Semi-supervised Feature Selection Based on Cost-Sensitive and Structural Information

Yiling Tao, Guangquan Lu(✉), Chaoqun Ma, Zidong Su, and Zehui Hu

Guangxi Key Lab of Multi-Source Information Mining and Security,
Guangxi Normal University, Guilin, Guangxi, China
lugq@mailbox.gxnu.edu.cn

Abstract. Feature selection is an important process of high-dimensional data analysis in data mining and machine learning. In the feature selection stage, the cost of misclassification and the structural information of paired samples on each feature dimension are often ignored. To overcome this, we propose semi-supervised feature selection based on cost-sensitive and structural information. First, cost-sensitive learning is incorporated into the semi-supervised framework. Second, the structural information between a pair of samples in each feature dimension is encapsulated into the feature graph. Finally, the correlation between the candidate feature and the target feature is added, which avoids the misunderstanding of the feature with low correlation as the salient feature. Furthermore, the proposed method also considers the redundancy between feature pairs, which can improve the accuracy of feature selection. The proposed method is more interpretable and practical than previous semi-supervised feature selection algorithms, because it considers the misclassification cost, structural relationship and the correlations between features and target features. Experimental results show that the promising performance of the proposed method outperforms the state-of-the-arts on eight data sets.

Keywords: Feature selection · Cost-sensitive · Structural relationship · Semi-supervised

1 Introduction

Big data has widely appeared in various fields, such as pattern recognition and machine learning [1,2]. A common problem in data processing is that the data often contains some unimportant features [3,4], which will increase the calculation cost and affect the effectiveness of model training [5,6]. Therefore, feature selection has become one of the important research fields of machine learning in recent years.

Feature selection is used to delete redundant features for conducting dimensionality reduction [1], which can help model training and reduce the impact of “dimension disaster” [7]. Depending on the availability of sample labels, feature

M. Qiao et al. (Eds.): ADC 2021, LNCS 12610, pp. 23–36, 2021.
https://doi.org/10.1007/978-3-030-69377-0_3

selection is divided into supervised, semi-supervised and unsupervised. Supervised feature selection [8] only uses labeled samples to train the model, and takes the structural relationship between labels and features to choose the important features, so as to explores the result of feature subset with the highest relevance to the label. Unsupervised feature selection [9–11] uses unlabeled sample training model, which selects the most representative features from the original feature set according to certain evaluation criteria. Semi-supervised feature selection [12,13] uses a small number of labeled samples and a lot of unlabeled samples to achieve the optimal feature subset. These types of methods are efficient, because they can not only mine the global and local structure of all samples, but also utilize the small number of labels that providing category information. Therefore, this paper focus on research on semi-supervised feature selection.

Various semi-supervised feature selection methods have been proposed recently. For example, the typical feature selection based on classifier [14] (semi-supervised support vector machine, S3VM), it uses support vector machine (SVM) to tag no label samples, and then fused the “soft” label samples for model training. Zhao and Liu [15] proposed a semi-supervised regularized feature selection framework based on spectral learning to evaluate the correlation of features. In addition, Ren et al. [16] proposed a forward semi-supervised feature selection framework based on wrapper type, which combines forward selection with wrapper type to obtain the optimal feature subset. Chen et al. [17] combined the traditional fisher-score method to obtain the global optimal feature subset by the “soft” label of unlabeled samples with label propagation technology.

However, the existing semi-supervised feature selection methods have some defects. First, many semi-supervised feature selection researches focus on the lowest classification error rate without considering the misclassification cost. It has assumed that different misclassifications owning the equal costs [18], which may lead the model pays attention to samples which causes high misclassification losses, resulting in biases in the features selected by the learning model. Second, some advanced semi-supervised feature selection algorithms do not consider the structural information of the paired samples in each feature dimension, which can improve the performance of feature selection [19]. In addition, researchers believe that the correlation of a single candidate features is equal to the correlation of selected features, without considering the joint correlation of a pair of features, which will regard low-relevance features as salient features. Therefore, some low-correlation features are regarded as salient features.

To solve the above problems, we propose semi-supervised feature selection based on cost-sensitive and structural information (SF_CSSI). The contributions of this paper are as follows:

- In practical applications, misclassification has always existed, however, the cost of misclassification is always ignored by researchers. The proposed method considers the misclassification cost and sets different penalty costs for different categories samples. In contrast to conventional feature selection methods, we try to minimize the total cost rather than the total error rate, aiming to prevent disasters caused by mistakes with high costs.

- In this paper, the proposed method converts each original feature vector into a structure-based feature graph representation, which contains structural information between sample pairs in each feature dimension, in order to preserve more meaningful information. Furthermore, the proposed method constructs feature information matrix to simultaneously maximize joint relevancy of different pairwise feature combinations in relation to the target feature graphs and minimize redundancy among selected features, so as to obtain feature subset with high correlation and low redundancy.
- The method proposed in this paper has rarely been studied, because it considers misclassification cost, structural information and information measurement of paired features. Experiments prove that the proposed method in this paper can achieve better feature selection results on real datasets.

2 Approach

2.1 Notations

In this paper, matrices are written as boldface uppercase letters, vectors are written as boldface lowercase letters and scalars are written as normal italic letters. For matrix $\mathbf{X}$, $x_{i,j}$ represents the element in the i-th row and j-th column of $\mathbf{X}$. The Frobenius norm of matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ is defined as $\|\mathbf{X}\|_F = \sqrt{\sum_{i,j} x_{i,j}^2}$. The $l_{2,1}$-norm of matrix $\mathbf{X}$ is defined as $\|\mathbf{X}\|_{2,1} = \sum_{i=1}^{n}\sqrt{\sum_{j=1}^{d} x_{ij}^2}$. For vector $\mathbf{x}$, its l_1-norm is defined $\|\mathbf{x}\|_1 = \sum_{i=1}^{n}|x_i|$. The symbol $\odot$ denotes multiplication of corresponding elements and $tr(\mathbf{X})$ represents the trace of matrix $\mathbf{X}$.

In semi-supervised learning, the data set consists of two parts: labeled data $\mathbf{X}_L = (x_1, x_2, \ldots, x_l)$ and unlabeled data $\mathbf{X}_U = (x_{l+1}, x_{l+2}, \ldots, x_{l+u})$, $u = n - l$, n represents the number of samples, l represents the number of labeled samples, u represents the number of unlabeled samples. The corresponding labels is $\mathbf{Y}_L = (y_1, y_2, \ldots, y_l)^T$ and the label of $\mathbf{Y}_U = (y_{l+1}, y_{l+2}, \ldots, y_{l+u})^T$ is unknown.

