Stable L2-Regularized Ensemble Feature Weighting

Yun Li[1], Shasha Huang[1], Songcan Chen[2], and Jennie Si[3]

[1] College of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
[2] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China
[3] School of Electronic Computer and Energy Engineering, Arizona State University, Tempe 85281, USA
liyun@njupt.edu.cn, s.chen@nuaa.edu.cn, si@asu.edu

Abstract. When selecting features for knowledge discovery applications, stability is a highly desired property. By stability of feature selection, here it means that the feature selection outcomes vary only insignificantly if the respective data change slightly. Several stable feature selection methods have been proposed, but only with empirical evaluation of the stability. In this paper, we aim at providing a try to give an analysis for the stability of our ensemble feature weighting algorithm. As an example, a feature weighting method based on L2-regularized logistic loss and its ensembles using linear aggregation is introduced. Moreover, the detailed analysis for uniform stability and rotation invariance of the ensemble feature weighting method is presented. Additionally, some experiments were conducted using real-world microarray data sets. Results show that the proposed ensemble feature weighting methods preserved stability property while performing satisfactory classification. In most cases, at least one of them actually provided better or similar tradeoff between stability and classification when compared with other methods designed for boosting the stability.

1 Introduction

High dimensional data poses challenges into 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 has been an active area in machine learning for decades. It is an important and frequently used technique in data mining for dimension reduction via removing irrelevant and redundant features. Various studies show that features can be removed without performance deterioration [1]. Moreover feature selection brings the immediate effects of speeding up a data mining algorithm, enhancing generalization performances and allowing insights into the problem through the interpretation of the most relevant features [2]. Feature selection has been widely applied to many research fields such as genomic analysis [3], text mining [4], etc. A feature selection algorithm is usually associated with two important aspects: search strategy and evaluation criterion. Algorithms designed with different strategies broadly fall into three categories: filter, wrapper and embedded models [5]. Alternatively, according to the outcomes, feature selection algorithms can be divided into either feature weighting (ranking) algorithms or subset selection algorithms. A comprehensive surveys of

Z.-H. Zhou, F. Roli, and J. Kittler (Eds.): MCS 2013, LNCS 7872, pp. 167–178, 2013.

existing feature selection techniques and a general framework for their unification can be found in [2, 5–7].

Various feature selection algorithms have been developed with a focus on improving classification accuracy while reducing dimensionality [2, 5, 6]. 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 [8–10]. This issue is particularly critical for applications where feature selection is used as a knowledge discovery tool for identifying characteristic markers to explain the observed phenomena. For example, in microarray analysis, biologists are interested in finding a small number of features (genes or proteins) that explain the mechanisms driving different behaviors 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 [10]. Such instability dampens the confidence of domain experts when experimentally validating the selected features. More concrete example, in analyzing cancer biomarkers such as leukemia, the available data sets usually are high dimensional yet with small sample size. Among the thousands of genetic expression levels, a critical subset is to be discovered that links to two leukemia labels. It is therefore necessary that the selected predictive genes are common to variations of training samples. Otherwise the results will lead to less confident diagnosis. Stable feature selection has been demonstrated in biomarker identification via empirical process [11, 12]. Moreover, stability is desirable in algorithm designing since it is believed to lead to good generalization ability [13, 14].

For the stable feature selection methods, ensemble technique is among the most powerful to improve the stability of feature selection [8, 10, 15, 16]. Similar to the case of supervised learning, the general idea is to repeat the feature selection process on many randomly perturbed training sets (e.g., by bootstrapping the samples in the original training set), and aggregate the outputs in this procedure. Indeed, in large feature/small sample size domains it is often reported that several different feature subsets may yield equally optimal results [17], and ensemble feature selection may reduce the risk of choosing an unstable subset. Furthermore, different feature selection algorithms may yield feature subsets that can be considered local optima in the feature subsets, and ensemble feature selection might give a better approximation 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 help in alleviating this problem by aggregating the outputs of several feature selectors [8].

However, the existing stable feature selection algorithms are always short of one important aspect, that is a theoretic stability analysis of the selection algorithm. It is imperative to go beyond simple and empirical evaluation. Therefore in this study, as an example, a special feature weighting algorithm is introduced, which is based on L2-regularized logistic regression loss for describing the local data structure. And its ensemble version using linear aggregation strategy is also proposed. Moreover, the proof for the stability of ensemble feature weighting about small changes (changing or removal of one sample) of data set is also presented based on uniform stability.

