Data Dependant Learners Ensemble Pruning

Gang Zhang[1, 2], Jian Yin[2], Xiaomin He[1] and Lianglun Cheng[1]
1 Faculty of Automation, GuangDong University of Technology, GuangZhou, China
2 Computer Science Department, SUN YAT-SEN University, GuangZhou, China
Email: ipx@gdut.edu.cn, issjyin@mail.sysu.edu.cn, teacher_smc@163.com, LLCheng@gdut.edu.cn

***Abstract*—Ensemble learning aims at combining several slightly different learners to construct stronger learner. Ensemble of a well selected subset of learners would outperform than ensemble of all. However, the well studied accuracy / diversity ensemble pruning framework would lead to over fit of training data, which results a target learner of relatively low generalization ability. We propose to ensemble with base learners trained by both labeled and unlabeled data, by adopting data dependant kernel mapping, which has been proved successful in semi-supervised learning, to get more generalized base learners. We bootstrap both training data and unlabeled data, namely point cloud, to build slight different data set, then construct data dependant kernel. With such kernels data point can be mapped to different feature space which results effective ensemble. We also proof that ensemble of learners trained by both labeled and unlabeled data is of better generalization ability in the meaning of graph Laplacian. Experiments on UCI data repository show the effectiveness of the proposed method.**

***Index Terms*—ensemble learning; generalization ability; data dependant kernel; kernel mapping; point cloud kernel**

I. INTRODUCTION

Ensemble learning combines base learners to get better result than individual ones. It is widely accepted that it works if base learners are well selected by some criteria, and ensemble of a small size but carefully selected subset could be better than ensemble of all [3]. Currently there are two main selection strategies: weighted ensemble and ensemble pruning. For the former, a weight is associated with each base learner and output of final ensemble is the weighted sum of output of all base learners. Often we confine all weights are non-negative and sum to 1. For the latter, there is also a weight associated with each learner but it is either 0 or 1. In another word, the learners with weights 0 would be discarded. We can say the latter is a special case of the former.

There are many efforts to find optimal weights in both cases, though it is a NP-hard problem in essence. The subensemble selection problem is a combinatorial optimization problem and it is computationally prohibitive to search for the best solution directly [2]. But it is possible to get near-optimal subensembles through some approximate algorithms. Examples include GASEN [3] which based on an evolution algorithm, and SDP [2] which relied on a quadratic optimization process by searching in whole solution space to find the optimal solution. By selecting ensemble members with some strategies, they construct an effective weighted subensemble with smaller size and stronger generalization ability than original ones.

If ensemble members are aggregated in an appropriate order, the error rate of a subensemble by top k members can still get better results. This is the so-called ensemble pruning. Many ensemble pruning approaches exist in ensemble learning research, such as Reduce-Error Pruning [4], Kappa Pruning [4], Orientation Ordering (OO) [5] and Ensemble Pruning via Individual Contribution (EPIC) [6]. Efficiency of algorithms in this family is highly relied on ranking metric of member learners. Such metrics include accuracy and diversity and their weighted combination. There is some work on the relationship between performance of ensemble and such metric of individual learners [9].

However, previous ensemble pruning framework does not consider generalization ability of the final subensemble which means the performance on unseen data. Recently, a semi-supervised consideration of ensemble pruning is proposed. [10] gave a general review of relationship between ensemble learning and semi-supervised learning, and stated that it would be helpful if they are put together properly. [11] showed such helpfulness by adopting unlabeled data to get maximized diversity of learners in ensemble learning.

Generalization ability of ensemble or subensemble is coming into consideration currently. In a series algorithms like RNCL [7] targeted at ANN, a regularization term formed by weights in ANN is added to the optimization problem to get relative small weights of trained networks. Generalization ability formularized as Graph Laplacian regularizer [13] of ensemble over the whole dataset is considered in [14] and it is naturally suitable for both semi-supervised and transductive settings.

However, the idea of this paper lies in a different direction. We notice that in ensemble pruning, a subensemble constructed by members with better generalization ability also performs better in generalization. It is necessary to integrate generalization ability consideration in the process of training individual learners, and it can guarantee that all member learners have relative better generalization ability before pruning, which forms the main intuition of our work.

In this paper, we first introduce a data dependant kernel and plug it into a kernel learner. And then propose an efficient method to comprehensively evaluate contribution of individual learners with better generalization ability, which synthetically considers accuracy and diversity. An algorithm named Ensemble

doi:10.4304/jsw.7.4.919-926

Pruning for Data Dependant Learners (DDLEP) that trains individual learners with good generalization ability and incorporates ensemble members with order of hybrid metric value into subensembles is proposed as the main algorithm of this paper.

We also note that the proposed method can be used in both single-instance and multi-instance learning. For the former case, a semi-supervised single-instance kernel is used while in the latter case, we introduce a multi-instance version of such kernel with Graph Laplacian of bags.