2.2 Cost-Sensitive Feature Selection

Given data set $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{n\times d}$, n represents the number of samples, d represents the features of each sample. The traditional feature selection imposes a sparsity penalty in the objective function, which makes the selected features more sparse and more discriminative. The objective function of traditional feature selection [9] is defined as:

$$\min_{\mathbf{W}} \|\mathbf{Y} - \mathbf{X}\mathbf{W}\|_F^2 + \lambda\|\mathbf{W}\|_{2,1} \tag{1}$$

However, cost-sensitive learning is embedded into feature selection framework because the misclassification problem often occurs in practical applications. Cost-sensitive learning assigns different cost parameters to different types of samples, without loss of generality, so the specified cost matrix is introduced into the feature selection framework. The traditional cost-sensitive feature selection objective function [20] is defined as:

$$\min_{\mathbf{W}} \left\|\left(\mathbf{X}^T\mathbf{W} - \mathbf{Y}\right) \odot \mathbf{C}\right\|_{2,1} + \lambda \|\mathbf{W}\|_{2,1}, \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{d\times m}$ represents the feature weight matrix, $\mathbf{Y} \in \mathbb{R}^{n\times m}$ represents labels, $\mathbf{C} \in \mathbb{R}^{n\times m}$ represents cost matrix, λ represents the penalty coefficient.

2.3 Feature Selection with Graph Structural Information

The structural information can provide more abundant representation but few researchers pay attention to these between the features in each pairs of samples.

Therefore, each feature vector is transformed into a feature graph structure, which encapsulates the pairwise relationship between samples. In addition, the information theory criterion of Jensen-Shannon divergence is used to measure the joint correlation between different paired feature combinations and target labels. The specific process is as follows.

Let $\mathbf{X} = \{\mathbf{f}_1, \ldots, \mathbf{f}_i, \ldots, \mathbf{f}_N\} \in \mathbb{R}^{M\times N}$ represents a data set of M samples and N features. Each original feature vector $\mathbf{f}_i = (f_{i1}, \ldots, f_{ia}, \ldots, f_{ib}, \ldots, f_{iM})^T$is transformed into a feature graph $\mathbf{G}_i\,(V_i, E_i)$, where vertex $v_{ia} \in V_i$ represents the a-th sample f_{ia} in feature $\mathbf{f}_i$ (i.e., each vertex represents a sample), edge $(v_{ia}, v_{ib}) \in E_i$ represents the weight of the a-th sample and the b-th sample (i.e., the edge represents the correlation between a pair of samples in the corresponding feature dimension). In addition, we also construct a graph structure for the target feature $\mathbf{Y}$. For classification problems, $\mathbf{Y}$ are discrete value $c \in \{1, 2, \ldots, m\}$. Therefore, we calculate the continuous value of each discrete target feature $\mathbf{f}_i$ as $\overset{\wedge}{\mathbf{f}}_i = \left(\hat{f}_{i1}, \ldots, \hat{f}_{ia}, \ldots, \hat{f}_{ib}, \ldots, \hat{f}_{iM}\right)^T$, $\hat{f}_{ia}$ represents the a-th sample in $\overset{\wedge}{\mathbf{f}}_i$. When the f_{ia} in $\mathbf{f}_i$ belongs to class m, $\hat{f}_{ia}$ is the mean value of all class m samples in $\mathbf{f}_i$. Similarly, we construct the graph structure of the target feature $\overset{\wedge}{\mathbf{f}}_i$ as $\overset{\wedge}{\mathbf{G}}_i\left(\hat{V}_i, \hat{E}_i\right)$. $\hat{v}_{ia} \in \hat{V}_i$ represents the a-th sample in target feature $\overset{\wedge}{\mathbf{f}}_i$, $\left(\hat{v}_{ia}, \hat{v}_{ib}\right) \in \hat{E}_i$ is the weighted edge connecting the a-th sample and the b-th sample of $\overset{\wedge}{\mathbf{f}}_i$. This paper uses Euclidean distance to calculate the relationship between pairs of feature samples, that is, the weight of f_{ia} and f_{ib} can be expressed as:

$$\omega\,(v_{ia}, v_{ib}) = \sqrt{\left(f_{ia} - f_{ib}\right)^2} \tag{3}$$

Similiarly, the weight of edge $\left(\hat{v}_{ia}, \hat{v}_{ib}\right) \in \hat{E}_i$ in $\overset{\wedge}{\mathbf{G}}_i\left(\hat{V}_i, \hat{E}_i\right)$ is expressed as follows:

$$\omega\left(\hat{v}_{ia}, \hat{v}_{ib}\right) = \sqrt{\left(\mu_{ia} - \mu_{ib}\right)^2}, \tag{4}$$

where μ_{ia} is the mean value of all samples in $\mathbf{f}_i$ from the same class m.

Jensen Shannon divergence (JSD) is used to measure the divergence between two probability distributions [21]. Give two (discrete) probability distributions $\mathcal{P} = (p_1, \ldots, p_a, \ldots p_A)$ and $\mathcal{K} = (k_1, \ldots, k_b, \ldots k_B)$. The JSD between $\mathcal{P}$ and $\mathcal{K}$ is defined as:

$$D_{\mathrm{JS}}(\mathcal{P},\mathcal{K}) = H_S\left(\frac{\mathcal{P}+\mathcal{K}}{2}\right) - \frac{1}{2}H_S(\mathcal{P}) - \frac{1}{2}H_S(\mathcal{K}), \tag{5}$$

where $H_S(\mathcal{P}) = \sum_{i=1}^{A} p_i \log p_i$ is the Shannon entropy of probability distribution $\mathcal{P}$. In the literature [22], the JSD has been used as a means of measuring the information theoretic dissimilarity between graphs associated with their probability distributions. In this paper, we focus on the similarity between graph-based feature representations. We use the negative exponent of $D_{\mathrm{JS}}(\mathcal{P},\mathcal{K})$ to calculate the similarity I_S between probability distributions $\mathcal{P}$ and $\mathcal{K}$, so:

$$I_S(\mathcal{P},\mathcal{K}) = \exp\{-D_{\mathrm{JS}}(\mathcal{P},\mathcal{K})\} \tag{6}$$