The introduced feature weighting algorithm is under embedded model and outputs a feature weights (measuring features' relevance) vector.

The paper is organized as follows, a feature weighting algorithm based on L2-regularized logistic loss and its ensemble version is introduced in section 2. Section 3 provides the proof for stability of ensemble feature weighting. The experimental results are shown in section 4, the paper ends with conclusion in section 5.

2 Ensemble Feature Weighting

To train a ensemble model for feature weighting, we are given a training sample set *D*, which contains *n* samples, $D = \{X, Y\} = \{x_i, y_i\}_{i=1}^{n}$, where x_i is the input for the *i*-th training sample, and y_i is the label, and each sample is represented by a d-dimensional vector $x_i = (x_{i1}, x_{i2}, \cdots, x_{id}) \in R^d$.

2.1 Feature Weighting Algorithm

In general, in order to achieve good generalization, the nearest neighbors with the same label to a sample (i.e., target samples) always should be closer to the sample, while other samples from different classes are separated by a large margin. Based on local learning, for sample x_i, it should be close to the nearest target sample (i.e., nearest hit sample $NH(x_i)$) and away from the nearest neighbor sample with different class label (i.e., near miss sample $NM(x_i)$). For the purposes of this paper, we use the Manhattan distance to define the nearest neighbors and their closeness, while other standard definitions may also be used. The logistic regression loss is adopted to model the fit of data for its simplicity and effectiveness. To prevent from overfitting and improve the robustness of feature weighting, the L2-regularization is used for its rotational invariance [1] and strong stability property [18]. Thus, the evaluation criterion for feature weighting is defined as follows,

$$L_D(w) = \frac{1}{n}\sum_{i=1}^{n}\log(1+\exp(-w^T z_i)) + \gamma.||w||^2 \tag{1}$$

where γ is the cost parameter balancing the importance of the two terms, T is the transpose, w is the feature weight vector, $z_i = |x_i - NM(x_i)| - |x_i - NH(x_i)|$ and $|.|$ is an element-wise absolute operator. The z_i can be considered as the mapping point of x_i [9].

In the Eqn. (1), $w^T z_i$ is the local margin for x_i, which belongs to hypothesis margin [19] and an intuitive interpretation of this margin is a measure of the proportion of the features in x_i that can be corrupted by noise (or how much x_i can move in the feature space) before being misclassified [20]. By the large margin theory [21], 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 w, so that a margin-based error function in the induced feature space is minimized. In the end, feature selection aims to find the target model w, which minimizes the loss function in Eqn.(1) through gradient descent-based techniques.

2.2 Weight-Based Ensemble Feature Weighting

Similar to the ensemble models for supervised learning, there are two essential steps in ensemble feature selection. The first step involves creating a set of different base feature selectors, each provides its output, while the second step aggregates the results of all feature selectors [8]. We adopt a subsampling based strategy and linear aggregation. Then *m* subsamples of size $\alpha n(0 < \alpha < 1)$ are drawn randomly from *D*, where the parameters *m* and α can be varied. Subsequently, feature weighting is performed on each of the *m* subsamples. Therefore, we obtain feature weighting results ensemble $En = \{w_1, w_2, \cdots, w_m\}$, where $w_t(t = 1, 2, \cdots, m)$ represents the outcome of the *t*-th base feature selector trained on *t*-th subsample. Specifically, in our case, each feature selection result $w_t(t = 1, 2, \cdots, m)$ is a feature weighting vector. And we obtain the final ensemble feature weighting result $w_e = \frac{1}{m}\sum_{t=1}^{m} w_t$, where $w_t \in En$. This ensemble method belongs to weight-based ensemble model (WEn).

The proposed ensemble feature weighting is also corresponding to the recognition that when estimating an unknown function from data, one needs to find a tradeoff between bias and variance [13]. Indeed, besides the regularization, another idea is to use statistical procedures to reduce the variance without altering the bias and lead to high stability. One such technique is the bagging approach [22], which consists in averaging several estimators built from random subsamples of the data.

