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 heuristic means. 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 = E − A, where E is the mean-square error of the ensemble, E is the average mean-square error of component learners, and A is the average difference between the ensemble and the component learners. This decomposition implies that, as long as E is fixed, a higher difference among the component learners leads to a better ensemble. Later results achieved by the bias-variance-covariance decomposition imply that, as long as E is fixed, a higher difference among the component learners leads to a better ensemble. Later results achieved by the bias-variance-covariance decomposition [Krogh and Vedelsby, 1995] for regression tasks as E = E − A, where E is the mean-square error of the ensemble, E is the average mean-square error of component learners, and A is the average difference between the ensemble and the component learners. This decomposition implies that, as long as 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 D over X. A hypothesis, or a learner, is a function h : X → {−1, 1}, and a concept c is an underlying hypothesis. A training set is a set of m examples S = {(x_i, y_i)}_{i=1}^m, where x_i ∈ X are drawn i.i.d. under D and y_i = c(x_i). A learning algorithm is to select a hypothesis h from a feasible hypothesis space H according to the given training set. For an integer number n, we denote [n] as the set {1, 2, …, n}.

2.1 Diversity Measure

We consider linear classification model w ∈ R^d, which classifies an instance by the inner product w^T φ(x), where φ(x)
is the feature mapping of the instance $x$. Note that the model can be viewed as equivalent to the commonly used model $w^\top \phi(x) + b$, since by extending the mapping $\phi(x)$ to one extra dimension with a constant value, the extended $w$ 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], $\kappa$ statistic measure [Dietterich, 2000], etc. Thus we also consider a form of diversity based on pairwise difference. Given a pairwise diversity measure $\text{div}$ 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

$$\text{div}_p(H) = \left( \sum_{1 \leq i \neq j \leq T} \text{div}(h_i, h_j) \right)^{1/p}.$$  

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 $w_1$ and $w_2$, we measure their diversity using the angle between them,

$$\text{div}(w_1, w_2) = 1 - \frac{w_1^\top w_2}{\|w_1\| \cdot \|w_2\|},$$

so that the larger value of $\text{div}$, 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. For the training $T$ linear learners, $w_1, \ldots, w_T$, can be described as:

$$\arg \min_{w_1, \ldots, w_T} \sum_{t=1}^{T} \sum_{i=1}^{m} \ell(y_t, w_t^\top \phi(x_i))$$

s.t. $\|w_t\| \leq \theta$ $(\forall t \in [T])$,

$$\text{div}_p(w_1, \ldots, w_T) \geq q,$$

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

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

$$\arg \min_{(w_t, \rho_t, \xi_t)} \sum_{t=1}^{T} (-\nu \frac{\rho_t}{T} + \frac{1}{m} \sum_{i=1}^{m} \xi_{t,i}^2 + \mu \sum_{1 \leq i < t \leq T} \frac{w_t^\top w_t'}{\|w_t\| \cdot \|w_t'\|})$$

s.t. $y_t w_t^\top \phi(x_i) \geq \rho_t - \xi_{t,i}$ $(\forall i \in [m], \forall t \in [T])$,

$$\|w_t\|_2 \leq 1$ $(\forall t \in [T]).$$

2.3 Optimization

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

$$\arg \min_{w_t, \rho_t, \xi_t} \sum_{t=1}^{T} (-\nu \frac{\rho_t}{T} + \frac{1}{m} \sum_{i=1}^{m} \xi_{t,i}^2 + \mu \sum_{1 \leq i \neq t \leq T} \|w_t + w_t'\|_2^2$$

s.t. $y_t w_t^\top \phi(x_i) \geq \rho_t - \xi_{t,i}$ $(\forall i \in [m], \forall t \in [T])$,

$$\|w_t\|_2^2 \leq 1$ $(\forall t \in [T]).$$

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 $\{w_t, \rho_t, \xi_t\}$ while fixing the other variables $\{w_t', \rho_t', \xi_t'\}$ for all $t' \neq t$ as constants. Mathematically, in each step we are solving the following small QCQP problem for each $t$:

