Cost-Sensitive Feature Selection via F-Measure Optimization Reduction

Meng Liu,[†] Chang Xu,[‡] Yong Luo,[§‡] Chao Xu,[†] Yonggang Wen,[§] Dacheng Tao[‡]
[†]Key Laboratory of Machine Perception (MOE), Cooperative Medianet Innovation Center,
School of Electronics Engineering and Computer Science, PKU, Beijing 100871, China
[‡]Centre for Artificial Intelligence, UTS, Sydney, NSW 2007, Australia
[§]School of Computer Science and Engineering, NTU, 639798, Singapore
mengliu@pku.edu.cn, chang.xu@uts.edu.au, yluo180@gmail.com,
xuchao@cis.pku.edu.cn, ygwen@ntu.edu.sg, dacheng.tao@uts.edu.au

Abstract

Feature selection aims to select a small subset from the high-dimensional features which can lead to better learning performance, lower computational complexity, and better model readability. The class imbalance problem has been neglected by traditional feature selection methods, therefore the selected features will be biased towards the majority classes. Because of the superiority of F-measure to accuracy for imbalanced data, we propose to use F-measure as the performance measure for feature selection algorithms. As a pseudo-linear function, the optimization of F-measure can be achieved by minimizing the total costs. In this paper, we present a novel cost-sensitive feature selection (CSFS) method which optimizes F-measure instead of accuracy to take class imbalance issue into account. The features will be selected according to optimal F-measure classifier after solving a series of cost-sensitive feature selection sub-problems. The features selected by our method will fully represent the characteristics of not only majority classes, but also minority classes. Extensive experimental results conducted on synthetic, multi-class and multi-label datasets validate the efficiency and significance of our feature selection method.

Introduction

Feature selection has been one of the most popular dimensionality reduction techniques. It is a process of choosing a subset of relevant features from the high-dimensional data according to certain performance measure (Tang, Alelyani, and Liu 2014). Feature selection can further benefit the machine learning tasks such as classification and cluster by speeding up the learning process, improving the model generalization capability, and alleviating the effect of the curse of dimensionality (Nie et al. 2010). A considerable amount of research has been done during the last decade (Villela, de Castro Leite, and Neto 2015; Wang, Tang, and Liu 2015; Luo et al. 2016; Qian and Zhai 2013; Zhao et al. 2010; Xu, Tao, and Xu 2015), which can be divided into three groups: filter methods, wrapper methods and embedded methods. Filter methods, such as ReliefF (Kononenko 1994), mRMR (Peng, Long, and Ding 2005), F-statistic (Liu and Motoda 2012) and Information Gain (Raileanu and Stoffel 2004), choose features only relying the characteristics of data. Wrapper methods utilize predefined classifiers as a black box to evaluate the selected features. Support vector machine recursive feature elimination (SVM-RFE) (Guyon et al. 2002) and correlation-based feature selection (CFS) (Hall and Smith 1999) are representative wrapper methods. Embedded methods embed the feature selection process into classifier training. Regularized regression-based feature selection methods (Nie et al. 2010; Han and Kim 2015) are typical embedded methods.

The methods mentioned above are demonstrated effective in most situations. However, almost all of these methods neglect the influence of class imbalance issue. They are designed under the implicit assumption that the data distribution or sampling are balanced, *i.e.*, the sample sizes for different classes are about the same. The class imbalance issue is quite common in real-world datasets, which will negatively impact the traditional feature selection methods since they are inclined to choose the features that characterize the majority classes rather than those describe the minority classes. The neglect of class imbalance issue will make it more difficult to obtain better results for the subsequent machine learning tasks since the selected features are already biased towards the majority classes.

Feature selection methods which are dependent on classifiers also have the class imbalance problem (Nie et al. 2010; Han and Kim 2015). Taking regularized regression-based feature selection for example, these regularization models aim to minimize the fitting errors of the objective functions where the misclassification costs for different classes are treated equally (Tang, Alelyani, and Liu 2014). Therefore the feature subset is chosen to achieve the highest classification accuracy, which is not an appropriate performance under the imbalanced setting. Consequently, these types of methods can be referred as cost-blind feature selection methods.

High and balanced pair values of precision and recall result in high F-measure performance (Parambath, Usunier, and Grandvalet 2014). Therefore F-measure is a more suitable measure compared with accuracy in the imbalanced classes scenario (Pillai, Fumera, and Roli 2012; Dembczynski et al. 2011). Besides F-measure in binary classification, its variants in multi-class and multi-label classification are receiving much attention recently (Dembczynski et al. 2013; 2011; Ye et al. 2012; Pillai, Fumera, and Roli 2012). There

is a great number of studies on optimizing these F-measures, which can be categorized into two paradigms: the decision-theoretic approaches (DTA) (Lewis 1995) and empirical utility maximization (EUM) approaches. DTA approaches first estimate a probability model, which will be utilized to compute the optimal predictions. EUM approaches (Jansche 2005; Tsochantaridis et al. 2005) follow the structured risk minimization principle to minimize the objective function. Directly optimizing F-measure is difficult since it is non-convex, so different approximation methods are used in practice, such as the algorithms for maximizing a convex lower bound of F-measure for support vector machines (Tsochantaridis et al. 2005), and maximizing the expected F-measure of a probabilistic classifier using a logistic regression model (Jansche 2005). A simple yet effective method is to threshold the scores obtained by classifiers to maximize the F-measure empirically (Parambath, Usunier, and Grandvalet 2014; Yang 2001). Recent developments (Ye et al. 2012; Koyejo et al. 2014; Parambath, Usunier, and Grandvalet 2014; Narasimhan, Vaish, and Agarwal 2014) investigate the pseudo-linear property of F-measures by formulating them as functions of per-class false negative/false positive rate. Through the reduction to cost-sensitive classification, the optimization of F-measures can be accomplished by solving a series of cost-sensitive classification sub-problems.

