Combination of Classification and Clustering Results with Label Propagation

Xu-Yao Zhang, Peipei Yang, Yan-Ming Zhang, Kaizhu Huang, and Cheng-Lin Liu

***Abstract*—This letter considers the combination of multiple classification and clustering results to improve the prediction accuracy. First, an object-similarity graph is constructed from multiple clustering results. The labels predicted by the classification models are then propagated on this graph to adaptively satisfy the smoothness of the prediction over the graph. The convex learning problem is efficiently solved by the label propagation algorithm. A semi-supervised extension is also provided to further improve the performance. Experiments on 11 tasks identify the validity of the proposed models.**

***Index Terms*—Classification, clustering, label propagation.**

I. Introduction

SUCCESSFUL classification algorithms (e.g. support vector machines and artificial neural networks) and clustering models (e.g. k-means and spectral clustering) have been proposed and widely used in practical applications. Ensemble learning of different models can further improve the performance due to the diversity and heterogeneity. Classifier ensemble (e.g. bagging [2], boosting [6], and error-correcting output coding [4]) can yield higher accuracy than the best individual classifier. On the other hand, for unsupervised learning, clustering ensemble [5], [9], [11] can produce better partition performance by combining different clustering algorithms.

Recently, the joint ensemble learning of multiple classification and clustering models is proposed by [7], [8] and [10]. In classification, the objects are usually classified one at a time under the assumption of independent and identical distribution (i.i.d.), therefore, the internal structure information among the objects is actually discarded. Complementarily, the clustering models can capture the object-relationship by partitioning them into different clusters, and the objects in the same cluster are more likely to receive the same label. Therefore, via combination of the classification and clustering results, higher prediction accuracy can be achieved.

To accomplish this goal, Gao *et al.* [7], [8] proposed a bipartite graph-based consensus maximization (BGCM) model by embedding both objects and groups (defined as classification and clustering results) into a fixed-dimensional cube. Ma *et al.* [10] further proposed an unconstrained probabilistic embedding (UPE) model by relaxing the constraint of embedding. Both BGCM and UPE are based on the idea of object-group embedding. In this letter, we propose a new model by first constructing an object-similarity graph from the multiple clustering results. The graph accurately captures the internal relationship among different objects. The labels (majority voting results) predicted by the supervised classification models are then propagated on this graph to adaptively improve the prediction accuracy. This is based on the idea of manifold regularization, and the manifold can be viewed as the object-similarity graph which is constructed from the clustering results. By combining all the information, we can achieve better performance than the single models (either classification or clustering) and the ensemble models (BGCM and UPE) consistently. To further improve the ensemble performance, a semi-supervised extension of our model is also proposed.

The rest of this letter is organized as follows: Section II describes the related works. Section III introduces our methods including model definition, optimization method, and theoretical analysis. Section IV reports the experimental results, and Section V draws the concluding remarks.

II. Related Work

Following [7] and [10], we first give a brief description of the problem. Given a data set $\mathcal{O} = \{o_1, o_2, \cdots, o_N\}$ belonging to C classes. Suppose M models provide the prediction results where the first r of them are classification and the remaining are clustering results. The concept of *group* is defined as the "classes" and "clusters" predicted by different models. For example, assume that clustering algorithm i partitions the data $\mathcal{O}$ into l_i clusters, the total number of *groups* is $G = C + \sum_{i=r+1}^{M} l_i$.[1]

All the information can be displayed in an *object-group co-occurrence matrix* [10] $\mathcal{M} \in \mathbb{R}^{N \times G}$, where $\mathcal{M}_{n,g}$ is the number of times that object n belongs to group g. Table I shows the co-occurrence matrix for a toy example used in [10] ($C = 2$, $G = 7$). The results are produced by two classifiers and two clustering models. The first C columns in $\mathcal{M}$ contain

[1] In [7], the total number of groups is defined as $G = Cr + \sum_{i=r+1}^{M} l_i$, where the class IDs given by different classifiers are viewed as different group.

Manuscript received December 21, 2013; revised March 01, 2014; accepted March 11, 2014. Date of publication March 14, 2014; date of current version March 21, 2014. This work was supported by National Basic Research Program of China (973 Program) under Grant 2012CB316302 and by the National Natural Science Foundation of China (NSFC) under Grant 61203296. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Kjersti Engan.

