Semi-supervised Feature Selection Based on Logistic I-RELIEF for Multi-classification

Baige Tang[1] and Li Zhang[1,2(✉)]

[1] School of Computer Science and Technology & Joint International Research Laboratory of Machine Learning and Neuromorphic Computing, Soochow University, Suzhou 215006, Jiangsu, China
bgtang@stu.suda.edu.cn, zhangliml@suda.edu.cn
[2] Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou 215006, Jiangsu, China

Abstract. The semi-supervised Logistic I-RELIEF (SLIR) algorithm has been proposed for feature selection, which can handle both labeled and unlabeled data to perform feature selection. However, SLIR can only deal with binary problems. To remedy it, this paper presents a multi-classification semi-supervised Logistic I-RELIEF (MSLIR) algorithm for feature selection. Based on SLIR, MSLIR designs a novel scheme to calculate the margin vectors of unlabeled samples. Experimental results demonstrate the efficiency and effectiveness of our algorithm.

Keywords: Logistic I-RELIEF · Feature selection
Multi-classification · Semi-supervised · Margin vector

1 Introduction

High-dimensional data, such as DNA microarray data, medical data, and satellite remote sensing data, may contain irrelevant information, which generally exists in machine learning and pattern recognition. It is meaningful to use feature selection or feature extraction as a pre-processing way for reducing the negative influence of irrelevant data. Feature selection, as a dimensionality reduction technology makes great contributions to saving storage space and reducing the computational cost. Therefore, feature selection is widely applied to many learning tasks.

Feature selection methods can be categorized into three groups: supervised, unsupervised and semi-supervised methods. Supervised feature selection methods include RELIEF-based methods [1–3], Fisher criterion-based methods [4,5], etc. RELIEF is one of the most effective feature selection algorithms, which finds a weight vector for features by maximizing the margin between the differences of given samples and their nearest neighbors in two classes. The features with larger weights could be selected to perform classification tasks. The variants of RELIEF have been proposed, such as Logistic I-RELIEF (LIR) [1] and

X. Geng and B.-H. Kang (Eds.): PRICAI 2018, LNAI 11012, pp. 719–731, 2018.
https://doi.org/10.1007/978-3-319-97304-3_55

RELIEF-F [3]. Supervised feature selection methods require a large amount of labeled data which is hard to obtain. It is difficult for supervised feature selection methods to choose features that are distinguishable from few labeled data in the training dataset. Therefore, many unsupervised feature selection methods have been proposed [6–8], which can effectively utilize unlabeled data. However, these methods ignore the information contained in labeled data, so they cannot identify the discriminative features well [7].

The semi-supervised feature selection methods can achieve better performance since they make full use of the labeled and unlabeled data. Common methods include the clustering-based method [9], locality sensitive-based method [10], local discriminative information-based method [11], semi-supervised Logistic I-RELIEF method (SLIR) [12], forward method [13,14], multi-objective optimization method [15], spectral analysis method [16], and so on. This paper focuses on the RELIEF-based methods. In SLIR, labeled data are used to maximize the distance between samples in different classes, while the unlabeled data samples are used to extract the geometric structure in a data space [12]. SLIR is the first and successful variant of RELIEF in semi-supervised learning. However, SLIR can only deal with binary problems and cannot effectively solve the multi-class problems.

To solve this issue, we propose a multi-classification semi-supervised Logistic I-RELIEF (MSLIR) algorithm for feature selection based on SLIR. Under the semi-supervised learning framework, RELIEF-based methods need to consider how to compute the margin vector of an unlabeled sample. MSLIR designs a novel scheme to calculate the margin vectors of unlabeled samples for multi-classification problems.

The rest of paper is arranged as follows. Section 2 briefly introduces related work include Logistic I-RELIEF and semi-supervised Logistic I-RELIEF. We propose multi-classification semi-supervised Logistic I-RELIEF algorithm in Sect. 3. Experimental results are presented in Sect. 4. The paper is concluded in Sect. 5.

2 Related Work

This section mainly introduces two algorithms: LIR and SLIR, where SLIR is an extended algorithm of LIR.

2.1 Logistic I-RELIEF

In I-RELIEF, neighbors of a given sample in the original feature space is inconsistent with the nearest neighbors in the weighted feature space, so LIR proposes a new probabilistic model to calculate the distance between samples and their neighbors.

A assume that the input dataset is $D = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N} \subset R^I \times \{+1, -1\}$, where $\mathbf{x}_n$ is a labeled sample and y_n is its label, I is the number of original

features, and N is the number of samples. Here, we discuss the binary problem: $+1$ and -1 represent positive and negative classes, respectively.

