Multi-Modal Curriculum Learning over Graphs

CHEN GONG and JIAN YANG, PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and Jiangsu Key Lab of Image and Video Understanding for Social Security, School of Computer Science and Engineering, Nanjing University of Science and Technology, China
DACHENG TAO, UBTECH Sydney Artificial Intelligence Centre and the School of Computer Science, Faculty of Engineering and Information Technologies, the University of Sydney, Australia

Curriculum Learning (CL) is a recently proposed learning paradigm that aims to achieve satisfactory performance by properly organizing the learning sequence from simple curriculum examples to more difficult ones. Up to now, few works have been done to explore CL for the data with graph structure. Therefore, this article proposes a novel CL algorithm that can be utilized to guide the Label Propagation (LP) over graphs, of which the target is to "learn" the labels of unlabeled examples on the graphs. Specifically, we assume that different unlabeled examples have different levels of difficulty for propagation, and their label learning should follow a simple-to-difficult sequence with the updated curricula. Furthermore, considering that the practical data are often characterized by multiple modalities, every modality in our method is associated with a "teacher" that not only evaluates the difficulties of examples from its own viewpoint, but also cooperates with other teachers to generate the overall simplest curriculum examples for propagation. By taking the curriculums suggested by the teachers as a whole, the common preference (i.e., commonality) of teachers on selecting the simplest examples can be discovered by a row-sparse matrix, and their distinct opinions (i.e., individuality) are captured by a sparse noise matrix. As a result, an accurate curriculum sequence can be established and the propagation quality can thus be improved. Theoretically, we prove that the propagation risk bound is closely related to the examples' difficulty information, and empirically, we show that our method can generate higher accuracy than the state-of-the-art CL approach and LP algorithms on various multi-modal tasks.

CCS Concepts: • **Computing methodologies → Generative and developmental approaches**;

Additional Key Words and Phrases: Curriculum learning, multi-modal learning, semi-supervised learning, label propagation

This research is supported by NSF of China (Nos: 61602246, U1713208), NSF of Jiangsu Province (No: BK20171430), the Fundamental Research Funds for the Central Universities (No: 30918011319), the open project of State Key Laboratory of Integrated Services Networks (Xidian University, ID: ISN19-03), the "Summit of the Six Top Talents" Program (No: DZXX-027), the "Innovative and Entrepreneurial Doctor" Program of Jiangsu Province, the "Young Elite Scientists Sponsorship Program" by Jiangsu Province, the "Young Elite Scientists Sponsorship Program" by CAST (No: 2018QNRC001), the Program for Changjiang Scholars, and the Australian Research Council Projects (Nos: FL-170100117, DP-180103424, IH-180100002).

Authors' addresses: C. Gong and J. Yang, PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and Jiangsu Key Lab of Image and Video Understanding for Social Security, School of Computer Science and Engineering, Nanjing University of Science and Technology, 200 Xiaolingwei St, Nanjing, 210094, China; emails: {chen.gong; csjyang}@njust.edu.cn; D. Tao, UBTECH Sydney Artificial Intelligence Centre and the School of Computer Science, Faculty of Engineering and Information Technologies, the University of Sydney, 6 Cleveland St, Darlington, NSW 2008, Sydney, Australia; email: dacheng.tao@sydney.edu.au.

2157-6904/2019/07-ART35 $15.00
https://doi.org/10.1145/3322122

ACM Reference format:
Chen Gong, Jian Yang, and Dacheng Tao. 2019. Multi-Modal Curriculum Learning over Graphs. *ACM Trans. Intell. Syst. Technol.* 10, 4, Article 35 (July 2019), 25 pages.
https://doi.org/10.1145/3322122

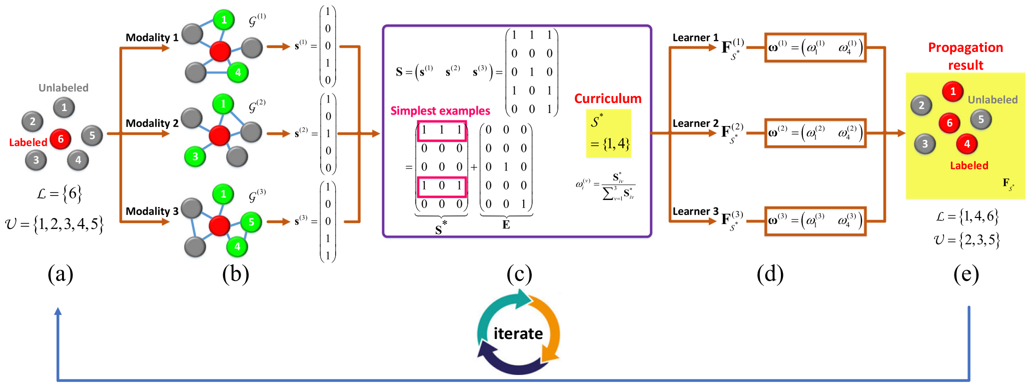
1 INTRODUCTION

Curriculum Learning (CL) is an emerging learning paradigm proposed by Bengio et al. [4], which mimics the cognitive process of humans and favors a learning algorithm to follow the logical learning sequence from simple examples to more difficult ones. In contrast to massive existing classifiers (e.g., Support Vector Machines and Naive Bayesian Classifier) that are trained on all examples at one time, CL establishes a sequence of curriculums so that only the optimal curriculum containing the simplest examples is employed to train the classifier in each learning round. Such "starting small" strategy is very similar to the human's knowledge acquisition process from childhood to adulthood, and has also been demonstrated to be effective in machine learning [14, 17, 20, 26, 43] and computer vision [3, 15, 31, 45, 57].

Since the notion of CL was proposed in [4], a series of explorations on CL algorithm design or phycological foundation have been made so far. Khan et al. [23] provide the cognitive evidence of CL by observing how humans teach a robot to grasp a new concept. Tu and Honavar [48] apply CL to probabilistic grammars learning by assuming that the structured probabilistic grammar can be decomposed as a sequence of grammatical rules. Pentina et al. [41] adapt CL to multi-task learning by advancing the learning of the most correlated tasks rather than dealing with all tasks jointly. Besides, Kumar et al. [26] propose the "Self-Paced Learning" (SPL), which employs latent SVMs as a learner and considers an example as simple if it lies far from the potential decision boundary. Jiang et al. [20] formulate SPL and CL into a unified framework by harnessing the dynamic example difficulty revealed during learning in addition to the estimation of difficulty before learning, and they term their algorithm as "Self-Paced Curriculum Learning".

However, existing CL methods are not applicable to the data with graph structure that is actually ubiquitous in many real-world situations. A graph is denoted by $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$, where $\mathcal{V}$ is the node set comprised of all data points and $\mathcal{E}$ is the edge set encoding the pairwise similarities between them. For example, in a social network, every user is a node and their interactions can be modeled by the edges indicating their personal relationship. In an image, pixels or compact regions can be regarded as a set of nodes and their similarities in the feature space are explicitly represented by the weighted edges. In an airline network, each airport constitutes a node and the flight route between two airports indicates the existence of an edge between them. Therefore, as a fundamental data structure, a graph has been widely utilized to analyze various practical data for the purpose of classification [59], regression [53], and clustering [36, 40].

Due to the popularity of graph structure, this article aims to develop a CL approach that is suitable for classifying the networked data. Specifically, we put our graph-based CL into the context of semi-supervised Label Propagation (LP) [13, 46, 60], which works on a graph $\mathcal{G}$ and aims to classify a massive number of unlabeled examples in the presence of very few labeled examples. LP follows the scheme of transduction originally proposed by Gammerman et al. [8] and Vapnik [49], which means that we are only interested in classifying a particular set of unlabeled examples rather than obtaining a general decision function for partitioning the entire example space. Mathematically, we have totally $n = l + u$ examples $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_l, \mathbf{x}_{l+1}, \ldots, \mathbf{x}_n\}$ represented by n nodes in $\mathcal{G}$, where the first l elements constitute the labeled set $\mathcal{L}$ and the remaining u examples form the unlabeled set $\mathcal{U}$ (see Figure 1(a)). The target of LP is to iteratively propagate the labels $\{y_i\}_{i=1}^{l}$ of $\mathcal{L}$ to the unlabeled set $\mathcal{U}$ along the edges of $\mathcal{G}$. To achieve this target, massive algorithms have


Unlabeled
Labeled
1
2
3
4
5
6
$\mathcal{L} = \{6\}$
$\mathcal{U} = \{1,2,3,4,5\}$
(a)
Modality 1
Modality 2
Modality 3
$\mathcal{G}^{(1)}$
$\mathcal{G}^{(2)}$
$\mathcal{G}^{(3)}$
$\mathbf{s}^{(1)} = (1\ 0\ 0\ 1\ 0)^T$
$\mathbf{s}^{(2)} = (1\ 0\ 1\ 0\ 0)^T$
$\mathbf{s}^{(3)} = (1\ 0\ 0\ 1\ 1)^T$
(b)
$\mathbf{S} = (\mathbf{s}^{(1)}\ \mathbf{s}^{(2)}\ \mathbf{s}^{(3)})$
Simplest examples
$\mathbf{S}^*$
$\mathbf{E}$
Curriculum
$\mathcal{S}^* = \{1,4\}$
$\omega_i^{(v)} = \frac{\mathbf{S}^*_{iv}}{\sum_{v=1}^{3} \mathbf{S}^*_{iv}}$
(c)
Learner 1
Learner 2
Learner 3
$\mathbf{F}^{(1)}_{\mathcal{S}^*}$
$\mathbf{F}^{(2)}_{\mathcal{S}^*}$
$\mathbf{F}^{(3)}_{\mathcal{S}^*}$
$\boldsymbol{\omega}^{(1)} = (\omega_1^{(1)}\ \omega_4^{(1)})$
$\boldsymbol{\omega}^{(2)} = (\omega_1^{(2)}\ \omega_4^{(2)})$
$\boldsymbol{\omega}^{(3)} = (\omega_1^{(3)}\ \omega_4^{(3)})$
(d)
Propagation result
Unlabeled
Labeled
$\mathbf{F}_{\mathcal{S}^*}$
$\mathcal{L} = \{1,4,6\}$
$\mathcal{U} = \{2,3,5\}$
(e)
iterate

$$\mathbf{S} = (\mathbf{s}^{(1)}\ \ \mathbf{s}^{(2)}\ \ \mathbf{s}^{(3)}) = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} = \underbrace{\begin{pmatrix} \boxed{1\ \ 1\ \ 1} \\ 0\ \ 0\ \ 0 \\ 0\ \ 0\ \ 0 \\ \boxed{1\ \ 0\ \ 1} \\ 0\ \ 0\ \ 0 \end{pmatrix}}_{\mathbf{S}^*} + \underbrace{\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{\mathbf{E}}$$

Fig. 1. The framework of our SMMCL algorithm. In (a), the examples in labeled set $\mathcal{L}$ and unlabeled set $\mathcal{U}$ are denoted by red and gray balls, respectively. In (b), the vth ($v = 1, 2, 3$ in this figure) teachers should choose the simplest examples (green balls) from their own modalities based on the graphs $\mathcal{G}^{(v)}$, and the selected curriculum examples are encoded in the selection vectors $\mathbf{s}^{(v)}$. In (c), the selection vectors produced by all teachers are put together to form a stacked matrix $\mathbf{S}$, which can be regarded as the sum of a row-sparse matrix $\mathbf{S}^*$ and a sparse noise matrix $\mathbf{E}$. The non-zero rows of $\mathbf{S}^*$ indicated by the magenta boxes correspond to the simplest examples that should be taken into the curriculum $\mathcal{S}^*$. Besides, the (i, v)-th element of $\mathbf{S}^*$ indicates the weight $\omega_i^{(v)}$ of the vth modality on deciding the label of the ith curriculum example. In (d), the learners (i.e., propagation algorithms) handle the examples in $\mathcal{S}^*$ by propagating the label information to them, and the resultant label matrices are $\mathbf{F}^{(1)}_{\mathcal{S}^*}, \mathbf{F}^{(2)}_{\mathcal{S}^*}, \mathbf{F}^{(3)}_{\mathcal{S}^*}$. In (e), a unified label matrix $\mathbf{F}_{\mathcal{S}^*}$ is obtained by adding $\mathbf{F}^{(1)}_{\mathcal{S}^*} \sim \mathbf{F}^{(3)}_{\mathcal{S}^*}$ weighted by $\boldsymbol{\omega}^{(1)} \sim \boldsymbol{\omega}^{(3)}$. The labeled set $\mathcal{L}$ and unlabeled set $\mathcal{U}$ are also updated accordingly. This process iterates until all the unlabeled examples are used up.

been proposed such as "Gaussian Field and Harmonic Functions" (GFHF) [59], "Local and Global Consistency" (LGC) [58], "Linear Neighbourhood Propagation" (LNP) [51], and "Dynamic Label Propagation" (DLP) [50], "Mean Teacher" [47], "Virtual Adversarial Training" [37], and the like.

There are two reasons for us to formulate the graph-based CL as a LP problem. First, LP is mathematically concise, widely used, and easy to implement. Second, the existing LP algorithms contain the defect that they are very likely to be misled by ambiguous examples such as "outliers" or "bridge points" [12]. This is because they ignore the propagation difficulty or reliability of unlabeled examples, so the unlabeled examples are classified in an imperfect order, leading to the error-prone classifications of difficult but critical examples [14]. To be specific, whether an unlabeled example $\mathbf{x}_i \in \mathcal{U}$ should be propagated is completely governed by the adjacency relationship between it and the labeled set $\mathcal{L}$. As long as $\mathbf{x}_i$ is directly connected to $\mathcal{L}$ by an edge, it will be classified in the current propagation by blindly receiving the labels from $\mathcal{L}$ without considering whether such propagation is difficult or dangerous. Consequently, the erroneous classifications will probably occur if the difficult examples are incorrectly activated to receive the propagated labels. At this time, CL can be a powerful tool for optimizing the traditional propagation sequence, so that an improved propagation performance can be obtained.

