Learning Preferences with Co-Regularized Least-Squares

Evgeni Tsivtsivadze[1], Fabian Gieseke[2], Tapio Pahikkala[1], Jorma Boberg[1], and Tapio Salakoski[1]

[1] Turku Centre for Computer Science (TUCS),
Department of Information Technology, University of Turku,
Joukahaisenkatu 3-5 B, 20520 Turku, Finland
{firstname.lastname}@it.utu.fi
[2] Faculty of Computer Science, Technische Universität Dortmund,
Otto-Hahn-Str. 14, 44227 Dortmund, Germany
{firstname.lastname@cs.uni-dortmund.de}

Abstract. Situations when only a limited amount of labeled data and a large amount of unlabeled data is available to the learning algorithm are typical for many real-world problems. In this paper, we propose a semi-supervised preference learning algorithm that is based on the multi-view approach. Multi-view learning algorithms operate by constructing a predictor for each view and by choosing such prediction hypotheses that minimize the disagreement among all of the predictors on the unlabeled data. Our algorithm, that we call *Sparse Co-RankRLS*, stems from the single-view preference learning algorithm RankRLS. It minimizes a least-squares approximation of the ranking error and is formulated within the co-regularization framework. The experiments demonstrate a significantly better performance of Sparse Co-RankRLS compared to the standard RankRLS algorithm. Moreover, our semi-supervised preference learning algorithm has a linear complexity in the number of unlabeled data items, making it applicable to large datasets.

1 Introduction

Semi-supervised learning algorithms have gained more and more attention in recent years as unlabeled data is typically much easier to obtain than labeled one. *Multi-view* learning algorithms, such as co-training [2], split the attributes into independent sets and an algorithm is learnt based on these different "views". The goal of the learning process consists in finding for every view a prediction function (for the learning task) performing well on the labeled data of the designated view such that all prediction functions agree on the unlabeled data. Closely related to this approach is the *co-regularization* framework described in [19], where the same idea of agreement maximization between the predictors is central. Briefly stated, algorithms based upon this approach search for hypotheses from different Reproducing Kernel Hilbert Spaces [18], namely views, such that the training error of each hypothesis on the labeled data is small and, at the

same time, the hypotheses give similar predictions for the unlabeled data. Within this framework, the disagreement is taken into account via a co-regularization term. Empirical results show that the co-regularization approach works well for classification [19], regression [3], and clustering [4] tasks. Moreover, theoretical investigations demonstrate that the co-regularization approach reduces the Rademacher complexity by an amount that depends on the "distance" between the views [17, 20].

We consider a problem of learning a function capable of arranging data points according to a given preference relation [8]. Training of existing kernel based ranking algorithms such as RankSVM [10] may be infeasible when the size of the training set is large. This is especially the case when nonlinear kernel functions are used. Recently, a sparse preference learning algorithm, called *Sparse RankRLS*, that can take advantage of a large amount of data in the training process, has been proposed [22]. In this paper, we will formulate a co-regularized version of RankRLS, called *Sparse Co-RankRLS*, and aim to improve the performance of RankRLS by making it applicable to situations when only a small amount of labeled data, but a large amount of unlabeled data is available.

We evaluate our algorithm on a *parse ranking task* [23] that is a common problem in natural language processing. In this task, the aim is to rank a set of parses associated with a single sentence, based on some goodness criteria. In our experiments, we consider the case when both labeled and a large amount of unlabeled data is available to the learning algorithm. We demonstrate that Sparse Co-RankRLS is computationally efficient when trained on large datasets and the obtained results are significantly better than the ones obtained with the standard RankRLS algorithm.

2 Problem Setting

Let $\mathcal{X}$ be a set of instances and $\mathcal{Y}$ be a set of labels. The learning scenario we consider is *label ranking* [6, 8], i.e., we want to predict for any instance $x \in \mathcal{X}$ (e.g., a person) a preference relation $\mathcal{P}_x \subseteq \mathcal{Y} \times \mathcal{Y}$ among the set of labels $\mathcal{Y}$, where each label $y \in \mathcal{Y}$ can be thought of as an alternative (e.g. a politician in an election). An element $(y, y') \in \mathcal{P}_x$ means that the instance x prefers the label y compared to y', also written as $y \succ_x y'$. [3] We assume that the (true) preference relation $\mathcal{P}_x$ is transitive and asymmetric for each instance $x \in \mathcal{X}$. As training information, we are given a finite set $\{(z_i, s_i)\}_{i=1}^m$ of m data points, where each data point $(z_i, s_i) = ((x_i, y_i), s_i) \in (\mathcal{X} \times \mathcal{Y}) \times \mathbb{R}$ consists of an instance-label tuple $z_i = (x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ and its score $s_i \in \mathbb{R}$. We say that two data points $((x, y), s)$ and $((x', y'), s')$ are *relevant*, iff $x = x'$. Considering two relevant data points $((x, y), s)$ and $((x, y'), s')$, we say that instance x *prefers* label y to y', if $s > s'$. If $s = s'$, the labels are called *tied*. Accordingly, we write $y \succ_x y'$ if $s > s'$ and $y \sim_x y'$ if $s = s'$.

[3] As described in [8], one can distinguish between *weak preference* ($\succeq$) and *strict preference* ($\succ$), where $y \succ_x y' \Leftrightarrow (y \succeq_x y') \wedge (y' \not\succeq_x y)$; furthermore, $y \sim_x y' \Leftrightarrow (y \succeq_x y') \wedge (y' \succeq_x y)$.