3 Stability Analysis

Now, we will firstly show the rotation invariance for our proposed feature weighting algorithm. Based on the Proposition 4.2 in [1], let *H* be a rotational matrix $\{H \in \mathcal{R}^{d\times d}, H^T H = HH^T = I, |H| = 1\}$, then Hx is *x* rotated through some angle around the origin. It is evident that loss function $L_D(w)$ is rotational invariance with respect to H. In other words, $L_D(w) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp(-(Hw)^T(Hz_i))) + \gamma.||Hw||^2$, and $Hz_i = H(|x_i - NM(x_i)| - |x_i - NH(x_i)|) = |Hx_i - H.NM(x_i)| - |Hx_i - H.NH(x_i)|$, which means the proposed feature weighting algorithm is rotational invariance for sample x_i. The linear aggregation strategy is employed, then intuitively the ensemble feature weighting is also rotational invariance.

3.1 Uniform Stability Definition

On the other hand, a stable algorithm is one whose output does not change significantly with small changes in the input. The stability of classification, regression and sample ranking has been deeply analyzed [13, 14, 23], however, the stability of feature selection has not been explicitly introduced in theory. Similarly, we also consider changes to such a sample that consist of replacing a single example in the sequence with a new example or the exclusion of the sample. For a given training set *D* of size *n*, we will denote $D^{\setminus i}$ as the training set obtained by removing point (x_i, y_i) for all $i \in \{1, \cdots, n\}$. And we denote by D^i the training set obtained by changing one point (x_i, y_i) into (x'_i, y'_i), which is assumed to be independent from *D*.

Definition 1. (Uniform weighting stability) For a feature selection algorithm *A* whose outputs on data set *D* and $D^{\setminus i}$ are weight vectors denoted by w_D and $w_{D^{\setminus i}}$,

respectively. Algorithm *A* has uniform weighting stability β ($\beta \geq 0$) if for all *D* and any $i \in \{1, \cdots, n\}$, we have

$$||w_D - w_{D^{\setminus i}}|| \leq \beta.$$

A smaller value of β corresponds to greater weighting stability. More formally, point (x_i, y_i) is replaced by the empty set which we assume the learning method treats as having this point simply removed, and the D^i can be regarded as one data set is firstly replaced by the empty set and then the empty set is replaced by point (x_i', y_i'). So an feature weighting algorithm with uniform stability β has also the following property: For all *D* and $i \in \{1, \cdots, n\}$, $||w_D - w_{D^i}|| \leq ||w_D - w_{D^{\setminus i}}|| + ||w_{D^i} - w_{D^{\setminus i}}|| \leq 2\beta$. In other words, stability with respect to the exclusion of one point implies stability with respect to changes of one point. Then we only focus on stability analysis for the exclusion case in the follows.

3.2 Stability for Ensemble Feature Weighting

For ensemble, bootstrap strategy is used to train the same feature weighting algorithm on a number *m* of different bootstrap sets of a training set *D* and by averaging the obtained solutions. We denote these bootstrap sets by $D(r_t)$ for $t = 1, \cdots, m$, where the $r_t \in R = \{1, \cdots, n\}^p (p < n)$ are instances of a random variable corresponding to sampling without replacement of *p* elements from the training set *D*. And R is a space containing elements r that model the randomization of the subsampling. We will use the shorthand $w_{D(r_t)}$ to denote the outcome of the feature weighting algorithm applied on the *t*-th bootstrap training set $D(r_t)$. And the ensemble result is $\frac{1}{m}\sum_{t=1}^{m} w_{D(r_t)}$. The uniform weighting stability of ensemble feature weighting is defined as follows: For all *D* and $i \in \{1, \cdots, n\}$.

$$\beta_e = \mathbb{E}_{r_1, \cdots, r_m}[||\frac{1}{m}\sum_{t=1}^{m} w_{D(r_t)} - \frac{1}{m}\sum_{t=1}^{m} w_{D^{\setminus i}(r_t)}||]$$

where $\mathbb{E}$ is the expectation and $r_1, \cdots, r_m$ are i.i.d. random variables modeling the random sampling and having the same distribution as r. Then

$$\begin{aligned}\beta_e &\leq \frac{1}{m}\sum_{t=1}^{m} \mathbb{E}_{r_t}[||w_{D(r_t)} - w_{D^{\setminus i}(r_t)}||] \\ &= \mathbb{E}_r[||w_{D(r)} - w_{D^{\setminus i}(r)}||] = \mathbb{E}_r[|| \triangle w_{D(r)}||]\end{aligned}$$

