Large-margin Weakly Supervised Dimensionality Reduction

Chang Xu CHANGXU1989@GMAIL.COM
Key Lab. of Machine Perception (Ministry of Education), Peking University, Beijing 100871, China
Centre for Quantum Computation and Intelligent Systems, University of Technology, Sydney 2007, Australia

Dacheng Tao DACHENG.TAO@UTS.EDU.CN
Centre for Quantum Computation and Intelligent Systems, University of Technology, Sydney 2007, Australia

Chao Xu XUCHAO@CIS.PKU.EDU.CN
Key Lab. of Machine Perception (Ministry of Education), Peking University, Beijing 100871, China

Yong Rui YONGRUI@MICROSOFT.COM
Microsoft Research, No. 5, Dan Ling Street, Haidian District, Beijing 10080, China

Abstract

This paper studies dimensionality reduction in a weakly supervised setting, in which the preference relationship between examples is indicated by weak cues. A novel framework is proposed that integrates two aspects of the large margin principle (*angle* and *distance*), which simultaneously encourage angle consistency between preference pairs and maximize the distance between examples in preference pairs. Two specific algorithms are developed: an alternating direction method to learn a linear transformation matrix and a gradient boosting technique to optimize a non-linear transformation directly in the function space. Theoretical analysis demonstrates that the proposed large margin optimization criteria can strengthen and improve the robustness and generalization performance of preference learning algorithms on the obtained low-dimensional subspace. Experimental results on real-world datasets demonstrate the significance of studying dimensionality reduction in the weakly supervised setting and the effectiveness of the proposed framework.

1. Introduction

High-dimensional data is encountered in many machine learning applications, is difficult to work with, and suffers deterioration in performance due to the "curse of dimensionality". Dimensionality reduction methods are therefore an important and increasingly inherent part of modern data analysis and processing. Working with a new low-dimensional space is more efficient and can often be of advantage for analyzing the intrinsic structure of the data in various applications, including compact data representations (Ye et al., 2004), high-dimensional data visualization (Roweis & Saul, 2000; Tenenbaum et al., 2000), and classification algorithms (Chen et al., 2005).

Conventional supervised dimensionality reduction algorithms employ explicit labels (e.g., the character of a handwritten digit or the identity of a face image) for learning. Here we propose a weakly supervised approach in which preference relationships between examples are extracted from prior information for dimensionality reduction. A weakly supervised approach is preferred to a fully supervised approach, since obtaining sufficient labeled data for large datasets can be expensive, especially when it involves procedures like human hand labeling, clinical trials, or experimentation. In some cases explicit labels that correspond to the data do not even exist; for example, in document retrieval it is difficult to establish whether a document is absolutely relevant or irrelevant with respect to the query without definite criteria. In the weakly supervised setting, we rely on the preference relationships between examples rather than their explicit labels, which can be obtained from both the explicit labels themselves or from other prior information, such as click count on a document indicating relevance to a query.

According to the quantity of supervised information used, existing dimensionality reduction methods can be roughly categorized into unsupervised (Tenenbaum et al., 2000;

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

Roweis & Saul, 2000; Hinton & Roweis, 2002), supervised (Fukumizu et al., 2003; Lacoste-Julien et al., 2008), or semi-supervised methods (Memisevic & Hinton, 2005; Jain et al., 2010; Rai & Daumé III, 2009). Traditionally, principal component analysis (PCA) and kernel PCA (Hastie et al., 2005) function by preserving the global covariance structure of data when the label information is unavailable. Several manifold learning methods, such as Locally Linear Embedding (LLE) (Roweis & Saul, 2000) and ISOMAP (Tenenbaum et al., 2000), perform well without prior information by assuming that the data lie on a low-dimensional manifold and by preserving the data locality. Fisher's Discriminant Analysis (FDA) (Fisher, 1936) and its extensions (Sugiyama, 2006; 2007) are known to work well and are useful in practice since they maximize between-class scatter and minimize within class scatter. By considering that unlabeled examples are readily available and labeled ones are fairly expensive to obtain, Yang *et. al.* (Yang et al., 2006) extended basic LLE and ISOMAP into semi-supervised versions. Kim *et. al.* (Kim et al., 2009) assumed that all data points are concentrated around a low-dimensional submanifold and derived the Hessian energy for semi-supervised dimensionality reduction.

These existing dimensionality reduction algorithms are not suitable for the weakly supervised setting where preference relationships between examples are required. Of relevance to this is preference learning (Fürnkranz & Hüllermeier, 2010; Chu & Ghahramani, 2005; Houlsby et al., 2012), which aims to learn a model that predicts the underlying preference relationships between examples. Here, however, we instead concentrate on how to conduct dimensionality reduction using preference relationships.

This paper proposes a new framework of weakly supervised dimensionality reduction that considers two different aspects of the large margin principle, the *angle level* and *distance level*. In the angle level, an ideal dimensionality reduction algorithm is capable of generating low-dimensional examples whose preference relationships can be linearly predicted. Therefore, the directions of preference vectors corresponding to preference pairs should be consistent. In the distance level, in order to clearly distinguish the preference relationship between two examples in a preference pair, their distance should be maximized. Based on these principles we present two weakly supervised dimensionality reduction algorithms, one of which learns a linear transformation matrix using an alternating direction method, while the other learns a non-linear transformation directly in function space based on the gradient boosting regression tree (GBRT), which offers the advantages of being insensitive to hyper-parameters, robust to overfitting, and scalable. We perform a theoretical analysis of the robustness and generalization error of preference learning algorithms on the obtained low-dimensional subspaces and suggest that using the large margin principle in weakly supervised dimensionality reduction is advantageous for both. Experimental results using both algorithms on real-world datasets demonstrate the value of studying dimensionality reduction in a weakly supervised setting and the effectiveness of the proposed algorithms.

