Neurocomputing 300 (2018) 70–79

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Feature selection in machine learning: A new perspective

Jie Cai, Jiawei Luo, Shulin Wang, Sheng Yang*

College of Computer Science and Electronic Engineering, Hunan University, Changsha, Hunan, China

ARTICLE INFO

Article history:
Received 6 August 2017
Revised 15 October 2017
Accepted 17 November 2017
Available online 9 March 2018

Keywords:
Feature selection
Dimensionality reduction
Machine learning
Data mining

ABSTRACT

High-dimensional data analysis is a challenge for researchers and engineers in the fields of machine learning and data mining. Feature selection provides an effective way to solve this problem by removing irrelevant and redundant data, which can reduce computation time, improve learning accuracy, and facilitate a better understanding for the learning model or data. In this study, we discuss several frequently-used evaluation measures for feature selection, and then survey supervised, unsupervised, and semi-supervised feature selection methods, which are widely applied in machine learning problems, such as classification and clustering. Lastly, future challenges about feature selection are discussed.

1. Introduction

With the rapid development of modern technology, tremendous new computer and internet applications have generated large amounts of data at an unprecedented speed, such as video, photo, text, voice, and data obtained from social relations and the rise of the Internet of things and cloud computing. These data often have the characteristics of high dimensions, which poses a high challenge for data analysis and decision-making. Feature selection has been proven in both theory and practice effective in processing high-dimensional data and in enhancing learning efficiency [1–3].

Feature selection is referred to the process of obtaining a subset from an original feature set according to certain feature selection criterion, which selects the relevant features of the dataset. It plays a role in compressing the data processing scale, where the redundant and irrelevant features are removed. Feature selection technique can pre-process learning algorithms, and good feature selection results can improve learning accuracy, reduce learning time, and simplify learning results [4–6]. Notably, feature selection and feature extraction [7–9] are two ways to dimensionality reduction. Unlike feature selection, feature extraction usually needs to transform the original data to features with strong pattern recognition ability, where the original data can be regarded as features with weak recognition ability.

Feature selection, which has been a research topic in methodology and practice for decades, is used in many fields, such as image recognition [10–14], image retrieval [15–17], text mining [18–20], intrusion detection [21–23], bioinformatic data analysis [24–31], fault diagnosis [32–34], and so on. According to the theoretical principle, feature selection methods can be based on statistics [35–39], information theory [40–45], manifold [46–48], and rough set [49–53], and can be categorized according to various standards.

(a) According to the utilized training data (labeled, unlabeled, or partially labeled), feature selection methods can be divided into supervised, unsupervised, and semi-supervised models. A unified framework for supervised, unsupervised and semi-supervised feature selection is shown in Fig. 1.
(b) According to their relationship with learning methods, feature selection methods can be classified into filter, wrapper, and embedded models.
(c) According to the evaluation criterion, feature selection methods can be derived from correlation, Euclidean distance, consistency, dependence, and information measure.
(d) According to the search strategies, feature selection methods can be divided into forward increase, backward deletion, random, and hybrid models.
(e) According to the type of the output, feature selection methods can be divided into feature rank (weighting) and subset selection models.

The filter model only considers the association between the feature and the class label. Compared with wrapper model, it has the less computational cost. The evaluation criterion is critical to the filter model. Meanwhile, the embedded model [54–56] selects feature in the training process of learning model, and the feature selection result outputs automatically while the training process is finished. Lars [55] as an embedded method is based on Least absolute shrinkage and selection operator (Lasso). Lasso minimizes the

* Corresponding author.
E-mail address: yangsh0506@sina.com (S. Yang).

https://doi.org/10.1016/j.neucom.2017.11.077
0925-2312/

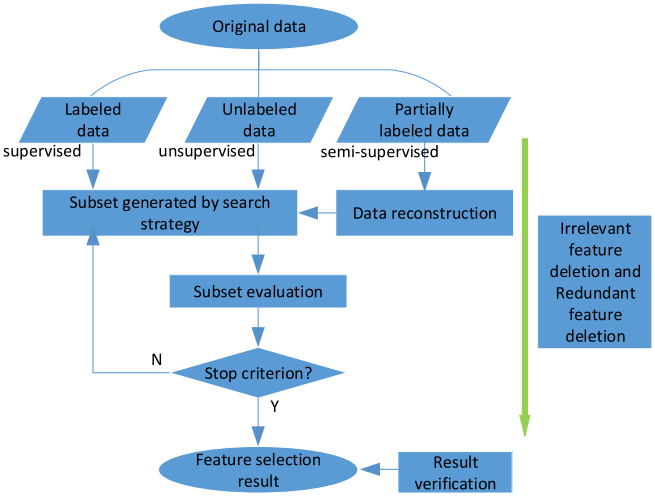

Fig. 1. A framework for feature selection.

sum of squares of residuals when the sum of the absolute values of the regression coefficients is less than a constant, which yields certain strict regression coefficients equal to 0. Then the AIC and the BIC criteria are used to truncate the variables, and the dimension reduction is realized. Lasso-based feature selection methods have well stability, including Lasso for regression model [57], Lars [55], Adaptive-lasso [58], elastic net [59], and so on. Lasso methods are prone to excessive computational overhead or over-fitting problems for high dimensional data.

The performance of the feature selection method is usually evaluated by the machine learning model. The commonly used machine learning models includes Naïve Bayes, KNN, C4.5, SVM, BP-NN, RBF-NN, K-means, Hierarchical clustering, Density based clustering and so on [60–65]. A good feature selection method should have high learning accuracy but less computational overhead (time complexity and space complexity). Although there have been solid reviews on feature selection [66–71], they are mainly focus on specific research fields in feature selection. Therefore, it is still worth to comprehensively survey recent advance on feature selection and discuss some future challenges. In the rest of the paper, we will introduce the evaluation measure for feature selection in Section 2, and then we will focus on supervised, unsupervised and semi-supervised feature selection methods in Section 3, Section 4 and Section 5, respectively. Section 6 concludes the paper and discusses some future challenges.

2. Measure for feature selection

The feature measure or evaluation criterion plays an important role in feature selection, which forms the basis of feature selection [2]. In the context of classification problems, optimal criterion would be the Bayesian error rate $E(S)$ shown in formula (1) under the continuous or discrete condition, where $c_i \in C$ is a class from all the possible classes C existing in the data [45].

$$E(S) = \int_S p(S)(1 - \max_i(p(c_i|S)))dS \quad \text{or} \quad \sum_S p(S)(1 - \max_i(p(c_i|S))) \tag{1}$$

As can be seen from formula (1), $E(S)$ is in the form of sum, and $p(S)(1 - \max_i(p(c_i|S)))$ is non-linear and non-negative. It has an upper bound shown in formula (2), where $H(C|S)$ is the conditional entropy of C given S.

$$E(S) \le H(C|S)/2 \tag{2}$$

It is hard to calculate $E(S)$ directly because S is the combination of features. Researchers prefer to use other measures, such as correlation, dependency and distance. Here, we show some general and representative evaluation measures as follows. Let two features are a and b.

Correlation coefficient:

$$r(a,b) = \frac{Cov(a,b)}{\sqrt{Var(a)}\sqrt{Var(b)}} \tag{3}$$

where $Cov(a, b)$ is the covariance of a and b, and $Var(.)$ is the variance.

Pearson correlation coefficient:

$$r(a,b) = \frac{N\sum a_i b_i - \sum a_i \sum b_i}{\sqrt{N\sum a_i^2 - (\sum a_i)^2}\sqrt{N\sum b_i^2 - (\sum b_i)^2}} \tag{4}$$

Mutual information:

$$I(a;b) = \sum_a \sum_b p(ab)\log\frac{p(ab)}{p(a)p(b)} \tag{5}$$

where $p(.)$ is the probability density function.

Symmetric uncertainty (SU):

$$SU(a;b) = \frac{2I(a;b)}{H(a)+H(b)} \tag{6}$$

where $H(.)$ is the entropy of a feature.

Information distance:

$$d(a,b) = \frac{H(a|b)+H(b|a)}{2} \tag{7}$$

where $H(a|b)$ is the conditional entropy of a given b.

Euclidean distance:

$$d(a,b) = \sqrt{\sum (a_i - b_i)^2} \tag{8}$$

These measures are often used in transformation for special applications. Moreover, there are other measures, such as Laplacian score, Fisher score, dependency index in rough set theory, and so on. Information measures require the feature in discrete type generally, thus the feature discretization should be implemented before feature selection. Discretization methods include Equal-Depth, Equal-Width and manual setting, and etc. [62]. Another frequently-used method is Minimum Description Length (MDL) [72].

