# Robust Graph-Based Semisupervised Learning for Noisy Labeled Data via Maximum Correntropy Criterion

Bo Du, *Senior Member, IEEE*, Xinyao Tang, Zengmao Wang, Lefei Zhang, *Member, IEEE*, and Dacheng Tao, *Fellow, IEEE*

***Abstract*—Semisupervised learning (SSL) methods have been proved to be effective at solving the labeled samples shortage problem by using a large number of unlabeled samples together with a small number of labeled samples. However, many traditional SSL methods may not be robust with too much labeling noisy data. To address this issue, in this paper, we propose a robust graph-based SSL method based on maximum correntropy criterion to learn a robust and strong generalization model. In detail, the graph-based SSL framework is improved by imposing supervised information on the regularizer, which can strengthen the constraint on labels, thus ensuring that the predicted labels of each cluster are close to the true labels. Furthermore, the maximum correntropy criterion is introduced into the graph-based SSL framework to suppress labeling noise. Extensive image classification experiments prove the generalization and robustness of the proposed SSL method.**

***Index Terms*—Graph semisupervised learning, half quadratic (HQ) optimization, maximum correntropy criterion, robust.**

## I. Introduction

COLLECTING labeled examples can be expensive and difficult in many data analysis and mining applications, such as medical diagnosis and analysis, human action recognition, image retrieval, and remote sensing image classification [1]–[5]. However, there is an abundance of available unlabeled data. To alleviate this issue, researchers have tried to use plentiful unlabeled data together with limited labeled data, and thus developed semisupervised learning (SSL) methods [6]. It has been confirmed that SSL not only avoids wasting data sources but also improves upon the generalization capability of traditional supervised learning methods.

Manuscript received June 25, 2017; revised November 6, 2017, January 11, 2018, and January 24, 2018; accepted January 30, 2018. Date of publication February 27, 2018; date of current version February 22, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61471274, Grant 41431175, Grant 61771349, and Grant 61711530239, and in part by the Australian Research Council Projects under Grant FL-170100117, Grant DP-180103424, Grant DP-140102164, and Grant LP-150100671. This paper was recommended by Associate Editor Y. Jin. *(Corresponding author: Lefei Zhang.)*

B. Du, X. Tang, Z. Wang, and L. Zhang are with the School of Computer, Wuhan University, Wuhan 430072, China (e-mail: zhanglefei@whu.edu.cn).

D. Tao is with the UBTECH Sydney Artificial Intelligence Centre and the School of Information Technologies, Faculty of Engineering and Information Technologies, University of Sydney, Darlington, NSW 2008, Australia.

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

Digital Object Identifier 10.1109/TCYB.2018.2804326

In the literature, some widely used SSL methods have been proposed, which can be categorized into the following major classes.

1) The generative model [7]–[9] assumes that all of the data is generated by an underlying model $P(x, y) = P(x|y)P(y)$, where $P(x|y)$ is an identifiable mixture distribution, such as the Gaussian mixture model. This assumption connects unlabeled samples with the learning target through the parameters of the underlying model.
2) Self-training [10]–[12] is another kind of commonly used SSL method. A self-training learner first uses a few labeled samples to train a model and classify the unlabeled samples, and then adds the most confident unlabeled samples together with their predicted labels into the training set. The model is retrained and the procedure is repeated. This method is simple but may cause error accumulation.
3) Co-training [13]–[15] is a simple but effective SSL method. The co-training learner assumes that the feature can be split into two subsets. The two subsets are "sufficient" and "conditionally independent" [6]. The co-training learner trains two models using labeled samples with two different feature subsets. The two models predict the labels for the unlabeled samples separately. Then, each model selects some of the most confident unlabeled samples (and the predicted labels) to be added into the training set of another model. The two models are retrained with the respective updated training sets. This process is repeated until the stop condition is reached.
4) Graph-based SSL [16]–[20] is one of the most influential fields in SSL. Given a dataset, this kind of method maps it to a graph, where the nodes are the examples (both labeled and unlabeled) in their feature space, and the strength of the edge is proportional to the similarity of each example pair. The main idea is label propagation over the graph with cluster consistency. This kind of method has the characteristics of a solid mathematical basis, high accuracy, and quick calculation speed.

