Diversity Regularized Machine*

Yang Yu and **Yu-Feng Li** and **Zhi-Hua Zhou**
National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210093, China
{yuy,liyf,zhouzh}@lamda.nju.edu.cn

Abstract

Ensemble methods, which train multiple learners for a task, are among the state-of-the-art learning approaches. The *diversity* of the component learners has been recognized as a key to a good ensemble, and existing ensemble methods try different ways to encourage diversity, mostly by heuristics. In this paper, we propose the *diversity regularized machine* (DRM) in a mathematical programming framework, which efficiently generates an ensemble of diverse support vector machines (SVMs). Theoretical analysis discloses that the diversity constraint used in DRM can lead to an effective reduction on its hypothesis space complexity, implying that the diversity control in ensemble methods indeed plays a role of regularization as in popular statistical learning approaches. Experiments show that DRM can significantly improve generalization ability and is superior to some state-of-the-art SVM ensemble methods.

1 Introduction

Ensemble methods, such as AdaBoost [Freund and Schapire, 1997], Bagging [Breiman, 1996] and Random Forests [Breiman, 2001], are among the state-of-the-art learning approaches. Ensemble methods train a number of component learners for a learning task, and combine them to achieve a strong generalization performance. It has been widely accepted that to achieve a good ensemble, the component learners should be accurate and *diverse*. This recognition was first theoretically justified by the *error-ambiguity decomposition* [Krogh and Vedelsby, 1995] for regression tasks as $E = \bar{E} - \bar{A}$, where E is the mean-square error of the ensemble, $\bar{E}$ is the average mean-square error of component learners, and $\bar{A}$ is the average difference between the ensemble and the component learners. This decomposition implies that, as long as $\bar{E}$ is fixed, a higher difference among the component learners leads to a better ensemble. Later results achieved by the *bias-variance-covariance* decomposition [Ueda and Nakano, 1996], the *strength-correlation* decomposition [Breiman, 2001], and the *information-theoretical* decompositions [Brown, 2009; Zhou and Li, 2010] all confirmed that the diversity among the component learners is a key to the ensemble performance.

Though it remains an open problem on how to measure and evaluate diversity [Brown, 2009; Zhou and Li, 2010], many effective ensemble methods have already been developed. These methods employ different mechanisms to create diverse component learners, mostly using randomization strategies by smart heuristics [Ho, 1998; Breiman, 2000; 2001; Dietterich, 2002; Zhou and Yu, 2005; Brown *et al.*, 2005].

In this paper, we propose to managing the diversity among component learners in a deterministic mathematical programming framework, resulting in the *diversity regularized machine* (DRM) which generates an ensemble of SVMs [Vapnik, 1995] with an imposed diversity constraint. Theoretical analysis in the PAC learning framework [Valiant, 1984] discloses that the diversity constraint used in DRM can effectively reduce the hypothesis space complexity. This implies that the diversity control in ensemble methods plays the role of regularization as in popular statistical learning approaches. Experiments show that DRM can improve both the training and generalization accuracy of SVM, and is superior to some state-of-the-art SVM ensemble methods.

The rest of this paper is organized as follows. Section 2 presents the DRM, which is then theoretically analyzed in Section 3 and experimented in Section 4. Section 5 concludes.

2 DRM

Consider an input space, X and an underlying distribution $\mathcal{D}$ over X. A hypothesis, or a learner, is a function $h : X \to \{-1, 1\}$, and a concept c is an underlying hypothesis. A training set is a set of m examples $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$, where $\boldsymbol{x}_i \in X$ are drawn i.i.d. under $\mathcal{D}$ and $y_i = c(\boldsymbol{x}_i)$. A learning algorithm is to select a hypothesis h from a feasible hypothesis space $\mathcal{H}$ according to the given training set. For an integer number n, we denote $[n]$ as the set $\{1, 2, \ldots, n\}$.

2.1 Diversity Measure

We consider linear classification model $\boldsymbol{w} \in \mathbb{R}^d$, which classifies an instance by the inner product $\boldsymbol{w}^\top \phi(\boldsymbol{x})$, where $\phi(\boldsymbol{x})$

*The research was supported by NSFC (60975043), JiangsuSF (BK2008018) and 973 Program (2010CB327903).

is a feature mapping of the instance $\boldsymbol{x}$. Note that the model can be viewed as equivalent to the commonly used model $\boldsymbol{w}^\top\phi(\boldsymbol{x})+b$, since by extending the mapping $\phi(\boldsymbol{x})$ to one extra dimension with a constant value, the extended $\boldsymbol{w}$ absorbs the functionality of b.

We note that, though there is no agreement on what form diversity should be defined in, the studied measures of diversity usually can be in a pairwise form, i.e., the total diversity is the sum of a pairwise difference measure, measuring classification effectiveness. Such diversity measures include *Q-statistics measure* [Kuncheva *et al.*, 2003], *correlation coefficient measure* [Kuncheva *et al.*, 2003], *disagreement measure* [Ho, 1998], *double-fault measure* [Giacinto and Roli, 2001], *κ statistic measure* [Dietterich, 2000], etc. Thus we also consider a form of diversity based on pairwise difference. Given a pairwise diversity measure `div` in a metric space of hypotheses, we consider the total diversity in norm p of the set of hypotheses $H=\{h_1,\ldots,h_T\}$ as

$$\mathtt{div}_p(H)=\Big(\sum\nolimits_{1\le i\ne j\le T}\mathtt{div}(h_i,h_j)^p\Big)^{1/p}.$$

Notice that each hypothesis h_i is a linear learner without the bias term, thus the direction of the linear learner effects the classification most. Thus, for a pair of linear learners $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$, we measure their diversity using the angle between them,

$$\mathtt{div}(\boldsymbol{w}_1,\boldsymbol{w}_2)=1-\frac{\boldsymbol{w}_1^\top\boldsymbol{w}_2}{\|\boldsymbol{w}_1\|\cdot\|\boldsymbol{w}_2\|},$$

so that the larger value of `div`, the larger angle between the linear learners.