The information theoretic function is used to evaluate the relevance between different feature combination and target labels to achieve the maximum correlation and minimum redundancy standards. For a set of N features $\mathbf{f}_1, \ldots, \mathbf{f}_i, \ldots, \mathbf{f}_N$ and related continuous target feature $\mathbf{Y}$, the correlation degree of feature pair $\{\mathbf{f}_i, \mathbf{f}_j\}$ is expressed as follows:

$$U_{f_i,f_j} = \frac{I_s\left(\mathbf{G}_i, \overset{\wedge}{\mathbf{G}}\right) + I_s\left(\mathbf{G}_j, \overset{\wedge}{\mathbf{G}}\right)}{I_s(\mathbf{G}_i,\mathbf{G}_j)}, \tag{7}$$

where I_s is the JSD based similarity measure of information theory defined in Eq. 6. $I_s\left(\mathbf{G}_i, \overset{\wedge}{\mathbf{G}}\right)$ represents the correlation measures of feature $\mathbf{f}_i$ with target feature $\mathbf{Y}$. $I_s\left(\mathbf{G}_j, \overset{\wedge}{\mathbf{G}}\right)$ represents the correlation measures of feature $\mathbf{f}_j$ with target feature $\mathbf{Y}$. $I_s(\mathbf{G}_i,\mathbf{G}_j)$ denotes the redundancy of paired feature $\{\mathbf{f}_i, \mathbf{f}_j\}$. Therefore, U_{f_i,f_j} is large if and only if $I_s\left(\mathbf{G}_i, \overset{\wedge}{\mathbf{G}}\right) + I_s\left(\mathbf{G}_j, \overset{\wedge}{\mathbf{G}}\right)$ is large and $I_s(\mathbf{G}_i,\mathbf{G}_j)$ is small. This indicates that the pairwise feature $\{\mathbf{f}_i, \mathbf{f}_j\}$ is informative and less redundant.

Given the feature information matrix $\mathbf{U}$ and d-dimensional feature vector $\mathbf{w}$. The feature subset is identified by solving the maximization problem of the following formula:

$$\max f(\mathbf{w}) = \max_{\mathbf{w}\in\mathbb{R}^d} \mathbf{w}^T\mathbf{U}\mathbf{w}, \tag{8}$$

where $\mathbf{w} \in \mathbb{R}^d$, $\mathbf{w} = (w_1, w_2, \cdots, w_i, \cdots, w_n)^T$, $w_i > 0$, w_i represents the correlation coefficient of the i-th feature.

2.4 Mathematical Formulation

The purpose of our proposed method is to improve the performance of feature selection through structural information and misclassification costs when the data does not have a large number of labels. Therefore, we combine cost-sensitive and Eq. 8 to propose semi-supervised feature selection based on cost-sensitive and structural information. The specific mathematical expression is as follows:

$$\min_{\mathbf{w}} \alpha_1 tr(\mathbf{w}^T\mathbf{X}^T\mathbf{L}\mathbf{X}\mathbf{w}) + \sum_{i=1}^{l} \|x_i\mathbf{w} - y_i\|_2^2 c_i + \alpha_2||\mathbf{w}||_1 - \alpha_3\mathbf{w}^T\mathbf{U}\mathbf{w} \tag{9}$$

The first term represents the learning of local proximity structure, which helps the model to select a representative feature subset by maintaining the local structure of the samples. $\mathbf{w}$ represents the feature coefficient vector, $\mathbf{L}$ is the Laplacian matrix, $\mathbf{L} = \mathbf{D} - \mathbf{A}$, where $\mathbf{D}$ is a diagonal matrix, the diagonal element satisfies $D_{ii} = \sum_{j=1}^{n} A_{ij}$ and $\mathbf{A}$ is the affinity matrix, if $i \neq j$, $A_{ij} = \exp(-\frac{||x_i - x_j||_2^2}{2\sigma^2})$; otherwise, $A_{ij} = 0$. The cost c_i represents the cost, the second term indicates that the loss of the original feature is combined with the cost to obtain the misclassification cost loss. Since we judge the misclassification result based on the label sample, we only use the labeled sample to calculate the misclassification loss. The third term $\|\mathbf{w}\|_1$ represents the sparse regular term, which uses the l_1-norm to shrink some coefficients to zero. The fourth term encourages the selected features to be jointly more relevant with the target while maintaining less redundancy between features, α_1, α_2 and α_3 are the penalty coefficients.

2.5 Optimization

In order to optimize, Eq. 9 can be rewritten as follows:

$$\begin{aligned} \min_{\mathbf{w},\mathbf{Q}} \alpha_1 tr\left(\mathbf{w}^T\mathbf{X}^T\mathbf{L}\mathbf{X}\mathbf{w}\right) + tr\left(\left(\mathbf{X}_L\mathbf{w} - \mathbf{Y}_L\right)^T\mathbf{C}\left(\mathbf{X}_L\mathbf{w} - \mathbf{Y}_L\right)\right) \\ + \alpha_2 tr\left(\mathbf{w}^T\mathbf{Q}\mathbf{w}\right) - \alpha_3\mathbf{w}^T\mathbf{U}\mathbf{w}, \end{aligned} \tag{10}$$

where $\mathbf{Q}$ is the diagonal matrix. We use the idea of iterative learning to optimize the objective function, that is, update $\mathbf{w}$ by fixing $\mathbf{Q}$ and update $\mathbf{Q}$ by fixing $\mathbf{w}$, until Eq. 9 converges, so that the optimal solution of weight vector $\mathbf{w}$ can be obtained.

- Update $\mathbf{w}$ by fixing $\mathbf{Q}$

  When $\mathbf{Q}$ is fixed, Eq. 10 can be regarded as a function of $\mathbf{w}$:

$$\begin{aligned} L\left(\mathbf{w}\right) = \alpha_1 tr\left(\mathbf{w}^T\mathbf{X}^T\mathbf{L}\mathbf{X}\mathbf{w}\right) + tr\left(\left(\mathbf{X}_L\mathbf{w} - \mathbf{Y}_L\right)^T\mathbf{C}\left(\mathbf{X}_L\mathbf{w} - \mathbf{Y}_L\right)\right) \\ + \alpha_2 tr\left(\mathbf{w}^T\mathbf{Q}\mathbf{w}\right) - \alpha_3\mathbf{w}^T\mathbf{U}\mathbf{w} \end{aligned} \tag{11}$$

  We take the derivative of $\mathbf{w}$ in Eq. 11 and make it equal to zero:

$$\frac{\partial L}{\partial \mathbf{w}} = 2\alpha_1\mathbf{X}^T\mathbf{L}\mathbf{X}\mathbf{w} + 2\mathbf{X}_L{}^T\mathbf{C}\mathbf{X}_L\mathbf{w} - 2\mathbf{X}_L{}^T\mathbf{C}\mathbf{Y}_L + 2\alpha_2\mathbf{Q}\mathbf{w} - 2\alpha_3\mathbf{U}\mathbf{w} = 0 \tag{12}$$