A *label ranking function* is a function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ mapping each instance-label tuple (x, y) to a real value representing the (predicted) relevance of the label y with respect to the instance x. This induces for any instance $x \in \mathcal{X}$ a transitive preference relation $\mathcal{P}_{f,x} \subseteq \mathcal{Y} \times \mathcal{Y}$ with $(y, y') \in \mathcal{P}_{f,x} \Leftrightarrow f(x, y) > f(x, y')$. Ties can be broken arbitrarily. Informally, the goal of our ranking task is to find a label ranking function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ such that the ranking $\mathcal{P}_{f,x} \subseteq \mathcal{Y} \times \mathcal{Y}$ induced by the function for any instance $x \in \mathcal{X}$ is a good "prediction" for the true (unknown) preference relation $\mathcal{P}_x \subseteq \mathcal{Y} \times \mathcal{Y}$.

To be able to incorporate the relevance information, we define a *preference graph* which is an undirected graph $G = (V, E)$ whose vertices are the m instance-label tuples appearing in the training set, i.e., $V = \{z_1, \ldots, z_m\}$. Furthermore, there exist an edge $(z_i, z_j) \in E$, iff z_i and z_j are relevant. Let $W \in \mathbb{R}^{m \times m}$ denote the adjacency matrix of G, i.e., $[W]_{i,j} = 1$ if $(z_i, z_j) \in E$ and $[W]_{i,j} = 0$ otherwise. To avoid loops, we set $[W]_{i,i} = 0$ for $i = 1, \ldots, m$, although an instance-label tuple is relevant to itself. Furthermore, let $Z = (z_1, \ldots, z_m)^t \in (\mathcal{X} \times \mathcal{Y})^m$ be the vector of instance-label training tuples and $S = (s_1, \ldots, s_m)^t \in \mathbb{R}^m$ the corresponding vector of scores. Given these definitions, our training set is the triple $T = (Z, S, W)$.

Let us define $\mathbb{R}^{\mathcal{Z}} = \{f : \mathcal{Z} \to \mathbb{R}\}$ with $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and let $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{Z}}$ be the hypothesis space of possible ranking functions. To measure how well a hypothesis $f \in \mathcal{H}$ is able to predict the preference relations $\mathcal{P}_x$ for all instances $x \in \mathcal{X}$, we consider the following cost function that captures the amount of incorrectly predicted pairs of relevant training data points:

$$d(f, T) = \frac{1}{2} \sum_{i,j=1}^{m} [W]_{i,j} \left| \operatorname{sign}\big(s_i - s_j\big) - \operatorname{sign}\big(f(z_i) - f(z_j)\big) \right|, \tag{1}$$

where $\operatorname{sign}(\cdot)$ denotes the signum function. It is well-known that the use of cost functions like (1) leads to intractable optimization problems. Therefore, we consider the following least squares approximation, which in fact regresses the differences $s_i - s_j$ with $f(z_i) - f(z_j)$ of relevant training data points z_i and z_j:

$$c(f, T) = \frac{1}{2} \sum_{i,j=1}^{m} [W]_{i,j} \Big((s_i - s_j) - (f(z_i) - f(z_j)) \Big)^2. \tag{2}$$

Note that the above cost function c also takes the extent of discrepancy between the predicted preference $(f(z_i) - f(z_j))$ and the training preference $(s_i - s_j)$ of pairs of relevant training data points into account.

3 Regularized Least Squares Ranking

The co-regularized ranking algorithm presented in this paper stems from the results developed in [11] and [22]. For completeness, we briefly review these results in this section.

We aim to construct an algorithm that selects a hypothesis f from $\mathcal{H}$ which minimizes (2) and which is, at the same time, not too "complex", i.e., which does not overfit at training phase and is therefore able to generalize to unseen data. We consider the framework of regularized kernel methods [18], in which $\mathcal{H}$ is a so-called *Reproducing Kernel Hilbert Space* (RKHS) defined by a positive definite kernel function.

3.1 Regularization Framework

Let $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be a positive definite kernel defined on the set $\mathcal{Z}$. Then we define $\mathcal{H}$ as

$$\mathcal{H} = \left\{ f \in \mathbb{R}^{\mathcal{Z}} \mid f(\cdot) = \sum_{j=1}^{\infty} \beta_j k(\cdot, z_j), \beta_j \in \mathbb{R}, z_j \in \mathcal{Z}, \|f\|_{\mathcal{H}} < \infty \right\}, \tag{3}$$

where $\|\cdot\|_{\mathcal{H}}$ denotes the norm in $\mathcal{H}$. Using the RKHS $\mathcal{H}$ as our hypothesis space, we consider the optimization problem

$$\mathcal{A}(T) = \operatorname*{argmin}_{f \in \mathcal{H}} J(f), \tag{4}$$

where $J(f) = c(f, T) + \lambda \|f\|_{\mathcal{H}}^2$ and where $\lambda \in \mathbb{R}^+$ is a regularization parameter controlling the tradeoff between the cost on the training set and the complexity of the hypothesis. By the generalized representer theorem [18], the minimizer of (4) has the form

$$f^*(\cdot) = \sum_{i=1}^{m} a_i k(\cdot, z_i) \tag{5}$$

with appropriate coefficients $a_i \in \mathbb{R}$. Hence, we can focus on functions $f \in \mathcal{H}$ having the above form. Defining the kernel matrix $K \in \mathbb{R}^{m \times m}$ with entries of the form $[K]_{i,j} = k(z_i, z_j)$ and $f(Z) = (f(z_1), \ldots, f(z_m))^t \in \mathbb{R}^m$, we can write $f(Z) = KA$ and $\|f\|_{\mathcal{H}}^2 = A^t K A$, where $A = (a_1, \ldots, a_m)^t \in \mathbb{R}^m$ is a corresponding coefficient vector.[4]

3.2 RankRLS

Let $\mathcal{L} = D - W$ be the Laplacian matrix [5] of G, where D denotes the diagonal matrix with elements of the form $[D]_{i,i} = \sum_{j=1}^{m} [W]_{i,j}$. Using a slightly different notation, it is shown in [11] that the cost function (2) can be rewritten as

$$c(f, T) = (S - KA)^t \mathcal{L} (S - KA). \tag{6}$$

[4] Unless stated otherwise, we assume that a kernel matrix K is positive definite, i.e., $B^t K B > 0$ for all $B \in \mathbb{R}^m, B \neq 0$. This can be ensured, for example, by performing a small diagonal shift.