3. Supervised feature selection

Supervised feature selection is often oriented to classification problem, and uses the relevance or correlation between the feature and the class label as its fundamental principle. The importance of the features can be evaluated by relevance measures. For a given dataset $D = (X, C)$, with a feature set $X = \{x_1, x_2, ..., x_n\}$ and class label C, the supervised model aims to find an optimal feature subset $\hat{S}(|\hat{S}| = \hat{k})$ that maximizes the classification accuracy.

Many supervised models have been proposed, such as CFS [73], ReliefF [74,75]. CFS is a subset selection method, and mainly uses heuristic approaches to evaluate the effect of single feature corresponding to each category in order to obtain an optimal feature subset. ReliefF [75] is extended from Relief [74] to support multi-class problem. Its main idea is to take Euclidean distance as correlation index and then weights features according to how well they differentiate instances of different classes. Recently, a supervised feature selection method was proposed based on a set of label-aided utility functions for multi-objective optimization feature selection [76].

Hilbert-Schmidt dependency criterion [77] shown in formula (9) is a general framework for feature selection, where $J(.)$ measures the dependency of a feature subset to C. The idea of this

framework is that the good feature subset should maximize $J(.)$, which turns the feature selection into an optimization problem.

$$D_{FS} = \arg\max_{\hat{S} \subseteq F} [J(\hat{S})] \tag{9}$$

If $J(\hat{S})$ is replaced by $I(\hat{S}; C)$, formula (9) is transformed into $\arg\max[I(\hat{S}; C)]$. Since $I(\hat{S}; C) = H(C) - H(C|\hat{S})$ and $H(C)$ is a definite value, maximizing $I(\hat{S}; C)$ means minimizing $H(C|\hat{S})$. According to formula (2), maximizing $I(\hat{S}; C)$ leads to a decline in $E(\hat{S})$. $\arg\max[I(\hat{S}; C)]$ is the core idea of feature selection method based on information theory. In fact, the dependency measure and the correlation measure are essentially the same.

Filter feature selection methods usually use evaluation criteria to enhance the correlation between the feature and the class label and to reduce correlation among features. Moreover, the correlation among features is often replaced by redundancy or diversity (distance). The relevance, redundancy, and diversity measures may be the same or different. Although some proposed methods do not explicitly present correlation, redundancy, or diversity analysis, these analyses are still implemented in their design. Next, we expand on the supervised model.

3.1. Relevance and redundancy based supervised filter model

Relevance (feature–class) and redundancy (feature–feature) analysis is the basis of this type of methods. These models use Euclidean distance, Pearson correlation, and information measures for relevance and redundancy analysis [2,78]. The features in an original set can be divided into four groups [79]: (a) completely irrelevant and noisy features, (b) weakly relevant and redundant features, (c) weakly relevant and non-redundant features, and (d) strongly relevant features. The strongly relevant feature is also called as essential attribute according to rough set theory, which constitutes the Core of conditional attribute set. Supervised feature selection results should include groups (c) and (d). Relevance and redundancy analysis can be transformed into two optimization problems: $\max(Relevance(\hat{S}; C))$ and $\min(Redundancy(\hat{S}))$.

A classical criterion for feature selection based on relevance and redundancy analysis is MRMR (Max-Relevance and Min-Redundancy), which uses mutual information as the evaluation measure [42]. In formula (10), the former item is the first-order expression of relevance analysis, and the later item belongs to redundancy analysis. This expression can be further extended into conditional mutual information form shown in formula (11).

$$\text{MRMR}: \max\left[\frac{1}{|S|}\sum_{x_i \in S} I(x_i; C) - \frac{1}{|S|^2}\sum_{x_i \in S}\sum_{x_j \in S} I(x_i; x_j)\right] \tag{10}$$

$$\text{CMRMR}: \tag{11}$$

$$\max\left[\frac{1}{|S|}\sum_{x_i \in S} I(x_i; C) - \frac{1}{|S|^2}\sum_{x_i \in S}\sum_{x_j \in S}\left(I(x_i; x_j) - I(x_i; x_j|C)\right)\right] \tag{11}$$

In the past, mutual information based feature selection methods have achieved great success. These methods [80–82] mainly focus on relevance analysis using mutual information. According to MRMR criterion, Peng et al. [42] proposed mRMR that implements this criterion by an first-order incremental search. However, mutual information based MRMR only minimizes feature–feature mutual information and ignores the classification performance of candidate features, which might be influenced by the selected features. Conditional mutual information analysis is then introduced to overcome this problem. Feature selection methods based on conditional mutual information have attracted significant attention [43,44,83–85]. Information theory-based feature selection methods select feature subsets that maximize information regarding the class label. The approximate expression and incremental heuristic search approaches are usually employed by these methods because directly calculating mutual information between the feature subset and the class label is difficult. For instances, the mRMR method adopts the incremental search approach, which is the first-order optimum in theory. Therefore, in many cases, the mRMR method could obtain superior performance over conditional mutual information-based feature selection methods.

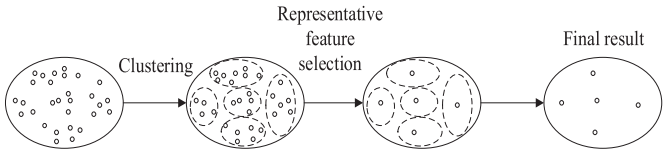

Fig. 2. The process of clustering-based feature selection.

3.2. Relevance and diversity based supervised filter model

Many researchers study feature selection methods with emphasis on diversity among features [86–88]. In literature [88], Kullback–Leibler divergence is used to build a feature selection framework, which focuses on separate class. This type of model could also be expressed by two optimization problems: $\max(Relevance(\hat{S}; C))$ and $\max(Diversity(\hat{S}))$. Dhillon et al. [89] proposed a global criterion for feature/word clustering, and presented a fast, divisive method for text classification. This method maximizes the Jensen–Shannon divergence between feature clusters.

The clustering-based feature selections [90–92] are typically performed in terms of maximizing diversity. The basic process (as shown in Fig. 2) in these methods consists of three steps: first, the appropriate distance measure is chosen to form feature space; next, features are grouped by clustering method; finally, the representative feature of each cluster is selected to generate the selection result. The representative feature often is the most relevance feature with the class label. Ienco and Meo [90] hierarchically clustered the feature sets according to their correlations, and selected the best feature from each cluster to form the final feature subset by using a packing method. Although this method does not need to adjust any parameters, but the introduction of packaging method not only increases the time cost and adds the learning method bias. Witten and Tibshirani [91] integrated sparse K-means and hierarchical clustering into a feature clustering framework. Firstly, the feature set is clustered and divided into multiple feature clusters. Then, Lasso-type penalty factors are used to select the representative features from each cluster, which constitutes the final feature subset. In the supervised learning environment, Liu [92] employed the agglomerative hierarchical clustering method to divide the feature set, and set up the final feature subset by removing the features that are far away from each other. Zhao [93] used the maximum information coefficient as the measure of feature correlation, carried out the affinity propagation clustering of the feature subset, and then selected the centroid from each cluster as the representative feature of the cluster.

Moreover, many clustering-based feature selection methods using information theory have been presented. Au et al. proposed an ACA method [94], which selects an information measure to measure the correlation between features, clusters the features by a K-means-like clustering method, and then selects the representative feature. This method is very effective for gene data classification. The clustering process might fall into a dead loop, and thus the number of iterations should be predefined to stop the clustering process. The FAST [26] method proposed by Song uses hierar-

chical clustering method. This method takes each minimum spanning tree as a cluster and uses the symmetric uncertainty as a metric. FAST is suitable for high dimensional data, but the number of selected features cannot be manually specified. Liu also proposed a clustering feature selection method MFC [95] based on the minimum spanning tree. Unlike FAST, the metric that MFC uses, is the information distance metric, variance of information. Sotoca presented a hierarchical clustering based feature selection method mRR [96], where the conditional mutual information is taken as the diversity measure. Generally, the smaller feature–feature redundancy, the greater feature–feature diversity. However, the redundancy measure often cannot be used as the diversity directly. The above methods use SU [26], variant of information [95] and conditional mutual information [96] as the diversity measures, respectively, but not mutual information as shown in MRMR criterion.

The clustering-based method might select an irrelevant feature since the irrelevant features are often clustered together. Therefore, it is better to delete the irrelevant features before feature clustering. Recently, Zou proposed the MRMD (Max-Relevance-Max-Distance) feature ranking method using a hybrid measure to rank features [97]. This method has not only a better stability for feature selection, but also a shorter running time compared to mRMR method.

