Efficient Co-Regularised Least Squares Regression

Ulf Brefeld BREFELD@INFORMATIK.HU-BERLIN.DE
Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany

Thomas Gärtner THOMAS.GAERTNER@AIS.FRAUNHOFER.DE
Fraunhofer AIS, Schloß Birlinghoven, 53754 Sankt Augustin, Germany

Tobias Scheffer SCHEFFER@INFORMATIK.HU-BERLIN.DE
Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany

Stefan Wrobel STEFAN.WROBEL@AIS.FRAUNHOFER.DE
Fraunhofer AIS and Department of Computer Science III, University of Bonn, Germany

Abstract

In many applications, unlabelled examples are inexpensive and easy to obtain. Semi-supervised approaches try to utilise such examples to reduce the predictive error. In this paper, we investigate a semi-supervised least squares regression algorithm based on the co-learning approach. Similar to other semi-supervised algorithms, our base algorithm has cubic runtime complexity in the number of unlabelled examples. To be able to handle larger sets of unlabelled examples, we devise a semi-parametric variant that scales *linearly* in the number of unlabelled examples. Experiments show a significant error reduction by co-regularisation and a large runtime improvement for the semi-parametric approximation. Last but not least, we propose a distributed procedure that can be applied without collecting all data at a single site.

1. Introduction

As unlabelled examples are much easier to obtain in most real-world learning applications than labelled ones, semi-supervised learning is gaining more and more popularity among machine learning researchers. Despite the increasing popularity of such approaches, so far they have almost exclusively been applied to classification problems. The empirical results of these papers indicate that indeed unlabelled data can be used to significantly improve the predictive performance of classification algorithms.

In this paper we develop the *semi-supervised regression* algorithm coRLSR (co-regularised least squares regression) and propose a semi-parametric variant with improved scalability. CoRLSR is based on casting co-learning as a regularised risk minimisation problem in Hilbert spaces. Similar to other kernel methods, the optimal solution in the Hilbert space can be described by a linear combination of kernel functions “centred” on the set of labelled and unlabelled instances. Similar to other semi-supervised approaches, the solution, i.e., the expansion coefficients, can be computed in time cubic in the size of the unlabelled data. As this does not reflect our intuition that semi-supervised learning algorithms should be able to process, and benefit from, huge amounts of unlabelled data, we furthermore develop a semi-parametric approximation that *scales linearly* with the amount of unlabelled data.

Our experiments on 32 data sets from UCI and on the KDD-Cup 1998 data set show that both variants of coRLSR significantly outperform supervised regression, parallelling the findings made for classification. Although non-parametric coRLSR outperforms semi-parametric coRLSR in terms of error rates, in terms of runtime our experiments confirm that semi-parametric coRLSR scales very well with the unlabelled data.

Last but not least we also consider co-regression in a distributed setting, that is, we assume that labelled data is available at different sites and must not be merged (the labels need not be on the same instances and there might be privacy concerns about moving the

Appearing in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, 2006. Copyright 2006 by the author(s)/owner(s).

data). In this setting, we propose a distributed iterative procedure that optimises the same objective function as for centralised co-regression. Assuming that (different views of) the same unlabelled data are available at the different sites, the only communication needed in each iteration is to share the predictions of each site about the unlabelled data.

In Section 2 we introduce co-learning and discuss related work. In Section 3 we derive coRLSR and its semi-parametric approximation. The distributed co-regularised least squares regression algorithm is then presented in Section 4. Finally, Section 5 reports on experimental results and Section 6 concludes.

2. Related Work

Co-classification (Blum & Mitchell, 1998; Nigam & Ghani, 2000) and co-clustering (Bickel & Scheffer, 2004) are two frameworks for classification and clustering in domains where independent views — i.e., distinct sets of attributes — of labelled and unlabelled data exist. Both are based on the observation that the rate of disagreement between independent hypotheses upper-bounds their individual error rates (de Sa, 1994). A common application of such approaches is hypertext classification where it can be assumed that the links and text of each web page present two independent views of the same data. However, minimising the rate of disagreement increases the dependency between the hypotheses and the original motivation for co-learning no longer holds. Nevertheless, the predictive performance of these approaches is often significantly better than for single-view approaches. More surprisingly even, in many domains splitting attributes at random into different views and applying a co-classification approach outperforms single-view learning algorithms (Brefeld & Scheffer, 2004).