Experimental results on 10 UCI data sets for single-instance setting and 4 data sets for multi-instance setting show that subensembles formed by DDLEP are of better performance than the original ensemble and traditional semi-supervised methods. In all experiments, bootstrap is chosen to be the method for constructing the original ensemble since it has been shown to be a safe and robust method [8]. Current state-of-the-art ensemble pruning methods EPIC [6] and traditional non-ensemble semi-supervised learning algorithm [15] are chosen to be peer methods. Our experiments show that DDLEP outperforms them on most data sets.

The paper is structured as following. Section II reviews some important related work of our topic. Section III presents some basic concept and theory on data dependant kernel mapping and the accuracy / diversity trade-off in ensemble learning. In Section IV we present the main algorithm of this paper. We report experimental results in Section V and conclude in Section VI.

II. Related Work

An ensemble is composed of multiple learners that work together in some way to make decisions. It has been found that ensembles are often much more accurate than their member learners. Dietterich [12] stated that an ensemble of classifiers is more accurate than any of its ingredient if the members are of better accuracy and diversity. Gavin Brown [9] has explained the effectiveness of ensemble learning of regression in a theoretical view. He pointed out that the key properties of individual learners that affected final ensemble were accuracy and diversity. [10] gave some analysis on semi-supervised learning and ensemble learning, and presented some methods to incorporate both in a united framework.

There is some work on regularization of ensemble learning. RNCL [7] introduced a regularization term of member ANN weights in the optimization problem to improve generalization of the overall ensemble. But it is confined to ensemble of ANNs and it is not able to make use of unlabeled data. [14] proposed a regularization framework which directly used unlabeled data to improve generalization ability of weighted ensemble. A quadratic optimization is launched to get the optimal solution. The method is naturally suitable for semi-supervised and transductive learning, and with a threshold for weights, it can achieve weighted ensemble and pruning. However, we point out that the method in [14] may not get the optimal solution with all seen data, since the unlabeled data does not appear in construction of member learners. [11] is a more recent work on semi-supervised ensemble learning. The main idea is to get a maximized diversity estimation of the ensemble on unlabeled data while keeping accuracy on labeled data. But the unlabeled data is not considered while building member classifiers in a semi-supervised manner.

Recent semi-supervised learning methods can greatly improve generalization ability of learners. We introduce the work presented in [15] to training member learners, which makes use of unlabeled data by a deformed kernel named point cloud kernel. It introduced a modified RKHS with a norm determined by a set of data points. By combining manifold assumption and cluster assumption, the reproducing kernel appropriately warps an RKHS and is adapted to the geometry of data distribution. The point cloud is irrelevant to labels and can be constructed by all seen data points, which provides a natural way to get learners of better generalization ability [16][17].

Pruning is an approach to find better ensemble based on learners ranking at relatively low computational cost. Famous ensemble pruning include Reduce-Error (RE) pruning [4], Kappa Pruning, Orientation Ordering (OO) [5] and Ensemble Pruning via Individual Contribution ordering (EPIC) [6]. Current algorithms mainly adopt accuracy and diversity as ranking metric. However, there is still no work in the literature considering generalization ability of individual learners in ensemble pruning. A doctor thesis [18] presented some empirical study on pruning with the concern of accuracy, diversity and generalization ability of individual learners and solved it as a multi-objective optimization problem. The weak point is that supervised learners of poor generalization ability would be generated by bootstrapping training data set only. As mentioned before, we tackle this problem by adopting semi-supervised learning with unlabeled data to guarantee better performance on unseen data before pruning.

III. Pruing of Semi-Supervised Learners

In this section we first present semi-supervised learning with data-dependant kernel, and then detail ensemble pruning with accuracy and re-defined diversity.

A. Data Dependant Kernel

This section briefly describes the data dependant kernel proposed in [15] [16]. Here the main idea is to warp an RKHS using a redefined norm determined by a set of points. Then a modified kernel is obtained by reproducing property and the point cloud dependant norm.

Formally, given a data set of both labeled and unlabeled data $X = X_l \cup X_u$, $x_i \in X$, where X_l is a labeled data set and X_u is an unlabeled data set. The learning task is to find a function f on X in a hypnosis space H such that:

$$f = \arg\min_{h \in H} \{ \frac{1}{l} \sum_{i=1}^{l} V(h, x_i, y_i) + \|h\|_H^2 \} \qquad (1)$$

Where V is a loss function and $\|\cdot\|_H$ is a norm defined on H. If $\|\cdot\|_H$ is standard norm on H, then we can get the solution of (1) by Representer Theorem.

$$f(x)=\sum_{i=1}^{l}\alpha_i\cdot k(x,x_i) \tag{2}$$

k is the kernel of H. Now we are interested in the case that $\|\cdot\|_H$ is determined by a point cloud. Named the new space $\tilde{H}$ with a different inner product:

$$<f,g>_{\tilde{H}}=<f,g>_H+<Sf,Sg>_V \tag{3}$$

$$S:H\to V=R^n \quad Sf=(f(x_1),...,f(x_l)) \tag{4}$$

$$F=(f(x_1),...,f(x_l)) \tag{5}$$

V is a linear space with norm $\|Sf\|_V^2=F^TMF$, where M is a symmetric positive semi-defined matrix reflecting the geometry structure of the space. [15] proved that $\tilde{H}$ is still an RKHS with kernel $\tilde{k}$ as following:

$$\tilde{k}(x,z)=k(x,z)-k_x^T(I+MK)^{-1}M\cdot k_z \tag{6}$$

Where $k_x=(k(x_1,x),...,k(x_l,x))$ and K is Gram matrix. With kernel $\tilde{k}$, unlabeled data can be incorporated into supervised kernel learning machine, such as SVM.