In our method, we assume that different unlabeled examples may have different levels of difficulty and deploy CL to re-arrange the propagation process from simple examples to more difficult ones. Furthermore, it has been intensively observed that an example in real life is often represented by multiple modalities [9, 27–30, 35, 52, 55]. For example, an image can be characterized by different features such as color and texture, and a webpage usually contains both image and text content. Therefore, a multi-modal CL approach is developed in this article to comprehensively exploit the information revealed by different modalities, so that the difficulty of every unlabeled example can be accurately evaluated. To be specific, we associate each modality with a "teacher", who first evaluates the difficulty of an unlabeled example from its own modality and then compromises its choice with other teachers' selections to a consistent result. Thanks to the efforts of multiple teachers, the well-organized propagation sequence enables the previously attained simple knowledge to facilitate the subsequent classification of complex examples, so all the unlabeled examples in $\mathcal{G}$ can be reliably labeled.

To the best of our knowledge, multi-modal CL was studied only once by our prior work [16]. However, Gong et al. [16] directly minimizes the error between the curriculum of every modality and the central optimal curriculum, which is a "hard" constraint suppressing the individuality possessed by every modality. Such imperfect fusion scheme degrades the curriculum quality and is unfavorable to obtaining satisfactory classification results. To address this defect, this article explicitly models the commonality among all the constituted modalities as well as their individualities to achieve a "soft" curriculum fusion, so our algorithm is dubbed "Soft Multi-Modal Curriculum Learning" (SMMCL). Specifically, we assume that the curriculums of multiple modalities as a whole can be decomposed as a row-sparse component plus a sparse noise component, in which the row-sparse component describes the commonality shared by multiple modalities and the noise component captures the individuality carried out by each of the modalities. As a result, the involved teachers are more easily to reach an agreement on selecting the simplest examples, and the selected curriculum examples are also more accurate than those produced by [16]. Besides, the byproducts of SMMCL are that in each learning round (i.e., propagation in our setting) the amount of selected examples and the fusion weights of different modalities on these examples are automatically determined, which are better than the ad-hoc or heuristic way employed by [16].

This submission is the extended version of our previous conference paper [11]. Compared with [11], this submission contains the following improvements:

- *Model:* (1) Instead of using the hard constraint to manually specify the number of curriculum examples in each iteration, here we introduce a novel regularization term to the curriculum selection model to encourage the teachers to adaptively select as many curriculum examples as possible in each learning round; (2) Instead of incorporating the $\{0, 1\}$-binary constraint to the objective function with a large weighting parameter, here we relax it to a real-valued constraint to make the solution numerically stable; and (3) We develop a novel regularized propagation model that is compatible with our multi-modal CL setting to classify the curriculum examples. These improvements enhance the propagation accuracy as well as decrease the number of learning rounds required by [11].
- *Theory*: We theoretically prove the convergence of the iterative optimization process for solving the curriculum selection model. Besides, we also present an error bound for SMMCL, which explicitly demonstrates the importance of considering the examples' levels of difficulty for LP problem.
- *Experiment*: More empirical studies are conducted including the verifications on the motivation of SMMCL, the experiments on more challenging image datasets, the applications on remote sensing, and the sensitivity analyses on free tuning parameters.

2 FRAMEWORK OF OUR METHOD

This section briefly introduces the framework of our proposed SMMCL algorithm. SMMCL contains two roles: the "teacher" that decides the optimal curriculum in every learning round, and the "learner" which is a propagation model for labeling the curriculum examples suggested by teacher.

For the multi-modal CL case studied in this article, we assume that each example $\mathbf{x}_i$ is characterized by V different modalities, so V graphs $\mathcal{G}^{(1)}, \ldots, \mathcal{G}^{(V)}$ are built with the vth ($v = 1, 2, \ldots, V$) graph corresponding to the vth modality (see Figure 1(b)). In these graphs, the nodes represent n examples and the edges depict the similarities between these examples. Furthermore, we associate each modality with a teacher and a learner, and in each learning round the V teachers should pick up the overall simplest examples (denoted by the set $\mathcal{S}^*$) for the stepwise learners. To this end, the vth ($v = 1, \ldots, V$) teacher should generate an optimal curriculum from its own viewpoint that is recorded by a $\{0, 1\}$-binary selection vector $\mathbf{s}^{(v)}$. Here, $\mathbf{s}_i^{(v)} = 1$ if the ith example is considered simple and is chosen by the vth teacher, and $\mathbf{s}_i^{(v)} = 0$ otherwise. After that, the decisions made by all V teachers are integrated into a unified curriculum $\mathcal{S}^*$ (see Figure 1(c)), during which the commonality of the teachers and their individualities are discovered by the row-sparse matrix $\mathbf{S}^*$ and sparse noise matrix $\mathbf{E}$, accordingly. Given $\mathcal{S}^*$, the V learners will classify the totally s examples in $\mathcal{S}^*$ by respectively propagating the labels from $\mathcal{L}$ to $\mathcal{S}^*$ from V modalities, and the obtained results are recorded in the label matrices $\mathbf{F}_{\mathcal{S}^*}^{(v)} \in \mathbb{R}^{s \times c}$ ($v = 1, \ldots, V$, and c is the number of classes) (see Figure 1(d)). The ith row of $\mathbf{F}_{\mathcal{S}^*}^{(v)}$ (denoted as $(\mathbf{F}_{\mathcal{S}^*}^{(v)})_{i,:}$) is the label vector of the example $\mathbf{x}_i$ with its jth element (i.e., $(\mathbf{F}_{\mathcal{S}^*}^{(v)})_{ij}$) encoding the probability of the ith curriculum example belonging to the jth ($j = 1, \ldots, c$) class. Finally, the propagation results $\mathbf{F}_{\mathcal{S}^*}^{(1)}, \ldots, \mathbf{F}_{\mathcal{S}^*}^{(V)}$ are fused into $\mathbf{F}_{\mathcal{S}^*}$ by considering their weights $\boldsymbol{\omega}^{(v)}$ on all the curriculum examples (see Figure 1(e)). The labeled set and unlabeled set are then updated by $\mathcal{L} := \mathcal{L} \cup \mathcal{S}^*$ and $\mathcal{U} := \mathcal{U} - \mathcal{S}^*$, respectively. Such teaching and learning process iterates until the unlabeled set becomes empty, namely $\mathcal{U} = \varnothing$.

3 CURRICULUM SELECTION

The most critical problem in our SMMCL algorithm is how to decide the simplest curriculum examples for the current propagation. This section first introduces the curriculum selection model for a single teacher, and then details how the multiple teachers work together to generate the overall optimal curriculum $\mathcal{S}^*$.

3.1 Single-Modal Curriculum Selection

For the vth (v takes a value from $1, 2, \ldots, V$) modality, we build a K Nearest Neighborhood (KNN) graph $\mathcal{G}^{(v)}$ via the technique of adaptive edge weighting [21], and the associated adjacency matrix is $\mathbf{W}^{(v)}$ with the (i, j)-th element $\mathbf{W}_{ij}^{(v)}$ representing the similarity between examples $\mathbf{x}_i$ and $\mathbf{x}_j$ on the vth modality. The graph Laplacian is $\mathbf{L}^{(v)} = \mathbf{D}^{(v)} - \mathbf{W}^{(v)}$ where $\mathbf{D}^{(v)}$ is the diagonal degree matrix with the diagonal elements $\mathbf{D}_{ii}^{(v)} = \sum_{j=1}^{n} \mathbf{W}_{ij}^{(v)}$. Based on $\mathcal{G}^{(v)}$, we may design a curriculum selection model to wisely determine the suitable $\mathcal{S}^{(v)} \subseteq \mathcal{U}$. For the simplicity of presentation, the superscript "(v)" of the related notations is temporarily dropped in this section.

Under single modality, the difficulty level of an unlabeled example $\mathbf{x}_i \in \mathcal{U}$ can be evaluated by its reliability and discriminability. By taking y_i as a random variable of $\mathbf{x}_i$ and treating the propagations on $\mathcal{G}^{(v)}$ as a Gaussian process over the random vector $\mathbf{y} = (y_1, \ldots, y_n)^\top$ [61], the reliability of $\mathbf{x}_i$ is modeled by the conditional entropy $H(y_i|\mathbf{y}_\mathcal{L})$, where $\mathbf{y}_\mathcal{L}$ is the subvector of $\mathbf{y}$ corresponding to the labeled set $\mathcal{L}$. Therefore, by using the property of multivariate Gaussian [5], we have

$$H(y_i|\mathbf{y}_\mathcal{L}) \propto |\Sigma_{i|\mathcal{L}}| = \Sigma_{i,i} - \Sigma_{i,\mathcal{L}} \Sigma_{\mathcal{L},\mathcal{L}}^{-1} \Sigma_{\mathcal{L},i}, \tag{1}$$

where Σ is the covariance matrix of the random vector $\mathbf{y}$, and $\Sigma_{i,i}, \Sigma_{i,\mathcal{L}}, \Sigma_{\mathcal{L},i}, \Sigma_{\mathcal{L},\mathcal{L}}$ are submatrices of Σ associated with the corresponding subscripts. The covariance matrix Σ is defined by $\Sigma = (\mathbf{L} + \mathbf{I}/\vartheta^2)^{-1}$ where $\mathbf{I}$ is identity matrix, and ϑ^2 is the parameter governing the "sharpness" of the

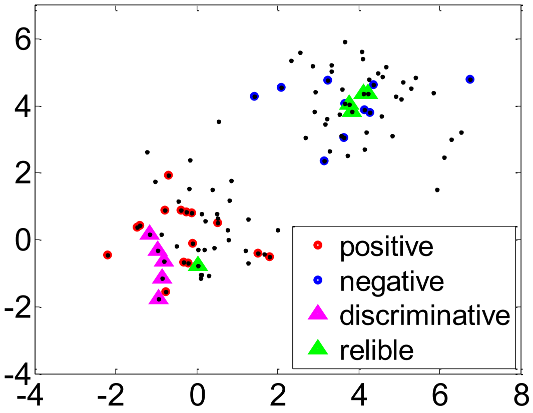


Fig. 2. An illustration of reliability and discriminability criteria. This dataset consists of two Gaussian centered at $(0, 0)$ and $(2.5, 2.5)$, respectively, and their covariance is $(0, 1; 1, 0)$. Each Gaussian represents a class with 50 data points, and the red and blue circles denote the labeled positive examples and labeled negative examples correspondingly. Five most discriminative points are displayed by magenta triangles, and the green triangles depict the five most reliable examples.

distribution that is fixed to 100 throughout this article. A small $H(y_i|\mathbf{y}_{\mathcal{L}})$ means that $\mathbf{x}_i$ comes as no "surprise" to the labeled set $\mathcal{L}$, so classifying $\mathbf{x}_i$ is reliable and it should be incorporated by the curriculum $\mathcal{S}^*$.

The discriminability of $\mathbf{x}_i$ depicts its tendency belonging to a certain class, which is modeled by the difference of average commute time [42] from $\mathbf{x}_i$ to its two nearest classes C_1 and C_2, namely

$$M(\mathbf{x}_i) = \bar{T}(\mathbf{x}_i, C_2) - \bar{T}(\mathbf{x}_i, C_1), \tag{2}$$

where $\bar{T}(\mathbf{x}_i, C_j)$ computes the average commute time between $\mathbf{x}_i$ and all the examples of class C_j $(j = 1, 2)$. A large $M(\mathbf{x}_i)$ means that $\mathbf{x}_i$ is significantly inclined to the class C_1 and thus it is ideal to be a curriculum example.

In Equation (2), the average commute time $\bar{T}(\mathbf{x}_i, C_j)$ $(j = 1, 2)$ is computed as

$$\bar{T}(\mathbf{x}_i, C_j) = \frac{1}{n_{C_j}} \sum_{\mathbf{x}_{i'} \in C_j} T(\mathbf{x}_i, \mathbf{x}_{i'}), \tag{3}$$

where n_{C_j} denotes the number of examples in the class C_j; $T(\mathbf{x}_i, \mathbf{x}_{i'})$ is the commute time between examples $\mathbf{x}_i$ and $\mathbf{x}_{i'}$, which is defined by [42] as

$$T(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_{k=1}^{n} h(\lambda_k)(u_{ki} - u_{ki'})^2, \tag{4}$$

where $0 = \lambda_1 \leq \cdots \leq \lambda_n$ are the eigenvalues of $\mathbf{L}$, and $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are the associated eigenvectors; u_{ki} denotes the ith element of $\mathbf{u}_k$; $h(\lambda_k) = 1/\lambda_k$ if $\lambda_k \neq 0$, and $h(\lambda_k) = 0$ otherwise.

Figure 2 visualizes the effects of reliability and discriminability criteria on selecting the simplest examples. We see that the reliability criterion tends to select the examples within the dense region of a class, while the discriminability criterion is more likely to choose the examples that are near to one class but far from another one. By taking account of $\mathbf{x}_i$'s reliability and discriminability together, the difficulty of $\mathbf{x}_i$ is represented by

$$R(\mathbf{x}_i) = \zeta \cdot H(y_i|\mathbf{y}_{\mathcal{L}}) + 1/M(\mathbf{x}_i), \tag{5}$$

where ζ is a trade-off parameter fixed to 1. The example with small $R(\mathbf{x}_i)$ is simple and is suitable for the current propagation conducted by the learner.