There are two basic assumptions to model the relations among instances, namely the clustering assumption [21] and the manifold assumption [22], [23]. The clustering assumption supposes that the instances which fall in the same cluster

should share the same label, while the manifold assumption suggests that the whole dataset is distributed on a manifold structure, in which the nearby samples should have similar outputs. Many existing graph-based SSL frames are designed based on the above two assumptions [16]–[20]. Generally speaking, the goal of graph-based SSL is to estimate a function $f$ on the graph where $f$ satisfies the following two conditions: 1) it is close to the given labels of labeled nodes and 2) it is smooth across the whole graph. Therefore, the overall optimization includes two terms.

1) The first one is a loss function, which forces the predicted labels unequal to true labels.
2) The second term is a regularizer, which smooths the whole graph utilizing the geometrical information reflected by the marginal distribution.

However, in many existing graph-based SSL methods [16]–[20], the regularizer only mines the structural information of the graph, but weakens the label constraint. This results in the predicted labels of unlabeled samples not certainly being identical to the ground truth. In order to address this issue, in this paper, we try to improve the regularization term by introducing supervised information to the regularizer.

Besides, many traditional graph-based SSL frames are not robust when faced with too much noisy labels (e.g., labeling mistakes from annotators or computing errors from extraction algorithms). Unfortunately, owing to various types of noise in reality, these labels are extremely unreliable. Some possible labeling noises are illustrated in Fig. 1. Confusion noise occurs when the two instances are too similar to distinguish. For example, "jackfruit" may be labeled as "durian." Context noise, which refers to the same labels in a different context, may have different meanings. For example, "Michael Jordan" may be a basketball player in sports news, while he may be a computer scientist in science news. Random noise is caused by the mismatch between an image and its surrounding text. For example, a "T-shirt" may be incorrectly labeled as "jeans" because of the mismatch between the picture of a T-shirt and its surrounding text jeans; likewise, a "sneaker" may be labeled as "stockings" because of the mismatch. This labeling noise introduces a bias to the supervised information and misleads the classification process. Many traditional graph SSL frames correspond to the L2-norm optimization problem [18], [19], while L2-norm is sensitive to noise.

In this paper, we try to solve this problem by using maximum correntropy criterion instead of the L2-norm. Correntropy belongs to the concept of information theoretic [24]. The maximum correntropy is introduced to measure the similarity, which is effective at reducing the adverse influence of noisy labels. The L2-norm extends the error by square when there are labeling noises. While correntropy measures the similarity between labels and predicted values with a kernel function. When the labels are wrong with noises, the labels differ greatly from predicted values. Then the correntropy between labels and predicted values would be small, thus have smaller impact on objective function. Hence, the influence of noisy labels will be restrained. Furthermore, the half quadratic (HQ) optimization technique [25] is adopted to handle the nonlinear kernel function optimization in the maximum correntropy criterion so that the computational complexity of optimization can be greatly reduced.

Graph-based SSL methods are transductive in nature [6]. This means that they cannot easily extend to a new test point outside of the labeled and unlabeled samples in the training set. The graph should be restructured when new samples need to be predicted. When a model can only predict the samples participating in the training process, but cannot predict new samples that are unseen in the training process, this is considered to be transductive learning. In this paper, we try to develop a kind of inductive learning, which can predict the labels for samples not seen in the training process. The structure of the graph is fixed in the training set (including labeled and unlabeled samples), and is not altered when there are new samples. The new samples are classified through their similarity to the labeled samples.

The main contributions of this paper can be summarized as follows.

1) It improves the discriminability of the graph-based SSL framework by introducing supervised information to the regularizer.
2) It introduces the maximum correntropy criterion into the proposed model for robust learning. To the best of our knowledge, it is the first time that the maximum correntropy criterion has been used in SSL for handling the labeling noise problem.
3) It proposes a robust SSL model for inductive learning. This avoids expensive graph computation every time new samples arrive.

The rest of this paper is organized as follows. In Section II, a review of the related work on graph-based SSL and correntropy is given. Section III presents the proposed robust SSL method. Section IV verifies the generalization and robustness of the proposed method through a series of image classification experiments. Finally, Section V draws the conclusions of this paper.