B. *Individual Contribution Assessment*

In ensemble pruning, diversity and accuracy of individual learners affect the quality of ensemble tremendously. Below we give the formal definition of both metrics used in our algorithm. For simplicity, we only consider bi-classification problem.

Let $D=\{(x_i,y_i)\,|\,i\in[1,n],x_i\in X,y_i\in[+1,-1]\}$ be a set of n data points with vectorical representation, x is properties associated with a data point and y is class label. $E=\{c_1,...,c_m\}$ is a set of m classifiers with $f_i(x_j)$ is the prediction of the ith classifier on the jth data point. The accuracy Acc_i of the ith ensemble member is defined as its probability of correct prediction on data set.

There are various diversity definitions of set of learners in the literature [10], with different background from problem domain to learning framework. In this paper, we employ a modified disagreement measure based on [9]'s definition. Intuitively speaking, only shift between member classifier and ensemble is considered in this diversity definition.

Definition 1: The continuous prediction of ensemble $E=\{c_1,..,c_m\}$ on a data point x is defined as following:

$$c_{ens\_cont}(x)=\frac{\sum_{k=1}^{m}c_k(x)}{m},x\in D \tag{7}$$

Definition 2: Diversity of classifier c_i to the ensemble, denoted as Div_i, is defined as following:

$$Div_i=\frac{\sum_{j=1}^{n}(c_i(x_j)-c_{ens\_cont}(x_j))^2\cdot\delta_k}{n^2} \tag{8}$$

δ in (8) is a penalty parameter such that:

$$\delta=\begin{cases}1 \text{ iff } c_i(x_i)=y_i\\ 0 \text{ otherwise}\end{cases}$$

The definition of δ only considers the contribution of correct classification samples. Thus the diversity in our framework is closely related to accuracy, which means accuracy and diversity would not contradict to each other. We show this point empirically.

59	59
56	56
129	155
155	129
152	161
161	152
165	173
173	165
118	118
106	106
64	64
149	149
85	85
31	31
194	194
123	166
166	123
47	47

107	107
44	197
60	60
62	164
85	117
115	175
117	190
118	85
126	196
164	118
175	62
183	183
190	126
196	44
197	115
2	167
27	153
36	45

Figure 1. Rank Order of Accuracy and Diversity

Fig. 1 show two base learner ranking of two UCI data sets. The left part in Fig. 1 is *harbman* and the right part is *glass*. In each part, the first column stands for the decent ranking by accuracy and the second column for diversity. The number in each column stands for the ID of learner. From Fig. 1 we see that both ranking are consistent in accuracy and diversity. The main reason lies in the label punishment in diversity definition.

An equal weighted ranking strategy is adopted to combine accuracy and diversity contribution of each learner. All learners are ranked twice, the first time is by accuracy and the second time is by diversity of both descend order. Two rankings are added together to get the final ranking.

IV. Main Algorithm

In this section we propose the main algorithm DDLEP (Data Dependent Learners Ensemble Pruning) of this paper, and give some detail discussion on applying the proposed algorithm to both single-instance (SI) and multi-instance (MI) learning.

A. DDLEP algorithm

As mentioned above, kernel based semi-supervised learning is adopted to generate base learner before ensemble pruning. Then accuracy and diversity of each learner is calculated according to Eq. (7) and Eq. (8).

Formally, the input of DDLEP is a training set D_{train}, an evaluation set D_{eval} ,a testing set D_{test} and a predefined parameter θ which is the percentage of the selected classifiers in original ensemble.

The original ensemble $C = \{c_1, c_2, ..., c_m\}$ is trained on D_{train} , and take $D_{train} + D_{eval}$ and $D_{train} + D_{eval} + D_{test}$ as point cloud for semi-supervised and transductive learning respectively. We also note that in transductive setting, the point cloud is made up of $D_{train} + D_{test}$ which would be very large in size. Thus a randomly selected subset is generated to substitute the original point cloud. It is empirically effective as shown in the experiment section. The algorithm is summarized in Fig. 2.