De Sa (1994) first observed the relationship between consensus of multiple hypotheses and their error rate and devised a semi-supervised learning method by cascading multi-view vector quantisation and linear classification. Blum and Mitchell (1998) introduced the co-training algorithm for semi-supervised learning that greedily augments the training sets of two classifiers. Alternatively, a variant of the AdaBoost algorithm has been suggested in (Collins & Singer, 1999) that boosts the agreement between two views on unlabelled data.

Dasgupta et al. (2001) and Leskes (2005) give bounds on the error of co-training in terms of the disagreement rate of hypotheses on unlabelled examples in two independent views. This allows the interpretation of the disagreement as an upper bound on the error solely on the basis of unlabelled examples and justifies the direct minimisation of the disagreement. The co-EM approach to semi-supervised learning probabilistically labels all unlabelled examples and iteratively exchanges those labels between two views (Nigam & Ghani, 2000). Recently, Hardoon et al. (2006) propose a fully supervised variant of a co-support vector machine that minimises the training error as well as the disagreement between two views.

Most studies on multi-view and semi-supervised learning consider classification problems, while regression remains largely under-studied. Generally, semi-supervised graph-based classification methods can be viewed as function estimators under smoothness constraints (see Zhu, 2005, for an overview). Zhou and Li (2005) apply co-training to kNN regression. Instead of utilising two disjoint attribute sets they use distinct distance measures for the two hypotheses. An approach similar to non-parametric coRLSR has been proposed by (Sindhwani et al., 2005) for classification.

3. Efficient Co-Regression

Given training data $\{(x, y(x))\}_{x\in X}, X \subseteq \mathcal{X}, y(x) \in \mathbb{R}$, the general approach of kernel methods is to find

$$\arg\min_{f(\cdot)\in\mathcal{H}} \sum_{x\in X} V\left(y(x), f(x)\right) + \nu\Omega[f(\cdot)] \tag{1}$$

where $\Omega[f(\cdot)]$ is a regularisation term, $\mathcal{H}$ is a Hilbert space of functions often called the hypothesis space, $V(y,\cdot)$ is a convex loss function, and $\nu \geq 0$ is a parameter. Often the regularisation term $\|f(\cdot)\|^2_{\mathcal{H}}$ is used.

For M-view learning we are essentially looking for M functions from different Hilbert spaces $\mathcal{H}_v$ (possibly defined by different instance descriptions — views — and/or different kernel functions) such that the error of each function on the training data and the disagreement between the functions on the unlabelled data is small. Note, we are considering a setting slightly more general than most other co-learning approaches: firstly, we directly consider $M \geq 1$ views and secondly, the instances described by different views may differ. Thus given M finite sets of training instances $X_v \subseteq \mathcal{X}$, $\left|\bigcup_{v=1}^{M} X_v\right|$ labels $y(x) \in \mathbb{R}$, and a finite set of instances $Z \subseteq \mathcal{X}$ for which the labels are unknown we want to find $f_1 : \mathcal{X} \to \mathbb{R}, \ldots, f_M : \mathcal{X} \to \mathbb{R}$, i.e., $\mathbf{f} = (f_1, \ldots, f_M) \in \mathcal{H}_1 \times \cdots \times \mathcal{H}_M$ that minimise

$$\begin{aligned} Q(\mathbf{f}) = &\sum_{v=1}^{M} \left[\sum_{x\in X_v} V\left(y(x), f_v(x)\right) + \nu \left\|f_v(\cdot)\right\|^2 \right] \\ &+ \lambda \sum_{u,v=1}^{M} \sum_{z\in Z} V\left(f_u(z), f_v(z)\right) \end{aligned} \tag{2}$$

where the norms are in the respective Hilbert spaces and λ is a new parameter that weights the influence of pairwise disagreements. To avoid cluttering the notation unnecessarily, we omit the obvious generalisation of allowing different ν for different views.

A simple application of the *representer theorem* (Wahba, 1990; Schölkopf et al., 2001) in this context shows that the solutions of (2) always have the form

$$f_v^*(\cdot) = \sum_{x \in X_v \cup Z} c_v(x) k_v(x, \cdot), \tag{3}$$

where $k_v(\cdot, \cdot)$ is the reproducing kernel of the Hilbert Space $\mathcal{H}_v$.

This allows us to express $(f_v(x_1), f_v(x_2), \ldots)^t_{x_i \in X_v \cup Z}$ as $K_v c_v$ and $\|f_v(\cdot)\|^2$ as $c_v^t K_v c_v$, where $[K_v]_{ij} = k_v(x_i, x_j)$ and $[c_v]_i = c_v(x_i)$. Here K_v forms a (strictly) positive definite kernel matrix, i.e., it is symmetric and has no negative (and no zero) eigenvalues. Similarly, we use the notation $y_v = (y(x_1), y(x_2), \ldots)^t_{x_i \in X_v}$.