2.2 Learning with Diversity Constraint

The training of multiple diverse linear learners can be formulated as an optimization framework that minimizes a loss function with a constraint of diversity. The framework for training T linear learners, $\boldsymbol{w}_1,\ldots,\boldsymbol{w}_T$, can be described as:

$$\begin{aligned}\mathop{\arg\min}_{\boldsymbol{w}_1,\ldots,\boldsymbol{w}_T}\ &\sum\nolimits_{t=1}^{T}\sum\nolimits_{i=1}^{m}\ell(y_i,\boldsymbol{w}_t^\top\phi(\boldsymbol{x}_i))) \qquad (1)\\ \text{s.t.}\ \ &\|\boldsymbol{w}_t\|_2\le\theta\ (\forall t\in[T]),\\ &\mathtt{div}_p(\{\boldsymbol{w}_1,\ldots,\boldsymbol{w}_T\})\ge q.\end{aligned}$$

where θ is the capacity control for each linear model, $\mathtt{div}_p$ is the p-norm diversity measure, q is the minimum amount of the diversity required, ϕ is a feature mapping induced by kernel k, ℓ is a loss function (e.g., hinge loss for classification problem or square loss for regression problem). After the training, the combined model is $\boldsymbol{w}_c=\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{w}_t$, which is still a linear learner.

Specifically, for classification problem, we implement the framework with the 1-norm diversity measure $\mathtt{div}_1$ and follow the ν-SVM framework [Schölkopf *et al.*, 2000] with square hinge loss. Thus, the framework is implemented as:

$$\begin{aligned}\mathop{\arg\min}_{\{\boldsymbol{w}_t,\rho_t,\boldsymbol{\xi}_t\}_{t=1}^T}\ &\sum_{t=1}^{T}(-\nu\frac{\rho_t}{T}+\frac{1}{m}\sum_{i=1}^{m}\xi_{t,i}^2)+\mu\sum_{1\le t<t'\le T}\frac{\boldsymbol{w}_t^\top\boldsymbol{w}_{t'}}{\|\boldsymbol{w}_t\|_2\|\boldsymbol{w}_{t'}\|_2}\\ \text{s.t.}\ \ &y_i\boldsymbol{w}_t^\top\phi(\boldsymbol{x}_i)\ge\rho_t-\xi_{t,i}(\forall i\in[m],\forall t\in[T]),\\ &\|\boldsymbol{w}_t\|_2\le 1\ (\forall t\in[T]). \qquad (2)\end{aligned}$$

where μ corresponds to q in Eq.(1), $\boldsymbol{\xi}_t=[\xi_{t,1},\ldots,\xi_{t,n}]^\top$ is a vector of slack variables, ν is the parameter to trade-off the ρ-margin and the slack variables.

2.3 Optimization

Under some conditions, which will be discussed later, the solution of Eq.(2) satisfies $\|\boldsymbol{w}_t\|_2=1$ for all t. Using this equation, the diversity term $\frac{\boldsymbol{w}_t^\top\boldsymbol{w}_{t'}}{\|\boldsymbol{w}_t\|\|\boldsymbol{w}_{t'}\|}$ is then simplified as $\boldsymbol{w}_t^\top\boldsymbol{w}_{t'}$. Further note that adding a constant $\|\boldsymbol{w}_t\|_2^2+\|\boldsymbol{w}_{t'}\|_2^2$ (that is 2) will not change the optimal solution of Eq.(2), thus the diversity term can be replaced by $\|\boldsymbol{w}_t+\boldsymbol{w}_{t'}\|_2^2$. We then have the following relaxed convex optimization problem:

$$\begin{aligned}\mathop{\arg\min}_{\{\boldsymbol{w}_t,\rho_t,\boldsymbol{\xi}_t\}_{t=1}^T}\ &\sum_{t=1}^{T}(-\nu\frac{\rho_t}{T}+\frac{1}{m}\sum_{i=1}^{m}\xi_{t,i}^2)+\mu\sum_{1\le t\ne t'\le T}\|\boldsymbol{w}_t+\boldsymbol{w}_{t'}\|_2^2\\ \text{s.t.}\ \ &y_i\boldsymbol{w}_t^\top\phi(\boldsymbol{x}_i)\ge\rho_t-\xi_{t,i}(\forall i\in[m],\forall t\in[T]),\\ &\|\boldsymbol{w}_t\|_2^2\le 1\ (\forall t\in[T]). \qquad (3)\end{aligned}$$

Since the constraints in Eq.(3) is naturally separable for each learner, instead of directly solving the large-scale *quadratically constrained quadratic program* (QCQP) problem of Eq.(3), we employ an efficient *alternating optimization* technique [Luo and Tseng, 1992]. The alternating optimization sequentially solves small QCQP problems with variables $\{\boldsymbol{w}_t,\rho_t,\boldsymbol{\xi}_t\}$ while fixing the other variables $\{\boldsymbol{w}_{t'},\rho_{t'},\boldsymbol{\xi}_{t'}\}$ for all $t'\ne t$ as constants. Mathematically, in each step we are solving the following small QCQP problem for each t:

$$\begin{aligned}\mathop{\arg\min}_{\boldsymbol{w}_t,\rho_t,\boldsymbol{\xi}_t}\ &-\nu\frac{\rho_t}{T}+\frac{1}{m}\sum\nolimits_{i=1}^{m}\xi_{t,i}^2+\mu\sum\nolimits_{t'\ne t}\|\boldsymbol{w}_t+\boldsymbol{w}_{t'}\|_2^2\\ \text{s.t.}\ \ &y_i\boldsymbol{w}_t^\top\phi(\boldsymbol{x}_i)\ge\rho_t-\xi_{t,i},\forall i\in[m], \qquad (4)\\ &\|\boldsymbol{w}_t\|_2^2\le 1.\end{aligned}$$