Algorithm 1 DDLEP

Input:

Training set D_{train}, point cloud P

testing set D_{test}, pruning rate θ

Output:

Subensemble $C_{pruning}$

1: for $k = 1$ to m
2: T = bootstrap(D_{train})
3: K_{train} = PointCloudKernel(T, T, P)
4: $c_k = svmtrain(K_{train}, T)$
5: Add c_k to C
6: End
7: $Acc = CalAccuracy(C, D_{train} + D_{eval})$
8: $Div = CalDiversity(C, D_{train} + D_{eval})$
9: $Order = sort(Acc + Div, 'descend')$
10: Return top θ learners according to $Order$ as $C_{pruning}$

Figure 2. DDLEP algorithm

Fig. 2 details the DDLEP algorithm. The procedure *PointCloudKernel* calculates the point cloud kernel on point cloud P , which is described in Fig. 3. Note that in SVM training and testing procedure, Gram matrix of associated data is needed. For training, Gram matrix is constructed between training data; for testing, Gram matrix is constructed between training data and testing data. The procedure *svmtrain* is a standard SVM training method with kernel K_{train} and training data T . *CalAccuracy* and *CalDiversity* are two methods for calculation of accuracy and diversity of all learners.

Graph Laplacian regularizer is calculated to measure smoothness. For SI learning, procedure 'GraphLaplacian' constructs KNN adjacency weighted graph and heat kernel [18] is used. Details can be found in [15]. For MI learning, we define a new bag level Graph Laplacian to evaluate the adjacency of the input samples. Details are described in subsection C.

B. On Single-Instance Learning

We show that the proposed algorithm is suitable for SI learning naturally. Note that M constructed by KNN adjacency graph really captures underlying manifold of the data set. There are two important parameters controlling the quality of M . The first one is p that controls degrees of manifold and the second is the parameter of original kernel K . In our work, radius basis function (RBF) is used with a parameter σ controlling kernel width. The meaning of σ is how far points in a manifold can be away from each other. Such parameters are data set dependant and a proper selection can reach relatively good result. We use a grid search procedure to obtain the best result in a manual defined scope.

Algorithm 2 Point Cloud Kernel Calculation

Input: Point Cloud P, Data set D_1, Data set D_2
Original Kernel Gram Matrix K, parameter h

Output: Point Cloud Kernel K_p

1: $L = GraphLaplacian(P)$
2: $M = L^h$
3: $DM_1 = D_1^T \cdot P \qquad DM_2 = (D_2^T \cdot P)^T$
4: $K_p = K - DM_1 \cdot (I + M \cdot K)^{-1} \cdot M \cdot DM_2$
5: Return K_p

Figure 3. Point Cloud Kernel Calculation

C. On Multi-Instance Learning

The proposed method is also suitable for multi-instance learning [19] with introduction of MI kernel [20] and smoothness penalty of bags. There are two key points to apply the proposed method to multi-instance setting. The first is MI kernel that is studied in many literature [20, 21]. In this paper, node kernel [20] and miGraph [21] are used as kernels in the main algorithm. We briefly describe the two kernels.

Suppose we have a multi-instance data set $D = \{(X_1, y_1), ..., (X_n, y_n)\}$, where $X_i = \{x_{i1}, .., x_{im_i}\}$, $y_i \in \{-1, +1\}$ and x_{ij} stands for the jth instance in bag i .

1) Node Kernel

Let k be a single-instance kernel, define node kernel as following:

$$k_{node}(X_i, X_j) = \sum_{p=1}^{m_i} \sum_{q=1}^{m_j} k(x_{ip}, x_{jq}) \qquad (9)$$

Node kernel provides a mapping between bags and some feature space.

2) miGraph

[21] proposed an effective but simplified MI kernel. It regarded a bag as a partial connected graph (KNN graph or ε –ball graph). Both nodes and edges that represent relationship between nodes are considered in the kernel function.

To construct miGraph kernel, define W^i as adjacency matrix for bag X_i:

$$W_{ab}^i = \begin{cases} 1 \text{ iff } \|x_{ia} - x_{ib}\| \le \delta \\ 0 \text{ otherwise} \end{cases} \tag{10}$$

(10) indicates that we only care whether there is relationship between nodes, but not how strong the relationship is. miGraph k_g is defined as (11):

$$k_g(X_i, X_j) = \frac{\sum_{a=1}^{m_i}\sum_{b=1}^{m_j} W_{ia} \cdot W_{jb} \cdot k_{node}(x_{ia}, x_{jb})}{\sum_{a=1}^{m_i} W_{ia} \cdot \sum_{b=1}^{m_j} W_{jb}} \tag{11}$$

where $W_{ia} = 1/\sum_{p=1}^{m_i} W_{ap}^i$, $W_{jb} = 1/\sum_{p=1}^{m_j} W_{bp}^j$.

[22] gave a theoretical analysis of the effectiveness of the kernel defined in (11).

The second point to be discussed is the Graph Laplacian of bags. Motivated by single instance semi-supervised kernel [13], we define semi-supervised multi-instance kernel to make use of labeled and unlabeled bags.