## II. Related Work

### A. Graph-Based Semisupervised Learning

Consider a dataset $\{\mathbf{x}_i\}_{i=1}^{n}$ with $l$ labeled samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$ and $u = n - l$ unlabeled samples $\{\mathbf{x}_i\}_{i=l+1}^{n}$, where $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{id}] \subseteq R^d$ denotes the $d$-dimensional feature space and $y \subseteq \{-1, 1\}$ denotes the class label. We now set $L = \{1, 2, \ldots, l\}$ as the index set of $l$ labeled instances in the dataset, $U = \{l + 1, l + 2, \ldots n\}$ as the index set of $u = n - l$ unlabeled samples in the dataset, and $N = L \cup U$ as the index set of all n instances. Graph-based SSL [6] defines a graph where the nodes are labeled and unlabeled examples in the dataset, and the strength of the edges is proportional to the similarity between examples. To exploit unsupervised information, SSL connects the data distribution information revealed by unlabeled samples with the class labels under clustering and manifold assumptions.

The goal of graph-based SSL is to estimate a classification function $f$ on the graph; $f_i = f(\mathbf{x}_i)$ is the classification function value of the $i$th sample $\mathbf{x}_i$ in the dataset and the predicted

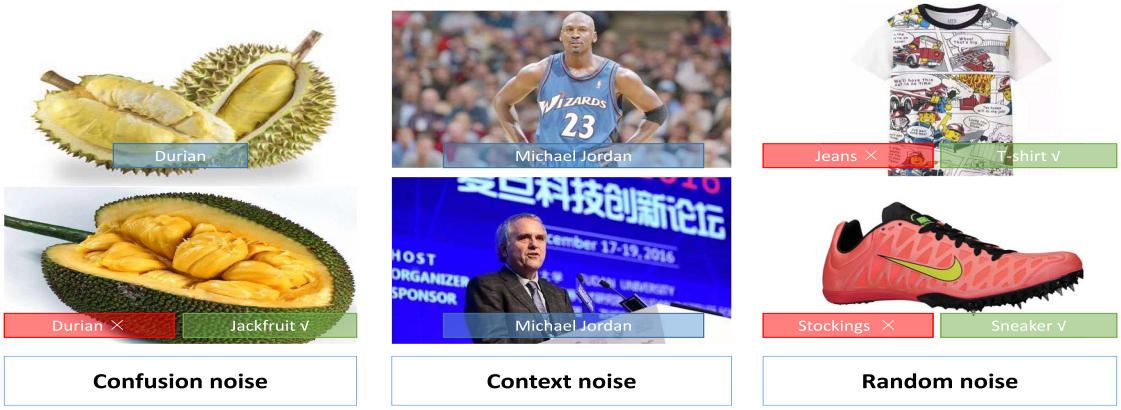



Fig. 1. Here are some of possible labeling noises. Confusion noise occurs when the two instances are too similar to distinguish. For example, *jackfruit* may be labeled as *durian*. Context noise, which refers to the same labels in a different context, may have different meanings. For example, *Michael Jordan* may be a basketball player in sports news, while he may be a computer scientist in science news. Random noise is caused by the mismatch between an image and its surrounding text. For example, a *T-shirt* may be incorrectly labeled as *jeans* because of the mismatch between the picture of a T-shirt and its surrounding text *jeans*; likewise, a *sneaker* may be labeled as *stockings* because of the mismatch.

labels can be obtained by $\widehat{y}_i = \text{sgn}(f_i)$. Depending on cluster assumption and manifold assumption, $f$ should satisfy two conditions: 1) it is close to the given labels of labeled nodes and 2) it is smooth over the whole graph. To satisfy these conditions, many traditional graph-based SSL algorithms can be formulated in the following framework [26]:

$$\min_f L(y,f) + \xi\,\Omega(f) \tag{1}$$

where the first term is a loss function, measuring the loss between the ground truth and prediction of the labeled samples, and making the prediction close to the ground truth. Common loss functions include quadratic loss function and logarithmic loss function. The second term is a regularizer, which smooths the classification $f$ (makes the $f$ value of neighboring samples close). Most of the graph-based SSL algorithms differ in the loss and regularization functions. Several classical graph-based SSL frames are briefly explained as follows.