3.1. Non-Parametric Least Squares Regression

In the remainder of this paper we will concentrate on squared loss $V(y, \hat{y}) = (y - \hat{y})^2$. For standard kernel methods (1), this is known as ridge regression (Saunders et al., 1998) or regularised least squares regression (RLSR). With n_v training examples in view v and m unlabelled examples, we can rephrase (2) and obtain the exact (non-parametric) coRLSR problem :

Definition 3.1 *Let for each view $v \in \{1, \ldots, M\}$ two matrices $L_v \in \mathbb{R}^{n_v \times (n_v + m)}$ and $U_v \in \mathbb{R}^{m \times (n_v + m)}$ be given, such that*

$$K_v = \begin{pmatrix} L_v \\ U_v \end{pmatrix}$$

is strictly positive definite. For fixed $\lambda, \nu \geq 0$ the coRLSR optimisation problem *is to minimise*

$$Q(\mathbf{c}) = \sum_{v=1}^{M} \left[\|y_v - L_v c_v\|^2 + \nu c_v^t K_v c_v \right] + \lambda \sum_{u,v=1}^{M} \|U_u c_u - U_v c_v\|^2$$

over $\mathbf{c} = (c_1, \ldots, c_M) \in \mathbb{R}^{n_1 + m} \times \cdots \times \mathbb{R}^{n_M + m}$.

This optimisation problem has been considered in (Sindhwani et al., 2005) for two-view classification. In the remainder of this section we propose a closed form solution and analyse its runtime complexity.

Proposition 3.1 *The solutions c_v of the* coRLSR optimisation problem *can be found in time $O\left(M^3 m^3\right)$ (assuming $m \geq n = \max_v n_v$).*

Proof With

$$G_v = L_v^t L_v + \nu K_v + 2\lambda(M-1) U_v^t U_v$$

we get

$$\nabla_{c_v} Q(\mathbf{c}) = 2 G_v c_v - 2 L_v^t y_v - 4\lambda \sum_{u: u \neq v} U_v^t U_u c_u \ .$$

At the optimum

$$(\nabla_{c_1} Q(\mathbf{c}), \nabla_{c_2} Q(\mathbf{c}), \ldots)^t = \mathbf{0}$$

holds and we can find the exact solution by solving

$$\begin{pmatrix} G_1 & -2\lambda U_1^t U_2 & \cdots \\ -2\lambda U_2^t U_1 & G_2 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \end{pmatrix} = \begin{pmatrix} L_1^t y_1 \\ L_2^t y_2 \\ \vdots \end{pmatrix} \ .$$

This requires the inversion of a strictly positive definite matrix as

$$\begin{pmatrix} G_1 - 2\lambda U_1^t U_1 & \mathbf{0} & \cdots \\ \mathbf{0} & G_2 - 2\lambda U_2^t U_2 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

is strictly positive definite for $M \geq 2$ and

$$\begin{pmatrix} \lambda U_1^t U_1 & -\lambda U_1^t U_2 & \cdots \\ -\lambda U_2^t U_1 & \lambda U_2^t U_2 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

is positive definite.[1] The solution can thus be found in time $O\left((Mm + Mn)^3\right)$. Using $m > n$ we obtain the bound as stated above. □

For 2-view co-regression we can use the partitioned inverse equations to obtain

$$c_1 = \left(G_1 - 4\lambda^2 U_1^t U_2 {G_2}^{-1} U_2^t U_1\right)^{-1} \left(L_1^t y_1 + 2\lambda U_1^t U_2 {G_2}^{-1} L_2^t y_2\right) \ .$$

3.2. Semi-Parametric Approximation

While cubic time complexity in the number of labelled examples appears generally acceptable (supervised algorithms like SVMs, RLSR, etc. all have cubic time complexity), cubic time complexity in the number of unlabelled examples renders most real-world problems infeasible as typically $m \gg n$ (still, most state-of-the-art semi-supervised or transductive learning algorithms have cubic or worse time complexity). To achieve lower complexity in the number of unlabelled

[1] For the case $M = 1$ the problem reduces to inverting G_1 which is strictly positive definite as K_1 is strictly positive definite by definition.

instances, we resort to a semi-parametric approximation. In particular we optimise over functions that can be expanded in terms of training instances only. With n_v training examples in view v and m unlabelled examples, we can phrase the semi-parametric approximation to the coRLSR optimisation problem as