X.-Y. Zhang, P. Yang, Y.-M. Zhang, and C.-L. Liu are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China (e-mail: xyz@nlpr.ia.ac.cn; ppyang@nlpr.ia.ac.cn; ymzhang@nlpr.ia.ac.cn; liucl@nlpr.ia.ac.cn).

K. Huang is with the Department of EEE, Xi'an Jiaotong-Liverpool University, Jiangsu 215123, China (e-mail: kaizhu.huang@xjtlu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2014.2312005

TABLE I
THE OBJECT-GROUP CO-OCCURRENCE MATRIX $\mathcal{M}$ SHOWN IN [10]

	classification		clustering 1			clustering 2	
	g_1	g_2	g_3	g_4	g_5	g_6	g_7
o_1	2	0	1	0	0	1	0
o_2	1	1	1	0	0	1	0
o_3	1	1	0	1	0	1	0
o_4	2	0	0	1	0	1	0
o_5	0	2	0	0	1	0	1
o_6	0	2	0	0	1	0	1

the classification information (e.g. o_1 has been classified to g_1 twice), while the last $G - C$ columns capture the clustering relationship among objects (e.g. o_1 and o_2 are grouped into the same cluster g_3).

The purpose is to combine all the information in $\mathcal{M}$ to make the final prediction. Given a matrix X, we use X_i to represent the vector of row i and X_{ij} to represent the (i, j)-th element of X. The vector with all the elements being one is $\mathbf{1}$ and the identity matrix is I.

1) BGCM: The purpose of BGCM [7], [8] is to jointly estimate the conditional probabilities of each object and group belonging to the C classes. Define $\mathcal{F} \in \mathbb{R}^{N\times C}$ and $\mathcal{Q} \in \mathbb{R}^{G\times C}$ as $\mathcal{F}_{iz} = P(o_i \in \text{ class } z)$ and $\mathcal{Q}_{jz} = P(g_j \in \text{ class } z)$. Let $\mathcal{Y} \in \{0, 1\}^{C\times C}$ denote the class labels where $\mathcal{Y}_{ii} = 1, \forall i$ and $\mathcal{Y}_{ij} = 0, \forall i \neq j$. The BGCM model is defined as:

$$\begin{aligned} \min_{\mathcal{F},\mathcal{Q}} \quad & \sum_{i=1}^{N}\sum_{j=1}^{G} \mathcal{M}_{ij}\|\mathcal{F}_i - \mathcal{Q}_j\|_2^2 + \alpha \sum_{j=1}^{C} \|\mathcal{Q}_j - \mathcal{Y}_j\|_2^2, \\ \text{s.t.} \quad & \mathcal{F} \geq 0, \mathcal{F}\cdot\mathbf{1} = \mathbf{1} \quad \text{and} \quad \mathcal{Q} \geq 0, \mathcal{Q}\cdot\mathbf{1} = \mathbf{1}. \end{aligned} \tag{1}$$

The first term encourages close estimation of the conditional probability if an object o_i is assigned to group g_j. The second term constrains the deviation of the first C groups from the initial class label. After the optimal $\mathcal{F}$ is obtained, the objects can be classified by $o_i \in \arg\max_{1\leq z\leq C}\mathcal{F}_{iz}$.

2) UPE: BGCM is a constraint embedding onto the C-dimensional cube. In the UPE model [10], the latent coordinate for objects is $\mathcal{F} \in \mathbb{R}^{N\times D}$ and for groups is $\mathcal{Q} \in \mathbb{R}^{G\times D}$, where D can be any positive integer and no constraint is adopted on $\mathcal{F}$ and $\mathcal{Q}$. The probability of an object belonging to a certain group is determined by:

$$P(g_j|o_i) = \frac{\exp(-\frac{1}{2}\|\mathcal{F}_i - \mathcal{Q}_j\|_2^2)}{\sum_{j'=1}^{G}\exp(-\frac{1}{2}\|\mathcal{F}_i - \mathcal{Q}_{j'}\|_2^2)}. \tag{2}$$

The coordinate of $\mathcal{F}$ and $\mathcal{Q}$ are learned by maximizing the posterior probabilities: $\max_{\mathcal{F},\mathcal{Q}}\sum_{i=1}^{N}\sum_{j=1}^{G}\mathcal{M}_{ij}\log P(g_j|o_i)$. A Gaussian prior with a zero mean and a spherical covariance for distribution of $\mathcal{F}$ and $\mathcal{Q}$ is also proposed as regularization. After we get the optimal $\mathcal{F}$ and $\mathcal{Q}$, the prediction process is: $o_i \in \arg\max_{1\leq j\leq C} P(g_j|o_i) = \arg\min_{1\leq j\leq C}\|\mathcal{F}_i - \mathcal{Q}_j\|_2^2$.