Considering this representation of the cost function c, we get the following optimization problem called *RankRLS* in [11]:

$$\mathcal{A}(T) = \operatorname*{argmin}_{A \in \mathbb{R}^m} J(A), \tag{7}$$

where $J(A) = (S - KA)^t \mathcal{L}(S - KA) + \lambda A^t K A$. Using the fact that $\mathcal{L}$ is positive semidefinite [14] and assuming that K is positive definite, it is easy to see that the Hessian matrix $H(J) = 2K^t \mathcal{L} K + 2\lambda K$ of J is positive definite. Thus, J is strictly convex and the global minimum of J can be obtained by setting the first derivative $\frac{d}{dA} J(A) = -2K^t \mathcal{L}(S - KA) + 2\lambda KA$ to zero and by solving the resulting system of equations with respect to A. As shown in [11], the optimal solution for (7) is

$$A = (K\mathcal{L}K + \lambda K)^{-1} K\mathcal{L}S = (\mathcal{L}K + \lambda I)^{-1} \mathcal{L}S, \tag{8}$$

where I denotes the identity matrix. The computational complexity of the matrix inversion in (8) is $\mathcal{O}(m^3)$.

Fact 1 *([11]) For fixed $\lambda \in \mathbb{R}^+$, the solution of the RankRLS optimization problem (7) can be found in $\mathcal{O}(m^3)$ time.*

3.3 Sparse RankRLS

Similarly to [13] and [21], an approximation algorithm aiming at reducing the cubic running time of the RankRLS approach is developed in [22]: The cost function c is evaluated over *all* points, but only a subset of the coefficients $a_1, \ldots, a_m$ is allowed to be non-zero, thus an approximation of the optimization problem is considered. Let $R = \{i_1, \ldots, i_r\} \subseteq \{1, \ldots, m\}$ be a subset of indices. Then, we only allow the coefficients $a_{i_1}, \ldots, a_{i_r}$ to be non-zero in (5), i.e., we search for minimizers $\widehat{f} \in \mathcal{H}$ having the form

$$\widehat{f}(\cdot) = \sum_{j=1}^{r} a_{i_j} k(\cdot, z_{i_j}). \tag{9}$$

By defining $\bar{K} \in \mathbb{R}^{m \times r}$ to be the submatrix of $K \in \mathbb{R}^{m \times m}$ that only contains the columns indexed by R and by defining $\widehat{K} \in \mathbb{R}^{r \times r}$ to be the submatrix of $\bar{K}$ only containing the rows indexed by R, we can express $\widehat{f}(Z) = (\widehat{f}(z_1), \ldots, \widehat{f}(z_m))^t \in \mathbb{R}^m$ as $\widehat{f}(Z) = \bar{K}\widehat{A}$ and $\|\widehat{f}\|^2_{\mathcal{H}} = \widehat{A}^t \widehat{K} \widehat{A}$, where $\widehat{A} = (a_{i_1}, \ldots, a_{i_r})^t \in \mathbb{R}^r$. Given these notations, the approximation presented in [22], called *Sparse RankRLS*, can be formulated as

$$\mathcal{A}(T) = \operatorname*{argmin}_{\widehat{A} \in \mathbb{R}^r} \widehat{J}(\widehat{A}), \tag{10}$$

where $\widehat{J}(\widehat{A}) = (S - \bar{K}\widehat{A})^t \mathcal{L}(S - \bar{K}\widehat{A}) + \lambda \widehat{A}^t \widehat{K} \widehat{A}$. Setting the derivative of $\widehat{J}$ to zero and solving the resulting system of equations with respect to $\widehat{A}$ leads to

$$\widehat{A} = (\bar{K}^t \mathcal{L} \bar{K} + \lambda \widehat{K})^{-1} \bar{K}^t \mathcal{L} S. \tag{11}$$

The overall training complexity of the Sparse RankRLS algorithm is $\mathcal{O}(mr^2)$, see [22] for more details.

Fact 2 *([22]) For fixed $\lambda \in \mathbb{R}^+$, the solution of the Sparse RankRLS optimization problem (10) can be found in $\mathcal{O}(mr^2)$ time.*

Hence, selecting r to be much smaller than m results in a significant acceleration of the training procedure. Clearly, the selection of the index set R may have an influence on results obtained by the above approximation approach. Different methods for selecting R are discussed, for example, in [16]. There, it is found that simply selecting the elements of R randomly performs no worse than more sophisticated methods. Hereafter, we refer to the data points contained in R as *basis vectors*.

3.4 Constructing Kernels with Subsets of Regressors

Considering the Sparse RankRLS algorithm, the label predictions for the training data points can be obtained by $\bar{K}\widehat{A}$. Using the Woodbury matrix identity [9] and (11) and by defining $\widetilde{K} = \frac{1}{\lambda}\bar{K}\widehat{K}^{-1}\bar{K}^t$, we can reformulate this expression as follows:

$$\begin{aligned}
\bar{K}\widehat{A} &= \bar{K}(\bar{K}^t\mathcal{L}\bar{K} + \lambda\widehat{K})^{-1}\bar{K}^t\mathcal{L}S \\
&= \bar{K}(\frac{1}{\lambda}\widehat{K}^{-1} - \frac{1}{\lambda}\widehat{K}^{-1}\bar{K}^t(\frac{1}{\lambda}\mathcal{L}\bar{K}\widehat{K}^{-1}\bar{K}^t + I)^{-1}\frac{1}{\lambda}\mathcal{L}\bar{K}\widehat{K}^{-1})\bar{K}^t\mathcal{L}S \\
&= (\widetilde{K} - \widetilde{K}(\mathcal{L}\widetilde{K} + I)^{-1}\mathcal{L}\widetilde{K})\mathcal{L}S \\
&= (\widetilde{K}(I - (\mathcal{L}\widetilde{K} + I)^{-1}\mathcal{L}\widetilde{K})\mathcal{L}S \\
&= (\widetilde{K}((\mathcal{L}\widetilde{K} + I)^{-1}(\mathcal{L}\widetilde{K} + I) - (\mathcal{L}\widetilde{K} + I)^{-1}\mathcal{L}\widetilde{K})\mathcal{L}S \\
&= \widetilde{K}(\mathcal{L}\widetilde{K} + I)^{-1}(\mathcal{L}\widetilde{K} + I - \mathcal{L}\widetilde{K})\mathcal{L}S \\
&= \widetilde{K}(\mathcal{L}\widetilde{K} + I)^{-1}\mathcal{L}S.
\end{aligned}$$

Note that because $\mathcal{L}$ and $\widetilde{K}$ are positive semidefinite, their product $\mathcal{L}\widetilde{K}$ contains only nonnegative eigenvalues [1]. Hence, $\mathcal{L}\widetilde{K} + I$ is invertible. Further, the last term can be rewritten as $\widetilde{K}(\mathcal{L}\widetilde{K} + I)^{-1}\mathcal{L}S = \check{K}(\mathcal{L}\check{K} + \lambda I)^{-1}\mathcal{L}S$, where $\check{K} = \bar{K}\widehat{K}^{-1}\bar{K}^t \in \mathbb{R}^{m\times m}$. These derivations show that the Sparse RankRLS algorithm operating with a kernel function k is essentially equivalent to the standard RankRLS algorithm operating with a modified kernel $\check{k}$. In the following section we will use this fact for constructing different Hilbert spaces by taking different sets of basis vectors.

4 Co-Regularized Least Squares Ranking

Both the RankRLS and the Sparse RankRLS algorithm, can only use labeled data points during the training phase. In this section, we present the algorithm that is applicable to situations when only a small amount of labeled, but a large amount of unlabeled data is available.

4.1 Co-Regularization Framework

The co-regularization approach is based on the idea of constructing M hypotheses from M different Hilbert spaces such that the error of each function on the labeled data is small and, at the same time, the functions give similar predictions for the unlabeled data.

As shown in Section 3.4, the solution of the Sparse RankRLS algorithm equals to the one obtained by the standard RankRLS algorithm with a modified kernel function. Hence, taking different subsets of the input set leads to different Reproducing Kernel Hilbert Spaces. These RKHSs can also stem from different data point descriptions (i.e., different features) and/or different kernel functions. In the following, we will consider M different RKHSs $\mathcal{H}_1, \ldots, \mathcal{H}_M$ and corresponding kernel functions $k_1, \ldots, k_M$ with $k_v : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$. Considering our ranking task, we have a training set $T = (Z, S, W)$ originating from a set $\{(z_i, s_i)\}_{i=1}^m$ of data points *with* scoring information, where $Z = (z_1, \ldots, z_m)^t \in \mathcal{Z}^m$, $S = (s_1, \ldots, s_m)^t \in \mathbb{R}^m$, and where W is the matrix incorporating the relevance information. Moreover, we have a training set $\widetilde{T} = (\widetilde{Z}, \widetilde{W})$ from a set $\{z_{m+i}\}_{i=1}^n$ of data points *without* scoring information, $\widetilde{Z} = (z_{m+1}, \ldots, z_{m+n})^t \in \mathcal{Z}^n$, and an appropriate adjacency matrix $\widetilde{W}$. To avoid misunderstandings with the definition of the label ranking task, we will use the terms "scored" instead of "labeled" and "unscored" instead of "unlabeled".

In the ranking task, we search for the functions $\mathbf{f} = (f_1, \ldots, f_M) \in \mathcal{H}_1 \times \ldots \times \mathcal{H}_M$ minimizing

$$J(\mathbf{f}) = \sum_{v=1}^{M} c(f_v, T) + \lambda \sum_{v=1}^{M} \|f_v\|_{\mathcal{H}_v}^2 + \nu \sum_{v,u=1}^{M} V(f_v, f_u, \widetilde{T}), \tag{12}$$

where $\lambda, \nu \in \mathbb{R}^+$ are regularization parameters and where V is the loss function measuring the disagreement between the prediction functions of the views on the unscored data:

$$V(f_v, f_u, \widetilde{T}) = \frac{1}{2} \sum_{i,j=1}^{n} \widetilde{[W]}_{i,j} \Big(\big(f_v(z_{m+i}) - f_v(z_{m+j})\big) - \big(f_u(z_{m+i}) - f_u(z_{m+j})\big) \Big)^2.$$

Applying the representer theorem [18] in this context shows that the minimizers $f_v^* \in \mathcal{H}_v$ of (12) for $v = 1, \ldots, M$ have the form

$$f_v^*(\cdot) = \sum_{i=1}^{m} a_i^{(v)} k_v(\cdot, z_i) + \sum_{i=1}^{n} a_{m+i}^{(v)} \, k_v(\cdot, z_{m+i}) \tag{13}$$

with adequate coefficients $a_1^{(v)}, \ldots, a_{m+n}^{(v)} \in \mathbb{R}$.