Further, the above QCQP problem can be addressed via *sequential quadratic programming* efficiently. Introducing the Lagrange multipliers α_t and λ_t for the constraints in Eq.4, we have:

$$\begin{aligned}&\boldsymbol{L}(\boldsymbol{w}_t,\rho_t,\boldsymbol{\xi}_t;\boldsymbol{\alpha}_t,\lambda_t)=\\ &-\nu\frac{\rho_t}{T}+\frac{1}{m}\sum\nolimits_{i=1}^{m}\xi_{t,i}^2+\mu\sum\nolimits_{t'\ne t}\|\boldsymbol{w}_t+\boldsymbol{w}_{t'}\|_2^2\\ &+\lambda_t(\|\boldsymbol{w}_t\|_2^2-1)-\sum\nolimits_{i=1}^{m}\alpha_{t,i}(y_i\boldsymbol{w}_t^\top\phi(\boldsymbol{x}_i)-\rho_t+\xi_{t,i}),\end{aligned}$$

where $\boldsymbol{\alpha}_t=[\alpha_{t,1},\ldots,\alpha_{t,n}]$. Setting the partial derivations *w.r.t.* $\{\boldsymbol{w}_t,\rho_t,\boldsymbol{\xi}_t\}$ to zeros, we have:

$$\begin{aligned}\frac{\partial\boldsymbol{L}}{\partial\boldsymbol{w}_t}&=2\mu\sum_{t'\ne t}(\boldsymbol{w}_t+\boldsymbol{w}_{t'})+2\lambda_t\boldsymbol{w}_t-\sum_{i=1}^{m}\alpha_{t,i}y_i\phi(\boldsymbol{x}_i)=0,\\ \frac{\partial\boldsymbol{L}}{\partial\rho_t}&=-\nu\frac{1}{T}+\sum\nolimits_{i=1}^{m}\alpha_{t,i}=0,\\ \frac{\partial\boldsymbol{L}}{\partial\xi_{t,i}}&=\frac{2}{m}\xi_{t,i}-\alpha_{t,i}=0.\end{aligned}$$

Let $\hat{\boldsymbol{w}}_t=\sum_{t'\ne t}\boldsymbol{w}_{t'}$, we then obtain the solution of $\boldsymbol{w}_t$ as:

$$\boldsymbol{w}_t=\frac{-2\mu\hat{\boldsymbol{w}}_t+\sum_{i=1}^{m}\alpha_{t,i}y_i\phi(\boldsymbol{x}_i)}{2((T-1)\mu+\lambda_t)}, \qquad (5)$$

and thus the dual of Eq.(4) can be cast as:

$$\arg\min_{\boldsymbol{\alpha}_t,\lambda_t} \frac{\| -2\mu\hat{\boldsymbol{w}}_t + \sum_{i=1}^m \alpha_{t,i} y_i \phi(\boldsymbol{x}_i)\|_2^2}{4((T-1)\mu+\lambda_t)} + \frac{m}{4}\sum_{i=1}^m \alpha_{t,i}^2 + \lambda_t$$
$$\text{s.t.} \quad \sum\nolimits_{i=1}^m \alpha_{t,i} = \nu\frac{1}{T}, \boldsymbol{\alpha}_t \geq 0, \lambda_t \geq 0.$$

which is jointly-convex for $\{\boldsymbol{\alpha}_t, \lambda_t\}$ [Boyd and Vandenberghe, 2004]. We further employ the alternating optimization technique to achieve the global optimal solution of the dual of Eq.(4) [Luo and Tseng, 1992]. Specifically, when λ_t is fixed, $\boldsymbol{\alpha}_t$ can be solved via:

$$\arg\min_{\boldsymbol{\alpha}_t} \frac{1}{2}\boldsymbol{\alpha}_t^\top \left(\frac{(\boldsymbol{K} \odot \boldsymbol{y}\boldsymbol{y}^\top)}{2(\lambda_t + (T-1)\mu)} + \frac{m}{2}\mathbf{I} \right) \boldsymbol{\alpha}_t - \boldsymbol{r}^\top \boldsymbol{\alpha}_t$$
$$\text{s.t.} \quad \sum\nolimits_{i=1}^m \alpha_{t,i} = \nu\frac{1}{T}, \boldsymbol{\alpha}_t \geq 0, \lambda_t \geq 0. \tag{6}$$

where $\boldsymbol{K}$ is the kernel matrix of $\phi(\boldsymbol{x})$ and $\odot$ is the entry-wise product, and $\boldsymbol{r}$ is a vector with components:

$$r_i = \frac{\mu y_i \hat{\boldsymbol{w}}_t^\top \phi(\boldsymbol{x}_i)}{(\lambda_t + (T-1)\mu)}.$$

Noted that Eq.6 is a convex quadratic programming problem involving only one equality constraint, this is similar to the dual problem of SVM which can be efficiently solved by state-of-the-art SVM solver, such as Libsvm using SMO algorithm [Chang and Lin, 2001].