According to Eq. 12, it is solved as follows:

$$\mathbf{w} = \left(2\alpha_1\mathbf{X}^T\mathbf{L}\mathbf{X} + 2{\mathbf{X}_L}^T\mathbf{C}\mathbf{X}_L + 2\alpha_2\mathbf{Q} - 2\alpha_3\mathbf{U}\right)^{-1} 2{\mathbf{X}_L}^T\mathbf{C}\mathbf{Y}_L \tag{13}$$

– Update **Q** by fixing **w**

When **w** is fixed, Eq. 10 can be regarded as:

$$\min_{\mathbf{Q}} \alpha_2 tr\left(\mathbf{w}^T\mathbf{Q}\mathbf{w}\right) \tag{14}$$

By setting the partial derivative of the above function with respect to **Q** as 0 and according to the article [23], it is solved as follows:

$$Q_{ii} = \frac{1}{2\left|w_i\right|}, \tag{15}$$

where **Q** is a diagonal matrix and $Q_{ii} = \frac{1}{2|w_i|}$ is the diagonal element.

Algorithm 1: The pseudo code of solving Eq. 9

```
Input:  Data matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, labeled data $\mathbf{X}_L \in \mathbb{R}^{l\times d}$, labels $\mathbf{Y}_L \in \mathbb{R}^{l}$, cost matrix $\mathbf{C} \in \mathbb{R}^{l\times l}$,
        control parameters $\alpha_1, \alpha_2, \alpha_3$;
Output:  $\mathbf{w} \in \mathbb{R}^d$;
1. Initialize $t = 0$ and $\mathbf{Q}^{(0)}$;
2. Build affinity matrix $\mathbf{A}$, $A_{ij} = \exp(-\frac{||x_i - x_j||_2^2}{2\sigma^2})$ or $A_{ij} = 0$;
3. Build diagonal matrix $\mathbf{D}$, $D_{ii} = \sum_{j=1}^{n} A_{ij}$;
4. Build Laplacian matrix $\mathbf{L}$, $\mathbf{L} = \mathbf{D} - \mathbf{A}$;
5. repeat:
  5.1 Update $\mathbf{w}^{(t+1)}$via Eq. 13;
  5.2 Update $\mathbf{Q}^{(t+1)}$ via Eq. 15;
  5.3 $t = t + 1$;
until converges;
```

2.6 Convergence Analysis

Let **w** and **Q** be $\mathbf{w}^{(\mathrm{t})}$ and $\mathbf{Q}^{(\mathrm{t})}$ in the t-th iteration, and Eq. 10 can be rewritten as:

$$\begin{aligned} E\left(\mathbf{w}^{(t)},\mathbf{Q}^{(t)}\right) = {} & \alpha_1 tr\left(\left(\mathbf{w}^{(t)}\right)^T\mathbf{X}^T\mathbf{L}\mathbf{X}\mathbf{w}^{(t)}\right) + tr\left(\left(\mathbf{X}_L\mathbf{w}^{(t)} - \mathbf{Y}_L\right)^T\mathbf{C}\left(\mathbf{X}_L\mathbf{w}^{(t)} - \mathbf{Y}_L\right)\right) \\ & + \alpha_2 tr\left(\left(\mathbf{w}^{(t)}\right)^T\mathbf{Q}^{(t)}\mathbf{w}^{(t)}\right) - \alpha_3\left(\mathbf{w}^{(t)}\right)^T\mathbf{U}\mathbf{w}^{(t)} \end{aligned} \tag{16}$$

Because the objective function $E\left(\mathbf{w}^{(t)},\mathbf{Q}^{(t)}\right)$ is a convex optimization problem about **w**, we have the following inequality:

$$E\left(\mathbf{w}^{(t+1)},\mathbf{Q}^{(t)}\right) \leq E\left(\mathbf{w}^{(t)},\mathbf{Q}^{(t)}\right) \tag{17}$$

According to the article [23], we know that Eq. 14 is convergent, so we can deduce that Eq. 10 is convergent about **Q**, so we express the convergence as the following inequality:

$$E\left(\mathbf{w}^{(t+1)},\mathbf{Q}^{(t+1)}\right) \leq E\left(\mathbf{w}^{(t+1)},\mathbf{Q}^{(t)}\right) \tag{18}$$

Combining Eq. 17 and Eq. 18, we can get the inequality:

$$E\left(\mathbf{w}^{(t+1)},\mathbf{Q}^{(t+1)}\right) \leq E\left(\mathbf{w}^{(t)},\mathbf{Q}^{(t)}\right) \tag{19}$$

Equation 16 is non-increasing at each iteration according to Eq. 19. Therefore, the proposed Algorithm 1 is convergence.

3 Experiments

In this section, we evaluated our proposed SF_CSSI and other six comparison methods on eight data sets. Specially, we first employed each feature selection method to choose the new feature subsets from original data sets, and then utilized support vector machine classification to evaluate the selected subsets.

3.1 Datasets and Comparison Methods

The data sets (i.e., madelon, SECOM, chess, isolet, Hill-with, Hill-without, musk and sonar) are from UCI Machine Learning Repository[1]. We summarized the detail of all data sets in Table 1.

Table 1. Summarization of data sets.

Datasets	Samples	Features	Classes
madelon	2000	500	2
SECOM	1967	590	2
chess	3196	36	2
isolet	1560	617	2
Hill-with	606	100	2
Hill-without	606	100	2
musk	486	166	2
sonar	208	360	2

We compared our proposed method with six comparison methods and the details of them are listed as follow:

- Cost-Sensitive Laplacian Score (CSLS [24]) uses Laplacian graphs and the cost of misclassification between classes to score each feature individually.

[1] http://archive.ics.uci.edu/ml/.

- Semi-supervised feature selection based on joint mutual information (Semi-JMI [25]) uses the redundancy between features and the correlation between features and labels to complete feature selection.
- Semi-supervised feature selection based on information theory method (Semi-IMIM [25]) only uses the correlation between features and labels to complete feature selection.
- Cost-Sensitive Feature Selection via F-Measure Optimization Reduction (CSFS [20]) introduces cost sensitivity to select features, which optimizes F-measure instead of accuracy to take class imbalance issue into account.
- Cost-sensitive feature selection via the $l_{2,1}$-norm (CSEFS [26]) combines $l_{2,1}$-norm minimization regularization and loss term of embedding misclassification cost to select feature subset.
- Semi-supervised Feature Selection via Rescaled Linear Regression (RLSR [17]) uses a set of scale factors to adjust regression coefficients, then uses regression coefficients to rank features.