By employing F-measure as the performance measure of selected features, we present an effective cost-sensitive feature selection (CSFS) method to handle the feature selection problem in the imbalanced data setting. Different from the existing embedded feature selection approaches (Nie et al. 2010; Han and Kim 2015), which focus on optimizing the accuracy, we encourage the feature selection solution to achieve the best F-measure. Motivated by the developments (Parambath, Usunier, and Grandvalet 2014; Ye et al. 2012) that F-measure optimization problem can be decomposed into a series of cost-sensitive classification problems, we further modify the classifiers of regularized regression-based feature selection methods into cost-sensitive. After solving a series of cost-sensitive feature selection problems, features will be selected according to the optimal classifier with the largest F-measure. Therefore, the class imbalance is taken into consideration, and selected features will fully represent both majority class and minority class. Experimental results on synthetic, multi-class and multi-label datasets have confirmed the efficiency of our method.

F-Measure Optimization Reduction

We first give a brief introduction of the notations used in this paper. We present matrices as bold uppercase letters and vectors as bold lowercase letters. Given a matrix $\mathbf{W} = [w_{ij}]$, we denote $\mathbf{w}^i$ as its i-th row and $\mathbf{w}_j$ as its j-th column. For $p > 0$, the ℓ_p-norm of the vector $\mathbf{b} \in \mathbb{R}^n$ is defined as $\|\mathbf{b}\|_p = (\sum_{i=1}^n |b_i|^p)^{\frac{1}{p}}$. The $\ell_{p,q}$-norm of the matrix $\mathbf{W} \in \mathbb{R}^{n\times m}$ is defined as $\|\mathbf{W}\|_{p,q} = (\sum_{i=1}^n \|\mathbf{w}^i\|_p^q)^{\frac{1}{q}}$, where $p > 0$ and $q > 0$. The symbol $\odot$ denotes the element-wise multiplication.

	Actual Positive	Actual Negative
Predicted Positive	tp	fp
Predicted Negative	fn	tn

(a) Confusion matrix

	Actual Positive	Actual Negative
Predicted Positive	0	r
Predicted Negative	$1+\beta^2-r$	0

(b) Cost matrix

Figure 1: Confusion matrix and associated cost matrix of binary classification.

For a given binary classifier, there are four possible outcomes: true positives tp, false positives fp, false negatives fn, and true negatives tn. They are represented as a confusion matrix in Figure 1(a). F-measure can be defined in terms of the marginal probabilities of classes and the per-class false negative/false positive probabilities. The marginal probability of label k is denoted by P_k, and the per-class false negative probability and false positive probability of a classifier h are denoted by $FN_k(h)$ and $FP_k(h)$, respectively (Parambath, Usunier, and Grandvalet 2014). These probabilities of a classifier h can be summarized by the error profile $\mathbf{e}(h)$:

$$\mathbf{e}(h) = (FN_1(h), FP_1(h), \ldots, FN_L(h), FP_L(h)), \quad (1)$$

where L is the number of labels, e_{2k-1} of $\mathbf{e}(h) \in \mathbb{R}^{2L}$ is the false negative probability of class k and e_{2k} is the false positive probability. In binary classification, we have $FN_2 = FP_1$. Thus, for any $\beta > 0$, F-measure can be written as a function of error profile $\mathbf{e}$:

$$F_\beta(\mathbf{e}) = \frac{(1+\beta^2)(P_1 - e_1)}{(1+\beta^2)P_1 - e_1 + e_2}. \quad (2)$$

There are several different definitions of F-measures in multi-class and multi-label classification. Specifically, we can transform the multi-class or multi-label classification into multiple binary classification problems, and the average over the F_β-measures of these binary problems is defined as the macro-F-measure. According to (Parambath, Usunier, and Grandvalet 2014), the micro-F-measure mlF_β for multi-label classification is defined as:

$$mlF_\beta(\mathbf{e}) = \frac{(1+\beta^2)\sum_{k=1}^L (P_k - e_{2k-1})}{\sum_{k=1}^L ((1+\beta^2)P_k + e_{2k} - e_{2k-1})}. \quad (3)$$

Multi-class classification differs from multi-label classification in that only a single class can be predicted for each example. According to (Kim, Wang, and Yasunori 2013), one definition of multi-class micro-F-measure, denoted as mcF_β can be written as:

$$mcF_\beta(\mathbf{e}) = \frac{(1+\beta^2)(1 - P_1 - \sum_{k=2}^L e_{2k-1})}{(1+\beta^2)(1-P_1) - \sum_{k=2}^L e_{2k-1} + e_1}. \quad (4)$$

The fractional-linear F-measures presented in Eqs. (2-4) are pseudo-linear functions with respect to $\mathbf{e}$. The important property of pseudo-linear functions is that their level sets, as function of the false negative rate and the false

