Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence

Ensemble Feature Weighting Based on Local Learning and Diversity

Yun Li and **Suyan Gao**
College of Computer Science
Nanjing University of Posts and Telecommunications
Nanjing, P.R.China
liyun@njupt.edu.cn

Songcan Chen
College of Computer Science and Technology
Nanjing University of Aeronautics and Astronautics
Nanjing, P.R.China
s.chen@nuaa.edu.cn

Abstract

Recently, besides the performance, the stability (robustness, i.e., the variation in feature selection results due to small changes in the data set) of feature selection is received more attention. Ensemble feature selection where multiple feature selection outputs are combined to yield more robust results without sacrificing the performance is an effective method for stable feature selection. In order to make further improvements of the performance (classification accuracy), the diversity regularized ensemble feature weighting framework is presented, in which the base feature selector is based on local learning with logistic loss for its robustness to huge irrelevant features and small samples. At the same time, the sample complexity of the proposed ensemble feature weighting algorithm is analyzed based on the VC-theory. The experiments on different kinds of data sets show that the proposed ensemble method can achieve higher accuracy than other ensemble ones and other stable feature selection strategy (such as sample weighting) without sacrificing stability.

Introduction

The high dimensionality of data poses challenges to learning tasks due to the curse of dimensionality. In the presence of many irrelevant features, learning models tend to overfit and become less comprehensible. Feature selection is an important and frequently used technique in data mining for dimension reduction via removing irrelevant and redundant features and has been an active area for decades. Various studies show that features can be removed without performance deterioration (Ng 2004). Then feature selection brings the immediate effects of speeding up a data mining algorithm, improving learning accuracy, and enhancing model comprehensibility (Zhao 2010), and it has been widely applied to many research fields such as genomic analysis (Inza et al. 2004), text mining (Forman 2003), etc. A comprehensive surveys of existing feature selection techniques and a general framework for their unification can be found in (Zhao 2010; Liu and Yu 2005; Guyon et al. 2006; Guyon and Elisseeff 2003).

Feature selection algorithms designed with different strategies broadly fall into three categories: filter, wrapper and embedded models (Liu and Yu 2005). Compared to wrapper and embedded models, feature selection algorithms under filter model rely on analyzing the general characteristics of data and evaluating features without involving any learning algorithm, therefore most of them do not have bias on specific learner models, which is believed to be one advantage of the filter model. Another advantage of the filter model is that it has very simple structure, generally consists of straightforward search strategy and feature evaluation criterion. Because of its simple structure, it is easy to design and understand for other researchers. This explains that why most feature selection algorithms are of filter model. Moreover, since its structure is simple, it is usually very fast (Zhao 2010), and is always appropriate for high dimensional data preprocessing. On the other hand, according to the type of the output, feature selection algorithms can be divided into either feature weighting algorithms or feature ranking algorithms, such as Relief (Kira and Rendell 1992; Kononenko 1994; Robnik-Sikonja and Kononenko 2003; Sun 2007), Lmba (Li and Lu 2009) and SQP-FW (Takeuchi and Sugiyama 2011), etc, or subset selection algorithms, such as SVM-RFE (Guyon et al. 2002) and MRSF (Zhao, Wang, and Liu 2010), etc.

Various feature selection algorithms have been developed with a focus on improving classification accuracy while reducing dimensionality (Zhao 2010; Liu and Yu 2005; Guyon et al. 2006; Wasikowski and Chen 2010). Besides high accuracy, another important issue is stability of feature selection - the insensitivity of the result of a feature selection algorithm to variations of the training set (Saeys, Abeel, and de Peer 2008; Han and Yu 2010; Loscalzo, Yu, and Ding 2009). This issue is very important for the applications where feature selection is used as a knowledge discovery tool to identify characteristic markers and to explain the observed phenomena. For example, in microarray analysis, biologists are interested in finding a small number of features (genes or proteins) that can explain the behavior mechanisms of microarray samples. A feature selection algorithm often selects largely different subsets of features under variations to the training data, although most of these subsets are as good as each other in terms of classification performance (Loscalzo, Yu, and Ding 2009). Such instability will be confusion, and dampen the confidence of domain experts in experimentally validating the selected features.

Copyright © 2012, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Similar to the case of supervised learning, ensemble techniques might be used to improve the robustness of feature selection techniques (Saeys, Abeel, and de Peer 2008; Loscalzo, Yu, and Ding 2009). Indeed, for sample samples with high dimension, it is often reported that several different feature subsets may yield equally optimal results (Saeys, Inza, and Larranaga 2007), and the risk of choosing an unstable subset may be reduced by ensemble feature selection. Furthermore, different feature selection algorithms may yield feature subsets that can be considered as local optima in the space of feature subsets, and result of ensemble feature selection might be closer to the optimal subset or ranking of features. Finally, the representational power of a particular feature selector might constrain its search space such that optimal subsets cannot be reached. Ensemble feature selection could alleviate this problem by integrating the outputs of several feature selectors (Saeys, Abeel, and de Peer 2008). Ensemble feature selection has been successfully applied in biomarker identification (Abeel et al. 2010).

It is important to note that robustness of feature selection results should not be considered independently, it always should combine with classification performance, as domain experts are not interested in a strategy that yields very robust feature subsets, but the returned subsets do not perform well. Hence, these two aspects (stability and performance) need always be investigated together. Then an ensemble feature weighting algorithm with high performance and stability is our aim. In this study, especially for one type of feature selection-feature weighting, we present a framework about diversity regularized ensemble feature weighting, and its sample complexity is also presented. The base feature selector in ensemble is based on local learning, which is under filter model and outputs a feature weights (measuring features' relevance) vector. .