The stability for the removal of $x_i (i \in \{1, 2, \cdots, n\})$ case is considered, then the random sampling containing sample x_i should be determined.

$$\begin{aligned}\beta_e &\leq \mathbb{E}_r[|| \triangle w_{D(r)}||(\mathbb{I}(i \in r) + \mathbb{I}(i \notin r))] \\ &= \mathbb{E}_r[|| \triangle w_{D(r)}||\mathbb{I}(i \in r)] + \mathbb{E}_r[|| \triangle w_{D(r)}||\mathbb{I}(i \notin r)] \\ &= \mathbb{E}_r[|| \triangle w_{D(r)}||\mathbb{I}(i \in r)]\end{aligned}$$

where $\mathbb{I}(.)$ is indicator function. Note that the second part of the last second equation is equal to zero because when *i* is not in *r*, which means point x_i does not belong to $D(r)$

and, thus, $D(r) = D^{\setminus i}(r)$. The size of subsample $D(r)$ is p, then $\mathbb{E}_r(\mathbb{I}(i \in r)) = \frac{p}{n}$ because this subsampling is done without replacement,

$$\beta_e \leq \frac{p}{n} || \triangle w_{D(r)} ||$$

where $|| \triangle w_{D(r)})||$ is the uniform stability of base feature weighting on bootstrap set $D(r)$ that contains sample x_i, and $\triangle w_{D(r)} = w_{D(r)} - w_{D^{\setminus i}(r)}$ where $w_{D(r)}$ and $w_{D^{\setminus i}(r)}$ is the minimizer for the convex objective function $L_{D(r)}(w)$ and $L_{D^{\setminus i}(r)}(w)$ respectively. According to Eqn.(1), these objective functions are defined as follows,

$$L_{D(r)}(w) = \frac{1}{p} \sum_{j=1}^{p} \log(1 + \exp(-w^T z_j)) + \gamma . ||w||^2$$

$$L_{D^{\setminus i}(r)}(w) = \frac{1}{p} \sum_{j=1, j \neq i}^{p} \log(1 + \exp(-w^T z_j)) + \gamma . ||w||^2$$

Due to the convexity of the objective functions, for any $a \in [0, 1]$, we get

$$\begin{aligned} L_{D(r)}(w_{D(r)}) - L_{D(r)}(w_{D(r)} - a \triangle w_{D(r)}) \leq 0 \\ L_{D^{\setminus i}(r)}(w_{D^{\setminus i}(r)}) - L_{D^{\setminus i}(r)}(w_{D^{\setminus i}(r)} + a \triangle w_{D(r)}) \leq 0 \end{aligned}$$

So summing the two equations above, we get that

$$\begin{aligned} & \frac{1}{p} \sum_{j=1, j \neq i}^{p} \log(1 + \exp(-w_{D(r)}^T z_j)) + \frac{1}{p} \log(1 + \exp(-w_{D(r)}^T z_i)) \\ & \quad - \frac{1}{p} \sum_{j=1, j \neq i}^{p} \log(1 + \exp(-(w_{D(r)} - a \triangle w_{D(r)})^T z_j)) \\ & \quad - \frac{1}{p} \log(1 + \exp(-(w_{D(r)} - a \triangle w_{D(r)})^T z_i)) \\ & \quad + \frac{1}{p} \sum_{j=1, j \neq i}^{p} \log(1 + \exp(-w_{D^{\setminus i}(r)}^T z_j)) \\ & \quad - \frac{1}{p} \sum_{j=1, j \neq i}^{p} \log(1 + \exp(-(w_{D^{\setminus i}(r)} + a \triangle w_{D(r)})^T z_j)) \\ & \quad + \gamma . ||w_{D(r)}||^2 - \gamma . ||w_{D(r)} - a \triangle w_{D(r)}||^2 \\ & \quad + \gamma . ||w_{D^{\setminus i}(r)}||^2 - \gamma . ||w_{D^{\setminus i}(r)} + a \triangle w_{D(r)}||^2 \\ & \leq 0 \end{aligned} \tag{2}$$

