LEARNING KERNEL COMBINATION FROM NOISY PAIRWISE CONSTRAINTS

Tianbao Yang, Rong Jin, Anil K. Jain

Department of Computer Science and Engineering, Michigan State University
{yangtia1, rongjin, jain}@msu.edu

ABSTRACT

We consider the problem of learning the combination of multiple kernels given noisy pairwise constraints, which is in contrast to most of the existing studies that assume perfect pairwise constraints. This problem is particularly important when the pairwise constraints are derived from side information such as hyperlinks and paper citations. We propose a probabilistic approach for learning the combination of multiple kernels and show that under appropriate assumptions, the combination weights learned by the proposed approach from the noisy pairwise constraints converge to the optimal weights learned from perfectly labeled pairwise constraints. Empirical studies on data clustering using the learned combined kernel verify the effectiveness of the proposed approach.

Index Terms— kernel learning, pairwise constraints

1. INTRODUCTION

Learning to combine multiple kernels has found its wide applications in various classification and clustering problems [2, 4]. The objective is to find the optimal combination of multiple kernels that minimizes a certain regularized empirical loss. A number of algorithms [5] have been proposed to learn kernel combination either from labeled data or from a mixture of labeled and unlabeled data. Given the difficulty in acquiring labeled examples, we consider the problem of learning the combination weights of multiple kernels from pairwise constraints in the context of data clustering. Two types of constraints are commonly used, i.e. positive pairs that are labeled to belong to the same cluster and negative pairs that are labeled to belong to different clusters. Learning Kernel Combination from Pairwise Constraints is to find the optimal combination of multiple kernels that is consistent with the given constraints. Pairwise constraints are typically derived automatically from side information, making it more suitable for learning the optimal combination of multiple kernels. The main challenge in learning kernel combination from pairwise constraints is the potential noise in the constraints. For instance, in the domain of document clustering, the pairwise constraints can be derived from the citation information by assuming that two articles belong to the same class if one article cites the other or are in different classes if there is no citation between them. Evidently, this assumption is not always true as two articles could discuss different subjects even though one cites the other.

Learning kernel combination from noisy pairwise constraints is a challenging problem, because we do not know a priori which pairs are incorrectly labeled. Simply optimizing the kernel combination with respect to the noisy constraints could lead to a poor performance, as shown in our empirical study. One way to address the noisy label problem is to take a Bayesian approach that explicitly models the stochastic process of generating noisy labels from hidden true labels [8, 10]. However, there are two limitations: (i) there is no guarantee for the solution; (ii) the related optimization problem is non-convex making it difficult to find globally optimal solution. In this paper, we propose a probabilistic approach for learning kernel combination from noisy pairwise constraints, leading to a convex programming problem. We show that under appropriate assumptions the combination weights learned from the noisy pairwise constraints converge to the optimal ones learned from the pairwise constraints with perfect labels. Empirical studies on clustering using the learned combined kernel demonstrate the efficacy of the proposed approach.

2. A PROBABILISTIC APPROACH

In this section, we present a probabilistic approach for learning kernel combination from noisy pairwise constraints. We also present a theoretic analysis of the proposed algorithm.

We first present some notations. Let $\mathcal{D} = \{\mathbf{x}_j \in \mathcal{X}, j = 1, \cdots, N\}$ be a collection of observed data, and $\mathcal{P} = \{(\mathbf{x}_i^1, \mathbf{x}_i^2, \widehat{y}_i) : \mathbf{x}_i^1, \mathbf{x}_i^2 \in \mathcal{D}, \widehat{y}_i \in \{1, -1\}, i = 1, \cdots, n\}$ be a collection of n pairwise constraints, where $\widehat{y}_i$ is the *noisy* label given to the pair $(\mathbf{x}_i^1, \mathbf{x}_i^2)$ indicating if the pair is a positive constraint ($\widehat{y}_i = 1$) or a negative constraint ($\widehat{y}_i = -1$). Let y_i denote the underlying true label for the pair $(\mathbf{x}_i^1, \mathbf{x}_i^2)$ which is unknown in our setting. Let $\{k_j(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}, j = 1, \cdots, m\}$ be a set of m base kernels. Our goal is to learn a combination of the multiple kernels $k(\cdot,\cdot) = \sum_{j=1}^m w_j k_j(\cdot,\cdot)$, given the noisy pairwise constraints.

