. Xiao, L. Yanan, and L. Jinling, "Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data," *Knowl.-Based Syst.*, vol. 94, pp. 88–104, Feb. 2015.
[159] J. Fan, Z. Niu, Y. Liang, and Z. Zhao, "Probability model selection and parameter evolutionary estimation for clustering imbalanced data without sampling," *Neurocomputing*, vol. 211, pp. 172–181, Oct. 2016.
[160] Z. Zhang, H. Chen, Y. Xu, J. Zhong, N. Lv, and S. Chen, "Multisensor-based real-time quality monitoring by means of feature extraction, selection and modeling for al alloy in arc welding," *Mech. Syst. Signal Process.*, vols. 60–61, pp. 151–165, Aug. 2015.
[161] S. Vluymans, D. S. Tarragó, Y. Saeys, C. Cornelis, and F. Herrera, "Fuzzy rough classifiers for class imbalanced multi-instance data," *Pattern Recognit.*, vol. 53, pp. 36–45, May 2016.
[162] G. Haixiang, L. Yijing, L. Yanan, L. Jinlinga, and L. Xiao, "BPSO-AdaBoost-KNN ensemble learning algorithm for multi-class imbalanced data classification," *Eng. Appl. Artif. Intell.*, vol. 49, pp. 176–193, Mar. 2016.
[163] W. C. Yeh, Y. T. Yang, and C. M. Lai, "A hybrid simplified swarm optimization method for imbalanced data feature selection," *Austral. Acad. Bus. Econ. Rev.*, vol. 2, no. 3, pp. 263–275, 2017.
[164] W. D. Fisher, T. K. Camp, and V. V. Krzhizhanovskaya, "Anomaly detection in earth dam and levee passive seismic data using support vector machines and automatic feature selection," *J. Comput. Sci.*, vol. 20, pp. 143–153, May 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877750316304185
[165] L. Yin, Y. Ge, K. Xiao, X. Wang, and X. Quan, "Feature selection for high-dimensional imbalanced data," *Neurocomputing*, vol. 105, no. 3, pp. 3–11, 2013.
[166] M. Alibeigi, S. Hashemi, and A. Hamzeh, "DBFS: An effective density based feature selection scheme for small sample size and high dimensional imbalanced data sets," *Data Knowl. Eng.*, vols. 81–82, no. 4, pp. 67–103, 2012.
[167] U. Mahdiyah, M. I. Irawan, and E. M. Imah, "Integrating data selection and extreme learning machine for imbalanced data," *Procedia Comput. Sci.*, vol. 59, pp. 221–229, Aug. 2015.
[168] J. A. Sáez, B. Krawczyk, and M. Woźniak, "Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets," *Pattern Recognit.*, vol. 57, pp. 164–178, Sep. 2016.
[169] J. F. Díez-Pastor, J. J. Rodríguez, C. García-Osorio, and L. I. Kuncheva, "Random balance: Ensembles of variable priors classifiers for imbalanced data," *Knowl.-Based Syst.*, vol. 85, pp. 96–111, Sep. 2015.
[170] Z. Sun, Q. Song, X. Zhu, H. Sun, B. Xu, and Y. Zhou, "A novel ensemble method for classifying imbalanced data," *Pattern Recognit.*, vol. 48, no. 5, pp. 1623–1637, 2015.
[171] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.
[172] H. Yu, J. Ni, and J. Zhao, "ACOSampling: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data," *Neurocomputing*, vol. 101, no. 3, pp. 309–318, 2013.
[173] A. K. Uysal, "On two-stage feature selection methods for text classification," *IEEE Access*, vol. 6, pp. 43233–43251, 2018.
[174] A. Onan and S. Korukoğlu, "A feature selection model based on genetic rank aggregation for text sentiment classification," *J. Inf. Sci.*, vol. 43, no. 1, pp. 25–38, 2017.