Suppose there are τ unlabeled examples that are directly connected to $\mathcal{L}$ in $\mathcal{G}$, and then the teacher should choose a certain amount of simplest examples from these τ examples. To this end, we utilize the binary selection vector $\mathbf{s} \in \{0, 1\}^{\tau \times 1}$ defined in Section 2, and thus the curriculum generation model for a single teacher is formulated as:

$$\begin{aligned} &\min_{\mathbf{s}} \ \mathbf{s}^{\top}\mathbf{R}\mathbf{s} - \gamma \mathbf{1}^{\top}\mathbf{s} \\ &s.t. \ \ \mathbf{s}_i \in \{0, 1\}, \forall i = 1, 2, \ldots, \tau, \end{aligned} \tag{6}$$

where $\mathbf{R}$ is a diagonal matrix with the ith diagonal element being $R(\mathbf{x}_i)$ defined in Equation (5), $\mathbf{1}$ is the all-one column vector with the same length as $\mathbf{s}$, and $\gamma > 0$ is a free parameter.

In Equation (6), minimizing the first term requires $\mathbf{s}_i$ to be small if the difficulty value $R(\mathbf{x}_i)$ of $\mathbf{x}_i$ is large, which means that $\mathbf{x}_i$ is not recommended to be a curriculum example. Minimizing the second term encourages the teacher to choose as many curriculum examples as possible for each learning round.

3.2 Multi-Modal Curriculum Selection

Given the individual decision $\mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(V)}$ made by each of the V teachers, we need to find a way to make the teachers maximally reach an agreement on determining the suitable $\mathcal{S}^*$. However, different teachers often have different selections on the simplest examples, as the difficulties of an example revealed by different modalities are distinct. Therefore, we should exploit both the underlying consensus of all V teachers and their individual opinions when fusing $\mathbf{s}^{(1)}, \mathbf{s}^{(2)}, \ldots, \mathbf{s}^{(V)}$. To this end, we put the binary selection vectors $\mathbf{s}^{(v)}$ $(v = 1, \ldots, V)$ of the V teachers altogether as a matrix $\mathbf{S} = (\mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(V)})$, and then its all-zero rows will indicate the difficult examples considered by all V teachers. Practically, the teachers can hardly draw the identical conclusion on deciding $\mathcal{S}^*$, so we assume that $\mathbf{S}$ implicitly contains a row-sparse component $\mathbf{S}^*$ representing the consensus of all teachers, and a sparse noise term $\mathbf{E}$ capturing the individuality of each modality. As a result, we have $\mathbf{S} = \mathbf{S}^* + \mathbf{E}$ with $\mathbf{S}$, $\mathbf{S}^*$ and $\mathbf{E}$ being $\{0, 1\}$-binary matrices (see Figure 1(c)). Thereby, our model is formulated as

$$\begin{aligned} &\min_{\mathbf{S}, \mathbf{S}^*, \mathbf{E}} \sum_{v=1}^{V} \left(\mathbf{s}^{(v)\top}\mathbf{R}^{(v)}\mathbf{s}^{(v)} - \gamma \mathbf{1}^{\top}\mathbf{s}^{(v)} \right) + \alpha \|\mathbf{S}^*\|_{2,1} + \beta \|\mathbf{E}\|_1 \\ &s.t. \ \ \mathbf{S} = \mathbf{S}^* + \mathbf{E}, \ \mathbf{S}_{ij} \in \{0, 1\}, \ \mathbf{S}^*_{ij} \in \{0, 1\}, \ \mathbf{E}_{ij} \in \{0, 1\}, \end{aligned} \tag{7}$$

where $\|\mathbf{S}^*\|_{2,1} = \sum_i \sqrt{\sum_j \mathbf{S}^{*2}_{ij}}$ computes $\mathbf{S}^*$'s $l_{2,1}$ norm, $\|\mathbf{E}\|_1 = \sum_{i,j} |\mathbf{E}_{ij}|$ is the l_1 norm of matrix $\mathbf{E}$, and $\alpha, \beta, \gamma > 0$ are tradeoff parameters.

In the objective function of Equation (7), the first summation term follows Equation (6) and considers the isolated example selection of every single modality. The second term utilizes the $l_{2,1}$ norm of $\mathbf{S}^*$ to discover the common decision made by all the teachers from different modalities. The third term models the unique opinion of each teacher and minimizing it drives all teachers to reach an agreement as possible as they can. Compared with the model in [16] that rashly forces all $\mathbf{s}^{(v)}$ $(v = 1, \ldots, V)$ to a compromised $\mathbf{s}^*$ by minimizing $\|\mathbf{s}^{(v)} - \mathbf{s}^*\|_2^2$, Equation (7) developed here tries to discover the underlying consensus among different teachers as well as explicitly preserves the individuality of every teacher, so it achieves "soft" fusion of multiple modalities without loosing their specialities. Therefore, our approach can produce better results than [16], which will be demonstrated in the experiments.

3.3 Solution of Equation (7)

Due to the binary constraints on S, S* and E in Equation (7), directly solving Equation (7) is difficult as it is NP-hard. To make Equation (7) tractable, we relax such binary constraints to continuous nonnegative constraints and achieve the following expression:

$$\begin{aligned} &\min_{\mathbf{S},\mathbf{S}^*,\mathbf{E}} \sum_{v=1}^{V} \left(\mathbf{s}^{(v)\top}\mathbf{R}^{(v)}\mathbf{s}^{(v)} - \gamma \mathbf{1}^\top \mathbf{s}^{(v)}\right) + \alpha \|\mathbf{S}^*\|_{2,1} + \beta\|\mathbf{E}\|_1 \\ &s.t.\;\; \mathbf{S} = \mathbf{S}^* + \mathbf{E},\; \mathbf{S} \geq \mathbf{O},\; \mathbf{S}^* \geq \mathbf{O},\; \mathbf{E} \geq \mathbf{O}, \end{aligned} \tag{8}$$

where **O** represents the all-zero matrix.

The problem (8) is not convex w.r.t. S, S*, and E altogether; however, it can be easily proved that the subproblem for optimizing each of them is convex. Therefore, Equation (8) can be solved via the Alternating Direction Method of Multipliers (ADMM), which alternatively optimizes one variable at one time with the other variables remaining fixed. To decouple the variables S* and S, we introduce an auxiliary variable **J** and a related constraint **J** = S*, and thus the optimization problem (8) is reformulated as

$$\begin{aligned} &\min_{\mathbf{S},\mathbf{S}^*,\mathbf{E},\mathbf{J}} \sum_{v=1}^{V} \left(\mathbf{s}^{(v)\top}\mathbf{R}^{(v)}\mathbf{s}^{(v)} - \gamma \mathbf{1}^\top \mathbf{s}^{(v)}\right) + \alpha \|\mathbf{J}\|_{2,1} + \beta\|\mathbf{E}\|_1 \\ &s.t.\;\; \mathbf{S} = \mathbf{S}^* + \mathbf{E},\; \mathbf{J} = \mathbf{S}^*,\; \mathbf{S} \geq \mathbf{O},\; \mathbf{S}^* \geq \mathbf{O},\; \mathbf{E} \geq \mathbf{O}. \end{aligned} \tag{9}$$

Therefore, by introducing the auxiliary variables $\mathbf{T}_1, \mathbf{T}_2, \mathbf{T}_3$ to cope with the inequality constraints, the augmented Lagrangian function is

$$\begin{aligned} &L(\mathbf{S},\mathbf{S}^*,\mathbf{E},\mathbf{J},\Lambda_1,\Lambda_2,\mathbf{Z}_1,\mathbf{Z}_2,\mathbf{Z}_3,\mu_{in},\mu_{eq}) \\ &= \sum_{v=1}^{V} \left(\mathbf{s}^{(v)\top}\mathbf{R}^{(v)}\mathbf{s}^{(v)} - \gamma \mathbf{1}^\top \mathbf{s}^{(v)}\right) + \alpha\|\mathbf{J}\|_{2,1} + \beta\|\mathbf{E}\|_1 + \mathrm{tr}\left(\Lambda_1^\top(\mathbf{S}-\mathbf{S}^*-\mathbf{E})\right) + \mathrm{tr}\left(\Lambda_2^\top(\mathbf{J}-\mathbf{S}^*)\right) \\ &\quad + \mathrm{tr}\left(\mathbf{Z}_1^\top(\mathbf{S}-\mathbf{T}_1)\right) + \mathrm{tr}\left(\mathbf{Z}_2^\top(\mathbf{S}^*-\mathbf{T}_2)\right) + \mathrm{tr}\left(\mathbf{Z}_3^\top(\mathbf{E}-\mathbf{T}_3)\right) \\ &\quad + \frac{\mu_{eq}}{2}\left(\|\mathbf{S}-\mathbf{S}^*-\mathbf{E}\|_F^2 + \|\mathbf{J}-\mathbf{S}^*\|_F^2\right) + \frac{\mu_{in}}{2}\left(\|\mathbf{S}-\mathbf{T}_1\|_F^2 + \|\mathbf{S}^*-\mathbf{T}_2\|_F^2 + \|\mathbf{E}-\mathbf{T}_3\|_F^2\right), \end{aligned} \tag{10}$$

where $\Lambda_1, \Lambda_2, \mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3$ are Lagrangian multipliers, and $\mu_{eq} > 0$ and $\mu_{in} > 0$ are the penalty coefficients related to the equality constraints and inequality constraint, respectively. Based on Equation (10), the variables S, S*, E, J can be sequentially updated via an iterative way.

Update **J**: The subproblem related to **J** is

$$\min_{\mathbf{J}} \; \alpha\|\mathbf{J}\|_{2,1} + \mathrm{tr}\left(\Lambda_2^\top(\mathbf{J}-\mathbf{S}^*)\right) + \frac{\mu_{eq}}{2}\|\mathbf{J}-\mathbf{S}^*\|_F^2. \tag{11}$$

By dropping the terms irrelevant to **J**, Equation (11) is equivalent to

$$\min_{\mathbf{J}} \; \frac{\alpha}{\mu_{eq}}\|\mathbf{J}\|_{2,1} + \frac{1}{2}\left\|\mathbf{J} - \left(\mathbf{S}^* - \frac{1}{\mu_{eq}}\Lambda_2\right)\right\|_F^2, \tag{12}$$

of which the optimal solution is [34]

$$\mathbf{J}_{i,:} = \begin{cases} \dfrac{\|\mathbf{T}_{i,:}\|_2 - \alpha/\mu_{eq}}{\|\mathbf{T}_{i,:}\|_2}\mathbf{T}_{i,:}, & \alpha/\mu_{eq} < \|\mathbf{T}_{i,:}\|_2 \\ 0, & \text{otherwise} \end{cases}, \tag{13}$$

where $\mathbf{T} = \mathbf{S}^* - \frac{1}{\mu_{eq}}\Lambda_2$ and $\mathbf{T}_{i,:}$ denotes the *i*th row of **T**.

Update $\mathbf{S}^*$: By ignoring the unrelated terms to $\mathbf{S}^*$ in Equation (10), the subproblem regarding $\mathbf{S}^*$ is

$$\begin{aligned}\min_{\mathbf{S}^*} \ & \mathrm{tr}\left(\Lambda_1^\top(\mathbf{S}-\mathbf{S}^*-\mathbf{E})\right) + \mathrm{tr}\left(\Lambda_2^\top(\mathbf{J}-\mathbf{S}^*)\right) \\ & + \frac{\mu_{eq}}{2}\left(\|\mathbf{S}-\mathbf{S}^*-\mathbf{E}\|_{\mathrm{F}}^2 + \|\mathbf{J}-\mathbf{S}^*\|_{\mathrm{F}}^2\right) + \mathrm{tr}\left(\mathbf{Z}_2^\top(\mathbf{S}^*-\mathbf{T}_2)\right) + \frac{\mu_{in}}{2}\|\mathbf{S}^*-\mathbf{T}_2\|_{\mathrm{F}}^2 .\end{aligned} \tag{14}$$

By calculating the derivative of above objective to $\mathbf{S}^*$, and then setting the result to 0, the closed-form solution for $\mathbf{S}^*$ is

$$\mathbf{S}^* = \frac{1}{\mu_{in}+2\mu_{eq}}\left[\Lambda_1 + \Lambda_2 - \mu_{eq}(\mathbf{E}-\mathbf{S}-\mathbf{J}) - \mathbf{Z}_2 + \mu_{in}\mathbf{T}_2\right]. \tag{15}$$

Update $\mathbf{E}$: The subproblem for optimizing $\mathbf{E}$ is

$$\min_{\mathbf{E}} \beta\|\mathbf{E}\|_1 + \mathrm{tr}\left(\Lambda_1^\top(\mathbf{S}-\mathbf{S}^*-\mathbf{E})\right) + \frac{\mu_{eq}}{2}\|\mathbf{S}-\mathbf{S}^*-\mathbf{E}\|_{\mathrm{F}}^2 + \mathrm{tr}\left(\mathbf{Z}_3^\top(\mathbf{E}-\mathbf{T}_3)\right) + \frac{\mu_{in}}{2}\|\mathbf{E}-\mathbf{T}_3\|_{\mathrm{F}}^2 . \tag{16}$$

After ignoring the constant variables, Equation (16) is equivalent to

$$\begin{aligned} & \min_{\mathbf{E}} \ \beta\|\mathbf{E}\|_1 + \frac{\mu_{eq}+\mu_{in}}{2}\left[\mathrm{tr}(\mathbf{E}^\top\mathbf{E}) - \frac{2}{\mu_{eq}+\mu_{in}}\mathrm{tr}(\widetilde{\mathbf{B}}^\top\mathbf{E})\right] \\ \Leftrightarrow & \min_{\mathbf{E}} \ \beta\|\mathbf{E}\|_1 + \frac{\mu_{eq}+\mu_{in}}{2}\left\|\mathbf{E} - \frac{1}{\mu_{eq}+\mu_{in}}\widetilde{\mathbf{B}}\right\|_{\mathrm{F}}^2 ,\end{aligned} \tag{17}$$

where $\widetilde{\mathbf{B}} = \Lambda_1 - \mu_{eq}(\mathbf{S}^*-\mathbf{S}) - \mathbf{Z}_3 + \mu_{in}\mathbf{T}_3$. By further denoting $\mathbf{B} = \frac{1}{\mu_{eq}+\mu_{in}}\widetilde{\mathbf{B}}$, Equation (17) is reformulated as

$$\min_{\mathbf{E}} \ \beta\|\mathbf{E}\|_1 + \frac{\mu_{eq}+\mu_{in}}{2}\|\mathbf{E}-\mathbf{B}\|_{\mathrm{F}}^2 , \tag{18}$$

of which the solution can be easily found by employing the soft-thresholding operator [33], namely

$$\mathbf{E}_{ij} = \begin{cases} \mathbf{B}_{ij} - \dfrac{\beta}{\mu_{eq}+\mu_{in}}, & \mathbf{B}_{ij} > \dfrac{\beta}{\mu_{eq}+\mu_{in}} \\ \mathbf{B}_{ij} + \dfrac{\beta}{\mu_{eq}+\mu_{in}}, & \mathbf{B}_{ij} < \dfrac{-\beta}{\mu_{eq}+\mu_{in}} \\ 0, & \text{otherwise.} \end{cases} \tag{19}$$