2. Motivation and Notation

Here we take web image searching as an example to illustrate the motivation behind using weakly supervised dimensionality reduction. Click count information encoding users' preferences on each query-image pair can be collected from the search engine and regarded as weakly supervised information. This weakly supervised information is intrinsically different to conventional supervised information; for example, given two images a and b and their click counts 10 and 100 for a specific query q, it would be inappropriate and inaccurate to assert that image b is relevant (positive) to the query and image a irrelevant (negative) simply by setting a threshold to determine the labels. A more reasonable assumption would be that image b is preferable to image a with regard to this query. However, it is not straightforward to introduce this weakly supervised information into conventional supervised dimensionality reduction methods because they require explicit class label information (i.e., the relevant/irrelevant information in web image ranking) for training. We thus propose a novel framework of weakly supervised dimensionality reduction for this setting.

Formally, given a training sample of n examples $\{x_1, \cdots, x_n\}$ where $x \in \mathbb{R}^D$, we use binary matrix $\eta_{ij} \in \{0, 1\}$ to indicate whether there exists a preference relationship between example x_i and example x_j. Moreover, for the preference pairs, if example x_i is preferred to example x_j, i.e., $x_i \succ x_j$, we define $y_{ij} = 1$, otherwise $y_{ij} = -1$. The task is to then learn a transformation $\phi : \mathbb{R}^D \to \mathbb{R}^d$ where $d \ll D$ in order to discover the low-dimensional representations of examples by considering this weak preference information. The framework of weakly supervised dimensionality reduction is formulated from two perspectives, as described above: the *angle level* and the *distance level* (see Figure 1).

Angle level. Considering a preference prediction task, an optimal dimensionality reduction algorithm would be capable of generating low-dimensional examples that are discriminative enough for preference learning (see Figure 2). For simplicity, a linear preference prediction model represented by a weight vector $\mathbf{w}$ is considered in the low-dimensional space. For a preference pair $(\phi(x_i), \phi(x_j))$, indicating that example-i is preferred to example-j, we have $\mathbf{w}(\phi(x_i) - \phi(x_j)) > 0$, which means that the angle between the weight vector $\mathbf{w}$ and preference vector

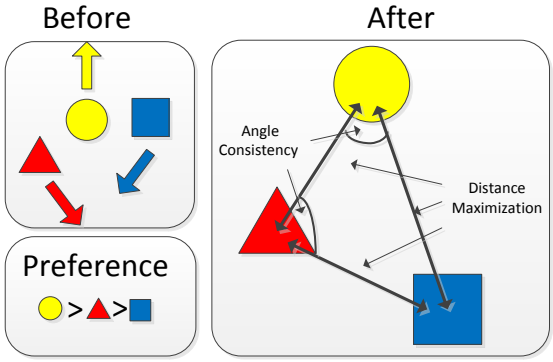

Figure 1. Schematic illustration of the large margin principle in the proposed weakly supervised dimensionality reduction framework.

$(\phi(x_i)-\phi(x_j))$ is in $(0,\pi/2)$. Given more preference pairs we can obtain the similar angle constraints (e.g., the included angles are in $(0,\pi/2)$). In order to accurately predict the preference pairs, the angle constraints between the weight vector $\mathbf{w}$ and all the preference vectors should be maximally satisfied. In other words, the optimal weight vector should be the one whose angles between preference vectors most meet this angle constraint. In dimensionality reduction this optimization principle can be transformed in order to discover the low-dimensional preference vectors whose angles between each other are, on average, small.

Formally, given two preference pairs $p_{ij}=(\phi(x_i),\phi(x_j))$ and $p_{kl}=(\phi(x_k),\phi(x_l))$, the included angle between them can be measured by the cosine function

$$\cos(p_{ij},p_{kl})=\frac{(\phi(x_i)-\phi(x_j))^T(\phi(x_k)-\phi(x_l))}{\|\phi(x_i)-\phi(x_j)\|\|\phi(x_k)-\phi(x_l)\|}, \quad (1)$$

and the corresponding loss can be computed by the hinge loss with margin γ:

$$h(\gamma-y_{ij}y_{kl}\cos(p_{ij},p_{kl}))=\max(\gamma-y_{ij}y_{kl}\cos(p_{ij},p_{kl}),0). \quad (2)$$

This can be seen as an extension of the large margin principle to the angle domain and an approach to improve generalization.

Distance level. Similarly, starting with a prediction weight vector $\mathbf{w}$ and a preference pair $(\phi(x_i),\phi(x_j))$, we have $\delta=\mathbf{w}(\phi(x_i)-\phi(x_j))>0$. The larger δ, the easier it is to distinguish example-i and example-j. The value of δ is largely determined by the distance between $\phi(x_i)$ and $\phi(x_j)$ because the angle level constraint has forced the weight vector to be parallel to the preference vectors. Therefore, in dimensionality reduction it is necessary to maximize the distance between two low-dimensional examples in one preference pair. Formally, in the large margin principle, this distance level constraint can be described by

$$h(\gamma-\|\phi(x_{ij})-\phi(x_j)\|)=\max(\gamma-\|\phi(x_i)-\phi(x_j)\|,0), \quad (3)$$

where γ is the constant margin.