When $\boldsymbol{\alpha}_t$ is fixed, λ_t can be solved in a closed-form, i.e.,

$$\lambda_t = \tag{7}$$
$$\max\{0, \frac{\| -2\mu\boldsymbol{w}_0 + \sum_{i=1}^m \alpha_{t,i} y_i \phi(\boldsymbol{x}_i)\|_2}{2} - (T-1)\mu\}.$$

Algorithm 1 presents the pseudocode of the DRM. It is worth noticing that, when the optimal solution of all λ_t's are non-zeros, according to KKT condition, the optimal solution $\{\boldsymbol{w}_t^*, \rho_t^*, \boldsymbol{\xi}_t^*\}_{t=1}^T$ obtained by DRM according to Eq.(3) satisfies $\|\boldsymbol{w}_t^*\|_2^2 = 1$, thus is also the optimal solution of Eq.(2).

Algorithm 1 DRM

Input: Training set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$ and kernel matrix $\boldsymbol{K}$, parameters T and μ

Process:

$\lambda_t \leftarrow 1$ and $\boldsymbol{w}_t \leftarrow \mathbf{0}, \forall t \in [T]$
while not converged yet **do**
  for $t = 1, \ldots, T$ **do**
    while not converged yet **do**
      $\boldsymbol{\alpha}_t \leftarrow$ solutions returned by Eq. (6)
      $\lambda_t \leftarrow$ solutions returned by Eq. (7)
    end while
    set $\boldsymbol{w}_t \leftarrow$ according to Eq. (5)
  end for
end while

Output: $\boldsymbol{w}_c = \frac{1}{T}\sum_{t=1}^T \boldsymbol{w}_t$

It is well known that, the generalization error of a learning algorithm A can be bounded using its empirical error and the complexity of its feasible hypothesis space. For linear learners, its hypothesis space is uncountable, thus we measure that using covering number as the definition below.

Definition 1 (Covering Number) Given m samples $S = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$ and a function space $\mathcal{F}$, characterize every $f \in \mathcal{F}$ using a vector $\boldsymbol{v}_S(f) = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_m)]$ in a metric space B^m with metric ρ. The covering number $\mathcal{N}_p(\mathcal{F}, \gamma, S)$ is the minimum number l of vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_l \in B^m$ such that, for all $f \in \mathcal{F}$ there exists $j \in \{1, \ldots, l\}$,

$$\|\rho(\boldsymbol{v}_S(f), \boldsymbol{u}_j)\|_p = \left(\sum\nolimits_{i=1}^m \rho(f(\boldsymbol{x}_i), u_{j,i})^p\right)^{\frac{1}{p}} \leq m^{\frac{1}{p}}\gamma,$$

and $\mathcal{N}_p(\mathcal{F}, \gamma, m) = \sup_{S:|S|=m} \mathcal{N}_p(\mathcal{F}, \gamma, S)$.

Lemma 1 [Bartelett, 1998] Consider the learning algorithm A selecting a hypothesis from space $\mathcal{H}$ according to m random examples. For all $\gamma > 0$, with probability at least $1-\delta$, the generalization error is bounded as

$$\epsilon_g(A) \leq \epsilon_e^\gamma(A) + \sqrt{\frac{2}{m}\left(\ln \mathcal{N}_\infty(\mathcal{H}, \gamma/2, 2m) + \ln\frac{2}{\delta}\right)},$$

where $\mathcal{N}_\infty$ is the covering number with infinity norm.

Lemma 1 indicates that the generalization error is bounded by two factors, one is the performance on the training set, and the other is the hypothesis space complexity. A good learning algorithm should balance the two factors well.

3 Theoretical Analysis

3.1 Preliminaries

Probabilistic Approximately Correct (PAC) learning [Valiant, 1984] is a powerful tool for analyzing learning algorithms. There has been much development of the theory, however, we choose to use the simple results for the clarity of presenting our core idea, instead of proving the tightest results. Comprehensive introductions to learning theory can be found in textbooks such as [Anthony and Bartlett, 1999].

Noted that $y \in \{-1, +1\}$ and $h \in [-1, +1]$, the margin of h on an instance is $yh(\boldsymbol{x})$. The training error with margin γ of a hypothesis h is defined as

$$\epsilon_e^\gamma(h) = \frac{1}{m}\sum\nolimits_{i=1}^m \mathbb{I}[h(\boldsymbol{x}_i)y_i < \gamma],$$

where $\mathbb{I}$ is the indicator function that outputs 1 if its inner expression is true and 0 otherwise. Define the generalization error, or true error, as

$$\epsilon_g(h) = E_{\boldsymbol{x}\sim\mathcal{D}}[\mathbb{I}[h(\boldsymbol{x})c(\boldsymbol{x}) < 0]].$$

3.2 Analysis of DRM

First, we look into the loss term of DRM in Eq.(2), which evaluates the loss of each hypothesis (linear learner) as:

$$\ell_T(\boldsymbol{w}_t) = (-\nu\frac{\rho_t}{T} + \frac{1}{m}\sum\nolimits_{i=1}^m \xi_{t,i}^2),$$

where $\xi_{t,i} \geq \rho_t - y_i\boldsymbol{w}_t^\top\phi(\boldsymbol{x}_i) (\forall i \in [m])$ by the constraints. We concern about the loss of the combined hypothesis $\boldsymbol{w}_c = \frac{1}{T}\sum_{t=1}^T \boldsymbol{w}_t$ with $\rho_c = \frac{1}{T}\sum_{t=1}^T \rho_t$:

$$\ell_c(\boldsymbol{w}_c) = (-\nu\rho_c + \frac{1}{m}\sum\nolimits_{i=1}^m \xi_i^2),$$

where $\xi_i \geq \rho_c - y_i \boldsymbol{w}_c^\top \phi(\boldsymbol{x}_i)(\forall i \in [m])$. We then have the following proposition.