Update $\mathbf{s}^{(v)}$: Note that the columns of $\mathbf{S}$ (i.e., $\mathbf{s}^{(v)}$, $v = 1, 2, \ldots, V$) are independent to each other in Equation (10), so they can be updated one by one, and then the objective regarding $\mathbf{s}^{(v)}$ is

$$\begin{aligned}\min_{\mathbf{s}^{(v)}} \ & \mathbf{s}^{(v)\top}\mathbf{R}^{(v)}\mathbf{s}^{(v)} - \gamma\mathbf{1}^\top\mathbf{s}^{(v)} + (\Lambda_1)_{:,v}^\top\mathbf{s}^{(v)} + (\mathbf{Z}_1)_{:,v}^\top\mathbf{s}^{(v)} \\ & + \frac{\mu_{eq}}{2}\left\|\mathbf{s}^{(v)} - \mathbf{S}_{:,v}^* - \mathbf{E}_{:,v}\right\|_2^2 + \frac{\mu_{in}}{2}\left\|\mathbf{s}^{(v)} - (\mathbf{T}_1)_{:,v}\right\|_2^2 ,\end{aligned} \tag{20}$$

where the subscript ":, v" denotes the vth column of the corresponding matrix. The solution of Equation (20) can be easily obtained by setting the derivative of Equation (20) to $\mathbf{s}^{(v)}$ to 0, which leads to

$$\mathbf{s}^{(v)} = \left[2\mathbf{R}^{(v)} + (\mu_{eq}+\mu_{in})\mathbf{I}\right]^{-1}\left[\gamma\mathbf{1} - (\Lambda_1)_{:,v} \ + \mu_{eq}(\mathbf{S}_{:,v}^* + \mathbf{E}_{:,v}) - (\mathbf{Z}_1)_{:,v} + \mu_{in}(\mathbf{T}_1)_{:,v}\right]. \tag{21}$$

Since $\mathbf{R}^{(v)}$ is a diagonal matrix with positive diagonal elements, the matrix $2\mathbf{R}^{(v)} + (\mu_{eq}+\mu_{in})\mathbf{I}$ in Equation (21) is positive definite and is always invertible.

The entire iterative process for solving Equation (9) is summarized in Algorithm 1. Its convergence will be theoretically proved in Section 5.1 and also empirically illustrated by the experiments. The indices of non-zero rows in the optimized $\mathbf{S}^*$ indicate the selected simplest examples for the

ALGORITHM 1: The ADMM process for solving Equation (9)

```
Input: $\mathbf{R}^{(v)}$, $\alpha$, $\beta$, $\gamma$, $\mu_{eq} = \mu_{in} = 1$, $\mu_{max} = 10^{10}$, $\rho = 1.2$, $\epsilon = 10^{-4}$, $MaxIter = 50$, initial $\mathbf{S}$, $\mathbf{E}$, $\mathbf{S}^*$.
set $\Lambda_1, \Lambda_2, \mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3$ to all-one matrices;
set $iter = 0$;
repeat
    // Update auxiliary matrices $\mathbf{T}_1, \mathbf{T}_2, \mathbf{T}_3$
    $\mathbf{T}_1 = \max(\mathbf{O}, \mathbf{S} + \mathbf{Z}_1/\mu_{in})$; $\mathbf{T}_2 = \max(\mathbf{O}, \mathbf{S}^* + \mathbf{Z}_2/\mu_{in})$; $\mathbf{T}_3 = \max(\mathbf{O}, \mathbf{E} + \mathbf{Z}_3/\mu_{in})$;
    Update $\mathbf{J}$ via Equation (13);
    Update $\mathbf{S}^*$ via Equation (15);
    Update $\mathbf{E}$ via Equation (19);
    // Update $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_V$ in parallel
    for $v = 1$ to $V$ do
        Update $\mathbf{s}^{(v)}$ via Equation (21);
    end for
    // Update Lagrangian multipliers
    $\Lambda_1 := \Lambda_1 + \mu_{eq}(\mathbf{S} - \mathbf{S}^* - \mathbf{E})$,   $\Lambda_2 := \Lambda_2 + \mu_{eq}(\mathbf{J} - \mathbf{S}^*)$;   $\mathbf{Z}_1 := \max(\mathbf{O}, \mathbf{Z}_1 - \mu_{in}\mathbf{S})$;   $\mathbf{Z}_2 := \max(\mathbf{O}, \mathbf{Z}_2 -$
    $\mu_{in}\mathbf{S}^*)$; $\mathbf{Z}_3 := \max(\mathbf{O}, \mathbf{Z}_3 - \mu_{in}\mathbf{E})$;
    // Update penalty coefficients
    $\mu_{in} := \min(\rho\mu_{in}, \mu_{max})$;
    $\mu_{eq} := \min(\rho\mu_{eq}, \mu_{max})$;
    $iter := iter + 1$;
until $\left\|\mathbf{S}^{*(iter)} - \mathbf{S}^{*(iter-1)}\right\|_\infty \le \epsilon$ or $iter = MaxIter$
Output: The optimal $\mathbf{S}, \mathbf{S}^*, \mathbf{E}$.
```

current propagation, and thus how many examples should be selected are also automatically determined (see Figure 1(c)).

4 LABEL PROPAGATION AND LABEL FUSION

Suppose there are totally s curriculum examples $\mathcal{S}^* = \{\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_s^*\}$ designated by the teachers, every learner will propagate the labels of $\mathcal{L}$ to these s examples from its associated modality. After that, the output label matrices $\mathbf{F}_{\mathcal{S}^*}^{(1)}, \ldots, \mathbf{F}_{\mathcal{S}^*}^{(V)}$ are properly integrated into a consistent $\mathbf{F}_{\mathcal{S}^*}$ as shown in Figures 1(d)(e).

The existing regularized propagation algorithms such as GFHF [59] and LGC [58] are not suitable to be our learner, as they only leverage the label information of the initial labeled examples, so the labels assigned to the curriculum examples during our CL-guided propagation process cannot be properly deployed. Therefore, we have to design a new regularized propagation model that fits our iterative CL setting. Moreover, we can also directly find the convergence result of the propagation process by solving the proposed regularization model. Suppose we are going to conduct the tth propagation to find the updated label matrix $\mathbf{F}^{(v)} \in [0,1]^{(l+s)\times c}$ on the vth modality, where the previously "learned" examples correspond to the first l rows of $\mathbf{F}^{(v)}$, and the last s rows record the labels of the curriculum examples in $\mathcal{S}^*$, and then given the unified label matrix $\widetilde{\mathbf{F}} \in [0,1]^{l\times c}$ output by the $(t-1)$-th learning round, the propagation model in our SMMCL is

$$\min_{\mathbf{F}^{(v)}} \mathcal{J}(\mathbf{F}^{(v)}) = \frac{1}{2}\left\{\sum_{i=1}^{l+s}\sum_{j=1}^{l+s}\left(\mathbf{W}_{\mathcal{Q},\mathcal{Q}}^{(v)}\right)_{ij}\left\|\mathbf{F}_{i,:}^{(v)} - \mathbf{F}_{j,:}^{(v)}\right\|_2^2 + \theta\left[\sum_{i=1}^{l}\sum_{j=1}^{c}\widetilde{\mathbf{F}}_{ij}\left(\mathbf{F}_{ij}^{(v)} - 1\right)^2 + (1-\widetilde{\mathbf{F}}_{ij})\mathbf{F}_{ij}^{(v)2}\right]\right\}, \quad (22)$$

where $\theta > 0$ is the tradeoff parameter, and c is the number of classes as defined in Section 2. The set $\mathcal{Q} = \mathcal{L} \bigcup \mathcal{S}^*$ and $\mathbf{W}_{\mathcal{Q},\mathcal{Q}}^{(v)}$ is the submatrix of $\mathbf{W}^{(v)}$ by preserving its rows and columns corresponding to $\mathcal{Q}$. The first term in Equation (22) is the *smoothness term*, which forces that the similar examples sharing a large edge weight are assigned similar labels after the propagation. The second term is the *soft fidelity term* which requires that the optimized $\mathbf{F}^{(v)}$ should be approach to the $(t-1)$-th classification on the previously labeled examples. This term is "soft" as we do not simply push $\mathbf{F}_{ij}^{(v)}$ to 1 or 0. Instead, the $(t-1)$-th real-valued soft label $\widetilde{\mathbf{F}}_{ij}$ has been taken into consideration to describe the confidence of $\mathbf{F}_{ij}^{(v)}$ to be 1 or 0. If $\widetilde{\mathbf{F}}_{ij}$ is very close to 1, which means that $\mathbf{x}_i$ belongs to the jth class with a high probability, the term $(\mathbf{F}_{ij}^{(v)}-1)^2$ will be assigned larger weight than $\mathbf{F}_{ij}^{(v)2}$, so that $\mathbf{F}_{ij}^{(v)}$ is preferred to be 1 rather than 0. In the first learning round (i.e., $t=1$), $\widetilde{\mathbf{F}}$ equals to the label matrix of initial labeled examples with $\widetilde{\mathbf{F}}_{ij}=1$ if $\mathbf{x}_i \in \mathcal{L}$ has the label j, and 0 otherwise.

To find the solution of Equation (22), we compute $\frac{d\mathcal{J}(\mathbf{F}^{(v)})}{d\mathbf{F}^{(v)}}$ and then set the result to zero matrix $\mathbf{O}_{(l+s)\times c}$, namely

$$\frac{d\mathcal{J}(\mathbf{F}^{(v)})}{d\mathbf{F}^{(v)}} = \mathbf{L}_{\mathcal{Q},\mathcal{Q}}^{(v)}\mathbf{F}^{(v)} + \theta\begin{pmatrix}\mathbf{F}_{\mathcal{L}}^{(v)}-\widetilde{\mathbf{F}}\\ \mathbf{O}_{s\times c}\end{pmatrix} = \mathbf{O}_{(l+s)\times c}, \tag{23}$$

where $\mathbf{F}_{\mathcal{L}}^{(v)}$ denotes the sub-matrix of $\mathbf{F}^{(v)}$, of which the rows correspond to the labeled set $\mathcal{L}$. Correspondingly, we partition the graph Laplacian $\mathbf{L}_{\mathcal{Q},\mathcal{Q}}^{(v)}$ as $\mathbf{L}_{\mathcal{Q},\mathcal{Q}}^{(v)} = \begin{pmatrix}\mathbf{L}_{\mathcal{L},\mathcal{L}}^{(v)} & \mathbf{L}_{\mathcal{L},\mathcal{S}^*}^{(v)}\\ \mathbf{L}_{\mathcal{S}^*,\mathcal{L}}^{(v)} & \mathbf{L}_{\mathcal{S}^*,\mathcal{S}^*}^{(v)}\end{pmatrix}$ where the blocks are sub-matrices of $\mathbf{L}^{(v)}$ indexed by the corresponding subscripts, and divide $\mathbf{F}^{(v)}$ as $\mathbf{F}^{(v)} = \begin{pmatrix}\mathbf{F}_{\mathcal{L}}^{(v)}\\ \mathbf{F}_{\mathcal{S}^*}^{(v)}\end{pmatrix}$, thereby we arrive at the following system of linear equations:

$$\begin{cases}\mathbf{L}_{\mathcal{L},\mathcal{L}}^{(v)}\mathbf{F}_{\mathcal{L}}^{(v)} + \mathbf{L}_{\mathcal{L},\mathcal{S}^*}^{(v)}\mathbf{F}_{\mathcal{S}^*}^{(v)} + \theta(\mathbf{F}_{\mathcal{L}}^{(v)}-\widetilde{\mathbf{F}}) = \mathbf{O}_{l\times c}\\ \mathbf{L}_{\mathcal{S}^*,\mathcal{L}}^{(v)}\mathbf{F}_{\mathcal{L}}^{(v)} + \mathbf{L}_{\mathcal{S}^*,\mathcal{S}^*}^{(v)}\mathbf{F}_{\mathcal{S}^*}^{(v)} = \mathbf{O}_{s\times c}\end{cases}. \tag{24}$$

After solving Equation (24), we obtain the labels of curriculum examples decided by the vth learner as

$$\mathbf{F}_{\mathcal{S}^*}^{(v)} = -\theta\left[\mathbf{L}_{\mathcal{S}^*,\mathcal{S}^*}^{(v)} - \mathbf{L}_{\mathcal{S}^*,\mathcal{L}}^{(v)}\left(\mathbf{L}_{\mathcal{L},\mathcal{L}}^{(v)}+\theta\mathbf{I}\right)^{-1}\mathbf{L}_{\mathcal{L},\mathcal{S}^*}^{(v)}\right]^{-1}\mathbf{L}_{\mathcal{S}^*,\mathcal{L}}^{(v)}\left(\mathbf{L}_{\mathcal{L},\mathcal{L}}^{(v)}+\theta\mathbf{I}\right)^{-1}\widetilde{\mathbf{F}}, \tag{25}$$

which is further row-normalized[1] so that every element falls into the range $[0,1]$.