Putting the above angle level and distance level optimization criteria together, we obtain the objective function

$$\min_{\phi}\sum_{ij}\eta_{ij}h(\gamma_1-\|\phi(x_i)-\phi(x_j)\|) +C\sum_{i,j,k,l}\eta_{ij}\eta_{kl}h(\gamma_2-y_{ij}y_{kl}\cos(p_{ij},p_{kl})), \quad (4)$$

where γ_1 and γ_2 are constant margins for two kinds of constraint and C is a positive constant. The first term in Eq. (4) aims to maximize the distance between low-dimensional examples in preference pairs and the second term encourages consistency between the directions of low-dimensional preference vectors. The constant C controls the relative importance of these two competing terms and it is carefully chosen via cross-validation.

3. Large-margin Weakly Supervised Dimensionality Reduction

We first study the linear transformation $L\in\mathbb{R}^{d\times D}$ for weakly supervised dimensionality reduction. The low-dimensional representation of each example can be computed by $\phi(x)=Lx$ and Eq. (4) is reformulated as

$$\min_{L\in\mathbb{R}^{d\times D}}\sum_{ij}\eta_{ij}h(\gamma_1-\|L(x_i-x_j)\|) +C\sum_{i,j,k,l}\eta_{ij}\eta_{kl}h\left(\gamma_2-y_{ij}y_{kl}\frac{(x_i-x_j)^TL^TL(x_k-x_l)}{\|L(x_i-x_j)\|\|L(x_k-x_l)\|}\right), \quad (5)$$

Since problem (5) is composed of two non-smooth objective functions (i.e., the distance level and angle level functions) that are difficult to jointly optimize, we solve these two sub-problems alternately using the alternating direction method.

Since the alternating direction method is usually based on variable splitting combined with the augmented Lagrangian, we initially split the variable L into two and transform the primal problem (5) as

$$\min_{L,K}\{f(L)+g(K)\ :\ L-K=0\}, \quad (6)$$

where $f(\cdot)$ corresponds to the first term in Eq. (5) and $g(\cdot)$ corresponds to the second term in Eq. (5). The augmented Lagrangian related to problem (6) is

$$\mathcal{L}(L,K,\lambda,\mu)=f(L)+g(K)+\langle\lambda,L-K\rangle+\frac{\mu}{2}\|L-K\|^2, \quad (7)$$

where λ is a Lagrangian multiplier related to the equality constraint and μ is a parameter weighting the quadratic penalty. After rearranging the terms, the augmented Lagrangian is rewritten as

$$\mathcal{L}(L,K,\beta)=f(L)+g(K)+\frac{\mu}{2}\|L-K+\beta\|^2, \quad (8)$$

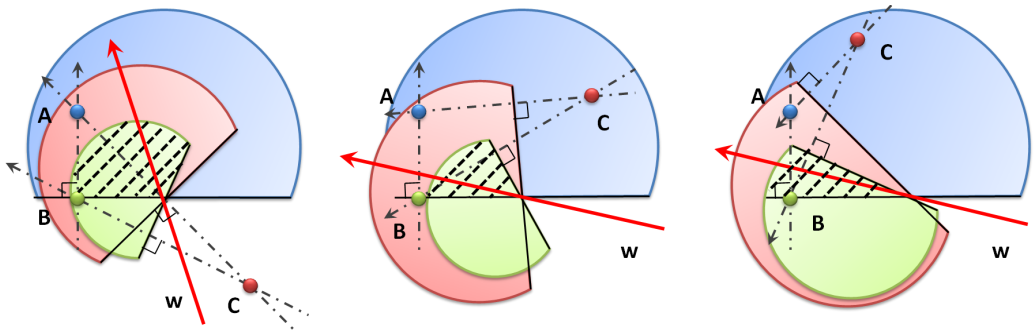

Figure 2. Illustration of the angle constraint. Given three examples A, B and C and their orders $A \succ B \succ C$, we can generate three preference pairs $(A, B), (B, C)$ and (A, C) and their corresponding preference vectors $(A-B), (B-C)$ and $(A-C)$. To ensure that the weight vector $\mathbf{w}$ correctly predicts the preferences of these examples, the direction of $\mathbf{w}$ is constrained in the overlapping orientations of these three vectors, i.e., the shadow regions. Furthermore, from the left to the right of the figure we observe that if the directions of preference vectors are more inconsistent the constraint on $\mathbf{w}$ is more rigorous and it is more difficult to find the optimal weight vector.

where $\beta = \frac{\lambda}{\mu}$. The alternating direction method that solves the original problem (5) looks for a saddle point of the augmented Lagrangian by alternately solving the following problems at iteration t:

$$L^{t+1} = \min_L \mathcal{L}(L, K^t, \beta^t) \quad (9)$$

$$K^{t+1} = \min_K \mathcal{L}(L^{t+1}, K, \beta^t) \quad (10)$$

$$\beta^{t+1} = \beta^t + L^{t+1} - K^{t+1}. \quad (11)$$

All the challenges of the algorithm now reside in the resolution of these three problems.

3.1. Solving problem (9)

The optimization problem related to L can be restated as

$$\min_L \sum_{i,j} \eta_{ij} h(\gamma_1 - \|L(x_i - x_j)\|) + \frac{\mu}{2}\|L - K + \beta\|^2. \quad (12)$$

The most challenging part comes from the non-smooth hinge loss function. Here we apply the smoothing technique introduced by (Nesterov, 2005) to approximate the hinge loss with smooth parameter $\sigma > 0$:

$$h_\sigma = \max_{z \in \mathcal{Q}} z_{ij}(\gamma_1 - \|L(x_i - x_j)\|) - \frac{\sigma}{2}\|x_i - x_j\|_\infty z_{ij}^2$$
$$\mathcal{Q} = \{z : 0 \le z_{ij} \le 1, z \in \mathbb{R}^{n \times n}\}, \quad (13)$$

where z_{ij} can be obtained by setting the gradient of this function as zero and then projecting z_{ij} in $\mathcal{Q}$, i.e.,

$$z_{ij} = \texttt{median}\{\frac{\gamma_1 - \|L(x_i - x_j)\|}{\sigma\|x_i - x_j\|_\infty}, 0, 1\} \quad (14)$$