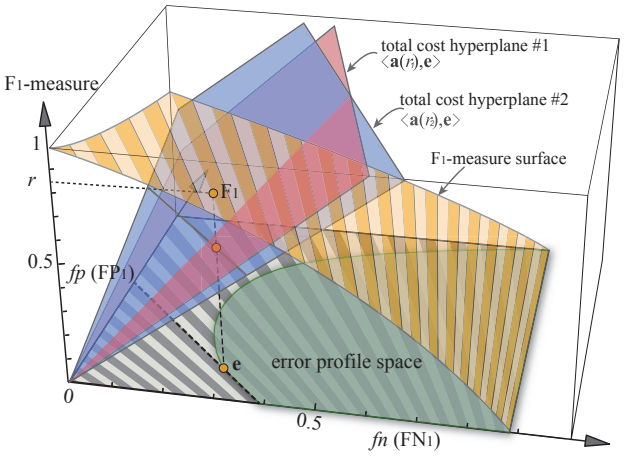

Figure 2: Illustration of F_1-measure surface with level sets and two total cost hyperplanes. Different $\mathbf{a}(r_i)$ will generate differnt costs for the misclassification errors. Error profile space contains all possible values of $\mathbf{e}$. We can notice that higher values of F-measure entail lower values of the total cost.

positive rate, are linear. Based on this observation, a recent work (Parambath, Usunier, and Grandvalet 2014) was proposed for F-measure maximization by reducing it into cost-sensitive classification, and proved that the obtained optimal classifier for a cost-sensitive classification problem with label dependent costs is also an optimal classifier for F-measure. This method can be separated into three steps. Firstly, the F-measure interval is discretized into a set of evenly spaced values $\{r_i\}$. F-measure is not invariant under label switching (Nie et al. 2010), *i.e.*, if the positive label is changed to negative, a different F-measure can be obtained. Therefore, the F-measure interval is discretized within the range $[0, 1 + \beta^2]$ rather than $[0, 1]$ in practice. Secondly, for each given F-measure value r_i, cost function $\mathbf{a} : \mathbb{R}^1_+ \rightarrow \mathbb{R}^{2L}_+$ generates a cost vector $\mathbf{a}(r_i)$ and assigns costs to the elements of error profile $\mathbf{e}$, more specifically, $1 + \beta^2 - r_i$ for false negative and r_i for false positive in binary classification. These costs are shown as a cost matrix in Figure 1(b) (Parambath, Usunier, and Grandvalet 2014). Therefore the goal of optimization is changed to minimize the total cost $\langle \mathbf{a}(r_i), \mathbf{e}(h) \rangle$, which is the inner product of cost vector and error profile (Parambath, Usunier, and Grandvalet 2014). Finally, cost-sensitive classifiers for each $\mathbf{a}(r_i)$ are learned to minimize the total cost $\langle \mathbf{a}(r_i), \mathbf{e}(h) \rangle$, and the one with largest F-measure on the validation set is selected as the optimal classifier. Figure 2 shows that the higher the F-measure value, the lower the total cost. This indicates that maximizing F-measure can be achieved by minimizing the corresponding total cost.

Cost-Sensitive Feature Selection

When the data sampling of different classes is imbalanced, it is difficult to discover a satisfactory feature selection solution to fully represent the properties of different classes. To deal with this problem, we propose to optimize F-measure instead of accuracy in the feature selection task. Motivated by the reduction of F-measure optimization to cost-sensitive classification (Parambath, Usunier, and Grandvalet 2014), we modify the classifiers used in traditional feature selection methods into cost-sensitive by adding properly generated costs with the in-depth theory guidance. Features are selected according to the classifier with the optimal F-measure performance. This leads to a novel cost-sensitive feature selection (CSFS) method. Figure 3 presents a systematic illustration of our method .

Problem Formulation

Given training data, let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ denote feature matrix with n samples and the feature dimension is d. The corresponding label matrix is given by $\mathbf{Y} = [\mathbf{y}^1; \ldots; \mathbf{y}^n] \in \{-1, 1\}^{n \times m}$ where $\mathbf{y}^i$ is a row vector of the labels for the i-th example, and m is the number of class labels. The general formulation of regularized regression-based feature selection methods (Nie et al. 2010; Han and Kim 2015), which aim to obtain a projection matrix $\mathbf{W} \in \mathbb{R}^{d \times m}$, can be summarized as follows:

$$\min_{\mathbf{W}} \mathcal{L}(\mathbf{X}^T\mathbf{W} - \mathbf{Y}) + \lambda \mathcal{R}(\mathbf{W}), \tag{5}$$

where $\mathcal{L}(\cdot)$ is the norm-based loss function of the prediction residual, $\mathcal{R}(\cdot)$ is the regularizer that introduces sparsity to make $\mathbf{W}$ applicable for feature selection, and λ is a trade-off parameter. For simplicity, the bias has been absorbed into $\mathbf{W}$ by adding a constant value 1 to the feature vector of each example. Such methods have been widely used in both multi-class and multi-label learning tasks (Nie et al. 2010; Kong and Ding 2014; Han and Kim 2015; Xu, Tao, and Xu 2016). However, they are designed to maximize the classification accuracy, which is unsuitable for highly imbalanced classes situations (Parambath, Usunier, and Grandvalet 2014), since equal costs are assigned to different classes.