To fuse the label matrices $\mathbf{F}_{\mathcal{S}^*}^{(1)},\ldots,\mathbf{F}_{\mathcal{S}^*}^{(V)}$ generated by V learners into a unified $\mathbf{F}_{\mathcal{S}^*}$, we should find the weights of these V modalities on deciding $\mathbf{F}_{\mathcal{S}^*}$. Specifically, we assume that the weights of one modality on different curriculum examples are different, and use $\omega_i^{(v)}$ to indicate the the weight of the vth learner's result $(\mathbf{F}_{\mathcal{S}^*}^{(v)})_{i,:}$ on determining the fused label of $\mathbf{x}_i \in \mathcal{S}^*$. In our case, we relate $\omega_i^{(v)}$ to the tendency of the vth teacher to choose $\mathbf{x}_i$ as a curriculum example, and consider that the examples strongly recommended by the vth teacher can be reliably "learned" by the vth learner. This is because the strong recommendation from the vth teacher indicates that these examples are quite simple for the vth learner, therefore the learning result $\mathbf{F}_{\mathcal{S}^*}^{(v)}$ is trustable and should be emphasized. Fortunately, the ith element of the selection vector $\mathbf{s}^{(v)}$ (i.e., the (i,v)-th element of matrix $\mathbf{S}^*$) exactly reflects the recommendation level of the vth teacher on the example $\mathbf{x}_i$. Therefore, the weight $\omega_i^{(v)}$ can be computed by

$$\omega_i^{(v)} = \frac{\mathbf{S}_{iv}^*}{\sum_{v=1}^{V}\mathbf{S}_{iv}^*}, \tag{26}$$

[1] Given a nonnegative matrix $\mathbf{X}$, the row-normalization is implemented as $\mathbf{X}_{ij} := \mathbf{X}_{ij}/\sum_j \mathbf{X}_{ij}$.

ALGORITHM 2: The pseudo-code of our SMMCL algorithm

```
Input: $l$ labeled examples $\mathcal{L} = \{\mathbf{x}_1, \ldots, \mathbf{x}_l\}$ with known labels $y_1, \ldots, y_l$ expressed in $V$ modalities; $u$ unlabeled examples $\mathcal{U} = \{\mathbf{x}_{l+1}, \ldots, \mathbf{x}_{l+u}\}$ with unknown labels $y_{l+1}, \ldots, y_{l+u}$; Parameters $\alpha$, $\beta$, $\gamma$, $\theta$;
// Pre-processing
Construct graphs $\mathcal{G}^{(v)}$ $(v = 1, \ldots, V)$ via [21]; Compute diagonal matrix $\mathbf{R}^{(v)}$ via Equation (5);
// Multi-modal curriculum generation and propagation
repeat
    Establish the optimal curriculum $\mathcal{S}^*$ by solving Equation (9) (Algorithm 1);
    Compute the label matrix $\mathbf{F}_{\mathcal{S}^*}^{(v)}$ of curriculum examples via Equation (25);
    Compute the weights $\omega_i^{(v)}$ $(v = 1, \ldots, V, i = 1, \ldots, s)$ via Equation (26);
    Fuse $V$ label matrices to $\mathbf{F}_{\mathcal{S}^*}$ via Equation (27);
    Expand $\mathbf{F}$ as $\mathbf{F} := (\mathbf{F}; \mathbf{F}_{\mathcal{S}^*})$;
    $\mathcal{L} := \mathcal{L} \cup \mathcal{S}^*$; $\mathcal{U} := \mathcal{U} - \mathcal{S}^*$;
until $\mathcal{U} = \varnothing$;
Classify every originally unlabeled example to the $j$th class via $j = \arg\max_{j' \in \{1, \ldots, c\}} \mathbf{F}_{ij'}$;
Output: Class labels $y_{l+1}, \ldots, y_{l+u}$;
```

based on which the integrated label vector of $\mathbf{x}_i \in \mathcal{S}^*$ is finally computed as

$$(\mathbf{F}_{\mathcal{S}^*})_{i,:} = \sum_{v=1}^{V} \omega_i^{(v)} \left(\mathbf{F}_{\mathcal{S}^*}^{(v)}\right)_{i,:}. \tag{27}$$

The above multi-modal teaching and learning process iterates until all the unlabeled examples are propagated. Let $\mathbf{F} \in [0,1]^{n \times s}$ be the final label matrix produced by our SMMCL algorithm when $\mathcal{U} = \varnothing$, the example $\mathbf{x}_i \in \mathcal{U}$ is then classified into the jth class that satisfies $j = \arg\max_{j' \in \{1, \ldots, c\}} \mathbf{F}_{ij'}$. The complete SMMCL algorithm is outlined in Algorithm 2.

5 THEORETICAL ANALYSES

In this section, we want to answer two questions from the theoretical aspects:

(1) Is the iterative optimization process introduced in Section 3.3 guaranteed to converge?
(2) How does the difficulty factor considered by CL influence the propagation result?

5.1 Convergence of Algorithm 1

In our algorithm, the optimization problem (9) should be solved via the iterative ADMM method as outlined in Algorithm 1. Therefore, it is necessary to prove that such iterative process will finally converge to a stationary point. Up to now, massive prior works [10, 32, 33] have been done to prove the convergence of ADMM with only two separable blocks of variables. However, our case contains four blocks of variables $\mathbf{S}$, $\mathbf{S}^*$, $\mathbf{E}$, and $\mathbf{J}$, so the existing convergence results are not directly applicable to our problem. Recently, Hong and Luo [18] provide a strict proof for the convergence of ADMM when the objective function contains more than two separable blocks of variables, which is

Theorem 1 [18]. *Given $x = (x_1^\top, \ldots, x_I^\top)^\top$ as a partition of the optimization variable x with I blocks, $\mathcal{D} = \prod_{i=1}^{I} \mathcal{D}_i$ is the feasible set for x, and $\mathbf{A} = (\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_I)$ is an appropriate partition of matrix $\mathbf{A}$, then with each f_i (i takes a value from $1, 2, \ldots, I$) being a nonsmooth convex function, the*

ADMM algorithm for solving the constrained optimization problem formatted as

$$\begin{aligned} \min_{x}\ & f(x) = f_1(x_1) + f_2(x_2) + \cdots + f_I(x_I) \\ s.t.\ & \mathbf{A}x = \mathbf{A}_1 x_1 + \mathbf{A}_2 x_2 + \cdots + \mathbf{A}_I x_I = q \\ & x_i \in \mathcal{D}_i,\ \ i = 1, 2, \ldots, I \end{aligned} \tag{28}$$

will converge linearly to an optimal solution if the following properties are satisfied:

(a) *The intersection $\mathcal{D} \bigcap \mathrm{dom}(f) \bigcap \{x|\mathbf{A}x = q\}$ is nonempty;*
(b) *Every $f_i(x_i)$ in the objective function is convex and continuous over its domain, and the epigraph of $f_i(x_i)$ is a polyhedral set;*
(c) *For any fixed and finite Lagrangian multiplier λ_L and a scalar ι, $\sum_i f_i(x_i)$ is finite for all $x \in \{x : L(x, \lambda_L) \le \iota\} \bigcap \mathcal{D}$ where $L(\cdot)$ is the Lagrangian function;*
(d) *Each sub-matrix $\mathbf{A}_i$ has full column rank;*
(e) *The feasible sets $\mathcal{D}_i$, $i = 1, \ldots, I$ are compact polyhedral sets.*

The above theorem can easily accommodate the general setting with both equality and inequality constraints [18]. One may simply add an extra nonnegative block to turn the inequality constraint into an equality constraint. Therefore, Theorem 1 can be used to demonstrate the convergence of our algorithm, which leads to

Theorem 2. *The iterative process in Algorithm 1 for solving Equation (9) converges linearly to an optimal solution.*

Proof. This theorem can be proved by verifying that our optimization problem (Equation (9)) satisfies the five properties mentioned in Theorem 1. To this end, we decompose Equation (9) as the formulation of Equation (28), and obtain $f_1(\mathbf{S}) = \sum_{v=1}^{V}(\mathbf{s}^{(v)\top}\mathbf{R}^{(v)}\mathbf{s}^{(v)} - \gamma\mathbf{1}^\top\mathbf{s}^{(v)})$, $f_2(\mathbf{S}^*) = 0$, $f_3(\mathbf{E}) = \beta\|\mathbf{E}\|_1$, and $f_4(\mathbf{J}) = \alpha\|\mathbf{J}\|_{2,1}$. Besides, by treating the stacked optimization matrix $(\mathbf{S}^\top, \mathbf{S}^{*\top}, \mathbf{E}^\top, \mathbf{J}^\top, \mathbf{T}_1^\top, \mathbf{T}_2^\top, \mathbf{T}_3^\top)^\top$ as the "x" in Equation (28), the partitioned matrix $\mathbf{A}$ for establishing the constraints in our case is

$$\mathbf{A} = (\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_7) = \begin{pmatrix} \mathbf{I} & -\mathbf{I} & -\mathbf{I} & \mathbf{O} & \mathbf{O} & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{I} & \mathbf{O} & -\mathbf{I} & \mathbf{O} & \mathbf{O} & \mathbf{O} \\ \mathbf{I} & \mathbf{O} & \mathbf{O} & \mathbf{O} & -\mathbf{I} & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{I} & \mathbf{O} & \mathbf{O} & \mathbf{O} & -\mathbf{I} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & \mathbf{I} & \mathbf{O} & \mathbf{O} & \mathbf{O} & -\mathbf{I} \end{pmatrix}, \tag{29}$$

where the size of either $\mathbf{I}$ or $\mathbf{O}$ is $\tau \times \tau$. Therefore, it is straightforward to see that the our optimization problem Equation (9) meets the five properties listed in Theorem 1. Consequently, Theorem 2 is proved and our method is guaranteed to converge linearly. □

5.2 Difficulty-Aware Transductive Error Bound

In this subsection, we derive the error bound of our SMMCL algorithm, and show that it is necessary to consider the difficulty levels of examples to achieve accurate classification.

Our algorithm assigns the most likely label to an unlabeled example based on the weighted voting result of V learners (see Lines 9 and 13 of Algorithm 2), so it is a *Bayes classifier* B_Q that relies on a posterior distribution Q over a hypothesis space $\mathcal{H}$ such that the weighted majority-voting classifier will have the smallest potential risk on the unlabeled examples [2]. In contrast, if the base classifiers of an ensemble are randomly selected or combined to yield the result, they altogether form a *Gibbs classifier* G_Q. The transductive risks of B_Q and G_Q on unlabeled set $\mathcal{U}$ are respectively defined by [2]:

$$\mathcal{R}(B_Q) = \frac{1}{u}\sum_{\mathbf{x}\in\mathcal{U}} [\![B_Q(\mathbf{x}) \neq y]\!] \tag{30}$$

and

$$\mathcal{R}(G_Q) = \frac{1}{u}\sum_{\mathbf{x}\in\mathcal{U}} \mathbb{E}_{f\sim Q}[\![f(\mathbf{x}) \neq y]\!], \tag{31}$$

where y is groundtruth label of $\mathbf{x}$, and $[\![\cdot]\!] = 1$ if the argument holds and 0 otherwise. To take the difficulty information of examples into consideration, we extend Equation (30) and define a joint Bayes risk with difficulty threshold κ, that is

$$\mathcal{R}_\kappa(B_Q) = \frac{1}{u}\sum_{\mathbf{x}\in\mathcal{U}} [\![B_Q(\mathbf{x}) \neq y \wedge R(\mathbf{x}) < \kappa]\!], \tag{32}$$

where $R(\mathbf{x})$ is the difficulty level of $\mathbf{x}$ as defined in Equation (5). Equation (32) means that we only care about the classification correctness of unlabeled examples whose difficulty levels are below a threshold κ. Besides, we define P_u as the uniform probability distribution over $\mathcal{U}$, and for any of its subset $\mathcal{U}'$ we have $P(\mathcal{U}') = \frac{1}{u}|\mathcal{U}'|$ with "$|\cdot|$" denoting the set size.

Given above definitions, the difficulty-aware transductive Bayes risk for our SMMCL is provided in the theorem below:

Theorem 3. *Suppose $B_Q(\mathbf{x}) = sgn[\mathbb{E}_{f\sim Q}f(\mathbf{x})]$ is our Bayes classifier, $G_Q(\mathbf{x})$ is the associated Gibbs classifier, κ is the difficulty threshold, and $H(\mathbf{x})$ is the reliability of $\mathbf{x}$ defined by Equation (1), then for all $\delta \in (0,1]$ with probability at least $1-\delta$, we have*

$$\mathcal{R}_\kappa(B_Q) \leq \inf_{0\leq\xi\leq\kappa}\left\{P_u(R_Q(\mathbf{x}) < \kappa) + \frac{1}{\xi}\left\lfloor \psi_u^\delta(Q) - \Omega(\xi)\right\rfloor_+\right\}, \tag{33}$$

where $\psi_u^\delta(Q) = \mathcal{R}^\delta(G_Q) - \frac{1}{2}[1-\mathbb{E}_u(R(\mathbf{x})) - \zeta\mathbb{E}_u(H(\mathbf{x}))]$, $\Omega(\xi) = \mathbb{E}_u(R(\mathbf{x})[\![R(\mathbf{x}) < \xi]\!])$, and "$\lfloor\cdot\rfloor_+$" preserves the positive value and sets the negative value to 0.