3.2 Experimental Settings

The experiment of this paper is implemented with the MATLAB 2018a under Windows 10 system. Referring to [27] article's method, we can divide the data set into three parts: labeled sample set (L), unlabeled sample set (U), and test sample set (T). For each of data sets, the labeled samples were randomly selected with the given ratio $\{10\%, 20\%, 30\%\}$.

We use 10-fold cross-validation to generate training sample set and test sample set, then randomly select L and U from the training sample set for training, and finally use T to test the performance of different methods. All algorithms perform 10 times 10-fold cross-validation and take the average of the 10 experimental results as the final total cost, which reduce the accidental occurrence. We set the parameters α_1, α_2 and α_3 in Eq. 9 in range of $\left\{10^{-3}, 10^{-1}, ..., 10^{1}, 10^{3}\right\}$. For other comparison methods, we set these according to their corresponding literature.

Table 2. Total cost (cost ± std) of misclassification on eight data sets. Bold numbers indicate the best results.

Cost	Data sets	CSLS	Semi-JMI	Semi-IMIM	CSEFS	CSFS	RLSR	Proposed
$cost_1 = 10$ $cost_2 = 25$	madelon	1399.40 ± 17.57	1396.65 ± 20.88	1394.35 ± 19.39	1540.60 ± 74.59	1462.65 ± 19.22	1405.23 ± 23.38	$\mathbf{1369.56} \pm 12.96$
	SECOM	273.65 ± 5.87	270.26 ± 5.51	269.02 ± 5.19	268.03 ± 5.41	268.88 ± 5.70	269.01 ± 6.60	$\mathbf{266.09} \pm 4.41$
	chess	2494.00 ± 9.98	2349.15 ± 20.86	2624.30 ± 12.03	2485.75 ± 15.27	2726.00 ± 24.67	2074.65 ± 10.29	$\mathbf{515.00} \pm 0.00$
	isolet	525.50 ± 19.97	553.00 ± 22.84	538.90 ± 22.34	413.20 ± 19.60	352.05 ± 28.52	349.35 ± 15.91	$\mathbf{346.60} \pm 10.53$
	Hill-with	167.50 ± 18.73	177.38 ± 18.68	147.05 ± 63.65	157.35 ± 23.30	105.10 ± 17.08	102.75 ± 21.07	$\mathbf{100.40} \pm 15.60$
	Hill-without	1.05 ± 0.72	141.40 ± 52.04	168.30 ± 1.35	1.35 ± 1.22	2.50 ± 5.52	4.05 ± 5.90	$\mathbf{0.65} \pm 0.89$
	musk	165.20 ± 7.77	133.10 ± 7.94	145.75 ± 10.33	151.50 ± 11.77	132.30 ± 11.33	141.00 ± 17.88	$\mathbf{130.00} \pm 6.10$
	Sonar	99.20 ± 11.40	94.20 ± 6.17	93.95 ± 6.96	93.55 ± 7.26	99.00 ± 9.52	98.20 ± 8.79	$\mathbf{89.35} \pm 4.77$
$cost_1 = 25$ $cost_2 = 10$	madelon	1675.57 ± 14.98	1678.53 ± 19.80	1652.81 ± 12.71	1536.18 ± 25.85	1577.72 ± 23.29	1526.05 ± 18.59	$\mathbf{1510.15} \pm 13.54$
	SECOM	168.90 ± 27.01	161.10 ± 26.70	165.75 ± 30.30	163.45 ± 22.17	167.45 ± 21.50	$\mathbf{154.80} \pm 19.11$	158.40 ± 20.37
	chess	1193.80 ± 4.52	964.65 ± 8.23	1060.85 ± 7.13	1222.60 ± 34.73	1175.15 ± 7.43	$\mathbf{921.00} \pm 7.32$	1222.60 ± 34.73
	isolet	350.35 ± 11.60	432.90 ± 15.18	435.85 ± 12.80	334.40 ± 11.16	331.35 ± 8.08	321.85 ± 6.33	$\mathbf{318.76} \pm 5.73$
	Hill-with	533.10 ± 32.68	499.20 ± 54.94	464.15 ± 42.81	482.80 ± 53.24	482.80 ± 53.34	454.45 ± 68.49	$\mathbf{411.95} \pm 43.52$
	Hill-without	1115.20 ± 34.18	428.10 ± 55.74	427.26 ± 53.62	43.60 ± 32.47	2.35 ± 1.98	9.30 ± 1.99	$\mathbf{1.60} \pm 2.50$
	musk	150.15 ± 7.01	129.50 ± 9.66	133.95 ± 6.83	137.40 ± 9.81	128.95 ± 11.14	129.75 ± 12.04	$\mathbf{128.65} \pm 11.09$
	Sonar	168.90 ± 27.01	161.10 ± 26.70	165.75 ± 30.30	163.45 ± 22.17	167.45 ± 21.50	154.80 ± 19.11	$\mathbf{153.40} \pm 20.37$

Table 3. The value of specificity on eight data sets. Bold numbers indicate the best results.

