SVM+ Regression and Multi-Task Learning

Feng Cai and Vladimir Cherkassky, Fellow IEEE

***Abstract*— Exploiting additional information to improve traditional inductive learning is an active research area in machine learning. In many supervised-learning applications, training data can be naturally separated into several groups, and incorporating this group information into learning may improve generalization. Recently, Vapnik [9] proposed general approach to formalizing such problems, known as Learning With Structured Data (LWSD) and its SVM-based optimization formulation called SVM+. Liang and Cherkassky [5,6] showed empirical validation of SVM+ for classification, and its connections to Multi-Task Learning (MTL) approaches in machine learning. This paper builds upon this recent work [5,6,9] and describes a new methodology for *regression problems*, combining Vapnik's SVM+ regression [9] and the MTL classification setting [6], for regression problems. We also show empirical comparisons between standard SVM regression, SVM+, and proposed SVM+MTL regression method. Practical implementation of new learning technologies, such as SVM+, is often hindered by their complexity, i.e. large number of tuning parameters (vs standard inductive SVM regression). To this end, we provide a practical scheme for model selection that combines analytic selection of parameters for SVM regression [3] and resampling-based methods for selecting model parameters specific to SVM+ and SVM+MTL.**

I. INTRODUCTION

There is a growing need for development of powerful and robust methods for estimating predictive models from data. Most supervised learning methods developed in statistical learning, pattern recognition and machine learning are based on standard inductive formulation of the learning problem [3,4,8]. Many challenging new applications with sparse high-dimensional data may require new (non-standard) learning formulations [3,9].

In this paper, we consider supervised learning applications where the training data includes additional (group) information about training samples. Examples include: (1) handwritten digit recognition where training examples are provided by several persons, (2) medical diagnosis where predictive (diagnostic) model, say for lung cancer, is estimated using a training data set of male and female patients, etc. Incorporating this additional information has lead to approaches known as Multi-Task Learning [1,2,6,10] and, more recently, to Learning with Structured Data (aka SVM+) [9], as discussed next.

Suppose that training data can be represented as a union of t related groups, i.e. each group $r \in [1,2,..,t]$ contains n_r samples independently and identically generated from a distribution P_r on $\mathbf{x} \times y$. Therefore, available data is a union of t>1 groups:

$$\{\{\mathbf{X}_r, \mathbf{Y}_r\}, r = 1,...,t\}, \{\mathbf{X}_r, \mathbf{Y}_r\} = \{\{\mathbf{x}_{r_1}, y_{r_1}\},...,\{\mathbf{x}_{r_{nr}}, y_{r_{nr}}\}\}$$

and can be thought as samples identically and independently generated from unknown distribution $P(\mathbf{x}, y) = \{P_r(\mathbf{x}, y), if \{\mathbf{x}, y\} \in \{\mathbf{X}_r, \mathbf{Y}_r\}\}$.

If the group labels of future test samples are not given, the problem is "**L**earning **W**ith **S**tructured **D**ata (LWSD)" formulation [9]. In this formulation, the goal is to find one best mapping function f such that the expected loss

$$R_{LWSD}(w) = \int L(f(\mathbf{x}, w), y) P(\mathbf{x}, y) d\mathbf{x} dy$$

is minimized. Here L is the loss function. Note that even though the expected loss is in the same form as in the supervised learning setting, the difference is that in supervised learning setting P is unknown, while in LWSD it is known that P is a union of t sub-distributions.

On the other hand, if the group labels of future test samples are given, the problem is **M**ulti-**T**ask **L**earning (MTL) problem [1,2,6,8]. The goal in multi-task learning is to find t related mapping functions $\{f_1, f_2,..., f_t\}$ so that the sum of expected losses for each task

$$R_{MTL}(w) = \sum_{r=1}^{t} (\int L(f_r(\mathbf{x}, w), y) P_r(\mathbf{x}, y) d\mathbf{x} dy)$$

is minimized.

From the application point of view, different learning settings (standard inductive learning, multi-task learning and learning with structured data) handle training and test data in different ways, as illustrated in Fig.1. That is, standard inductive setting does not use (ignores) group information in the training data; MTL setting estimates separate predictive models (for each task), and LWSD estimates a single model that utilizes group information in the training data. Note that under LWSD test inputs do not have group information, whereas under MTL test inputs are assumed to have group labels.