Ensemble Feature Weighting

Components of Ensemble Feature Selection

Same to the ensemble models for supervised learning, there are two essential steps in ensemble feature selection: creating a set of different feature selectors with outputs and aggregating the results of all feature selectors (Saeys, Abeel, and de Peer 2008).

To measure the effect of ensemble feature selection, we adopt a subsampling based strategy. Consider a training set $\mathbf{X}$ contains n samples, $\mathbf{X} = \{\mathbf{x}_i, y_i\}_{i=1}^n$, and each sample $\mathbf{x}_i$ is represented by an d-dimensional vector $\mathbf{x}_i \in \mathcal{R}^d$ and discrete class labels y_i. Then m subsamples of size $\beta n (0 < \beta < 1)$ are drawn randomly from $\mathbf{X}$, where the parameters m and β can be varied. Subsequently, feature selection is performed on each of the m subsamples.

Similar to the ensemble learning, the basic idea of our ensemble feature weighting analysis is to maximize the fit of the feature weighting vector, while maximizing the diversity between vectors. Therefore, ensemble feature selection generates the feature weighting results ensemble $\mathbf{E} = \{\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_m\}$, where $\mathbf{w}_k (k = 1, 2, \cdots, m)$ represents the outcome of the k-th base feature selector trained on k-th subsample. Specifically, in our case, each feature selection result $\mathbf{w}_k (k = 1, 2, \cdots, m)$ is a feature weighting vector. The results ensemble $\mathbf{E}$ can be obtained by minimizing the following loss function:

$$L(\mathbf{E}) = L_{emp}(\mathbf{E}) + \gamma . L_{div}(\mathbf{E}) \tag{1}$$

Here, the first term $L_{emp}(\mathbf{E})$ corresponds to the empirical loss of $\mathbf{E}$; the second term $L_{div}(\mathbf{E})$ corresponds to the diversity loss of $\mathbf{E}$. Furthermore, γ is the cost parameter balancing the importance of the two terms.

In this paper, we employ local learning-based logistic regression to implement the base feature selectors for its high efficiency on the huge irrelevant features (Sun, Todorovic, and Goodison 2010), then it is effective to improve the stability of feature selection. Thus, the first term $L_{emp}(\mathbf{E})$ in Eq.(1) is set to measure the empirical loss of logistic regression for feature weighting:

$$L_{emp}(\mathbf{E}) = \sum_{k=1}^{m} \sum_{x_i \in k} \log(1 + \exp(\frac{-\mathbf{w}_k^T}{m} \mathbf{z}_i)) \tag{2}$$

where $\mathbf{z}_i = |\mathbf{x}_i - NM(\mathbf{x}_i)| - |\mathbf{x}_i - NH(\mathbf{x}_i)|$, and $|.|$ is an element-wise absolute operator. $\mathbf{x}_i$ is a sample in k-th subsample. And two nearest neighbors of sample $\mathbf{x}_i$, one from the same class is called as nearest hit(NH), and the other from the different class is called as nearest miss (NM). $\mathbf{w}_k^T \mathbf{z}_i$ is the local margin for $\mathbf{x}_i$, which belongs to hypothesis margin (Crammer et al. 2002) and an intuitive interpretation of this margin is a measure as to how much the features of $\mathbf{x}_i$ can be corrupted by noise (or how much $\mathbf{x}_i$ can move in the feature space) before being misclassified. Margin (Schapire et al. 1998; Cortes and Vapnik 1995) is a geometric measure for evaluating the confidence of a classifier with respect to its decision. Margin is used both for theoretic analysis of generalization bounds and as guidelines for algorithm designs. By the large margin theory (Schapire et al. 1998), a classifier that minimizes a margin-based error function usually generalizes well on unseen test data. Then one natural idea is to scale each feature, and thus obtain a weighted feature space parameterized by a vector $\mathbf{w}_k$, so that a margin-based error function in the induced feature space is minimized.

For the purposes of this paper, we use the Manhattan distance to define the margin and nearest neighbors, while other standard definitions may also be used. Note that the defined margin only requires the information about the neighborhood of $\mathbf{x}_i$, while no assumption is made about the underlying data distribution. This means that we can transform an arbitrary nonlinear problem into a set of locally linear ones by local learning (Sun, Todorovic, and Goodison 2010). On the other hand, the optimization problem of Eq.(2) has an interesting interpretation: if $\mathbf{x}_i$ is correctly classified if and only if margin $\mathbf{w}_k^T \mathbf{z}_i \geq 0$ (i.e., on average, $\mathbf{x}_i$ is closer to the samples with the same label in the training data excluding itself than to those from other classes), then $\sum_{x_i \in k} I(\mathbf{w}_k^T \mathbf{z}_i < 0)$ is the leave-one-out (LOO) classification error induced by $\mathbf{w}_k$, where $I(.)$ is the indicator function. Since the logistic loss function is an upper bound of the misclassification loss function, up to a difference of a constant factor, the physical meaning of this base feature selector is to find a feature weight vector that can minimize the

upper bound of the LOO classification error in the induced feature space (Sun, Todorovic, and Goodison 2010).