Cost	Data sets	CSLS	Semi-JMI	Semi-IMIM	CSEFS	CSFS	RLSR	Proposed
$cost_1 = 10$ $cost_2 = 25$	madelon	61.34 ± 0.70	60.98 ± 0.86	60.41 ± 0.96	56.31 ± 0.18	58.19 ± 0.62	57.99 ± 0.38	$\mathbf{63.93} \pm 1.96$
	SECOM	98.30 ± 0.23	98.90 ± 0.13	98.91 ± 0.14	97.88 ± 0.10	93.65 ± 0.39	93.61 ± 0.39	$\mathbf{99.59} \pm 0.12$
	chess	1.04 ± 0.08	26.70 ± 2.03	78.81 ± 2.50	18.23 ± 0.41	1.12 ± 0.93	26.01 ± 1.39	$\mathbf{90.83} \pm 0.02$
	isolet	80.51 ± 0.73	79.68 ± 1.10	80.39 ± 0.96	84.80 ± 1.03	86.00 ± 1.35	86.61 ± 1.26	$\mathbf{86.99} \pm 0.54$
	Hill-with	89.09 ± 2.45	86.02 ± 2.22	88.85 ± 2.25	89.04 ± 2.95	89.83 ± 2.48	90.22 ± 2.76	$\mathbf{93.12} \pm 4.20$
	Hill-without	99.96 ± 0.10	84.00 ± 1.51	81.78 ± 9.61	99.86 ± 0.10	99.68 ± 0.93	99.83 ± 0.22	$\mathbf{99.87} \pm 0.19$
	musk	81.27 ± 1.30	85.48 ± 1.14	83.87 ± 1.43	82.90 ± 1.70	85.39 ± 1.43	84.35 ± 2.39	$\mathbf{88.63} \pm 1.13$
	Sonar	83.46 ± 6.01	81.35 ± 6.57	82.50 ± 7.25	84.94 ± 5.33	86.06 ± 5.55	82.96 ± 5.84	$\mathbf{96.06} \pm 1.93$
$cost_1 = 25$ $cost_2 = 10$	madelon	56.78 ± 0.48	56.59 ± 1.00	56.18 ± 1.18	56.46 ± 0.83	56.47 ± 1.12	55.35 ± 0.77	$\mathbf{61.35} \pm 0.54$
	SECOM	98.86 ± 0.19	98.89 ± 0.22	98.93 ± 0.17	98.94 ± 0.22	99.86 ± 0.18	98.93 ± 0.17	$\mathbf{99.89} \pm 0.06$
	chess	38.39 ± 0.30	$\mathbf{48.04} \pm 0.43$	45.35 ± 0.46	44.26 ± 2.05	41.77 ± 1.22	46.97 ± 0.65	46.67 ± 1.12
	isolet	82.53 ± 0.44	78.44 ± 0.61	75.77 ± 6.46	80.76 ± 0.61	81.10 ± 0.64	82.04 ± 0.56	$\mathbf{83.17} \pm 0.26$
	Hill-with	82.29 ± 3.21	87.89 ± 2.25	87.71 ± 3.06	87.65 ± 3.21	87.65 ± 3.21	91.62 ± 2.32	$\mathbf{92.89} \pm 1.25$
	Hill-without	67.16 ± 1.32	36.61 ± 2.00	80.55 ± 2.20	77.77 ± 2.64	77.77 ± 2.64	79.31 ± 1.52	$\mathbf{82.46} \pm 7.30$
	musk	81.02 ± 1.32	84.72 ± 1.63	83.57 ± 1.52	83.01 ± 1.22	84.35 ± 1.39	84.72 ± 1.86	$\mathbf{85.40} \pm 1.88$
	Sonar	87.08 ± 3.28	87.10 ± 3.71	86.93 ± 3.70	87.06 ± 3.51	86.59 ± 3.66	85.67 ± 2.55	$\mathbf{87.23} \pm 4.98$

The total cost, specificity and sensitivity are used as evaluation indicators to evaluate the performance of all methods on eight data sets.

The total cost is calculated as follows:

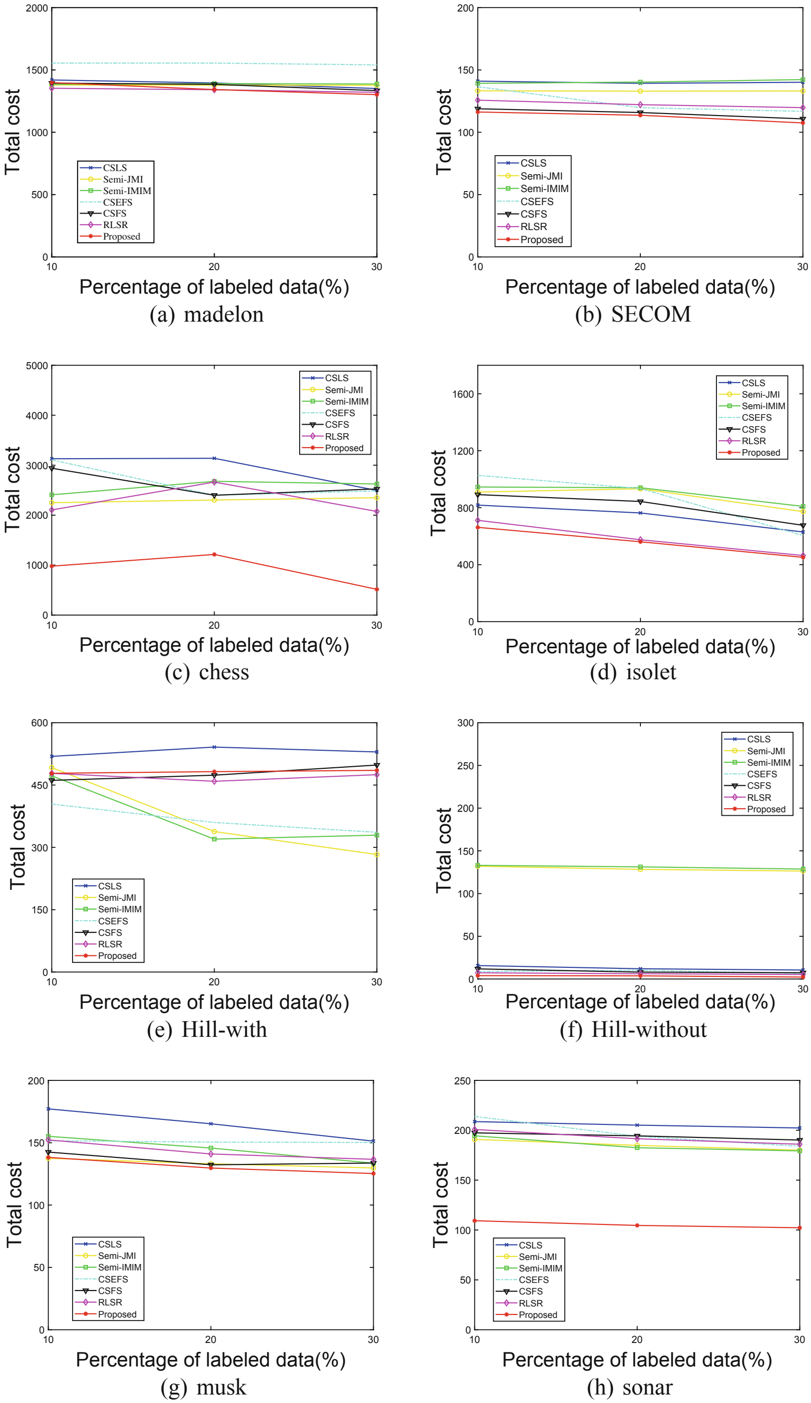
$$Total \quad Cost = \mathrm{sum}\left(c_i\right), \tag{20}$$

$$c_i = \begin{cases} cost_1 \quad \text{or} \quad \text{cost}_2 & , \text{predicted} \quad \text{label} \neq \text{true} \quad \text{label} \\ 0 & , \text{otherwise} \end{cases}, \tag{21}$$

where c_i represents the misclassification cost of a sample. If the predicted label is equal to the true label, the cost of c_i is 0, otherwise, the cost of c_i is equal

Table 4. The value of sensitivity on eight data sets. Bold numbers indicate the best results.