Since logistic loss is the convex function, then by Jensen's inequality,

$$\begin{aligned} & \log(1 + \exp(-(w_{D(r)} - a \triangle w_{D(r)})^T z_j) \\ & \quad = \log(1 + \exp(-((1 - a) w_{D(r)}^T z_j + a w_{D^{\setminus i}(r)}^T z_j)) \\ & \quad \leq \log(1 + \exp(-w_{D(r)}^T z_j)) \\ & \qquad - a(\log(1 + \exp(-w_{D(r)}^T z_j)) - \log(1 + \exp(-w_{D^{\setminus i}(r)}^T z_j))) \end{aligned}$$

Similarly, we also can get

$$\begin{aligned}&\log(1+\exp(-(w_{D^{\setminus i}(r)}+a\triangle w_{D(r)})^T z_j))\\&\quad\le \log(1+\exp(-w^T_{D^{\setminus i}(r)}z_j)\\&\qquad + a(\log(1+\exp(-w^T_{D(r)}z_j))-\log(1+\exp(-w^T_{D^{\setminus i}(r)}z_j)))\end{aligned}$$

The two equations above are plugged into (2), then

$$\begin{aligned}&||w_{D(r)}||^2-||w_{D(r)}-a\triangle w_{D(r)}||^2-||w_{D^{\setminus i}(r)}+a\triangle w_{D(r)}||^2+||w_{D^{\setminus i}(r)}||^2\\&\le\frac{a}{p\gamma}(\log(1+\exp(-w^T_{D^{\setminus i}(r)}z_i))-\log(1+\exp(-w^T_{D(r)}z_i)))\\&\le\frac{a}{p\gamma}|\triangle w^T_{D(r)}z_i|\end{aligned}$$

and the last line above is gotten because it is proved in [24] that the logistic loss function is a Lipschitz function with Lipschitz constant 1. If we set $a = 1/2$, the left side of previous equation approximately amounts to

$$\begin{aligned}&||w_{D(r)}||^2+||w_{D^{\setminus i}(r)}||^2-\frac{1}{2}||w_{D(r)}+w_{D^{\setminus i}(r)}||^2\\&=\frac{1}{2}||w_{D(r)}||^2+\frac{1}{2}||w_{D^{\setminus i}(r)}||^2-w^T_{D(r)}w_{D^{\setminus i}(r)}=\frac{1}{2}||\triangle w_{D(r)}||^2\end{aligned}$$

Thus,

$$||\triangle w_{D(r)}||^2\le\frac{1}{p\gamma}|\triangle w^T_{D(r)}z_i|$$

and based on Cauchy-Schwarz inequality

$$|\triangle w^T_{D(r)}z_i|\le||\triangle w_{D(r)}||||z_i||$$

Then combine the two equations above and the samples are normalized, it can be shown $||z_i||\le 2$, we obtain the stability for our base feature weighting.

$$||w_{D(r)}-w_{D^{\setminus i}(r)}||=||\triangle w_{D(r)}||\le\frac{2}{p\gamma}$$

and the uniform stability for ensemble feature weighting is

$$\beta_e\le\frac{2}{n\gamma}$$

3.3 Remarks and Discussions

The analysis results show that a larger regularization parameter γ leads to better stability. And the ensemble feature weighting owns better stability bounds than base feature weighting. The ensemble algorithm has a uniform stability bound goes to zero as $\frac{1}{n\gamma}$, and it is stable because of the wide acceptance that the algorithms have a uniform stability bound that decreases as $O(\frac{1}{n})$, and are hence stable [13, 14, 23, 18]. To our best

knowledge, this work provides the first uniform stability-style analysis on the stability of feature selection. Although in [2], the analysis for robustness of spectral feature selection against noise is presented, and in [9], the experimental results also show the variance reduction leads to stable feature selection. It is obvious that our work is significantly different because formal stability notion is considered explicitly, and we mainly focus on sampling randomness instead of noise, and we thus are interested in how changes to the training data influence the result of feature weighting algorithm.

Moreover, the L2-norm is used as stability metric in the paper, this is only for ease of presentation. Other norm can be adopted, such as L∞-norm, which is employed to measure the uniform stability of classification and regression algorithm [13, 14]. Certainly, the proof for the stability also makes sense because of $L\infty - norm \leq L2 - norm$ in most cases. And it should be noted that the stability bound is loose, and we only like to prove that the proposed ensemble feature weighting algorithm is stable because its stability scales like $\frac{1}{n}$.