$$\arg \min_{w_t, \rho_t, \xi_t} -\nu \frac{\rho_t}{T} + \frac{1}{m} \sum_{i=1}^{m} \xi_{t,i}^2 + \mu \sum_{t' \neq t} \|w_t + w_{t'}\|_2^2$$

s.t. $y_t w_t^\top \phi(x_i) \geq \rho_t - \xi_{t,i}$ $(\forall i \in [m], \forall t \in [T])$,

$$\|w_t\|_2^2 \leq 1.$$

Further, the above QCQP problem can be addressed via sequential quadratic programming efficiently. Introducing the Lagrange multipliers $\alpha_t$ and $\lambda_t$ for the constraints in Eq.4, we have:

$L(w_t, \rho_t, \xi_t; \alpha_t, \lambda_t) = -\nu \frac{\rho_t}{T} + \frac{1}{m} \sum_{i=1}^{m} \xi_{t,i}^2 + \mu \sum_{t' \neq t} \|w_t + w_{t'}\|_2^2 + \lambda_t (\|w_t\|_2^2 - 1) - \sum_{i=1}^{m} \alpha_t \xi_{t,i} (y_t w_t^\top \phi(x_i) - \rho_t + \xi_{t,i})$,

where $\alpha_t = [\alpha_{t,1}, \ldots, \alpha_{t,n}]$. Setting the partial derivations w.r.t. $\{w_t, \rho_t, \xi_t\}$ to zeros, we have:

$$\frac{\partial L}{\partial w_t} = 2 \mu \sum_{t' \neq t} (w_t + w_{t'}) + 2 \lambda_t w_t - \sum_{i=1}^{m} \alpha_t \xi_{t,i} \phi(x_i) = 0,$$

$$\frac{\partial L}{\partial \rho_t} = -\nu \frac{1}{T} + \sum_{i=1}^{m} \alpha_t \xi_{t,i} = 0,$$

$$\frac{\partial L}{\partial \xi_{t,i}} = 2 \xi_{t,i} - \alpha_t \xi_{t,i} = 0.$$

Let $w_t = \sum_{t' \neq t} w_{t'}$, we then obtain the solution of $w_t$ as:

$$w_t = \frac{2 \mu w_t + \sum_{i=1}^{m} \alpha_t \xi_{t,i} \phi(x_i)}{2((T-1)\mu + \lambda_t)}.$$
and thus the dual of Eq.(4) can be cast as:
\[
\arg\min_{\alpha_t, \lambda_t} \frac{1}{2} \alpha_t^T \left( \frac{K \odot yy^T}{2(\lambda_t + (T-1)\mu)} + \frac{m}{2} \right) \alpha_t - r^T \alpha_t
\]
\[
\text{s.t. } \sum_{i=1}^{m} \alpha_{t,i} = \nu \frac{1}{T}, \quad \alpha_t \geq 0, \lambda_t \geq 0.
\]
which is jointly-convex for \( \{\alpha_t, \lambda_t\} \) [Boyd and Vandenbergh, 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 \( \lambda_t \) is fixed, \( \alpha_t \) can be solved via:
\[
\arg\min_{\alpha_t} \frac{1}{2} \alpha_t^T \left( \frac{K \odot yy^T}{2(\lambda_t + (T-1)\mu)} + \frac{m}{2} \right) \alpha_t - r^T \alpha_t
\]
\[
\text{s.t. } \sum_{i=1}^{m} \alpha_{t,i} = \nu \frac{1}{T}, \quad \alpha_t \geq 0, \lambda_t \geq 0.
\]
where \( K \) is the kernel matrix of \( \phi(x) \) and \( \odot \) is the entry-wise product, and \( r \) is a vector with components:
\[
r_i = \frac{\nu y_i \phi(x_i)^T \phi(x_i)}{(\lambda_t + (T-1)\mu)}.
\]
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 \( \alpha_t \) is fixed, \( \lambda_t \) can be solved in a closed-form, i.e.,
\[
\lambda_t = \frac{\| - 2\mu w_0 + \sum_{i=1}^{m} \alpha_{t,i} y_i \phi(x_i) \|_2^2}{2} - (T - 1)\mu
\]
\[
= \max \{0, \frac{\| - 2\mu w_0 + \sum_{i=1}^{m} \alpha_{t,i} y_i \phi(x_i) \|_2^2}{2} - (T - 1)\mu\}.
\]
Algorithm 1 presents the pseudocode of the DRM. It is worth noticing that, when the optimal solution of all \( \lambda_t \)'s are non-zeros, according to KKT condition, the optimal solution \( \{w_t^*, \rho_t^*, \xi_t^*\}_{t=1}^{T} \) obtained by DRM according to Eq.(3) satisfies \( \|w_t^*\|_2^2 = 1 \), thus is also the optimal solution of Eq.(2).