Cost	Data sets	CSLS	Semi-JMI	Semi-IMIM	CSEFS	CSFS	RLSR	Proposed
$cost_1 = 10$ $cost_2 = 25$	madelon	59.51 ± 0.68	59.93 ± 0.62	60.27 ± 0.54	55.42 ± 2.89	56.36 ± 1.22	55.32 ± 0.38	$\mathbf{61.56} \pm 0.96$
	SECOM	80.46 ± 3.01	81.35 ± 3.57	82.50 ± 3.25	84.94 ± 4.33	84.06 ± 3.55	82.96 ± 4.84	$\mathbf{86.06} \pm 0.30$
	chess	94.58 ± 0.13	$\mathbf{99.28} \pm 0.10$	84.12 ± 1.65	94.52 ± 0.15	97.87 ± 0.21	97.44 ± 0.16	90.10 ± 0.02
	isolet	81.33 ± 0.76	78.89 ± 0.48	79.97 ± 0.81	84.95 ± 1.01	89.81 ± 0.62	90.03 ± 0.37	$\mathbf{92.11} \pm 0.68$
	Hill-with	73.56 ± 1.68	77.81 ± 2.09	80.93 ± 1.06	77.01 ± 1.81	90.68 ± 2.82	90.92 ± 2.28	$\mathbf{91.84} \pm 1.45$
	Hill-without	99.15 ± 0.18	92.31 ± 4.24	90.10 ± 4.75	99.46 ± 0.22	99.49 ± 0.42	99.10 ± 0.42	$\mathbf{99.58} \pm 0.45$
	musk	81.19 ± 1.78	83.54 ± 1.58	81.54 ± 2.17	82.29 ± 1.33	83.22 ± 1.97	82.60 ± 2.69	$\mathbf{86.40} \pm 1.88$
	Sonar	38.38 ± 12.61	38.61 ± 11.28	3.91 ± 12.62	38.07 ± 11.91	38.49 ± 12.74	36.63 ± 13.26	$\mathbf{58.10} \pm 2.31$
$cost_1 = 25$ $cost_2 = 10$	madelon	56.36 ± 0.93	56.04 ± 0.89	56.11 ± 0.93	56.66 ± 1.45	57.30 ± 0.84	56.22 ± 1.08	$\mathbf{60.16} \pm 0.54$
	SECOM	85.08 ± 3.02	86.10 ± 3.71	86.93 ± 2.70	28.63 ± 2.04	87.06 ± 3.51	83.59 ± 3.10	$\mathbf{87.22} \pm 3.32$
	chess	98.25 ± 0.66	97.87 ± 0.25	98.35 ± 0.18	97.50 ± 0.88	98.78 ± 0.39	98.05 ± 0.22	$\mathbf{98.98} \pm 0.02$
	isolet	89.02 ± 0.51	86.49 ± 0.62	87.37 ± 0.59	90.36 ± 0.45	90.61 ± 0.41	90.71 ± 0.30	$\mathbf{90.96} \pm 0.18$
	Hill-with	67.16 ± 1.32	76.61 ± 2.00	80.55 ± 2.20	77.77 ± 2.64	74.77 ± 2.65	79.31 ± 1.52	$\mathbf{82.46} \pm 7.30$
	Hill-without	90.91 ± 3.13	58.55 ± 9.90	56.17 ± 9.71	96.26 ± 3.41	99.71 ± 0.26	98.90 ± 0.25	$\mathbf{99.86} \pm 0.12$
	musk	80.06 ± 1.50	83.00 ± 1.28	82.48 ± 1.28	82.29 ± 1.71	83.22 ± 1.97	80.87 ± 1.70	$\mathbf{84.23} \pm 1.27$
	Sonar	39.46 ± 12.24	39.85 ± 12.75	39.21 ± 11.71	38.86 ± 11.06	38.37 ± 9.50	44.99 ± 8.24	$\mathbf{46.92} \pm 11.92$



(a) madelon

(b) SECOM

(c) chess

(d) isolet

(e) Hill-with

(f) Hill-without

(g) musk

(h) sonar

Fig. 1. The total cost of different methods under different labeled samples, at eight data sets while $cost_1 = 10$, $cost_2 = 25$.

to $cost_1$ or $cost_2$ ($cost_1$ and $cost_2$ represent the costs of being judged as positive and negative samples, respectively).

For a binary classification, there are four possible results: TP (True Positive) is positive instances correctly classified and TN (True Negative) is negative instances correctly classified. FP (False Positive) is negative instances incorrectly classified and FN (False Negative) is positive instances misclassified.

Specificity refers to the proportion of samples that are actually negative which are judged to be negative. It can be calculated by the following formula:

$$specificity = \frac{TN}{FP + TN} \tag{22}$$

Sensitivity refers to the proportion of samples that are actually positive which are judged to be positive. It can be calculated by the following formula:

$$sensitivity = \frac{TP}{TP + FN} \tag{23}$$

3.3 Experiment Results and Analysis

In this experiment, we reported the cost, specificity and sensitivity of all methods on eight UCI datasets in Table 2, Table 3 and Table 4 under different cost value settings and listed our observations as follows. In addition, we use a line chart to show the changing trend of the total cost under different proportions of labeled samples. It can be seen from Fig. 1.

From Table 2, we can know that the proposed SF_CSSI method outperformed other methods on most cases. Especially, on the chess data set, the total cost of SF_CSSI has reduced by 75% compared with the second best approach Semi-JMI, when $cost_1 = 10$ and $cost_2 = 25$. When $cost_1 = 25$ and $cost_2 = 10$, 31% reduction was achieved by the proposed method SF_CSSI on the Hill-without data set, compared to the second best approach CSFS.

From Table 3 and Table 4, the proposed model has high specificity and sensitivity. The highest specificity was obtained on SECOM and Hill-without data sets. The highest sensitivity was obtained on isolet and Hill-without data sets compared with other methods. In addition, specificity and sensitivity are commonly used diagnostic methods in clinical practice. The higher the value is, the more real, reliable and practical the diagnosis result will be.

From Fig. 1, the more labeled data we have, the lower cost we can achieve, in most cases. We also notice that SF_CSSI outperformed other CSLS, CSFS and CSEFS methods on almost all cases, which indicates that CSLS, CSFS and CSEFS can be improved with unlabeled data. This verifies the effectiveness of the semi-supervised feature selection method. In addition, the proposed method has the minimum total cost on most cases, especially in Hill-without data set.

