Learning from Noisy Side Information by Generalized Maximum Entropy Model

Tianbao Yang YANGTIA1@CSE.MSU.EDU
Rong Jin RONGJIN@CSE.MSU.EDU
Anil K. Jain JAIN@CSE.MSU.EDU
Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, 48824, USA

Abstract

We consider the problem of learning from noisy side information in the form of pairwise constraints. Although many algorithms have been developed to learn from side information, most of them assume perfect pairwise constraints. Given the pairwise constraints are often extracted from data sources such as paper citations, they tend to be noisy and inaccurate. In this paper, we introduce the generalization of maximum entropy model and propose a framework for learning from noisy side information based on the generalized maximum entropy model. The theoretic analysis shows that under certain assumption, the classification model trained from the noisy side information can be very close to the one trained from the perfect side information. Extensive empirical studies verify the effectiveness of the proposed framework.

1. Introduction

Learning from side information has been studied extensively and has found its application in distance metric learning (Xing et al., 2003), kernel learning (Hoi et al., 2007), and constrained clustering (Basu et al., 2004b). The side information is usually cast in the form of pairwise constraints, including the pairs in the same class, called the *positive (pairwise) constraints*, and the pairs in different classes, called the *negative (pairwise) constraints*. The side information can often be derived from data, making it more attractive than the standard setup of supervised learning. For instance, in classifying research articles, we can derive the pairwise constraints based on the citations between papers.

Although various algorithms have been proposed for learning from side information, most of them assume *perfect* side information. In contrast, in this study, we focus on the problem of **learning from noisy side information** in which some of the pairwise constraints are labeled incorrectly. This is important because the pairwise constraints extracted from data tend to be noisy and inaccurate. In the example of classifying research articles with pairwise constraints constructed from paper citations, the cited paper may not share the same research topic as the citing paper.

In order to handle the noisy side information, we introduce the generalization of maximum entropy model and propose a framework for learning from noisy side information based on the generalized maximum entropy model. We show that under certain assumptions, the conditional probabilistic model trained from the noisy side information converges to that trained from the perfect side information. Extensive experimental results verify the efficacy of the proposed framework for learning from noisy side information.

The remainder of this paper is organized as follows. In section 2, we review the related work. In section 3, we present the generalized maximum entropy learning from noisy side information. We present experimental results in section 4, and conclude our study in section 5.

2. Related Work

In this section, we review the related work of learning from side information with its application to distance metric learning, kernel learning, and constrained clustering. We also discuss the relation between this work and previous work on learning from noisy information.

Appearing in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010.

2.1. Learning from Side Information

The objective of learning from side information is to learn a statistical model that is consistent with the given pairwise constraints. There are three major applications of learning from side information, i.e., distance metric learning, kernel learning, and constrained clustering. Below we briefly review each of the three applications.

Distance metric learning The objective is to learn a distance metric so that data points in positive constraints are separated by a short distance while data points in negative constraints are separated by a large distance. Many algorithms have been developed for distance metric learning, such as distance metric learning by convex optimization (Xing et al., 2003), relevance component analysis (Shental et al., 2002), discriminative component analysis (Hoi et al., 2006a), nearest neighbor component analysis (Goldberger et al., 2004), local distance metric learning (Yang et al., 2006), large margin nearest neighbor classifier (Weinberger et al., 2006), information theoretic metric learning (Davis et al., 2007), and learning a Bregman distance function (Wu et al., 2009). More work on distance metric learning with side information can be found in the survey (Yang & Jin, 2006) and references therein.

Kernel learning The objective is to learn an appropriate kernel matrix/function for a given data set from a set of pairwise constraints. Kernel learning methods can be broadly classified into two categories: parametric kernel learning and nonparametric kernel learning. The parametric methods make parametric assumption about the target kernel function. Exemplar approaches include cluster kernels (Chapelle et al., 2003), diffusion kernel (Kondor & Lafferty, 2002), Gaussian random field (Zhu et al., 2003), spectral kernel (Zhu et al., 2005; Zhang & Ando, 2006), and kernel alignment (Cristianini et al., 2001). For nonparametric approaches, the goal is to find a kernel matrix, instead of a kernel function, that is consistent with pairwise constraints (Cristianini et al., 2001; Lanckriet et al., 2002; Hoi et al., 2007). These approaches are usually computationally expensive because they often require solving a Semi-Definite Programming(SDP) problem.

Constrained clustering The objective is to improve the accuracy of data clustering by exploiting the pairwise constraints. Based on how the constraints are used, the constrained clustering methods can be cast into three categories: constrained-based (Wagstaff et al., 2001; Davidson & Ravi, 2005; Basu et al., 2004a), distance-based (Bilenko & Mooney, 2003; Klein et al., 2002; Cohn et al., 2003), and a mixture of these two (Basu et al., 2004b). More work on constrained clustering with side information can be found in the (Davidson & Sugato, 2007; Liu et al., 2007), and the references therein.

Most studies on learning from side information assume side information is noise-free. However, in many applications, the pairwise constraints are derived from data, making them prone to errors. Although some work (Basu et al., 2004b; Pelleg & Baras, 2007) claims that their approaches are robust to noise in side information, they do not have systematic approaches for handling noisy constraints. In contrast, we present a principled framework for learning from noisy side information based on the generalized maximum model. We also provide theoretic analysis to further justify the proposed approach for learning from noisy side information.