4.2 Sparse Co-RankRLS

Using matrix notations we can reformulate (12) as

$$J(\mathbf{A}) = \sum_{v=1}^{M} (S - L_v A_v)^t \mathcal{L}_L (S - L_v A_v) + \lambda \sum_{v=1}^{M} {A_v}^t K_v A_v \qquad (14)$$
$$+ \nu \sum_{v,u=1}^{M} (U_v A_v - U_u A_u)^t \mathcal{L}_U (U_v A_v - U_u A_u),$$

where $A_v = (a_1^{(v)}, \ldots, a_{m+n}^{(v)})^t \in \mathbb{R}^{m+n}$ and $\mathbf{A} = (A_1^t, \ldots, A_M^t)^t \in \mathbb{R}^{M(m+n)}$. The matrix $L_v \in \mathbb{R}^{m \times (m+n)}$ has entries of the form $[L_v]_{i,j} = k_v(z_i, z_j)$ and the matrix $U_v \in \mathbb{R}^{n \times (m+n)}$ has entries of the form $[U_v]_{i,j} = k_v(z_{m+i}, z_j)$. Stacking both matrices up gives the matrix K_v:

$$K_v = \begin{pmatrix} L_v \\ U_v \end{pmatrix} \in \mathbb{R}^{(m+n) \times (m+n)}.$$

Further, $\mathcal{L}_L \in \mathbb{R}^{m \times m}$ and $\mathcal{L}_U \in \mathbb{R}^{n \times n}$ denote the Laplacian matrices corresponding to W and $\widetilde{W}$, respectively. Hence, we have the following optimization problem:

$$\mathcal{A}(T, \widetilde{T}) = \operatorname*{argmin}_{\mathbf{A} \in \mathbb{R}^{M(m+n)}} J(\mathbf{A}). \qquad (15)$$

Although the Hilbert spaces $\mathcal{H}_1, \ldots, \mathcal{H}_M$ can stem from different data point descriptions and/or different kernel functions, we consider the case when they are obtained with different subsets of the input set $\mathcal{Z}$. Considering the minimizers in (13), we only allow a subset of the coefficients to be non-zero for each view. As in Section 3, this corresponds to taking submatrices of the original matrices, i.e., for each view v we define $\bar{L}_v \in \mathbb{R}^{m \times r}$ to be the submatrix of L_v that only contains the columns corresponding to r selected basis vectors $z_{c_v(1)}, \ldots, z_{c_v(r)}$. Here, the number $c_v(i) \in \{1, \ldots, m+n\}$ denotes the index (column) of the i-th selected vector of view v. Accordingly, we define $\bar{U}_v \in \mathbb{R}^{n \times r}$ to be the submatrix of U_v that only contains the columns corresponding to $z_{c_v(1)}, \ldots, z_{c_v(r)}$. Finally, we define $\widehat{K}_v \in \mathbb{R}^{r \times r}$ to be the kernel matrix with elements $\left[\widehat{K}_v\right]_{i,j} = k_v(z_{c_v(i)}, z_{c_v(j)})$. Hence, we obtain the following optimization problem, which we call *Sparse Co-RankRLS*:

$$\mathcal{A}(T, \widetilde{T}) = \operatorname*{argmin}_{\widehat{\mathbf{A}} \in \mathbb{R}^{Mr}} \widehat{J}(\widehat{\mathbf{A}}), \qquad (16)$$

where

$$\widehat{J}(\widehat{\mathbf{A}}) = \sum_{v=1}^{M} \left(S - \bar{L}_v \widehat{A}_v\right)^t \mathcal{L}_L \left(S - \bar{L}_v \widehat{A}_v\right) + \lambda \sum_{v=1}^{M} \widehat{A}_v^t \widehat{K}_v \widehat{A}_v \qquad (17)$$
$$+ \nu \sum_{v,u=1}^{M} \left(\bar{U}_v \widehat{A}_v - \bar{U}_u \widehat{A}_u\right)^t \mathcal{L}_U \left(\bar{U}_v \widehat{A}_v - \bar{U}_u \widehat{A}_u\right),$$

$\widehat{A}_v = (a^{(v)}_{c_v(1)}, \ldots, a^{(v)}_{c_v(r)})^t \in \mathbb{R}^r$ and $\widehat{\mathbf{A}} = (\widehat{A}_1^t, \ldots, \widehat{A}_M^t)^t \in \mathbb{R}^{Mr}$. For ease of notation, we consider the same number of basis vectors for each view. It should be noted that the above co-regularization setting is valid if *different* basis vectors are selected for each view.

Given this matrix formulation of our optimization problem, we can follow the framework described in [3] to find a closed form for the solution: Taking the partial derivative of $\widehat{J}(\widehat{\mathbf{A}})$ with respect to $\widehat{A}_v$ we get

$$\frac{d}{d\widehat{A}_v}\widehat{J}(\widehat{\mathbf{A}}) = -2\bar{L}_v^t\mathcal{L}_L(S - \bar{L}_v\widehat{A}_v) + 2\lambda\widehat{K}_v\widehat{A}_v \\ -4\nu\sum_{u=1,u\neq v}^{M}\bar{U}_v^t\mathcal{L}_U(\bar{U}_u\widehat{A}_u - \bar{U}_v\widehat{A}_v).$$

By defining $G_v^\nu = 2\nu(M-1)\bar{U}_v^t\mathcal{L}_U\bar{U}_v$, $G_v^\lambda = \lambda\widehat{K}_v$ and $G_v = \bar{L}_v^t\mathcal{L}_L\bar{L}_v$, we can rewrite the above term as

$$\frac{d}{d\widehat{A}_v}\widehat{J}(\widehat{\mathbf{A}}) = 2(G_v + G_v^\nu + G_v^\lambda)\widehat{A}_v - 2\bar{L}_v^t\mathcal{L}_L S \\ -4\nu\sum_{u=1,u\neq v}^{M}\bar{U}_v^t\mathcal{L}_U\bar{U}_u\widehat{A}_u.$$