Following the idea of Citation-KNN [23], we define the smallest Euclidean distance between instances from two bags as bag instance, denoted as D_{Min}. Formally, we have:

$$D_{Min}(X_i, X_j) = \min \left\| x_i - x_j \right\|^2 \quad x_i \in X_i, x_j \in X_j \tag{12}$$

Since the calculation of Graph Laplacian and kernel is based on certain distance definition, it is acceptable to plug D_{Min} into Algorithm 2 to get a point cloud kernel for MI problem.

D. Time Complexity

An approximate evaluation of time complexity of the proposed method is presented here. Let m be the size of learners, n_1 be the size of training data, n_2 be the size of evaluation data and n_3 be the size of testing data.

$$T_{K\_train} = O(n_1^2 + (n_2 + n_3)^2) \tag{13}$$

$$T_{DIV} = O(n_1^2) \tag{14}$$

$$T_{pruning} = O(m) \tag{15}$$

T_{K_train} denotes time of point cloud kernel calculation for training. T_{DIV} stands for time of calculation of diversity of each member learner and $T_{pruning}$ for pruning time. T_{SVM} is the standard SVM training time complexity. Then the overall time complexity for the proposed method is:

$$T = m \cdot (T_{SVM} + T_{K\_train} + T_{DIV}) + T_{pruning} \tag{16}$$

IV. EVALUATION

A. Experiment

Experimental data in this paper were chosen from the UCI machine learning data repository [23] and Corel Image Gallery Magic. We choose 10 data sets from UCI repository to evaluate the performance of DDLEP on the single-instance learning setting, and 2 data sets from Corel Image Gallery Magic on the multi-instance learning setting. We also test the proposed algorithm on the famous multi-instance learning data set Musk1 and Musk2 [19]. We verify the efficiency of the proposed method on both single-instance learning and multi-instance learning.

Each data set is divided into 3 parts: training, evaluation and testing with a size ratio 3:3:4. Each data set is randomly partitioned 10 times and 10 independent trials of experiments are launched on each partition. Therefore a total of 100 trials of experiments were conducted on each data set. We take the average of all experimental results as the final result. TABLE I summarizes the characteristics of the data sets.

TABLE I. SINGLE INSTANCE DATA SET PROPERTIES

Data Set	Attributes	Classes	Size
balance-scale	4	3	625
breast-w	9	2	683
ecoli	8	8	336
glass	9	6	214
haberman	3	2	306
ionosphere	34	2	351
segment	19	7	2310
transfusion	4	2	748
wine	13	3	178
yeast	8	10	1484

TABLE II. MULTI INSTANCE DATA SET PROPERTIES

Data Set	Pos. Bags	Neg. Bags	Attributes
Musk1	47	45	476
Musk2	39	63	6598
Fox	100	100	9
Tiger	100	100	9

We compare the proposed method DDLEP to two peer methods: semi-supervised kernel learning method and EPIC [6]. The former is a semi-supervised learning without ensemble, while the latter is a supervised ensemble learning method. For MI data set *Fox* and *Tiger*, method described in [24] is used to covert each image into a bag with 9-attribute regions as instances.

We demonstrate the overall prediction accuracy on test data for all data sets of both SI and MI settings as TABLE III and TABLE IV. For DDLEP and EPIC, top 15% of the learners are used to construct subensemble.

TABLE III. ERROR RATE OF SI SETTING

Data Set	DDLEP	EPIC	SSLK
balance-scale	7.4%±8.4e-5	7.4%±8.4e-5	**5.7%± 5.7e-3**
breast-w	5%±2e-4	4.7%±5.7e-5	**12.6%±2e-2**
ecoli	**4.9%±2e-4**	4.7%±5.7e-5	5%±2e-4
glass	**8.5%±9e-5**	10.4%±5e-4	10%±5e-4
haberman	**11.5%±4.2e-3**	12%±4.4e-3	15.6%±9e-4
ionosphere	**25.4%±5.3e-4**	26.1%±7e-4	28.6%±2.6e-3
segment	**6.6%±1e-4**	6.7%±1.4e-4	6.7%±1.3e-4
spambase	**3.7%±4.4e-5**	3.9±3.9e-5	4%±4e-5
wine	5.3%±4.7e-5	5.5%±2.2e-4	**5.2%±2.8e-4**
yeast	**8.4%±4.5e-5**	8.6%±6.8e-5	8.6%±4.9e-5

TABLE IV. ERROR RATE OF MI SETTING

Data Set	MI-Kernel	EPIC	DDLEP
Musk1	8.2%±2.5e-4	8.2%± 2.5e-4	**7.4%± 8.38e-5**
Musk2	5%±2e-4	4.7%±5.7e-5	**4.9%±2e-4**
Tiger	10%±5e-4	10.4%±5e-4	**8.5%±9e-5**
Fox	15.6%±9e-4	12%±4.4e-3	**11.5%±4.2e-3**

We use the proposed algorithm into MI learning by defining MI point cloud kernels, making use of SI kernel learner. However, the successfulness of this method is highly relied on the definition of point cloud norm, and such norm should be consistent in manifold. Another two point cloud norms are also proposed here.