Proof. To prove Theorem 3, we need the following useful lemma of which the proof has been put into the supplementary material: □

Lemma 4. *Let $\{\xi_1,\ldots,\xi_N\}$ be N possible difficulty thresholds on $\mathcal{U}$, namely $\{\xi_i\}_{i=1}^N = \{R(\mathbf{x})|\mathbf{x}\in\mathcal{U}\wedge R(\mathbf{x}) > 0\}$ with $\xi_i < \xi_{i+1}$, $\forall i = 1,2,\ldots,N-1$, and $\{h_1,\ldots,h_N\}$ are the corresponding values of $H(\mathbf{x})$. Moreover, we denote $b_i = P_u(B_Q(\mathbf{x}) \neq y \wedge R(\mathbf{x}) = \xi_i)$ and $\pi_i = P_u(B_Q(\mathbf{x}) = y \wedge H(\mathbf{x}) = h_i)$ for $i = 1,2,\ldots,N$. The risks of Gibbs classifier $\mathcal{R}(G_Q)$ and Bayes classifier $\mathcal{R}_\kappa(B_Q)$ satisfy*

$$\mathcal{R}(G_Q) \geq \sum_{i=1}^N b_i\xi_i + \frac{1}{2}\left[1-\mathbb{E}_u(R(\mathbf{x})) - \zeta\mathbb{E}_u(H(\mathbf{x}))\right] \tag{34}$$

and

$$\mathcal{R}_\kappa(B_Q) = \sum_{i=1}^{k-1} b_i, \text{ where } k = \{i|\xi_i > \kappa\}. \tag{35}$$

According to Lemma 4, we know that for a fixed κ, $\{b_i\}_{i=1}^N$ satisfy $0 \leq b_i \leq P_u(R(\mathbf{x}) = \xi_i)$. Besides, by assuming that we can obtain an upper bound $\mathcal{R}^\delta(G_Q)$ of $\mathcal{R}(G_Q)$ with a probability $1-\delta$

over the random choices of $\mathcal{L}$ and $\mathcal{U}$, and recalling the definition of $\psi_u^\delta(Q)$ in Theorem 3, we have

$$\begin{aligned}\sum_{i=1}^{N} b_i \xi_i &\le \mathcal{R}(G_Q) - \frac{1}{2}\left[1 - \mathbb{E}_u(R(\mathbf{x})) - \zeta \mathbb{E}_u(H(\mathbf{x}))\right] \\ &\le \mathcal{R}^\delta(G_Q) - \frac{1}{2}\left[1 - \mathbb{E}_u(R(\mathbf{x})) - \zeta \mathbb{E}_u(H(\mathbf{x}))\right] \\ &= \psi_u^\delta(Q).\end{aligned} \tag{36}$$

Therefore, we arrive at the upper bound of $\mathcal{R}_\kappa(B_Q)$ with a probability $1-\delta$ based on an optimization regarding $\{b_1, \ldots, b_N\}$, namely:

$$\mathcal{R}_\kappa(B_Q) \le \max_{b_1,\ldots,b_N} \sum_{i=1}^{k-1} b_i \quad s.t.\ 0 \le b_i \le P_u(R(\mathbf{x}) = \xi_i),\ \sum_{i=1}^{N} b_i \xi_i \le \psi_u^\delta(Q),\ \forall i. \tag{37}$$

It can be observed that the optimization in Equation (37) is a linear programming problem, and its analytic solution can be attained based on the lemma below:

Lemma 5. *Suppose $g_1, g_2, \ldots, g_N$ are N constants satisfying $0 < g_i < g_{i+1}$ $(g = 1, \ldots, N-1)$, $p_1, p_2, \ldots, p_N$ and Θ are nonnegative numbers, then the optimization problem on $x_1, x_2, \ldots, x_N$*

$$\max_{x_1,\ldots,x_N} \sum_{i=1}^{k-1} x_i \quad s.t.\ 0 \le x_i \le p_i,\ \sum_{i=1}^{N} g_i x_i \le \Theta,\ \forall i \tag{38}$$

has a closed-form solution:

$$x_i^* = \begin{cases} 0, & i \ge k \\ \min\left(p_i, \frac{1}{g_i}\left\lfloor \Theta - \sum_{j<i} x_j^* g_i \right\rfloor_+\right), & 1 \le i < k \end{cases}, \tag{39}$$

where the optimal $x_1^*, x_2^*, \ldots, x_{k-1}^*$ are decided successively.

The fact illustrated by Lemma 5 is quite natural and its explanation can be found in the supplementary material. Based on Lemma 5, we know that the analytic solution of problem (37) is

$$b_i = \begin{cases} 0, & i \ge k \\ \min\left(P_u\left(R(\mathbf{x}) = \xi_i\right), \frac{1}{\xi_i}\left\lfloor \psi_u^\delta(Q) - \sum_{j<i} \xi_j P_u\left(R(\mathbf{x}) = \xi_j\right) \right\rfloor_+\right), & 1 \le i < k. \end{cases} \tag{40}$$

Note that the term $\sum_{j<i} \xi_j P_u(R(\mathbf{x}) = \xi_j)$ in Equation (40) can be rewritten as $\mathbb{E}_u(R(\mathbf{x})[\![R(\mathbf{x}) < \xi_i]\!])$, which is exactly the $\Omega(\xi_i)$ defined in Theorem 3. By defining the index $\hat{I} = \max\{i | \psi_u^\delta(Q) - \Omega(\xi_i) > 0\}$, we know $b_{\hat{I}} = \frac{1}{\xi_{\hat{I}}}(\psi_u^\delta(Q) - \Omega(\xi_{\hat{I}}))$ from Equation (40). As a result, an inequality can be reached as

$$b_1 + \cdots + b_{\hat{I}-1} \le P_u\left(R(\mathbf{x}) < \xi_i\right) < P_u\left(R(\mathbf{x}) < \kappa\right), \tag{41}$$

where the second inequality holds because the ξ_i corresponding to the nonzero $b_1, \ldots, b_{\hat{I}-1}$ are less than κ according to Equation (35). Therefore, through the objective of Equation (37) and Equation (40), we immediately get

$$\begin{aligned}\mathcal{R}_\kappa(B_Q) &\le \max_{b_1,\ldots,b_N} \sum_{i=1}^{k-1} b_i = (b_1 + \cdots + b_{\hat{I}-1}) + b_{\hat{I}} \le P_u(R(\mathbf{x}) < \kappa) + \frac{\psi_u^\delta(Q) - \Omega(\xi_{\hat{I}})}{\xi_{\hat{I}}} \\ &\le \inf_{\xi \in (0,\kappa]} \left\{ P_u\left(R(\mathbf{x}) < \kappa\right) + \frac{1}{\xi}\left\lfloor \psi_u^\delta(Q) - \Omega(\xi) \right\rfloor_+ \right\},\end{aligned} \tag{42}$$

which is exactly the Equation (33) in Theorem 3.

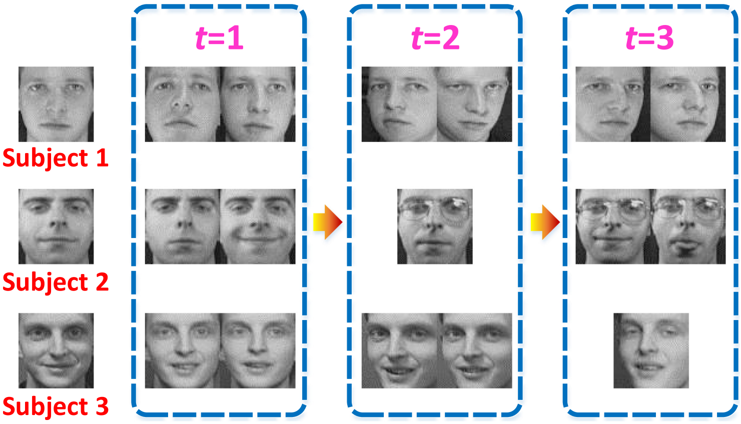


Fig. 3. Visualization of propagation sequence generated by our SMMCL.

From Theorem 3, we have two main observations: first, the error risk of our SMMCL is upper bounded, so SMMCL is able to obtain satisfactory classification results with a high probability; and second, this risk bound is closely related to the difficulty threshold κ, therefore considering the difficulty information is essential for our algorithm to achieve encouraging performances. Furthermore, we see that κ that governs the amount of simplest examples should be a suitable value such that the minimal risk bound can be reached. This is naturally achieved by SMMCL which automatically decides the number of curriculum examples according to the non-zero rows in the optimized $\mathbf{S}^*$.

6 EXPERIMENTAL RESULTS

This section verifies the effectiveness of our motivation (Section 6.1), compares the proposed SMMCL with state-of-the-art methods on various practical tasks (Sections 6.2~6.5), and studies the parametric sensitivity of involved tuning parameters (Section 6.6).

6.1 Algorithm Validation

There are mainly two innovations in our SMMCL algorithm, one is the learning sequence from simple examples to more difficult ones, and the other is the multi-modal strategy for boosting the performance. Apart from demonstrating the usefulness of these two innovations, we also want to show that the improvements in this article can bring about better performance than our previous conference work [11].

Specifically, we extract totally 30 face images of three subjects from the *ORL*[2] face recognition dataset, and the resolution of every image is 64 × 64. For each subject, we only select one labeled example which depicts the frontal face with normal expression (see the left panel in Figure 3), and the remaining 27 face images with various angles, expressions, illuminations and appearances are treated as unlabeled. Our task is to classify these unlabeled faces based on only three labeled examples. Generally, the images that are similar to the labeled examples are easier to classify than the faces with abnormal expressions or appearances, so we aim to test whether the proposed algorithm is able to correctly identify the difficulty levels of these examples and yield a proper learning sequence.

[2] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html.

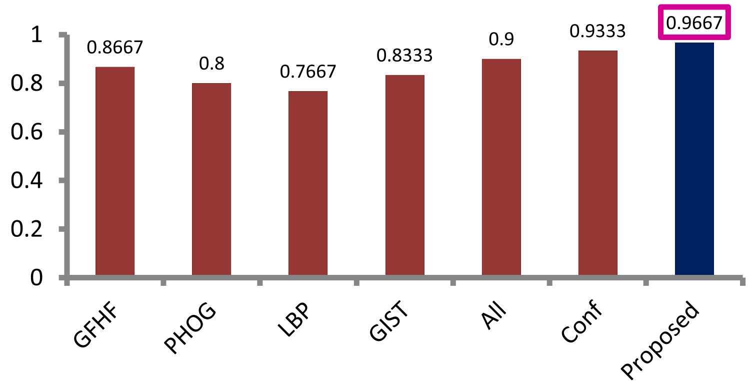


Fig. 4. The comparison of different settings on *ORL* dataset.

All face images are represented by the 72-dimensional Pyramid Histogram Of Gradients (PHOG) [6], 512-dimensional GIST [38], and 256-dimensional Local Binary Patterns (LBP) [1] features. Therefore, three different feature modalities are formed to serve as the input of all the compared settings. First, we run the Single-Modal CL (SMCL) model Equation (6) on this *ORL* dataset with one of the LBP, PHOG, and GIST feature descriptors (denoted as "LBP", "PHOG", and "GIST", respectively). Then we concatenate these three different feature vectors to a long vector, and apply SMCL again to simultaneously utilize the three feature modalities (denoted "All"). Next we implement two versions of the SMMCL algorithm proposed in this article (denoted "Proposed") and in our earlier conference paper [11] (denoted "Conf"). Finally, if CL strategy is not utilized for propagation, our propagation model Equation (22) will degenerate into the existing GFHF [59] as in this case $\widetilde{\mathbf{F}}_{ij}$ in Equation (22) will be either 0 or 1. Therefore, we also report the performance of GFHF with the concatenated long feature vector, in which the propagation sequence has not been optimized by a teacher.

The recognition accuracies of all the compared settings are illustrated in Figure 4, which clearly indicate that the proposed SMMCL in this article achieves the best result. Apart from this observation, we have several other interesting findings:

(1) By comparing "GFHF" with "All", we see that the plain propagation method "GFHF" performs inferior to "All" that employs the single-teacher curriculum learning, which verifies that the examples' difficulty information is important and the propagation quality can be enhanced under the supervision of a teacher.
(2) By comparing the performances of "PHOG", "LBP", "GIST", "All" with "Conf" and "Proposed", we note that utilizing multiple modalities for selecting the simplest examples can bring about more suitable curriculums than those output by single modality. As a result, the recognition accuracies of "Conf" and "Proposed" are higher than "PHOG", "LBP", "GIST", and "All".
(3) The result of "All" indicates that directly concatenating different feature modalities into a long feature vector is not a good idea to handle multi-modal cases, as its performance is much worse than "Conf" or "Proposed" that fuses the curriculums of different modalities via a specific way.
(4) By comparing "Conf" and "Proposed", we see that the algorithmic improvements mentioned in Section 1 lead to better results of the proposed method in this article than our previous conference work [11].

Fig. 5. Example images of *Architecture* dataset.

Although there are only three labeled examples at the beginning, the proposed algorithm is still able to achieve very impressive recognition accuracy that is 96.67%, and this is attributed to the proper learning sequence arranged by the teachers. To show the strength of this teaching committee, in Figure 3 we present the curriculum examples chosen by the teachers in each learning round. We see that most of the frontal faces of three subjects are recognized in the first learning round. In the second learning round, some examples contain profile faces (Subjects 1 and 3) or glasses (Subject 2), which are apparently more difficult than the curriculum images in the first learning round. In the last round, the Subject 2 puts out his tongue, and the expression and illumination of Subject 3 are quite different from those in the labeled image, so their classifications have been postponed by our SMMCL algorithm. This result reflects that the teachers in SMMCL accurately assess the difficulty of examples and properly design a simple-to-difficult curriculum sequence for the stepwise learners.

6.2 Architecture Style Recognition

In this section, we apply our SMMCL algorithm to recognize the styles of architectures as they appeared in images. To this end, a recently released *Architecture* dataset [56] is adopted here which consists of 1000 building images belonging to 25 styles such as “Baroque”, “Greek Revival”, “Post-modern”, and so on (see Figure 5). Besides, five representative graph-based label propagation algorithms are taken as baselines for comparison including: (1) Gaussian Field and Harmonic Functions (GFHF) [59], which is a classical algorithm and has been introduced in Section 6.1; (2) Dynamic Label Propagation (DLP) [50], which is a recently proposed single-modal propagation methodology; (3) Sparse Multiple Graph Integration (SMGI) [22] that is a competitive multi-modal graph transduction method; (4) Adaptive Multi-Modal Semi-Supervised classifier (AMMSS) [7] which is based on multiple graphs and also automatically learns the weight of each modality like our SMMCL; and (5) Multi-Modal Curriculum Learning (MMCL) [16] which is the state-of-the-art CL algorithm and is very relevant to the proposed method. Moreover, our previous conference work (Conf) [11] is also compared throughout this article. Similar to Section 6.1, here we also employ GIST, LBP, and PHOG features as different modalities. For the single-modal methods like GFHF and DLP, the GIST, LBP, and PHOG feature vectors are directly concatenated into a long feature vector as their inputs.