The optimization problem of LIR can be described as:

$$\begin{aligned} &\min_{\mathbf{w}} \sum_{n=1}^{N} \log(1+\exp(-\mathbf{w}^T\mathbf{z}_n)) + \lambda\|\mathbf{w}\|_1 \\ &s.t. \quad \mathbf{w} \geq 0 \end{aligned} \tag{1}$$

where $\|\cdot\|_1$ is the 1-norm, $\mathbf{w}$ is the feature weight vector which represents the importance of features, $\lambda \geq 0$ is a regularization parameter to avoid overfitting, and $\mathbf{z}_n$ is the margin vector of the sample $\mathbf{x}_n$ which can be written as follows:

$$\begin{aligned} \mathbf{z}_n = &\sum_{\mathbf{x}_i \in M_n} P(\mathbf{x}_i = NM(\mathbf{x}_n)|\mathbf{w})|\mathbf{x}_n - \mathbf{x}_i| \\ &- \sum_{\mathbf{x}_i \in H_n} P(\mathbf{x}_i = NH(\mathbf{x}_n)|\mathbf{w})|\mathbf{x}_n - \mathbf{x}_i| \end{aligned} \tag{2}$$

where $M_n = \{\mathbf{x}_i|(\mathbf{x}_i, y_i) \in D, y_i \neq y_n, i = 1, \ldots, N\}$, $H_n = \{\mathbf{x}_i|(\mathbf{x}_i, y_i) \in D, y_i = y_n, i = 1, \ldots, N\}$, $P(\mathbf{x}_i = NM(\mathbf{x}_n)|\mathbf{w})$ and $P(\mathbf{x}_i = NH(\mathbf{x}_n)|\mathbf{w})$ are the probabilities that the sample $\mathbf{x}_i$ is the nearest miss and the nearest hit of $\mathbf{x}_n$, respectively, $NM(\mathbf{x}_n)$ represents the nearest miss of sample $\mathbf{x}_n$ and $NH(\mathbf{x}_n)$ represents the nearest hit of sample $\mathbf{x}_n$.

2.2 Semi-supervised Logistic I-RELIEF

On the basis of Logistic I-RELIEF algorithm, Sun et al. [12] extended Logistic I-RELIEF to semi-supervised learning. Due to the introduction of unlabeled samples, the calculation for the margin vectors of unlabeled samples needs to be revised.

We are given a labeled sample set $D_l = \{(\mathbf{x}_i^l, y_i^l)\}_{i=1}^L (y_i \in \{\pm 1\})$ and an unlabeled sample set $D_u = \{\mathbf{x}_i^u\}_{i=1}^U$. For a given labeled sample $\mathbf{x}^l$ and unlabeled sample $\mathbf{x}^u$, let their margin vector be $\mathbf{z}^l$ and $\mathbf{z}^u$, respectively. Since $\mathbf{x}^l$ is a labeled sample, its margin vector $\mathbf{z}^l$ can be computed as Eq. (2). Fortunately, $\mathbf{z}^u$ has the same absolute value, regardless of which class the sample $\mathbf{x}^u$ belongs to. Therefore, $\mathbf{z}^u$ can use the same way of computing $\mathbf{z}_n$. SLIR can be cast into the following optimization problem:

$$\begin{aligned} &\min_{\mathbf{w}} \|\mathbf{w}\|_1 + \alpha \sum_{i=1}^{L} \log(1+\exp(-\mathbf{w}^T\mathbf{z}_i^l)) + \beta \sum_{i=1}^{U} \exp(-(\mathbf{w}^T\mathbf{z}_i^u)^2/\delta) \\ &s.t. \quad \mathbf{w} \geq 0 \end{aligned} \tag{3}$$

where δ is the kernel width, $\mathbf{z}_i^l$ is margin vector of sample $\mathbf{x}_i^l$, $\mathbf{z}_i^u$ is margin vector of sample $\mathbf{x}_i^u$, α and β represent the contribution of labeled and unlabeled samples to the cost function (3), respectively. Then (3) can be solved by using the gradient descent method.

3 Semi-supervised Logistic I-RELIEF for Multi-classification

SLIR can only deal with bianry problems. To solve this issue, we propose the MSLIR algorithm for multi-classification based on SLIR. Given a labeled sample set $D_l = \{(\mathbf{x}_i^l, y_i^l)\}_{i=1}^L$ with $y_i^l \in \{1, 2, \ldots, c\}$ and the class number c, and an unlabeled sample set $D_u = \{\mathbf{x}_i^u\}_{i=1}^U$, the goal of MSLIR is to find feature weights for multi-classification tasks under the semi-supervised situation. There are $L+U$ samples in total ($L \ll U$). MSLIR designs a novel scheme to calculate the margin vectors of unlabeled samples, based on which MSLIR can be formulated as an optimization problem similar to SLIR. In the following, we discuss MSLIR in detail.