3.3. Supervised wrapper model

Wrapper models take the classification error or accuracy rate as the feature evaluation standard. The feature selection result is often produced simultaneously as that of the learning model because the learning method is included in feature selection. In comparison with the filter model, the wrapper model could achieve higher classification accuracy and tend to have a smaller subset size; however, it has poor generalization capability and high time complexity. The SVM-RFE method, which was proposed by Guyon et al. [98], is a widely studied and applied wrapper method that uses support vector machines to measure the performance of features, and constructs a classifier with high performance scores [99,100]. Michalak and Kwaśnicka [101] proposed a relationship-based dual-strategy wrapper feature selection method. A rank criteria system based on class densities for binary data is presented in this method. In literature [102], a two-stage method utilizing a low-cost filter method to rank features and a costly wrapper method to further eliminate irrelevant variables is used.

The genetic algorithm and particle swarm optimization, two random search strategies, are often applied in wrapper models. Hsu [103] uses decision tree to select features, and genetic algorithm is used to find a set of feature subsets that minimize the misclassification rate of decision trees. Chiang and Pell [104] combined the Fisher discriminant analysis with the genetic algorithm. The key variables achieve good results in chemical fault identification. In literature [105], kNN classification accuracy is taken as the evaluation function, and the feature weights of MPEG-7 image are calculated optimally by a real coded chromosome genetic algorithm. An improved binary particle swarm optimization is used to implement feature selection for gene selection, where the kNN accuracy is taken as the evaluator [106]. Xue et al. [107] proposed a PSO feature selection model with new benefit mechanisms that take into consideration both the classification accuracy and the number of selected features.

The filter model is often combined with the wrapper model to form the hybrid feature selection method. Generally, the hybrid feature selection consists of two stages: the first stage is to use the filter to reduce the feature space size by removing the irrelevant and the noisy features; the second stage is to use the wrapper to find the optimal feature subset from the retained features. Akadi et al. [108] combined the mRMR method and genetic algorithm to obtain a Filter-Wrapper hybrid method. Cadenas et al. [109] proposed a hybrid Filter-Wrapper method using Fuzzy Random Forest, called FRF-fs.

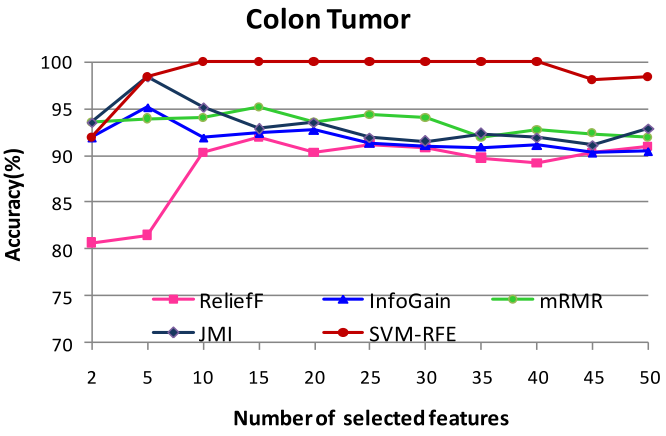

Fig. 3. Classification accuracy of Colon.

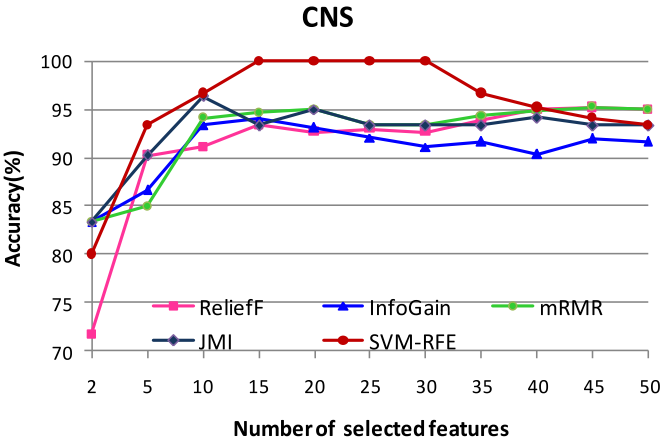

Fig. 4. Classification accuracy of CNS.

3.4. Some state-of-the-art feature selection methods: an experimental analysis

For better and more intuitively understanding the feature selection method, we give an experimental analysis. Five rank feature selection methods, ReliefF [75], Information Gain (InfoGain), mRMR [42], JMI [44] and SVM-RFE [98], are applied on two high dimensional gene expression datasets (Colon and CNS), which are often used to validate the performances of the feature selection methods. The two datasets can be downloaded from http://datam.i2r.a-star.edu.sg/datasets/krbd/. Colon includes 2000 features, 62 samples and 2 classification labels, and CNS includes 7129 features, 60 samples and 2 classification labels. The machine learning model is SVM since SVM-RFE as a wrapper model uses SVM. Furthermore, ten times of 10-fold cross validation are implemented to obtain classification accuracy, and the results are shown as Figs. 3 and 4. The *x*-axis represents the number of features selected, while the *y*-axis shows the average classification accuracy obtained by each feature selection method.

As shown in Figs. 3 and 4, we can reach the following three conclusions.

(a) Those state-of-the-art feature selection methods can obtain the classification accuracy greater than 90%, which shows

that feature selection methods are effective by reducing the data processing size.

(b) The classification accuracy is improved with the increasing number of selected features. When the number of selected features reaches a certain range, the classification accuracy enters in a relatively steady state. In fact, the optimal number of selected features is usually small.

(c) SVM-RFE reaches the best accuracy of 100%, which is better than other filter methods. Generally, the wrapper method obtains much better classification accuracy.

The time complexity of Information Gain is $O(nm)$, and the time complexity of ReliefF is $O(nm^2)$, where n is the feature number and m is the sample size of dataset. The time complexities of JMI and mRMR are both $O(n^2m)$, and the time complexity of SVM-RFE is $O(\max(n, m)n^2)$.

4. Unsupervised feature selection

Unsupervised feature selection methods aim to cover the natural classification of data and improve the clustering accuracy by finding a feature subset based on either clustering or evaluation criteria. Unsupervised feature selection methods could be unsupervised filter or wrapper feature selection methods, depending on whether they rely on cluster algorithms.

4.1. Unsupervised filter model

The unsupervised filter feature selection methods select features according to the nature of the data features. The clustering and learning algorithms are not applied in the feature selection process, thereby reducing clustering time and algorithm complexity. The unsupervised filter feature selection method directly utilizes the statistical performance of all the training data as the evaluation measure, which is highly versatile and suitable for large-scale datasets. However, the clustering performance of the selected feature subset is usually lower than that of the wrapper model because the evaluation criteria are independent of the specific clustering algorithm.

Dash and Liu [110] proposed to use entropy to evaluate the importance of features, and they used the trace criterion to select the most relevant feature subset. Vandenbroucke et al. [111] proposed an unsupervised filter feature selection method. It uses a competitive learning algorithm to classify the samples and ascertain the number of clusters, and then divides the original feature set into several feature subsets. A judgment function is designed for the average dispersion within class and the average scatter distance among classes. The value of the judgment function in each feature subset is calculated, and the feature subset that maximizes the judgment function value is selected to determine the candidate feature. Finally, the correlation coefficient between the candidate feature and the selected feature is calculated. If the correlation coefficient is greater than 0.75, then the candidate feature is abandoned. Alibeigi et al. [112] analyzed the distribution of the feature data by use of the probability density of different feature spaces in an unsupervised environment. The feature is selected by the data distribution relation among the features.

Mitra et al. [113] developed an unsupervised feature selection method that uses the maximum information compression index to measure the similarity between features. This method runs rapidly and is applicable to datasets with different sizes because it does not require a search. Zhou and Chan [114] proposed an unsupervised attribute clustering algorithm, together with an unsupervised feature selection method. It first calculates the maximal information coefficient for each attribute pair to construct an attribute distance matrix, and then clusters all attributes using the optimal K-mode clustering method to find K-mode attributes as features of each cluster. Meanwhile, the number of clusters is determined automatically. Li et al. [27] proposed a clustering-based unsupervised feature selection method called FSFC, which follows the same process as the clustering-based supervised feature selection models. FSFC also works well for high-dimensional datasets.

Unsupervised feature selection techniques [115–117], including filter techniques based on the Laplacian score, have also been proposed. These techniques concern the local topology of the data clusters. He et al. [116] proposed a method based on the idea that data within the same class should be close to one another. Besides, the Laplacian Score is used to evaluate the importance of the features. Saxena et al. [117] selected features using a genetic algorithm with Sammon's stress function, thereby preserving the topology structure of the original data in the reduced feature space.

4.2. Unsupervised wrapper model

The unsupervised wrapper feature selection method uses a clustering-based algorithm to adjust the validity of feature selection [118,119]. The feature subset with the best clustering performance will be considered as the final optimal feature subset. The clustering performance of the feature subset selected by the wrapper method is usually better than that of the feature subset selected by the filter method. However, since each feature subset needs to be evaluated by the clustering algorithm, this method has high computational complexity and it might be a problem when dealing with large-scale data. In addition, the unsupervised feature selection method could be divided into the global and the local wrapper models depending on whether the feature selection is performed in all clusters or in a single cluster.