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 \( y_h(x) \). The training error with margin \( \gamma \) of a hypothesis \( h \) is defined as
\[
\epsilon^t_g(h) = \frac{1}{m} \sum_{i=1}^{m} I[h(x_i)y_i < \gamma],
\]
where \( I \) is the indicator function that outputs 1 if its inner expression is true and 0 otherwise. Define the generalization error, or true error, as
\[
\epsilon_g(h) = \mathbb{E}_{x \sim D}[I[h(x)c(x) < 0]].
\]

Algorithm 1 DRM

Input: Training set \( S = \{(x_i, y_i)\}_{i=1}^{m} \) and kernel matrix \( K \), parameters \( T \) and \( \mu \).

Process:

1: \( \lambda_t \leftarrow 1 \) and \( w_t \leftarrow 0 \), \( \forall t \in [T] \)
2: while not converged yet do
3: \( for \ t = 1, \ldots, T \) do
4: \( while \ not \ converged \ yet \ do \)
5: \( \alpha_t \leftarrow \) solutions returned by Eq. (6)
6: \( \lambda_t \leftarrow \) solutions returned by Eq. (7)
7: \( end \ while \)
8: \( set \ w_t \leftarrow \) according to Eq. (5)
9: \( end \ for \)
10: end while

Output: \( w_c = \frac{1}{T} \sum_{t=1}^{T} w_t \)

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 \( S = \{x_1, \ldots, x_m\} \) and a function space \( F \), characterize every \( f \in F \) using a vector \( v_S(f) = [f(x_1), \ldots, f(x_m)] \) in a metric space \( B^m \) with metric \( \rho \). The covering number \( N_p(F, \gamma, S) \) is the minimum number \( l \) of vectors \( u_1, \ldots, u_l \in B^m \) such that, for all \( f \in F \) there exists \( j \in \{1, \ldots, l\} \):
\[
\|\rho(v_S(f), u_j)\|_p \leq \left( \sum_{i=1}^{m} \rho(f(x_i), u_j)^p \right)^{\frac{1}{p}} \leq m^{\frac{1}{p}},
\]
and
\[
N_p(F, \gamma, m) = \sup_{S:|S|=m} N_p(F, \gamma, S).
\]

Lemma 1 [Bartlett, 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
\[
\epsilon_g(A) \leq \epsilon^c_g(A) + \sqrt{\frac{2}{m} \left( \ln N_{\infty}(H, \gamma/2, 2m) + \ln \frac{2}{\delta} \right)},
\]
where \( N_{\infty} \) 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.