Bag Center Point Norm is defined as Euclidean distance between two bags central points, as Eq. (17):

$$D_{Center}(X_i, X_j) = \left\| \bar{x}_i - \bar{x}_j \right\|^2 \tag{17}$$

Eq. (17) assumes that the central point of a bag represents this bag. Though it is not consistent with the original MI assumption presented in [19], it works well in a few cases.

Set Kernel Norm is defined as node kernel between two bags, as Eq. (18):

$$D_{Set}(X_i, X_j) = \frac{\sum_{m,n} \left\| x_{im} - x_{jn} \right\|^2}{\left| X_i \right| \cdot \left| X_j \right|} \tag{18}$$

We adopt a set kernel to evaluate the bags smoothness. The motivation of this norm is from MI kernel mapping. It maps each bag to a feature space with vectorial representation. The traditional Graph Laplacian is meanful in such space. It is feasible to plug the norm defined in Eq. (18) in to the point cloud kernel calculation.

We also use very few training samples (5% of the whole data set) to construct base learners and test the proposed algorithm. In Fig. 4-9, we launch experiments on UCI data sets *breast-w*, *balance-scale*, *ecoli*, *wine*, *ionosphere* and *glass*.

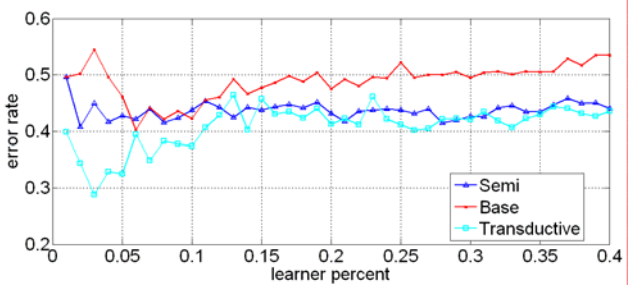

Figure 4. *breast-w* prediction result

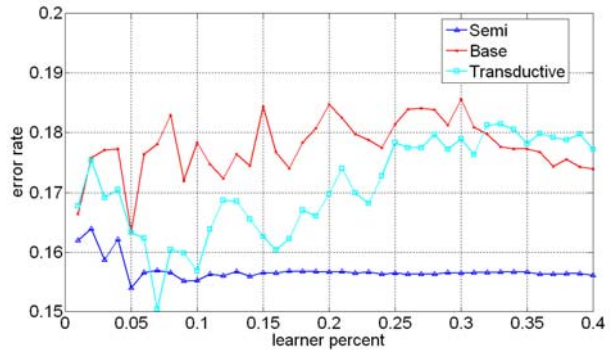

Figure 5. *balance-scale* prediction result

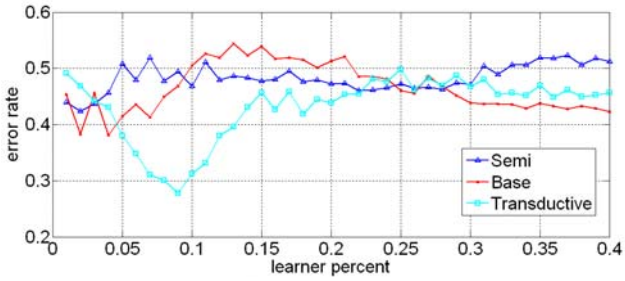

Figure 6. *ecoil* prediction result

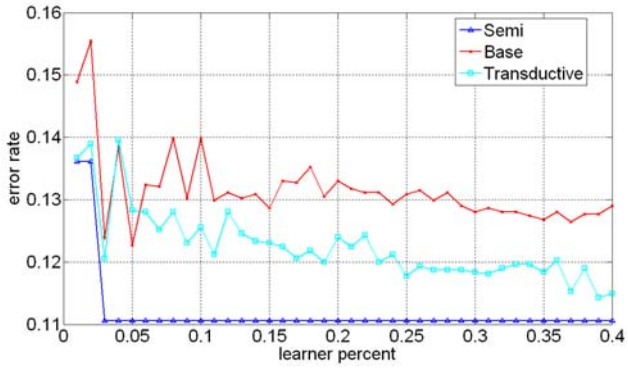

Figure 7. *wine* prediction result

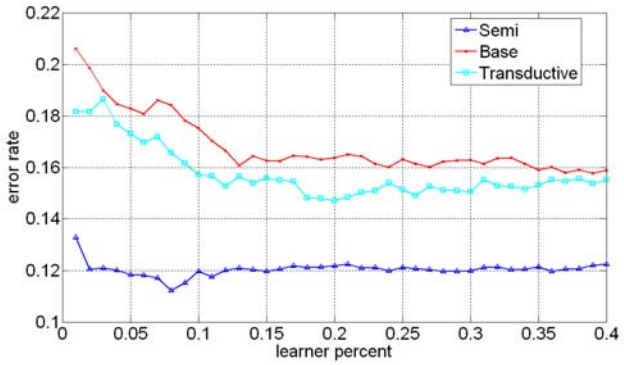

Figure 8. *ionosphere* prediction result

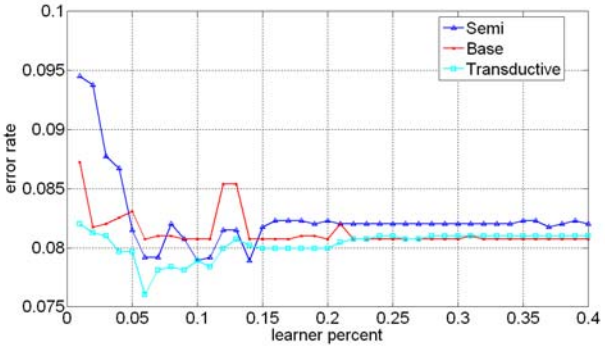

Figure 9. *glass* prediction result

In Fig. 4-9, *Base* stands for error rate of ensemble of supervised kernel learners; *Semi* stands for error rate of ensemble of the proposed data dependant kernel learner and *Transductive* also stands for the proposed method but with a point cloud construction of both evaluation and testing data set. We bootstrap the training set and then train 200 learners for each case. *Semi* and *Transductive* cases use different point clouds. We can conclude that in many cases the *Transductive* setting outperforms the other two. And ensemble of data dependant learner yields smoother change of classification error rate, especially when ensemble size is small. Our remark on this point is that the point cloud improves generalization ability of individual learners.

However, there are also cases that *Transductive* setting yields relatively large error rate compared to *Semi* setting. We regard such phenomenon as point cloud inconsistence. From previous semi-supervised learning study, it is well known that semi-supervised learning works only some conditions are satisfied. More unlabeled data may not really help improve the target learner. In fact we don't have any knowledge about the data to be predicted. If we pull all seen data into the training procedure, side effect would appear because the point cloud includes conflicted data sets.

B. Discussion