To solve the class imbalance problem, we present a new feature selection method, which optimizes F-measure by modifying the classifiers of regularized regression-based feature selection into cost-sensitive. Without loss of generality, we start with the illustration on the cost-sensitive feature selection under a binary-class setting, where the label vector is $[y_1; y_2; \ldots; y_n] \in \{-1, 1\}^{n \times 1}$. As mentioned previously, the cost for positive class is $1 + \beta^2 - r$ and the cost for negative class is r. Thus for each class, we obtain a cost vector $\mathbf{c} = [c_1, \ldots, c_n]^T \in \mathbb{R}^n$, where $c_i = 1 + \beta^2 - r$ if $y_i = 1$, and $c_i = r$ if $y_i = -1$. The formulation of total cost for all samples can be given as follows:

$$\min_{\mathbf{w}} \sum_{i=1}^{n} \mathcal{L}((\mathbf{x}_i^T\mathbf{w} - y_i) \cdot c_i) + \lambda \mathcal{R}(\mathbf{w}), \tag{6}$$

where $\mathbf{w} \in \mathbb{R}^{d \times 1}$ is the projection vector. In multi-class and multi-label scenarios, the cost vector $\mathbf{c}_i \in \mathbb{R}^n$ for the i-th class can be obtained according to their per-class false negative/false negative cost generated by corresponding cost function $\mathbf{a}(r)$. Denoting the cost matrix as $\mathbf{C} =$

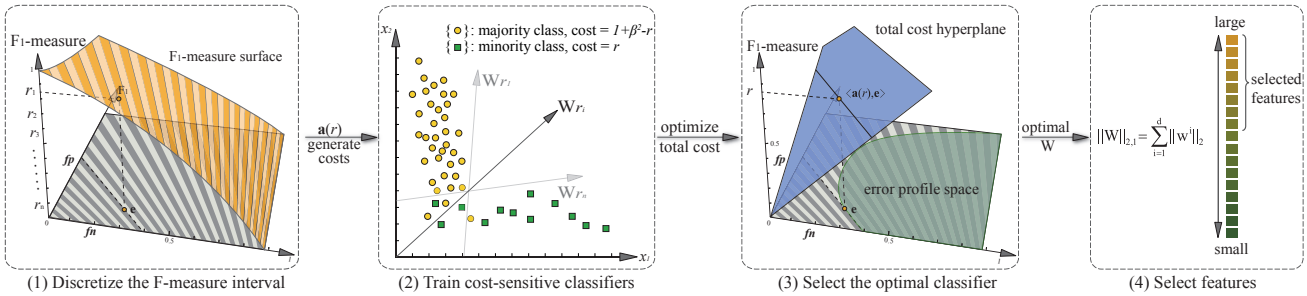

Figure 3: System diagram of the proposed cost-sensitive feature selection (CSFS) model in the case of binary classification. This model can be divided into four stages. (1) Discretize the F-measure interval to obtain a set of evenly spaced values $\{r_1, \ldots, r_n\}$. (2) For a given r_i, cost function $\mathbf{a}(r_i)$ generates costs $1+\beta^2-r_i$ for the false negative and r_i for the false positive, thus we can get a series of cost-sensitive classifiers. (3) Select the optimal classifier with the largest F-measure value on the validation set. (4) Select the top-ranking features according to the projection matrix $\mathbf{W}$ of the optimal classifier by sorting $\|\mathbf{w}^i\|$ $(1 \leq i \leq d)$ in descending order.

$[\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_m] \in \mathbb{R}^{n \times m}$, we obtain the following formulation:

$$\min_{\mathbf{W}} \sum_{i=1}^{n} \mathcal{L}((\mathbf{x}_i^T\mathbf{W} - \mathbf{y}^i) \odot \mathbf{c}^i) + \lambda \mathcal{R}(\mathbf{W}), \tag{7}$$

where $\mathbf{c}^i$ is the i-th row of $\mathbf{C}$ corresponding to the i-th example. Due to the rotational invariant property and robustness to outliers (Nie et al. 2010), we adopt ℓ_2-norm based loss function as the specific form of $\mathcal{L}(\cdot)$ and the optimization problem becomes:

$$\min_{\mathbf{W}} \sum_{i=1}^{n} \|(\mathbf{x}_i^T\mathbf{W} - \mathbf{y}^i) \odot \mathbf{c}^i\|_2 + \lambda \mathcal{R}(\mathbf{W}). \tag{8}$$

By further considering that

$$\sum_{i=1}^{n} \|(\mathbf{x}_i^T\mathbf{W} - \mathbf{y}^i) \odot \mathbf{c}^i\|_2 = \|(\mathbf{X}^T\mathbf{W} - \mathbf{Y}) \odot \mathbf{C}\|_{2,1}, \tag{9}$$

and taking the commonly used $\ell_{2,1}$-norm as regularization (Nie et al. 2010), we obtain the following compact form of the cost-sensitive feature selection (CSFS) optimization problem:

$$\min_{\mathbf{W}} \|(\mathbf{X}^T\mathbf{W} - \mathbf{Y}) \odot \mathbf{C}\|_{2,1} + \lambda \|\mathbf{W}\|_{2,1}. \tag{10}$$

As shown in Figure 3, we can get a series of cost-sensitive feature selection problems with different cost matrix $\mathbf{C}$ corresponding to each F-measure value r. After obtaining the optimal $\mathbf{W}$, features can be selected by sorting $\|\mathbf{w}^i\|$ $(1 \leq i \leq d)$ in descending order. If $\|\mathbf{w}^i\|$ shrinks to zero, the i-th feature is less important and will not be selected.