Throughout the paper, we use capital letters $\mathrm{X}, \mathrm{Y}, \widehat{\mathrm{Y}}$ for the corresponding random variables, and small letters $\mathbf{x}, y, \widehat{y}$ for the observed data. To simplify presentation, we define

Table 1. notations used throughout the paper

$\mathbf{k}(\mathrm{X}^1,\mathrm{X}^2) = (\kappa_1(\mathrm{X}^1,\mathrm{X}^2),\cdots,\kappa_m(\mathrm{X}^1,\mathrm{X}^2))^\top$
$\mathrm{E}[\mathbf{k}] = \mathrm{E}_{\mathrm{X}^1,\mathrm{X}^2}\left[\mathbf{k}(\mathrm{X}^1,\mathrm{X}^2)\right]$
$\mathrm{E}^o[\mathbf{k}] = \mathrm{E}_{\mathrm{X}^1,\mathrm{X}^2,\mathrm{Y}}\left[\mathrm{Y}\mathbf{k}(\mathrm{X}^1,\mathrm{X}^2)\right]$
$\mathrm{E}^c_y[\mathbf{k}] = \mathrm{E}_{\mathrm{X}^1,\mathrm{X}^2\vert\mathrm{Y}=y}\left[\mathbf{k}(\mathrm{X}^1,\mathrm{X}^2)\right]$
$\widehat{\mathrm{E}}^c_y[\mathbf{k}] = \mathrm{E}_{\mathrm{X}^1,\mathrm{X}^2\vert\widehat{\mathrm{Y}}=y}\left[\mathbf{k}(\mathrm{X}^1,\mathrm{X}^2)\right]$

notations for different expectations in Table 1, where the superscript ^{o} and ^{c} are used to indicate the expectation with respect to the joint distribution $\Pr(\mathrm{X}^1,\mathrm{X}^2,\mathrm{Y})$ and to the conditional distribution $\Pr(\mathrm{X}^1,\mathrm{X}^2|\mathrm{Y})$ or $\Pr(\mathrm{X}^1,\mathrm{X}^2|\widehat{\mathrm{Y}})$, respectively. We use subscript $_+$ and $_-$ for conditions $_{y=1}$ and $_{y=-1}$, respectively, and $\delta(y,z)$ for Kronecker delta function.

To learn the combination of multiple kernels, we construct a conditional model $\Pr(y|\mathbf{x}^1,\mathbf{x}^2)$ for computing the pairwise classification probability, where $y \in \{1,-1\}$ is the underlying true label to indicate whether the pair $(\mathbf{x}^1,\mathbf{x}^2)$ belongs to the same class or not. Given multiple kernels $k_j(\cdot,\cdot), j = 1,\ldots,m$, we represent $\Pr(y|\mathbf{x}^1,\mathbf{x}^2)$ by a logistic function

$$\Pr(y|\mathbf{x}^1,\mathbf{x}^2) = \frac{1}{1+\exp(-y\mathbf{w}^\top\mathbf{k}(\mathbf{x}^1,\mathbf{x}^2))} = \frac{\exp(y\mathbf{w}^\top\mathbf{k}(\mathbf{x}^1,\mathbf{x}^2)/2)}{\exp(-\mathbf{w}^\top\mathbf{k}(\mathbf{x}^1,\mathbf{x}^2)/2)+\exp(\mathbf{w}^\top\mathbf{k}(\mathbf{x}^1,\mathbf{x}^2)/2)},$$