We present some analysis of the effectiveness of the proposed method. As mentioned before, the proposed method tends to ensemble learners with better generalization ability. We put this analysis in a regularization framework, and propose a theorem.

Theorem 1: Given two sets of binary classifiers $E_1 = \{c_1, c_2, ..., c_{ne_1}\}$ and $E_2 = \{c_1, c_2, ..., c_{ne_2}\}$, We define the regularization penalty S of individual classifier c as $S_c = \dfrac{\sum_{i,j=1}^{N} (f_c(x_i) - f_c(x_j))^2 W_{ij}}{N^2}$.

If $S_c \le S_d$ holds for each c in E_1 and d in E_2, then $S_{c_{ens\_E1}} \le S_{c_{ens\_E2}}$ holds, where

$$c_{ens\_E_i}(x) = sign\left(\frac{\sum_{k=1}^{nei} c_k(x)}{ne_i}\right), i = 1,2, c_k \in E_i$$

Theorem 1 tells us that the generalization ability measured by S of individual learners will affect that of ensembles by majority voting in a bi-classification problem. We omit the rigorous proof due to the page limit. Another thing we should point out here is that for supervised learning, the training accuracy would be significantly reflected by bootstrapping if the size of training set is too small. However, the selective ensemble would greatly reduce the side effect of bootstrapping. We observe it also appears for semi-supervised learning. There may be some inner relationship between these two famous frameworks.

V. Conclusion

We have introduced ensemble pruning of data dependant learners for the improvement of generalization ability. The key idea is to train individual learners in a semi-supervised manner and prune ensembles with proper selected metrics. Point cloud kernel is adopted to incorporate unlabeled data, which greatly improves generalization ability of individual learners. And accuracy / diversity ranking is used for pruning. The proposed method is naturally suitable for both SI and MI framework, with special defined kernel and smoothness penalty term. Using 10 UCI and 4 MI data sets, we could show that the DDLEP can significantly improve the performance of ensemble learning compared to some famous methods.

Further work includes incorporating of other semi-supervised learning methods into ensemble learning, finding data set specific ranking metrics and theoretical analysis of inner relationship between semi-supervised learning and ensemble learning.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (61033010, 60673132，U0935002), Research Foundation of National Science and Technology Plan Project (2008ZX10005-013), Research Foundation of Science and Technology Plan Project in Guangdong Province (2009A080207005, 2009B090300450, 2010A040303004), GuangDong NSF (07117421, 8351009001000002), GDUT Creative Experiment Project (2010037).

References

[1] A. Chandra, X. Yao, Evolving hybrid ensembles of learning machines for better generalisation. Neurocomput. 69, 7-9, pp. 686-700, 2006.G.

[2] Y. Zhang, S. Burer, W. N. Street, Ensemble pruning via semi-definite programming. JMLR. 7, pp. 1315-1338, 2006.

[3] Z.-H. Zhou, J. Wu, W. Tang, Ensembling neural networks: many could be better than all, Artificial Intelligence, 137, pp. 239–263, 2002.