As shown in Eq.(1), the second term $L_{div}(\mathbf{E})$ is used to characterize the diversity loss among the base feature selectors. We note that, though there is no agreement on what form of diversity should be defined, the diversity measures usually can be defined in a pairwise form, i.e., the total diversity is the sum of a pairwise difference measure. Thus we also consider a form of diversity based on pairwise difference, and then the form of diversity loss is defined as pairwise similarity. The more similar all outputs are, the higher the diversity loss measure will be. The overall diversity loss can be defined as the average over all pairwise similarity between the outputs of different feature selectors:

$$L_{div}(\mathbf{E}) = \frac{1}{m(m-1)} \sum_{k=1}^{m-1} \sum_{k'=k+1}^{m} Sim(\mathbf{w}_k, \mathbf{w}_{k'}) \quad (3)$$

where $Sim(\mathbf{w}_k, \mathbf{w}_{k'})$ represents a similarity measure between feature weighting vector $\mathbf{w}_k$ and $\mathbf{w}_{k'}$. $\mathbf{w}_k$ is a vector of length *d*, $\mathbf{w}_k = (w_k^1, w_k^2, \cdots, w_k^d)$, where $w_k^t (t = 1, 2, \cdots, d)$ represents the weight for feature *t* in *k*-th base feature selector output. Notice that the feature weighting vector is direct related to the classification error based on the margin as described above, and each feature weighting vector $\mathbf{w}_k$ is linear without the bias term, thus the direction of vector is the most important factor for the classification performance. In the paper, the cosine similarity measure is adopted with normalized feature weights to calculate the similarity between weighting vector $\mathbf{w}_k$ and $\mathbf{w}_{k'}$, then $Sim(\mathbf{w}_k, \mathbf{w}_{k'}) = \mathbf{w}_k^T \mathbf{w}_{k'}$. Note that the adding of a constant $||\mathbf{w}_k||_2^2 + ||\mathbf{w}_{k'}||_2^2$ (its value is 2) does not change the optimal solution (Yu, Li, and Zhou 2011). In this case, the diversity loss can be replaced by $||\mathbf{w}_k + \mathbf{w}_{k'}||_2^2$, i.e.

$$L_{div}(\mathbf{E}) = \frac{1}{m(m-1)} \sum_{k=1}^{m-1} \sum_{k'=k+1}^{m} ||\mathbf{w}_k + \mathbf{w}_{k'}||_2^2 \quad (4)$$

and a relaxed convex optimization problem is obtain for ensemble feature weighting loss in Eq.(1). Furthermore, the diversity loss can be considered as a *l*2-norm regularization for logistic regression, which can obtain the stable feature weighting vectors for its robustness to the rotational variation (Ng 2004). Then the proposed diversity loss term has positive effect on feature selection stability besides the classification performance.

In the end, ensemble feature selection aims to find the target model $\mathbf{E}^*$, which minimizes the loss function in Eq.(1):

$$\mathbf{E}^* = argmin_{\mathbf{w}_k} L(\mathbf{E}) \quad (5)$$

and the final ensemble feature weighting result $\mathbf{w}_e = \frac{1}{m} \sum_{k=1}^{m} \mathbf{w}_k$, where $\mathbf{w}_k \in \mathbf{E}^*$.

The target model $\mathbf{E}^*$ is found by employing gradient descent-based techniques. Accordingly, the gradients of $L(\mathbf{E})$ w.r.t the model parameters $\Theta = \{\mathbf{w}_k | 1 \le k \le m\}$ are determined as follows:

$$\frac{\partial \mathbf{L}}{\partial \Theta} = [\frac{\partial \mathbf{L}}{\partial \mathbf{w}_1}, \cdots, \frac{\partial \mathbf{L}}{\partial \mathbf{w}_k}, \cdots, \frac{\partial \mathbf{L}}{\partial \mathbf{w}_m}] \quad (6)$$

where

$$\frac{\partial \mathbf{L}}{\partial \mathbf{w}_k} = \frac{1}{\beta n} \sum_{x_i \in k} \frac{\partial \log(1 + \exp(\frac{-\mathbf{w}_k^T}{m} \mathbf{z}_i))}{\partial \mathbf{w}_k} + \frac{2\gamma}{m(m-1)} \sum_{k'=1, k' \neq k}^{m} \frac{\partial Sim(\mathbf{w}_k, \mathbf{w}_{k'})}{\partial \mathbf{w}_k} \quad (7)$$

and

$$\frac{\partial \log(1 + \exp(\frac{-\mathbf{w}_k^T}{m} \mathbf{z}_i))}{\partial \mathbf{w}_k} = -\frac{1}{m} \frac{\exp(\frac{-\mathbf{w}_k^T}{m} \mathbf{z}_i)}{1 + \exp(\frac{-\mathbf{w}_k^T}{m} \mathbf{z}_i)} \mathbf{z}_i \quad (8)$$

$$\frac{\partial Sim(\mathbf{w}_k, \mathbf{w}_{k'})}{\partial \mathbf{w}_k} = 2(\mathbf{w}_{k'} + \mathbf{w}_k) \quad (9)$$

To initialize the ensemble, each feature selector is learned from a bootstrapped sample of $\mathbf{X}$. Specifically, the corresponding feature weighting $\mathbf{w}_k$ is obtained by minimizing the objective function $\mathbf{w}_k = \min_{\mathbf{w}_k} \sum_{x_i \in k} \log(1 + \exp(-\mathbf{w}_k^T \mathbf{z}_i))$. Note that the ensemble can also be initialized in other ways, such as instantiating each $\mathbf{w}_k$ with random values, etc.

Theoretic Analysis

Now, we will firstly show the optimization of the base feature weighting logistic loss tends to optimize the upper bound of the ensemble logistic loss. Then we have the following proposition.