Definition 3.2 *Given for each view $v \in \{1, \dots, M\}$ a strictly positive definite matrix $L_v \in \mathbb{R}^{n_v \times n_v}$ and an arbitrary matrix $U_v \in \mathbb{R}^{m \times n_v}$. For fixed $\lambda, \nu \geq 0$ the* semi-parametric coRLSR optimisation problem *is to minimise*

$$\begin{aligned} Q(\mathbf{c}) = &\sum_{v=1}^{M} \left[\|y_v - L_v c_v\|^2 + \nu c_v^t L_v c_v \right] \\ &+ \lambda \sum_{u,v=1}^{M} \|U_u c_u - U_v c_v\|^2 \end{aligned}$$

over $\mathbf{c} = (c_1, \dots, c_M) \in \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_M}$.

Typically, L_v and U_v are computed from a strictly positive definite kernel function and form a positive definite kernel matrix $K_v \in \mathbb{R}^{(n_v+m)\times(n_v+m)}$ as

$$K_v = \begin{pmatrix} L_v & U_v^t \\ U_v & * \end{pmatrix}$$

where the part marked by $*$ is not needed.

Proposition 3.2 *The solutions c_v of the* semi-parametric coRLSR optimisation problem *can be found in time $O\left(M^3 n^2 m\right)$ (assuming $m \geq n = \max_v n_v$).*

Note that the matrices L_v, U_v, and G_v in the following proof are different from the corresponding matrices in the proof of Theorem 3.1. The symbols are overloaded as they play corresponding roles in either proof. Furthermore, this enables us to prove two theorems at once in the next section.

Proof With

$$G_v = L_v^2 + \nu L_v + 2(M-1)\lambda U_v^t U_v$$

we get

$$\nabla_{c_v} Q(\mathbf{c}) = 2G_v c_v - 2L_v y_v - 4\lambda \sum_{u:u \neq v} U_v^t U_u c_u \ .$$

At the optimum

$$(\nabla_{c_1} Q(\mathbf{c}), \nabla_{c_2} Q(\mathbf{c}), \dots)^t = \mathbf{0}$$

holds and we can find the exact solution by solving

$$\begin{pmatrix} G_1 & -2\lambda U_1^t U_2 & \cdots \\ -2\lambda U_2^t U_1 & G_2 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \end{pmatrix} = \begin{pmatrix} L_1 y_1 \\ L_2 y_2 \\ \vdots \end{pmatrix} \ .$$

This requires the inversion of a strictly positive definite matrix as

$$\begin{pmatrix} G_1 - 2\lambda U_1^t U_1 & \mathbf{0} & \cdots \\ \mathbf{0} & G_2 - 2\lambda U_2^t U_2 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

is strictly positive definite for $M \geq 2$ and

$$\begin{pmatrix} \lambda U_1^t U_1 & -\lambda U_1^t U_2 & \cdots \\ -\lambda U_2^t U_1 & \lambda U_2^t U_2 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

is positive definite. The solution can thus be found in time $O\left((Mn)^3 + M^2 m\right)$. Using $m > n$ we obtain the bound as stated above. □

For 2-view co-regression we can again make use of the partitioned inverse equations to obtain

$$\begin{aligned} c_1 = &\left(G_1 - 4\lambda^2 U_1^t U_2 {G_2}^{-1} U_2^t U_1\right)^{-1} \\ &\left(L_1 y_1 + 2\lambda U_1^t U_2 {G_2}^{-1} L_2 y_2\right) \ . \end{aligned}$$

3.3. Relation to RLSR

It turns out that the above two optimisation problems from Definitions 3.1 and 3.2 are natural generalisations of regularised least squares regression. In both cases for $m = 0$ we obtain M independent regularised least squares solutions. In the semi-parametric case we also obtain M independent regularised least squares solutions for $\lambda = 0$. For $M = 1$ the agreement term (the second part of the objective function in Definition 3.2) vanishes and we recover a single regularised least squares solution. In the non-parametric case for $\lambda = 0$ or $M = 1$ the optimisation problem still appears different from the regularised least squares optimisation problem as the regularisation term for each view includes a regularisation over the unlabelled data. However, applying the representer theorem to this case shows immediately that all components of c_v corresponding to unlabelled data will be zero for the minimiser of the optimisation problem. This shows that non-parametric as well as semi-parametric coRLSR contain traditional RLSR as a special case and can hence both be seen as natural generalisations.