Proposition 1 Let $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_T$ be the component hypotheses solved by DRM, and $\boldsymbol{w}_c = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{w}_t$ be the combined hypothesis, the loss of $\boldsymbol{w}_c$ is bounded as

$$\ell_c(\boldsymbol{w}_c) \leq \sum\nolimits_{t=1}^{T} \ell_T(\boldsymbol{w}_t).$$

Proof. Notice that, for a training instance $\boldsymbol{x}$, the classification margin $y\boldsymbol{w}_c^\top \phi(\boldsymbol{x})$ can be expanded as $y\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{w}_t^\top \phi(\boldsymbol{x})$, which is no larger than $\frac{1}{T}\sum_{t=1}^{T} \rho_t - \frac{1}{T}\sum_{t=1}^{T} \xi_t$ by the constraints in Eq.(2). Since $\rho_c = \frac{1}{T}\sum_{t=1}^{T} \rho_t$, the proposition is then proved by that $\sum_{t=1}^{T} \xi_t^2 \geq (\frac{1}{T}\sum_{t=1}^{T} \xi_t)^2$. □

The proposition shows that, as we optimize the loss of component hypotheses, we also optimize an upper bound of the loss of the combined hypothesis. Keeping the proposition in mind, we then focus on the hypothesis space complexity.

We conceptively distinguish the hypothesis space of the component hypothesis and that of the combined hypothesis. Let $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_T$ be hypotheses from space $\mathcal{H}$, and the combined $\boldsymbol{w}_c$ be in space $\mathcal{H}_c$. We will use Lemma 2.

Lemma 2 [Zhang, 2002] If $\mathcal{H}$ is a space such that for all $\boldsymbol{w} \in \mathcal{H}$, it holds $\|\boldsymbol{w}\|_2 \leq a$, then for any $\epsilon > 0$,

$$\log_2 \mathcal{N}_\infty(\mathcal{H}, \epsilon, m) \leq C_1 \frac{a^2}{\epsilon^2} \log_2(1 + m(C_2 \frac{a}{\epsilon} + 4)),$$

where C_1 and C_2 are constants.

We then consider if maximizing the diversity, can lead to constraining the norm of $\boldsymbol{w}_c$, which results Theorem 1.

Theorem 1 Let $\mathcal{H}$ be a space such that for all $\boldsymbol{w} \in \mathcal{H}$, it holds $\|\boldsymbol{w}\|_2 \leq a$. If $\mathcal{H}_c$ is a space such that for all $\boldsymbol{w}_c \in \mathcal{H}_c$ there is a set $H = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_T\} \in \mathcal{H}^T$ satisfying $\boldsymbol{w}_c = \frac{1}{T}\sum_{i=1}^{T} \boldsymbol{w}_i$ and $\mathtt{div}_\infty(H) > q$, then for any $\epsilon > 0$,

$$\log_2 \mathcal{N}_\infty(\mathcal{H}_c, \epsilon, m) \leq C_1 \frac{1}{\epsilon^2}\left(\frac{a^2}{T} + (1-q)a^2\right) \cdot \log_2(1 + m(C_2 \frac{1}{\epsilon}\sqrt{\frac{a^2}{T} + (1-q)a^2} + 4)),$$

where C_1 and C_2 are constants.

Proof. Since $\boldsymbol{w}_c = \frac{1}{T}\sum_{i=1}^{T} \boldsymbol{w}_i$, we explicitly write $\boldsymbol{w}_c$ as $\boldsymbol{w} = [\frac{1}{T}\sum_{i=1}^{T} w_{i,1}, \ldots, \frac{1}{T}\sum_{i=1}^{T} w_{i,d}]$. By the constraint $\mathtt{div}_\infty(H) > q$, we have, for all i, j, $1 - \frac{\boldsymbol{w}_i^\top \boldsymbol{w}_j}{\|\boldsymbol{w}_i\| \cdot \|\boldsymbol{w}_j\|} \geq q$, so that

$$\boldsymbol{w}_i^\top \boldsymbol{w}_j \leq (1-q)\|\boldsymbol{w}_i\| \cdot \|\boldsymbol{w}_j\| \leq (1-q)a^2.$$

Therefore,

$$\|\boldsymbol{w}_c\|_2^2 = \boldsymbol{w}_c^\top \boldsymbol{w}_c = \sum_{j=1}^{d}(\frac{1}{T}\sum_{i=1}^{T} w_{i,j})(\frac{1}{T}\sum_{k=1}^{T} w_{k,j})$$
$$= \frac{1}{T^2}\sum\nolimits_{i=1}^{T} \|\boldsymbol{w}_i\|_2^2 + \frac{2}{T^2}\sum\nolimits_{i=1}^{T-1}\sum\nolimits_{j=i}^{T} \boldsymbol{w}_i{}^\top \boldsymbol{w}_j$$
$$\leq \frac{1}{T}a^2 + \frac{2}{T^2}\frac{T(T-1)}{2}(1-q)a^2 \leq \frac{1}{T}a^2 + (1-q)a^2$$

Applying Lemma 2, the theorem is proved. □

From the proof, it can be observed that, by summing up vectors, the norm of the combined vector is decomposed into the sum of norm of every vector and the inner product between every different vectors, which is interestingly in a similar form to the error-ambiguity decomposition.

It is more interesting to connect maximum diversity principle to maximum entropy principle. The entropy of a vector defined as follows can lead to a bound of covering number.