where $\mathbf{w} = (w_1,\cdots,w_m)^\top \in \mathbb{R}^m_+$ are the non-negative weights that need to be learned. We constrain $\mathbf{w}$ to be non-negative to ensure that the resulting combined kernel is positive semi-definite. By maximizing the log-likelihood of the observed data, we obtain the optimal weights for kernel combination. More specifically, we need to solve the following optimization problem

$$\max_{\mathbf{w}\in\mathbb{R}^m_+} \frac{1}{n}\sum_{i=1}^n \ln p(y_i|\mathbf{x}^1_i,\mathbf{x}^2_i) - \frac{\lambda}{2}\|\mathbf{w}\|_2^2, \tag{1}$$

or equivalently,

$$\max_{\mathbf{w}\in\mathbb{R}^m_+} \frac{1}{2}\mathbf{w}^\top a^o[\mathbf{k}] - \frac{\lambda}{2}\|\mathbf{w}\|_2^2 - \frac{1}{n}\sum_{i=1}^n L(\mathbf{w},\mathbf{k}_i), \tag{2}$$

where $L(\mathbf{w},\mathbf{k}_i) = \ln\sum_{z=\{1,-1\}}\exp\left(\frac{1}{2}z\mathbf{w}^\top\mathbf{k}(\mathbf{x}^1_i,\mathbf{x}^2_i)\right)$, and $a^o[\mathbf{k}]$ is computed by $a^o[\mathbf{k}] = \frac{1}{n}\sum_i y_i\mathbf{k}(\mathbf{x}^1_i,\mathbf{x}^2_i)$, which is what we call sufficient statistics. The main challenge is that the true labels of the observed pairs are unknown, making it difficult to apply the maximum likelihood estimation. Below, we present an approach that effectively approximates the sufficient statistics using the noisy labels.

Our key observation is that (i) sufficient statistics $a^o[\mathbf{k}]$ can be viewed as a sample average of $\mathrm{E}^o[\mathbf{k}]$, and (ii) $\mathrm{E}^o[\mathbf{k}]$ can be well approximated using the noisy labels $\widehat{y}_i$ under appropriate conditions. To assist the estimation of sufficient statistics from noisy labels, we assume the following information is available: (1) $\Pr(\mathrm{Y}=y) = p_y$, i.e. how likely a pair is truly positive or negative, and (2) $\Pr(\mathrm{Y}=y|\widehat{\mathrm{Y}}=y) = d_y$, i.e. how likely a positively (negatively) labeled pair is indeed positive (negative). We will discuss how to obtain p_y and d_y in the experimental section. Besides p_y and d_y, we assume that $\Pr(\mathrm{X}^1,\mathrm{X}^2|\mathrm{Y},\widehat{\mathrm{Y}}) = \Pr(\mathrm{X}^1,\mathrm{X}^2|\mathrm{Y})$, i.e. given the true label Y, the pair $(\mathrm{X}^1,\mathrm{X}^2)$ is independent of the noisy label $\widehat{\mathrm{Y}}$. Note that this assumption alleviates the assumption made in [13], where the noise label $\widehat{\mathrm{Y}}$ is assumed to be independent of pair $(\mathrm{X}^1,\mathrm{X}^2)$ given the true label Y. Given the knowledge of p_y, we rewrite the expectation $\mathrm{E}^o[\mathbf{k}]$ by

$$\begin{aligned}\mathrm{E}^o[\mathbf{k}] &= \mathrm{E}_{\mathrm{X}^1,\mathrm{X}^2,\mathrm{Y}}\left[\mathrm{Y}\mathbf{k}(\mathrm{X}^1,\mathrm{X}^2)\right]\\ &= \mathrm{E}_{\mathrm{Y}}\left[\mathrm{Y}\mathrm{E}_{\mathrm{X}^1,\mathrm{X}^2|\mathrm{Y}}\mathbf{k}(\mathrm{X}^1,\mathrm{X}^2)\right] = p_+\mathrm{E}^c_+[\mathbf{k}] - p_-\mathrm{E}^c_-[\mathbf{k}].\end{aligned}$$