Therefore, the smoothed hinge loss h_σ is a piece-wise approximation of h according to different choices of z_{ij} in Eq. (14)

$$h_\sigma = \begin{cases} 0 & z_{ij} = 0 \\ \gamma_1 - \|L(x_i - x_j)\| - \frac{\sigma}{2}\|x_i - x_j\|_\infty & z_{ij} = 1 \\ \frac{(\gamma_1 - \|L(x_i - x_j)\|)^2}{2\sigma\|x_i - x_j\|_\infty} & else. \end{cases} \quad (15)$$

whose gradient is calculated by

$$\frac{\partial h_\sigma}{\partial L} = \begin{cases} 0 & z_{ij} = 0 \\ -\frac{L(x_i - x_j)(x_i - x_j)^T}{\|L(x_i - x_j)\|} & z_{ij} = 1 \\ -\frac{(\gamma_1 - \|L(x_i - x_j)\|)}{\sigma\|x_i - x_j\|_\infty} \frac{L(x_i - x_j)(x_i - x_j)^T}{\|L(x_i - x_j)\|} & else. \end{cases}$$

The gradient is now continuous and gradient descent type methods can be efficiently applied to solve the objective function and find the optimal L. Problem (10) can also be similarly solved using this approach by replacing the distance level hinge loss with the angle level hinge loss. Finally, problem (5) can be efficiently optimized using the alternating direction method.

3.2. Non-linear Extension

To handle the non-linear input features we propose a gradient boosting approach to directly optimize the objective function (4) in the function space. In contrast to conventional kernel tricks for non-linear extension, such as by considering $\phi(x) = L\psi(x)$ and the corresponding kernel $K(x_i, x_j) = \phi(x_i)^T\phi(x_j) = \psi(x_i)^T L^T L \psi(x_j)$, we base our non-linear algorithm on the gradient boosting regression tree (GBRT) (Friedman, 2001), which is suitable for large scale applications.

Given the unspecified transformation $\phi(\cdot)$ the resulting objective function (4) becomes

$$\min_\phi \mathcal{J} = \sum_{ij} \eta_{ij} h(\gamma_1 - \|\phi(x_i) - \phi(x_j)\|)$$
$$+C \sum_{i,j,k,l} \eta_{ij}\eta_{kl} h\left(\gamma_2 - y_{ij}y_{kl}\frac{(\phi(x_i) - \phi(x_j))^T(\phi(x_k) - \phi(x_l))}{\|\phi(x_i) - \phi(x_j)\|\|\phi(x_k) - \phi(x_l)\|}\right). \quad (16)$$

Optimization. The transformation $\phi(\cdot)$ learned directly in the function space can be constructed as an ensemble of multivariate regression trees selected by gradient boosting. Formally, we can represent the transformation as an additive function

$$\phi = \phi_0 + \alpha \sum_{t=1}^{T} h_t, \quad (17)$$

where ϕ_0 is the initialization, h_t is an iteratively added regression tree of limited depth p, and α is the learning rate. Instead of training many high variance trees that are then averaged to avoid overfitting, small trees with high bias are better suited for efficient computing and generalization. In each iteration, a new tree to be added is greedily selected to best minimize the objective function on its addition to the ensemble:

$$h_t = \min_h \mathcal{J}(\phi_{t-1} + \alpha h). \tag{18}$$

This optimal tree h_t can be found in the steepest step in the function space by approximating the negative gradient g_t of the objective function $\mathcal{J}(\phi_{t-1})$ with respect to the transformation learned in the previous iteration ϕ_{t-1}. Thus, subgradients are computed with respect to each training example x_i and the tree h_t is approximated in the least squares sense:

$$h_t = \min_h \sum_{i=1}^{n} (g_t(x_i) - h_t(x_i))^2, \tag{19}$$

where

$$g_t(x_i) = \frac{\partial \mathcal{J}(\phi_{t-1})}{\partial \phi_{t-1}(x_i)}. \tag{20}$$

In each iteration the p-th depth tree splits the input space into 2^p regions and a constant vector consequently translates the inputs falling into the same region. Furthermore, in order to improve the efficiency of the solution, we perform region splitting by different threads in parallel during the boosting procedure.

4. Theoretical Analysis

In this section we study the robustness and generalization error of preference learning algorithms on the low-dimensional subspace. We suggest that our proposed large margin optimization criteria are effective for reducing the dimensionality of preference pairs. Throughout the theoretical analysis we consider the examples as preference pairs on the learned low-dimensional space. The loss of the linear preference prediction model is bounded by a constant B. Detailed proofs of the following theoretical results are given in the supplementary materials.

4.1. Robustness Analysis

If a test example and a training example are close to each other then their associated losses are also close. This property is formalized as "robustness" in (Xu & Mannor, 2012), and the precise definition is given below:

Definition 1. *An algorithm $\mathcal{A}$ is $(K, \epsilon(\cdot))$ robust for $K \in \mathbb{N}$ and $\epsilon(\cdot) : \mathcal{Z} \to \mathbb{R}$ if the sample $\mathcal{Z}$ can be partitioned into K disjoint sets, denoted by $\{C_i\}_{i=1}^K$, so that the following holds for all $\mathbf{s} \in \mathcal{Z}$, given the loss function $l(\mathcal{A}_s, z)$ of the algorithm $\mathcal{A}_s$ trained on s:*

$$\forall s \in \mathbf{s}, \forall z \in \mathcal{Z}, \forall i = 1, \cdots, K :$$
$$if\ s, z \in C_i, then\ |l(\mathcal{A}_s, s) - l(\mathcal{A}_s, z)| \leq \epsilon(\mathbf{s}).$$