At the optimum we have $\frac{d}{d\widehat{A}_v}\widehat{J}(\widehat{\mathbf{A}}) = 0$ for all views, thus we get the exact solution by solving

$$\begin{pmatrix} \bar{G}_1 & -2\nu\bar{U}_1^t\mathcal{L}_U\bar{U}_2\ldots \\ -2\nu\bar{U}_2^t\mathcal{L}_U\bar{U}_1 & \bar{G}_2 & \ldots \\ \vdots & \vdots & \ddots \end{pmatrix}\begin{pmatrix}\widehat{A}_1 \\ \widehat{A}_2 \\ \vdots\end{pmatrix} = \begin{pmatrix}\bar{L}_1^t\mathcal{L}_L S \\ \bar{L}_2^t\mathcal{L}_L S \\ \vdots\end{pmatrix}$$

with respect to $\widehat{A}_1, \ldots, \widehat{A}_M$, where $\bar{G}_v = G_v + G_v^\nu + G_v^\lambda$. The left-hand side matrix is positive definite and therefore invertible (see Appendix). By defining

$$B = \begin{pmatrix} G_1 & 0 & \ldots \\ 0 & G_2 & \ldots \\ \vdots & \vdots & \ddots \end{pmatrix} \quad D = \begin{pmatrix} G_1^\lambda & 0 & \ldots \\ 0 & G_2^\lambda & \ldots \\ \vdots & \vdots & \ddots \end{pmatrix} E = \begin{pmatrix}\bar{L}_1^t\mathcal{L}_L S \\ \bar{L}_2^t\mathcal{L}_L S \\ \vdots\end{pmatrix}$$

$$C = \begin{pmatrix} G_1^\nu & -2\nu\bar{U}_1^t\mathcal{L}_U\bar{U}_2\ldots \\ -2\nu\bar{U}_2^t\mathcal{L}_U\bar{U}_1 & G_2^\nu & \ldots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

we can formulate the solution of the system as follows:

$$\widehat{\mathbf{A}} = (B + C + D)^{-1}E. \tag{18}$$

The computational complexity of constructing the vector E is $\mathcal{O}(Mmr)$. Further, the matrices B, C, and D can be constructed in $\mathcal{O}(Mr^2m)$, $\mathcal{O}(M^2r^2n)$, and $\mathcal{O}(Mr^2)$, respectively. The resulting matrix $(B+C+D) \in \mathbb{R}^{Mr \times Mr}$ can be inverted in $\mathcal{O}(M^3r^3)$. Hence, our algorithm scales linearly in the number of unscored data items. Note that the multiplications involving the Laplacian matrices $\mathcal{L}_L$ and $\mathcal{L}_U$ can be accelerated using the approach described in [22]. Assuming $n \geq m$ we have shown the following theorem:

Theorem 1. *For fixed parameters $\lambda, \nu \in \mathbb{R}^+$ and assuming $n \geq m$, the solution of the Sparse Co-RankRLS optimization problem (16) can be found in $\mathcal{O}(M^3r^3 + M^2r^2n)$ time.*

4.3 Efficient Regularization Parameter Selection

When performing experiments, the recurrent matrix inversion in (18) for each combination of the regularization parameters λ and ν could be time-consuming. Therefore, we propose a procedure which accelerates this parameter selection process. Writing D as $D = \lambda \acute{D}$ with an appropriate (positive definite) matrix $\acute{D}$ and rewriting $\acute{D}$ as $\acute{D} = GG^t$ using the Cholesky decomposition [9], we obtain

$$\begin{aligned}(B+C+D)^{-1} &= (B+C+\lambda\acute{D})^{-1} \\ &= (GG^{-1}(B+C)(G^t)^{-1}G^t + \lambda GG^t)^{-1} \\ &= (G^t)^{-1}(G^{-1}(B+C)(G^t)^{-1} + \lambda I)^{-1}G^{-1}.\end{aligned}$$

Further, the matrix $G^{-1}(B+C)(G^t)^{-1}$ can be eigen decomposed to $V\Lambda V^t$, where Λ is a diagonal matrix containing the eigenvalues and V is matrix composed of the eigenvectors [9]. Hence, we get

$$\begin{aligned}(B+C+D)^{-1} &= (G^t)^{-1}(V\Lambda V^t + \lambda I)^{-1}G^{-1} \\ &= (G^t)^{-1}V(\Lambda + \lambda I)^{-1}V^tG^{-1}\end{aligned}$$

and the solution in (18) can be rewritten as

$$\widehat{\mathbf{A}} = (G^t)^{-1}V(\Lambda + \lambda I)^{-1}V^tG^{-1}E.$$

Thus, by fixing the parameter ν we can efficiently search for the second regularization parameter λ. The decompositions and the inversion of G can be calculated in $\mathcal{O}(M^3r^3)$ time, and hence, the overall training complexity is not increased. The computational cost of calculating $(\Lambda+\lambda I)^{-1}$ is $\mathcal{O}(Mr)$, since it is a diagonal matrix. If the matrices $V^tG^{-1}E \in \mathbb{R}^{Mr \times 1}$ and $(G^t)^{-1}V \in \mathbb{R}^{Mr \times Mr}$ are stored in memory, the subsequent training with different values of λ can be performed in $\mathcal{O}(M^2r^2)$ time.

5 Experiments

We evaluate the performance of the Sparse Co-RankRLS algorithm on the parse ranking task, namely the task of ranking given parses for an unseen sentence. For this purpose, we use the BioInfer corpus [15] which consists of 1100 manually annotated sentences. A detailed description of the parse ranking problem and the data used in the experiments is given in [23]. Each sentence is associated with a set of candidate parses. The manual annotation of the sentence, present in the corpus, provides the correct parse. Further, each candidate parse is associated with a goodness score that indicates how close to the correct parse it is. The correct ranking of the parses associated with the same sentence is determined by this score. While the scoring induces a total order over the whole set of parses, the preferences between parses associated with different sentences are not considered in the parse ranking task.