4. Distributed coRLSR

Machine learning traditionally considers application scenarios where the data is available at a single site (computer/cluster) to a single machine learning algorithm. Novel problems and challenges arise whenever this is not the case and the data is distributed over many sites and must not be collected at a single site, e.g., for privacy reasons. In this section we devise a distributed coRLSR algorithm for this scenario.

Algorithm 1 Distributed CoRLSR

Require: Matrices as in Definition 3.1 and Proposition 4.1, or matrices as in Definition 3.2 and Proposition 4.2. At each site $\hat{y}_u = \mathbf{0}$.

Ensure: c_v become the optimal solution of the respective coRLSR optimisation problem

```
repeat
  for each view $v$ sequentially do
    $c_v \leftarrow G_v^{-1}\left[L_v^t y_v + 2\lambda U_v^t \sum_{u \neq v} \hat{y}_u\right]$
    $\hat{y}_v \leftarrow U_v c_v$
    send $\hat{y}_v$ to all
  end for
until convergence
```

4.1. Motivation

Consider a situation in which different companies have similar prediction problems and could greatly benefit from better predictive accuracy. This is for example the case for different loan providers each trying to prevent fraud using some prediction technique. Another example is companies trying to protect their computers from attacks over the internet using a learned model of internet connections.

In both cases sharing the data or their models could increase the quality of the predictions but the companies are rather unlikely to do that. In this section we consider the case that the different companies, however, agree on a set of unlabelled data and to exchange their predictions on this unlabelled data. As the unlabelled data may even be (appropriately generated) synthetic data, it is realistic to assume that companies do this.

4.2. Block Coordinate Descent CoRLSR

In this section we show that the above non-parametric and semi-parametric coRLSR optimisation problems can be solved by an iterative, distributed algorithm that only communicates the predictions of each site about the unlabelled data.

Proposition 4.1 *The* non-parametric coRLSR optimisation problem *can be solved by Algorithm 1 with* $G_v = L_v^t L_v + \nu K_v + 2(M-1)\lambda U_v^t U_v$.

Proposition 4.2 *The* semi-parametric coRLSR optimisation problem *can be solved by Algorithm 1 with* $G_v = L_v^2 + \nu L_v + 2(M-1)\lambda U_v^t U_v$.

With all variables defined as in the corresponding non-parametric and semi-parametric coRLSR definitions and proofs, we can prove both propositions together. Note, however, the slight notational difference between the gradient in the following proof and the gradient in the proof of Proposition 3.1. In the following we use the symmetry of L_v to replace it by its transpose L_v^t for notational harmony with the gradient in the proof of Proposition 3.2.

Proof From the respective proofs we have

$$\nabla_{c_v} Q(\mathbf{c}) = 2G_v c_v - 2L_v^t y_v - 4\lambda \sum_{u: u \neq v} U_v^t U_u c_u .$$

Now, we can compute the gradient directions using predictions ($\hat{y}_u = U_u c_u$) on the unlabelled data as

$$\nabla_{c_v} Q_v\left(c_v, y_v, \{\hat{y}_u\}_u\right) = \\ 2G_v c_v - 2L_v y_v - 4\lambda U_v^t \sum_{u: u \neq v} \hat{y}_u .$$

While the gradient direction itself is only given jointly

$$-\left(\nabla_{c_1} Q_1\left(c_1, y_1, \{\hat{y}_u\}_u\right), \nabla_{c_2} Q_2\left(c_2, y_2, \{\hat{y}_u\}_u\right), \ldots\right)^t ,$$

the global minimum can also be found by block coordinate descent (Bertsekas, 1999) over each view v. This only requires setting the block gradient to zero, i.e., solving

$$G_v c_v = L_v^t y_v + 2\lambda U_v^t \sum_{u: u \neq v} \hat{y}_u .$$

As G_v is strictly positive definite and the objective function is convex, block coordinate descent converges. □

4.3. Analysis of Distributed CoRLSR

Block coordinate descent has similar convergence properties as steepest descent (Bertsekas, 1999) which reduces the error rate in each iteration by a factor depending on the largest and the smallest eigenvalue of the Hessian. Assuming that this factor is $1/\Delta$, the error after $N \in \mathbb{N}$ iterations is reduced by a factor $1/\Delta^N$. Let $n = \max_v n_v$. Given that all labels are from the interval $[-1, 1]$, we can upper bound the starting error $Q(\mathbf{0}) - Q(\mathbf{c}^*) \leq Mn$, where $\mathbf{c}^*$ is the optimal solution. Let $\mathbf{c}^{(N)}$ be the solution of Algorithm 1 after N iterations. To achieve an error reduction factor of at least ϵ, i.e., an upper bound on the error of $Q\left(\mathbf{c}^{(N)}\right) - Q(\mathbf{c}^*) \leq Mn\epsilon$, we must have $N \geq \log_{1/\Delta} \epsilon = \log_\Delta \frac{1}{\epsilon}$ iterations.