Estimating $\mathrm{E}^o[\mathbf{k}]$ is therefore reduced to estimating $\mathrm{E}^c_y[\mathbf{k}]$. Given the independence assumption, we have for $y \in \{1,-1\}$

$$\mathrm{E}_{\mathrm{X}^1,\mathrm{X}^2|\widehat{\mathrm{Y}}=y}[\mathbf{k}(\mathrm{X}^1,\mathrm{X}^2)] = \sum_{z\in\{1,-1\}}\mathrm{E}_{\mathrm{X}^1,\mathrm{X}^2|\mathrm{Y}=z}[\mathbf{k}(\mathrm{X}^1,\mathrm{X}^2)]\Pr(\mathrm{Y}=z|\widehat{\mathrm{Y}}=y).$$

Writing in the matrix form, we have

$$\left(\widehat{\mathrm{E}}^c_+[\mathbf{k}],\widehat{\mathrm{E}}^c_-[\mathbf{k}]\right) = \left(\mathrm{E}^c_+[\mathbf{k}],\mathrm{E}^c_-[\mathbf{k}]\right)\mathbf{B}, \tag{3}$$

where $\mathbf{B} = \begin{pmatrix} d_+ & 1-d_- \\ 1-d_+ & d_- \end{pmatrix}$. Equation (3) allows us to estimate $\mathrm{E}^c_y[\mathbf{k}]$ (and therefore $\mathrm{E}^o[\mathbf{k}]$) from $\widehat{\mathrm{E}}^c_y[\mathbf{k}]$, a quantity that can be computed from the noisy labels. In particular, we approximate $\widehat{\mathrm{E}}^c_y[\mathbf{k}]$ by its sample average $\widehat{a}^c_y[\mathbf{k}]$ which is computed by $\widehat{a}^c_y[\mathbf{k}] = \frac{\sum_i \delta(\widehat{y}_i,y)\mathbf{k}(\mathbf{x}^1_i,\mathbf{x}^2_i)}{\sum_i\delta(\widehat{y}_i,y)}$. By replacing $\widehat{\mathrm{E}}^c_y[\mathbf{k}]$ with $\widehat{a}^c_y[\mathbf{k}]$ in (3), we obtain the estimation for $\mathrm{E}^c_y[\mathbf{k}]$, denoted by $b^c_y[\mathbf{k}]$, by solving the following least square problem

$$\min_{b^c_y[\mathbf{k}]} \quad \left\|\left(b^c_+[\mathbf{k}],b^c_-[\mathbf{k}]\right)\mathbf{B} - \widehat{\mathbf{A}}\right\|_F^2, \tag{4}$$

where $\widehat{\mathbf{A}} = \left(\widehat{a}^c_+[\mathbf{k}],\widehat{a}^c_-[\mathbf{k}]\right)$. To obtain a more robust estimation of $\mathrm{E}^c_y[\mathbf{k}]$, we note that $\sum_y p_y\mathrm{E}^c_y[\mathbf{k}] = \mathrm{E}[\mathbf{k}]$, and therefore add the following constraint for the estimator $b^c_y[\mathbf{k}]$

$$\sum_y p_y b^c_y[\mathbf{k}] = \frac{1}{n}\sum_{i=1}^n \mathbf{k}(\mathbf{x}^1_i,\mathbf{x}^2_i) = a[\mathbf{k}]. \tag{5}$$

Combining equation (4) and (5), we have the following constrained least square problem for $b^c_y[\mathbf{k}]$, an estimator of $\mathrm{E}^c_y[\mathbf{k}]$

$$\min_{b^c_y[\mathbf{k}]} \quad \left\|\left(b^c_+[\mathbf{k}],b^c_-[\mathbf{k}]\right)\mathbf{B} - \widehat{\mathbf{A}}\right\|_F^2, \quad s.t.\ \sum_y p_y b^c_y[\mathbf{k}] = a[\mathbf{k}].$$