Recently, Vapnik [9] proposed general approach for solving such problems, known as **L**earning **W**ith **S**tructured **D**ata (LWSD) and its SVM-based optimization formulation called SVM+. Liang and Cherkassky [5,6] showed empirical validation of SVM+ for *classification*, and showed its connection to Multi-Task Learning (MTL) classifiers in machine learning [1,2,6,10]. This paper builds upon this recent work and describes a new methodology for *regression problems*, combining Vapnik's SVM+ approach [9] and the MTL classification scheme [6]. We also show empirical

Feng Cai is with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis MN 55455, USA. (e-mail: caixx043@ umn.edu).

Vladimir Cherkassky is with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455, USA. (e-mail: cherk001@umn.edu) .

comparisons between standard SVM regression, SVM+, and the proposed SVM+MTL regression method.

The paper is organized as follows. Section II gives a brief overview of standard SVM regression and Vapnik's SVM+ regression formulation, followed by the proposed optimization formulation for SVM+MTL regression. Section III describes different approaches to handling group information in the training data, including standard (single) SVM regression, multiple SVM regression models (trained independently), Vapnik's SVM+ regression, and proposed SVM+MTL regression. Section IV shows empirical comparisons between these approaches. Conclusions and discussion are presented in Section V.

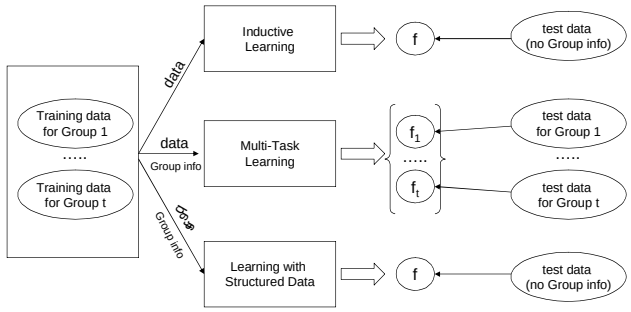

Figure 1: Inductive learning, Multi-task learning and Learning with structured data handle training and test data in different ways.

II. SVM+ and SVM+MTL Regression Formulation

This section describes first standard (linear) SVM regression [8], using ε-insensitive loss function, in order to introduce basic notations and terminology. Then we describe Vapnik's SVM+ regression formulation following [9], and finally present the proposed SVM+MTL regression formulation.

A. Standard SVM Regression

Given iid training data $\{\{\mathbf{x}_i, y_i\}\}_{1\le i\le n}$, $\mathbf{x}_i \in R^d$, $y_i \in (-\infty, \infty)$, standard SVM will first map input vectors $\mathbf{x}$ onto the feature space Z ($\mathbf{z} = \phi(\mathbf{x}) \in Z$) and then approximate the regression by a linear function: $f(\mathbf{x}) = (\mathbf{w}, \phi(\mathbf{x})) + b$. To this end, standard SVM solves the following optimization problem:

$$\min_{\mathbf{w},b} \frac{1}{2}(\mathbf{w},\mathbf{w}) + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \qquad \text{(OP1)}$$

Subject to:

$$y_i - (\mathbf{w}, \mathbf{z}_i) - b \le \varepsilon + \xi_i^*, \xi_i^* \ge 0, i = 1,...,n$$

$$(\mathbf{w}, \mathbf{z}_i) + b - y_i \le \varepsilon + \xi_i, \xi_i \ge 0, i = 1,...,n$$

Where $\xi_i, \xi_i^*, i = 1,\ldots, n$ are called slack variables, measuring the deviation from ε-insensitive tube. And C is the regularization parameter.

B. SVM+ Regression

Suppose the training data are the union of $t > 1$ related groups. Let us denote the indices from group r by $T_r = \{i_{n1},\ldots,i_{nr}\}, r = 1,\ldots,t$. Then all training samples can be represented as: $\{\{\mathbf{X}_r, \mathbf{Y}_r\}, r = 1,...,t\}, \{\mathbf{X}_r, \mathbf{Y}_r\} = \{\{\mathbf{x}_{r_1}, y_{r_1}\},...,\{\mathbf{x}_{r_{nr}}, y_{r_{nr}}\}\}$. Similar to standard SVM, SVM+ will map vectors in each group $\mathbf{x}_i, i \in T_r$ simultaneously *into two different Hilbert spaces* Z ($\mathbf{z}_i = \phi_z(\mathbf{x}_i) \in Z$) and Z_r ($\mathbf{z}_i^r = \phi_{z_r}(\mathbf{x}_i) \in Z_r$). To account for the group information, Vapnik [9] defines the slack variables as follows:

$$\xi_i = (\mathbf{w}_r, \mathbf{z}_i) + d_r, i \in T_r, r = 1,...,t$$

$$\xi_i^* = (\mathbf{w}_r^*, \mathbf{z}_i) + d_r^*, i \in T_r, r = 1,...,t.$$

Compared to standard SVM regression, here the slack variables are restricted by the correcting functions, and the correcting functions represent additional information about the data. The goal is to find the regression function in decision space Z,

$$f(\mathbf{x}) = (\mathbf{w}, \phi_Z(\mathbf{x})) + b$$

Note that data of different groups are mapped into the same decision space, and they are all used to construct the regression function. However, there are different correcting functions for different groups. Correcting functions are defined in the correcting space. Different correcting functions can be defined either in the same correcting space or different correcting function spaces. Of course, if data of different groups are mapped to different correcting spaces, the correcting functions for different groups are different. If data of different groups are mapped to the same correcting space, we still can construct different correcting functions for different groups. The importance is that correcting functions are different, not correcting space.

Correcting functions represent a unique way that SVM+ handles group information. Since correcting functions represent slack variables, there are some unique characteristics:

(1) All slack variables are non-negative, so $\xi_r(\mathbf{x}_i) = (\mathbf{x}_i, \mathbf{w}_r) + d_r \ge 0, r = \{1,...,t\}$.
Therefore mapping samples in the correcting space have to lie on one side of the corresponding correcting function. Correcting function also has to pass through some points with slack variables being zero.

(2) Like regression function, correcting function is also chosen from a set of correcting functions, and

$(\mathbf{w}_r, \mathbf{w}_r)$ reflects the capacity of the set of correcting functions; but this term does not have meaning of the size of margin (in classification).

(3) Correcting functions are not used to assign a sample a group membership.

In the case of two groups, mapping of the training data by SVM+ regression is schematically shown in Fig. 2.

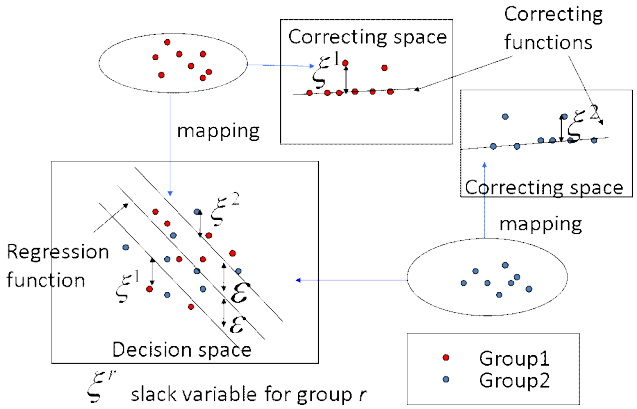

Figure 2: SVM+ maps data simultaneously into decision space and correcting spaces. Regression function is found in decision space. Slack variables are represented by correcting functions which are defined in correcting space.

Formally, SVM+ regression approach estimates regression model $f(\mathbf{x}) = (\mathbf{w}, \phi_z(\mathbf{x})) + b$ by solving the following optimization problem:

$$\min_{\mathbf{w},\mathbf{w}_1,\ldots,\mathbf{w}_r,\mathbf{w}_1^*,\ldots,\mathbf{w}_r^*,b,d_1,\ldots,d_r,d_1^*,\ldots,d_r^*} \frac{1}{2}(\mathbf{w},\mathbf{w}) + \frac{\gamma}{2}\left(\sum_{r=1}^{t}(\mathbf{w}_r,\mathbf{w}_r) + \sum_{r=1}^{t}(\mathbf{w}_r^*,\mathbf{w}_r^*)\right)$$
$$+C\sum_{r=1}^{t}\sum_{i\in T_r}(\xi_i^r + \xi_i^{r*}) \qquad \text{(OP2)}$$

Subject to:

$$y_i - (\mathbf{w},\mathbf{z}_i) - b \le \varepsilon + \xi_i^{r*}, \xi_i^{r*} \ge 0, i \in T_r, r = 1,\ldots,t$$
$$(\mathbf{w},\mathbf{z}_i) + b - y_i \le \varepsilon + \xi_i^r, \xi_i^r \ge 0, i \in T_r, r = 1,\ldots,t$$
$$\xi_i^r = (\mathbf{w}_r, \mathbf{z}_i^r) + d_r, i \in T_r, r = 1,\ldots,t$$
$$\xi_i^{r*} = (\mathbf{w}_r^*, \mathbf{z}_i^r) + d_r^*, i \in T_r, r = 1,\ldots,t$$