The matrices G_v^{-1} in Algorithm 1 can be computed in time $O(m^3)$ and $O(mn^2)$ for non-parametric and semi-parametric coRLSR, respectively. It needs to be computed only once and can be computed at the same time for all sites. Step 3 of Algorithm 1 can then be computed in time $O\left(M(m+n)^2\right)$

and $O(Mmn)$, respectively. As each step has to be performed at each site, each iteration takes $O\left(M^2(m+n)^2\right)$ and $O(M^2mn)$ time, respectively. Thus to achieve an error reduction factor of at least ϵ, Algorithm 1 takes $O\left(m^3 + M^2(m+n)^2 \left\lceil \log_\Delta \frac{1}{\epsilon} \right\rceil\right)$ and $O\left(mn^2 + M^2mn \left\lceil \log_\Delta \frac{1}{\epsilon} \right\rceil\right)$ time, respectively.

Similarly, in each iteration, Mm numbers have to be broadcasted. If we consider the machine precision a constant, this requires broadcasting $O\left(Mm \left\lceil \log_\Delta \frac{1}{\epsilon} \right\rceil\right)$ bits to achieve an error reduction by the factor ϵ.

5. Empirical Evaluation

In this section we summarise experiments comparing regular RLSR with non-parametric and semi-parametric coRLSR on benchmark regression datasets.

In all experiments we use a Gaussian kernel $k(x, x') = \exp(-\|x - x'\|^2/\sigma)$ with $\sigma = 1/n^2 \sum_{i,j=1}^n \|x_i - x_j\|^2$ and $\nu = (\sum_{i=1}^n \|x_i\|/n)^{-1}$ as regularisation parameter. Note, that σ and ν depend only on the labelled examples; in case of multiple views, σ_v and ν_v are computed from the attributes in the respective view v. We report scaled root mean square errors (rmse)

$$\text{rmse}(f) = \frac{1}{\max y_i} \sqrt{\frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2}.$$

which allows viewing all results in the same figure.

5.1. UCI Experiments

The UCI repository (Newman et al., 1998) contains 63 data sets with continuous target attributes. We omit data sets containing less than 50 examples and/or less than 4 attributes. We leave out the largest 20 data sets because of memory problems in Matlab with inverting the matrices for the non-parametric case. On the remaining 32 data sets we perform a 10-fold 'inverse' cross validation, i.e., in each run we use one fold as labelled examples and the other 9 folds as unlabelled and holdout examples. In each run the available attributes are split randomly into two disjoint sets. The results are averages over 20 such runs. In all experiments we use $\lambda = 1/10$. The results are shown in Figure 1 where error bars indicate the standard error.

In Figure 1 (left) we plot the rmse of regular RLSR for all 32 UCI problems (x-axis) against the corresponding rmse values of non-parametric coRLSR (y-axis). Thus, each point refers to a UCI problem. The dashed line marks the threshold where both methods perform equally well. Points below this line indicate that non-parametric coRLSR has a lower rmse for these data sets compared to regular RLSR. Figure 1 (middle) shows the analogue for regular RLSR and semi-parametric coRLSR. Both comparisons show that the multi-view algorithms outperform the baseline in most of the 32 problems. Figure 1 (right) compares the two multi-view methods. Semi-parametric coRLSR performs slightly worse than non-parametric coRLSR.

While the Figures indicate that coRLSR outperforms the baseline RLSR method over all datasets, we want to confirm this hypothesis in a sound statistical test. We use the null hypotheses that the algorithms perform equally well. As suggested recently by Demšar (2006) we use the Wilcoxon signed ranks test.

The Wilcoxon signed ranks test is a nonparametric test to detect shifts in populations given a number of paired samples. The underlying idea is that under the null hypothesis the distribution of differences between the two populations is symmetric about 0. It proceeds as follows: (*i*) compute the differences between the pairs, (*ii*) determine the ranking of the absolute differences, and (*iii*) sum over all ranks with positive and negative difference to obtain W_+ and W_-, respectively. The null hypothesis can be rejected if W_+ (and W_- depending on whether we need a one-sided or a two-sided test) is located in the tail of the null distribution which has sufficiently small probability.