3.2 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:
\[
\ell_T(w_t) = (-\nu \rho_t + \frac{1}{m} \sum_{i=1}^{m} \xi_{t,i}^2),
\]
where \( \xi_{t,i} \geq \rho_t - y_i w_t \phi(x_i) \forall i \in [m] \) by the constraints. We concern about the loss of the combined hypothesis \( w_c = \frac{1}{T} \sum_{t=1}^{T} w_t \) with \( \rho_c = \frac{1}{T} \sum_{t=1}^{T} \rho_t \):
\[
\ell_c(w_c) = (-\nu \rho_c + \frac{1}{m} \sum_{i=1}^{m} \xi_i^2),
\]
where $\xi_i \geq \rho_c - y_i w_i^\top \phi(x_i) (\forall i \in [m])$. We then have the following proposition.

**Proposition 1** Let $w_1, \ldots, w_T$ be the component hypotheses solved by DRM, and $w_c = \frac{1}{T} \sum_{t=1}^T w_t$ be the combined hypothesis, the loss of $w_c$ is bounded as

$$\ell_c(w_c) \leq \sum_{t=1}^T \ell_T(w_t).$$

**Proof.** Notice that, for a training instance $x$, the classification margin $y w_i^\top \phi(x)$ can be expanded as $y \frac{1}{T} \sum_{t=1}^T \xi_t - \frac{1}{T} \sum_{t=1}^T \xi_t$ by the constraints in Eq.(2). Since $\rho_c = \frac{1}{T} \sum_{t=1}^T \rho_c$, the proposition is then proved by that $\sum_{t=1}^T \xi_t^2 \leq \left(\frac{T}{T} \sum_{t=1}^T \xi_t^2\right)^2$.

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 conceptually distinguish the hypothesis space of the component hypothesis and that of the combined hypothesis. Let $w_1, \ldots, w_T$ be hypotheses from space $H_t$, and the combined $w_c$ be in space $H_c$. We will use Lemma 2.

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

$$\log_2 N_\infty(H, \epsilon, m) \leq C_1 \frac{2}{\epsilon^2} \log_2 (1 + m(C_2 - \epsilon + 4)),$$

where $C_1$ and $C_2$ are constants.

We then consider if maximizing the diversity, can lead to constraining the norm of $w_c$, which results Theorem 1.

**Theorem 1** Let $H$ be a space such that for all $w \in H$, it holds $||w||_2 \leq a$. If $H_c$ is a space such that for all $w_c \in H_c$ there is a set $H = \{w_1, \ldots, w_T\} \in H_T$ satisfying $w_c = \frac{1}{T} \sum_{t=1}^T w_t$ and $\text{div}_c(H) > q$, then for any $\epsilon > 0$,

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

where $C_1$ and $C_2$ are constants.

**Proof.** Since $w_c = \frac{1}{T} \sum_{i=1}^T w_i$, we explicitly write $w_c$ as $w = \left[\frac{1}{T} \sum_{i=1}^T w_i, \ldots, \frac{1}{T} \sum_{i=1}^T w_i, d\right]$. By the constraint $\text{div}_c(H) > q$, we have, for all $i, j, k \in \{1, \ldots, T\}$ satisfying $w_c = \frac{1}{T} \sum_{i=1}^T w_i$, we have $\text{div}_c(H) > q$, then for any $\epsilon > 0$,

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

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 $[0, 1]$. 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 $w$ is defined as

$$\text{entropy}(w) = \sum_{i=1}^d (w_i \ln |w_i| - \frac{1}{2} \|w\|_1).$$

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

$$\log_2 N_\infty(H, \epsilon, m) \leq C_1 \frac{2}{\epsilon^2} \log_2 (1 + m(C_2 - \epsilon + 4)),$$

where $C_1$ and $C_2$ are constants.

**Lemma 4** Given $H = \{w_1, \ldots, w_T\}$ be a set of $T$ vectors in $\mathbb{R}^d$, let $w_c = \frac{1}{T} \sum_{i=1}^T w_i$. If $\text{div}_c(H) > q$, then

$$\text{entropy}(w_c) \leq \|w_c\|_1 \ln d \leq \sqrt{T} a + (1-q)a^2 \cdot \sqrt{\ln d}.$$
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.