Using the dual optimization technique (similar to standard SVM), one can show that $\mathbf{w}$, $\mathbf{w}_r$ and $\mathbf{w}_r^*$ can be expressed in terms of training samples:

$$\mathbf{w} = \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)\mathbf{z}_i,$$

$$\mathbf{w}_r = \frac{1}{\gamma}\sum_{i\in T_r}(\alpha_i + \mu_i - C)\mathbf{z}_i^r,$$

$$\mathbf{w}_r^* = \frac{1}{\gamma}\sum_{i\in T_r}(\alpha_i^* + \mu_i^* - C)\mathbf{z}_i^r,$$

where the coefficients $\alpha_i, \alpha_i^*, \mu_i$ and μ_i^* are solution of the following dual optimization problem:

$$\max_{\alpha,\alpha^*,\mu,\mu^*} -\varepsilon\sum_{i=1}^{n}(\alpha_i^* + \alpha_i) + \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)y_i$$
$$-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)(\mathbf{z}_i,\mathbf{z}_j)$$
$$-\frac{1}{2\gamma}\sum_{r=1}^{t}\sum_{i,j\in T_r}(\alpha_i + \mu_i - C)(\alpha_j + \mu_j - C)(\mathbf{z}_i^r,\mathbf{z}_j^r)$$
$$-\frac{1}{2\gamma}\sum_{r=1}^{t}\sum_{i,j\in T_r}(\alpha_i^* + \mu_i^* - C)(\alpha_j^* + \mu_j^* - C)(\mathbf{z}_i^r,\mathbf{z}_j^r)$$

Subject to:

$$\sum_{i=1}^{n}(\alpha_i^* - \alpha_i) = 0$$
$$\sum_{i\in T_r}(\alpha_i + \mu_i) = C\,|T_r|, r = 1,\ldots,t$$
$$\sum_{i\in T_r}(\alpha_i^* + \mu_i^*) = C\,|T_r|, r = 1,\ldots,t$$
$$\alpha_i \ge 0, \alpha_i^* \ge 0, \mu_i \ge 0, \mu_i^* \ge 0, i = 1,\ldots,n$$

C. SVM+MTL Regression

Similar to SVM+, SVM+MTL also makes good use of group information. However, the goal of this approach is to estimate t regression models (one model per group). So instead of incorporating group information into slack variables, SVM+MTL incorporates group information into estimated regression functions.

We specify the following parameterization for t regression models:

$$f_r(\mathbf{x}) = (\phi_z(\mathbf{x}), \mathbf{w}) + b + (\phi_{z_r}(\mathbf{x}), \mathbf{w}_r) + d_r, r = 1,\ldots,t.$$

Here $(\phi_z(\mathbf{x}), \mathbf{w}) + b$ is the common estimation function while $(\phi_{z_r}(\mathbf{x}), \mathbf{w}_r) + d_r$ is the unique correction function for each group. The proposed method SVM+MTL formulation solves the following optimization problem:

$$\min_{w,w_1,\ldots w_t,b,d_1,\ldots,d_t} \frac{1}{2}(\mathbf{w},\mathbf{w}) + \frac{\gamma}{2}\sum_{r=1}^{t}(\mathbf{w}_r,\mathbf{w}_r)$$
$$+C\sum_{r=1}^{t}\sum_{i\in T_r}(\xi_i^r + \xi_i^{r*}) \qquad \text{(OP3)}$$

Subject to:

$$y_i^r - ((\mathbf{w},\mathbf{z}_i) + b + (\mathbf{w}_r,\mathbf{z}_i^r) + d_r) \le \varepsilon + \xi_i^{r*}, i \in T_r, r = 1,\ldots,t$$
$$((\mathbf{w},\mathbf{z}_i) + b + (\mathbf{w}_r,\mathbf{z}_i^r) + d_r) - y_i^r \le \varepsilon + \xi_i^r, i \in T_r, r = 1,\ldots,t$$
$$\xi_i^r \ge 0, \xi_i^{r*} \ge 0, i \in T_r, r = 1,\ldots,t.$$

The dual form of the above optimization problem is as follows:

$$\max_{\alpha,\alpha^*} -\varepsilon\sum_{i=1}^{n}(\alpha_i^* + \alpha_i) + \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)y_i$$
$$-\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)(\mathbf{z}_i,\mathbf{z}_j)$$

$$-\frac{1}{2\gamma}\sum_{r=1}^{t}\sum_{i,j\in T_r}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)(\mathbf{z}_i^r,\mathbf{z}_j^r)$$