It can be shown that the solution for $b^c_y[\mathbf{k}]$ is given by

$$\left(b^c_+[\mathbf{k}],b^c_-[\mathbf{k}]\right) = \left(\frac{a[\mathbf{k}]\mathbf{p}^\top}{\mathbf{p}^\top(\mathbf{B}\mathbf{B}^\top)^{-1}\mathbf{p}} + \widehat{\mathbf{A}}\mathbf{B}\left[\mathbf{I} - \frac{(\mathbf{B}\mathbf{B}^\top)^{-1}\mathbf{p}\mathbf{p}^\top}{\mathbf{p}^\top(\mathbf{B}\mathbf{B}^\top)^{-1}\mathbf{p}}\right]\right)(\mathbf{B}\mathbf{B}^\top)^{-1}, \tag{6}$$

where $\mathbf{p} = (p_+, p_-)^\top$. Note that the solution in (6) requires $\mathbf{B}$ to be non-singular, implying $d_+ + d_- \neq 1$. Given $\mathrm{E}^c_y[\mathbf{k}]$ is estimated by $b^c_y[\mathbf{k}]$, we have $\mathrm{E}^o[\mathbf{k}]$ estimated as follows

$$\mathrm{E}^o[\mathbf{k}] \approx b^o[\mathbf{k}] = p_+ b^c_+[\mathbf{k}] - p_- b^c_-[\mathbf{k}],$$

leading to the following the approximate log-likelihood of the true labels for the observed pairs to be maximized

$$\max_{\mathbf{w}\in\mathbb{R}^m_+} \frac{1}{2}\mathbf{w}^\top b^o[\mathbf{k}] - \frac{\lambda}{2}\|\mathbf{w}\|_2^2 - \frac{1}{n}\sum_i L(\mathbf{w}, \mathbf{k}_i). \quad (7)$$

Next, we present a convergence analysis showing that the combination weights learned by the proposed approach converge to the optimal ones learned from perfectly labeled pairs. Due to space limit, we omit the proofs of the key inequalities.

Comparing (7) with (2), we see that the only difference between the two optimization problems is that $a^o[\mathbf{k}]$ in (2) is replaced with $b^o[\mathbf{k}]$ in (7). The following lemma shows the bound for $\mathbf{w}$ when replacing $a^o[\mathbf{k}]$ with $b^o[\mathbf{k}]$.

Lemma 1 (Lemma 6[9]). *Let $\mathbf{w}^*_a$ be the solution to (2) with $a^o[\mathbf{k}]$, and $\mathbf{w}^*_b$ be the solution to (7) with $b^o[\mathbf{k}]$. We have*

$$\|\mathbf{w}^*_a - \mathbf{w}^*_b\|_2 \leq \frac{1}{\lambda}\left\|a^o[\mathbf{k}] - b^o[\mathbf{k}]\right\|_2.$$

Next, we try to bound the difference between $a^o[\mathbf{k}]$ and $b^o[\mathbf{k}]$. We assume $|\kappa_j(\mathbf{x}^1, \mathbf{x}^2)| \leq R, j \in [m]$. Note that

$$\|a^o[\mathbf{k}] - b^o[\mathbf{k}]\|_2 \leq \|a^o[\mathbf{k}] - \mathrm{E}^o[\mathbf{k}]\|_2 + \|b^o[\mathbf{k}] - \mathrm{E}^o[\mathbf{k}]\|_2.$$

We can bound the first difference by a concentration inequality, i.e. with a probability $1-\delta$,

$$\|a^o[\mathbf{k}] - \mathrm{E}^o[\mathbf{k}]\|_2 \leq \sqrt{\frac{2mR^2}{n}\ln\left(\frac{2m}{\delta}\right)}.$$