III. METHODOLOGY

Both BGCM and UPE are based on the idea of learning latent embedding coordinates for objects and groups either in a constrained or unconstrained way. In this letter, we propose a new method to fuse the information in $\mathcal{M}$. The basic idea is to construct an object-similarity graph from clustering results and then propagate the classification results on this graph to adaptively improve the prediction accuracy.

The majority voting results $\mathcal{F}_0 \in \mathbb{R}^{N\times C}$ of the classification models can be obtained via

$$[\mathcal{F}_0]_i = \frac{1}{\sum_{j=1}^{C}\mathcal{M}_{ij}}\sum_{j=1}^{C}\mathcal{M}_{ij}\mathcal{Y}_j.$$

The purpose now is to learn the conditional probability matrix $\mathcal{F} \in \mathbb{R}^{N\times C}$ of each object belonging to C classes. First, the prediction should agree with the base classification results, implying that we should minimize the differences of $\|\mathcal{F} - \mathcal{F}_0\|_F^2$. Second, the objects in the same cluster are more likely to have the same prediction. If o_i and $o_{i'}$ are in the same clustering group g_j, then $\mathcal{M}_{ij}\mathcal{M}_{i'j} = 1$ else 0 (see Table I). Therefore, we can minimize $\mathcal{M}_{ij}\mathcal{M}_{i'j}\|\mathcal{F}_i - \mathcal{F}_{i'}\|_2^2$ for $\forall C+1 \leq j \leq G$ to satisfy the clustering constraints.

A. Object Similarity Graph

From the above analysis, we can define an object-similarity graph $S \in \mathbb{R}^{N\times N}$ based on the clustering results:

$$S_{ij} = \begin{cases} \sum_{k=C+1}^{G}\mathcal{M}_{ik}\mathcal{M}_{jk} & \text{if} \quad i \neq j, \\ 0 & \text{if} \quad i = j. \end{cases} \tag{3}$$

The adjacency matrix S is an ensemble of different clustering algorithms which helps to construct the neighborhood relationship among different objects. To normalize the degree, a doubly-stochastic similarity matrix $W \in \mathbb{R}^{N\times N}$ is learned by minimizing the relative entropy[2] between W and S:

$$\begin{aligned} \min_{W} \quad & \text{dist}(W, S) = \sum_{i,j=1}^{N}\left[W_{ij}\log\frac{W_{ij}}{S_{ij}} - W_{ij} + S_{ij}\right], \\ \text{s.t.} \quad & W\cdot\mathbf{1} = \mathbf{1}, W = W^\top, W \geq 0. \end{aligned} \tag{4}$$

This convex problem can be efficiently solved by projecting S onto the constraints [3]. By initializing $W = S$, the following procedures should be repeated until convergence:

$$d_i = \sum_{j=1}^{N} W_{ij}^{\text{old}}, \quad W_{ij}^{\text{new}} = W_{ij}^{\text{old}}/d_i, \quad \forall i, j \tag{5}$$

$$W_{ij}^{\text{new}} = W_{ji}^{\text{new}} = \sqrt{W_{ij}^{\text{old}} \quad W_{ji}^{\text{old}}}, \quad \forall i, j \tag{6}$$

where (5) is for projecting onto $W\cdot\mathbf{1} = \mathbf{1}$ and (6) is for projecting onto $W = W^\top$. In this way, a degree-normalized symmetrical object-similarity graph W can be obtained.

B. The Learning Problem

The combination of classification and clustering results can be formulated as an optimization problem:

$$\begin{aligned} \min_{\mathcal{F}\in\mathbb{R}^{N\times C}} \quad & (1-\beta)\|\mathcal{F} - \mathcal{F}_0\|_F^2 + \frac{\beta}{2}\sum_{i,j=1}^{N} W_{ij}\|\mathcal{F}_i - \mathcal{F}_j\|_2^2, \\ \text{s.t.} \quad & \mathcal{F} \geq 0, \ \mathcal{F}\cdot\mathbf{1} = \mathbf{1}. \end{aligned} \tag{7}$$

[2]Relative entropy is better than Frobenius norm [12] in measuring the difference of doubly-stochastic matrices.