The first graph-based SSL method is mincut (also known as st-cut) [26]. In mincut, the positive labels are seen as sources and the negative labels are viewed as sinks. First find a minimal set of edges that satisfies the following condition: after removing all the edges of this set, there is no connection between sources and sinks. Then, the points connecting to sources are marked as positive and the points connecting to sinks are labeled as negative. The loss function of mincut is a quadratic loss function with infinite weight

$$L(y,f) = \infty \sum_{i\in L} \|y_i - f_i\|_2^2 \tag{2}$$

so that the $f$ values of labeled data are in fact their given labels. The regularizer is

$$\Omega(f) = \frac{1}{2}\sum_{i\in N}\sum_{j\in N}\|f_i - f_j\|_2^2 W_{ij} \tag{3}$$

where $W_{ij}$ is the weight coefficient between $\mathbf{x}_i$ and $\mathbf{x}_j$, evaluating the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$. Generally, a Gaussian kernel function is used to calculate the weight between two samples

$$W_{ij} = \exp\left(-\mu\|\mathbf{x}_i - \mathbf{x}_j\|_2^2\right) \tag{4}$$

where $\mu$ is the kernel parameter. Putting (2) and (3) together, the objective function of mincut is

$$\begin{aligned} &\infty \sum_{i\in L}\|y_i - f_i\|_2^2 + \frac{1}{2}\sum_{i\in N}\sum_{j\in N}\|f_i - f_j\|_2^2 W_{ij} \\ &\text{s.t.} \qquad f_i \in \{0,1\}, i \in N \end{aligned} \tag{5}$$

owing to the integer constraint in (5), one problem of mincut is only giving hard classification without confidence (or classification probabilities).

The Gaussian random fields method (also called the harmonic function method) [27] is a continuous relaxation of the difficulty discrete Markov random fields (or Boltzmann machines). Similar to mincut, this method also uses a quadratic function with infinity weight in (2) as the loss function, and uses a regularizer based on a combinatorial Laplacian

$$\Omega(f) = \frac{1}{2}\sum_{i\in N}\sum_{j\in N}\|f_i - f_j\|_2^2 W_{ij} = \mathbf{f}^T\mathbf{L}\mathbf{f}. \tag{6}$$

The objective function is

$$\infty \sum_{i\in L}\|y_i - f_i\|_2^2 + \mathbf{f}^T\mathbf{L}\mathbf{f} \tag{7}$$

notice that $f_i \in R$, which is the key relaxation to mincut. $\mathbf{f} = [f_1, f_2, \ldots, f_n]^T$ is the classification function vector. $\mathbf{L} \subseteq R^{n\times n}$ is the combinatorial Laplacian matrix which is widely used [28]

$$\mathbf{L} = \mathbf{D} - \mathbf{W}. \tag{8}$$

$\mathbf{W} = [W_{ij}] \subseteq R^{n\times n}$ is the weight (adjacency) matrix of the graph. On large datasets, we can enforce the sparsity on $\mathbf{W}$ by only computing the $k$-nearest neighbors, and preserve the symmetry of $\mathbf{W}$ by $\mathbf{W} \leftarrow (1/2)(\mathbf{W} + \mathbf{W}^T)$. $\mathbf{D} \subseteq R^{n\times n}$ is an $n$th-order diagonal matrix where $D_{ii} = \sum_{j=1}^{n} W_{ij}$. This

method adequately considers the classification probabilities of examples, so that the disadvantage of mincut is overcome.

The local and global consistency method [29] uses a quadratic loss function and normalized Laplacian in the regularizer

$$\xi\Omega(f) = \frac{1}{2}\sum_{i\in N}\sum_{j\in N}\left\|f_i/\sqrt{D_{ii}} - f_j/\sqrt{D_{jj}}\right\|_2^2 W_{ij} = \mathbf{f}^T\mathbf{S}\mathbf{f}. \quad (9)$$

And the objective function is

$$\sum_{i\in L}\|y_i - f_i\|_2^2 + \mathbf{f}^T\mathbf{S}\mathbf{f} \quad (10)$$

while $\mathbf{S}$ is the normalized Laplacian matrix

$$\mathbf{S} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}. \quad (11)$$

The normalized Laplacian is also widely adopted in many SSL studies [29]–[31].