[175] D. Agnihotri, K. Verma, and P. Tripathi, "Variable global feature selection scheme for automatic classification of text documents," *Expert Syst. Appl.*, vol. 81, pp. 268–281, Sep. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417417302208
[176] M. Labani, P. Moradi, F. Ahmadizar, and M. Jalili, "A novel multivariate filter method for feature selection in text classification problems," *Eng. Appl. Artif. Intell.*, vol. 70, pp. 25–37, Apr. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0952197617303172
[177] B. Parlak and A. K. Uysal, "On feature weighting and selection for medical document classification," in *Developments and Advances in Intelligent Systems and Applications*, Á. Rocha and L. Reis, Eds. Cham, Switzerland: Springer, 2018, pp. 269–282.
[178] T. Nakanishi, "A feature selection method for comparision of each concept in big data," in *Proc. IEEE/ACIS 14th Int. Conf. Comput. Inf. Sci. (ICIS)*, Jun./Jul. 2015, pp. 229–234.
[179] S. Wang, D. Li, Y. Wei, and H. Li, "A feature selection method based on Fisher's discriminant ratio for text sentiment classification," in *Web Information Systems and Mining*. 2009, pp. 88–97.
[180] K. Kesorn and S. Poslad, "An enhanced bag-of-visual word vector space model to represent visual content in athletics images," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 211–222, Feb. 2012.
[181] I. Jain, V. K. Jain, and R. Jain, "Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification," *Appl. Soft Comput.*, vol. 62, pp. 203–215, Jan. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S156849461730577X
[182] Y. Li, T. Li, and H. Liu, "Recent advances in feature selection and its applications," *Knowl. Inf. Syst.*, vol. 53, no. 3, pp. 551–577, Dec. 2017. [Online]. Available: https://doi.org/10.1007/s10115-017-1059-8
[183] S. Turgut, M. Dağtekin, and T. Ensari, "Microarray breast cancer data classification using machine learning methods," in *Proc. Electr. Electron., Comput. Sci., Biomed. Eng. Meeting (EBBT)*, Apr. 2018, pp. 1–3.

[184] J. Apolloni, G. Leguizamón, and E. Alba, "Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments," *Appl. Soft Comput.*, vol. 38, pp. 922–932, Jan. 2016.
[185] S. Tabakhi, A. Najafi, R. Ranjbar, and P. Moradi, "Gene selection for microarray data classification using a novel ant colony optimization," *Neurocomputing*, vol. 168, pp. 1024–1036, Nov. 2015.
[186] G. Taşkın, H. Kaya, and L. Bruzzone, "Feature selection based on high dimensional model representation for hyperspectral images," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2918–2928, Jun. 2017.
[187] H. Lang, J. Zhang, X. Zhang, and J. Meng, "Ship classification in SAR image by joint feature and classifier selection," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 2, pp. 212–216, Feb. 2016.
[188] P. Bolourchi, H. Demirel, and S. Uysal, "Entropy-score-based feature selection for moment-based SAR image classification," *Electron. Lett.*, vol. 54, no. 9, pp. 593–595, May 2018. [Online]. Available: https://digital-library.theiet.org/content/journals/10.1049/el.2017.4419
[189] I. Garali, M. Adel, S. Bourennane, and E. Guedj, "Histogram-based features selection and volume of interest ranking for brain PET image classification," *IEEE J. Transl. Eng. Health Med.*, vol. 6, 2018, Art. no. 2100212.
[190] L. Zhang, Q. Zhang, B. Du, X. Huang, Y. Y. Tang, and D. Tao, "Simultaneous spectral-spatial feature selection and extraction for hyperspectral images," *IEEE Trans. Cybern.*, vol. 48, no. 1, pp. 16–28, Jan. 2018.
[191] C. Shang and D. Barnes, "Fuzzy-rough feature selection aided support vector machines for Mars image classification," *Comput. Vis. Image Understand.*, vol. 117, no. 3, pp. 202–213, 2013.
[192] A. Vavilin and K.-H. Jo, "Automatic context analysis for image classification and retrieval based on optimal feature subset selection," *Neurocomputing*, vol. 116, no. 10, pp. 201–207, 2013.