Finally, it is evident the theoretical analysis results still hold true for other ensemble feature weighting algorithms where base feature weighting algorithm is based on L2-regularized convex Lipschitz loss functions and linear aggregation strategy is employed.

4 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. The chosen data sets are Sonar, Arcene, Musk, Ionosphere, which are taken from UCI ML repository [25], and Colon cancer diagnosis data set is introduced in [26] and Lung cancer is introduced in [27]. Colon, Arcene and Lung owns small samples (62,200,203) with extremely high dimensionality (2000,10000,12600). The small sample problem is one of the most challenging problem for feature selection, particularly on its output stability.

Note that feature weighting is almost never directly used to measure the stability of feature selection, and instead converted to a ranking based on the weights [8]. Because the feature weights are always changed to feature ranks, then another ensemble strategy should be considered: instead of the feature weights linear combination, the feature weight vectors outputted from the m subsamples are firstly changed to feature rank vectors (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), then linear combination of these feature rank vectors is adopted to obtain the ensemble ranking results as in [8]. And we call this ensemble strategy as rank-based ensemble(REn). Other chosen stable feature weighting algorithms for comparison are ensemble Relief (En-Relief) [8] 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 [9]. We combine the sample weighting strategy with the newly proposed feature weighting algorithm-Lmba [28] and named as VR-Lmba.

4.1 Experimental Results for Stability

To measure the stability of feature weighting algorithms, we also adopt a subsampling based strategy-bootstrap without replacement. For a data set, 10 subsamples containing 90% of the data are randomly drawn without replacement. This percentage was chosen as in [8] to assess robustness with respect to relatively small changes in the data set. Of course, the sampling rate and the number of subsamples can be varied. Subsequently, the proposed ensemble algorithms (WEn and REn) with $\alpha = 0.9$ and $\gamma = 1$, En-Relief and VR-Lmba is performed on each subsample, which is considered as the data set D described in section 2, and output a feature rank vector (if the output is a feature weight vector, it should be changed to feature rank vector). Then the similarity between feature ranking result pairs is calculated using Spearman rank correlation coefficient [8], and the stability is the average similarity over all pairwise similarity between the different feature ranking results [8]. The stability of these feature weighting algorithms for

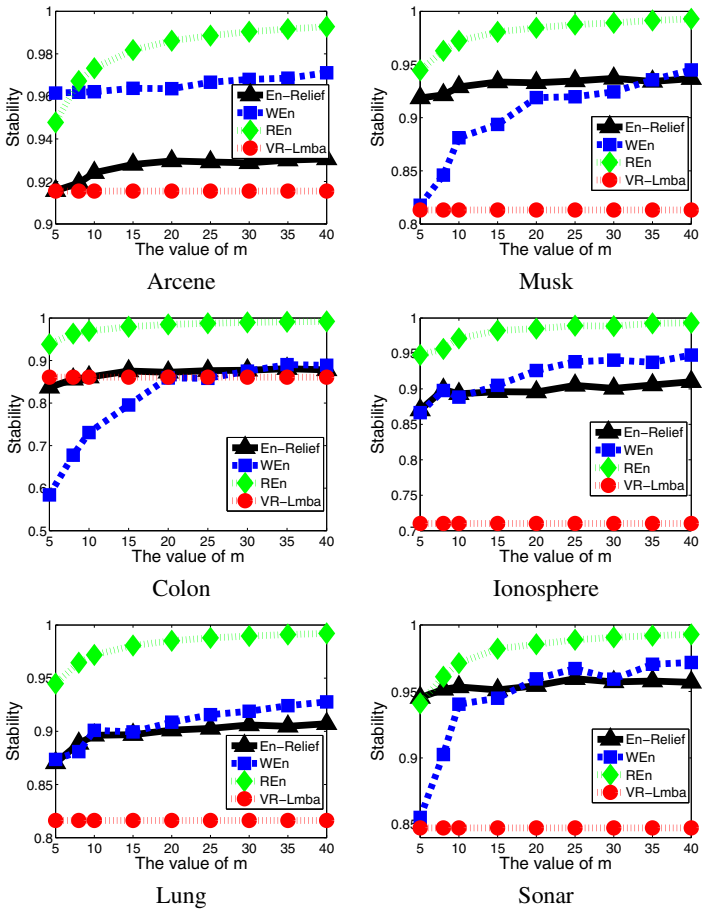

Fig. 1. Experimental results of stability

different data sets is shown in Fig.1. The X-axis is the number of base feature selectors m. Note that VR-Lmba is not a ensemble method, then its stability does not change along with m.