Given two preference pairs (z_1, z_2) and (s_1, s_2) on the learned low-dimensional subspace through $\phi(\cdot)$, we assume that for both pairs the first example is always preferred to the second example, i.e., $z_1 \succ z_2$ and $s_1 \succ s_2$. When each example in the first pair falls into the same subset of the partition of $\mathcal{Z}$ as the corresponding example of the other pair, the "closeness" between two preference pairs is defined by the angle aspect: if the angle between two preference vectors is small, i.e., $\arccos(z_1 - z_2, s_1 - s_2) \leq \theta$, we suggest that these two preference pairs are close. A preference learning algorithm is said to be robust if two preference pairs are close to each other, i.e., their losses are close. This robustness can be measured by the following theorem.

Theorem 1. *Fix $\theta \geq 0$. For any preference pair (z_1, z_2) in low-dimensional space $\mathcal{Z} \subset \mathbb{R}^d$ that can be partitioned into K disjoint sets, denoted by $\{C_{i=1}\}_{i=1}^K$, assume that $\|z_1 - z_2\| \in [a, b]$. Given a linear preference learning algorithm $\mathcal{A}$ $\{w : z \to \mathbb{R}\}$ and $\|w\| \leq W$, we have for any $\mathbf{s} \subset \mathcal{Z}$:*

$$|\ell(\mathcal{A}_s, z_1, z_2) - \ell(\mathcal{A}_s, s_1, s_2)| \leq W\sqrt{2b^2 - 2a^2\cos(\theta)}$$
$$\forall\, i, j = 1, \cdots, K\ : s_1, z_1 \in C_i \text{ and } s_2, z_2 \in C_j,$$
$$\cos(z_1 - z_2, s_1 - s_2) \geq \cos(\theta).$$

Hence $\mathcal{A}$ is $(K, W\sqrt{2b^2 - 2a^2\cos(\theta)})$-robust.

Robustness is a fundamental property and ensures that a learning algorithm performs well. According to Theorem 1, it is instructive to suggest that if we force the angle between preference vectors to be small, i.e., enlarge $\cos\theta$, or encourage the distance of examples in preference pairs to be large, i.e., increase the lower bound a, the robustness of the preference learning algorithm will be improved. These are exactly the two objectives of the weakly supervised dimensionality reduction framework.

4.2. Generalization Analysis

Based on the robustness analysis we now give a PAC generalization bound for the preference learning algorithm that fulfills the property of robustness. We first present a concentration inequality (Van Der Vaart & Wellner, 1996) that will help us to derive the bound.

Proposition 1. *Let $(|N_1, \cdots, |N_K||)$ be an i.i.d. multinomial random variable with parameters n and $(\mu(C_1), \cdots, \mu(C_K))$. By the Breteganolle-Huber-Carol inequality we have $Pr\{\sum_{i=1}^K |\frac{N_i}{n} - \mu(C_i)| \geq \lambda\} \leq 2^K \exp(\frac{-n\lambda^2}{2})$, hence with probability at least $1 - \delta$,*

$$\sum_{i=1}^{K} |\frac{N_i}{n} - \mu(C_i)| \leq \sqrt{\frac{2K\ln 2 + 2\ln(1/\delta)}{n}}.$$

The generalization error bound for preference learning algorithms is presented in the following theorem.

Theorem 2. *If a preference learning algorithm $\mathcal{A}$ is $(K, \epsilon(\cdot))$-robust and the training sample $\mathbf{s}$ is composed of n preference pairs $\{p_i = (s_1, s_2)\}_{i=1}^n$ whose examples are generated from μ, then for any $\delta > 0$, with probability at least $1 - \delta$, we have,*

$$|\mathcal{L}(\mathcal{A}_s) - \ell_{emp}(\mathcal{A}_s)| \leq \epsilon(s) + 2B\sqrt{\frac{2K \ln 2 + 2\ln(1/\delta)}{n}}.$$

By combining the results of Theorem 1 and Theorem 2 we can easily bound the generalization error of preference learning on the obtained low-dimensional subspace. Robustness is sufficient and necessary for generalization. Discovering the low-dimensional examples based on the distance and angle aspects of the large margin principle can simultaneously strengthen the robustness of preference learning and improve the generalization error

5. Experiments

In this section we qualitatively and quantitatively evaluate the proposed WSDR algorithm on two toys and two real-world applications. The proposed algorithm and its non-linear extension GB-WSDR are compared with Locality Preserving Projections (LPP, He & Niyogi, 2004), Local Fisher's Discriminant Analysis (LFDA, Sugiyama, 2006) and Large Margin Component Analysis (LMCA, Torresani & Lee, 2007). The effectiveness of the learned low-dimensional examples is assessed using the ranking error from RankSVM (Joachims, 2002).

5.1. Toy Examples

We first conducted an experiment using data samples from the UMIST face dataset (Graham & Allinson, 1998). Suppose that we have three kinds of face images: a man with glasses, a man without glasses, and a woman without glasses; we are required to find the face images of a man with glasses. Clearly, these three classes can be ranked as follows: the images of a man with glasses are ranked first since they are an absolute match; the images of a man without glasses are ranked second since they partially match; and the images of a woman without glasses are ranked last since they do not match at all. The dataset used in this experiment is composed of these three classes of human faces from UMIST face dataset, with 48, 25 and 19 images in rank 1, 2 and 3, respectively.