2.2. Learning from Noisy Label Information

Learning from noisy label information (not noisy pairwise constraints) was considered in several studies. (Lawrence & Schölkopf, 2001) estimates a kernel Fisher Discriminant in the presence of label noise. Their work is based on two assumptions: (a) a Gaussian distribution for the input patterns, and (b) a known noise model for the corruption of class labels. (Pal et al., 2007) presents a probabilistic model for extracting location information for events by using the noisy training labels. Our work differs from these studies in that we deal with noisy pairwise constraints, not noisy label information. In addition, we present theoretic analysis showing that under certain assumption, the solution found by our algorithm using noisy side information converges to the one trained from perfect side information.

Finally, it is worthwhile pointing out that our approach is closely relate to the learning framework based on divergence minimization (Altun & Smola, 2006). However, unlike the divergence minimization framework that is designed for the standard setup of supervised learning, our work focuses on learning from noisy side information.

3. Learning from Noisy Side Information

We start with the basic formulation for maximum entropy learning from perfect side information, followed by its generalization. We then extend the generalized maximum entropy learning to the case of noisy side

information. For the purpose of presentation, we first introduce the notations that are used throughout this article.

3.1. Notations

Let $\mathcal{D} = \{\mathbf{x}_i \in \mathcal{X}, i = 1, \cdots N\}$ be a collection of data points, $\mathcal{P} = \{(\mathbf{x}_i^1, \mathbf{x}_i^2, \widehat{y}_i) | \mathbf{x}_i^1, \mathbf{x}_i^2 \in \mathcal{D}, i = 1, \ldots, n, \widehat{y}_i \in \{+1, -1\}\}$ be a collection of observed labeled pairs. We slightly abuse the terminology of labeled and unlabeled examples by referring to the examples in $\mathcal{D}$ that also occur in $\mathcal{P}$ as labeled examples, and to the remaining examples in $\mathcal{D}$ as unlabeled examples. We denote by y_i the true label for the pair $(\mathbf{x}_i^1, \mathbf{x}_i^2)$. We refer to the pairs with $y_i = +1$ as *perfect positive constraints* and the pairs with $y_i = -1$ as *perfect negative constraints*. Similarly, we refer to the pairs with $\widehat{y}_i = +1$ as *noisy positive constraints* and the pairs with $\widehat{y}_i = -1$ as *noisy negative constraints*. We use $\bar{y} = -y$ for complement of y. We use $K_j(\mathbf{x}^1, \mathbf{x}^2)$ for the $j^{\text{th}} \in \{1, \cdots, m\}$ feature function defined on $\mathcal{X} \times \mathcal{X}$. We denote by $\mathbf{k}_i = (K_1(\mathbf{x}_i^1, \mathbf{x}_i^2), \cdots, K_m(\mathbf{x}_i^1, \mathbf{x}_i^2))^\top$ the feature vector for pair $(\mathbf{x}_i^1, \mathbf{x}_i^2)$. Throughout the paper, we use capital letters $X, Y, \widehat{Y}$ for the corresponding random variables. We define $\Pr(\widehat{Y} = y | Y = y) = c_y$, and use $c_+ = c_{+1}, c_- = c_{-1}$ for short. Also, in the sequel, we use the following notations:

$$a_\delta^j[y] = \frac{1}{n} \sum_{i=1}^{n} \delta(y_i, y) K_j(\mathbf{x}_i^1, \mathbf{x}_i^2)$$

$$\widehat{a}_\delta^j[y] = \frac{1}{n} \sum_{i=1}^{n} \delta(\widehat{y}_i, y) K_j(\mathbf{x}_i^1, \mathbf{x}_i^2)$$

where $\delta(y_i, y)$ is the Kronecker delta function that outputs 1 if $y_i = y$ and zero, otherwise.

3.2. Learning from Perfect Side Information by Generalized Maximum Entropy Model

We first consider the maximum entropy model for learning from perfect side information. We cast the problem of learning from side information into a binary classification problem where the objective is to classify each pair $(\mathbf{x}_i^1, \mathbf{x}_i^2)$ into two categories, i.e., a positive pair ($y_i = +1$) and a negative pair ($y_i = -1$). Using maximum entropy model, we aim to learn the conditional distribution $\Pr(Y = y | X^1, X^2)$, which leads to the following optimization problem:

$$\begin{aligned} \max \quad & \sum_{i=1}^{n} H(p | \mathbf{x}_i^1, \mathbf{x}_i^2) \qquad (1) \\ s.t. \quad & \frac{1}{n} \sum_{i=1}^{n} p(y | \mathbf{x}_i^1, \mathbf{x}_i^2) K_j(\mathbf{x}_i^1, \mathbf{x}_i^2) = a_\delta^j[y], \forall y, j \end{aligned}$$

where $H(p | \mathbf{x}_i^1, \mathbf{x}_i^2) = -\sum_y p(y | \mathbf{x}_i^1, \mathbf{x}_i^2) \ln p(y | \mathbf{x}_i^1, \mathbf{x}_i^2)$. The solution to (1) is given by

$$p(y | \mathbf{x}_i^1, \mathbf{x}_i^2) = \frac{1}{1 + \exp(-y \lambda^\top \mathbf{k}_i)}$$

where $\lambda \in \mathbb{R}^m$ are the dual variables and are obtained by solving the following optimization problem,

$$\min_{\lambda \in \mathbb{R}^m} \sum_{i=1}^{n} \ln \left(1 + \exp(-y_i \lambda^\top \mathbf{k}_i)\right)$$

One major problem with the maximum entropy model in (1) is the equality constraint, which is unlikely to hold if for each pair $(\mathbf{x}_i^1, \mathbf{x}_i^2)$, y_i is a random sample from the distribution $p(y | \mathbf{x}_i^1, \mathbf{x}_i^2)$. We denote by $a_p^j[y]$ the left side of equality constraint in problem (1), i.e.

$$a_p^j[y] = \frac{1}{n} \sum_{i=1}^{n} p(y | \mathbf{x}_i^1, \mathbf{x}_i^2) K_j(\mathbf{x}_i^1, \mathbf{x}_i^2)$$

The following theorem shows that $a_p^j[y]$ and $a_\delta^j[y]$ could differ significantly if n is small. The difference between the two quantities will diminish only when n approaches infinity.