4.2 Balance between Stability and Classification

Besides the stability, classification performance is another important issues for feature selection. In order to validate the tradeoff between the stability and classification accuracy, a F-Measure is employed, which is defined as $\frac{2\times stability\times accuracy}{stability+accuracy}$ [8]. In this part of experiments, the number of base selectors for ensemble feature weighting is constant and set as 20 for all ensemble algorithms, i.e., $m = 20$. 10-cross validation is used and the linear SVM with C=1 and 3-nearest neighbors(3NN) classifier is adopted. The experimental results are shown in Fig. 2 and 3 corresponding to 3NN and SVM. For space constraints, only the experimental results of two data sets for each classifier are shown in the figures.

4.3 Observations and Discussions

From the results, we can observe that the stability value of rank-based ensemble-REn is the highest among all stable feature weighting algorithms, and weight-based ensemble-WEn always gets higher or similar stability to En-Relief and VR-Lmba. In addition,

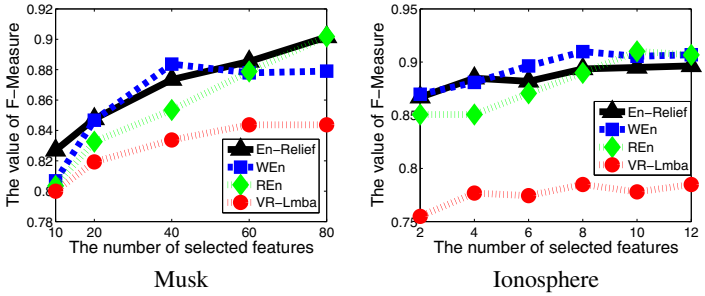

Fig. 2. Experimental results of F-Measure for 3NN classifier

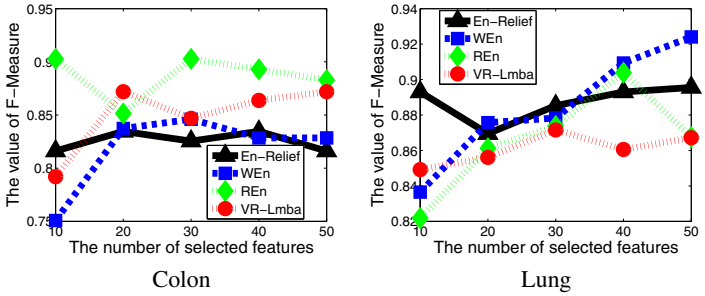

Fig. 3. Experimental results of F-Measure for SVM

at least one of our proposed ensemble methods (REn or WEn) always obtain higher or similar balance between the stability and classification accuracy to other stable ones.

For the higher stability of rank-based ensemble-REn than weight-based ensemble-WEn, this can be explained intuitively by the fact that the stability is measured based on the feature ranks. Consider the scenarios if the feature weights produced by base feature weighting algorithm change due to the data variation, however, their ranks may not change, which leads to higher stability for rank-based ensemble than weight-based ensemble. Of course, if the feature weights do not change, then the feature ranks surely stable. Thus it means that the stable weight-based ensemble leads to stable rank-based ensemble, then the theoretic analysis of stability for weight-based ensemble hold true for the rank-based ensemble. And the above analysis also give some reasons for the high efficiency of En-Relief, which also belongs to rank-based ensemble model.

5 Conclusion

The stability of feature selection is attracted much attention. Our major contribution is presenting the theoretical analysis for the uniform stability of ensemble feature weighting algorithm. In the paper, as an example, a logistic loss-based feature weighting algorithm via L2-regularization is introduced. And its weight-based ensemble version-WEn is presented and is formally analyzed on the stability. The experimental results on some real-world data sets including microarray data (small sample size problem) have also shown the proposed ensemble feature weighting algorithms (weight-based ensemble-WEn or rank-based ensemble-REn) get higher stability and better or comparable tradeoff between classification and stability to other stable algorithms in most cases. 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 61105082) and Jiangsu Government Scholarship.