Here $0 \leq \beta < 1$ is a balance between classifier consistence and clustering consistence. Because $W \cdot \mathbf{1} = \mathbf{1}$ and $W = W^\top$, we can transform the second term of (7) as:

$$\begin{aligned}\frac{1}{2}\sum_{i,j=1}^{N} W_{ij}\|\mathcal{F}_i - \mathcal{F}_j\|_2^2 &= \sum_{i=1}^{N}\mathcal{F}_i^\top\mathcal{F}_i - \sum_{i,j=1}^{N}\mathcal{F}_i^\top\mathcal{F}_j W_{ij} \\ &= \operatorname{tr}\left(\mathcal{F}^\top\mathcal{F}\right) - \operatorname{tr}\left(\mathcal{F}^\top W\mathcal{F}\right) = \operatorname{tr}\left(\mathcal{F}^\top L\mathcal{F}\right),\end{aligned}$$

where $L = I - W$ is the normalized graph Laplacian. Therefore, (7) is equivalent to

$$\begin{aligned}\min_{\mathcal{F}\in\mathbb{R}^{N\times C}} \quad & \mathcal{J} = (1-\beta)\left\|\mathcal{F} - \mathcal{F}_0\right\|_F^2 + \beta\operatorname{tr}\left(\mathcal{F}^\top L\mathcal{F}\right), \\ \text{s.t.} \quad & \mathcal{F} \geq 0, \mathcal{F}\cdot\mathbf{1} = \mathbf{1}.\end{aligned} \tag{8}$$

The first term is to constrain the deviation of the prediction from the results of majority voting, while the second term is a graph-based manifold regularization to constrain the smoothness of the prediction over the graph. After we get the optimal $\mathcal{F}$, the final prediction is: $o_i \in \arg\max_{1\leq z\leq C}\mathcal{F}_{iz}$.

C. Optimization

The problem (8) is a convex quadratic programming problem and can be solved by the *label propagation* algorithm [13]. By initializing $\mathcal{F} = \mathcal{F}_0$, the following updating is repeated until convergence

$$\mathcal{F}_{\text{new}} = (1-\beta)\mathcal{F}_0 + \beta W\mathcal{F}_{\text{old}}. \tag{9}$$

The constraints $\mathcal{F} \geq 0$ and $\mathcal{F}\cdot\mathbf{1} = \mathbf{1}$ are naturally satisfied in the iteration, since $\mathcal{F}_0$ satisfies this constraint and W constructed by (4) is a degree-normalized symmetrical matrix. Following [13], we can prove the following theorems:

Theorem 1: The iteration of (9) will converge to:

$$\mathcal{F} = (1-\beta)(I - \beta W)^{-1}\mathcal{F}_0. \tag{10}$$

Proof: From the iteration of Eq. (9), we get:

$$\mathcal{F} = \lim_{n\to\infty}\left\{(\beta W)^n\widetilde{\mathcal{F}} + (1-\beta)\sum_{i=1}^{n}(\beta W)^{i-1}\mathcal{F}_0\right\},$$

where $\widetilde{\mathcal{F}}$ is the initialization. Because W is the normalized graph and $0 \leq \beta < 1$, we have $\lim_{n\to\infty}(\beta W)^n = 0$. Moreover, $\lim_{n\to\infty}\sum_{i=1}^{n}(\beta W)^{i-1} = (I-\beta W)^{-1}$. We get $\mathcal{F} = (1-\beta)(I-\beta W)^{-1}\mathcal{F}_0$. ∎

Theorem 2: Eq. (10) is the optimal solution for model (8).