In Figure 3 (a) we show the 2D subspace discovered by the WSDR algorithm. It is clear that the angles between different low-dimensional preference pairs are very close due to the angle constraint through the large margin principle. On the other hand, the two examples in a particular preference pair are obviously separated, when considering the large margin principle from the distance level. Hence, based on this optimal subspace, it would be a trivial task to obtain a linear function that accurately predicts their preference relationships.

The second toy example is based on the USPS digit dataset (Hull, 1994), which is composed of 16×16 grayscale images of hand written digital characters from 0 to 9. We sampled 20 examples from five classes (0 to 4) and treated the true digits shown in the images as their corresponding weak labels. Therefore, the preference relationships between examples could be determined by comparing their attached digits. The low-dimensional subspace for these digit images was generated by GB-WSDR using a non-linear approach and the results are presented in Figure 3 (b).

In Figures 3 (b) and (c) it can be seen that the low-dimensional examples of different labels are separated from each other, while the examples bearing the same labels are closely distributed. The low-dimensional examples generated by LMCA are suitable for multi-class classification, whereas it is difficult to find a linear function that predicts their preference relationships. On the other hand, WSDR extracts the preference relationship from the weak labels and then integrates it into the large margin principle from the distance and angle aspects simultaneously. The preference relationships of examples are therefore preserved in the low-dimensional subspace and it can be expected that preference learning on this subspace would be more accurate.

5.2. Collaborative Filtering

We next present results on a collaborative filtering task for movie recommendations on the MovieLens dataset [1] which contains approximately 1 million ratings for 3592 movies by 6040 users; ratings are on a scale of 1 to 5. For each user, a different predictive model was derived. We used 70% of the movies rated by each user for training and the remaining 30% for testing. The features for each movie consisted of the ranking provided by D reference users. Missing rating values in the input features were populated with the median rating score of the given reference users. Figure 4 shows the pairwise accuracies of different algorithms:

$$\frac{\sum_{i,j} 1_{y_i > y_j \& w(\phi(x_i) - \phi(x_j)) > 0}}{\sum_{i,j} 1_{y_i > y_j}}. \quad (21)$$

In each subfigure in Figure 4, the horizontal axis corresponds to the number of the reference user (i.e., the dimension D of the original feature x), while the vertical axis indicates the dimension d of the projected examples. We find that both WSDR and its non-linear extension GB-WSDR outperform other competitors on nearly all different d and D settings. This is mainly due to the fact that WSDR is constructed by considering the preference relationships of

[1] http://grouplens.org/datasets/movielens/

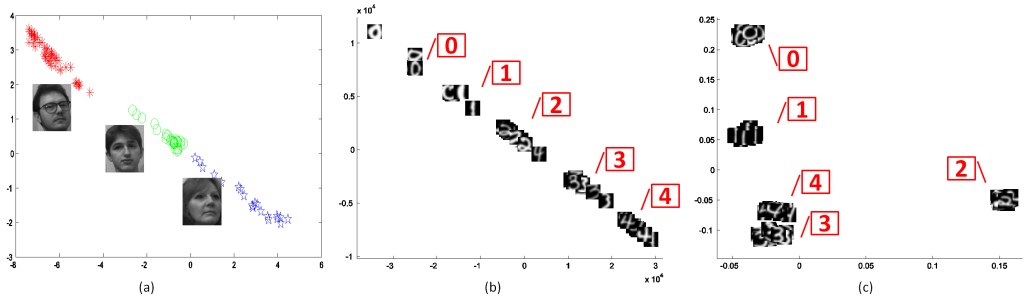

Figure 3. (a) The 2D subspace discovered by WSDR on the UMIST dataset and the 2D subspaces discovered by the GB-WSDR (b) and LMCA (c) algorithms on the USPS dataset.

examples and its elaborate use of the large margin principle to improve performance.

The book-crossing dataset [2] contains 278,858 users and 1,149,780 ratings for 271,379 books. The low density of ratings makes predictions very noisy. Thus, we required users to have reviewed at least 200 books and then kept those books with at least 10 reviews; a dataset of 89 books and 131 users remained. For this dataset, each of the 131 users was selected in turn as a test user and the other 130 users served as input features. The pairwise accuracies for different algorithms are reported for these 131 leave-one-out experiments in Figure 5 (a).

Both LFDA and LMCA achieve satisfactory experimental results based on Fisher's principle and the large margin principle, respectively, but their performance is limited because of their inaccurate interpretation of the weakly supervised labels. Instead of straightforwardly using these labels, WSDR infers pairwise relationships from them and then exploits a general large margin principle to formulate the dimensionality reduction problem. Therefore, the preference learning on the low-dimensional examples generated by WSDR is more accurate. The alternating direction method reduces the burden of solving the combination of two complex non-smooth functions in WSDR, and gradient boosting is an ideal technique for extending WSDR into a non-linear version. Both of these two optimization methods are effective and converge fast. Figures 5 (b) and (c) show the convergence curves for the WSDR and GB-WSDR algorithms on the book-crossing dataset.

5.3. Ranking

Three datasets [3] (Table 1) were used that have previously been used to evaluate ranking and ordinal regression (Fung et al., 2006). Since the target values are continuous, we discretized them into S equal sized bins (Table 1). The pairwise accuracies for different algorithms were evaluated on each dataset in a five-fold cross-validation experiment; these results are presented in Table 2. Both WSDR and GB-WSDR stably outperform other competitors. For example, when the low-dimension is set as five on the Triazines dataset, the baseline algorithms have pairwise accuracies of around 60%, whereas the accuracy of WSDR exceeds 70% and GB-WSDR obtains a further performance improvement via the non-linear approach.