There are some recent excellent works based on graph SSL. Zhang *et al.* [32] proposed a graph-based constrained SSL framework. This framework increase supervised information by constructing and enrich pairwise constraints sets. Nie *et al.* [33] introduced a general graph-based SSL algorithm based on the normalized weights evaluated on data graph, which can output the probabilities of data points belonging to the labeled classes or the novel class. Nie *et al.* [34] proposed a unified manifold learning framework for SSL. An unsupervised dimension reduction has been provided as well, which using a linear regression function to map the new data points. Zhang *et al.* [35] proposed an enhanced version of [34], by fully considering the robust and sparse properties of L2,1-norm. Zhang *et al.* [36] proposed a inductive semisupervised classification model termed adaptive embedded label propagation with weight learning (AELP-WL). This model is based on embedded representation, which has robust characteristics. Zoidi *et al.* [37] improved the advanced label propagation framework by propagation of negative labels. Zhang *et al.* [38] extended the existing neighborhood propagation by regularizing the L2,1-norm on the soft labels during optimization for a better prediction performance. He *et al.* [39] provided a nonnegative sparse algorithm for graph-based SSL. This paper uses approximated algorithm over the $l^0 - l^1$ equivalence theory to compute the nonnegative sparse weights of a graph. Lu and Wang [17] effectively solved the problem of noise by formulating Laplacian regularization as an L1-norm term. Dornaika and Traboulsi [16] provided a discriminative semisupervised dimensionality reduction method which can be viewed as an exponential version of semisupervised discriminant embedding. Yu *et al.* [20] proposed a classification method of SSL, using subspace sparse representation to handle noise and redundant features in high-dimensional data. Zhang *et al.* [19] provided a scale-up graph-based SSL called prototype vector machine (PVM). This method utilize a set of sparse prototypes derived from the full set of data to estimate the graph-based regularizer and the predictive model, thus decreasing the problem size and making sure the smoothness of the output. Wang *et al.* [18] introduced a scalable graph-based SSL algorithm called efficient anchor graph regularization (EAGR) to "reduce" the graph Laplacian for a large size dataset; in order to evaluate the local weights between data points and close anchor nodes, a fast local anchor embedding method has been provided, as well as a normalized graph Laplacian over anchors has been employed.

### B. Correntropy

Recently, correntropy [24] has drawn much attention in obtaining robust analysis [40], [41] in information theoretic learning [42] and effectively handling outliers and non-Gaussian noise [24]. Correntropy is directly associated with the probability of similarity between two random variables in the neighborhood of the joint space controlled by the kernel bandwidth, and could be considered as a generalized similarity measure between two random variables $A$ and $B$

$$V_\sigma(A, B) = E(k_\sigma(A - B)) \quad (12)$$

where $E(.)$ is the expectation operator, and $k_\sigma(.)$ is the kernel function that satisfies the Mercer theory [43]. In practice, the joint probability distribution function is unavailable, and finite samples $\{(a_i, b_i)\}_{i=1}^m$ are given, resulting in the sample estimator of correntropy

$$\widehat{V}_{m,\sigma}(A, B) = \frac{1}{m}\sum_{i=1}^{m} k_\sigma(a_i - b_i). \quad (13)$$

Using the kernel function, correntropy maps the input space into the high-dimensional space. Different from the conventional kernel-based methods, correntropy works independently with pairwise samples and is symmetric, positive and bounded [24]. Furthermore, correntropy is a localized similarity measure, that is, correntropy between $A$ and $B$ is primarily dictated by the kernel function along the $a = b$ line; and correntropy bears a close relationship with $M$-estimation [44]. Correntropy has been applied to deal with non-Gaussian noise or large outliers in some recent works in pattern recognition filed [45]–[47], and obtained robust results. However, correntropy has been not applied to handling SSL with labeling noise since optimization with the form including correntropy is difficult.

## III. Proposed Method

In the following discussion, the dataset symbols used are the same as those described in the setup in Section II-A.

### A. Robust Graph-Based Semisupervised Learning for Noisy Labeled Data

A typical graph-based SSL frame using a quadratic loss function and the regularizer in (6) can be reformulated as

$$\min_f \sum_{i\in L}\|f_i - y_i\|_2^2 + \xi\sum_{i\in N}\sum_{j\in N}\|f_i - f_j\|_2^2 W_{ij}. \quad (14)$$