Dy and Brodley [120] investigated the wrapper framework through feature subset selection using EM clustering. The EM algorithm is applied to estimate the maximum likelihood parameters of a finite Gaussian mixture. Then, scatter separability and maximum likelihood are used to evaluate candidate feature subsets. Gennari [121] incorporated feature selection into the CLASSIT, which is a conceptual information hierarchical clustering algorithm. This unsupervised feature selection method searches for the optimal feature subset from the most significant features according to the clustering capability of features. The feature search continues until the new selected feature can no longer change the existing clustering results. The purpose of this method is to improve the validity and prediction accuracy of feature selection.

Devaney and Ram [122] applied sequential forward and sequential backward searches to find feature subsets. A clustering algorithm is used to evaluate the feature subset, and the optimal feature subset is selected by clustering accuracy. Vaithyanathan and Dom [123] conceived an object evaluation function to select the feature subset and used the Bayesian statistical evaluation model in the document clustering problem to find the optimal number of clusters. They built a polynomial model for each cluster and extended the concept of hierarchical clustering. Huang et al. [124] also proposed an improved K-means clustering algorithm called W-K-means for feature selection. Feature weighting is used to guide the mean clustering process such that it is better clustered on important features rather than depending equally on each feature. The W-K-means algorithm is used to cluster the datasets to produce a weight set for each feature. The feature set is then selected according to the weight of the feature, and the selected features are removed from the dataset. Finally, W-K-means or other clustering algorithms are used to cluster the datasets to obtain the final clustering results. Deepthi and Thampi [125] presented an unsupervised feature selection approach to implement sample-based clustering on gene expression data. The proposed work uses PSO to search for the best subset and evaluates the subsets using the K-means algorithm.

5. Semi-supervised feature selection

Given the dataset $D=\{D_l, D_u\}$, where D_l is the sample set with class labels, and D_u is the sample set without class labels, semi-supervised learning model uses D_u to improve the learning performance of learning model trained by D_l. Semi-supervised feature selection methods, which are mainly filter models, play an important role in semi-supervised learning. Score functions are applied in most semi-supervised feature selection methods and can be divided into four categories: variance score [126], Laplacian score [127–130], Fisher score [131–136], and Constraint score [137–139].

Semi-supervised feature selection methods based on the Laplacian score combine Laplacian criteria and output information for feature selection. These are graph-based methods that construct the neighborhood graph and evaluate the features according to their capability in preserving the local data structure. Semi-supervised feature selection methods based on the Fisher score use the properties of Fisher criterion and the local structure and distribution information of labeled and unlabeled data to select the features with the best discriminant and locality-preserving capabilities. Semi-supervised feature selection methods based on pairwise constraints use pairwise constraints and the local properties of the labeled and unlabeled data to evaluate the relevance of features according to their constraint and locality-preserving power.

The methods mentioned above are usually scored against individual features. Existing researches emphasize the redundancy analysis among features in designing the semi-supervised feature selection methods [140–142]. Benabdeslem et al proposed a filter approach based on a constrained Laplacian score, in which the redundancy is removed after the relevant features are selected [140]. Wang presented a semi-supervised filter feature selection method called SRFS based on information theory, where the unlabeled data are utilized in the Markov blanket as the labeled data through the relevance gain [141]. Meanwhile, research is needed on utilizing semi-supervised feature selection methods for regression problems [71].

6. Conclusions and future challenges

Feature selection is a typical optimization problem in which the optimal solution can only be obtained by an exhaustive search given an evaluation or search criterion. Therefore, researchers still apply the heuristic method with polynomial time complexity for high-dimensional problems. In this paper, we have surveyed some representative supervised, unsupervised, and semi-supervised feature selection methods and their recent applications in machine learning. Although fruitful achievements have been obtained in this area, some challenges still remain.

6.1. Extreme data for machine learning

Extreme data are revealed in small or large samples with high dimensions and imbalanced class labels. Big data with large samples and high dimensionality have appeared in various data mining fields, such as text mining and information retrieval [143,144]. For example, Weinberger studied a collaborative spam email filtering task with 16 trillion (10^{13}) unique features [144,145]. In gene expression analysis, the dimension might be high (e.g., up to 20,000) while the sample size is small (e.g., approximately 50) [146,147]. Many researchers focus mainly on imbalanced classification and related problems [148,149].

Novel feature selection methods need to be developed for extreme datasets, which often invalidate traditional methods. Moreover, the emphasis of the researches should be different for different types of extreme datasets. For example, the time complexity of an algorithm is nearly as important as its accuracy when working with high-dimensional datasets or large samples. Meanwhile, feature selection methods should focus on the accuracy of their recognition capabilities for minority classes when working with imbalanced datasets.

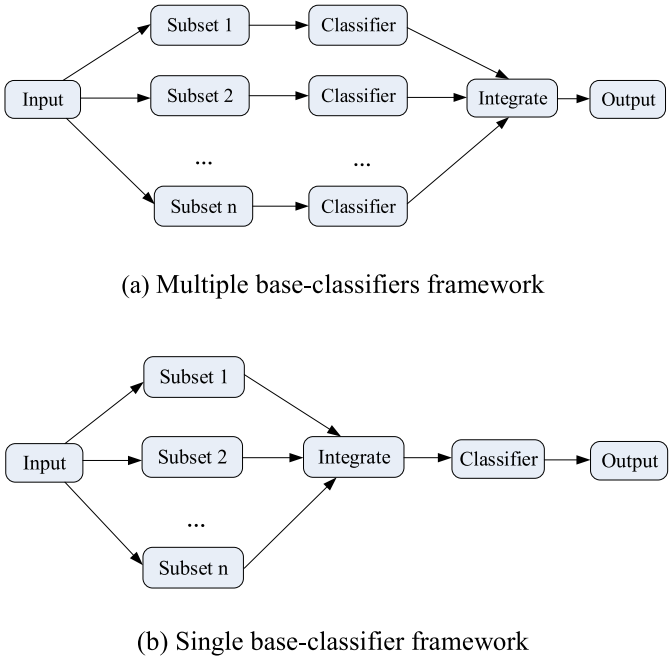

Fig. 5. Ensemble feature selection framework.

Furthermore, when designing feature selection method, we should not only consider the accuracy of the method but also the scalability and stability. The scalability is defined as the sensitivity of the computational performance of feature selection method to the data scale. The computational performance includes accuracy, time complexity and space complexity. The feature selection method with good scalability should be adapted to datasets of various sizes and have polynomial time complexity. The stability is defined as the sensitivity of the feature selection results to training set variations. When the feature selection results change after adding or reducing some samples, the stability of the feature selection method is considered to be poor. The stability indexes to measure feature selection algorithm include Hamming distance, average Tanimoto index, weighted consistency and etc. These indexes need to calculate the size of the intersection of two subsets or specify the size of object subset, which often lead to failure in stability evaluation. Good stability evaluation methods need: (1) better feature subset similarity measure and (2) insensitivity to the size difference between feature subsets.

6.2. Ensemble feature selection

Ensemble feature selection is the ensemble of base classifiers trained by different feature spaces obtained by feature selection. Fig. 5 illustrates the overall workflow of ensemble feature selection. This system can improve feature selection stability and obtain diverse training data to improve the performance of the ensemble classifiers [150–152]. The random subspace method (RSM) [153,154] and the random forest method [155] could be regarded as ensemble feature selection methods since they train base classifiers on different feature subsets. Bock [156] introduced generalized additive models (GAMs) as base classifiers for binary ensemble classification using RSM and/or Bagging. GAMbag, GAMrsm, and GAMens, which use GAMs as base classifiers, are proposed as alternative ensemble selection strategies. These methods are

especially useful when working with microarray data, which generally have small samples and suffer from degraded base classifier performance caused by using different training sample sets. Ensemble feature selection can manipulate multiple gene sets simultaneously without affecting the performance of base classifiers. Liu et al. [157] proposed a new ensemble gene selection method called EGS based on information theory. Wang et al. [158] used a heuristic breadth-first search to find as many optimal gene subsets as possible, and trains SVMs based on those subsets as the base classifiers. The ensemble classifier is then constructed for robust tumor classification by majority voting. Zhang and Suganthan [159] proposed a novel transformation-based method to increase the diversity of base classifiers, thereby improving the ensemble accuracy.

Another form of ensemble feature selection is the single base-classifier framework. In this framework, multiple feature subsets are generated first by filters, then those subsets are combined into one ensemble subset by the intersection strategy [160,161]. This framework usually obtains an ensemble feature subset with higher accuracy than single subset selected by filter, and also does not require a combiner method for base classifiers like multiple base-classifiers framework. However, the redundancy between features in the ensemble feature subset might be high.