Subject to:

$$\sum_{i\in T_r}(\alpha_i^*-\alpha_i)=0, r=1,\ldots,t$$

$$0\le\alpha_i\le C, 0\le\alpha_i^*\le C, i=1,\ldots,n$$

Based on KKT condition, we can express $\mathbf{w}$, $\mathbf{w}_r$ in terms of training samples:

$$\mathbf{w}=\sum_{i=1}^{n}(\alpha_i^*-\alpha_i)\mathbf{z}_i\ ,$$

$$\mathbf{w}_r=\frac{1}{\gamma}\sum_{i\in T_r}(\alpha_i^*-\alpha_i)\mathbf{z}_i^r\ .$$

Thus,

$$f_r(\mathbf{x})=\sum_{i=1}^{n}(\alpha_i^*-\alpha_i)(\mathbf{z}_i,\phi_z(\mathbf{x}))+b+\frac{1}{\gamma}\sum_{i\in T_r}(\alpha_i^*-\alpha_i)(\mathbf{z}_i^r,\phi_{z^r}(\mathbf{x}))+d_r, r=1,\ldots,t$$

III. APPROACHES for MODELING HETEROGENEOUS DATA

"Learning with structured data" formulation and multi-task learning formulation are similar in the sense that they both try to exploit the group information hidden in the data. Such 'group information' is common in many applications with *heterogeneous* data. For example, in medical diagnostic applications, patients' data may include clinical, genetic and demographic input features. Then certain inputs, for example demographic features, such as Gender or Age, can be used to separate labeled training data into several groups. Proper selection of such a *group variable* is specific to each application at hand (see examples in Section 4).

Assuming that available training data can be partitioned (in a meaningful way) into several groups, we can identify several learning approaches for utilizing the group information, as described next:

- *Single SVM* inductive model which estimates standard SVM regression by pooling together training samples from different groups (i.e. group information is ignored);
- *multiple SVM* approach where a separate SVM regression model is estimated for each group (independently);
- *SVM+* approach where a single regression model, utilizing available group information, is estimated from all data;
- *SVM+MTL* implementing multi-task learning, which estimates several related regression models.

Various approaches for incorporating group data into learning process are shown in Fig. 3.

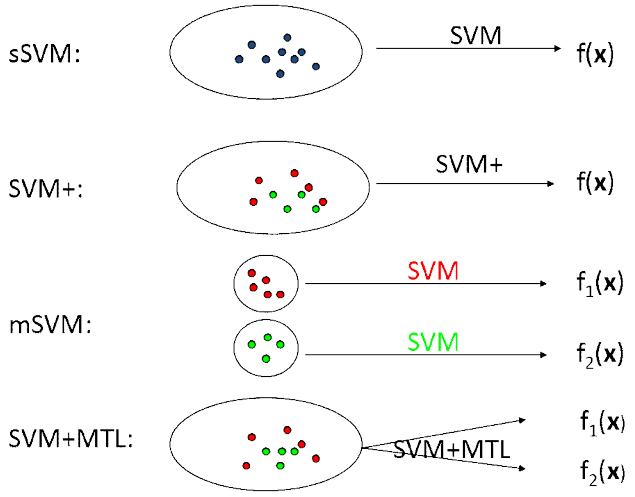

Figure 3: Different ways of using group information in learning: (a) sSVM ~ Single SVM regression, (b) SVM+ regression, (c) mSVM ~ multiple (independent) SVMs, and (d) SVM+MTL ~ proposed SVM+ Multi-Task Learning.

In this paper, we use SVM as an underlying technology for implementing different approaches utilizing group information. However, arguably one can employ other learning techniques, such as MLP networks or AdaBoost, for implementing standard inductive learning, Multi-Task Learning etc. Theoretically, one can expect more sophisticated modeling approaches (utilizing the group information), i.e., SVM+ and SVM+MTL, to yield better generalization than single inductive SVM and multiple (independent) SVM's, respectively. In practice, the trade-off is not so clear, because more advanced approaches (SVM+ and SVM+MTL) have more tunable parameters (than standard SVM), and their potential advantages can be easily offset by more complex model selection.

Next we consider tunable parameters for various learning approaches:

- Single SVM regression: parameters C, ε (width of insensitive zone) and σ (if RBF kernel is used);
- Multiple SVMs: parameters C, ε and σ (if RBF kernel is used) for each task;
- SVM+ regression, where same kernel as in standard SVM is used for the common space, and RBF kernel is used for correcting space, requires following parameters C, ε and σ_{common} (as in standard SVM), γ and $\sigma_{correction}$ (RBF width);
- SVM+MTL regression: requires following parameters C, ε and σ_{common} (as in standard SVM), γ and $\sigma_{correction}$ (RBF width parameter).