<table>
<thead>
<tr>
<th>data</th>
<th>DRM21</th>
<th>DRM31</th>
<th>DRM101</th>
<th>SVM</th>
<th>Bag21</th>
<th>Bag51</th>
<th>Bag101</th>
<th>Ada21</th>
<th>Ada51</th>
<th>Ada101</th>
</tr>
</thead>
<tbody>
<tr>
<td>austria</td>
<td>0.141</td>
<td>0.103</td>
<td>0.140</td>
<td>0.141</td>
<td>0.140</td>
<td>0.138</td>
<td>0.138</td>
<td>0.138</td>
<td>0.138</td>
<td>0.138</td>
</tr>
<tr>
<td>australian</td>
<td>0.140</td>
<td>0.103</td>
<td>0.140</td>
<td>0.140</td>
<td>0.140</td>
<td>0.138</td>
<td>0.138</td>
<td>0.138</td>
<td>0.138</td>
<td>0.138</td>
</tr>
<tr>
<td>breastw</td>
<td>0.063</td>
<td>0.039</td>
<td>0.063</td>
<td>0.063</td>
<td>0.063</td>
<td>0.063</td>
<td>0.063</td>
<td>0.063</td>
<td>0.063</td>
<td>0.063</td>
</tr>
<tr>
<td>clean</td>
<td>0.358</td>
<td>0.029</td>
<td>0.420</td>
<td>0.333</td>
<td>0.427</td>
<td>0.424</td>
<td>0.512</td>
<td>0.512</td>
<td>0.512</td>
<td>0.512</td>
</tr>
<tr>
<td>diabetes</td>
<td>0.249</td>
<td>0.016</td>
<td>0.251</td>
<td>0.020</td>
<td>0.250</td>
<td>0.250</td>
<td>0.243</td>
<td>0.243</td>
<td>0.243</td>
<td>0.243</td>
</tr>
<tr>
<td>e10h</td>
<td>0.149</td>
<td>0.057</td>
<td>0.173</td>
<td>0.078</td>
<td>0.138</td>
<td>0.500</td>
<td>0.323</td>
<td>0.004</td>
<td>0.004</td>
<td>0.004</td>
</tr>
<tr>
<td>german</td>
<td>0.280</td>
<td>0.017</td>
<td>0.282</td>
<td>0.017</td>
<td>0.281</td>
<td>0.286</td>
<td>0.287</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
</tr>
<tr>
<td>haberman</td>
<td>0.303</td>
<td>0.044</td>
<td>0.286</td>
<td>0.022</td>
<td>0.299</td>
<td>0.300</td>
<td>0.380</td>
<td>0.048</td>
<td>0.048</td>
<td>0.048</td>
</tr>
<tr>
<td>heart</td>
<td>0.196</td>
<td>0.018</td>
<td>0.199</td>
<td>0.021</td>
<td>0.198</td>
<td>0.190</td>
<td>0.190</td>
<td>0.17</td>
<td>0.17</td>
<td>0.17</td>
</tr>
<tr>
<td>house-votes</td>
<td>0.097</td>
<td>0.020</td>
<td>0.097</td>
<td>0.020</td>
<td>0.097</td>
<td>0.20</td>
<td>0.190</td>
<td>0.17</td>
<td>0.17</td>
<td>0.17</td>
</tr>
<tr>
<td>house</td>
<td>0.063</td>
<td>0.019</td>
<td>0.063</td>
<td>0.018</td>
<td>0.063</td>
<td>0.019</td>
<td>0.063</td>
<td>0.019</td>
<td>0.019</td>
<td>0.019</td>
</tr>
<tr>
<td>ionosphere</td>
<td>0.337</td>
<td>0.014</td>
<td>0.334</td>
<td>0.015</td>
<td>0.326</td>
<td>0.025</td>
<td>0.337</td>
<td>0.014</td>
<td>0.014</td>
<td>0.014</td>
</tr>
<tr>
<td>liver-disorders</td>
<td>0.364</td>
<td>0.035</td>
<td>0.361</td>
<td>0.035</td>
<td>0.365</td>
<td>0.369</td>
<td>0.433</td>
<td>0.433</td>
<td>0.433</td>
<td>0.433</td>
</tr>
<tr>
<td>vehicle</td>
<td>0.226</td>
<td>0.020</td>
<td>0.218</td>
<td>0.019</td>
<td>0.221</td>