3.4 Conclusion

This paper considers the misclassification and the structural information of the paired samples on each feature dimension. In addition, the information theory method is used to introduce a feature information matrix to simultaneously maximize joint relevancy of different pairwise feature combinations in relation to the target feature graphs and minimize redundancy among selected features. Compared with previous research on semi-supervised feature selection, this paper comprehensively considers the cost of misclassification, the structure information of paired samples on the feature dimension, and the information relationship of paired features. In general, it is more interpretable and generalizable for our method than others in this paper. Experiments on 8 real data sets show that the proposed method has better feature selection results.

In future work, we will try to extend our method to conduct a cost-sensitive multi-class classification.

Acknowledgment. This work was supported by the National Natural Science Foundation of China (Grant No: 81701780); the Research Fund of Guangxi Key Lab of Multi-source Information Mining and Security (No. 20-A-01-01); the Guangxi Natural Science Foundation (Grant No: 2017GXNSFBA198221); the Project of Guangxi Science and Technology (GuiKeAD20159041,GuiKeAD19110133); the Innovation Project of Guangxi Graduate Education (Grants No: JXXYYJSCXXM-011).

References

1. Zhang, S., Li, X., Zong, M., Zhu, X., Wang, R.: Efficient KNN classification with different numbers of nearest neighbors. IEEE Trans. Neural Netw. Learn. Syst. **29**(5), 1774–1785 (2017)
2. Gao, L., Guo, Z., Zhang, H., Xu, X., Shen, H.T.: Video captioning with attention-based LSTM and semantic consistency. IEEE Trans. Multimed. **19**(9), 2045–2055 (2017)
3. Shen, H.T., et al.: Heterogeneous data fusion for predicting mild cognitive impairment conversion. Inf. Fusion **66**, 54–63 (2021)
4. Zhu, X., Song, B., Shi, F., Chen, Y., Shen, D.: Joint prediction and time estimation of COVID-19 developing severe symptoms using chest CT scan. Med. Image Anal. **67**, 101824 (2021)
5. Lei, C., Zhu, X.: Unsupervised feature selection via local structure learning and sparse learning. Multimed. Tools Appl. **77**(22), 2960–2962 (2018)
6. Zhu, X., Zhang, S., Hu, R., Zhu, Y., Song, J.: Local and global structure preservation for robust unsupervised spectral feature selection. IEEE Trans. Knowl. Data Eng. **30**(99), 517–529 (2018)
7. Zhu, X., Li, X., Zhang, S.: Block-row sparse multiview multilabel learning for image classification. IEEE Trans. Cybern. **46**(46), 450 (2016)
8. Wu, X., Xu, X., Liu, J., Wang, H., Nie, F.: Supervised feature selection with orthogonal regression and feature weighting. IEEE Trans. Neural Netw. Learn. Syst. **99**, 1–8 (2020)
9. Zheng, W., Zhu, X., Wen, G., Zhu, Y., Yu, H., Gan, J.: Unsupervised feature selection by self-paced learning regularization. Pattern Recogn. Lett. **132**, 4–11 (2020)
10. Zhu, X., Zhang, S., Zhu, Y., Zhu, P., Gao, Y.: Unsupervised spectral feature selection with dynamic hyper-graph learning. IEEE Trans. Knowl. Data Eng. (2020). https://doi.org/10.1109/TKDE.2020.3017250

11. Shen, H.T., Zhu, Y., Zheng, W., Zhu, X.: Half-quadratic minimization for unsupervised feature selection on incomplete data. IEEE Trans. Neural Netw. Learn. Syst. (2020). https://doi.org/10.1109/TNNLS.2020.3009632
12. Cai, J., Luo, J., Wang, S., Yang, S.: Feature selection in machine learning: a new perspective. Neurocomputing **300**(jul.26), 70–79 (2018)
13. Shi, C., Duan, C., Gu, Z., Tian, Q., An, G., Zhao, R.: Semi-supervised feature selection analysis with structured multi-view sparse regularization. Neurocomputing **330**, 412–424 (2019)
14. Bennett, K.P., Demiriz, A.: Semi-supervised support vector machines. In: Advances in Neural Information Processing Systems, pp. 368–374 (1999)
15. Zhao, Z., Liu, H.: Semi-supervised feature selection via spectral analysis. In: Proceedings of the 2007 SIAM International Conference on Data Mining, pp. 641–646 (2007)
16. Ren, J. Qiu, Z., Fan, W., Cheng, H., Philip, S.Y.: Forward semi-supervised feature selection. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 970–976 (2008)
17. Chen, X., Yuan, G., Nie, F., Huang, J.Z.: Semi-supervised feature selection via rescaled linear regression. In: IJCAI, pp. 1525–1531 (2017)
18. Moosavi, M.R., Jahromi, M.Z., Ghodratnama, S., Taheri, M., Sadreddini, M.H.: A cost sensitive learning method to tune the nearest neighbour for intrusion detection. Iran. J. Sci. Technol. - Trans. Electr. Eng. **36**, 109–129 (2012)
19. Bai, L., Cui, L., Wang, Y., Yu, P.S., Hancock, E.R.: Fused lasso for feature selection using structural information. Trans. Knowl. Data Eng. 16–27 (2019)
20. Liu, M., Xu, C., Luo, Y., Xu, C., Wen, Y., Tao, D.: Cost-sensitive feature selection by optimizing F-measures. IEEE Trans. Image Process. **27**(3), 1323–1335 (2017)
21. Lin, J.: Divergence measures based on the shannon entropy. IEEE Trans. Inf. Theory **37**(1), 145–151 (1991)
22. Bai, L., Hancock, E.R.: Graph kernels from the Jensen-Shannon divergence. J. Math. Imaging Vis. **47**(1), 60–69 (2013)
23. Wang, H., et al.: Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance. In: 2011 International Conference on Computer Vision, pp. 557–562 (2011)
24. Miao, L., Liu, M., Zhang, D.: Cost-sensitive feature selection with application in software defect prediction. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR 2012), pp. 967–970 (2012)
25. Sechidis, K., Brown, G.: Simple strategies for semi-supervised feature selection. Mach. Learn. **107**(2), 357–395 (2018)
26. Zhao, H., Yu, S.: Cost-sensitive feature selection via the $l_{2,1}$-norm. Int. J. Approx. Reason. **104**(1), 25–37 (2019)
27. Melacci, S., Belkin, M.: Laplacian support vector machines trained in the primal. J. Mach. Learn. Res. **12**(3), 1149–1184 (2011)