Ensemble feature selection aims to reduce the influence of high dimensions on learning algorithms while producing ensembles of diverse base classifiers and building effective ensemble learning systems that suit classification problems. Ensemble feature selection methods can effectively handle high-dimensional data because of how they combine classifier ensemble and feature selection [157,162]. However, the performance of ensemble methods is unstable because the feature subset is divided randomly and diversity among the selected feature subsets is not guaranteed.

6.3. Online feature selection

Traditional feature selection is based on static feature space, that is, the input for feature selection does not change. However, in many fields such as video data stream and network data stream, the data changes over time. As a result, the feature obtained is changing constantly. Feature selection under this dynamic feature space is called online feature selection. Online feature selection is independent of online learning model. It can be used in dynamic and unknown feature space, but online learning model is applicable to certain feature space.

The state-of-the-art online feature selection methods include Grafting [163], Alpha-investing [164], OSFS [165], Fast-OSFS [166], and SAOLA [167]. SAOLA (towards Scalable and Accurate OnLine Approach) achieves the redundancy analysis by feature–feature correlation, which replaces the feature subset search in OSFS and FAST-OSFS to remove redundant, thus greatly improving the time efficiency. An online feature selection open-source library called LOFS is freely available at https://github.com/kuiy/LOFS.

The performance of online feature selection methods can be further improved in three ways: the stability of the method, the redundancy analysis and the time efficiency. Specifically, for the real-time processing system, time efficiency is especially important. Excellent algorithm design can implement fast online feature selection. Meanwhile, high performance hardware can also improve the processing speed with good result.

6.4. Feature selection and deep learning

Deep learning combines low-level features to form more abstract high-level features, and discovers distributed representations of data. It is one of the most important advances in machine learning field in recent years, and widely used in speech recognition, image processing and recognition, information understanding and game intelligence. Deep learning consists of many models, such as DBN, DNN, CNN, RNN, GAN and etc.

The connection between feature selection and deep learning could be summarized as follows. The irrelevant feature might cost a great deal of resources during the training process of neural networks. Features should be accurately selected to save the training time for deep learning methods [168–171]. In literature [170], Individual Training Error Reduction Ranking is used to select nodes in each hidden layer, which simplifies the structure of DNN. In literature [171], reconstruction error is used to select the input feature for a DNN, which is equivalent to a network pruning algorithm. Furthermore, feature selection at the input level, is also helpful to understand the nature of a complex system or a trained model.

When training a neural network, the weight of irrelevant feature will be approximately zero. Therefore, deep learning method can also be applied to feature selection. A deep-learning-based feature selection method is proposed for remote sensing scene classification/recognition, which formulates the feature selection problem as a feature reconstruction problem [172]. Feature selection-based deep learning and deep-learning-based feature selection should be explored further. For example, should we consider removing irrelevant and noisy features from the original features, and using only relevant features to train the depth learning model?

Acknowledgment

This work was supported by the grants of the National Science Foundation of China (Grant nos. 61472467, 61672011 and 61572180).

References

[1] A.L. Blum, P. Langley, Selection of relevant features and examples in machine learning, Artif. Intell. 97 (1997) 245–271.
[2] H. Liu, H. Motoda, Feature Selection for Knowledge Discovery and Data Mining, Springer Science & Business Media, 2012.
[3] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[4] Z. Zhao, F. Morstatter, S. Sharma, S. Alelyani, A. Anand, H. Liu, Advancing Feature Selection Research, ASU Feature Selection Repository (2010) 1–28.
[5] P. Langley, Selection of relevant features in machine learning, in: Proceedings of the AAAI Fall Symposium on Relevance, 1994, pp. 245–271.
[6] P. Langley, Elements of Machine Learning, Morgan Kaufmann, 1996.
[7] J.L. Crowley, A.C. Parker, A representation for shape based on peaks and ridges in the difference of low pass transform, IEEE Trans. Pattern Anal. Mach. Intell. 6 (1984) 156–170.
[8] Z.L. Sun, D.S. Huang, Y.M. Cheun, Extracting nonlinear features for multispectral images by FCMC and KPCA, Digit. Signal Process. 15 (2005) 331–346.
[9] Z.L. Sun, D.S. Huang, Y.M. Cheung, J. Liu, G..B. Huang, Using FCMC, FVS, and PCA techniques for feature extraction of multispectral images, IEEE Geosci. Remote Sens. Lett. 2 (2005) 108–112.
[10] A. Khotanzad, Y.H. Hong, Rotation invariant image recognition using features selected via a systematic method, Pattern Recognit. 23 (1990) 1089–1101.
[11] N. Vasconcelos, Feature selection by maximum marginal diversity: optimality and implications for visual recognition, in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003, pp. 762–769.
[12] N. Vasconcelos, M. Vasconcelos, Scalable discriminant feature selection for image retrieval and recognition, in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2004.
[13] J.Y. Choi, Y.M. Ro, K.N. Plataniotis, Boosting color feature selection for color face recognition, IEEE Trans. Image Process. 20 (2011) 1425–1434.
[14] A. Goltsev, V. Gritsenko, Investigation of efficient features for image recognition by neural networks„ Neural Netw. 28 (2012) 15–23.
[15] D.L. Swets, J.J. Weng, Efficient content-based image retrieval using automatic feature selection, in: Proceedings of International Symposium on Computer Vision, 1995.
[16] D.L. Swets, J.J. Weng, Using discriminant eigenfeatures for image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 18 (1996) 831–836.
[17] E. Rashedi, H. Nezamabadi-Pour, S. Saryazdi, A simultaneous feature adaptation and feature selection method for content-based image retrieval systems, Knowl.-Based Syst. 39 (2013) 85–94.
[18] D.D. Lewis, Y. Yang, T.G. Rose, F. Li, Rcv1: a new benchmark collection for text categorization research, J. Mach. Learn. Res. 5 (2004) 361–397.