<td>0.433</td>
<td>0.369</td>
<td>0.433</td>
<td>0.433</td>
<td>0.433</td>
</tr>
<tr>
<td>wdbc</td>
<td>0.031</td>
<td>0.008</td>
<td>0.031</td>
<td>0.009</td>
<td>0.031</td>
<td>0.009</td>
<td>0.031</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
</tr>
<tr>
<td>average rank</td>
<td>4.87</td>
<td>4.80</td>
<td>4.93</td>
<td>6.4</td>
<td>6.00</td>
<td>6.07</td>
<td>5.60</td>
<td>5.87</td>
<td>5.87</td>
<td>5.10</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Bag21</td>
<td>0/150</td>
<td>0/141</td>
<td>0/141</td>
<td>0/141</td>
<td>0/27</td>
<td>0/27</td>
<td>0/27</td>
<td>0/27</td>
<td>0/27</td>
<td>0/27</td>
<td>0/27</td>
<td>0/27</td>
<td>0/27</td>
<td>0/27</td>
</tr>
<tr>
<td>Bag51</td>
<td>0/150</td>
<td>0/141</td>
<td>0/141</td>
<td>0/141</td>
<td>0/27</td>
<td>0/27</td>
<td>0/27</td>
<td>0/27</td>
<td>0/27</td>
<td>0/27</td>
<td>0/27</td>
<td>0/27</td>
<td>0/27</td>
<td>0/27</td>
</tr>
<tr>
<td>Bag101</td>
<td>0/150</td>
<td>0/141</td>
<td>0/141</td>
<td>0/141</td>
<td>0/27</td>
<td>0/27</td>
<td>0/27</td>
<td>0/27</td>
<td>0/27</td>
<td>0/27</td>
<td>0/27</td>
<td>0/27</td>
<td>0/27</td>
<td>0/27</td>
</tr>
<tr>
<td>Ada21</td>
<td>1/100</td>
<td>1/100</td>
<td>1/100</td>
<td>1/100</td>
<td>1/100</td>
<td>1/100</td>
<td>1/100</td>
<td>1/100</td>
<td>1/100</td>
<td>1/100</td>
<td>1/100</td>
<td>1/100</td>
<td>1/100</td>
<td>1/100</td>
</tr>
<tr>
<td>Ada51</td>
<td>0/110</td>
<td>0/110</td>
<td>0/110</td>
<td>0/110</td>
<td>0/110</td>
<td>0/110</td>
<td>0/110</td>
<td>0/110</td>
<td>0/110</td>
<td>0/110</td>
<td>0/110</td>
<td>0/110</td>
<td>0/110</td>
<td>0/110</td>
</tr>
</tbody>
</table>

SVM and DRM share the same RBF kernel with the width being the average distance among training instances, and the same parameter $\nu$ determined through 5-fold cross-validation. However, the role of the parameter $\nu$ is different, and $T$, the number of component learners. $\nu$ is also selected by 5-fold cross-validation. SVM and DRM have the same parameter $\nu$. The diversity constraint in SVM significantly reduces the generalization error. The diversity constraint in DRM contributes to the diversity constraint. The diversity constraint in DRM can lead to much larger improvement than SVM. In contrast to previous methods which may 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.

5 Conclusion

Diversity has been recognized as the key to the success of ensemble methods. In contrast to previous methods which may 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.
Figure 1: Diversity-controlling parameter $\mu$ 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