To evaluate the performances of above-compared methods, a number of the examples in *Architecture* dataset are selected to form the labeled set, and the rest are taken as the unlabeled examples. We study the classification accuracies of above-mentioned methodologies under different sizes of labeled set, and the experiment under each size is conducted ten times with different initial labeled examples. At least one example in each class is incorporated to the labeled set. The reported accuracies are then obtained by averaging the outputs of these ten independent runs.

To achieve fair comparison, an identical graph with $K = 30$ is established for all the compared methods except DLP. In DLP, the 30-NN graph is built by leveraging the Gaussian kernel as required. Besides, the key parameters in SMGI are optimally tuned to $\lambda_1 = 0.01$ and $\lambda_2 = 0.1$ via searching the grid $\{0.01, 0.1, 1, 10\}$, and r and λ in AMMSS are adjusted to 0.5 and 10, respectively. As recommended by the authors, the parameters α and λ in DLP are set to 0.05 and 0.1. In MMCL,

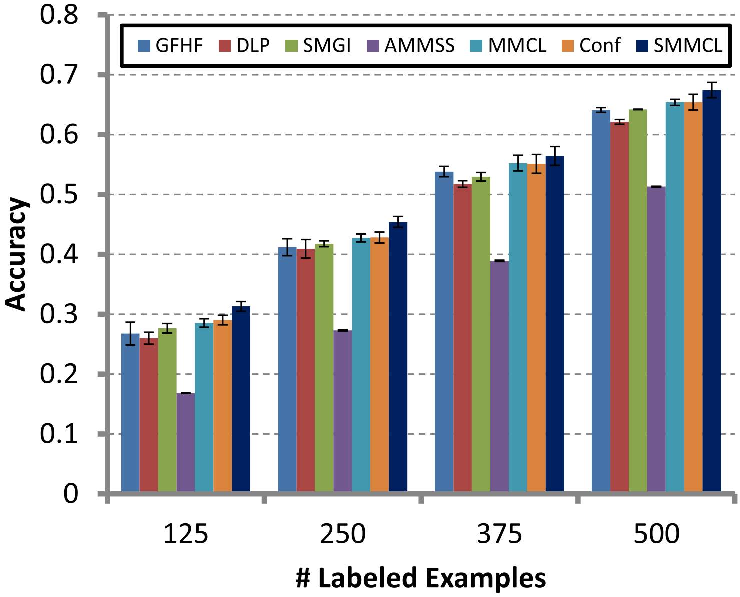


Fig. 6. The comparison of various algorithms on *Architecture* dataset.

we set $\beta = 10$, $\gamma = 3$, and $\eta = 1.1$, as indicated in [16]. For our proposed SMMCL, the tradeoff parameters in Equation (9) are optimally tuned to $\alpha = 1$, $\beta = 0.1$ and $\gamma = 1$ via cross validation, and the weighting parameter θ in Equation (22) is set to 0.01 by following [58]. In Section 6.6, we will show that the performance of SMMCL is generally not sensitive to the choices of α, β, γ, and θ.

Figure 6 presents the results, which clearly indicate that the proposed SMMCL consistently outperforms other baseline algorithms when the number of labeled examples l changes from 125 to 500. Besides, we observe that SMMCL achieves higher accuracy than MMCL, so our “soft” strategy for fusing the opinions of different teachers is better than the “hard” one adopted by [16]. Furthermore, it can be noted that MMCL, Conf, and SMMCL perform favorably to the “plain” learner GFHF that is not equipped with a teacher, which again demonstrates that the curriculum sequence from simple to difficult arranged by the teachers does contribute to an improved propagation quality.

Apart from the accuracy, we also examine the convergence of the iterative process in Algorithm 1 for optimizing Equation (9). Specifically, we plot the values of $\|\mathrm{S}^{*(iter)} - \mathrm{S}^{*(iter-1)}\|_{\infty}$ under different iterations when the numbers of initial labeled examples are $l = 250$ and $l = 500$, respectively. From the results in Figure 7, we see that the difference of S^* between successive iterations gradually vanishes when the iteration proceeds. Therefore, the convergence of the ADMM for solving Equation (9) is empirically verified, which also confirms our theoretical results in Section 5.1.

6.3 Scene Categorization

This section tests the ability of our SMMCL algorithm on the problem of scene categorization. A typical *Scene UNderstanding* (*SUN*) database [54] is adopted here for algorithm evaluation, which contains 108754 color image examples belonging to 397 natural scene categories such as “garage”, “kitchen”, “stadium”, and so on. For our experiment, we extract the first 100 images in each scene category, and thus the constituted subset contains 39,700 examples across 397 classes. From the example images shown in Figure 8, we see that this dataset is quite challenging for scene classification. The compared methods in this experiment include GFHF, DLP, AMMSS, DLP, Conf, and MMCL as in Section 6.2.

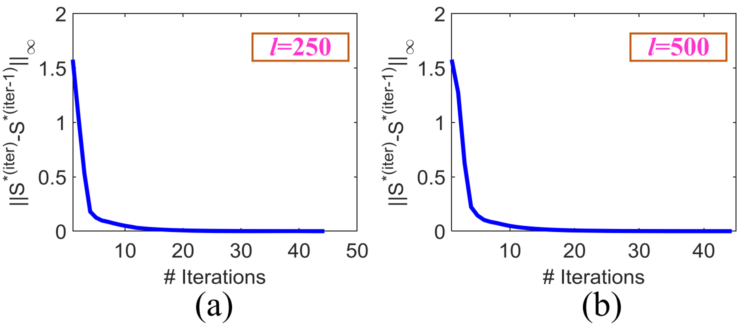


Fig. 7. The convergence curve of ADMM process in Algorithm 1. (a) shows the result when $l = 250$, and (b) presents the result under $l = 500$.

Fig. 8. Typical examples of *SUN* dataset.

Table 1. Classification Accuracy on *SUN* Dataset

	l = 11,910	l = 23,820
GFHF [59]	0.530 ± 0.001 √	0.751 ± 0.002 √
DLP [50]	0.573 ± 0.003 √	0.756 ± 0.001 √
SMGI [22]	0.478 ± 0.002 √	0.718 ± 0.002 √
AMMSS [7]	0.401 ± 0.003 √	0.701 ± 0.001 √
MMCL [16]	0.579 ± 0.002 √	0.774 ± 0.001 √
Conf [11]	0.580 ± 0.002 √	0.778 ± 0.002 √
SMMCL	0.614 ± 0.001	0.792 ± 0.001

The best result under each l is marked in red. "√" means that SMMCL is significantly better than the corresponding method.

To obtain satisfactory performance on this dataset, we use the features extracted by VGGNet-16 [44] and AlexNet [25] to serve as the two modalities for all the compared algorithms. Both deep neural networks are trained on ImageNet, and we employ the output of their first fully-connected layer to form a 4096-dimensional feature vector. All methods are tested on 10-NN graphs, and their parameters are set to the same values as those in Section 6.2. Algorithms here are implemented ten times independently under different l with different selected labeled examples, and the reported accuracies and standard deviations are calculated as the mean value of the outputs of these runs.

The classification accuracies of various methods are evaluated when 30 and 60 labeled examples are selected from every class, and thus l for this experiment ranges from 11,910 to 23,820. The experimental results are presented in Table 1. We see that the classification performances of all the methods can be improved with the expansion of labeled set. Among the compared methods, AMMSS is significantly worse than the remaining algorithms. In contrast, the CL-based methods SMMCL, Conf, and MMCL achieve top-level performances which consistently outperform GFHF, DLP, and SMGI. Furthermore, it can be observed that SMMCL generates higher accuracies than MMCL. Therefore, the "soft" fusion of curriculums adopted by SMMCL that explicitly considers

Table 2. Classification Accuracy on *CIFAR100* Dataset

	l = 20,000	l = 40,000
GFHF [59]	0.599 ± 0.001 √	0.801 ± 0.001 √
DLP [50]	0.471 ± 0.005 √	0.750 ± 0.005 √
SMGI [22]	0.609 ± 0.001 √	0.826 ± 0.001 √
AMMSS [7]	0.341 ± 0.002 √	0.671 ± 0.001 √
MMCL [16]	0.605 ± 0.005 √	0.820 ± 0.001 √
Conf [11]	0.625 ± 0.003	0.825 ± 0.002 √
SMMCL	0.626 ± 0.001	0.847 ± 0.005

The best result under each l is marked in red. "√" means that SMMCL is significantly better than the corresponding method.

the commonality and individuality of the involved teachers is better than the "hard" one employed by MMCL. The superiority of the proposed SMMCL over other comparators is also statistically verified by the paired t-test with the significance level 0.1.

6.4 Natural Image Classification

In this section, we evaluate our SMMCL on classifying general images from everyday life. Specifically, *CIFAR100* dataset [24] is employed here, which contains 60,000 32 × 32 color images across 100 classes with 600 images per class.

We investigate the classification accuracies of compared methods when we have l = 20,000 and l = 40,000 labeled examples, and the results are presented in Table 2. We see that GFHF obtains the highest records among the single-modal approaches including GFHF and DLP. However, it is slightly inferior to the multi-modal methodologies such as SMGI, MMCL, Conf, and SMMCL. This is due to the fact that the multi-modal algorithms can better leverage the complementary information carried out by every modality than the single-modal methods. For multi-modal methods, our SMMCL achieves the comparable performance with Conf when l = 20,000 and the highest accuracy when l = 40,000. Furthermore, we see that SMMCL is able to obtain the accuracy as high as 84.7% when l = 40,000. Considering that the employed *CIFAR100* dataset is quite challenging for image classification, this is a very encouraging result which again demonstrates the strength of our SMMCL approach.

6.5 Hyperspectral Image Classification

Considering that label propagation has been widely adopted for tackling the remotely sensed hyperspectral image classification, this section tests the performance of all the compared methods on a typical hyperspectral dataset called *JapserRidge*. This dataset is formed by a remotely sensed image with 100 × 100 pixels, and each pixel (i.e., example) is recorded by totally 198 spectral channels ranging from 380nm to 2,500nm. Our target is to classify the image pixels into one of the four categories including "tree", "soil", "water", and "road". In this application, only a small fraction of pixels can be manually annotated because the entire image contains too many pixels and the land types in the remotely sensed image are also quite complicated (see the groundtruth image in Figure 9). For example, the *JapserRidge* dataset only provides 1,797 labeled pixels, and the remaining 8,203 pixels are unlabeled. Therefore, our SMMCL is quite suitable for accomplishing this task.

To run SMMCL, we use the values in the channels 1~100 and 101~198 as two modalities to characterize each pixel. The number of nearest neighbors for various methods is set to $K = 30$, and the tradeoff parameters of SMMCL are set to $\alpha = 0.01$ and $\beta = \gamma = 1$. The classification results of SMMCL and other comparators are displayed in Figure 9. We observe that DLP and AMMSS

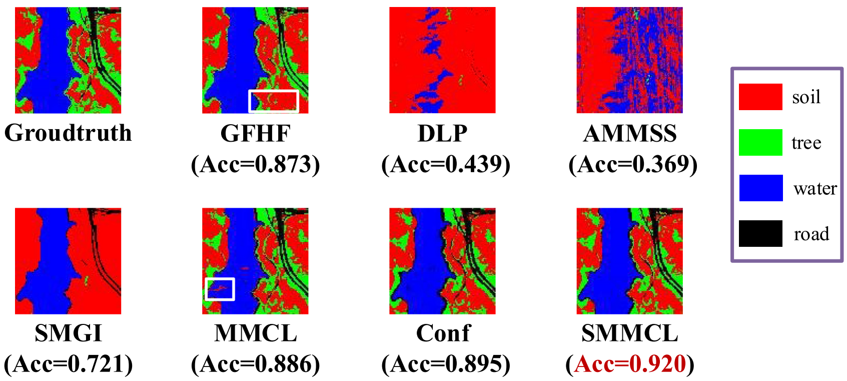


Fig. 9. The classification results of compared methods on *JasperRidge* dataset. The groundtruth is displayed in the left-top corner. The accuracies of various compared methods are indicated in the brackets below the images, and the highest record obtained by our SMMCL is marked in red. The imperfect image regions output by GFHF and MMCL are highlighted by the white boxes.