[19] L.P. Jing, H.K. Huang, H.B. Shi, Improved feature selection approach TFIDF in text mining, in: Proceedings of International Conference on Machine Learning and Cybernetics, 2002, pp. 944–946.
[20] S. Van Landeghem, T. Abeel, Y. Saeys, Y. Van de Peer, Discriminative and informative features for biomolecular text mining with ensemble feature selection, Bioinformatics 26 (2010) 554–560.
[21] G. Stein, B. Chen, A.S. Wu, K.A. Hua, Decision tree classifier for network intrusion detection with GA-based feature selection, in: Proceedings of the 43rd ACM Southeast Conference, 2005, pp. 136–141.
[22] F. Amiri, M.R. Yousefi, C. Lucas, A. Shakery, N. Yazdani, Mutual information-based feature selection for intrusion detection systems, J. Netw. Comput. Appl. 34 (2011) 1184–1199.
[23] A. Alazab, M. Hobbs, J. Abawajy, M. Alazab, Using feature selection for intrusion detection system, in: Proceedings of International Symposium on Communications and Information Technologies (ISCIT), 2012, pp. 296–301.
[24] H. Liu, J. Li, L. Wong, A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns, Genome Inform. 13 (2002) 51–60.
[25] H. Liu, H. Han, J. Li, L. Wong, Using amino acid patterns to accurately predict translation initiation sites, In Silico Biol. 4 (2004) 255–269.
[26] Q. Song, J. Ni, G. Wang, A fast clustering-based feature subset selection algorithm for high-dimensional data, IEEE Trans. Knowl. Data Eng. 25 (2013) 1–14.
[27] G. Li, X. Hu, X. Shen, X. Chen, Z. Li, A novel unsupervised feature selection method for bioinformatics data sets through feature clustering, in: Proceedings of IEEE International Conference on Granular Computing, 2008, pp. 41–47.
[28] Y.F. Gao, B.Q. Li, Y.D. Cai, K.Y. Feng, Z.D. Li, Y. Jiang, Prediction of active sites of enzymes by maximum relevance minimum redundancy (mRMR) feature selection, Mol. Biosyst. 9 (2013) 61–69.
[29] D.S. Huang, C.H. Zheng, Independent component analysis-based penalized discriminant method for tumor classification using gene expression data, Bioinformatics 22 (2006) 1855–1862.
[30] C.H. Zheng, D.S. Huang, L. Zhang, X.Z. Kong, Tumor clustering using nonnegative matrix factorization with gene selection, IEEE Trans. Inf. Technol. Biomed. 13 (2009) 599–607.
[31] H.J. Yu, D.S. Huang, Normalized feature vectors: a novel alignment-free sequence comparison method based on the numbers of adjacent amino acids, IEEE/ACM Trans. Comput. Biol. Bioinform. 10 (2013) 457–467.
[32] L. Wang, J. Yu, Fault feature selection based on modified binary PSO with mutation and its application in chemical process fault diagnosis, Adv. Nat. Comput. 3612 (2005) 832–840.
[33] T.W. Rauber, F. de Assis Boldt, F.M. Varejão, Heterogeneous feature models and feature selection applied to bearing fault diagnosis, IEEE Trans. Ind. Electron. 62 (2015) 637–646.
[34] K. Zhang, Y. Li, P. Scarf, A. Ball, Feature selection for high-dimensional machinery fault diagnosis data using multiple models and Radial Basis Function networks, Neurocomputing 74 (2011) 2941–2952.
[35] M. Vasconcelos, N. Vasconcelos, Natural image statistics and low-complexity feature selection, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2009) 228–244.
[36] T. Khoshgoftaar, D. Dittman, R. Wald, A. Fazelpour, First order statistics based feature selection: a diverse and powerful family of feature seleciton techniques, in: Proceedings of 11th International Conference on Machine Learning and Applications (ICMLA), 2012, pp. 151–157.
[37] J. Gibert, E. Valveny, H. Bunke, Feature selection on node statistics based embedding of graphs, Pattern Recognit. Lett. 33 (2012) 1980–1990.
[38] M.C. Lane, B. Xue, I. Liu, M. Zhang, Gaussian based particle swarm optimisation and statistical clustering for feature selection, in: Proceedings of European Conference on Evolutionary Computation in Combinatorial Optimization, 2014, pp. 133–144.
[39] H. Li, C.J. Li, X.J. Wu, J. Sun, Statistics-based wrapper for feature selection: an implementation on financial distress identification with support vector machine, Appl. Soft Comput. 19 (2014) 57–67.
[40] L. Shen, L. Bai, Information theory for Gabor feature selection for face recognition, EURASIP J. Appl. Signal Process. (2006) 1–11.
[41] B. Morgan, Model selection and inference: a practical information – theoretic approach, Biometrics 57 (2001) 320.
[42] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 1226–1238.
[43] F. Fleuret, Fast binary feature selection with conditional mutual information, J. Mach. Learn. Res. 5 (2004) 1531–1555.
[44] H.H. Yang, J.E. Moody, Data visualization and feature selection: new algorithms for nongaussian data, Adv. Neural Inf. Process. Syst. 12 (1999) 687–693.
[45] B. Bonev, Feature Selection Based on Information Theory, Universidad de Alicante, 2010.
[46] Z. Xu, I. King, M.R.T. Lyu, R. Jin, Discriminative semi-supervised feature selection via manifold regularization, IEEE Trans. Neural Netw. 21 (2010) 1033–1047.
[47] B. Jie, D. Zhang, B. Cheng, D. Shen, Manifold regularized multi-task feature selection for multi-modality classification in Alzheimer's disease, in: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention, 2013, pp. 275–283.
[48] B. Li, C.H. Zheng, D.S. Huang, Locally linear discriminant embedding: an efficient method for face recognition, Pattern Recognit. 41 (2008) 3813–3821.
[49] R.W. Swiniarski, A. Skowron, Rough set methods in feature selection and recognition, Pattern Recognit. Lett. 24 (2003) 833–849.
[50] Y. Chen, D. Miao, R. Wang, A rough set approach to feature selection based on ant colony optimization, Pattern Recognit. Lett. 31 (2010) 226–233.
[51] W. Shu, H. Shen, Incremental feature selection based on rough set in dynamic incomplete data, Pattern Recognit. 47 (2014) 3890–3906.
[52] J. Derrac, C. Cornelis, S. García, F. Herrera, Enhancing evolutionary instance selection algorithms by means of fuzzy rough set based feature selection, Inf. Sci. 186 (2012) 73–92.
[53] J. Wang, K. Guo, S. Wang, Rough set and Tabu search based feature selection for credit scoring, Procedia Comput. Sci. 1 (2010) 2425–2432.
[54] J.R. Quinlan, C4. 5: Programs for Machine Learning, Elsevier, 2014.
[55] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, Ann. Stat. 32 (2004) 407–499.
[56] A. Mirzaei, Y. Mohsenzadeh, H. Sheikhzadeh, Variational relevant sample-feature machine: a fully Bayesian approach for embedded feature selection, Neurocomputing 241 (2017) 181–190.
[57] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc.: Ser. B 58 (1996) 267–288.
[58] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc.: Ser. B 67 (2005) 301–320.
[59] H. Zou, The adaptive lasso and its oracle properties, J. Am. Stat. Assoc. 101 (2006) 1418–1429.
[60] D.S. Huang, Radial basis probabilistic neural networks: Model and application, Int. J. Pattern Recognit. Artif. Intell. 13 (1999) 1083–1101.
[61] A. Rodriguez, A. Laio, Clustering by fast search and find of density peaks, Science 344 (2014) 1492–1496.
[62] J. Han, J. Pei, M. Kamber, Data Mining: Concepts and Techniques, Elsevier, 2011.
[63] D.S. Huang, Systematic Theory of Neural Networks for Pattern Recognition, Publishing House of Electronic Industry of China, 1996.
[64] D.S. Huang, J.X. Du, A constructive hybrid structure optimization methodology for radial basis probabilistic neural networks, IEEE Trans. Neural Netw. 19 (2008) 2099–2115.
[65] J.R. Zhang, J. Zhang, T.M. Lok, M.R. Lyu, A hybrid particle swarm optimization–back-propagation algorithm for feedforward neural network training, Appl. Math. Comput. 185 (2007) 1026–1037.
[66] J. Tang, S. Alelyani, H. Liu, Feature selection for classification: a review, Data Classif.: Algorithms Appl. (2014) 37–64 CRC Press.
[67] S. Alelyani, J. Tang, H. Liu, Feature selection for clustering: a revie, Data Clust.: Algorithms Appl. 29 (2013) 110–121.
[68] J.R. Vergara, P.A. Estévez, A review of feature selection methods based on mutual information, Neural Comput. Appl. 24 (2014) 175–186.
[69] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, Recent advances and emerging challenges of feature selection in the context of big data, Knowl.-Based Syst. 86 (2015) 33–45.
[70] J.C. Ang, A. Mirzal, H. Haron, H.N.A. Hamed, Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection, IEEE/ACM Trans. Comput. Biol. Bioinf. 13 (2016) 971–989.
[71] R. Sheikhpour, M.A. Sarram, S. Gharaghani, M.A.Z. Chahooki, A survey on semi-supervised feature selection methods, Pattern Recognit. 64 (2017) 141–158.
[72] U. Fayyad, K. Irani, Multi-interval discretization of continuous-valued attributes for classification learning, in: Proceedings of the 13th International Joint Conference on Artificial Intelligence, 1993, pp. 1022–1027.
[73] M.A. Hall, Correlation-based feature selection of discrete and numeric class machine learning, in: Proceedings of 17th International Conference on Machine Learning, 2000, pp. 359–366.
[74] K. Kira, L.A. Rendell, A practical approach to feature selection, in: Proceedings of the Ninth International Workshop on Machine Learning, 1992, pp. 249–256.
[75] I. Kononenko, Estimating attributes: analysis and extensions of RELIEF, in: Proceedings of European Conference on Machine Learning, 1994, pp. 171–182.
[76] P. Martín-Smith, J. Ortega, J. Asensio-Cubero, J.Q. Gan, A. Ortiz, A supervised filter method for multi-objective feature selection in EEG classification based on multi-resolution analysis for BCI, Neurocomputing 250 (2017) 45–56.
[77] L. Song, A. Smola, A. Gretton, K.M. Borgwardt, J. Bedo, Supervised feature selection via dependence estimation, in: Proceedings of the 24th International Conference on Machine Learning, 2007, pp. 823–830.
[78] H. Liu, L. Yu, Toward integrating feature selection algorithms for classification and clustering, IEEE Trans. Knowl. Data Eng. 17 (2005) 491–502.
[79] L. Yu, H. Liu, Efficient feature selection via analysis of relevance and redundancy, J. Mach. Learn. Res. 5 (2004) 1205–1224.
[80] R. Battiti, Using mutual information for selecting features in supervised neural net learning, IEEE Trans. Neural Netw. 5 (1994) 537–550.
[81] N. Kwak, C.H. Choi, Improved mutual information feature selector for neural networks in supervised learning, in: Proceedings of International Joint Conference on Neural Networks (IJCNN), 1999, pp. 1313–1318.
[82] J. Novovičová, P. Somol, M. Haindl, P. Pudil, Conditional mutual information based feature selection for classification task, in: Proceedings of the 12th Iberoamerican Conference on Congress on Pattern Recognition, 2007, pp. 417–426.
[83] Y. Zhang, Z. Zhang, Feature subset selection with cumulate conditional mutual information minimization, Expert Syst. Appl. 39 (2012) 6078–6088.
[84] G. Herman, B. Zhang, Y. Wang, G. Ye, F. Chen, Mutual information-based