[4] D.D. Margineantu and T.G. Dietterich, Pruning adaptive boosting, In Proceedings of the 14th ICML, pp. 211-218, 1997.
[5] G. M. Muñoz and A. Suárez, Pruning in ordered bagging ensembles, In Proceedings of the 23rd ICML, pp. 609-616, 2006.
[6] Z. Lu, X. Wu, X. Zhu, J. Bongard, Ensemble pruning via individual contribution ordering, In Proceedings of the 16th ACM SIGKDD (KDD '10), pp. 871-880, 2010.
[7] H. Chen, X. Yao. Regularized Negative Correlation Learning for Neural Network Ensembles. IEEE Transactions on Neural Networks, Vol. 20, No. 12, pp. 1962-1979, 2009.
[8] R. Caruana and A. Niculescu-Mizil, An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd ICML, pp. 161-168, 2006.
[9] G. Brown, J. L. Wyatt, P. Tiňo, Managing diversity in regression ensembles, JMLR, 6, pp. 1621-1650, 2005.
[10] Zhi-Hua Zhou. Research Article: When Semi-Supervised Learning Meets Ensemble Learning. Front. Electr. Electron. Eng. China, 2010, 5(3).
[11] Min-Ling Zhang, Zhi-Hua Zhou. Exploiting Unlabeled Data to Enhance Ensemble Diversity. ICDM 2010.
[12] T. G. Dietterich. Ensemble methods in machine learning. In Proceedings of the 1st IWMCS, pp. 1-15, 2000.
[13] M. Belkin, P. Niyogi, V. Sindhwani, Manifold regularization: A geometric framework for learning from labeled and unlabeled examples, JMLR, 7, pp. 2399-2434, 2006.
[14] N. Li and Z.-H. Zhou. Selective ensemble under regularization framework. In Proceedings of the 8th IWMCS (MCS'09), Reykjavik, Iceland, LNCS 5519, 2009, pp.293-303.
[15] V. Sindhwani, P. Niyogi, M. Belkin. Beyond the point cloud:from transductive to semi-supervised learning. In: Proceed-ings of the 22nd ICML. Bonn, Germany: MIT Press, 2005, pp.824-831.
[16] D. Rosenberg, V. Sindhwani, P. Bartlett, P. Niyogi. A Kernel for Semi-supervised Learning with Multi-view Point Cloud Regularization. IEEE Signal Processing Magazine, 2009.
[17] M. Belkin, P. Niyogi, V. Sindhwani . On Manifold Regularization. Artificial Intelligence and Statistics (AISTATS), 2005.
[18] X. Zhu, Z. Ghahramani, and J. Lafferty. Semisupervised learning using gaussian fields and harmonic functions. In Proceddings of the 20th ICML, pp. 912-919, 2003.
[19] Dietterich, T. G., Lathrop, R. H., & Lozano-P´erez, T. Solving the multiple-instance problem with axis-parallel rectangles. Artif. Intell., 89, 31-71, 1997.
[20] T. Gartner, A.P. Flach, A. Kowalczyk, A. Smola. Multi-instance kernels. In Proceedings of the 19th ICML. 179–186, 2002.
[21] Zhi-Hua Zhou, Yu-Yin Sun, Yu-Feng Li. Multi-instance learning by treating instances as non-I.I.D. samples. In Proceedings of the 26th ICML. Montreal, Quebec, Canada, 2009: 1249-1256.
[22] Partha Niyogi. Manifold regularization and semi-supervised learning: Some theoretical analyses. Technical report, Department of Computer Science, University of Chicago, 2008.
[23] UCI Repository: http://archive.ics.uci.edu/ml/
[24] Chen, Y., & Wang, J. Z. Image categorization by learning and reasoning with regions. JMLR, 5, 913–939, 2004.

Gang Zhang, was born in 1979, Guangzhou, received the Master's Degree in computer software and theory in SUN YAT-SEN University (Guangzhou, China) in 2005. Now he is a Ph.D candidate in SUN YAT-SEN University, majored in data mining and machine learning.

He is a lecturer of Faculty of Automation, GuangDong University of Technology (Guangzhou, China), teaching Java Programming, C++ Programming, Computer Network, Software Technology Foundation and Web Site Design and Development. His current research interests include structural data mining, semi-supervised learning, multi-instance learning, manifold learning, ensemble learning and metric learning. He has published several papers in international conferences and journals on his research fields, some of which are indexed by SCI, EI and ISTP.

Yin Jian, was born in 1967. He has been professor of SUN YAT-SEN University since 1994. His main research interests are data mining, machine learning, advanced database, data warehouse and artificial intelligence.

Xiaomin He, was born in 1961, received the Master's Degree from South China University of Technology (Guangzhou, China).

She is an associate professor of Faculty of Automation, GuangDong University of Technology (Guangzhou, China), teaching Ensemble Language, Configuration Software, and Principles of Computer Organization. She is professional in research of Control Science and Control Engineering. Her current research interests include SCM, e-learning and operation system. She has published several papers in her research fields.

Lianglun Cheng, was born in 1965. Now He is professor of Guangdong University of Technology. His main research interests are automation equipment, intelligent computer network, sensor network, RFID network.