3.1 Calculation of Margin Vectors

MSLIR requires to calculate margin vector for each sample including labeled and unlabeled one. The margin vector $\mathbf{z}_i^l$ of the labeled sample $\mathbf{x}_i^l$ can be expressed as follows:

$$\begin{aligned}\mathbf{z}_i^l = &\sum_{\mathbf{x}_k^l \in M_i} P(\mathbf{x}_k^l = NM(\mathbf{x}_i^l)|\mathbf{w})|\mathbf{x}_i^l - \mathbf{x}_k^l| \\ &- \sum_{\mathbf{x}_k^l \in H_i} P(\mathbf{x}_k^l = NH(\mathbf{x}_i^l)|\mathbf{w})|\mathbf{x}_i^l - \mathbf{x}_k^l| \end{aligned} \tag{4}$$

where the set $M_i = \{\mathbf{x}_k^l | (\mathbf{x}_k^l, y_k^l) \in D_l, y_k^l \neq y_i^l, i = 1, \ldots, L, y_i^l = \{1, \ldots, c\}\}$ contains all labeled samples that have different labels from $\mathbf{x}_i^l$, the set $H_i = \{\mathbf{x}_k^l | (\mathbf{x}_k^l, y_k^l) \in D_l, y_k^l = y_i^l, i = 1, \ldots, L\}$ contains all labeled samples that have the same labels as $\mathbf{x}_i^l$.

For the multi-classification problem, the sample $\mathbf{x}_i^u$ in the data set D_u may belong to any class, hence, the calculation of the margin vectors of unlabeled samples needs to be redefined. Suppose that we assign a temporary label to the unlabeled sample $\mathbf{x}_i^u$. Then we can calculate the candidate margin vector $(\mathbf{z}_i^u)^j$ of the unlabeled sample $\mathbf{x}_i^u$ with the assigned label j as follows:

$$\begin{aligned}(\mathbf{z}_i^u)^j = &\sum_{\mathbf{x}_k^l \in \overline{M}_i^j, y_k \neq j} P(\mathbf{x}_k^l = NM^j(\mathbf{x}_i^u)|\mathbf{w})|\mathbf{x}_i^u - \mathbf{x}_k^l| \\ &- \sum_{\mathbf{x}_k^l \in \overline{H}_i^j, y_k = j} P(\mathbf{x}_k^l = NH^j(\mathbf{x}_i^u)|\mathbf{w})|\mathbf{x}_i^u - \mathbf{x}_k^l| \end{aligned} \tag{5}$$

where $\overline{M}_i^j = \{\mathbf{x}_k^l | (\mathbf{x}_k^l, y_k^l) \in D_l, y_k^l \neq j, k = 1, \ldots, L, j = 1, \ldots, c\}$ is the sample set that contains all labeled samples whose label is not equal to j, and $\overline{H}_i^j = \{\mathbf{x}_k^l | (\mathbf{x}_k^l, y_k^l) \in D_l, y_k^l = j, k = 1, \ldots, L\}$ is the sample set that contains all labeled samples whose label is equal to j.

Which is the possible margin vector of the sample $\mathbf{x}_i^u$ among the c candidate margin vectors? Without any apriori information, we consider the candidate

margin vector which has the greatest inner product with the feature weight vector $\mathbf{w}$ as the possible one. Then $\mathbf{z}_i^u$ can be determined by:

$$\mathbf{z}_i^u = \arg \max_{j=1,\ldots,c} \mathbf{w}^T (\mathbf{z}_i^u)^j \tag{6}$$

3.2 Optimization Problem and Algorithm Description

After the margin vectors of labeled and unlabeled samples are defined, MSLIR is to solve the following optimization problem:

$$\begin{aligned} \min_{\mathbf{w}} \|\mathbf{w}\|_1 + \alpha \sum_{i=1}^{L} \log(1 + \exp(-\mathbf{w}^T \mathbf{z}_i^l)) \\ + \beta \sum_{i=1}^{U} \log(1 + \exp(-\mathbf{w}^T \mathbf{z}_i^u)) \\ s.t. \quad \mathbf{w} \geq 0 \end{aligned} \tag{7}$$

where $\alpha \geq 0$ and $\beta \geq 0$ are the regularization parameters that control the importance of labeled and unlabeled samples, respectively. By comparing (3) and (7), we find that there are two differences between them. First, the calculation way of $\mathbf{z}_i^u$ is different. In (3), calculating $\mathbf{z}_i^u$ does not consider the label of $\mathbf{x}_i^u$. Contrary, the calculation of $\mathbf{z}_i^u$ takes into account the possible label of $\mathbf{x}_i^u$. Second, (7) uses the same logical regression form for both labeled and unlabeled samples to ensure the symmetry of labeled and unlabeled samples.