Clearly, application of SVM+ and SVM+MTL regression requires practical strategies for tuning parameters of these

methods for a given data set. Next we discuss such a strategy that will be used in empirical comparisons presented later in Section 4. For standard SVM regression, we use analytic prescription for parameters C following [3] and resampling for ε and σ. The same approach is used for multiple SVMs, where parameters C, ε and σ are selected for each group (task). For SVM+ and SVM+MTL, parameters are tuned in a two-stage manner: first set parameters parameters C, ε and σ_{common} same as for standard SVM regression, and second, select parameters γ and $\sigma_{correction}$ via resampling. Empirical comparisons presented next use 5-fold cross-validation for estimating γ and σ.

IV. EMPIRICAL COMPARISONS

This section describes empirical comparisons of various modeling approaches such as single SVM (sSVM), multiple SVM (mSVM), SVM+ and SVM+MTL. All comparisons for synthetic data use linear parameterization for sSVM and mSVM, and RBF(Gaussian) kernel is used for real data. The common estimation space for SVM+ and SVM+MTL use linear kernel for synthetic data and RBF kernel for real data while the unique correction space use RBF kernel. All comparisons use the following experimental procedure:

(a) Select a group variable (from a list of input variables).

(b) Partition available data into several groups (tasks) corresponding to different values (or range of values) of group variable. Each group should be roughly of similar size.

(c) Within each group, order data samples by increasing value of the group variable.

(d) For estimating prediction error of a particular method, use 5-fold cross-validation, so that 80% of data samples are used for training and 20% of the data are used as test data. Note that conditions (b) and (c) ensure that each fold has approximately equal number of samples from all groups (tasks).

(e) For each training fold, parameter tuning (model selection) for different methods is performed as specified in Section 3. That is, parameter C is estimated first following [3], and then parameters ε, γ and σ are tuned via resampling within the training fold.

For each modeling method, we present the predicted mean squared error (MSE) for each of the five folds, as well as the mean and standard deviation of the MSEs.

A. Boston Housing Dataset

This dataset is available at UCI machine learning repository. It includes 14 variables (13 continuous and 1 Boolean) and 506 instances. The goal is to estimate the median value of owner-occupied homes in $1000's from 13 attributes. We present two different comparisons for this dataset. *First,* variable 'RAD' is selected to separate data into 3 groups: group 1($RAD < 5$, 192 instances), group 2($5 \le RAD < 7.5$, 158 instances) and group 3($RAD \ge 7.5$, 156 instances). *Second,* we separate data into 3 groups by another variable 'DIS': group 1($DIS < 2.5$, 188 instances), group 2($2.5 \le DIS < 4.5$, 163 instances) and group 3($DIS \ge 4.5$, 155 instances). Therefore, sSVM, mSVM, SVM+ and SVM+MTL all use 13 attributes for prediction. Possible choices of parameters for SVM+ and SVM+MTL are: $\gamma = [10\ 1\ 0.1\ 0.01\ 0.001]$, $\sigma = [0.25\ 0.5\ 1\ 3\ 4]$. Results are shown in Table 1 and Table 2.

B. Auto MPG Dataset

This is another dataset from UCI machine learning repository. There are 398 instances, each of which has 8 input variables and 1 output variable (mpg). The input variable 'horsepower' has 6 missing values. After removing 6 samples with missing values, we are left with 392 samples used for modeling. The goal is to estimate the real-valued output 'mpg' for each car using 7 input variables (the input variable 'car name' is not used for modeling). We choose variable 'cylinders' to separate the data into 3 groups: group1(cylinders< 6, with 206 instances), group2(cylinders=6, with 83 instances), group3(cylinders>6, with 103 instances). Possible choices of parameters for SVM+ and SVM+MTL are: $\gamma = [10\ 1\ 0.1\ 0.01\ 0.001]$, $\sigma = [0.25\ 0.5\ 1\ 3\ 4]$. Modeling results are shown in Table 3. In both cases, the SVM+MTL method shows some improvement in MSE prediction error, over all other methods.

C. Synthetic Dataset

This data set was artificially generated, in order to illustrate the effect of training set size and the noise level, on relative performance of different learning methods.
Synthetic data set was generated as follows:

(1) Let number of input features be $d = 30$, and number of tasks(groups) be $t = 3$.