To bound the second difference, we assume that there are a significantly large number n_+ of pairs (possibly nosily) labeled as positive and large number n_- of pairs labeled as negative, i.e., there exists some positive constant $\rho > 0$ such that $\min(n_+/n, n_-/n) \geq \rho$. Then our analysis shows that with probability at least $1-\delta$ for any $\delta > 0$

$$\|b^o[\mathbf{k}] - \mathrm{E}^o[\mathbf{k}]\|_2 \leq \sqrt{C\frac{2mR^2}{n}\ln\left(\frac{6m}{\delta}\right)},$$

where $C = \frac{4f^4}{\|\mathbf{p}\|_2^2 e^4} + \frac{32(e^6 f + f^7)}{\rho^2 e^8}$, $e = d_+ + d_- - 1$, and $f = 1 + |d_+ - d_-|$. Combining the above results with Lemma 1, we have the following theorem.

Theorem 2. *The following inequality holds with a probability at least $1-\delta$ for any $\delta > 0$*

$$\|\mathbf{w}^*_a - \mathbf{w}^*_b\|_2 \leq \frac{1}{\lambda}\sqrt{\frac{2mR^2}{n}}\left(\sqrt{C\ln\frac{12m}{\delta}} + \sqrt{\ln\frac{4m}{\delta}}\right).$$

Theorem 2 indicates $\mathbf{w}^*_b$, the optimal solution to problem (7) converges to $\mathbf{w}^*_a$, the optimal solution to problem (2), as the number of the pairs n approaches infinity.

3. EXPERIMENTS

In this section, we validate the proposed approach by clustering of linked documents using the combined kernel learned from noisy pairwise constraints.

Two paper citation data sets, Cora and Citeseer, are used in our study [1]. The two data sets have been used in previous studies in the context of data clustering [12, 13]. Besides the attributes and class assignments of papers, the citations between papers are also available in these two datasets. The statistics of the two data sets are summarized in Table 2. We compare the proposed algorithm (LKCnpc) to the following baseline algorithms: (i) a simple baseline that combines kernels with equal weights, referred to as LK; (ii) two kernel learning algorithms from pairwise constraints: SKL [7], a spectral kernel learning algorithm, and NPK [6], a non-parametric kernel learning algorithm; (iii) two distance metric learning algorithms from pairwise constraints: GDM [11], a global distance metric learning algorithm, and ITML [3], a information-theoretic based metric learning algorithm; and (iv) a constrained clustering algorithm from pairwise constraints: MPCK [1], a state-of-the-art algorithm for constrained clustering.

For the proposed approach and baseline kernel learning methods, each candidate kernel k_j is the linear kernel on the j^{th} attribute. Given the learned combination of kernel matrices, the same spectral clustering algorithm will be used to cluster the documents. In all the experiments, we set the regularization parameter $\lambda = 0.01/n$. To evaluate the clustering results, we compute the normalized mutual information (NMI) [13] by comparing the cluster assignments predicted by the clustering algorithms to the class assignments given in the data set.

We construct the pairwise constraints from citations between papers. In particular, a positive pairwise constraint is created if two papers are linked by a citation, and a negative pairwise constraint is created when two papers do not cite each other. Since the number of unlinked paper pairs is much larger than that of linked pairs, for computational efficiency, we randomly sample the same number of unlinked paper pairs as that of linked paper pairs. Finally, we have on average about 20%-25% pairwise constraints being incorrect. To obtain p_y, d_y we randomly sample 1% of document pairs and label them according to their class assignments, and compute p_y, d_y from the noisy labels and the correct labels for these sampled pairs. For a fair comparison, the correct labels of these sampled pairs are used by the baselines.