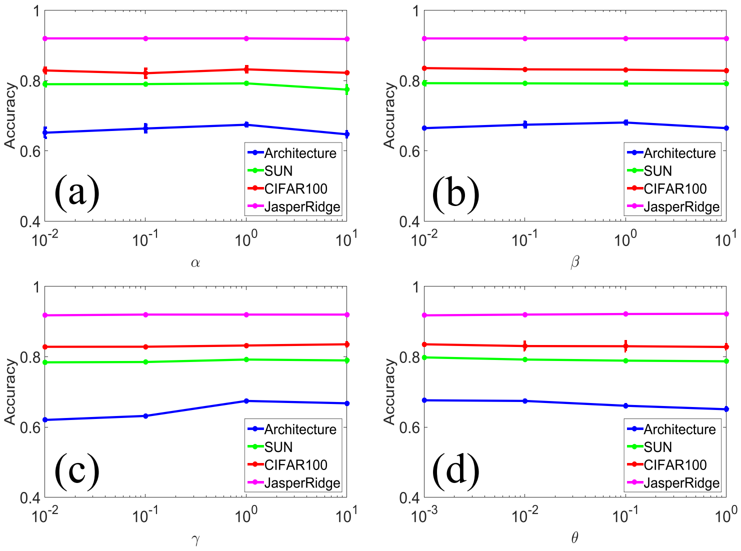


Fig. 10. Empirical studies on the parametric sensitivity of SMMCL. (a), (b), (c), (d) plot the accuracy of SMMCL w.r.t. the variation of α, β, γ and θ, respectively.

are significantly confused by the complicated landforms and most of the pixels are erroneously classified. SMGI obtains relatively better performance than DLP and AMMSS, but the tree regions are seldom identified, which leads to the accuracy as low as 72.1%. By comparing the outputs of GFHF and SMMCL, we see that the tree regions in the bottom right corner of the image are mistakenly classified as "soil" by GFHF (see the white box in the output of GFHF). In contrast, in this region our SMMCL makes a clearer distinguish between "tree" and "soil" than GFHF, which shows the strength of SMMCL on handling complex and difficult examples. By comparing the outputs of MMCL and SMMCL, it can be observed that a small part of water regions near the bank are incorrectly recognized as "soil" by MMCL (see the white box in the output of MMCL), so its accuracy is lower than SMMCL with a gap of 3.4%. Furthermore, our previous conference work

[11] generates very similar classification result with the groundtruth. However, its result can still be improved by SMMCL which achieves 92.0% classification accuracy.

6.6 Parametric Sensitivity

Our method contains four parameters that should be manually tuned, including α, β, and γ in Equation (7), and θ in Equation (22). Therefore, in this section, we discuss whether their selection will significantly influence the performance of SMMCL method.

To this end, we examine the classification accuracies of SMMCL by varying one of α, β, γ, and θ, and meanwhile fixing the remaining parameters to constant values. The four practical datasets from Sections 6.2~6.5 are adopted here including *Architecture*, *SUN*, *CIFAR100*, and *JasperRidge*. We change α, β, and γ from $10^{-2} \sim 10^{1}$, and increase θ from $10^{-3} \sim 10^{0}$. The results on the four datasets are shown in Figure 10. From the experimental results, we learn that the performance of SMMCL is generally not sensitive to the change of α, β, γ, and θ; therefore, the involved four parameters are not difficult to tune before implementing our SMMCL algorithm.

7 CONCLUSION

This article proposes a novel CL methodology for multi-modal networked data which is termed “Soft Multi-Modal Curriculum Learning” (SMMCL). By explicitly considering the commonality and individuality of the incorporated teachers, their different opinions are flexibly fused into an unbiased simplest curriculum in every learning round, which leads to the trustable and accurate LP results as revealed by the experiments. We have theoretically proved the convergence of the ADMM optimization process, and also verified that the difficulty factor plays an important role in regulating the propagation error bound.

In the future, we plan to develop a strategy to monitor the propagation process, and simply invoke the curriculum generation step when the propagation quality significantly decreases. As such, the computational burden of SMMCL can be further reduced in the presence of relatively large datasets. Besides, it is worthwhile to apply SMMCL to other unsupervised graph-based scenarios such as spectral clustering [39] and dimensionality reduction on manifolds [19].

REFERENCES

[1] T. Ahonen, A. Hadid, and M. Pietikäinen. 2004. Face recognition with local binary patterns. In *European Conference on Computer Vision*. Springer, 469–481.

[2] M. Amini, F. Laviolette, and N. Usunier. 2008. A transductive bound for the voted classifier with an application to semi-supervised learning. In *Advances in Neural Information Processing Systems*. 65–72.

[3] Y. Bai, Y. Lou, F. Gao, S. Wang, Y. Wu, and L. Duan. 2018. Group-sensitive triplet embedding for vehicle reidentification. *IEEE Transactions on Multimedia* 20, 9 (2018), 2385–2399.

[4] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. 2009. Curriculum learning. In *International Conference on Machine Learning*. 41–48.

[5] C. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer.

[6] A. Bosch, A. Zisserman, and X. Munoz. 2007. Representing shape with a spatial pyramid kernel. In *6th ACM International Conference on Image and Video Retrieval*. 401–408.

[7] X. Cai, F. Nie, W. Cai, and H. Huang. 2013. Heterogeneous image features integration via multi-modal semi-supervised learning model. In *IEEE International Conference on Computer Vision*. 1737–1744.

[8] A. Gammerman, V. Vovk, and V. Vapnik. 1998. Learning by transduction. In *14th Conference on Uncertainty in Artificial Intelligence*. 148–155.

[9] Y. Gao, H. Zhang, X. Zhao, and S. Yan. 2017. Event classification in microblogs via social tracking. *ACM Transactions on Intelligent Systems and Technology* 8, 3 (2017), 35.

[10] T. Goldstein, B. O’Donoghue, S. Setzer, and R. Baraniuk. 2014. Fast alternating direction optimization methods. *SIAM Journal on Imaging Sciences* 7, 3 (2014).

[11] C. Gong. 2017. Exploring commonality and individuality for multi-modal curriculum learning. In *AAAI Conference on Artificial Intelligence*. 1926–1933.

[12] C. Gong, T. Liu, D. Tao, K. Fu, E. Tu, and J. Yang. 2015. Deformed graph laplacian for semisupervised learning. *IEEE Transactions on Neural Networks and Learning Systems* 26, 10 (Oct. 2015), 2261–2274.

[13] C. Gong, D. Tao, K. Fu, and J. Yang. 2015. Fick's law assisted propagation for semisupervised learning. *IEEE Transactions on Neural Networks and Learning Systems* 26, 9 (2015), 2148–2162.

[14] C. Gong, D. Tao, W. Liu, L. Liu, and J. Yang. 2017. Label propagation via teaching-to-learn and learning-to-teach. *IEEE Transactions on Neural Networks and Learning Systems* 28, 6 (2017), 1452–1465.

[15] C. Gong, D. Tao, W. Liu, S. Maybank, M. Fang, K. Fu, and J. Yang. 2015. Saliency propagation from simple to difficult. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2531–2539.

[16] C. Gong, D. Tao, S. Maybank, W. Liu, G. Kang, and J. Yang. 2016. Multi-modal curriculum learning for semi-supervised image classification. *IEEE Transactions on Image Processing* 25, 7 (2016), 3249–3260.

[17] B. Han, I. Tsang, L. Chen, C. Yu, and S. Fung. 2018. Progressive stochastic learning for noisy labels. *IEEE Transactions on Neural Networks and Learning Systems* (2018), 1–13.

[18] M. Hong and Z. Luo. 2017. On the linear convergence of the alternating direction method of multipliers. *Mathematical Programming* 162, 1–2 (2017), 165–199.

[19] Z. Huang, H. Zhu, J. T. Zhou, and X. Peng. 2018. Multiple marginal Fisher analysis. *IEEE Transactions on Industrial Electronics* (2018).

[20] L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. Hauptmann. 2015. Self-paced curriculum learning. In *AAAI Conference on Artificial Intelligence*. 2078–2086.

[21] M. Karasuyama and H. Mamitsuka. 2013. Manifold-based similarity adaptation for label propagation. In *Advances in Neural Information Processing Systems*. 1547–1555.

[22] M. Karasuyama and H. Mamitsuka. 2013. Multiple graph label propagation by sparse integration. *IEEE Transactions on Neural Networks and Learning Systems* 24, 12 (2013), 1999–2012.

[23] F. Khan, B. Mutlu, and X. Zhu. 2011. How do humans teach: On curriculum learning and teaching dimension. In *Advances in Neural Information Processing Systems*. 1449–1457.

[24] A. Krizhevsky and G. Hinton. 2009. *Learning Multiple Layers of Features from Tiny Images*. Technical Report. U. Toronto.

[25] A. Krizhevsky, I. Sutskever, and G. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. 1097–1105.

[26] M. Kumar, B. Packer, and D. Koller. 2010. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*. 1189–1197.

[27] Xiangyuan Lan, Andy Jinhua Ma, and Pong Chi Yuen. 2014. Multi-cue visual tracking using robust feature-level fusion based on joint sparse representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1194–1201.

[28] Xiangyuan Lan, Andy Jinhua Ma, Pong C. Yuen, and Rama Chellappa. 2015. Joint sparse representation and robust feature-level fusion for multi-cue visual tracking. *IEEE Transactions on Image Processing* 24, 12 (2015), 5826–5841.

[29] Xiangyuan Lan, Mang Ye Shengping Zhang, Huiyu Zhou, and Pong C. Yuen. 2018. Modality-correlation-aware sparse representation for RGB-infrared object tracking. *Pattern Recognition Letters* (2018).

[30] Xiangyuan Lan, Shengping Zhang, Pong C. Yuen, and Rama Chellappa. 2018. Learning common and feature-specific patterns: A novel multiple-sparse-representation-based tracker. *IEEE Transactions on Image Processing* 27, 4 (2018), 2022–2037.

[31] Y. Lee and K. Grauman. 2011. Learning the easy things first: Self-paced visual category discovery. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1721–1728.

[32] X. Li, G. Cui, and Y. Dong. 2018. Discriminative and orthogonal subspace constraints-based nonnegative matrix factorization. *ACM Transactions on Intelligent Systems and Technology* 9, 6 (2018), 65.

[33] Z. Lin, M. Chen, and Y. Ma. 2010. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv:1009.5055v3* 9 (2010).

[34] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. 2012. Robust recovery of subspace structures by low-rank representation.*IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 1 (2012), 171–184.

[35] Y. Lou, Y. Bai, J. Liu, S. Wang, and L. Duan. 2019. Embedding adversarial learning for vehicle re-identification. *IEEE Transactions on Image Processing* (2019).

[36] U. Luxburg. 2007. A tutorial on spectral clustering. *Statistics and Computing* 17, 4 (2007), 395–416.

[37] Takeru Miyato, Shin Ichi Maeda, Masanori Koyama, and Shin Ishii. 2018. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018).

[38] A. Oliva and A. Torralba. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision* 42, 3 (2001), 145–175.

[39] Xi Peng, Jiashi Feng, Shijie Xiao, Wei-Yun Yau, Joey Tianyi Zhou, and Songfan Yang. 2018. Structured autoencoders for subspace clustering. *IEEE Transactions on Image Processing* 27, 10 (2018), 5076–5086.

[40] Xi Peng, Canyi Lu, Zhang Yi, and Huajin Tang. 2018. Connections between nuclear-norm and Frobenius-norm-based representations. *IEEE Transactions on Neural Networks and Learning Systems* 29, 1 (2018), 218–224.

[41] A. Pentina, V. Sharmanska, and C. H. Lampert. 2015. Curriculum learning of multiple tasks. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5492–5500.

[42] H. Qiu and E. Hancock. 2007. Clustering and embedding using commute times. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 11 (2007), 1873–1890.

[43] Z. Ren, D. Dong, H. Li, and C. Chen. 2018. Self-paced prioritized curriculum learning with coverage penalty in deep reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems* (2018), 1–11.

[44] K. Simonyan and A. Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv: 1409.1556v6* (2014).

[45] J. Supancic and D. Ramanan. 2013. Self-paced learning for long-term tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2379–2386.

[46] J. Tang, R. Hong, S. Yan, T. Chua, G. Qi, and R. Jain. 2011. Image annotation by Knn-sparse graph-based label propagation over noisily tagged web images. *ACM Transactions on Intelligent Systems and Technology* 2, 2 (2011), 14.

[47] Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*.

[48] K. Tu and V. Honavar. 2011. On the utility of curricula in unsupervised learning of probabilistic grammars. In *International Joint Conference on Artificial Intelligence*. 1523–1528.

[49] V. Vapnik. 1998. Statistical learning theory. *Encyclopedia of the Sciences of Learning* 41, 4 (1998), 3185–3185.

[50] B. Wang, Z. Tu, and J. Tsotsos. 2013. Dynamic label propagation for semi-supervised multi-class multi-label classification. In *IEEE International Conference on Computer Vision*. 425–432.

[51] J. Wang, F. Wang, and C. Zhang. 2009. Linear neighborhood propagation and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 9 (2009), 1600–1615.

[52] P. Wang, L. Ji, J. Yan, D. Dou, N. Silva, Y. Zhang, and L. Jin. 2018. Concept and attention-based CNN for question retrieval in multi-view learning. *ACM Transactions on Intelligent Systems and Technology* 9, 4 (2018), 41.

[53] Y. Wang, J. Sharpnack, A. Smola, and R. Tibshirani. 2016. Trend filtering on graphs. *Journal of Machine Learning Research* 17 (2016), 1–41.

[54] J. Xiao, A. Ehinger, J. Hays, A. Torralba, and A. Oliva. 2016. SUN database: Exploring a large collection of scene categories. *International Journal of Computer Vision* 119, 1 (2016), 3–22.

[55] C. Xu, D. Tao, and C. Xu. 2013. A survey on multi-view learning. *arXiv:1304.5634* (2013).

[56] Z. Xu, Z. Hong, Y. Zhang, J. Wu, A. Tsoi, and D. Tao. 2016. Multinomial latent logistic regression for image understanding. *IEEE Transactions on Image Processing* 25, 2 (2016), 973–987.

[57] J. Liu S. Wang L. Duan Y. Lou, Y. Bai. 2019. Feature distance adversarial network for vehicle re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[58] D. Zhou and O. Bousquet. 2003. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*. 321–328.

[59] X. Zhu, Z. Ghahramani, and J. Lafferty. 2003. Semi-supervised learning using gaussian fields and harmonic functions. In *International Conference on Machine Learning*. 912–919.

[60] X. Zhu and B. Goldberg. 2009. *Introduction to Semi-Supervised Learning*. Morgan & Claypool Publishers.

[61] X. Zhu, J. Lafferty, and Z. Ghahramani. 2003. *Semi-supervised Learning: From Gaussian Fields to Gaussian Processes*. Technical Report. CMU-CS-03-175.

Received November 2018; revised March 2019; accepted March 2019