Definition 2 [Zhang, 2002] The (uniform) entropy of a vector $\boldsymbol{w}$ is defined as

$$\mathtt{entropy}(\boldsymbol{w}) = \sum\nolimits_i^d |w_i| \ln \frac{|w_i|}{\frac{1}{d}\|\boldsymbol{w}\|_1}.$$

Lemma 3 [Zhang, 2002] If $\mathcal{H}$ is a space such that for all $\boldsymbol{w} \in \mathcal{H}$, it holds $\boldsymbol{w}$ has no negative entries, $\|\boldsymbol{w}\|_1 \leq a$, and $\mathtt{entropy}(\boldsymbol{w}) \leq c$, then for any $\epsilon > 0$,

$$\log_2 \mathcal{N}_\infty(\mathcal{H}, \epsilon, m) \leq C_1 \frac{a^2 + ac}{\epsilon^2} \log_2(1 + m(C_2 \frac{a}{\epsilon} + 4)),$$

where C_1 and C_2 are constants.

Lemma 4 Given $H = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_T\}$ be a set of T vectors in $\mathbb{R}^d$, let $\boldsymbol{w}_c = \frac{1}{T}\sum_{i=1}^{T} \boldsymbol{w}_i$. If $\mathtt{div}_\infty(H) > q$, then $\mathtt{entropy}(\boldsymbol{w}_c) \leq \|\boldsymbol{w}_c\|_1 \ln d \leq \sqrt{\frac{1}{T}a + (1-q)a^2} \cdot \sqrt{d} \ln d.$

Lemma 4 is proved by the relationship between 1-norm and the diversity. Besides, we can have an intuitive explanation. The entropy measures the uniformness of a vector. When two vectors are with a large angle, their large components are likely to be in different locations, so that their average leads to a more uniform vector, i.e., large entropy. The proof of Theorem 2 follows Lemma 3 and Lemma 4, and is ignored due to limited page space.

Theorem 2 Let $\mathcal{H}$ be a space such that for all $\boldsymbol{w} \in \mathcal{H}$, it holds $\boldsymbol{w}$ has no negative entries and $\|\boldsymbol{w}\|_1 \leq a$. If $\mathcal{H}_c$ is a space such that for all $\boldsymbol{w}_c \in \mathcal{H}_c$ there is a set $H = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_T\} \in \mathcal{H}^T$ satisfying $\boldsymbol{w}_c = \frac{1}{T}\sum_{i=1}^{T} \boldsymbol{w}_i$ and $\mathtt{div}_\infty(H) > q$, then for any $\epsilon > 0$,

$$\log_2 \mathcal{N}_\infty(\mathcal{H}_c, \epsilon, m) \leq C_1 \frac{d(1 + \ln d)}{\epsilon^2}\left(\frac{a^2}{T} + (1-q)a^2\right) \cdot \log_2(1 + m(C_2 \frac{\sqrt{d}}{\epsilon}\sqrt{\frac{a^2}{T} + (1-q)a^2} + 4)),$$

where C_1 and C_2 are constants.

Remark. Both Theorems 1 and 2 disclose that constraining a large diversity can lead to a small hypothesis space complexity. Notice that the training performance is optimized. According to Lemma 1, a good generalization performance can be expected. It should also be noted that Theorems 1 and 2 do not mean that diversity maximization equals norm constraint or entropy maximization, here we only use them as the bridging techniques for the proofs.

4 Experiments

Fifteen UCI data sets are employed to conduct the experiments. All features are normalized into the interval $[0, 1]$.

Table 1: Comparison of test errors. An entry of DRM is bolded (or italic) if it is significantly better (or worse) than SVM. An entry of Bagging and AdaBoost is marked bullet (or circle) if it is significantly worse (or better) than the DRM with the same component number.

data	DRM_{21}	DRM_{51}	DRM_{101}	SVM	Bag_{21}	Bag_{51}	Bag_{101}	Ada_{21}	Ada_{51}	Ada_{101}
austra	0.141±0.006	0.141±0.005	0.141±0.005	0.138±0.004	0.157±0.008•	0.154±0.007•	0.154±0.007•	0.206±0.016•	0.206±0.015•	0.207±0.014•
australian	0.140±0.006	0.141±0.007	0.140±0.007	0.141±0.006	0.145±0.013	0.142±0.009	0.141±0.010	0.192±0.013•	0.191±0.012•	0.185±0.015•
breastw	0.063±0.039	0.063±0.039	0.063±0.039	0.064±0.039	0.046±0.012	0.046±0.011	0.045±0.009	0.045±0.013	0.044±0.011	0.045±0.012
clean1	**0.358**±0.029	**0.420**±0.033	**0.427**±0.044	0.512±0.013	0.500±0.005•	0.498±0.008•	0.499±0.010•	0.379±0.031	0.314±0.030◦	0.282±0.022◦
diabetes	0.249±0.016	0.251±0.020	0.250±0.019	0.243±0.014	0.246±0.007	0.249±0.009	0.248±0.008	0.275±0.011•	0.278±0.011•	0.281±0.014•
ethn	**0.149**±0.057	**0.173**±0.078	**0.138**±0.050	0.323±0.004	0.321±0.004•	0.320±0.003•	0.320±0.004•	0.041±0.008◦	0.039±0.007◦	0.039±0.007◦
german	**0.280**±0.017	**0.282**±0.017	**0.281**±0.016	0.287±0.008	0.302±0.006•	0.303±0.007•	0.303±0.008•	0.316±0.010•	0.314±0.014•	0.319±0.019•
haberman	**0.303**±0.044	**0.286**±0.022	**0.299**±0.030	0.380±0.048	0.360±0.036•	0.361±0.039•	0.358±0.035•	0.311±0.022	0.303±0.029	0.299±0.026
heart	0.196±0.018	0.199±0.021	0.198±0.021	0.190±0.017	0.210±0.020•	0.212±0.026	0.213±0.026	0.248±0.032•	0.256±0.029•	0.259±0.032•
house-votes	0.097±0.020	0.097±0.020	0.097±0.020	0.097±0.020	0.081±0.006	0.081±0.005◦	0.082±0.006◦	0.079±0.008◦	0.079±0.009◦	0.078±0.010◦
house	0.063±0.019	0.063±0.018	0.063±0.019	0.063±0.019	0.051±0.006	0.050±0.006◦	0.052±0.007	0.064±0.016	0.066±0.015	0.066±0.015
ionosphere	0.337±0.014	0.334±0.015	0.326±0.025	0.337±0.014	0.324±0.006◦	0.324±0.006◦	0.324±0.005	0.258±0.020◦	0.242±0.016◦	0.217±0.018◦
liver-disorders	0.364±0.035	0.361±0.035	0.365±0.039	0.369±0.043	0.362±0.030	0.359±0.029	0.359±0.030	0.338±0.035◦	0.336±0.034◦	0.332±0.030◦
vehicle	**0.226**±0.020	**0.218**±0.019	0.221±0.043	0.243±0.012	0.250±0.016•	0.248±0.019•	0.248±0.016	0.040±0.009◦	0.034±0.011◦	0.027±0.009◦
wdbc	0.031±0.008	0.031±0.009	0.031±0.009	0.031±0.008	0.033±0.007	0.033±0.008	0.032±0.008	0.040±0.010•	0.040±0.009•	0.040±0.010•
average rank	4.87	4.80	4.93	6.4	6.00	6.07	5.60	5.87	5.37	5.10