Theorem 1. *Assume $(\mathbf{x}_i^1, \mathbf{x}_i^2, y_i)$ are i.i.d. samples from an unknown distribution $P(X^1, X^2, Y)$, the equality constraint in (1) for any j and y holds with probability 1 when the number of instances approaches infinity. In particular, for any $\epsilon > 0$ we have*

$$\Pr\left(\left|a_p^j[y] - a_\delta^j[y]\right| \geq \epsilon\right) \leq 4 \exp\left(-\frac{\epsilon^2 n}{8\kappa_j^2}\right)$$

where $\kappa_j = \max_{\mathbf{x}^1, \mathbf{x}^2} |K_j(\mathbf{x}^1, \mathbf{x}^2)|$.

The theorem can be proved by noting that $\mathrm{E}[a_\delta^j[y]] = \mathrm{E}[a_p^j[y]]$ and applying McDiarmid's inequality. Details are provided in the supplementary material. To address the case that $a_p^j[y]$ and $a_\delta^j[y]$ could be different, we propose a generalization to the traditional maximum entropy model in (1). Given the finite number of training data, we relax the equality constraints in (1) into inequality ones, leading to the following formulation for learning from side information

$$\begin{aligned} \max \quad & \frac{1}{n} \sum_{i=1}^{n} H(p | \mathbf{x}_i^1, \mathbf{x}_i^2) - \frac{1}{2\gamma} \sum_y \|\epsilon_y\|^2 \qquad (2) \\ s.t. \quad & \frac{1}{n} \sum_i p(y | \mathbf{x}_i^1, \mathbf{x}_i^2) K_j(\mathbf{x}_i^1, \mathbf{x}_i^2) \geq a_\delta^j[y] - \epsilon_{yj}, \ \forall y, j \end{aligned}$$

where $\epsilon_y = (\epsilon_{y1}, \ldots, \epsilon_{ym})^\top$ and $\|\cdot\|$ is a norm that measures the length of vector ϵ_y. The key features of the generalized maximum entropy model in (2) are:

- Replacing equality constraints with inequality ones. As a result, we have

$$a^j_\delta[y] - \epsilon_{yj} \le a^j_p[y] \le a^j_\delta[y] + \epsilon_{\bar{y}j}.$$

  Note that although only one side inequality is included in (2), the upper bound of $a^j_p[y]$ can be easily derived by using the relation $a^j_p[y] + a^j_p[\bar{y}] = a^j_\delta[y] + a^j_\delta[\bar{y}]$.
- The positive dummy variables ϵ are introduced to account for the difference between the two empirical means $a^j_p[y]$ and $a^j_\delta[y]$. A regularization term $\|\epsilon_y\|^2/(2\gamma)$ is introduced into the objective in order to determine these variables automatically.

We further justify the generalized maximum entropy model by showing it is equivalent to the regularized logistic regression model.

Proposition 2. *When $\|\cdot\| = \|\cdot\|_2$, the dual problem of (2) is equivalent to the regularized logistic regression model, i.e.,*

$$\min_{\lambda\in\mathbb{R}^d} \frac{\gamma}{2}\|\lambda\|_2^2 + \frac{1}{n}\sum_{i=1}^n \ln\left(1+\exp(-y_i\lambda^\top \mathbf{k}_i)\right)$$

3.3. Learning from Noisy Side Information by Generalized Maximum Entropy Model

We extend the framework of generalized maximum entropy learning to the case when pairwise constraints are noisy, i.e., $\widehat{y}_i \neq y_i$ for some pairs. The advantage of extending the maximum entropy model for noisy side information is that the label information y_i is only utilized in computing $a^j_\delta[y]$. Hence, if we have an alternative approach to estimate $a^j_\delta[y]$ without having to know which labels are incorrect, we can construct the maximum entropy model for noisy side information.

In order to estimate $a^j_\delta[y]$ in the case of noisy side information, we make the following assumptions.

Assumption 1. *We assume (1.a)* $\Pr(\widehat{Y}|X^1, X^2, Y) = \Pr(\widehat{Y}|Y)$, *(1.b)* $\Pr(\widehat{Y} = y|Y = y) = c_y$ *is given and (1.c)* $c_y + c_{\bar{y}} - 1 > 0$.

In the above assumption, (1.a) assumes $\widehat{Y}$ is conditionally independent of (X^1, X^2) given Y, (1.b) assumes the group-level knowledge about the noise in the pairwise constraints and (1.c) essentially assumes that the noise level of the side information is not too significant. With these three assumptions, the following theorem shows that it is possible to express empirical mean $a^j_\delta[y]$ in terms of $\widehat{a}^j_\delta[y]$, i.e., the empirical mean estimated from the noisy side information.