Proof: Because (10) is the convergence of (9), it satisfies the constraints in (8). Let $\partial\mathcal{J}/\partial\mathcal{F} = 0$, we get:

$$\begin{aligned}\frac{\partial\mathcal{J}}{\partial\mathcal{F}} &= 2(1-\beta)(\mathcal{F}-\mathcal{F}_0) + \beta L\mathcal{F} + \beta L^\top\mathcal{F} \\ &= 2\mathcal{F} - 2\beta W\mathcal{F} - 2(1-\beta)\mathcal{F}_0 = 0.\end{aligned}$$

Hence, we have $\mathcal{F} = (1-\beta)(I-\beta W)^{-1}\mathcal{F}_0$. ∎

In the label propagation process of (9), the majority voting result $\mathcal{F}_0$ is propagated to different objects according to the object-similarity graph W to adaptively satisfy both the classification and the clustering results.

D. Incorporating Labeled Information

In previous sections, the true labels of all the objects are unknown. In practice, the labeled information can be used to further improve the accuracy in the sense of semi-supervised learning [7]. Suppose we have K additional labeled objects at the last rows of $\mathcal{M}$. Their ground truth is given as $\mathcal{T} \in \mathbb{R}^{K\times C}$ (0-1 matrix indicating the true labels). Similar to W, we can now define a bipartite graph between the unlabeled objects and the labeled objects $\widehat{B} \in \mathbb{R}^{N\times K}$ as:

$$\widehat{B}_{ij} = \sum_{k=C+1}^{G}\mathcal{M}_{ik}\mathcal{M}_{N+j,k} \quad i = 1\cdots N,\ j = 1\cdots K.$$

Since $\widehat{B}$ is a bipartite graph, we can normalize it to have unit row-degree. Let $D \in \mathbb{R}^{N\times N}$ be a diagonal matrix with $D_{ii} = \sum_{j=1}^{K}\widehat{B}_{ij}$. The normalized bipartite graph is now $B = D^{-1}\widehat{B}$.

Now we should seek a balance between the original objective (8) and the consistence of the clustering relationship between the labeled objects and the unlabeled objects, which leads to a semi-supervised extension of (8):

$$\begin{aligned}\min_{\mathcal{F}\in\mathbb{R}^{N\times C}} \quad & (1-\gamma)\mathcal{J} + \gamma\sum_{i=1}^{N}\sum_{j=1}^{K}B_{ij}\|\mathcal{F}_i - \mathcal{T}_j\|_2^2 \\ \text{s.t.} \quad & \mathcal{F} \geq 0, \mathcal{F}\cdot\mathbf{1} = \mathbf{1},\end{aligned} \tag{11}$$

where $0 \leq \gamma \leq 1$ is a tradeoff parameter. A simple modification of (9) can be used to solve (11):

$$\mathcal{F}_{\text{new}} = (1-\gamma)\left[(1-\beta)\mathcal{F}_0 + \beta W\mathcal{F}_{\text{old}}\right] + \gamma B\mathcal{T}. \tag{12}$$

In this process, the prediction is adaptively propagated to satisfy the classification results $\mathcal{F}_0$, clustering constraint W, B, and the given ground truth $\mathcal{T}$ of some labeled objects.

IV. Experiments

We evaluate the proposed methods on the datasets from [7] and [10], including 11 classification tasks from three real world applications: 20 Newsgroups categorization, Cora research paper classification, and DBLP network. The classification models include logistic regression and SVM models while the clustering models are k-means and min-cuts. All the classification and clustering results can be displayed in the object-group co-occurrence matrix $\mathcal{M}$ and the purpose is to fuse the information to make the final prediction.

On each of the 11 tasks, the test set is partitioned into two parts. One part is used to evaluate prediction accuracy and another part is used for semi-supervised models (i.e., their ground-truth is given as described in Section III-D). The experiments are repeated 50 times with random partition of test set. The models for comparison include: two classification and two clustering models (denoted by M1 to M4), BGCM and semi-supervised BGCM models [7], [8], UPE model [10], C3E model [1]. The proposed model of (9) is denoted as LP and (12) is denoted as semi-LP. We also give the results of the baselines in [7] including clustering ensemble models of MCLA [11] and HBGF [5].