method for selecting informative feature sets, Pattern Recognit. 46 (2013) 3315–3327.
[85] H. Cheng, Z. Qin, W. Qian, W. Liu, Conditional mutual information based feature selection, in: Proceedings of International Symposium on Knowledge Acquisition and Modeling, 2008, pp. 103–107.
[86] J. Novovicová, P. Pudil, J. Kittler, Divergence based feature selection for multimodal class densities, IEEE Trans. Pattern Anal. Mach. Intell. 18 (1996) 218–223.
[87] C. Lee, G.G. Lee, Information gain and divergence-based feature selection for machine learning-based text categorization, Inf. Process. Manag. 42 (2006) 155–165.
[88] Y. Zhang, S. Li, T. Wang, Z. Zhang, Divergence-based feature selection for separate classes, Neurocomputing 101 (2013) 32–42.
[89] I.S. Dhillon, S. Mallela, R. Kumar, A divisive information-theoretic feature clustering algorithm for text classification, J. Mach. Learn. Res. 3 (2003) 1265–1287.
[90] D. Ienco, R. Meo, Exploration and reduction of the feature space by hierarchical clustering, in: Proceedings of SIAM International Conference on Data Mining, 2008, pp. 577–587.
[91] D.M. Witten, R. Tibshirani, A framework for feature selection in clustering, J. Am. Stat. Assoc. 105 (2010) 713–726.
[92] H. Liu, X. Wu, S. Zhang, Feature selection using hierarchical feature clustering, in: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, 2011, pp. 979–984.
[93] X. Zhao, W. Deng, Y. Shi, Feature selection with attributes clustering by maximal information coefficient, Procedia Comput. Sci. 17 (2013) 70–79.
[94] W.H. Au, K.C. Chan, A.K. Wong, Y. Wang, Attribute clustering for grouping, selection, and classification of gene expression data, IEEE/ACM Trans. Comput. Biol. Bioinf. 2 (2005) 83–101.
[95] Q. Liu, J. Zhang, J. Xiao, H. Zhu, Q. Zhao, A supervised feature selection algorithm through minimum spanning tree clustering, in: Proceedings of International Conference on Tools with Artificial Intelligence (ICTAI), 2014, pp. 264–271.
[96] J.M. Sotoca, F. Pla, Supervised feature selection by clustering using conditional mutual information-based distances, Pattern Recognit. 43 (2010) 2068–2081.
[97] Q. Zou, J. Zeng, L. Cao, R. Ji, A novel features ranking metric with application to scalable visual and bioinformatics data classification, Neurocomputing 173 (2016) 346–354.
[98] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (2002) 389–422.
[99] C. Furlanello, M. Serafini, S. Merler, G. Jurman, Semisupervised learning for molecular profiling, IEEE/ACM Trans. Comput. Biol. Bioinf. 2 (2005) 110–118.
[100] J. Zhong, J. Wang, W. Peng, Z. Zhang, M. Li, A feature selection method for prediction essential protein, Tsinghua Sci. Technol. 20 (2015) 491–499.
[101] K. Michalak, H. Kwaśnicka, Correlation-based feature selection strategy in classification problems, Int. J. Appl. Math. Comput. Sci. 16 (2006) 503–511.
[102] K. Javed, H.A. Babri, M. Saeed, Feature selection based on class-dependent densities for high-dimensional binary data, IEEE Trans. Knowl. Data Eng. 24 (2012) 465–477.
[103] W.H. Hsu, Genetic wrappers for feature selection in decision tree induction and variable ordering in Bayesian network structure learning, Inf. Sci. 163 (2004) 103–122.
[104] L.H. Chiang, R.J. Pell, Genetic algorithms combined with discriminant analysis for key variable identification, J. Process Control 14 (2004) 143–155.
[105] J. Lu, T. Zhao, Y. Zhang, Feature selection based-on genetic algorithm for image annotation, Knowl.-Based Syst. 21 (2008) 887–891.
[106] L.Y. Chuang, H.W. Chang, C.J. Tu, C.H. Yang, Improved binary PSO for feature selection using gene expression data, Comput. Biol. Chem. 32 (2008) 29–38.
[107] B. Xue, M. Zhang, W.N. Browne, Particle swarm optimisation for feature selection in classification: novel initialisation and updating mechanisms, Appl. Soft Comput. 18 (2014) 261–276.
[108] A. El Akadi, A. Amine, A. El Ouardighi, D. Aboutajdine, A two-stage gene selection scheme utilizing MRMR filter and GA wrapper, Knowl. Inf. Syst. 26 (2011) 487–500.
[109] J.M. Cadenas, M.C. Garrido, R. MartíNez, Feature subset selection filter–wrapper based on low quality data, Expert Syst. Appl. 40 (2013) 6241–6252.
[110] M. Dash, H. Liu, Handling large unsupervised data via dimensionality reduction, in: Proceedings of SIGMOD Research Issues in Data Mining and Knowledge Discovery Workshop, 1999.
[111] N. Vandenbroucke, L. Macaire, J.G. Postaire, Unsupervised color texture feature extraction and selection for soccer image segmentation, in: Proceedings of International Conference on Image Processing, 2000, pp. 800–803.
[112] M. Alibeigi, S. Hashemi, A. Hamzeh, Unsupervised feature selection based on the distribution of features attributed to imbalanced data sets, Int. J. Artif. Intell. Expert Syst. 2 (2011) 14–22.
[113] P. Mitra, C. Murthy, S.K. Pal, Unsupervised feature selection using feature similarity, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2002) 301–312.
[114] P.Y. Zhou, K.C. Chan, An unsupervised attribute clustering algorithm for unsupervised feature selection, in: Proceedings of IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2015, pp. 1–7.
[115] P. Padungweang, C. Lursinsap, K. Sunat, Univariate filter technique for unsupervised feature selection using a new laplacian score based local nearest neighbors, in: Proceedings of Asia-Pacific Conference on Information Processing, 2009, pp. 196–200.
[116] X. He, D. Cai, P. Niyogi, Laplacian score for feature selection, in: Proceedings of Advances in Neural Information Processing Systems, 2005, pp. 507–514.
[117] A. Saxena, N.R. Pal, M. Vora, Evolutionary methods for unsupervised feature selection using Sammon's stress function, Fuzzy Inf. Eng. 2 (2010) 229–247.
[118] R. Agrawal, J. Gehrke, D. Gunopulos, P. Raghavan, Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications, ACM, 1998.
[119] B. Mirkin, Concept learning and feature selection based on square-error clustering, Mach. Learn. 35 (1999) 25–39.
[120] J.G. Dy, C.E. Brodley, Feature selection for unsupervised learning, J. Mach. Learn. Res. 5 (2004) 845–889.
[121] J. Gennari, Concept formation and attention, in: Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society, 1991, pp. 724–728.
[122] M. Devaney, A. Ram, Efficient feature selection in conceptual clustering, in: Proceedings of International Conference on Machine Learning (ICML), 1997, pp. 92–97.
[123] S. Vaithyanathan, B. Dom, Model selection in unsupervised learning with applications to document clustering, in: Proceedings of International Conference on Machine Learning, 1999, pp. 433–443.
[124] J.Z. Huang, J. Xu, M. Ng, Y. Ye, Weighting method for feature selection in k-means, Computational Methods of Feature Selection, CRC Press, 2008, pp. 193–209.
[125] P. Deepthi, S.M. Thampi, Unsupervised gene selection using particle swarm optimization and k-means, in: Proceedings of the Second ACM IKDD Conference on Data Sciences, 2015, pp. 134–135.
[126] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, 1995.
[127] H. Cheng, W. Deng, C. Fu, Y. Wang, Z. Qin, Graph-based semi-supervised feature selection with application to automatic spam image identification, Computer Science for Environmental Engineering and EcoInformatics, Springer, 2011, pp. 259–264.
[128] J. Zhao, K. Lu, X. He, Locality sensitive semi-supervised feature selection, Neurocomputing 71 (2008) 1842–1849.
[129] G. Doquire, M. Verleysen, Graph Laplacian for semi-supervised feature selection in regression problems, in: Proceedings of International Work-Conference on Artificial Neural Networks, 2011, pp. 248–255.
[130] G. Doquire, M. Verleysen, A graph Laplacian based approach to semi-supervised feature selection for regression problems, Neurocomputing 121 (2013) 5–13.
[131] L. Chen, R. Huang, W. Huang, Graph-based semi-supervised weighted band selection for classification of hyperspectral data, in: Proceedings of International Conference on Audio Language and Image Processing (ICALIP), 2010, pp. 1123–1126.
[132] M. Yang, Y.J. Chen, G.L. Ji, Semi_Fisher score: a semi-supervised method for feature selection, in: Proceedings of International Conference on Machine Learning and Cybernetics (ICMLC), 2010, pp. 527–532.
[133] L. Sunzhong, H. Jiang, L. Zhao, D. Wang, M. Fan, Manifold based fisher method for semi-supervised feature selection, in: Proceedings of International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2013, pp. 664–668.
[134] W. Yang, C. Hou, Y. Wu, A semi-supervised method for feature selection, in: Proceedings of International Conference on Computational and Information Sciences (ICCIS), 2011, pp. 329–332.
[135] Y. Liu, F. Nie, J. Wu, L. Chen, Efficient semi-supervised feature selection with noise insensitive trace ratio criterion, Neurocomputing 105 (2013) 12–18.
[136] Y. Liu, F. Nie, J. Wu, L. Chen, Semi-supervised feature selection based on label propagation and subset selection, in: Proceedings of International Conference on Computer and Information Application (ICCIA), 2010, pp. 293–296.
[137] M. Kalakech, P. Biela, L. Macaire, D. Hamad, Constraint scores for semi-supervised feature selection: a comparative study, Pattern Recognit. Lett. 32 (2011) 656–665.
[138] K. Benabdeslem, M. Hindawi, Constrained laplacian score for semi-supervised feature selection, in: Proceedings of Machine Learning and Knowledge Discovery in Databases, 2011, pp. 204–218.
[139] D. Zhang, S. Chen, Z.H. Zhou, Constraint Score: A new filter method for feature selection with pairwise constraints, Pattern Recognit. 41 (2008) 1440–1451.
[140] K. Benabdeslem, M. Hindawi, Efficient semi-supervised feature selection: constraint, relevance, and redundancy, IEEE Trans. Knowl. Data Eng. 26 (2014) 1131–1143.
[141] Y. Wang, J. Wang, H. Liao, H. Chen, An efficient semi-supervised representatives feature selection algorithm based on information theory, Pattern Recognit. 61 (2017) 511–523.
[142] X.K. Yang, L. He, D. Qu, W.Q. Zhang, Semi-supervised minimum redundancy maximum relevance feature selection for audio classification, Multimedia Tools and Applications, Springer, 2016, pp. 1–27.
[143] J. Deng, A.C. Berg, L. Fei-Fei, Hierarchical semantic indexing for large scale image retrieval, in: Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 785–792.
[144] M. Tan, I.W. Tsang, L. Wang, Towards ultrahigh dimensional feature selection for big data, J. Mach. Learn. Res. 15 (2014) 1371–1429.
[145] K. Weinberger, A. Dasgupta, J. Langford, A. Smola, J. Attenberg, Feature hashing for large scale multitask learning, in: Proceedings of the 26th Annual International Conference on Machine Learning, 2009, pp. 1113–1120.
[146] T. Hastie, R. Tibshirani, Efficient quadratic regularization for expression arrays, Biostatistics 5 (2004) 329–340.
[147] L. Wang, Y. Wang, Q. Chang, Feature selection methods for big data bioinformatics: a survey from the search perspective, Methods 111 (2016) 21–31.
[148] N. Japkowicz, Learning from imbalanced data sets: a comparison of various