[2] http://grouplens.org/datasets/book-crossing/

[3] http://www.dcc.fc.up.pt/ ltorgo/Regression/DataSets.html

Table 1. Datasets used in the ranking experiments. N is the size of the dataset, D is the number of attributes, S is the number of classes, and M is the average total number of pairwise relations per fold of the training set.

DATA SET	N	D	S	M
PYRIMIDINES	74	28	3	1113
TRIAZINES	186	61	4	7674
WISCONSIN BREAST CANCER	194	33	4	8162

We next sampled a set of queries from the query logs of a commercial search engine and generated a certain number of query-image pairs for each of the queries. In total, $10,000$ queries and $41,021$ query-image pairs were available with click counts greater than zero. From this dataset, we randomly sampled a test set containing about $2,000$ queries and $9,300$ query-image pairs and used the remaining data for model training. Each query-image pair was represented as the combination of a 4000-dim PHOW feature of the image and a 7394-dim BOW feature of the text query. For WSDR and GB-WSDR, preference relationships could easily be inferred from the click counts for the images in the training set and used as the weakly supervised information for dimensionality reduction and rank model learning. For the comparison algorithms we assumed that an image was relevant to a query if its click count was greater than a particular threshold; otherwise, it was deemed irrelevant. We manually labeled the query-image pairs in the test set for ranking performance evaluation.

Figure 6 summarizes the performances of RankSVM in mAP on the low-dimensional subspaces learned by the five algorithms over different-dimensional subspaces. WSDR and GB-WSDR produce considerably better ranking accuracies than the other algorithms on the dataset, and the gradient boosting approach always outperforms the linear approach. The difference in accuracy between our algorithms and the competitors is particularly dramatic when

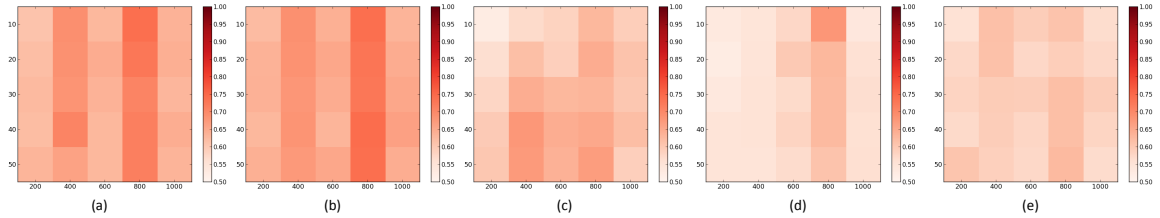

Figure 4. Pairwise accuracies of different algorithms on the MovieLens dataset: (a) GB-WSDR, (b) WSDR, (c) LMCA, (d) LFDA, and (e) LPP.

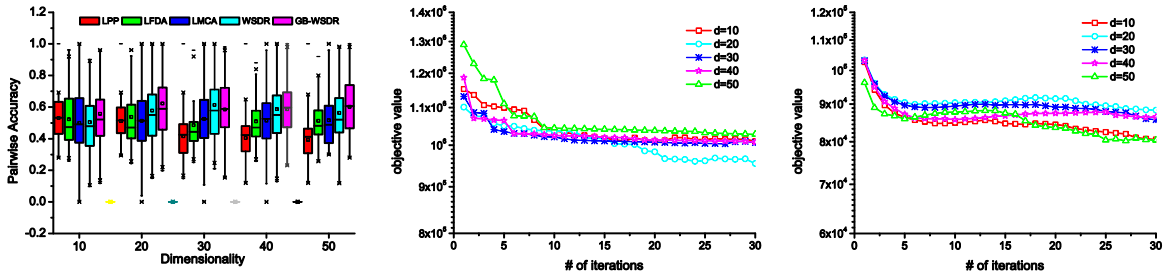

Figure 5. (a) Pairwise accuracies of different algorithms and convergence curves for the WSDR (b) and GB-WSDR (c) algorithms on the book-crossing dataset.

Table 2. Pairwise accuracies of different algorithms on three datasets

	d=5			d=10			d=15			d=20		
	Pyrim	Triazines	Wiscoin	Pyrim	Triazines	Wiscoin	Pyrim	Triazines	Wiscoin	Pyrim	Triazines	Wiscoin
LPP	78.93± 1.87	63.93± 0.42	63.31± 0.30	82.85± 2.10	69.47± 1.84	64.82± 0.51	83.33± 2.15	70.82± 2.56	64.37± 0.33	83.66± 2.83	71.36± 2.72	64.17± 0.41
LFDA	88.68± 0.42	56.18± 0.52	65.41± 1.02	88.49± 0.74	57.48± 1.83	65.34± 0.46	88.60± 0.67	58.57± 1.39	67.36± 0.92	88.23± 1.29	53.57± 1.01	**70.85± 2.53**
LMCA	88.10± 1.58	67.01± 5.59	67.01± 6.24	89.17± 1.35	72.29±3.19	66.88± 6.46	87.85± 1.79	71.97± 1.89	69.47± 6.90	88.24± 1.40	71.90± 1.67	69.48± 4.97
WSDR	**90.83± 1.69**	72.67± 2.26	69.40± 2.12	91.10± 2.00	72.62± 1.33	**69.71± 1.36**	90.44± 1.85	72.73± 2.03	69.52± 1.73	90.16± 1.28	73.13± 3.82	69.42± 1.58
GB-WSDR	90.00± 1.22	**74.11± 2.23**	**69.89± 0.23**	**91.31± 1.13**	**74.08± 1.67**	69.64± 0.11	**90.87± 0.82**	**74.32± 1.85**	**69.53± 1.86**	**90.81± 1.56**	**76.48± 1.29**	69.57± 1.49

a small number of projection dimensions is used. In such cases, LPP, LFDA and LMCA cannot find appropriate low-dimensional subspaces because they are not concerned with the subsequent ranking task or influenced by the noise from inferring explicit labels from click count information. In contrast, WSDR and GB-WSDR solve the low-dimensional subspaces by optimizing the ranking-related objective function of Eq. (4) and therefore achieve stable performance even when projecting onto a very low-dimensional subspace.