References

1. Ng, A.Y.: Feature selection, l1 vs. l2 regularization, and rotational invariance. In: Proceedings of International Conference on Machine Learning, Banff, Canada (2004)
2. Zhao, Z.: Spectral Feature Selection for Mining Ultrahigh Dimensional Data. PhD thesis, Arizona State University (2010)
3. Inza, I., Larranaga, P., Blanco, R., Cerrolaza, A.J.: Filter versus wrapper gene selection approaches in dna microarray domains. Artificial Intelligence in Medicine 31, 91–103 (2004)
4. Forman, G.: An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research 3, 1289–1305 (2003)
5. Liu, H., Yu, L.: Toward integrating feature selection algorithms for classification and clustering. IEEE Trans. Knowledge and Data Engineering 17, 494–502 (2005)
6. Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.: Feature Extraction, Foundations and Applications. Springer, Physica-Verlag, New York (2006)
7. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research 31, 1157–1182 (2003)

8. Saeys, Y., Abeel, T., Van de Peer, Y.: Robust feature selection using ensemble feature selection techniques. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part II. LNCS (LNAI), vol. 5212, pp. 313–325. Springer, Heidelberg (2008)
9. Han, Y., Yu, L.: A variance reduction for stable feature selection. In: Proceedings of the International Conference on Data Mining, pp. 206–215 (2010)
10. Loscalzo, S., Yu, L., Ding, C.: Consensus group stable feature selection. In: Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 567–575 (2009)
11. Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., Saeys, Y.: Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics 26, 392–398 (2010)
12. Yu, L., Han, Y., Berens, M.E.: Stable gene selection from microarray data via sample weighting. IEEE/ACM Trans. Computational Biology and Bioinformatics 9, 262–272 (2012)
13. Bousquet, O., Elisseeff, A.: Stability and generalization. Journal of Machine Learning Research 2, 499–526 (2002)
14. Elisseeff, A., Evgeniou, T., Pontil, M.: Stability of randomized learning algorithm. Journal of Machine Learning Research 6, 55–79 (2005)
15. Li, Y., Gao, S.Y., Chen, S.C.: Ensemble feature weighting based on local learning and diversity. In: AAAI Conference on Artificial Intelligence, pp. 1019–1025 (2012)
16. Woznica, A., Nguyen, P., Kalousis, A.: Model mining for robust feature selection. In: Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 913–921 (2012)
17. Saeys, Y., Inza, I., Larranaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics 23, 2507–2517 (2007)
18. Xu, H., Caramanis, C., Mannor, S.: Sparse algorithm are not stable: A no-free-lunch theorem. IEEE Trans. Pattern Analysis and Machine Intelligence 34, 187–193 (2012)
19. Crammer, K., Bachrach, R.G., Navot, A., Tishby, N.: Margin analysis of the lvq algorithm. In: Advances in Neural Information Processing Systems, pp. 462–469 (2002)
20. Sun, Y.J., Todorovic, S., Goodison, S.: Local learning based feature selection for high dimensional data analysis. IEEE Trans. Pattern Analysis and Machine Intelligence 32, 1–18 (2010)
21. Schapire, R.E., Freud, Y., Bartlett, P., Lee, W.S.: Boosting the margin: A new explanation for the effectiveness of voting methods. The Annals of Statistics 26, 1651–1686 (1998)
22. Breiman, L.: Bagging predictors. Machine Learning 26, 123–140 (1996)
23. Agarwal, S., Niyogi, P.: Generalization bounds for ranking algorithm via algorithmic stability. Journal of Machine Learning Research 10, 441–474 (2009)
24. Anthony, M., Bartlett, P.L.: Neural Network Learning: Theoretical Foundations. Cambridge University Press, Cambridge (1999)
25. Frank, A., Asuncion, A.: UCI machine learning repository (2010), `http://archive.ics.uci.edu/ml`
26. Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon cancer tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences of the United States of America, 6745–6750 (1999)
27. Bhattacharjee, A., Richards, W.G., Staunton, J., Li, C., Monti, S.: Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. Proceedings of the National Academy of Sciences of the United States of America 98, 13790–13795 (2001)
28. Li, Y., Lu, B.L.: Feature selection based on loss margin of nearest neighbor classification. Pattern Recognition 42, 1914–1921 (2009)