Theorem 3. *Assume $(\mathbf{x}^1_i, \mathbf{x}^2_i, \widehat{y}_i), i = 1, \ldots, n$ are i.i.d samples, then for any $\epsilon > 0$ we have*

$$\Pr\left(\left|a^j_\delta[y] - \widehat{b}^j_\delta[y]\right| \ge \epsilon\right) \le 4\exp\left(-\frac{\epsilon^2(c_y + c_{\bar{y}} - 1)^2 n}{8\kappa_j^2}\right)$$

where

$$\widehat{b}^j_\delta[y] = \frac{\widehat{a}^j_\delta[y]}{(c_y + c_{\bar{y}} - 1)} - \frac{1}{n}\frac{(1 - c_{\bar{y}})}{c_y + c_{\bar{y}} - 1}\sum_{i=1}^n K_j(\mathbf{x}^1_i, \mathbf{x}^2_i)$$

The proof can be found in the supplementary material. As indicated by Theorem 3, under Assumption 1, we can approximate $a^j_\delta[y]$ by $\widehat{b}^j_\delta[y]$. It is interesting to note that the convergence rate is $O\left(1/[(c_y + c_{\bar{y}} - 1)\sqrt{n}]\right)$, not $O(1/\sqrt{n})$. Thus, when the noise level of pairwise constraints is high, i.e., $c_y + c_{\bar{y}} - 1$ is small, the two quantities could still differ significantly even with modest number of training pairs. Similar to Theorem 3, we can have the following corollary to bound the difference between $a^j_p[y]$ and $\widehat{b}^j_\delta[y]$.

Corrolary 4. *Assume $(\mathbf{x}^1_i, \mathbf{x}^2_i, \widehat{y}_i), i = 1, \ldots, n$ are i.i.d samples, then for any $\epsilon > 0$ we have*

$$\Pr\left(\left|a^j_p[y] - \widehat{b}^j_\delta[y]\right| \ge \epsilon\right) \le 4\exp\left(-\frac{\epsilon^2(c_y + c_{\bar{y}} - 1)^2 n}{8\kappa_j^2}\right)$$

With theorem 3 and corollary 4, we finally arrive at the following formulation for generalized maximum entropy learning from noise side information

$$\begin{aligned} \max \quad & \frac{1}{n}\sum_{i=1}^n H(p|\mathbf{x}^1_i, \mathbf{x}^2_i) - \frac{1}{2\gamma}\sum_y \|\epsilon_y\|^2 \\ s.t. \quad & \frac{1}{n}\sum_{i=1}^n p(y|\mathbf{x}^1_i, \mathbf{x}^2_i) K_j(\mathbf{x}^1_i, \mathbf{x}^2_i) \ge \widehat{b}^j_\delta[y] - \epsilon_{yj} \end{aligned} \tag{3}$$

3.4. Optimization, Analysis and Application to Kernel Learning

We present the dual formulation to the generalized maximum entropy learning from noisy side information in (3). To simplify our presentation, we define

$$\widehat{\mathbf{b}}_1 = (\widehat{b}^1_\delta[y], \ldots, \widehat{b}^m_\delta[y])^\top_{|y=1}, \widehat{\mathbf{b}}_0 = (\widehat{b}^1_\delta[y], \ldots, \widehat{b}^m_\delta[y])^\top_{|y=-1}$$

The dual problem to (3) is given by

$$\begin{aligned} \max_{\lambda_1, \lambda_0 \in \mathbb{R}^m_+} \quad & \lambda_1^\top \widehat{\mathbf{b}}_1 + \lambda_0^\top \widehat{\mathbf{b}}_0 - \frac{\gamma}{2}(\|\lambda_1\|_*^2 + \|\lambda_0\|_*^2) \\ & - \frac{1}{n}\sum_i \ln\left(\exp(\lambda_1^\top \mathbf{k}_i) + \exp(\lambda_0^\top \mathbf{k}_i)\right) \end{aligned} \tag{4}$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$. The resulting conditional distribution $p(y|\mathbf{x}^1, \mathbf{x}^2)$ is given by

$$p(y = 1|\mathbf{x}^1, \mathbf{x}^2) = \frac{\exp((\lambda_1 - \lambda_0)^\top \mathbf{k}(\mathbf{x}^1, \mathbf{x}^2))}{1 + \exp((\lambda_1 - \lambda_0)^\top \mathbf{k}(\mathbf{x}^1, \mathbf{x}^2))} \tag{5}$$

Next, we show how the solution λ_1 and λ_0 will be affected when replacing $a^j_\delta[y]$ with $\widehat{b}^j_\delta[y]$, i.e., the empirical mean computed from the noisy pairwise constraints.

Theorem 5. *Assume ℓ_2 norm is in the generalized maximum entropy model, i.e., $\|\cdot\| = \|\cdot\|_2$. Let $\lambda^* = (\lambda^*_1, \lambda^*_0)$ be the solution to (4) with $\widehat{\mathbf{b}}^* = \left(\widehat{\mathbf{b}}^*_1, \widehat{\mathbf{b}}^*_0\right)$, and $\lambda^o = (\lambda^o_1, \lambda^o_0)$ be the solution with $\widehat{\mathbf{b}}^o = \left(\widehat{\mathbf{b}}^o_1, \widehat{\mathbf{b}}^o_0\right)$. We have*

$$\|\lambda^* - \lambda^o\|_F \leq \frac{2}{\gamma}\|\widehat{\mathbf{b}}^* - \widehat{\mathbf{b}}^o\|_F$$

The proof for the theorem as well as the following theorems can be found in the supplementary material. Combining Theorem 5 and Theorem 3, we have the following theorem showing the impact of replacing $a^j_\delta[y]$ with $\widehat{b}^j_\delta[y]$.