Proposition 1 Let $\mathbf{w}_1, \cdots, \mathbf{w}_m$ be the base feature weighting results, and $\mathbf{w}_e = \frac{1}{m} \sum_{k=1}^{m} \mathbf{w}_k$ be the ensemble feature weighting result, the loss of $\mathbf{w}_e$ is bounded as

$$l(\mathbf{w}_e) \le \sum_{k=1}^{m} l(\mathbf{w}_k) \quad (10)$$

Proof. For a training sample $\mathbf{x}_i$, the loss of base feature selector is $l(\mathbf{w}_k) = \log(1 + \exp(\frac{-\mathbf{w}_k^T}{m} \mathbf{z}_i)) \le (1 + \frac{1}{m} \mathbf{w}_k^T \mathbf{z}_i)$ (Sun, Todorovic, and Goodison 2010), then $\sum_{k=1}^{m} l(\mathbf{w}_k) = \sum_{k=1}^{m} \log(1 + \exp(\frac{-\mathbf{w}_k^T}{m} \mathbf{z}_i)) \le (\frac{1}{m} \sum_{k=1}^{m} \mathbf{w}_k^T \mathbf{z}_i + m)$. While the ensemble loss is $l(\mathbf{w}_e) = \log(1 + \exp(\frac{-\sum_{k=1}^{m} \mathbf{w}_k^T}{m} \mathbf{z}_i))$, which is less than or equal to $(\frac{1}{m} \sum_{k=1}^{m} \mathbf{w}_k^T \mathbf{z}_i + 1)$. Since in ensemble feature weighting, the number of base selectors is generally larger than 2, i.e., $m \ge 2$, the proposition is then proved.

On the other hand, this subsection presents a theoretical study of our algorithm's sample complexity. The main purpose of the analysis is to show the dependence of the generalization performance of the proposed algorithm on input data dimensionality and the diversity between base feature weighting vectors. As one can see from Eq. (1), the algorithm finds an ensemble feature weight vector and aims at minimizing an empirical logistic loss and diversity loss. Hence, it is a learning problem and the analysis can be performed under the VC-theory framework.

Let $\{\mathbf{x}, y\}$ be a pair of instance and target (class) value, which is sampled from a fixed but unknown joint distribution $p(\mathbf{x}, y)$. And y is absorbed into $\mathbf{x}$ for notation simplicity.

Given a set of real valued mapping functions $F = \{f(\mathbf{x}|\alpha) : \alpha \in \Theta\}$ parameterized by α and a loss function $L(f(\mathbf{x}|\alpha))$, we like to seek a parameter α to minimize the expected loss: $R(\alpha) = E[L(f(\mathbf{x}|\alpha))] = \int L(f(\mathbf{x}|\alpha))p(x)dx$. In real applications, the true distribution is rarely known, and only a limited number of instances $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ are available, which are independently drawn from the unknown distribution. A natural method to find a parameter α through minimizing the empirical loss: $R(\alpha, \mathbf{X}) = \frac{1}{n}\sum_{i=1}^n L(f(\mathbf{x}_i|\alpha))$. We like to know how well a learning algorithm trained on a limited number of samples will perform on unseen data. This can be studied based on the VC theory, which depends on the uniform convergence of the empirical loss to the expected loss (Sun, Todorovic, and Goodison 2010). It has been proved by (Vapnik 1998; Anthony and Bartlett 1999) that if the bound $\sup_{\alpha\in\Omega}|R(\alpha, \mathbf{X}) - R(\alpha)|$ is tight, then the function that minimizes the empirical loss is likely to have an expected loss that is close to the best in the function class. A theorem provides an upper bound on the rate of the uniform convergence of a class of functions in terms of its covering number (Pollard 1984). Before we present the theorem, we first give the concept of covering number.

Definition 1 (Covering Number) Given n samples $\mathbf{X} = \{\mathbf{x}_i\})_{i=1}^n$ and a function space F, characterized $f \in F$ using a vector $\upsilon_X(f) = [f(\mathbf{x}_1), \cdots, f(\mathbf{x}_n)]$ in a metric space $\mathbf{B}^n$ with metric κ. The covering number $N_p(F, \epsilon, \mathbf{X})$ is the minimum number g of vectors $\mathbf{u}_1, \cdots, \mathbf{u}_g \in \mathbf{B}^n$ with margin ϵ such that, for all $f \in F$ there exists $j \in \{1, \cdots, g\}$,

$$||\kappa(\upsilon_X(f), \mathbf{u}_j)||_p = (\sum_{i=1}^n \kappa(f(\mathbf{x}_i), u_{ji})^p)^{1/p} \leq n^{1/p}\epsilon \tag{11}$$

and $N_p(F, \epsilon, n) = \sup_X N_p(F, \epsilon, \mathbf{X})$

Lemma 1 ((Pollard 1984)) For all $\epsilon > 0$ and distribution $p(x)$, we have

$P[\sup_{\alpha\in\Theta}|R(\alpha, \mathbf{X}) - R(\alpha)| > \epsilon]$

$$\leq 8E[N_1(L, \epsilon/8, \mathbf{X})]\exp(\frac{-n\epsilon^2}{128M^2}) \tag{12}$$

where $M = \sup_{\alpha,x} L(\alpha, \mathbf{x}) - \inf_{\alpha,x} L(\alpha, \mathbf{x})$ and N_1 is the 1-norm covering number of function class L.

Lemma 1 indicates that the bound of generalization error is related to the performance on the training set and the space complexity defined by covering number.

In general, it is very difficult to estimate the covering number of an arbitrary function class. Fortunately, there exists a tight bound for linear function class as described in (Zhang 2002), which can be used for estimating the covering number for our purposes.