In order to facilitate calculation, we convert the optimization problem (7) to an unconstraint optimization problem. Let $\mathbf{w} = [v_1^2, \ldots, v_I^2]^T$ and $\mathbf{v} = [v_1, \ldots, v_I]$. The optimization formula (7) can be rewritten as:

$$\begin{aligned} \min_{\mathbf{v}} J = \|\mathbf{v}\|_2^2 + \alpha \sum_{i=1}^{L} \log(1 + \exp(-\sum_{d=1}^{I} v_d^2 z_{id}^l)) \\ + \beta \sum_{i=1}^{U} \log(1 + \exp(-\sum_{d=1}^{I} v_d^2 z_{id}^u)) \end{aligned} \tag{8}$$

where $\mathbf{z}_i^l = [z_{i1}^l, \ldots, z_{iI}^l]$. In order to solve formula (8), we use the gradient descent method. The derivation of J to $\mathbf{v}$ can be written as follows:

$$\begin{aligned} \frac{\partial J}{\partial v_k} = 2v_k - \alpha \sum_{i=1}^{L} \frac{\exp(-\sum_{d=1}^{I} v_d^2 z_{id}^l)(2v_k z_{ik}^l)}{1 + \exp(-\sum_{d=1}^{I} v_d^2 z_{id}^l)} \\ - \beta \sum_{i=1}^{U} \frac{\exp(-\sum_{d=1}^{I} v_d^2 z_{id}^u)(2v_k z_{ik}^u)}{1 + \exp(-\sum_{d=1}^{I} v_d^2 z_{id}^u)} \end{aligned} \tag{9}$$

Let $Q_1 = \sum_{i=1}^{L} \frac{\exp(-\sum_{d=1}^{I} v_d^2 z_{id}^l)(v_k z_{ik}^l)}{1+\exp(-\sum_{d=1}^{I} v_d^2 z_{id}^l)}$ and $Q_2 = \sum_{i=1}^{U} \frac{\exp(-\sum_{d=1}^{I} v_d^2 z_{id}^u)(v_k z_{ik}^u)}{1+\exp(-\sum_{d=1}^{I} v_d^2 z_{id}^u)}$, the update formulation for each dimension can be described as:

$$v_k \leftarrow v_k - \eta(v_k - \alpha Q_1 - \beta Q_2) \tag{10}$$

Algorithm 1. MSLIR

Input: Labeled dataset $D_l = \{(\mathbf{x}_i^l, y_i^l)\}_{i=1}^L \subset R^I \times \{1, 2, \ldots, c\}$; unlabeled dataset $D_u = \{\mathbf{x}_i^u\}_{i=1}^U \subset R^I$, regularization parameters α and β, the iteration number T, and the stop criterion θ.

Output: Feature weight $\mathbf{w}$.

```
begin
    Initialization: Set w_(0) = [1, 1, ..., 1]^T, t = 1, ρ = 1;
    while t ≤ T && ρ > θ do
        Compute z_i^l by (4), i = 1, ..., L;
        Compute (z_i^u)^j by (5), j = 1, ..., c and find z_i^u by (6), i = 1, ..., U;
        Solve formula (7) using the gradient descent method to find v;
        Compute w_(t) = [v_1^2, ..., v_I^2]^T;
        ρ = ||w_(t) − w_(t−1)||;
        t = t + 1;
    end
    w = w_(t);
    Output w.
end
```

where η is the learning rate. The pseudo-code of MSLIR is shown in Algorithm 1. The algorithm alternatively modifies the weight vector until convergence.

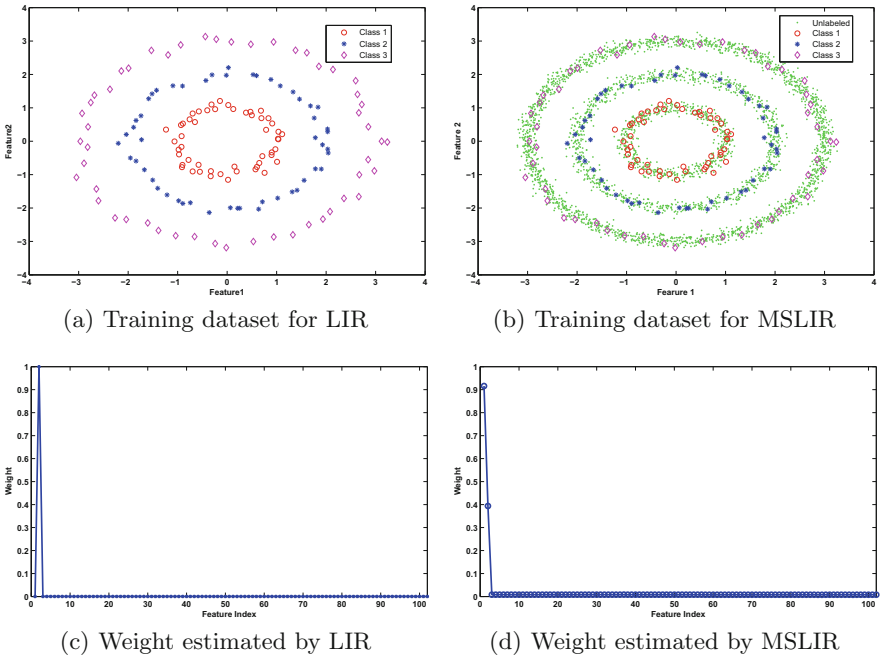
The computational complexity of MSLIR includes three parts: the calculation of the margin vectors for both labeled and unlabeled samples, and the solution to the optimization problem (8). In MSLIR, the computational complexity of solving (8) and calculating the margin vectors for labeled samples are identical to SLIR. The computational complexity of calculating the margin vectors for unlabeled samples is $O(cdLU)$ in MSLIR. For the same task, SLIR has the computational complexity of $O(dLU)$. In a nutshell, MSLIR has a compared complexity with SLIR.