Theorem 6. *Let $\|\cdot\| = \|\cdot\|_2$. Let $\widehat{p}(y|\mathbf{x}^1, \mathbf{x}^2)$ be the conditional model derived from noisy side information using (3), and $p(y|\mathbf{x}^1, \mathbf{x}^2)$ be the conditional model derived from the perfect side information using (2). Under Assumption 1, with probability $1-\delta$, for any $\mathbf{x}^1$, $\mathbf{x}^2$ and y, we have*

$$\left|\widehat{p}(y|\mathbf{x}^1, \mathbf{x}^2) - p(y|\mathbf{x}^1, \mathbf{x}^2)\right| \leq \frac{4m\kappa^2}{\gamma c}\sqrt{\frac{8}{n}\ln\left(\frac{8m}{\delta}\right)}$$

where $\kappa = \max_{1\leq j\leq m}\kappa_j$, and $c = c_+ + c_- - 1$.

As indicated by the above theorem, the difference between two conditional models will be reduced at the rate of $1/[(c_+ + c_- - 1)\sqrt{n}]$. Finally, since our algorithm depends on the knowledge of c_+ and c_-, we further analyze the behavior of the proposed algorithm with inaccurate estimation of c_+ and c_-. We denote by $\widehat{c}_+$ and $\widehat{c}_-$ the estimates of c_+ and c_-, respectively. We define $\widetilde{p}(y|\mathbf{x}^1, \mathbf{x}^2)$ the conditional model derived from the noisy side information using $\widehat{c}_+$ and $\widehat{c}_-$. We measure the difference (c_+, c_-) and their estimates by $\Delta = \max(|c_+ - \widehat{c}_+|, |c_- - \widehat{c}_-|)$. The next theorem shows the difference between $\widetilde{p}(y|\mathbf{x}^1, \mathbf{x}^2)$ and $p(y|\mathbf{x}^1, \mathbf{x}^2)$.

Theorem 7. *Let $\|\cdot\| = \|\cdot\|_2$. Let $\widetilde{p}(y|\mathbf{x}^1, \mathbf{x}^2)$ be the conditional model derived from noisy side information with $\widehat{c}_+$ and $\widehat{c}_-$, and $p(y|\mathbf{x}^1, \mathbf{x}^2)$ be the conditional model derived from the perfect side information using (2). Assume $c_+ + c_- - 1 \geq \rho$ with $\rho \geq 0$ and $\Delta \leq \rho/4$. Under Assumption 1, with probability $1-\delta$, for any $\mathbf{x}^1$, $\mathbf{x}^2$ and y, we have*

$$\left|\widetilde{p}(y|\mathbf{x}^1, \mathbf{x}^2) - p(y|\mathbf{x}^1, \mathbf{x}^2)\right| \leq \frac{4m\kappa^2}{\gamma c}\sqrt{\frac{8}{n}\ln\frac{8m}{\delta}} + \frac{32\kappa m\Delta}{\gamma\rho^2}$$

We finally discuss the application of the generalized maximum entropy learning from side information to kernel learning. In this case, $K_j(\mathbf{x}^1, \mathbf{x}^2)$ is a candidate kernel function, and the solution to (5) could provide us the way to linearly combine kernels, i.e., $\sum_j(\lambda^j_1 - \lambda^j_0)K_j$. However, the combined kernel may not be positive semi-definite, because some weights $\lambda^j_1 - \lambda^j_0$ are negative. To ensure the combined kernel to be valid, we introduce one more constraint $\lambda^j_1 \geq \lambda^j_0, j = 1, \ldots, m$ to the optimization problem in (3). The optimization problems are solved by Nesterov method (Nemirovski, 1994).

4. Experiments

We evaluate the proposed algorithm by clustering the linked documents. We first present the experiments on clustering linked documents with noisy pairwise constraints derived from the link information. We then examine the behavior of the proposed algorithm in more details. Before presenting the experimental results, we first introduce the data sets, baselines and evaluation metric.

Data Sets We select three linked document data sets, i.e. Cora, Citeseer, Terrorist Attacks(TeAt) [1] for our evaluation. They were processed by the research group of Lise Getoor [2]. Each data set contains (1) a set of documents described by binary vectors indicating the presence and absence of words from a dictionary, (2) links among documents (e.g., citations between research articles), and (3) the class assignment for each document. The statistics of these three data sets are summarized in Table 1. In the experiments, the attributes for each document are normalized by first dividing the sum of the attributes and then taking the square root (Jebara et al., 2004).

Evaluation In order to evaluate the proposed algorithm, we apply it to kernel learning as described at the end of Section 3. In particular, for each attribute j, we construct a linear kernel matrix $K_j(\mathbf{x}^1, \mathbf{x}^2) = \mathbf{x}^1[j]\mathbf{x}^2[j]$ for paired documents $(\mathbf{x}^1, \mathbf{x}^2)$, where $\mathbf{x}[j]$ is the j^{th} normalized attribute of document $\mathbf{x}$. The proposed algorithm will be applied to learn the combination of multiple kernel matrices from the noisy pairwise constraints derived from links. The ℓ_2 norm is used in the proposed algorithm. Given the learned kernel matrix, a spectral clustering algorithm (Shi & Malik, 1997) is applied for document clustering. We evaluate

[1] We choose these data sets because they have relatively low noise (20% $\sim$ 35%) in their pairwise constraints derived from links that satisfies the condition $c_+ + c_- - 1 > 0$ in Assumption 1.