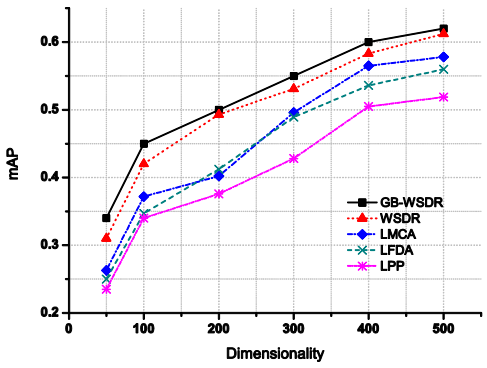

Figure 6. mAP scores on different low-dimensional subspaces.

6. Conclusion

In this paper we study the weakly supervised dimensionality reduction problem, where the preference relationships between examples are provided rather than explicit class labels. By extending the large margin principle into the angle domain we encourage angle consistency between preference pairs while simultaneously maximizing the distance between the two examples in one particular preference pair. Theoretical analysis of these two objectives show that they are beneficial for strengthening robustness and improving the generalization error bound of preference learning algorithms on the obtained low-dimensional subspace. We introduce two new practical algorithms: a linear algorithm for learning a transformation matrix and a non-linear algorithm utilizing the gradient boosting technique to learn the transformation directly in the function space. Both algorithms are efficiently optimized and demonstrate promising experimental results on real-world datasets.

7. Acknowledgements

The work was supported in part by Australian Research Council Projects FT-130101457 and DP-140102164, NBRPC 2011CB302400, NSFC 61121002, 61375026 and JCYJ 20120614152136201.

References

Chen, Hwann-Tzong, Chang, Huang-Wei, and Liu, Tyng-Luh. Local discriminant embedding and its variants. In *CVPR*, 2005.

Chu, Wei and Ghahramani, Zoubin. Preference learning with gaussian processes. In *ICML*, 2005.

Fisher, Ronald A. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

Friedman, Jerome H. Greedy function approximation: a gradient boosting machine. *Ann. Statist*, 29(5):1189–1232, 2001.

Fukumizu, Kenji, Bach, Francis R, and Jordan, Michael I. Kernel dimensionality reduction for supervised learning. In *NIPS*, 2003.

Fung, Glenn, Rosales, Romer, and Krishnapuram, Balaji. Learning rankings via convex hull separation. *NIPS*, 2006.

Fürnkranz, Johannes and Hüllermeier, Eyke. *Preference learning*. Springer, 2010.

Graham, D and Allinson, N. Characterizing virtual eigensignatures for general purpose face recognition, face recognition from theory to applications, ed. h. wechsler, et al. *Computer and Systems Sciences*, 1998.

Hastie, Trevor, Tibshirani, Robert, Friedman, Jerome, and Franklin, James. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

He, Xiaofei and Niyogi, Partha. Locality preserving projections. In *NIPS*, 2004.

Hinton, Geoffrey and Roweis, Sam. Stochastic neighbor embedding. In *NIPS*, 2002.

Houlsby, Neil, Hernandez-Lobato, Jose Miguel, Huszar, Ferenc, and Ghahramani, Zoubin. Collaborative gaussian processes for preference learning. In *NIPS*, 2012.

Hull, Jonathan J. A database for handwritten text recognition research. *TPAMI*, 16(5):550–554, 1994.

Jain, Prateek, Kulis, Brian, and Dhillon, Inderjit S. Inductive regularized learning of kernel functions. In *NIPS*, 2010.

Joachims, Thorsten. Optimizing search engines using clickthrough data. In *SIGKDD*, 2002.

Kim, Kwang In, Steinke, Florian, and Hein, Matthias. Semi-supervised regression using hessian energy with an application to semi-supervised dimensionality reduction. In *NIPS*, 2009.

Lacoste-Julien, Simon, Sha, Fei, and Jordan, Michael I. Disclda: Discriminative learning for dimensionality reduction and classification. In *NIPS*, 2008.

Memisevic, Roland and Hinton, Geoffrey. Multiple relational embedding. In *NIPS*, 2005.

Nesterov, Yu. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

Rai, Piyush and Daumé III, Hal. Multi-label prediction via sparse infinite cca. In *NIPS*, 2009.

Roweis, Sam T and Saul, Lawrence K. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500): 2323–2326, 2000.

Sugiyama, Masashi. Local fisher discriminant analysis for supervised dimensionality reduction. In *ICML*, 2006.

Sugiyama, Masashi. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *The Journal of Machine Learning Research*, 8:1027–1061, 2007.

Tenenbaum, Joshua B, De Silva, Vin, and Langford, John C. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Torresani, Lorenzo and Lee, Kuang-chih. Large margin component analysis. In *NIPS*, 2007.

Van Der Vaart, Aad W and Wellner, Jon A. *Weak Convergence*. Springer, 1996.

Xu, Huan and Mannor, Shie. Robustness and generalization. *Machine learning*, 86(3):391–423, 2012.

Yang, Xin, Fu, Haoying, Zha, Hongyuan, and Barlow, Jesse. Semi-supervised nonlinear dimensionality reduction. In *ICML*, 2006.

Ye, Jieping, Janardan, Ravi, and Li, Qi. Gpca: an efficient dimension reduction scheme for image compression and retrieval. In *SIGKDD*, 2004.