Table 2: Pairwise win/tie/loss counts of rows against columns.

	DRM_{51}	DRM_{101}	SVM	Bag_{21}	Bag_{51}	Bag_{101}	Ada_{21}	Ada_{51}	Ada_{101}
DRM_{21}	1/14/0	1/14/0	5/10/0	7/7/1	6/6/3	7/6/2	6/4/5	6/3/6	6/3/6
DRM_{51}		6/3/6	5/10/0	6/8/1	6/6/3	6/7/2	7/4/4	6/3/6	6/3/6
DRM_{101}			4/11/0	5/10/0	5/8/2	5/9/1	6/3/6	6/3/6	6/3/6
SVM				3/9/3	3/8/4	3/9/3	6/2/7	6/2/7	6/2/7
Bag_{21}					0/15/0	0/14/1	7/2/6	7/2/6	7/2/6
Bag_{51}						0/15/0	6/3/6	5/4/6	6/3/6
Bag_{101}							6/3/6	6/3/6	6/3/6
Ada_{21}								1/10/4	1/10/4
Ada_{51}									0/11/4

SVM and DRM share the same RBF kernel with the width being the average distance among training instances, and the same parameter ν determined through 5-fold cross-validation on the training set. DRM has two more parameter, μ, controlling the amount of diversity, and T, the number of component learners. μ is also selected by 5-fold cross validation on training sets. We consider three ensemble sizes, that is, 21, 51 and 101. For comparison, we also evaluate the performance of Bagging (abbreviated as Bag) and AdaBoost (abbreviated as Ada). We perform 20 times hold-out tests with random data partitions, half data for training and the other half for testing. Table 1 summarizes the comparison results. On each data set, we assign ranks to methods. The best method receives the rank 1, and the worst the rank 10. The last row of Table 1 presents the average ranks. A two-tail pair-wise t-test with 95% significance level is employed to compare every pair of the evaluated approaches, and the win/tie/loss counts are summarized in Table 2.

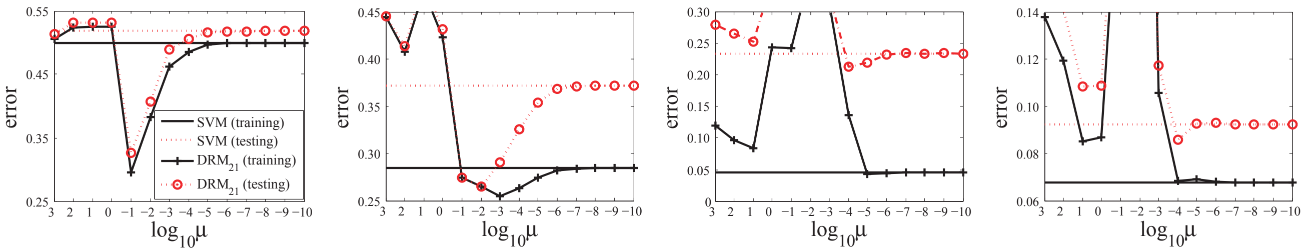
It can be observed from the last row of Table 1 that the average ranks of DRM methods are all lower than 5, while the ranks of the other methods are above 5. From Table 2, by the t-test, DRM is never worse than SVM, while Bagging and AdaBoost are worse than SVM on some data sets. Comparing with Bagging which always makes marginal improvement, DRM can have much larger improvement. Comparing with AdaBoost which is, sometimes, much worse than SVM, DRM is more robust. The t-test shows that DRM has more wins than losses comparing to Bagging, and is comparable with AdaBoost. Moreover, by comparing DRM with difference ensemble size, it can be observed that the size does not have a significant impact on its performance. The observations suggest that DRM has a comparable overall performance with Bagging and AdaBoost, but is more robust.