[2] http://www.cs.umd.edu/projects/linqs/projects/lbc/

Table 1. Statistics of Data sets

name	#examples	#words	#links	#classes
Cora	2708	1433	5429	7
Citeseer	3312	3703	4732	6
TeAt	1293	106	571	6

the clustering result by comparing it to the class assignment information provided in each data set. *Normalized mutual information*(NMI) (Yang et al., 2009) is used as our evaluation metric. For all the experiments, we set γ in the proposed algorithm to be $0.01/c^2$, where $c = c_+ + c_- - 1$.

Baseline We compare the proposed algorithm to the following metric/kernel learning algorithms: (a) **GDM**, the global distance metric learning algorithm (Xing et al., 2003), (b) **DCA**, the discriminative component analysis algorithm (Hoi et al., 2006a), (c) **ITML**, the information theoretic metric learning algorithm proposed by (Davis et al., 2007), and (d) **SKL**, the spectral kernel learning algorithm (Hoi et al., 2006b). For fair comparison, the distance metric A learned by the metric learning algorithms will be used to construct a kernel matrix $K = XAX^\top$, where X is the data matrix, and the same spectral clustering algorithm will be applied to K for document clustering. We also evaluate the proposed algorithm against the metric pairwise constrained K-means clustering algorithm (Basu et al., 2004b), referred to as **MPCK**. In order to improve the robustness of **MPCK** to noisy constraints, we follow (Liu et al., 2007) and weight the noisy positive constraints and the noisy negative constraints by c_+, c_- respectively to reduce their impact on the clustering results. As the reference point, we compute a linear kernel for both labeled and unlabeled examples, without using the provided pairwise constraints. We refer to this baseline as **base**. Finally, we refer to as **GMEns** the proposed generalized maximum entropy model for learning from noisy side information, and as **GMEs** the generalized maximum entropy model without considering the noise in side information. All the experiments are run five times and the clustering accuracy averaged over five runs is reported in our study.

4.1. Experiments on Real Noisy Constraints

We conduct experiments of document clustering with the noisy pairwise constraints derived from the links between documents. In particular, we use all the linked document pairs as the positive constraints. The same number of document pairs without link are sampled to construct the negative constraints. To obtain the noise levels of the pairwise constraints, we sample a total of 100 pairwise constraints and estimate c_+ and c_- based on the correctness of the sampled constraints [3]. Figures 1(a), 2(a) and 3(a) show the clustering accuracy measured in NMI for the three data sets. The mean values of the estimated c_+ and c_- are listed under each figure. We observe that given the noisy pairwise constraints, all the algorithms except **ITML** perform significantly worse on at least one data set than the reference method **base**. In contrast, the proposed algorithm for learning from noisy pairwise constraints outperforms the reference method significantly for all three data sets. We thus conclude that the proposed algorithm is overall more robust to noise in the side information.

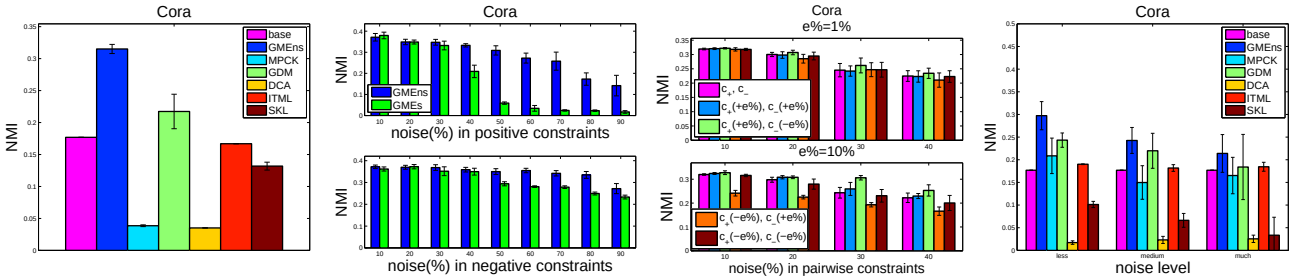
4.2. Controlled Experiments with Synthetic Noisy Constraints

In this section, we examine the robustness of the proposed algorithm to (a) different noise levels in synthetically generated pairwise constraints, and (b) the estimated values for c_+ and c_-.

Robustness to the Noise We first sample $10,000$ pairwise constraints from each data set, with $5,000$ positive constraints and $5,000$ negative constraints. Random noise is introduced to the synthetic constraints by randomly flipping the label of a pair with a probability $p\%$, where $p\%$ specifies the noise level. We set c_+ and c_- to be $1 - p\%$, with the assumption that the knowledge of noise level is perfect. To examine the impact of noisy positive constraints and noisy negative constraints separately, for each data set, with a given noise level $p\%$, we conduct two experiments, one with corrupted positive constraints but perfect negative constraints, and the other with corrupted negative constraints but perfect positive constraints. Figures 1(b), 2(b) and 3(b) compare the clustering results for **GMEns** and **GMEs** with the noise levels in the synthetic pairwise constraints varied from 10% to 90% on the three data sets. We observe that **GMEns**, the generalized maximum entropy model for noisy side information, is significantly more robust to the noise in the pairwise constraints than **GMEs** which does not take into account the noise in side information. We also observe that the noisy positive constraints have significantly higher adverse impact on the clustering results than the noisy negative constraints.