Lemma 2 ((Zhang 2002)) Let $F = \{\mathbf{w}^T\mathbf{x}, ||\mathbf{w}||_2 \leq a, ||\mathbf{x}||_2 \leq b, \mathbf{x} \in \Re^d\}$. Then we have

$$log_2^{N_2(F,\epsilon,n)} \leq \lceil\frac{a^2b^2}{\epsilon^2}\rceil \log_2^{(2d+1)} \tag{13}$$

where $\lceil\psi\rceil$ is the nearest integers of ψ towards infinity.

In our case, for a given data set $\mathbf{X}$, we can obtain a transformed data set $\mathbf{Z} = [\mathbf{z}_1, \cdots, \mathbf{z}_n], \mathbf{z} \in \Re^d$ based on the definition of $\mathbf{z}$ above. Define a class of linear functions $G = \{g(\mathbf{z}) = \mathbf{w}^T\mathbf{z}, ||\mathbf{w}||_2 \leq a, ||\mathbf{z}||_2 \leq b, \mathbf{z} \in \Re^d\}$. And $\sum_{x_i\in k} I(\mathbf{w}_k^T\mathbf{z}_i < \epsilon)$ is the training error with margin ϵ induced by $\mathbf{w}_k$, where $I(.)$ is the indicator function. By Definition 1 and Lemma 2, the covering number $log_2^{N_2(G,\epsilon,n)} \leq \lceil\frac{a^2b^2}{\epsilon^2}\rceil \log_2^{(2d+1)}$.

Moreover, for the ensemble feature weighting results $\mathbf{w}_e$, we consider if the diversity can constrain the norm of $\mathbf{w}_e$, and then affect the covering number. Define a class of ensemble linear functions $G_e = \{g_e(\mathbf{z}) = {\mathbf{w}_e}^T\mathbf{z}\}$ satisfying $\mathbf{w}_e = \frac{1}{m}\sum_{k=1}^m \mathbf{w}_k$ and diversity is larger than q. We measure their diversity using the angle between them, then for all base feature weighting vectors, we have $1 - \frac{\mathbf{w}_k^T\mathbf{w}_{k'}}{||\mathbf{w}_k||||\mathbf{w}_{k'}||} \geq q$ $(k, k' = 1, \cdots, m)$, so that, $\mathbf{w}_k^T\mathbf{w}_{k'} \leq (1-q)||\mathbf{w}_k||||\mathbf{w}_{k'}|| \leq (1-q)a^2$. Since $\mathbf{w}_e = \frac{1}{m}\sum_{k=1}^m \mathbf{w}_k$ and $||\mathbf{w}_e||_2^2 = {\mathbf{w}_e}^T\mathbf{w}_e$, then according to the Theorem 1 in (Yu, Li, and Zhou 2011)

$$\begin{aligned} ||\mathbf{w}_e||_2^2 &= \frac{1}{m^2}\sum_{k=1}^m ||\mathbf{w}_k||_2^2 + \frac{2}{m^2}\sum_{k=1}^{m-1}\sum_{k'=k+1}^m \mathbf{w}_k^T\mathbf{w}_{k'} \\ &\leq \frac{1}{m}a^2 + (1-q)a^2 \end{aligned} \tag{14}$$

And then the covering number for the G_e is

$$\log_2^{N_2(G_e,\epsilon,n)} \leq \lceil\frac{(\frac{1}{m}a^2 + (1-q)a^2)b^2}{\epsilon^2}\rceil \log_2^{(2d+1)} \tag{15}$$

From the definition of the covering number and Jensen's inequality, we have $N_1 \leq N_2$. Now let us consider the function class $L = \{l(g_e(\mathbf{z})) : g_e \in G_e\}$. In the proposed algorithm, $l(g_e(\mathbf{z})) = \log(1 + \exp(-g_e(\mathbf{z})))$ is a logistic loss function. It is proved in (Anthony and Bartlett 1999) that the logistic loss function is a Lipschitz function with Lipschitz constant 1, hence

$$\begin{aligned} E[N_1(L, \epsilon, \mathbf{X})] &\leq N_1(L, \epsilon, n) \leq N_1(G_e, \epsilon, n) \\ &\leq N_2(G_e, \epsilon, n) \leq (2d+1)^{\lceil\frac{(1+\frac{1}{m}-q)a^2b^2}{\epsilon^2}\rceil} \end{aligned} \tag{16}$$

By using Holder's inequality, $|l(g_e(\mathbf{z}))| = |\log(1 + \exp(-g_e(\mathbf{z})))| \leq |{\mathbf{w}_e}^T\mathbf{z}| + 1 \leq ||\mathbf{w}_e||_2||\mathbf{z}||_2 \leq \sqrt{(\frac{1}{m}a^2 + (1-q)a^2)}b$. Hence, $M = \sup_{w_e,z} L(\mathbf{w}_e, \mathbf{z}) - \inf_{w_e,z} L(\mathbf{w}_e, \mathbf{z}) \leq 2b\sqrt{(\frac{1}{m}a^2 + (1-q)a^2)}$

The covering number of ensemble function and the M value is plugged into Lemma 1, we can obtain the bound of proposed ensemble feature weighting method as follows.