4 Experiments

In this section, we evaluate the performance of the proposed MSLIR algorithm on one artificial dataset and eight UCI datasets [17]. The compared methods include LIR, SLIR and RELEIF-F. In each dataset, we added 100 irrelevant features to each sample, which are independently sampled from zero-mean, unit-variance Gaussian distribution, then we normalize these irrelevant features with the original data. All the experiments are implement in MATLAB R2013a on a PC with an Inter Core I5 processor with 4 GB RAM.

4.1 Artificial Dataset

We conduct experiments on the "ThreeCircles" dataset to verify the ability of the proposed algorithm in feature selection. There are 3 classes in the "ThreeCircles"



(a) Training dataset for LIR (b) Training dataset for MSLIR

(c) Weight estimated by LIR (d) Weight estimated by MSLIR

Fig. 1. "ThreeCircles" dataset (a) training dataset for LIR, (b) training dataset for MSLIR, feature weights estimated by (c) LIR and (d) MSLIR.

dataset, as shown in Fig. 1(a) and (b). In Fig. 1(a), each class has 51 labeled samples. In Fig. 1(b), each class also has 51 labeled samples as Fig. 1(a), and additionally 3450 unlabeled ones. LIR and MSLIR are conducted to compare the performance in feature selection. In MSLIR, let regularization parameters $\alpha = 6$ and $\beta = 3$. Let kernel width $\delta = 8$ for MSLIR and LIR.

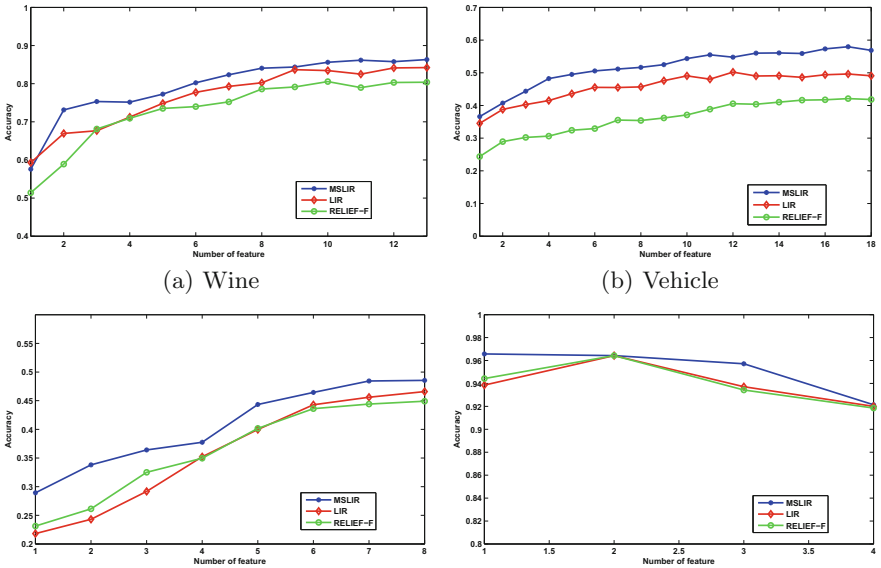
In Fig. 1(c), LIR fails to identify the useful feature since the weight of the first feature is equal to zero. Figure 1(d) shows that the first two features selected by MSLIR has the highest weights and the weights of other noisy features are nearly zero. The results in Fig. 1(c) and (d) indicate that MSLIR can successfully identify the first two useful features using labeled and unlabeled data.

4.2 UCI Datasets

In this section, we use eight UCI datasets to verify the performance of algorithms. The eight UCI datasets include Wine, Vehicle, Yeast, Iris, Wdbc, Heart, Pima and Sonar datasets, of which the first four datasets are multi-class ones, and the rest are two-class ones. The datasets are randomly divided into independent training and test subsets, where training subsets contain labeled and unlabeled data. Semi-supervised methods use both labeled and unlabeled data, and supervised methods only use labeled data. The data information is summarized in Table 1, where "#Training" and "#Test" represent the number of

Table 1. Information of four UCI datasets.