(2) Generate $\mathbf{x} \in R^{30}$ with each component $x_i \sim uniform(0,1), i = 1,...,30$.

(3) The coefficient vectors of three (linear) regression tasks are specified as:

$$\beta_1 = [1,1,2,3,3,1,1,1,1,0,2,0,2,2,0,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0],$$
$$\beta_2 = [1,1,2,3,3,1,1,1,1,0,2,0,2,2,0,0,0,0,2,0,2,0,0,0,0,0,0,0,0,0],$$
$$\beta_3 = [1,1,2,3,3,1,1,1,0,1,0,0,3,0,0,2,0,2,0,2,0,0,0,0,0,0,0,0,0,0].$$

(4) For each task and each data vector, $y = \beta_i \mathbf{x} + n$, where n is Gaussian noise with $\sigma_{noise} = 1$.

For each task, we generate N samples for training, N samples for validation, and 1000 samples for testing, and repeat the experiment 5 times. We use the same four methods described in Section III. We follow the ways of tuning parameters in Section III except that γ and σ for SVM+ and SVM+MTL are now tuned using validation data. Possible choices of parameters for SVM+ and SVM+MTL are $\gamma = [100\ 10\ 1\ 0.1\ 0.01]$, $\sigma = [0.25\ 0.5\ 1\ 4\ 6\ 8]$.

Comparison results, for different values of N (number of training samples per task), are shown in Table 4 and Table 5. Note that SVM+MTL shows superior prediction accuracy vs other learning approaches.

In the next experiment, we reduced the input dimensionality ($d = 20$) and noise level($\sigma_{noise} = 0.5$), and increased the sample size used for training and validation ($N = 100$). Also, new coefficient vectors used to generate target functions for three tasks are specified as follows:

$\beta_1 = [1,1,2,3,3,1,1,1,1,0,4,0,4,4,0,0,0,0,0,0]$,

$\beta_2 = [1,1,2,3,3,1,1,1,0,4,0,2,4,0,0,0,0,0,0,0]$,

$\beta_3 = [1,1,2,3,3,1,1,0,1,0,0,4,0,0,4,0,0,0,0,0]$.

Comparison results shown in Table 6 indicate that SVM+MTL underperformed mSVM.

Careful examination of results in Tables 4-6 makes it possible to relate performance of different learning approaches to statistical characteristics of synthetic data sets, as discussed below:

- Synthetic data set 1 (see Table 4) is small and very noisy, so one can expect that methods emphasizing utilization of group information (such as SVM+ and SVM+MTL) yield better prediction performance;
- Synthetic data set 2 (see Table 5) has more training samples (N=50 vs 20 for set 1). Therefore, separating training data into several independent groups as in mSVM approach can now be beneficial, and this accounts for relative improvement in the prediction accuracy of mSVM;
- Synthetic data set 3 (see Table 6) has largest number of training samples, and very low noise level. Hence, we can expect that independent estimation of individual regression models, as in mSVM, would yield the best prediction accuracy.

Table 1 Prediction MSE for Boston housing dataset (group variable: RAD)

Folds	1	2	3	4	5	Mean(st.dev)
sSVM	8.9	26.1	8.5	5.9	10.9	12.1(8.0)
mSVM	12.1	27.2	10.4	6.2	15.1	14.2(7.9)
SVM+	8.9	23.5	9.5	6.1	8.8	11.4(6.9)
SVM+MTL	**7.6**	**15.6**	**8.0**	**4.9**	**8.7**	**9.0(4.0)**

Table 2 Prediction MSE for Boston housing dataset (group variable: DIS)

Folds	1	2	3	4	5	Mean(st.dev)
sSVM	8.9	8.3	11.1	9.0	18.4	11.1(4.2)
mSVM	10.2	8.9	10.3	11.1	20.1	12.1(4.5)
SVM+	8.1	8.7	10.7	**7.9**	**16.5**	10.4**(3.6)**
SVM+MTL	**7.1**	**8.2**	**8.6**	8.4	17.0	**9.9**(4.0)

Table 3 Prediction MSE for auto mpg dataset (group variable: cylinders)

Folds	1	2	3	4	5	Mean(st.dev)
sSVM	6.4	**6.8**	5.9	10.6	6.6	7.3(1.9)
mSVM	5.9	7.4	6.5	10.8	8.2	7.8(1.9)
SVM+	6.7	6.9	5.8	10.3	6.6	7.3**(1.8)**
SVM+MTL	**5.5**	6.9	**5.7**	**10.3**	**4.5**	**6.6**(2.2)