Optimization

For a given F-measure r, the corresponding cost matrix $\mathbf{C}$ is fixed and thus $\mathbf{W}$ is the only variable in Eq. (10). Taking the derivative of the objective function with respect to $\mathbf{w}_k (1 \leq k \leq m)$ and setting it to zero, we obtain[1]:

$$\mathbf{X}\mathbf{U}_k\mathbf{G}\mathbf{U}_k\mathbf{X}^T\mathbf{w}_k - \mathbf{X}\mathbf{U}_k\mathbf{G}\mathbf{U}_k\mathbf{y}_k + \lambda\mathbf{D}\mathbf{w}_k = 0, \tag{11}$$

where diagonal matrix $\mathbf{U}_k = diag(\mathbf{c}_k)$, $\mathbf{D}$ is a diagonal matrix with the i-th diagonal element as $d_{ii} = \frac{1}{2\|\mathbf{w}^i\|_2}$ and $\mathbf{G}$ is a diagonal matrix with the i-th diagonal element as $g_{ii} = \frac{1}{2\|((\mathbf{X}^T\mathbf{W} - \mathbf{Y}) \odot \mathbf{C})^i\|_2}$. Each $\mathbf{w}_k$ can thus be solved in the closed form:

$$\mathbf{w}_k = (\lambda\mathbf{D} + \mathbf{X}\mathbf{U}_k\mathbf{G}\mathbf{U}_k\mathbf{X}^T)^{-1}(\mathbf{X}\mathbf{U}_k\mathbf{G}\mathbf{U}_k)\mathbf{y}_k. \tag{12}$$

Since the solution of $\mathbf{W}$ is dependent on $\mathbf{D}$ and $\mathbf{G}$, we develop an iterative algorithm to obtain the ideal $\mathbf{D}$ and $\mathbf{G}$. The whole optimization procedure is described in Algorithm 1. In each iteration, $\mathbf{D}$ and $\mathbf{G}$ are calculated with current $\mathbf{W}$, and then each column vector $\mathbf{w}_k$ of $\mathbf{W}$ is updated based on the newly solved $\mathbf{D}$ and $\mathbf{G}$. The iteration procedure is repeated until the convergence criterion is reached. The convergence of Algorithm 1 is guaranteed by the following theorem:

Theorem 1. *Algorithm 1 monotonically decreases the objective value of Eq. (10) in each iteration, that is,*

$$\begin{aligned} &\|(\mathbf{X}^T\mathbf{W}_{t+1} - \mathbf{Y}) \odot \mathbf{C}\|_{2,1} + \lambda\|\mathbf{W}_{t+1}\|_{2,1} \leq \\ &\|(\mathbf{X}^T\mathbf{W}_t - \mathbf{Y}) \odot \mathbf{C}\|_{2,1} + \lambda\|\mathbf{W}_t\|_{2,1}. \end{aligned} \tag{13}$$

Due to the limited space, the proof of Theorem 1 is not presented here. In a nutshell, according to (Nie et al. 2010), the objective value of Eq. (10) monotonically decreases in each iteration.

[1]When $\|\mathbf{w}^i\|_2 = 0$, Eq. (10) is not differnetiable. This problem can be solved by introducing a small perturbation to regularize d_{ii} as $\frac{1}{2\sqrt{\|\mathbf{w}^i\|_2^2+\zeta}}$. Similarly, the i-th diagonal element g_{ii} of $\mathbf{G}$ can be regularized as $\frac{1}{2\sqrt{\|((\mathbf{X}^T\mathbf{W}-\mathbf{Y})\odot\mathbf{C})^i\|_2^2+\zeta}}$. It can be verified that the derived algorithm minimizes the following problem: $\sum_{i=1}^{n}\sqrt{\|((\mathbf{X}^T\mathbf{W}-\mathbf{Y})\odot\mathbf{C})^i\|_2^2+\zeta} + \lambda\sum_{i=1}^{d}\sqrt{\|\mathbf{w}^i\|_2^2+\zeta}$, which is apparently reduced to Eq. (10) when $\zeta \to 0$.

Algorithm 1 An iterative algorithm to solve the optimization problem in Eq. (10).

Input: $\mathbf{X} \in \mathbb{R}^{d\times n}$, $\mathbf{Y} \in \mathbb{R}^{n\times m}$ and $\mathbf{C} \in \mathbb{R}^{n\times m}$.
Output: $\mathbf{W} \in \mathbb{R}^{d\times m}$.

1: Initialize $\mathbf{W}_0$ as a random matrix, $t = 0$.
2: **while** not converging **do**
3: **update** diagonal matrix $\mathbf{D}_{t+1}$ where the i-th diagonal element is $\frac{1}{2\|\mathbf{w}_t^i\|_2}$.
4: **update** diagonal matrix $\mathbf{G}_{t+1}$ where the i-th diagonal element is $\frac{1}{2\|((\mathbf{X}^T\mathbf{W}_t-\mathbf{Y})\odot\mathbf{C})^i\|_2}$.
5: **for** $k \leftarrow 1$ **to** m **do**
6: $\mathbf{U}_k = diag(\mathbf{c}_k)$.
7: $(\mathbf{w}_{t+1})_k = (\lambda\mathbf{D}_{t+1} + \mathbf{X}\mathbf{U}_k\mathbf{G}_{t+1}\mathbf{U}_k\mathbf{X}^T)^{-1}$
8: $\cdot(\mathbf{X}\mathbf{U}_k\mathbf{G}_{t+1}\mathbf{U}_k)\mathbf{y}_k$.
9: **end for**
10: $t = t + 1$.
11: **end while**