TABLE II
PREDICTION ACCURACY (%) OF DIFFERENT MODELS. THE BEST RESULTS FOR EACH TASK ARE IN BOLDFACE

Methods	20 Newsgroup						Cora				DBLP	Average	Rank
	1	2	3	4	5	6	1	2	3	4	1		
M1	79.67	88.55	85.57	88.26	87.65	88.80	77.45	88.58	86.71	88.41	93.37	86.64	8.91
M2	77.21	86.11	81.34	86.76	83.58	85.63	77.97	85.94	85.08	88.79	87.66	84.19	10.50
M3	80.56	87.96	86.58	89.83	87.16	90.20	77.79	88.33	86.46	88.13	93.82	86.98	8.64
M4	77.70	85.71	81.49	84.67	85.43	85.78	74.76	85.94	78.10	90.16	79.49	82.66	10.68
MCLA	75.92	81.73	82.53	86.86	82.95	85.46	87.03	83.88	88.92	87.16	89.53	84.72	10.45
HBGF	81.99	92.44	88.11	91.52	89.91	91.25	78.34	91.11	84.81	89.43	93.57	88.41	6.91
BGCM	81.28	91.01	86.08	91.25	88.64	90.88	86.87	91.55	89.65	90.90	94.17	89.30	6.45
semi-BGCM	83.16	91.97	88.59	92.40	90.16	91.77	88.91	91.81	92.46	92.06	94.80	90.74	4.09
C3E	85.01	93.64	89.64	93.80	91.22	91.80	88.54	91.71	90.60	91.49	94.38	91.08	3.82
UPE	84.44	93.89	89.83	94.23	92.11	92.59	88.39	90.69	91.74	91.28	**95.41**	91.33	3.50
LP	84.86	93.92	91.32	94.23	92.09	92.38	89.31	91.36	**92.87**	92.36	94.09	91.71	2.91
semi-LP	**85.44**	**94.13**	**91.52**	**94.25**	**93.11**	**92.89**	**89.35**	**92.00**	**92.87**	**92.67**	95.36	**92.14**	**1.14**
std	0.50	0.22	0.25	0.21	0.22	0.24	0.49	0.56	0.18	0.28	0.09	–	–

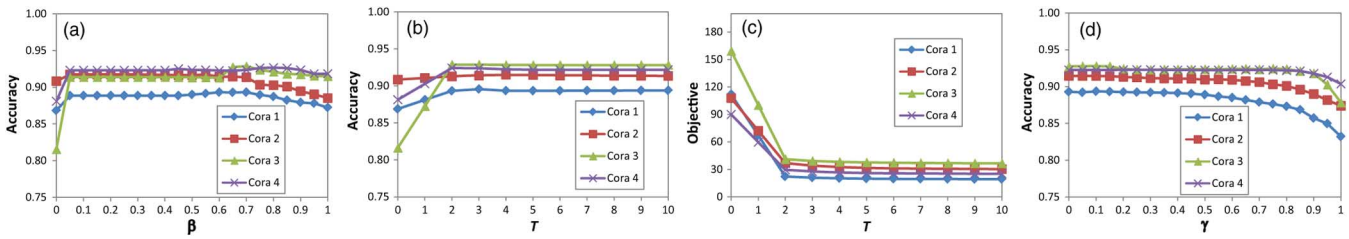

Fig. 1. (a) The sensitiveness of β. (b) The accuracy w.r.t. T. (c) The objective function w.r.t. T. (d) The sensitiveness of γ.

Table II summarizes the classification accuracies of all the baselines and the proposed methods on 11 tasks. We only show the standard deviation for semi-LP model in Table II. The highest accuracy of each task is bolded. The average accuracies over the 11 tasks are shown under the "Average" column, and the average ranks of different models are shown under the "Rank" column. We can find that: the two single classifiers (M1 and M2) and the two single clustering models (M3 and M4) usually have low accuracy. The clustering ensemble methods (MCLA and HBGF) can improve the performance over each single model. The accuracies of the classification and clustering ensemble models (BGCM, semi-BGCM, C3E, UPE, LP, and semi-LP) are significantly better than the base models and clustering ensembles. This demonstrates the power of combining classification and clustering results in accuracy improvements. Moreover, the proposed LP model shows superior performance consistently to the other models. By incorporating a small portion (around 10%) of labeled objects, the semi-LP model further improves the performances and shows significantly higher accuracy.