Theorem 1 For the proposed ensemble algorithm, let $\mathbf{x} \in \Re^d$ and corresponding transformed data $\mathbf{z}$, which holds $||\mathbf{z}||_2 \leq b$, $\mathbf{E}$ is a feature weighting space $\{\mathbf{w}_1, \cdots, \mathbf{w}_m\}$ such that for all $\mathbf{w}_k \in \mathbf{E}$ $(k = 1, 2, \cdots, m)$ with $||\mathbf{w}_k||_2 \leq a$ and the diversity between feature weighting results is larger than q. If $\mathbf{W}$ is an ensemble feature weighting space that for a $\mathbf{w}_e \in \mathbf{W}$ satisfying $\mathbf{w}_e = \frac{1}{m}\sum_{k=1}^m \mathbf{w}_k$. For all $\epsilon > 0$ and distribution $p(x)$, we have

$P[\sup_{w_e}|R(\mathbf{w}_e, \mathbf{X}) - R(\mathbf{w}_e)| > \epsilon]$

$$\leq 8(2d+1)^{\lceil\frac{64(1+\frac{1}{m}-q)a^2b^2}{\epsilon^2}\rceil}\exp(\frac{-n\epsilon^2}{512((1+\frac{1}{m}-q)a^2b^2)}) \tag{17}$$

Stability Analysis

To measure the stability of our proposed ensemble feature weighting algorithm and the effect of diversity loss term on stability, we also adopt a subsampling based strategy. Consider the data set $\boldsymbol{S}$ with Q instances and d features. Then c subsamples of size $\mu Q(0 < \mu < 1)$ are drawn randomly from $\boldsymbol{S}$, where the parameters c and μ also can be varied. Subsequently, ensemble feature weighting is performed on each of the c subsamples, and a measure of stability or robustness is calculated.

Consider a feature weighting vector set $\mathbf{W} = \{\mathbf{w}_{e_1}, \mathbf{w}_{e_2}, \cdots, \mathbf{w}_{e_c}\}$, $\mathbf{w}_{e_j} = (w_{e_j}^1, w_{e_j}^2, \cdots, w_{e_j}^d)(j = 1, 2, \cdots, c)$ is the feature weighting result of our proposed ensemble feature weighting algorithm on j-th subsample. However, feature weighting is almost never directly used to measure the stability of feature selection, and instead converted to a ranking based on the weights (Saeys, Abeel, and de Peer 2008). Then we can obtain the corresponding ranking set $\{\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_c\}$, $\mathbf{r}_j = (r_j^1, r_j^2, \cdots, r_j^d)(j = 1, 2, \cdots, c)$, $r_j^t(t = 1, 2, \cdots, d)$ represents the rank of feature t in j-th feature ranking vector. Noting that the ranking value for a feature is set as follows: The best feature with the largest weight is assigned rank 1, and the worst one rank d.

Here, feature stability is measured by comparing similarity of the outputs of ensemble feature selection on the c subsamples. The more similar all outputs are, the higher the stability measure will be. The overall stability can be defined as the average similarity over all pairwise similarity between the different ensemble feature ranking results:

$$R_{sta} = \frac{1}{c(c-1)} \sum_{j=1}^{c-1} \sum_{j'=j+1}^{c} Sim(\mathbf{r}_j, \mathbf{r}_{j'}) \qquad (18)$$

where $Sim(\mathbf{r}_j, \mathbf{r}_{j'})$ represents a similarity measure between feature ranking results $\mathbf{r}_j$ and $\mathbf{r}_{j'}$. For feature ranking, the Spearman rank correlation coefficient (Saeys, Abeel, and de Peer 2008; Kalousis, Prados, and Hilario 2007) can be used to calculate the similarity:

$$Sim(\mathbf{r}_j, \mathbf{r}_{j'}) = 1 - 6 \sum_{t=1}^{d} \frac{(r_j^t - r_{j'}^t)^2}{d(d^2-1)} \qquad (19)$$

Note that, for the ensemble feature weighting process, we like to minimize the similarity (diversity loss) between the outputs of base feature selectors , i.e.,$\mathbf{w}_k(k = 1, \cdots, m)$, in each ensemble learning. While the computation of stability is on the outputs of ensemble feature weighting (i.e.,$\mathbf{w}_{e_j}(j = 1, \cdots, c)$) and corresponding ranking vector (i.e., $\mathbf{r}_j(j = 1, \cdots, c)$) on c subsamples. On the other hand, each of the c subsamples is used as the training data set $\mathbf{X}$ for the ensemble feature weighting and the size n is equal to $\mu Q(0 < \mu < 1)$, i.e., $n = \mu Q$. Then there is no confusion for improving ensemble learning performance through minimizing similarity between base feature weighting vector and the stability is also kept because of the mechanism of ensemble learning.

Experiments

In order to validate the performance of our ensemble algorithm, the experiments are conducted on several real-world data sets to show its stability and classification power. The data sets consist of small samples with high dimension, medium samples and large samples with low dimension, such as Colon, Prostate, Wisconsin, Sonar, Arcene, Diabetes, Waveforms and Ionosphere, which are taken from UCI ML repository (Frank and Asuncion 2010) and Colon cancer diagnosis data set is introduced in (Alon et al. 1999). They are described in Table 1. In this paper, the parameter γ is set to the value 0.001 based on cross-validation.

Table 1: Description of experimental data sets

Data set	♯ Samples	♯ Features	♯ Classes
Arcene	200	10000	2
Colon	62	2000	2
Prostate	136	12600	2
Wisconsin	699	10	2
Ionosphere	351	34	2
Sonar	208	60	2
Waveform	5000	21	3
Pima Diabetes	768	8	2