Complexity Analysis

In Algorithm 1, step 3 and step 4 calculate the diagonal elements which are computationally trivial, so the complexity mainly depends on the matrix multiplication and inversion in step 7. By using sparse matrix multiplication and avoiding dense intermediate matrices, the complexity of updating each $(\mathbf{w}_{t+1})_k$ is $O(d^2(n+d))$. Thus the complexity of the proposed algorithm is $O(Ttmd^2(n+d))$, where t is the number of iterations, and T is the number of discretized values of F-measure. Empirical results show that the convergence of Algorithm 1 is rapid and t is usually less than 50. Besides, T is usually less than 20. Therefore, the proposed algorithm is quite efficient.

Experiments

Extensive experiments are conducted on synthetic, multi-class and multi-label datasets. For multi-class classification, we use two datasets: handwritten digit dataset USPS[2] and face image dataset YaleB[2]. For multi-label classification, we use MSVCv2[3] and TRECVID2005[4] datasets. Following the previous works (Kong and Ding 2014; Kong et al. 2012), the 384-dimensional color moment features are extracted on MSRC, and the 512-dimensional GIST features on TRECVID. For each dataset, we randomly select $1/3$ of the training samples for validation to tune the hyper-parameters. For datasets that do not have a separate test set, the data is first split to keep $1/4$ for testing. A summary of multi-class and multi-label datasets is shown in Table 1.

During the training process, the parameter λ in our method is optimized in the range of $\{10^{-6}, 10^{-5}, \ldots, 10^{6}\}$, and the number of selected features is set as $\{20, 30, \ldots, 120\}$. To fairly compare all different feature selection methods, classification experiments are conducted on all datasets using 5-fold cross validation SVM with linear kernel and parameter $C = 1$. We repeat the experiments 10 times with random seeds for generating the validation sets. Both mean and standard deviation of the accuracy and F_1-measures are reported.

[2]http://www.cad.zju.edu.cn/home/dengcai/Data/MLData.html

[3]http://research.microsoft.com/en-us/projects/objectclassrecognition/

[4]http://www-nlpir.nist.gov/projects/tv2005/

Table 1: Information of multi-class and multi-label datasets.

Datasets		Classes	Samples	Features
Multi-class	USPS	10	9258	256
	YaleB	38	2414	1024
Multi-label	MSRC	23	591	384
	TRECVID	39	3721	512

Synthetic Data

To demonstrate the advantage of cost-sensitive feature selection over traditional cost-blind feature selection, a toy experiment is performed to show the influence of the costs on the selected features. We construct a two-dimensional binary-class synthetic dataset based on two different uniform distributions, as shown in Figure 4. The ratio of majority class to minority class is $3:1$. In this experiment, majority class is treated as the positive class, and minority class as the negative class.

For a given linear classifier, each coefficient of its projection vector $\mathbf{w}$ corresponds to one feature weight such as w_1 for x_1, then the features with larger coefficients will be selected. The projection vector $\mathbf{w}$ varies with the costs assigned to both classes. In Figure 4(a), the cost of majority class is larger than the cost of minority class when $r < 1$. In Figure 4(b), the costs for both classes are the same when $r = 1$. In this case, the cost-sensitive feature selection degenerates to the cost-blind feature selection. When $r > 1$, as shown in Figure 4(c), the cost of majority class is smaller than the cost of minority class. It is worth noting that the weight of feature x_1 is larger than the weight of feature x_2, which is different from the first two examples. Therefore, different features will be selected from different cost-sensitive feature selection problems.

Multi-Class Datasets

On multi-class datasets, CSFS is compared with several popular and representative multi-class feature selection methods, such as ReliefF (Kononenko 1994), Information Gain (IG) (Raileanu and Stoffel 2004), mRMR (Peng, Long, and Ding 2005), F-statistic (Liu and Motoda 2012) and RFS (Nie et al. 2010).

The multi-class classification results in terms of micro-F_1-measure and accuracy is shown in Figure 5. Table 2 shows the results of different feature selection methods on their best dimensions. We observe that: (1) the proposed CSFS is superior to other multi-class feature selection methods consistently in terms of the micro-F_1-measure on both USPS and YaleB datasets; (2) in terms of accuracy, CSFS outperforms other methods on most of the feature subsets.

Multi-Label Datasets

On each multi-label dataset, CSFS is compared with five competitive multi-label feature selection methods: multi-

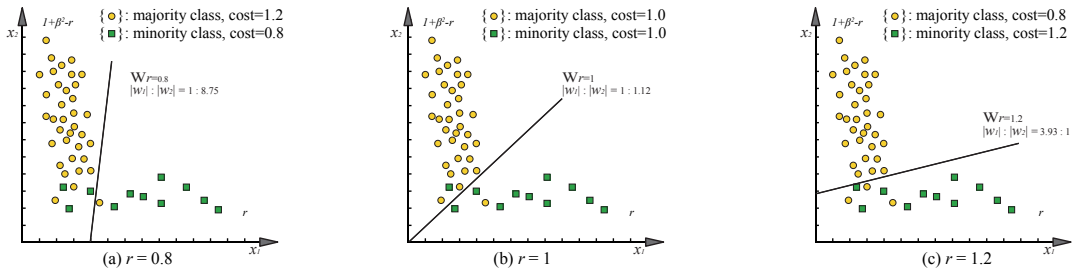

Figure 4: Illustration of how costs influence the feature weights on a two-dimensional synthetic dataset.