Data sets	#Training		#Test	#Feature	#Class
	#Labeled	#Unlabeled			
Wine	10	40	128	13(100)	3
Vehicle	30	170	646	18(100)	4
Yeast	178	250	1056	8(100)	10
Iris	20	30	100	4(100)	3
Heart	40	130	100	13(100)	2
Wdbc	20	149	400	30(100)	2
Pima	20	148	600	8(100)	2
Sonar	20	38	150	61(100)	2



(a) Wine (b) Vehicle

(c) Yeast (d) Iris

Fig. 2. Classification accuracy of MSLIR, RELIEF-F and LIR on UCI datasets, (a) Wine, (b) Vehicle, (c) Yeast and (d) Iris.

training samples and test samples, respectively. The total number of "#Labeled" and "#Unlabeled" data is equal to the number of training samples. "#Feature" represents the dimension of dataset.

The classifier utilized in our experiments is support vector machine (SVM). The Gaussian kernel parameter and regularization parameter of SVM are determined by the grid search method, where both parameters vary from 2^{-10} to 2^{10}.

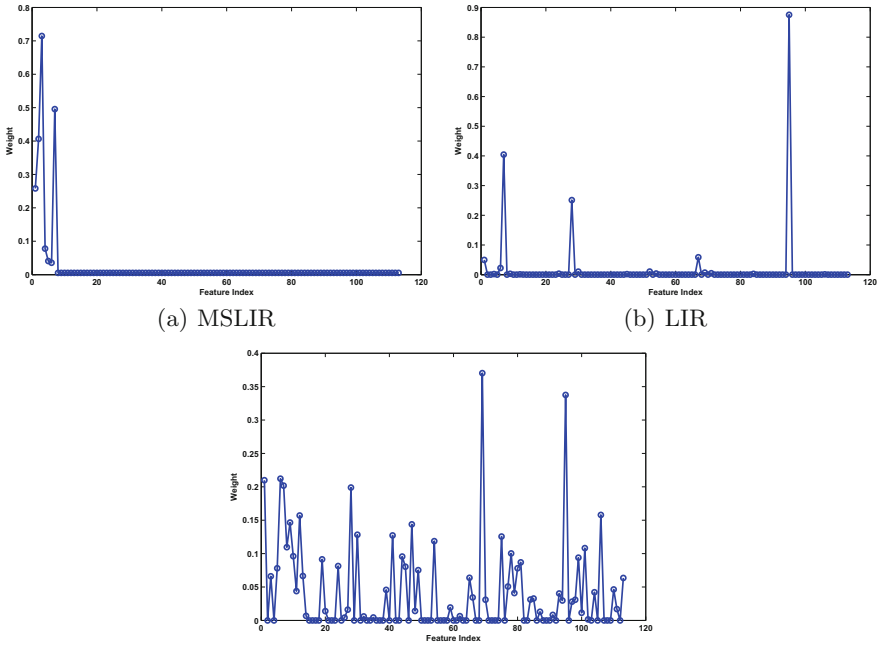
Multi-classification Datasets. We compare the proposed feature selection algorithm with RELIEF-F and LIR, which both are supervised algorithms for solving multi-classification problems. We use the cross-validation method to select parameter α and β in MSLIR. The regularization parameter and learning rate in LIR follow the setting in [1]: $\lambda = 10$ and $\eta = 0.03$. Experiments are implemented on the Wine, Vehicle, Yeast and Iris datasets, each of which is randomly partitioned 10 times. We report the average results. The curves of the average classification accuracy vs. the n top-ranked features are shown in Fig. 2, where n is the original feature dimension of datasets. In Fig. 2(a), (b) and (c), the classification accuracy of the three algorithms also gradually increases as the number of selected features increases, which indicates that the useful features are gradually found. We can see that SMLIR is always better than LIR and RELIEF-F, which may indicate LIR and RELIEF-F algorithms contain noise features. In Fig. 2(d), MSLIR achieves the best classification accuracy when one feature is used, and LIR and RELIEF-F need two features. Obviously, MSLIR has a significantly higher performance than the other two algorithms. The best average accuracy and the corresponding standard deviation are given in Table 2. We can observe that MSLIR has a great improvement on Wine and Vehicle datasets.

Table 2. Classification accuracy and standard deviations (%) of SVM using features selected by MSLIR, LIR and RELIEF-F

Data sets	MSLIR	LIR	RELIEF-F
Wine	**88.59** $\pm$ 9.07	84.14 $\pm$ 6.38	82.11 $\pm$ 8.62
Vehicle	**57.97** $\pm$ 4.62	50.22 $\pm$ 9.11	42.11 $\pm$ 10.69
Yeast	**48.55** $\pm$ 14.17	46.59 $\pm$ 14.95	44.91 $\pm$ 17.41
Iris	**96.57** $\pm$ 1.21	96.43 $\pm$ 1.21	96.43 $\pm$ 1.21

In Fig. 3, we give the feature weights of the three algorithms on the Wine dataset. We can observe that MSLIR correctly selects relevant features, while LIR assigns higher weights to some irrelevant features, and RELEIF-F fails to identify noise data. In summary, MSLIR can handle feature selection problems for multi-classification under the semi-supervised framework. The supervised approaches do not remove the noise features well with few labeled data. Because of the introduction of unlabeled data, MSLIR can effectively eliminate noise features.