Table 4 Prediction MSE for synthetic data set 1 ($d = 30$, $N = 20$, $\sigma_{noise} = 1$)

Trials	1	2	3	4	5	Mean(st.dev)
sSVM	4.2	3.7	4.2	4.0	3.8	4.0**(0.21)**
mSVM	5.1	4.9	4.5	5.1	4.9	4.9(0.27)
SVM+	4.3	3.3	3.5	3.9	4.0	3.8(0.37)
SVM+MTL	**2.9**	**2.7**	**2.9**	**3.0**	**3.4**	**3.0**(0.25)

Table 5 Prediction MSE for synthetic data set 2 ($d = 30$, $N = 50$, $\sigma_{noise} = 1$)

Trials	1	2	3	4	5	Mean(st.dev)
sSVM	2.5	2.4	2.6	2.5	2.5	2.5**(0.08)**
mSVM	2.7	1.8	2.8	2.1	2.2	2.3(0.44)
SVM+	2.5	2.4	2.4	3.4	3.0	2.7(0.44)
SVM+MTL	**1.9**	**1.7**	**1.8**	**1.7**	**1.9**	**1.8**(0.11)

Table 6 Prediction MSE for synthetic data set 3 ($d = 20$, $N = 100$, $\sigma_{noise} = 0.5$)

Trials	1	2	3	4	5	Mean(st.dev)

sSVM	7.8	7.8	8.1	7.8	7.6	7.8(0.15)
mSVM	**0.3**	**0.4**	**0.3**	**0.3**	**0.3**	**0.3(0.01)**
SVM+	7.8	7.9	7.5	7.6	7.7	7.7(0.13)
SVM+MTL	1.0	1.0	1.1	1.1	1.1	1.1(0.06)

V. Conclusions and Discussion

This paper presents new methodology called SVM+MTL regression, for utilizing group information in regression problems. This approach extends original Vapnik's SVM+ regression technology to multi-task learning. Empirical comparisons using several data sets show that the proposed SVM+MTL regression can provide significant improvement in prediction accuracy vs standard inductive SVM regression. Further, we discussed several approaches for utilizing the group information available in real-life data sets, including standard inductive SVM, multiple SVMs, SVM+ and SVM+MTL. Whereas presented empirical comparisons may suggest the advantages of SVM+MTL, we strongly warn against making such over-reaching conclusions. Relative performance of learning methods is always strongly affected by the properties of application data at hand. To this end, we also presented a synthetic data set where the proposed method shows inferior generalization performance. New learning settings, such as SVM+ regression and SVM+MTL regression, are more complex than standard SVM, and have more tuning parameters. Therefore, practical application of these new learning methodologies requires robust strategies for model selection. This paper describes such a practical strategy that combines analytic tuning of two parameters for standard SVM regression, followed by resampling approach for tuning parameters specific to SVM+ and SVM+MTL regression formulations.

Acknowledgements: This work was supported, in part, by NSF grant ECCS-0802056, by the A. Richard Newton Breakthrough Research Award from Microsoft Research, and by the BICB grant from the University of Minnesota, Rochester.

References

[1] Ando, R. and Zhang, T. A Framework for Learning predictive structures from multiple tasks and unlabeled data, Journal of Machine Learning Research, 2005.

[2] Ben-David, S., Gehrke, J. and Schuller, R. A theoretical framework for learning from a pool of disparate data sources. ACM KDD, 2002.

[3] Cherkassky, V. and Mulier, F. Learning from Data, John Wiley & Sons, New York, second edition, 2007.

[4] Hastie, T. , Tibshirani, R. and Friedman, J. The Elements of Statistical Learning. Data Mining, Inference and Prediction, Springer, 2001.

[5] Liang, L. and Cherkassky, V. Learning using Structured Data: Application to fMRI Data Analysis, IJCNN, 2007.

[6] Liang, L. and Cherkassky, V. Connection between SVM+ and Multi-Task Learning, IJCNN, 2008.

[7] Vapnik, V. Estimation of Dependencies Based on Empirical Data, Springer Verlag, New York, 1982.

[8] Vapnik, V. Statistical Learning Theory, Wiley, New York, 1998.

[9] Vapnik, V. Empirical Inference Science Afterword of 2006, Springer, 2006.

[10] Evgeniou, T. and Pontil, M.. Regularized multi--task learning. In Proc. 17th SIGKDD Conf. on Knowledge Discovery and Data Mining, 2004.