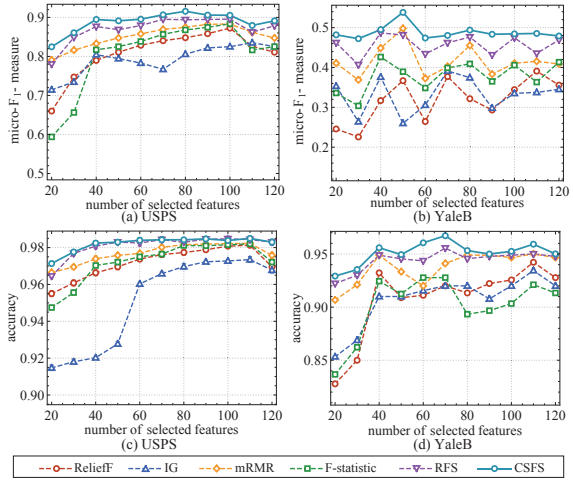

Figure 5: Multi-class classification results using SVM in terms of multi-class micro-F_1-measure and accuracy.

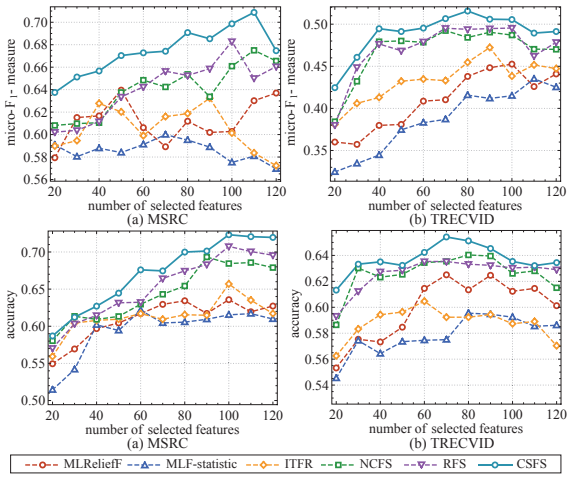

Figure 6: Multi-label classification results using SVM in terms of multi-label micro-F_1-measure and accuracy.

Table 2: Multi-class micro-F_1-measure (%± std) and accuracy (%± std) of multi-class feature selection methods.

	Micro-F_1-measure		Accuracy	
	USPS	YaleB	USPS	YaleB
ReliefF	87.25±0.71	39.09±6.55	98.13±1.39	94.22±0.34
IG	83.51±1.06	39.13±1.47	97.35±2.96	93.44±1.22
mRMR	88.30±0.87	49.80±9.44	98.23±0.76	95.00±0.85
F-statistic	88.36±0.84	42.64±1.13	98.20±1.69	92.78±2.70
RFS	89.54±0.62	48.68±8.54	**98.50±0.95**	95.56±1.51
CSFS	**91.56±0.56**	**53.83±1.63**	**98.50±0.56**	**96.72±0.41**

Table 3: Multi-label micro-F_1-measure (%± std) and accuracy (%± std) of multi-label feature selection methods.

	Micro-F_1-measure		Accuracy	
	MSRC	TRECVID	MSRC	TRECVID
MLReliefF	63.96±0.49	45.25±0.71	63.57±1.41	62.46±0.88
MLF-statistic	59.99±1.87	43.51±0.36	62.15±1.50	59.55±4.27
ITFR	63.24±1.07	47.23±0.59	65.70±0.32	60.46±3.75
NCFS	67.50±1.96	49.23±0.80	69.32±1.64	64.04±3.57
RFS	68.29±0.93	49.54±0.62	70.75±1.02	63.53±0.77
CSFS	**70.88±0.77**	**51.56±0.56**	**72.32±0.46**	**65.42±0.43**

label ReliefF (MLReliefF) (Kong et al. 2012), multi-label F-statistic (MLF-statistic) (Kong et al. 2012), information-theoretic feature ranking (ITFR) (Lee and Kim 2015), non-convex feature selection (NCFS) (Kong and Ding 2014) and RFS (Nie et al. 2010). Particularly, RFS can be extended for multi-label feature selection task (Kong and Ding 2014).

Figure 6 shows the classification results in terms of multi-label micro-F_1-measure and accuracy on MSRC and TRECVID datasets. Table 3 shows the results of each feature selection method on its best performing dimension. From the results, we can observe that: (1) the methods using joint sparse regularization, such as CSFS, NCFS and RFS, show better performances than other feature selection methods that only use the statistical information of the original features. This is because the projection matrices of these methods are determined at the same time during the optimization procedure, corresponding features are selected to prevent high correlation (Han and Kim 2015); (2) Our method outperforms these methods significantly under the F-measure criterion, and does not lead to obvious decrement to accuracy. In particular, our method outperforms other methods by a relative improvement between 3%-10% in terms of micro-F_1-measure.