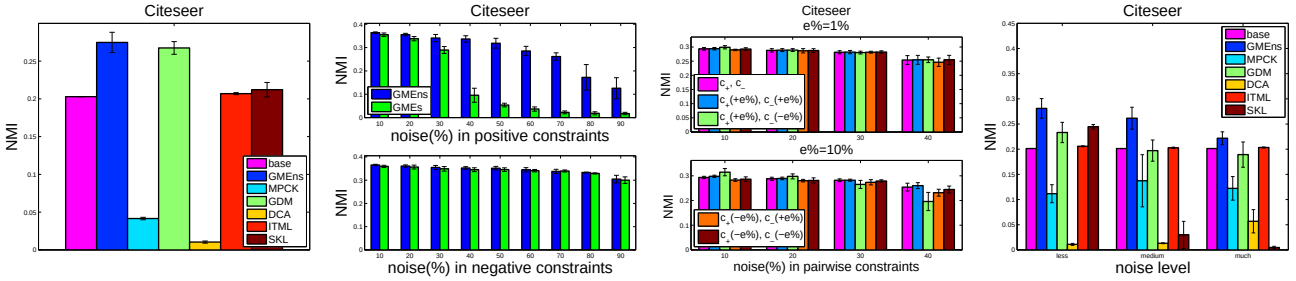
Sensitivity to c_+, c_- We use the same set of $10,000$ randomly sampled pairwise constraints for this study.

[3] These validated pairwise constraints are also used by the other baseline methods for computing distance metrics and kernel matrices



(a) $\bar{c}_+ = 0.8192$, $\bar{c}_- = 0.8150$ (b) (c) (d)

Figure 1. Experimental results on Cora data set (a) Comparison with baselines on real noisy constraints; (b) Robustness to the noise in synthetic constraints; (c) Sensitivity to c_+, c_-; (d) Comparison with baselines on synthetic noisy constraints. The same placement of subfigures applies to Figure 2 and Figure 3.



(a) $\bar{c}_+ = 0.8077$, $\bar{c}_- = 0.7596$ (b) (c) (d)

Figure 2. Experimental results on Citeseer data set

We add the same noise level to both positive constraints and negative constraints. To investigate the sensitivity to c_+, c_-, instead of setting them to be $1-p\%$, we perturb these parameters by setting them to be $(1-p\%)(1\pm e\%)$. Figures 1(c), 2(c) and 3(c) show the results of **GMEns** on the three data sets with four noise levels $p\% = 10\% \sim 40\%$ for $e\% = 1\%, 10\%$. We observe that **GMEns** is overall robust to modest perturbation level, making the proposed algorithm applicable even when the assumed noise levels are inaccurate.

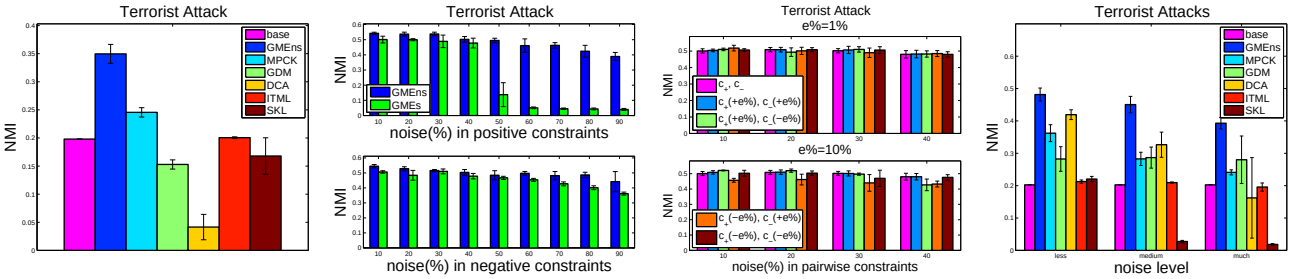
Finally, we compare the proposed algorithm to the baselines on the synthetic noisy constraints by varying the level of noise. Due to the fact that some of the baseline algorithms are time consuming, one thousand pairs are sampled for positive constraints and negative constraints, respectively. We show the results on the noise added to the positive constraints due to its stronger effect on the performance. Figures 1(d), 2(d) and 3(d) show the clustering results of all algorithms at three noise levels: low(10%), medium(40%), high(70%) on the three data sets. We observe that the proposed algorithm is able to outperform all the baseline algorithms for all the cases.

5. Conclusions

We have proposed a generalized maximum entropy model for learning from noisy side information, and discussed its application to kernel learning. Our theoretical analysis shows that the model trained from the noisy side information converges to the model trained from the perfect side information. Extensive experimental results verify the efficacy of the proposed model. In the future, we plan to apply the proposed approach to problems in other domains, including visual object recognition and gene expression pattern prediction.

Acknowledgement

The work was supported in part by National Science Foundation (IIS-0643494), Office of Naval Research (N00014-09-1-0663), and Army Research Office (W911NF-09-1-0421). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF, ONR and ARO. Part of Anil Jain's research was supported by WCU (World Class University) program through the National Research



(a) $\bar{c}_+ = 0.6590$, $\bar{c}_- = 0.6643$ (b) (c) (d)

Figure 3. Experimental results on Terrorist Attack data set

Foundation of Korea funded by the Ministry of Education, Science and Technology(R31-2008-000-10008-0).