The performance of different algorithms measured in terms of NMI is shown in Table 2. The results are averaged over 5 trials by randomly selecting the non-linked pairs as negative pairwise constraints. We observe that the proposed approach outperforms all the baseline methods on both the

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

Table 2. Statistics of data sets (left) and performance of algorithms as measured in NMI (right). Numbers in the parenthesis show the standard deviation of NMI over 5 trials.

data	#papers	#attr	#citations	#classes	LK	SKL	NPK	GDM	ITML	MPCK	**LKCnpc**
Cora	2708	1433	5429	7	0.1769 (0)	0.1296 (0.005)	0.0161 (0.001)	0.2306 (0.010)	0.1668 (0.002)	0.0352 (0.005)	**0.3107** (0.025)
Citeseer	3312	3703	4591	6	0.2029 (0)	0.2153 (0.010)	0.0072 (0.0005)	0.2669 (0.007)	0.2064 (0.002)	0.0563 (0.001)	**0.2902** (0.013)

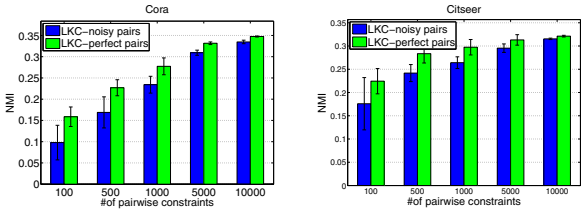

Fig. 1. Convergence behavior.

data sets used here. This shows that the proposed approach is effective for learning from noisy pairwise constraints.

We also verified the convergence behavior of the proposed approach with respect to the number of pairwise constraints n. We sample an equal number of pairs in the same class and pairs in different classes. The noisy label for these pairs are obtained by random flipping the labels with a probability 0.2. We compare the clustering performance of the proposed approach with noisy labels to the same approach but with perfect labels by increasing the total number of pairwise constraints from 100 to $10,000$. The results in Figure 1 show that the difference between using noisy labels and perfect labels decreases as the number of constraints increases, which verifies our analysis.

4. CONCLUSIONS

We have considered the problem learning kernel combination from noisy pairwise constraints. We proposed a probabilistic approach to learn the optimal combination weights. Our theoretical analysis shows that the combination weights learned from the noisily labeled pairs converge to the optimal weights learned from the pairs with perfect labels. Empirical studies demonstrate the efficacy of the proposed approach. In future work, we plan to compare the proposed approach to [13] by an appropriately designed experiment, since the two approaches require different knowledge of the noise.

5. REFERENCES

[1] S. Basu, M. Bilenko, and R. J. Mooney. A probabilistic framework for semi-supervised clustering. In *KDD*, pages 59–68, 2004.

[2] S. S. Bucak, R. Jin, and A. K. Jain. Multi-label multiple kernel learning by stochastic approximation: Application to visual object recognition. In *NIPS*, pages 325–333, 2010.

[3] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *ICML*, pages 209–216, 2007.

[4] M. Gönen and E. Alpaydin. Localized multiple kernel learning. In *ICML*, pages 352–359, 2008.

[5] M. Gönen and E. Alpaydin. Multiple kernel learning algorithms. *JMLR*, 12:2211–2268, 2011.

[6] S. C. H. Hoi, R. Jin, and M. R. Lyu. Learning nonparametric kernel matrices from pairwise constraints. In *ICML*, pages 361–368, 2007.

[7] S. C. H. Hoi, M. R. Lyu, and E. Y. Chang. Learning the unified kernel machines for classification. In *KDD*, pages 187–196, 2006.

[8] C. Pal, G. Mann, and R. Minerich. Putting semantic information extraction on the map: Noisy label models for fact extraction. In *AAAI Workshop*, 2007.

[9] N. Quadrianto, A. J. Smola, T. S. Caetano, and Q. V. Le. Estimating labels from label proportions. In *ICML*, pages 776–783, 2008.

[10] G. Ramakrishnan, K. P. Chitrapura, R. Krishnapuram, and P. Bhattacharyya. A model for handling approximate, noisy or incomplete labeling in text classification. In *ICML*, pages 681–688, 2005.

[11] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *NIPS*, pages 505–512, 2003.

[12] T. Yang, R. Jin, Y. Chi, and S. Zhu. Combining link and content for community detection: a discriminative approach. In *KDD*, pages 927–936, 2009.

[13] T. Yang, R. Jin, and A. K. Jain. Learning from noisy side information by generalized maximum entropy model. In *ICML*, pages 1199–1206, 2010.