Considering the graph-based SSL framework in (14), the first term guarantees the minimum loss between true labels and the predicted labels of labeled samples, and the second term guarantees the smoothness of the whole graph, i.e., the output values $f$ of the neighboring data points are similar. The

second term mines the structure information of the graph; the cluster structure is formed by data points, and the samples in the same clusters have similar labels. However, the predicted labels of each cluster are constrained by the first term indirectly, so the label constraint is weak. This framework cannot effectively ensure that the predicted labels of unlabeled samples are similar to the true labels of the similar labeled samples in the same clusters. To strengthen the label constraint, we add supervised information to the regularizer. The improved graph-based SSL framework is formulated as follows:

$$\min_f \sum_{i\in L}\|f_i - y_i\|_2^2 + \beta_1\frac{1}{u}\sum_{i\in U}\sum_{j\in U}\|f_i - f_j\|_2^2 W_{ij} + \beta_2\frac{1}{l}\sum_{i\in L}\sum_{j\in U}\|y_i - f_j\|_2^2 W_{ij} \tag{15}$$

the first term is the same quadratic loss function as in (14). The second and third terms are the regularizer. The second term guarantees that the outputs of the adjacent samples are similar. The third term guarantees that the predicted labels of the unlabeled samples are similar to the true labels of the neighboring labeled samples. The meaning of the third term that minimizes the weighted difference between $y_i$ and $f_j$ is intuitive. If the weight $W_{ij}$ between labeled sample $\mathbf{x}_i$ (corresponding to label $y_i$) and the unlabeled samples $\mathbf{x}_j$ (corresponding to classification function value $f_i$) is smaller, $\mathbf{x}_i$ and $\mathbf{x}_j$ are more similar. Thus, we minimize the weighted difference between $y_i$ and $f_j$ to make the predicted label value of $\mathbf{x}_j$ ($f_j$) close to the label of $\mathbf{x}_i$ ($y_i$). $\beta_1$ and $\beta_2$ are the tradeoff coefficients to balance the three terms, and the numerators of the second term and the third term eliminate the influence of the data scale of the unlabeled samples and the labeled samples from different datasets. However, the L2-norm in (15) is sensitive to noisy labels. When there are noisy labels, the L2-norm of the difference between noisy labels and predicted value may enlarge the error by square. Taking advantages of the robustness to noise of correntropy, we replace the two-norm in (15) with the correntropy and use Gaussian kernel function to calculate the correntropy. The robust model is as follows:

$$\max_f \sum_{i\in L}\exp\left(-\frac{\|f_i - y_i\|_2^2}{2\sigma^2}\right) + \beta_1\frac{1}{u}\sum_{i\in U}\sum_{j\in U}\exp\left(-\frac{\|f_i - f_j\|_2^2}{2\sigma^2}\right)W_{ij} + \beta_2\frac{1}{l}\sum_{i\in L}\sum_{j\in U}\exp\left(-\frac{\|y_i - f_j\|_2^2}{2\sigma^2}\right)W_{ij} \tag{16}$$

where $\sigma$ is the kernel parameter. Based on (16), we have made the following observations. First, we observe the first term. If the $y_i$ is incorrect with noise and the $f_i$ is predicted correctly, the difference between $y_i$ and $f_i$ would be large. That is to say, the correntropy between $y_i$ and $f_i$ would be small. Thus this term have a tiny contribution to the objective function. The labeling noises could be suppressed. Then, for the second term, if $\mathbf{x}_i$ is close to $\mathbf{x}_j$, the weight $W_{ij}$ should be paid more attention in the objective function. So the correntropy between $f_i$ and $f_j$ should be closed to the maximum value 1. Namely the value of $f_i$ is identical with $f_j$. This fits the manifold structure, the nearby samples should have similar outputs. Finally, the third term is similar to the second term. If a labeled sample $\mathbf{x}_i$ is close to a unlabeled samples $\mathbf{x}_j$, the value of weight $W_{ij}$ would be large, and the objective function would put more emphasis on this term. So the correntropy between $y_i$ and $f_j$ is closed to the maximum value 1. This results in that the value of $y_i$ is closed to $f_j$.

For the classification function $f(\mathbf{x})$, we define a linear regression model in the kernel space as $f(\mathbf{x}) = \phi(\mathbf{x})\boldsymbol{\omega}$, where $\boldsymbol{\omega} = [\omega_1, \omega_2, \ldots, \omega_d]^T$ and $\phi(\mathbf{x})$ is the nonlinear mapping to the kernel space. We use the auxiliary variable $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_l]^T$ to express $\boldsymbol{\omega}$ in the kernel space as $\boldsymbol{\omega} = \sum_{e\in L}\theta_e\phi(\mathbf{x}_e)^T$. Then the classification function $f(\mathbf{x})$ can be expressed as the weighted linear sum of the kernel function value between data point $\mathbf{x}$ and labeled points

$$f(\mathbf{x}) = \sum_{e\in L}\theta_e\phi(\mathbf{x})\phi(\mathbf{x}_e)^T = \sum_{e\in L}\theta_e k(\mathbf{x}, \mathbf{x}_e). \tag{17}$$