strategies, in: Proceedings of AAAI Workshop on Learning from Imbalanced Data Sets, 2000, pp. 10–15.
[149] W.J. Lin, J.J. Chen, Class-imbalanced classifiers for high-dimensional data, Brief. Bioinform. 14 (2013) 13–26.
[150] D.W. Opitz, Feature selection for ensembles, in: Proceedings of 16th National Conference on Artificial Intelligence, 1999, pp. 379–384.
[151] Y. Saeys, T. Abeel, Y. Van de Peer, Robust feature selection using ensemble feature selection techniques, in: Proceedings of Machine Learning and Knowledge Discovery in Databases, 2008, pp. 313–325.
[152] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. de Schaetzen, R. Duque, H. Bersini, A. Nowe, A survey on filter techniques for feature selection in gene expression microarray analysis, IEEE/ACM Trans. Comput. Biol. Bioinf. 9 (2012) 1106–1119.
[153] T.K. Ho, The random subspace method for constructing decision forests, IEEE Trans. Pattern Anal. Mach. Intell. 20 (1998) 832–844.
[154] H. Ahn, H. Moon, M.J. Fazzari, N. Lim, J.J. Chen, R.L. Kodell, Classification by ensembles from random partitions of high-dimensional data, Comput. Stat. Data Anal. 51 (2007) 6166–6179.
[155] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[156] K.W. De Bock, K. Coussement, D. Van den Poel, Ensemble classification based on generalized additive models, Comput. Stat. Data Anal. 54 (2010) 1535–1546.
[157] H. Liu, L. Liu, H. Zhang, Ensemble gene selection for cancer classification, Pattern Recognit. 43 (2010) 2763–2772.
[158] S.L. Wang, X.L. Li, J. Fang, Finding minimum gene subsets with heuristic breadth-first search algorithm for robust tumor classification, BMC Bioinform. 13 (2012) 178.
[159] L. Zhang, P.N. Suganthan, Random forests with ensemble of feature spaces, Pattern Recognit. 47 (2014) 3429–3437.
[160] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, Y. Saeys, Robust biomarker identification for cancer diagnosis with ensemble feature selection methods, Bioinformatics 26 (2010) 392–398.
[161] D. Álvarez-Estévez, N. Sánchez-Maroño, A. Alonso-Betanzos, V. Moret-Bonillo, Reducing dimensionality in a database of sleep EEG arousals, Expert Syst. Appl. 38 (2011) 7746–7754.
[162] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, An ensemble of filters and classifiers for microarray data classification, Pattern Recognit. 45 (2012) 531–539.
[163] S. Perkins, J. Theiler, Online feature selection using grafting, in: Proceedings of the 20th International Conference on Machine Learning (ICML), 2003, pp. 592–599.
[164] J. Zhou, D.P. Foster, R.A. Stine, L.H. Ungar, Streamwise feature selection, J. Mach. Learn. Res. 7 (2006) 1861–1885.
[165] X. Wu, K. Yu, H. Wang, W. Ding, Online streaming feature selection, in: Proceedings of the 27th International Conference on Machine Learning (ICML), 2010, pp. 1159–1166.
[166] X. Wu, K. Yu, W. Ding, H. Wang, X. Zhu, Online feature selection with streaming features, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013) 1178–1192.
[167] K. Yu, X. Wu, W. Ding, J. Pei, Towards scalable and accurate online feature selection for big data, in: Proceedings of IEEE International Conference on Data Mining (ICDM), 2014, pp. 660–669.
[168] P. Ruangkanokmas, T. Achalakul, K. Akkarajitsakul, Deep belief networks with feature selection for sentiment classification, in: Proceedings of the 7th International Conference on Intelligent Systems, Modelling and Simulation, 2016.
[169] Y. Li, C.Y. Chen, W.W. Wasserman, Deep feature selection: theory and application to identify enhancers and promoters, in: Proceedings of International Conference on Research in Computational Molecular Biology, 2015, pp. 205–217.
[170] V. Singh, N. Baranwal, R.K. Sevakula, N.K. Verma, Y. Cui, Layerwise feature selection in stacked sparse auto-encoder for tumor type prediction, in: Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2016, pp. 1542–1548.
[171] A. Antoniades, C.C. Took, Speeding up feature selection: a deep-inspired network pruning algorithm, in: Proceedings of International Joint Conference on Neural Networks (IJCNN), 2016, pp. 360–366.
[172] Q. Zou, L. Ni, T. Zhang, Q. Wang, Deep learning based feature selection for remote sensing scene classification, IEEE Geosci. Remote Sens. Lett. 12 (2015) 2321–2325.

Jie Cai received the Master degree in computer science from Hunan University and is currently a Ph.D. candidate in College of Computer Science and Electronic Engineering, Hunan university. Her research interests include data mining and computational biology.

Jiawei Luo received the Ph.D. degree in computer science from Hunan University in 2008. She is currently a professor in College of Computer Science and Electronic Engineering, Hunan University. She has published about 50 research papers in various international journals and proceedings of conferences. Her research interests include graph theory, data mining, computational biology, and bioinformatics.

Shulin Wang received the B.Sc. degree in computer application from China University of Geosciences in 1989, the M.Sc. degree in computer application from the National University of Defense Technology, China, in 1997, and the Ph.D. degree in computer science and technology from the National University of Defense Technology, China, in 2008 (Advisor: Prof. Huowang Chen, Academician, and Prof. Ji Wang). Currently, he is working at Hunan University as a professor. His current research interests include bioinformatics, software engineering, and complex system.

Sheng Yang received the Ph.D. degree in pattern recognition and intelligent system from Shanghai Jiaotong University in 2005. Now, he is working at Hunan University as an associate professor. His current research interests include feature selection, data mining and machine learning.