The critical value of the one-sided Wilcoxon signed ranks test for 32 samples on a 0.5% significance level is 128. The test statistic for comparing non-parametric coRLSR against RLSR is $54 < 128$, the test statistic for comparing semi-parametric coRLSR against RLSR is $66 < 128$, and finally the test statistic for comparing parametric coRLSR against semi-parametric coRLSR is $63 < 128$. Hence on this significance level we can reject all three null hypotheses. We conclude that the multi-view algorithms significantly outperform regular RLSR and that non-parametric coRLSR is the best regression method of our study.

5.2. KDD Cup Experiments

In the KDD Cup data set, the task is to predict the amount of money donated to a charity. The original data set comes with 479 attributes. We use a binary encoding of nominal attributes with less than 200 distinct values. Since there are many missing values we add an indicator attribute for each continuous attribute. The indicator equals 1 if the actual value is missing and 0 otherwise. The modified data set contains 95412 training examples with 5551 attributes. We use a resampling approach to adjust λ. For a fixed λ we draw a specified number of labelled and unlabelled examples and distinct holdout examples at random in each iteration. We average the rmse on the holdout set over 25

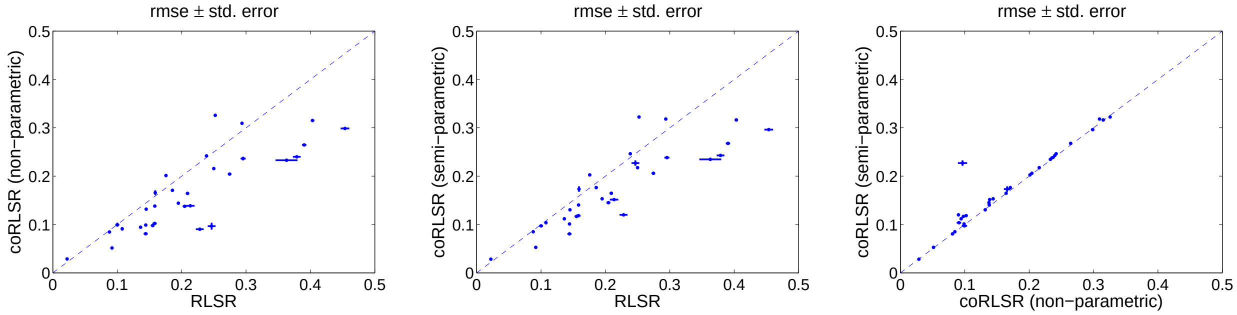

Figure 1. Pairwise rmse for non-parametric coRLSR, semi-parametric coRLSR, and regular RLSR over 32 UCI data sets.

runs with distinct, randomly drawn attribute splits. We compare equally many parameter values for all methods. For each problem we fix the apparently optimal λ for all methods and reevaluate the rmse for these parameter settings by again averaging over 25 runs with distinct resampled training and holdout sets.

In order to explore the influence of unlabelled examples we use 100 labelled and 200 holdout examples and vary the number of unlabelled examples. The results are shown in Figure 2. For 100 labelled and no unlabelled examples both multi-view algorithms have lower rmse compared to the baseline by simply averaging the predictions of the two views. As the number of unlabelled examples increases, the advantage of multi-view over single-view regression increases further. Again, non-parametric coRLSR turns out to be the best regression method.

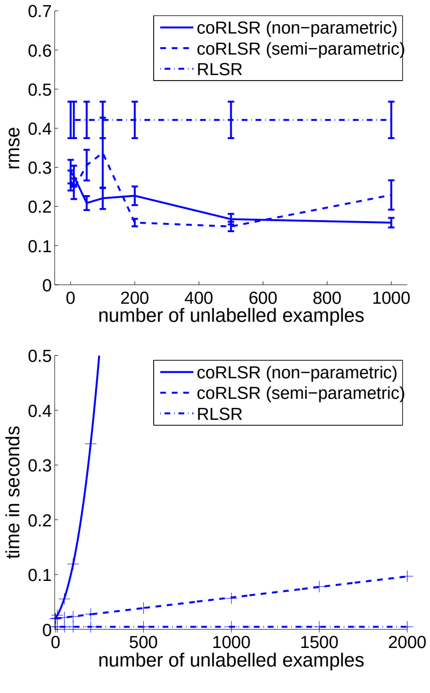

Figure 2. Results on the KDD Cup 98 data set with 100 labelled instances and varying numbers of unlabelled ones.

The performance of semi-parametric coRLSR can be further improved by increasing the number of unlabelled instances. We observe average rmse values of 0.1312 ± 0.006 for 10,000 unlabelled instances, 0.1078 ± 0.004 for 50,000 unlabelled instances, and 0.1253 ± 0.006 for 90,000 unlabelled instances. Note, that non-parametric coRLSR is not feasible for these sample sizes.