The hyper-parameters used in the proposed models are the trade-off price β in (8) and the number of iteration T used for (9). In our experiments, we set $\beta = 0.7$ and $T = 2$ for all the eleven tasks empirically. We now conduct experiments to analyze the sensitiveness of the parameters. We report the examination result on the data set of Cora, while the results are simply omitted on the other data sets due to the highly similar phenomenon. First, we fix $T = 2$ and change β from 0 to 1 with 0.05 as interval, the accuracy w.r.t. β can be found in Fig. 1(a). We can observe that the LP algorithm is not sensitive to β. When $\beta = 0$, LP reduces to the majority voting model; when $\beta > 0$, the clustering information is incorporated to adaptively adjust the classification results. We can see even when $\beta = 0.05$, the accuracy is greatly improved, which indicates the benefit of combining clustering results into classification models. When $0.05 \leq \beta \leq 0.7$, the performance does not change too much. When $\beta \geq 0.75$, the accuracy is decreased, since the clustering results dominate the base classification models. Second, we fix $\beta = 0.7$ and show the accuracy w.r.t. T in Fig. 1(b), the objective function in (8) w.r.t. T in Fig. 1(c). It is observed that the accuracy and objective function are almost not changed after $T = 2$. This indicates the LP algorithm (9) is a good method to find the optimal solution of model (8).

In the semi-LP model of (11), we have another hyper-parameter $\gamma \in [0, 1]$ to tradeoff between the unlabeled data and labeled data. When fixing $\beta = 0.7, T = 2$, we show the accuracy w.r.t γ in Fig. 1(d). We can find that the accuracy is not sensitive when $\gamma < 0.5$. However, the improvement when $\gamma > 0$ is not significant. Therefore, for the semi-LP model, we search the best β, γ from $[0, 1] \times [0, 1]$ for each task via cross-validation. Specifically, for each task we randomly partition the dataset into two parts (one for evaluation and the ground-truth of another one is given). Then the average accuracy of 10 times random partition is used to select the best β and γ for each task. After that, with the selected β and γ being fixed, we report the average accuracy and standard deviation (Table II) of another 50 times random partition of data.

V. CONCLUSION

By combining the outputs from multiple classification and clustering models, we can take advantage of the complementary information to derive a consolidated prediction. In this letter, a degree-normalized symmetrical object-similarity graph is constructed from multiple clustering results, and the classification results are then propagated on this graph. A semi-supervised propagation model is also proposed by incorporating a small portion of labeled objects. Experimental results on real applications identify the benefits of combining classification and clustering results, and the proposed models outperform other existing alternatives.

REFERENCES

[1] A. Acharya, E. Hruschka, J. Ghosh, and S. Acharyya, "C3E: A framework for combining ensembles of classifiers and clusters," in *Int. Workshop Multiple Classifier Systems*, 2011.

[2] L. Breiman, "Bagging predictors," *Mach. Learn.*, 1996.

[3] I. S. Dhillon and J. A. Tropp, "Matrix nearness problems with Bregman divergences," *SIAM J. Matrix Anal. Applicat.*, 2007.

[4] T. G. Dietterich and G. Bakiri, "Solving multiclass learning problems via error-correcting output codes," *J. Artif. Intell. Res.*, 1995.

[5] X. Z. Fern and C. E. Brodley, "Solving cluster ensemble problems by bipartitie graph partitioning," in *Int. Conf. Machine Learning*, 2004.

[6] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Int. Conf. Machine Learning*, 1996.

[7] J. Gao, F. Liang, W. Fan, Y. Sun, and J. Han, "Graph-based consensus maximization among multiple supervised and unsupervised models," *Adv. Neural Inf. Process. Syst.*, 2009.

[8] J. Gao, F. Liang, W. Fan, Y. Sun, and J. Han, "A graph-based consensus maximization approach for combining multiple supervised and unsupervised models," *IEEE Trans. Knowl. Data Eng.*, 2013.

[9] A. Gionis, H. Mannila, and P. Tsaparas, "Clustering aggregation," *ACM Trans. Knowl. Disc. Data*, 2007.

[10] X. Ma, P. Luo, F. Zhuang, Q. He, Z. Shi, and Z. Shen, "Combining supervised and unsupervised models via unconstrained probabilistic embedding," in *Int. Joint Conf. Artificial Intelligence*, 2011.

[11] A. Strehl and J. Ghosh, "Cluster ensembles–a knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, 2003.

[12] R. Zass and A. Shashua, "Doubly stochastic normalization for spectral clustering," *Adv. Neural Inf. Process. Syst.*, 2006.

[13] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Scholkopf, "Learning with local and global consistency," *Adv. Neural Inf. Process. Syst.*, 2003.