Binary Classification Datasets. We conduct experiments to compare MSLIR with SLIR and prove that MSLIR is also suitable for binary problems. SLIR is a semi-supervised feature selection method for solving binary problems. In MSLIR, the way of parameter setting is the same as the multi-classification case. The parameters α and β are determined by the cross-validation method in SLIR. Experiments are implemented on the Heart, Wdbc, Pima and Sonar datasets,



(a) MSLIR (b) LIR

(c) RELIEF-F

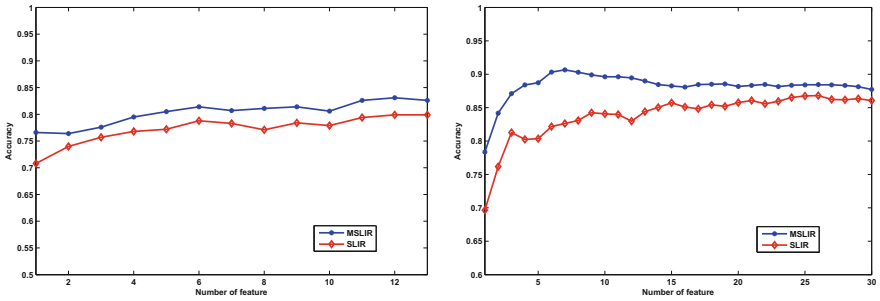
Fig. 3. Feature weights obtained by (a) MSLIR, (b) LIR and (c) RELIEF-F on the Wine dataset

each of which is randomly divided 10 times and the average accuracy is taken as the final results.

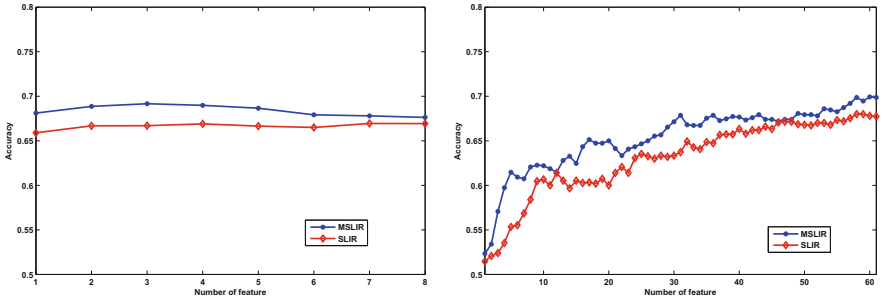
The results are shown in Fig. 4. In Figs. 4(a) and (c), we can observe that classification accuracy of MSLIR and SLIR are steady. However, MSLIR is much better than SLIR, which indicates MSLIR does not choose noise features. Figure 4(b) shows that SMLIR reaches the best accuracy when the number of features is 6, and SLIR when the number of features is 25, which implies that MSLIR chooses few features to reach the better accuracy. The best average classification accuracy and corresponding standard deviations are listed in Table 3. Compared to

Table 3. Classification accuracy and standard deviations (%) of SVM using features selected by MSLIR and SLIR

Data sets	MSLIR	SLIR
Heart	**83.1** $\pm$ 3.93	79.9 $\pm$ 4.04
Wdbc	**90.65** $\pm$ 3.15	86.8 $\pm$ 5.14
Pima	**69.17** $\pm$ 3.90	66.9 $\pm$ 3.25
Sonar	**70.13** $\pm$ 5.44	67.33 $\pm$ 8.38

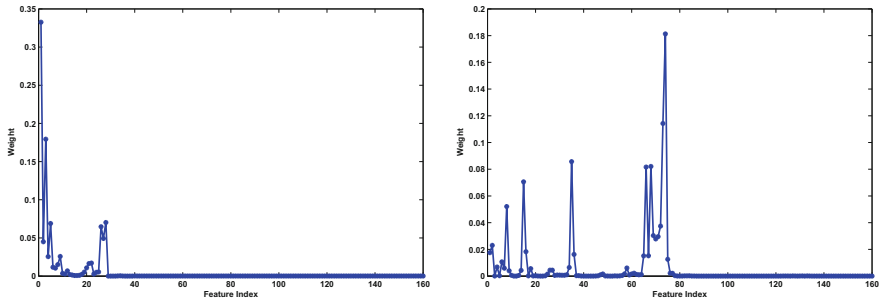


(a) Heart (b) Wdbc



(c) Pima (d) Sonar

Fig. 4. Classification accuracy of MSLIR and SLIR on UCI datasets, (a) Heart, (b) Wdbc, (c) Pima and (d) Sonar.