[193] C.-Y. Chang, S.-J. Chen, and M.-F. Tsai, "Application of support-vector-machine-based method for feature selection and classification of thyroid nodules in ultrasound images," *Pattern Recognit.*, vol. 43, no. 10, pp. 3494–3506, Oct. 2010.
[194] X. Zhou, X. Gao, J. Wang, H. Yu, Z. Wang, and Z. Chi, "Eye tracking data guided feature selection for image classification," *Pattern Recognit.*, vol. 63, pp. 56–70, Mar. 2017.
[195] Z. Shi, Z. Zou, and C. Zhang, "Real-time traffic light detection with adaptive background suppression filter," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 3, pp. 690–700, Mar. 2016.
[196] C.-C. Shen and J.-H. Yan, "High-order Hadamard-encoded transmission for tissue background suppression in ultrasound contrast imaging: Memory effect and decoding schemes," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 66, no. 1, pp. 26–37, Jan. 2019.
[197] T. M. Nguyen, Q. M. J. Wu, and D. Mukherjee, "An online unsupervised feature selection and its application for background suppression," in *Proc. IEEE 12th Conf. Comput. Robot Vis. (CRV)*, Jun. 2015, pp. 161–168.
[198] Y. Zhang *et al.*, "Parallel processing systems for big data: A survey," *Proc. IEEE*, vol. 104, no. 11, pp. 2114–2136, Nov. 2016.
[199] R. Trichet and F. Bremond, "Dataset optimization for real-time pedestrian detection," *IEEE Access*, vol. 6, pp. 7719–7727, 2018.
[200] S. Lekha and M. Suchetha, "A novel 1-D convolution neural network with SVM architecture for real-time detection applications," *IEEE Sensors J.*, vol. 18, no. 2, pp. 724–731, Jan. 2018.

**MIAO RONG** received the B.S. degree in electrical engineering and automation from the China University of Mining and Technology, Xuzhou, China, in 2014, where she is currently pursuing the Ph.D. degree in control theory and control engineering. Her main research interests include data mining and multiobjective optimization.

**DUNWEI GONG** received the B.S. degree in applied mathematics from the China University of Mining and Technology, Xuzhou, China, in 1992, the M.S. degree in control theory and its applications from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 1995, and the Ph.D. degree in control theory and control engineering from the China University of Mining and Technology, in 1999. He is currently a Professor and the Ph.D. Advisor of the School of Information and Electrical Engineering, China University of Mining and Technology. He is also with the School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao, China. His main research interests include evolutionary computation, intelligence optimization, and data mining.



**XIAOZHI GAO** received the B.Sc. and M.Sc. degrees from the Harbin Institute of Technology, China, in 1993 and 1996, respectively, and the D.Sc. (Tech.) degree from the Helsinki University of Technology (now Aalto University), Finland, in 1999. He was with the Helsinki University of Technology (now Aalto University) from 2004 to 2018. He is currently with the School of Computing, University of Eastern Finland, Kuopio, Finland. He is also a Guest/Visiting Professor with the Harbin Institute of Technology, Beijing Normal University, and Shanghai Maritime University, China. He has published over 290 technical papers in refereed journals and international conferences. His current research interests are nature-inspired computing methods (e.g., neural networks, fuzzy logic, evolutionary computing, swarm intelligence, and artificial immune systems) with their applications in optimization, data mining, control, signal processing, and industrial electronics. He is an invited plenary speaker at the 2014 International Workshop on Synchro-Phasor Measurements for Smart Grid, the 2006 International Workshop on Nature Inspired Cooperative Strategies for Optimization, and the 2001 NATO Advanced Research Workshop on Systematic Organization of Information in Fuzzy Systems. He is the General Chair of the 2005 IEEE Mid-Summer Workshop on Soft Computing in Industrial Applications. He is an Associate Editor of *Applied Soft Computing*, the *International Journal of Machine Learning and Cybernetics*, the journal of *Intelligent Automation and Soft Computing*, *International Journal of Innovative Computing*, *Information and Control*, and the *International Journal of Swarm Intelligence and Evolutionary Computation*. He also serves on the editorial boards for few international journals.