Experimental Results for Stability

To estimate the robustness (stability) of ensemble feature weighting algorithm, the strategy explained above was used with $c = 5$ subsamples of size 0.9Q (i.e. $\mu = 0.9$ and each subsample contains 90% of the data). This percentage was chosen because we want to assess robustness with respect to relatively small changes in the data set. Then, the proposed ensemble algorithm and ensemble one without diversity term with $\beta = 0.9$ was run on each subsample, the features are been ranking based on their weights, and then the similarity between feature ranking results pairs and stability is calculated using Eq.(19) and (18) respectively. The ensemble feature selection have been proved that it can improve the stability of feature selection (Saeys, Abeel, and de Peer 2008; Abeel et al. 2010), then we only show the stability of our ensemble feature weighting algorithm(EFW) and the proposed algorithm without diversity loss term(EFWWD) on two real-world data sets (such as Wisconsin and Sonar) w.r.t different numbers of base feature selectors, i.e. the value of m. The robustness of the ensemble feature weighting is shown in Fig.1 for Wisconsin and Sonar. The X-axis is the number of base feature selectors m and Y-axis is the stability value. From the results, we can observe that the stability value of these two ensemble feature weighting algorithms are very close and the difference between them does not exceed 0.01. Moreover, the stability is approaching to 1 for all values of m. For other data sets, we only list the stability value of EFW and EFWWD for $m = 5$ as follows, Arcene (0.9587, 0.9505), Pima Diabetes (0.9933, 0.9943), Colon (0.9661, 0.9592), Ionosphere (0.9725, 0.9733), Prostates (0.9687, 0.9619) and Waveform

(0.9959, 0.9966). The feature ranking stability for these data sets is also very high and close to 1, and then our proposed algorithms can obtain superior or at least equivalent stable results to other ensemble methods.

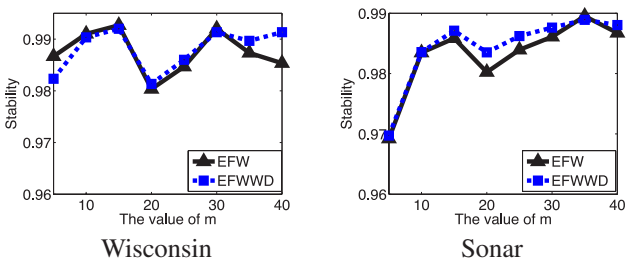

Figure 1: Experimental results of stability

Experimental Results for Classification

Now we will pay more attention to validate the classification performance of our proposed EFW, and comparing with other ensemble feature weighting algorithms, such as the the proposed EFWWD, ensemble-Relief (E-Relief) (Saeys, Abeel, and de Peer 2008) and newly proposed stable feature selection strategy based on variance reduction, which is to assign different weights to different samples based on margin, and then to obtain high stability for feature selection (Han and Yu 2010). We combine the sample weighting strategy with the newly proposed feature weighting algorithm-Lmba (Li and Lu 2009) and named as VR-Lmba. In this part of experiments, the number of base selectors for ensemble feature weighting is constant and set as 5 for all algorithms, i.e., $m = 5$. 5-cross validation is used and the linear SVM is adopted as classifier with C=1 (Chang and Lin 2002). The experimental results for these data sets are shown in Fig. 2, the X-axis is the number of selected features according to ensemble ranking results and Y-axis is the classification accuracy.

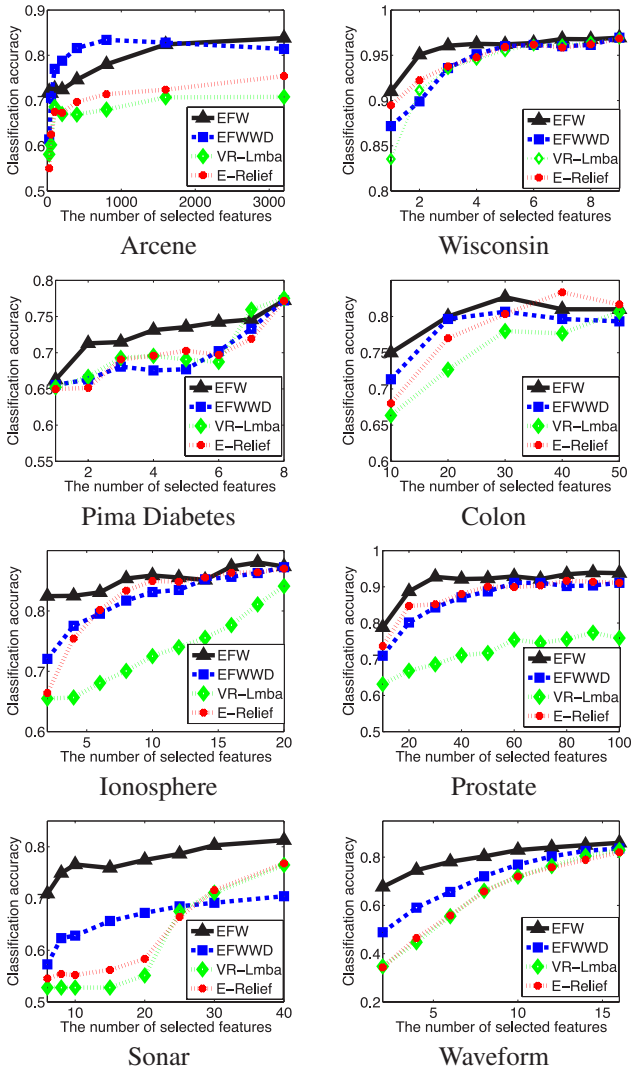

Figure 2: Experimental results of classification accuracy

Observations

From the experimental results, we can observe that our proposed ensemble algorithms, especially for EFW, can obtain higher classification accuracy than other ones in most cases, and the stability of our algorithms is also very high, then the diversity loss term in our proposed ensemble feature weighting algorithm is effective to improve the performance without deceasing the stability. Thus we achieve the goal that designing a ensemble feature weighting with high classification accuracy and stability.