Figure 1 plots the effect of μ against the training and testing errors on four data sets, where, for a clear view, we don't show the curves in a high error range. First, it can be observed that, when the training error of DRM is not larger than that of SVM, DRM generally has a smaller testing error. This also validates our setting of μ value according to the training error. Second, plots (a) and (b) show that DRM can reduce both the training and test error from SVM, while plots (c) and (d) show that DRM can still reduce the test error even when the training error is higher than SVM. Moreover, it can be observed from plots (b) to (d) that DRM can have a smaller generalization gap (i.e., the gap between training and testing error) than SVM. Particularly, in plot (b), the generalization gap closes to zero when $\log_{10} \mu$ is around -2.

5 Conclusion

Diversity has been recognized as the key to the success of ensemble methods. In contrast to previous methods which encourage diversity in a heuristic way, in this paper, we propose the DRM approach based on mathematical programming framework, which explicitly controls the diversity among the component learners. Theoretical analysis on DRM discloses that the hypothesis space complexity can be effectively reduced by the diversity constraint. The analysis suggests that the diversity control in ensemble methods plays a role similar to the regularization; this provides an explanation to why diversity is important for ensemble methods. Experiment shows that DRM can significantly improve the generalization performance, and is superior to some state-of-the-art ensemble methods. Comparing with Bagging which always makes marginal improvement, DRM can lead to much larger improvement; comparing with AdaBoost which is sometimes much worse than single learner, DRM never loses to single learner in our experiments.

In our analysis, we try to perform simple derivations in order to clarify the main idea. Incorporating more elaborate treatments (e.g. [Smola *et al.*, 2000]) may result in tighter bounds. Moreover, we only concern diversity and leave the consideration of the combination scheme for future work.



(a) *clean1* (b) *haberman* (c) *heart* (d) *house-votes*

Figure 1: Diversity-controlling parameter μ against errors of DRM$_{21}$. Legends of plots (b), (c) and (d) are the same as that of plot (a).

Acknowledgement

The authors want to thank the reviewers for their helpful comments, and thank Min-Ling Zhang and Sheng-Jun Huang for their suggestions to improve the paper.

References

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

[Bartelett, 1998] P.L. Bartelett. The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998.

[Boyd and Vandenberghe, 2004] S.P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, 2004.

[Breiman, 1996] L.E. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[Breiman, 2000] L.E. Breiman. Randomizing outputs to increase prediction accuracy. *Machine Learning*, 40(3):229–242, 2000.

[Breiman, 2001] L.E. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[Brown *et al.*, 2005] G. Brown, J. Wyatt, R. Harris, and X. Yao. Diversity creation methods: A survey and categorisation. *Information Fusion*, 6(1):5–20, 2005.

[Brown, 2009] G. Brown. An information theoretic perspective on multiple classifier systems. In *Proceedings of the 8th International Workshop Multiple Classifier Systems (MCS'09)*, pages 344–353, Reykjavik, Iceland, 2009.

[Chang and Lin, 2001] C.C. Chang and C.J. Lin. LIBSVM: A library for support vector machines, 2001. *Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm*, 2001.

[Dietterich, 2000] T.G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, 40(2):139–157, 2000.

[Dietterich, 2002] T.G. Dietterich. Ensemble learning. In M.A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks, 2nd edition*. MIT Press, Cambridge, MA, 2002.

[Freund and Schapire, 1997] Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[Giacinto and Roli, 2001] G. Giacinto and F. Roli. Design of effective neural network ensembles for image classification purposes. *Image and Vision Computing*, 19(9-10):699–707, 2001.

[Ho, 1998] T.K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998.

[Krogh and Vedelsby, 1995] A. Krogh and J. Vedelsby. Neural network ensembles, cross validation, and active learning. In G. Tesauro, D.S. Touretzky, and T.K. Leen, editors, *Advances in Neural Information Processing Systems 7 (NIPS'94)*, pages 231–238. MIT Press, 1995.

[Kuncheva *et al.*, 2003] L.I. Kuncheva, C.J. Whitaker, C. Shipp, and R. Duin. Limits on the majority vote accuracy in classifier fusion. *Pattern Analysis and Applications*, 6(1):22–31, 2003.

[Luo and Tseng, 1992] Z. Luo and P. Tseng. On the convergence of the coordinate descent method for convex differentiable minimization. *Journal of Optimization Theory and Applications*, 72(1):7–35, 1992.

[Schölkopf *et al.*, 2000] B. Schölkopf, A.J. Smola, R.C. Williamson, and P.L. Bartlett. New support vector algorithms. *Neural Computation*, 12(5):1207–1245, 2000.

[Smola *et al.*, 2000] Alex J. Smola, John Shawe-Taylor, B. Schölkopf, and R. C. Williamson. The entropy regularization information criterion. In S.A. Solla, T.K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12 (NIPS'99)*, pages 342–348. The MIT Press, 2000.

[Ueda and Nakano, 1996] P. Ueda and R. Nakano. Generalization error of ensemble estimators. In *Proceedings of International Conference on Neural Networks (ICNN'96)*, pages 90–95, Washington, DC, 1996.

[Valiant, 1984] L. Valiant. A theory of the learnable. *Communication of the ACM*, 27(11):1134–1142, 1984.

[Vapnik, 1995] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, NY, 1995.

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

[Zhou and Li, 2010] Z.-H. Zhou and N. Li. Multi-information ensemble diversity. In *Proceedings of the 9th International Workshop on Multiple Classifier Systems (MCS'10)*, pages 134–144, Cairo, Egypt, 2010.

[Zhou and Yu, 2005] Z.-H. Zhou and Y. Yu. Ensembling local learners through multi-modal perturbation. *IEEE Transaction on System, Man, And Cybernetics - Part B: Cybernetics*, 35(4):725–735, 2005.