(a) MSLIR (b) SLIR

Fig. 5. Feature weights obtained by (a) MSLIR and (b) SLIR on the Sonar dataset

SLIR, the accuracy of MSLIR is improved 3.2%, 3.85%, 2.27% and 2.80% on Heart, Wdbc, Pima and Sonar datasets, respectively.

Figure 5 shows the distribution of feature weights on the Sonar dataset, we can see that MSLIR selects relevant features, while SLIR evaluates noise features higher weights and fails to distinguish relevant features.

Compared to SLIR, MSLIR can effectively solve the binary problems, and improve the classification accuracy.

5 Conclusion

In this paper, we propose MSLIR based on SLIR. MSLIR overcomes the drawback of SLIR, which can make full use of unlabeled information to select relevant features on multi-class of data. The results on one artificial dataset demonstrate that MSLIR can effectively handle noise data. When dealing with multi-class classification tasks, MSLIR can extract effective features under the semi-supervised framework compared to supervised methods. For binary classification problems, MSLIR can achieve better performance than SLIR despite the fact that both methods are semi-supervised learning ones.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China under Grants No. 61373093, No. 61402310, No. 61672364 and No. 61672365, by the Soochow Scholar Project of Soochow University, by the Six Talent Peak Project of Jiangsu Province of China.

References

1. Sun, Y.: Iterative RELIEF for feature weighting: algorithms, theories, and applications. IEEE Trans. Pattern Anal. Mach. Intell. **29**, 1035–1051 (2007)
2. Kira, K., Rendell, L.A.: The feature selection problem: traditional methods and a new algorithm. In: Tenth National Conference on Artificial Intelligence, pp. 129–134 (1992)
3. Kononenko, I.: Estimating attributes: analysis and extensions of RELIEF. In: Bergadano, F., De Raedt, L. (eds.) ECML 1994. LNCS, vol. 784, pp. 171–182. Springer, Heidelberg (1994). https://doi.org/10.1007/3-540-57868-4_57
4. Cheng, Z., Zhang, Y., Fan, X., Zhu, B.: Study on discriminant matrices of commonly used fisher discriminant functions. Acta Automatica Sin. **36**(10), 1361–1370 (2010)
5. Chen, L.F., Liao, H.Y.M., Ko, M.T., Lin, J.C., Yu, G.J.: A new LDA based face recognition system which can solve the small sample size problem. Pattern Recognit. **33**(10), 1713–1726 (2000)
6. Mitra, P., Murthy, C.A., Pal, S.K.: Unsupervised feature selection using feature similarity. IEEE Trans. Pattern Anal. Mach. Intell. **24**(3), 301–312 (2002)
7. He, X., Cai, D., Niyogi, P.: Laplacian score for feature selection. In: International Conference on Neural Information Processing Systems, vol. 18, pp. 507–514 (2005)
8. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2006)
9. Quinzn, I., Sotoca, J.M., Pla, F.: Clustering-based feature selection in semi-supervised problems. In: International Conference on Intelligent Systems Design and Application, pp. 535–540 (2009)
10. Zhao, J., Lu, K., He, X.: Locality sensitive semi-supervised feature selection. Neurocomputing **71**(10), 1842–1849 (2008)
11. Zeng, Z., Wang, X.D., Zhang, J., Wu, Q.: Semi-supervised feature selection based on local discriminative information. Neurocomputing **173**(P1), 102–109 (2016)

12. Cheng, Y., Cai, Y., Sun, Y., Li, J.: Semi-supervised feature selection under the Logistic I-RELIEF framework. In: International Conference on Pattern Recognition, pp. 1–4 (2008)
13. Ren, J., Qiu, Z., Fan, W., Cheng, H., Yu, P.S.: Forward semi-supervised feature selection. In: Washio, T., Suzuki, E., Ting, K.M., Inokuchi, A. (eds.) PAKDD 2008. LNCS (LNAI), vol. 5012, pp. 970–976. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-68125-0_101
14. Wang, B., Jia, Y., Yang, S.: Forward semi-supervised feature selection based on relevant set correlation. Int. Conf. Comput. Sci. Softw. Eng. **4**, 210–213 (2008)
15. Handl, J., Knowles, J.: Semi-supervised feature selection via multi-objective optimization. In: International Joint Conference on Neural Networks, pp. 3319–3326 (2006)
16. Zhao, Z., Liu, H.: Semi-supervised feature selection via spectral analysis. In: SIAM International Conference on Data Mining, SIAM-2007, SIAM, Minneapolis, Minnesota, USA, pp. 641–646 (2007)
17. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/datasets.html