Figure 2 also compares the execution time of regular RLSR, non-parametric, and semi-parametric coRLSR. The figure shows execution time for a fixed number of labelled and different numbers of unlabelled examples and fitted polynomials. The empirical results confirm our theoretical findings. Non-parametric coRLSR is costly in terms of computation time (the degree of the fitted polynomial is 3). Its approximation is considerably faster (the fit of semi-parametric coRLSR is a linear function of the number of unlabelled examples as shown in Proposition 3.2). For any number of unlabelled datapoints, the runtime of semi-parametric coRLSR is comparable to that of regular RLSR.

6. Conclusions

In this paper we proposed co-regularised least squares regression (coRLSR), a semi-supervised regression algorithm based on the co-learning framework. While coRLSR like many other semi-supervised or transductive approaches has cubic runtime complexity in the amount of unlabelled data, we proposed a semi-parametric approximation of coRLSR which scales linearly in the amount of unlabelled data.

Both non-parametric and semi-parametric coRLSR have closed form solutions in the usual centralised learning setting. Additionally, both can be solved in the less common distributed learning setting where the labelled data must not be joined at a single site. This can be achieved by an iterative distributed algorithm that only communicates the predictions about the unlabelled data at each iteration.

Empirical results showed that coRLSR as well as its semi-parametric approximation clearly outperform traditional regularised least squares regression even on problems where there is no given obvious feature split. The observed improvements were achieved by applying co-learning based on a random feature split and thus might even be more pronounced when natural groups of features are available. While non-parametric coRLSR outperforms its semi-parametric approximation in predictive accuracy, in terms of runtime semi-parametric coRLSR is clearly more desireable than the exact, non-parametric, version.

Acknowledgements

This work has partially been funded by the German Science Foundation DFG under grants SCHE540/10-2 and WR40/2-2. The authors are grateful to Michael Brückner, Tamás Horváth, Heinz Mühlenbein, Stefan Rüping, and Alex J Smola for valuable discussions and comments.

References

Bertsekas, D. (1999). *Nonlinear programming.* Athena Scientific.

Bickel, S., & Scheffer, T. (2004). Multi-view clustering. *Proceedings of the IEEE International Conference on Data Mining.*

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data. *Proceedings of the Conference on Computational Learning Theory.*

Brefeld, U., & Scheffer, T. (2004). Co-EM support vector learning. *Proceedings of the International Conference on Machine Learning.*

Collins, M., & Singer, Y. (1999). Unsupervised models for named entity classification. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.*

Dasgupta, S., Littman, M., & McAllester, D. (2001). PAC generalization bounds for co-training. *Proceedings of Neural Information Processing Systems.*

de Sa, V. (1994). Learning classification with unlabeled data. *Proceedings of Neural Information Processing Systems.*

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*.

Hardoon, D., Farquhar, J. D. R., Meng, H., Shawe-Taylor, J., & Szedmak, S. (2006). Two view learning: SVM-2K, theory and practice. *Advances in Neural Information Processing Systems.*

Leskes, B. (2005). The value of agreement, a new boosting algorithm. *Proceedings of the Conference on Learning Theory.*

Newman, D. J., Hettich, S., Blake, C. L., & Merz, C. J. (1998). UCI repository of machine learning databases.

Nigam, K., & Ghani, R. (2000). Analyzing the effectiveness and applicability of co-training. *Proceedings of Information and Knowledge Management.*

Saunders, C., Gammerman, A., & Vovk, V. (1998). Ridge regression learning algorithm in dual variables. *Proceedings of the Fifteenth International Conference on Machine Learning.* Morgan Kaufmann.

Schölkopf, B., Herbrich, R., & Smola, A. J. (2001). A generalized representer theorem. *Proceedings of the 14th annual conference on learning theory.*

Sindhwani, V., Niyogi, P., & Belkin, M. (2005). A co-regularized approach to semi-supervised learning with multiple views. *Proceedings of the ICML Workshop on Learning with Multiple Views.*

Wahba, G. (1990). *Spline models for observational data*, vol. 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics.* Philadelphia: SIAM.

Zhou, Z.-H., & Li, M. (2005). Semi-supervised regression with co-training. *Proceedings of the International Joint Conference on Artificial Intelligence.*

Zhu, X. (2005). *Semi-supervised learning in literature survey* (Technical Report 1530). Computer Sciences, University of Wisconsin-Madison.