References

Altun, Yasemin and Smola, Alexander J. Unifying divergence minimization and statistical inference via convex duality. In *COLT*, 2006.

Basu, Sugato, Banjeree, A., Mooney, ER., Banerjee, Arindam, and Mooney, Raymond J. Active semi-supervision for pairwise constrained clustering. In *SDM*, 2004a.

Basu, Sugato, Bilenko, Mikhail, and Mooney, Raymond J. A probabilistic framework for semi-supervised clustering. In *KDD*, 2004b.

Bilenko, Mikhail and Mooney, Raymond J. Adaptive duplicate detection using learnable string similarity measures. In *KDD*, 2003.

Chapelle, O., Weston, J., and Schölkopf, B. Cluster kernels for semi-supervised learning. In *NIPS*, 2003.

Cohn, D., Caruana, R., and McCallum, A. Semi-supervised clustering with user feedback. Technical report, 2003.

Cristianini, Nello, Shawe-taylor, John, Elisseeff, Andr, and Kandola, Jaz. On kernel-target alignment. In *NIPS*, 2001.

Davidson, Ian and Ravi, S. S. Hierarchical clustering with constraints: Theory and practice. In *PKDD*, 2005.

Davidson, Ian and Sugato, Basu. A survey of clustering with instance level constraints. Technical report, 2007.

Davis, Jason V., Kulis, Brian, Jain, Prateek, Sra, Suvrit, and Dhillon, Inderjit S. Information-theoretic metric learning. In *ICML*, 2007.

Goldberger, Jacob, Roweis, Sam, Hinton, Geoff, and Salakhutdinov, Ruslan. Neighbourhood components analysis. In *NIPS*, 2004.

Hoi, Steven C. H., Liu, Wei, Lyu, Michael R., and Ma, Wei-Ying. Learning distance metrics with contextual constraints for image retrieval. In *CVPR*, 2006a.

Hoi, Steven C. H., Lyu, Michael R., and Chang, Edward Y. Learning the unified kernel machines for classification. In *KDD*, 2006b.

Hoi, Steven C. H., Jin, Rong, and Lyu, Michael R. Learning nonparametric kernel matrices from pairwise constraints. In *ICML*, 2007.

Jebara, Tony, Kondor, Risi, Howard, Andrew, Bennett, Kristin, and Cesa-bianchi, Nicol. Probability product kernels. *JMLR*, 5:819–844, 2004.

Klein, Dan, Kamvar, Sepandar, and Manning, Christopher. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In *ICML*, 2002.

Kondor, Risi Imre and Lafferty, John. Diffusion kernels on graphs and other discrete structures. In *ICML*, 2002.

Lanckriet, Gert, Cristianini, Nello, Bartlett, Peter, and Ghaoui, Laurent El. Learning the kernel matrix with semi-definite programming. *JMLR*, 5:2004, 2002.

Lawrence, Neil D. and Schölkopf, Bernhard. Estimating a kernel fisher discriminant in the presence of label noise. In *ICML*, 2001.

Liu, Yi, Jin, Rong, and Jain, Anil K. Boostcluster: boosting clustering by pairwise constraints. In *KDD*, 2007.

Nemirovski, A. Efficient methods in convex programming. 1994.

Pal, Chris, Mann, Gideon, and Minerich, Richard. Putting semantic information extraction on the map: Noisy label models for fact extraction. In *AAAI Workshop on Information Integration on the Web*, 2007.

Pelleg, Dan and Baras, Dorit. K-means with large and noisy constraint sets. In *ECML*, 2007.

Shental, Noam, Hertz, Tomer, Weinshall, Daphna, and Pavel, Misha. Adjustment learning and relevant component analysis. In *ECCV*, 2002.

Shi, Jianbo and Malik, Jitendra. Normalized cuts and image segmentation. *PAMI*, 22:888–905, 1997.

Wagstaff, Kiri, Cardie, Claire, Rogers, Seth, and Schrödl, Stefan. Constrained k-means clustering with background knowledge. In *ICML*, 2001.

Weinberger, K., Blitzer, J., and Saul, L. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, 2006.

Wu, Lei, Jin, Rong, Hoi, Steven C.H., Zhuang, Jinfeng, and Yu, Nenghai. Simplenpkl: Simple non-parametric kernel learning. In *NIPS*, 2009.

Xing, Eric P., Ng, Andrew Y., Jordan, Michael I., and Russell, Stuart. Distance metric learning, with application to clustering with side-information. In *NIPS*, 2003.

Yang, Liu and Jin, Rong. Distance metric learning: A comprehensive survey. Technical report, 2006.

Yang, Liu, Jin, Rong, Sukthankar, Rahul, and Liu, Yi. An efficient algorithm for local distance metric learning. In *AAAI*, 2006.

Yang, Tianbao, Jin, Rong, Chi, Yun, and Zhu, Shenghuo. Combining link and content for community detection: a discriminative approach. In *KDD*, 2009.

Zhang, Tong and Ando, Rie. Analysis of spectral kernel design based semi-supervised learning. In *NIPS*. 2006.

Zhu, Xiaojin, Ghahramani, Zoubin, and Lafferty, John. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, 2003.

Zhu, Xiaojin, Kandola, Jaz, Ghahramani, Zoubin, Lafferty, John, and K, Xiaojin Zhu Jaz. Nonparametric transforms of graph kernels for semi-supervised learning. In *NIPS*, 2005.