Conclusion

In this paper, we proposed a cost-sensitive feature selection method by optimizing F-measure instead of accuracy to tackle the class imbalance problem. Due to the neglect of class imbalance issue, traditional feature selection methods such as regularized regression-based methods usually select the feature subset by maximizing the classification accuracy to choose the features. Thus the selected features are biased towards the majority classes. Under the imbalanced classes setting, F-measure is a more suitable performance measure than accuracy. Motivated by the reduction of F-measure optimization to cost-sensitive classification, we modify the classifiers of regularized regression-based feature selection into cost-sensitive by generating and assigning different costs to each class. Features will be selected according to the classifier with optimal F-measure. Therefore, the selected features will fully represent for all classes. Extensive experiments have been performed on synthetic, multi-class and multi-label datasets. The results demonstrate the effectiveness of our method.

Acknowledgements

This research is partially supported by grants from NSFC 61375026 and 2015BAF15B00, and ARC FT-130101457, DP-140102164 and LE-140100061.

References

Dembczynski, K. J.; Waegeman, W.; Cheng, W.; and Hüllermeier, E. 2011. An exact algorithm for F-measure maximization. In *NIPS*, 1404–1412.

Dembczynski, K. J.; Jachnik, A.; Kotlowski, W.; Waegeman, W.; and Hüllermeier, E. 2013. Optimizing the F-measure in multi-label classification: Plug-in rule approach versus structured loss minimization. In *ICML*, 1130–1138.

Guyon, I.; Weston, J.; Barnhill, S.; and Vapnik, V. 2002. Gene selection for cancer classification using support vector machines. *Machine Learning* 46(1-3):389–422.

Hall, M. A., and Smith, L. A. 1999. Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper. In *FLAIRS*, 235–239.

Han, D., and Kim, J. 2015. Unsupervised simultaneous orthogonal basis clustering feature selection. In *CVPR*, 5016–5023.

Jansche, M. 2005. Maximum expected F-measure training of logistic regression models. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 692–699.

Kim, J.; Wang, Y.; and Yasunori, Y. 2013. The genia event extraction shared task, 2013 edition-overview. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, 8–15.

Kong, D., and Ding, C. 2014. Non-convex feature learning via $\ell_{p,\infty}$ operator. In *AAAI*, 1918–1924.

Kong, D.; Ding, C.; Huang, H.; and Zhao, H. 2012. Multi-label ReliefF and F-statistic feature selections for image annotation. In *CVPR*, 2352–2359.

Kononenko, I. 1994. Estimating attributes: analysis and extensions of RELIEF. In *ECML*, 171–182.

Koyejo, O. O.; Natarajan, N.; Ravikumar, P. K.; and Dhillon, I. S. 2014. Consistent binary classification with generalized performance metrics. In *NIPS*, 2744–2752.

Lee, J., and Kim, D. 2015. Fast multi-label feature selection based on information-theoretic feature ranking. *Pattern Recognition* 48(9):2761–2771.

Lewis, D. D. 1995. Evaluating and optimizing autonomous text classification systems. In *SIGIR*, 246–254.

Liu, H., and Motoda, H. 2012. *Feature selection for knowledge discovery and data mining*, volume 454. Springer Science & Business Media.

Luo, Y.; Wen, Y.; Tao, D.; Gui, J.; and Xu, C. 2016. Large margin multi-modal multi-task feature extraction for image classification. *IEEE Transactions on Image Processing* 25(1):414–427.

Narasimhan, H.; Vaish, R.; and Agarwal, S. 2014. On the statistical consistency of plug-in classifiers for non-decomposable performance measures. In *NIPS*, 1493–1501.

Nie, F.; Huang, H.; Cai, X.; and Ding, C. H. 2010. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *NIPS*, 1813–1821.

Parambath, S. P.; Usunier, N.; and Grandvalet, Y. 2014. Optimizing F-measures by cost-sensitive classification. In *NIPS*, 2123–2131.

Peng, H.; Long, F.; and Ding, C. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *TPAMI* 27(8):1226–1238.

Pillai, I.; Fumera, G.; and Roli, F. 2012. F-measure optimisation in multi-label classifiers. In *ICPR*, 2424–2427.

Qian, M., and Zhai, C. 2013. Robust unsupervised feature selection. In *IJCAI*, 1621–1627.

Raileanu, L. E., and Stoffel, K. 2004. Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence* 41(1):77–93.

Tang, J.; Alelyani, S.; and Liu, H. 2014. Feature selection for classification: A review. *Data Classification: Algorithms and Applications* 37.

Tsochantaridis, I.; Joachims, T.; Hofmann, T.; and Altun, Y. 2005. Large margin methods for structured and interdependent output variables. *JMLR* 6:1453–1484.

Villela, S. M.; de Castro Leite, S.; and Neto, R. F. 2015. Feature selection from microarray data via an ordered search with projected margin. In *IJCAI*, 3874–3881.

Wang, S.; Tang, J.; and Liu, H. 2015. Embedded unsupervised feature selection. In *AAAI*, 470–476.

Xu, C.; Tao, D.; and Xu, C. 2015. Large-margin multi-label causal feature learning. In *AAAI*, 1924–1930.

Xu, C.; Tao, D.; and Xu, C. 2016. Robust extreme multi-label learning. In *ACM SIGKDD Conferences on Knowledge Discovery and Data Mining (KDD)*, 1275–1284.

Yang, Y. 2001. A study of thresholding strategies for text categorization. In *SIGIR*, 137–145.

Ye, N.; Chai, K. M. A.; Lee, W. S.; and Chieu, H. L. 2012. Optimizing F-measures: a tale of two approaches. In *ICML*.

Zhao, Z.; Wang, L.; Liu, H.; et al. 2010. Efficient spectral feature selection with minimum redundancy. In *AAAI*, 673–678.