Substituting (17) into (16), the model is translated into an optimization problem of the variable $\theta$

$$\max_\theta \sum_{i\in L}\exp\left(-\frac{\|\sum_{e\in L}\theta_e k(\mathbf{x}_i, \mathbf{x}_e) - y_i\|_2^2}{2\sigma^2}\right) + \beta_1\frac{1}{u}\sum_{i\in U}\sum_{j\in U}\exp\left(-\frac{\|\sum_{e\in L}\theta_e k(\mathbf{x}_i, \mathbf{x}_e) - \sum_{e\in L}\theta_e k(\mathbf{x}_j, \mathbf{x}_e)\|_2^2}{2\sigma^2}\right)W_{ij} + \beta_2\frac{1}{l}\sum_{i\in L}\sum_{j\in U}\exp\left(-\frac{\|y_i - \sum_{e\in L}\theta_e k(\mathbf{x}_j, \mathbf{x}_e)\|_2^2}{2\sigma^2}\right)W_{ij}. \tag{18}$$

When $\theta$ is obtained through solving the optimization problem in (18), the classification function $f$ can be calculated by (17).

### B. Optimization Solution

Owing to the introduction of correntropy, the optimization problem in (18) is exponential. Solving this kind of optimization problem is complex and difficult. HQ optimization technology is used to obtain an approximate solution of this exponential optimization problem in this paper, which was proved to rapidly converge [25].

Depending on the property of the convex conjugate function [48], there would be a convex conjugate function $\varphi$ to make sure

$$g(x) = \exp\left(-\frac{\|x\|_2^2}{2\sigma^2}\right) = \arg\max_{p'}\left(p'\frac{\|x\|_2^2}{\sigma^2} - \varphi(p')\right) \tag{19}$$

where $p'$ is the auxiliary variable. With a fixed variable $x$, the maximum of $g(x)$ is achieved at $p' = -g(x)$. By substituting (19) into (18), the augmented objective function in an enlarged parameter space is acquired

$$E(\theta, p, q, r) = \sum_{i\in L}\left(p_i\frac{\|\sum_{e\in L}\theta_e k(\mathbf{x}_i, \mathbf{x}_e) - y_i\|_2^2}{\sigma^2} - \varphi(p_i)\right)$$

$$+\beta_1\frac{1}{u}\sum_{i\in U}\sum_{j\in U}\left(q_{ij}\frac{\left\|\sum_{e\in L}\theta_e k(\mathbf{x}_i,\mathbf{x}_e)-\sum_{e\in L}\theta_e k\left(\mathbf{x}_j,\mathbf{x}_e\right)\right\|_2^2}{\sigma^2}-\varphi(q_{ij})\right)W_{ij}$$
$$+\beta_2\frac{1}{l}\sum_{i\in L}\sum_{j\in U}\left(r_{ij}\frac{\left\|y_i-\sum_{e\in L}\theta_e k\left(\mathbf{x}_j,\mathbf{x}_e\right)\right\|_2^2}{\sigma^2}-\varphi(r_{ij})\right)W_{ij} \tag{20}$$

where $p$, $q$, and $r$ are the auxiliary variables of HQ optimization. According to (19), $\theta$ can be calculated by solving the following optimization problem:

$$\theta=\arg\max_{\theta,p,q,r}\sigma^2E(\theta,p,q,r). \tag{21}$$

Then we maximize the augmented objective function $E$ by the following two steps alternately.

1) Compute the auxiliary variables $p$, $q$, and $r$ at $t$+1th iteration

$$p_i^{t+1}=-\exp\left(-\frac{\left\|\sum_{e\in L}\theta_e^t k(\mathbf{x}_i,\mathbf{x}_e)-y_i\right\|_2^2}{2\sigma^2}\right),\ i\in L \tag{22}$$

$$q_{ij}^{t+1}=-\exp\left(-\frac{\left\|\sum_{e\in L}\theta_e^t k(\mathbf{x}_i,\mathbf{x}_e)-\sum_{e\in L}\theta_e^t k\left(\mathbf{x}_j,\mathbf{x}_e\right)\right\|_2^2}{2\sigma^2}\right)\quad i,j\in N \tag{23}$$

$$r_{ij}^{t+1}=-\exp\left(-\frac{\left\|y_i-\sum_{e\in L}\theta_e^t k\left(\mathbf{x}_j,\mathbf{x}_e\right)\right\|_2^2}{2\sigma^2}\right),\ i\in L,j\in U. \tag{24}$$

Obviously, $q$ and $r$ are both symmetrical.