Using the definitions presented in Section 2, we consider each sentence as an instance and the parses generated for the sentence as the labels associated with it. The score of an input indicates how well the parse included in the input matches the correct parse of the sentence. We have previously demonstrated that the RankRLS algorithm performs comparably to some state-of-the-art ranking methods [11]. In this section, we will compare the performance of the Sparse Co-RankRLS algorithm with that of the RankRLS algorithm.

5.1 Experimental Setup

From the 1100 sentences of the BioInfer corpus we randomly select 600 and 500 sentences for the training and final validation phase, respectively. To simulate a semi-supervised setting, we consider that only 50 sentence-parse pairs in the training set are scored, while the remaining 550 sentences do not have the scoring information associated with them. For the evaluation of the Sparse Co-RankRLS method we set the number M of views to 2. Further, we randomly select 20 sentences and their associated parses from the unscored data set as basis vectors for the first view and repeat this procedure for the second view. According to Section 4 we select different basis vectors for each view.

Both of the algorithms have the regularization parameter λ that controls the tradeoff between the minimization of the training error and the complexity of the learnt function(s). In addition, the Sparse Co-RankRLS algorithm has the regularization parameter ν that controls the agreement between the predictions of the different views. As a similarity measure for parses, we use the best performing graph kernel with the appropriate parameter considered in [12]. The values of the regularization parameters for RankRLS as well as for Sparse Co-RankRLS are estimated during a 10-fold cross-validation procedure, with the splits being performed on the sentence level ensuring that all parses associated with the same sentence are present in the same fold. In the semi-supervised setting each fold consists of one tenth of labeled and unlabeled data present in the training set. For the cross-validation phases, we randomly select 7 parses for each sentence to be associated with it, out of which 2 parses are used for training the model

Standard RankRLS	Sparse Co-RankRLS
0.373	0.344

Table 1. Comparison of the parse ranking performances of the standard RankRLS and the Sparse Co-RankRLS algorithms using a normalized version of the disagreement error (1) as performance evaluation measure.

and 5 for testing. Finally, we use 5 parses per sentence for the final validation procedure.

5.2 Results

The normalized version of the disagreement error (1) is used to measure the performance of the ranking algorithms. The error is calculated for each sentence separately and the performance is averaged over all sentences.

The algorithms are trained on the whole parameter estimation data set with the best found parameter values and tested with the 500 sentences reserved for the final validation. The results of the validation are presented in Table 1. They show that the Sparse Co-RankRLS algorithm notably outperforms the RankRLS method. Here, the results of the Sparse Co-RankRLS algorithm are obtained by averaging the predictions of the two views.

Furthermore, to test the statistical significance of the performance difference between the Sparse Co-RankRLS and RankRLS algorithms, we conduct the Wilcoxon signed-ranks test [7]. The sentences reserved for the final validation are considered as independent trials. We observe that the performance differences are statistically significant ($p < 0.05$).

6 Conclusions

We propose Sparse Co-RankRLS, a semi-supervised regularized least-squares algorithm for learning preference relations. The computational complexity of the algorithm is $\mathcal{O}(M^3r^3 + M^2r^2n)$, where n is the number of unlabeled training examples. We formulate the algorithm within the co-regularization framework, which aims at improving the prediction performance by minimizing the disagreement of all prediction hypotheses on the unlabeled data. In our experiments, we consider a parse ranking task and show that the Sparse Co-RankRLS algorithm significantly outperforms the standard RankRLS algorithm on this task.

Due to the fact that our semi-supervised preference learning algorithm has a linear complexity in the number of unlabeled examples, it is primarily applicable in cases when only a small amount of labeled but a large amount unlabeled data is available for training. In the future, we aim to evaluate our Sparse Co-RankRLS algorithm on various tasks where labeled data is scarce.

Acknowledgments

This work has been supported by Tekes, the Finnish Funding Agency for Technology and Innovation. We would like to thank CSC, the Finnish IT center for science, for providing us extensive computing resources.

References

1. Ravindra B. Bapat and T. E. S. Raghavan. *Nonnegative Matrices and Applications (Encyclopedia of Mathematics and its Applications)*. Cambridge University Press, March 1997.
2. Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100, New York, NY, USA, 1998. ACM.
3. Ulf Brefeld, Thomas Gärtner, Tobias Scheffer, and Stefan Wrobel. Efficient co-regularised least squares regression. In *Proceedings of the 23rd International Conference on Machine learning*, pages 137–144, New York, NY, USA, 2006. ACM.
4. Ulf Brefeld and Tobias Scheffer. Co-em support vector learning. In *Proceedings of the 21st International Conference on Machine learning*, page 16, New York, NY, USA, 2004. ACM.
5. Richard A. Brualdi and Herbert J. Ryser. *Combinatorial Matrix Theory*. Cambridge University Press, 1991.
6. Ofer Dekel, Christopher D. Manning, and Yoram Singer. Log-linear models for label ranking. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 497–504, Cambridge, MA, 2004. MIT Press.
7. Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30, 2006.
8. Johannes Fürnkranz and Eyke Hüllermeier. Preference learning. *Künstliche Intelligenz*, 19(1):60–61, 2005.
9. Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 1996.
10. Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Support vector learning for ordinal regression. In *Proceedings of the Ninth International Conference on Articial Neural Networks*, pages 97–102, London, 1999. Institute of Electrical Engineers.
11. Tapio Pahikkala, Evgeni Tsivtsivadze, Antti Airola, Jorma Boberg, and Tapio Salakoski. Learning to rank with pairwise regularized least-squares. In Thorsten Joachims, Hang Li, Tie-Yan Liu, and ChengXiang Zhai, editors, *SIGIR 2007 Workshop on Learning to Rank for Information Retrieval*, pages 27–33, 2007.
12. Tapio Pahikkala, Evgeni Tsivtsivadze, Jorma Boberg, and Tapio Salakoski. Graph kernels versus graph representations: a case study in parse ranking. In Thomas Gärtner, Gemma C. Garriga, and Thorsten Meinl, editors, *Proceedings of the ECML/PKDD'06 workshop on Mining and Learning with Graphs (MLG'06)*, 2006.
13. Tomaso Poggio and Frederico Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78(9):1481–1497, 1990.
14. Alex Pothen, Horst D. Simon, and Kan-Pu Liou. Partitioning sparse matrices with eigenvectors of graphs. *SIAM Journal on Matrix Analysis Applications*, 11(3):430–452, 1990.

15. Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. BioInfer: A corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8:50, 2007.
16. Ryan Rifkin, Gene Yeo, and Tomaso Poggio. Regularized least-squares classification. In J.A.K. Suykens, G. Horvath, S. Basu, C. Micchelli, and J. Vandewalle, editors, *Advances in Learning Theory: Methods, Model and Applications*, pages 131–154, Amsterdam, 2003. IOS Press.
17. David Rosenberg and Peter L. Bartlett. The rademacher complexity of co-regularized kernel classes. In Marina Meila and Xiaotong Shen, editors, *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, pages 396–403, 2007.
18. Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A generalized representer theorem. In David P. Helmbold and Bob Williamson, editors, *Proceedings of the 14th Annual Conference on Computational Learning Theory*, pages 416–426, London, 2001. Springer.
19. Vikas Sindhwani, Partha Niyogi, and Mikhail Belkin. A co-regularization approach to semi-supervised learning with multiple views. In *Proceedings of ICML Workshop on Learning with Multiple Views*, 2005.
20. Vikas Sindhwani and David Rosenberg. An rkhs for multi-view learning and manifold co-regularization. In Andrew McCallum and Sam Roweis, editors, *Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008)*, pages 976–983, Helsinki, Finland, 2008. Omnipress.
21. Alex J. Smola and Bernhard Schölkopf. Sparse greedy matrix approximation for machine learning. In Pat Langley, editor, *Proceedings of the 17th International Conference on Machine Learning*, pages 911–918, San Francisco, Ca, USA, 2000. Morgan Kaufmann Publishers Inc.
22. Evgeni Tsivtsivadze, Tapio Pahikkala, Antti Airola, Jorma Boberg, and Tapio Salakoski. A sparse regularized least-squares preference learning algorithm. In Anders Holst, Per Kreuger, and Peter Funk, editors, *10th Scandinavian Conference on Artificial Intelligence (SCAI 2008)*, volume 173, pages 76–83. IOS Press, 2008.
23. Evgeni Tsivtsivadze, Tapio Pahikkala, Sampo Pyysalo, Jorma Boberg, Aleksandr Mylläri, and Tapio Salakoski. Regularized least-squares for parse ranking. In A. Fazel Famili, Joost N. Kok, José Manuel Peña, Arno Siebes, and A. J. Feelders, editors, *Advances in Intelligent Data Analysis VI*, pages 464–474. Springer, 2005.

Appendix

We will show that the matrix

$$\begin{pmatrix} \bar{G}_1 & -2\nu\bar{U}_1^t\mathcal{L}_U\bar{U}_2 \ldots \\ -2\nu\bar{U}_2^t\mathcal{L}_U\bar{U}_1 & \bar{G}_2 \quad \ldots \\ \vdots & \vdots \quad \ddots \end{pmatrix}$$

is positive definite. To prove that, we decompose the above matrix into a sum of matrices

$$X_1 = \begin{pmatrix} \bar{G}_1 - 2\nu(M-1)\bar{U}_1^t\mathcal{L}_U\bar{U}_1 & 0 & \dots \\ 0 & \bar{G}_2 - 2\nu(M-1)\bar{U}_2^t\mathcal{L}_U\bar{U}_2 & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

and

$$X_2 = \begin{pmatrix} \nu(M-1)\bar{U}_1^t\mathcal{L}_U\bar{U}_1 & -\nu\bar{U}_1^t\mathcal{L}_U\bar{U}_2 & \dots \\ -\nu\bar{U}_2^t\mathcal{L}_U\bar{U}_1 & \nu(M-1)\bar{U}_2^t\mathcal{L}_U\bar{U}_2 & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}.$$

The matrix X_1 is positive definite as each block matrix is positive definite (we require the matrix $\widehat{K}_v$ to be positive definite). Further, the matrix X_2 is positive semidefinite as we can write it as a sum of positive semidefinite matrices of the form

$$\begin{pmatrix} 0 \cdots & 0 & \cdots & 0 & \cdots 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 \cdots & \nu\bar{U}_i^t\mathcal{L}_U\bar{U}_i & \cdots & -\nu\bar{U}_i^t\mathcal{L}_U\bar{U}_j & \cdots 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 \cdots & -\nu\bar{U}_j^t\mathcal{L}_U\bar{U}_i & \cdots & \nu\bar{U}_j^t\mathcal{L}_U\bar{U}_j & \cdots 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 \cdots & 0 & \cdots & 0 & \cdots 0 \end{pmatrix} = X_{(i,j)}^t X_{(i,j)},$$

where $X_{(i,j)} = \big(0, \dots, 0, \sqrt{\nu}P\bar{U}_i, 0, \dots, 0, -\sqrt{\nu}P\bar{U}_j, 0, \dots, 0\big)$. Here, the positive semidefinite matrix $\mathcal{L}_U$ is decomposed as $\mathcal{L}_U = P^tP$ using the Cholesky decomposition [9].