Conclusion

Both the stability and performance of feature selection are attracted much attention. Our work is motivated by the recognition that diversity is the key to the success of ensemble methods and the ensemble learning can effectively improve the robustness of models. Our main contribution is that we provide the ensemble feature weighting framework for stable feature selection with high performance (classification accuracy), and present the theoretic analysis about the sample complexity with diversity constraints. In the paper, an ensemble feature weighting algorithm based on local learning and diversity is introduced, and the theoretical results about the sample complexity based on VC-theory with diversity constraints are presented. The experiments on many kinds of real-world data sets have shown its higher accuracy and at least similar stability to other ensemble ones. Then we can conclude that the diversity is also very useful for the ensemble feature selection. In our analysis, the linear combination is adopted in ensemble feature weighting, other combination scheme is our future work.

Acknowledgments

This work is, in part, supported by NSFC Grant (60973097, 61035003, 61073114 and 61100135).

References

Abeel, T.; Helleputte, T.; de Peer, Y. V.; Dupont, P.; and Saeys, Y. 2010. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics* 26:392–398.

Alon, U.; Barkai, N.; Notterman, D. A.; Gish, K.; Ybarra, S.; Mack, D.; and Levine, A. J. 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon cancer tissues probed by oligonucleotide arrays. In *Proceedings of the National Academy of Sciences of the United States of America*, 6745–6750.

Anthony, M., and Bartlett, P. L. 1999. *Neural Network Learning: Theoretical Foundations*. Cambridge, UK: Cambridge University Press.

Chang, C. C., and Lin, C. J. 2002. Libsvm: a library for support vector machines. In *http://www.csie.ntu.edu.tw/ cjlin/papers/libsvm.ps.gz*.

Cortes, C., and Vapnik, V. 1995. Support-vector networks. *Machine Learning* 20:273–297.

Crammer, K.; Bachrach, R. G.; Navot, A.; and Tishby, N. 2002. Margin analysis of the lvq algorithm. In *Advances in Neural Information Processing Systems*, 462–469.

Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research* 3:1289–1305.

Frank, A., and Asuncion, A. 2010. UCI machine learning repository. In *http://archive.ics.uci.edu/ml*.

Guyon, I., and Elisseeff, A. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research* 31:1157–1182.

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

Guyon, I.; Gunn, S.; Nikravesh, M.; and Zadeh, L. 2006. *Feature Extraction, Foundations and Applications*. Physica-Verlag, New York: Springer.

Han, Y., and Yu, L. 2010. A variance reduction for stable feature selection. In *Proceedings of the International Conference on Data Mining*, 206–215.

Inza, I.; Larranaga, P.; Blanco, R.; and Cerrolaza, A. J. 2004. Filter versus wrapper gene selection approaches in dna microarray domains. *Artificial Intelligence in Medicine* 31:91–103.

Kalousis, A.; Prados, J.; and Hilario, M. 2007. Stability of feature selection algorithms: a study on high dimensional spaces. *Knowledge and Information Systems* 12:95–116.

Kira, K., and Rendell, L. 1992. A practical approach to feature selection. In *Proceedings of International Conference on Machine Learning*, 249–256.

Kononenko, I. 1994. Estimating attributes: Analysis and extension of relief. In *Proceedings of European Conference on Machine Learning*, 171–182.

Li, Y., and Lu, B. L. 2009. Feature selection based on loss margin of nearest neighbor classification. *Pattern Recognition* 42:1914–1921.

Liu, H., and Yu, L. 2005. Toward integrating feature selection algorithms for classification and clustering. *IEEE Trans. Knowledge and Data Engineering* 17:494–502.

Loscalzo, S.; Yu, L.; and Ding, C. 2009. Consensus group stable feature selection. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 567–575.

Ng, A. Y. 2004. Feature selection, l1 vs. l2 regularization, and rotational invariance. In *Proceedings of International Conference on Machine Learning*.

Pollard, D. 1984. *Convergence of stochastic processes*. New York: Springer-Verlag.

Robnik-Sikonja, M., and Kononenko, I. 2003. Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning* 53:23–69.

Saeys, Y.; Abeel, T.; and de Peer, Y. V. 2008. Robust feature selection using ensemble feature selection techniques. In *Proceedings of the 25th European Conference on Machine Learning and Knowledge Discovery in Databases*, 313–325.

Saeys, Y.; Inza, I.; and Larranaga, P. 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics* 23:2507–2517.

Schapire, R. E.; Freud, Y.; Bartlett, P.; and Lee, W. S. 1998. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics* 26:1651–1686.

Sun, Y. J.; Todorovic, S.; and Goodison, S. 2010. Local learning based feature selection for high dimensional data analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence* 32:1–18.

Sun, Y. J. 2007. Iterative relief for feature weighting: Algorithms, theories, and applications. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 29:1035–1051.

Takeuchi, I., and Sugiyama, M. 2011. Target neighbor consistent feature weighting for nearest neighbor classification. In *Advances in Neural Information Processing Systems*, 1–9.

Vapnik, V. 1998. *Statistical Learning Theory*. New York: Wiley.

Wasikowski, M., and Chen, X. 2010. Combating the small sample class imbalance problem using feature selection. *IEEE Trans. on Knowledge and Data Engineering* 22:1388–1400.

Yu, Y.; Li, Y. F.; and Zhou, Z. H. 2011. Diversity regularized machine. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, 1603–1608.

Zhang, T. 2002. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research* 2:527–550.

Zhao, Z.; Wang, L.; and Liu, H. 2010. Efficient spectral feature selection with minimum redundancy. In *AAAI Conference on Artificial Intelligence*, 673–678.

Zhao, Z. 2010. *Spectral Feature Selection for Mining Ultrahigh Dimensional Data*. Ph.D. Dissertation, Arizona State University.