2) Solve $\theta$ at $t+1$th iteration according to (25)

$$\theta^{t+1}=\arg\max_{\theta^{t+1}}\sum_{i\in L}p_i^{t+1}\left\|\sum_{e\in L}\theta_e k(\mathbf{x}_i,\mathbf{x}_e)-y_i\right\|_2^2+\beta_1\frac{1}{u}\sum_{i\in U}\sum_{j\in U}q_{ij}^{t+1}\left\|\sum_{e\in L}\theta_e k(\mathbf{x}_i,\mathbf{x}_e)-\sum_{e\in L}\theta_e k\left(\mathbf{x}_j,\mathbf{x}_e\right)\right\|_2^2W_{ij}+\beta_2\frac{1}{l}\sum_{i\in L}\sum_{j\in U}r_{ij}^{t+1}\left\|y_i-\sum_{e\in L}\theta_e k\left(\mathbf{x}_j,\mathbf{x}_e\right)\right\|_2^2W_{ij}. \tag{25}$$

The quadratic programming in (25) can be translated into a standard form

$$\min_{\theta}\frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{H}\boldsymbol{\theta}+\boldsymbol{c}^T\boldsymbol{\theta} \tag{26}$$

where

$$\mathbf{H}=-\left(\mathbf{K}_{L,L}^T\mathbf{P}\mathbf{K}_{L,L}+\frac{1}{u}\beta_1\mathbf{K}_{U,L}^T\mathbf{L}'_{U,U}\mathbf{K}_{U,L}+\frac{1}{l}\beta_2\mathbf{K}_{U,L}^T\mathbf{L}''_{U,U}\mathbf{K}_{U,L}\right) \tag{27}$$

$$\mathbf{c}^T=\mathbf{y}_L^T\mathbf{P}\mathbf{K}_{L,L}-\frac{1}{l}\beta_2\mathbf{y}_L^T\mathbf{L}''_{L,U}\mathbf{K}_{U,L}. \tag{28}$$

The various symbols in (27) and (28) are given as follows.

1) $\mathbf{K}=[K_{ij}]\in R^{n\times n}$ is the kernel Gram Matrix over the whole $n$ samples in the dataset where
$$K_{ij}=\exp\left(-\mu\|\mathbf{x}_i-\mathbf{x}_j\|_2^2\right). \tag{29}$$
Then $\mathbf{K}_{L,L}=\mathbf{K}(L,L)$ and $\mathbf{K}_{U,L}=\mathbf{K}(U,L)$ is the submatrix of $\mathbf{K}$.
2) $\mathbf{P}=\mathrm{diag}(p_i)\in R^{l\times l}$ is the auxiliary matrix.
3) $\mathbf{L}'_{U,U}\in R^{u\times u}$ can be seen as the analogous graph Laplacian matrix, where $\mathbf{L}'_{U,U}=\mathbf{D}'_{U,U}-\mathbf{W}'_{U,U}$. $\mathbf{W}'=[W'_{ij}]\in R^{n\times n}$ is the analogous weight matrix and $W'_{ij}=W_{ij}q_{ij}$. $\mathbf{W}'_{U,U}=\mathbf{W}'(U,U)$ is the submatrix of $\mathbf{W}'$. $\mathbf{D}'_{U,U}=\mathrm{diag}(\sum_{j\in U}W'_{ij})\in R^{u\times u}, i\in U$ is a $u$th-order diagonal matrix.
4) $\mathbf{L}''\in R^{n\times n}$ can also be considered as the analogous graph Laplacian matrix where $\mathbf{L}''=\mathbf{D}''-\mathbf{W}''$. $\mathbf{W}''=[W''_{ij}]\in R^{n\times n}$ is another analogous weight matrix and $\mathbf{W}''=W_{ij}r'_{ij}$. $D''=\mathrm{diag}(\sum_{j\in N}W''_{ij})\in R^{n\times n}$ is an $n$th-order diagonal matrix. The computation of auxiliary variable $r'_{ij}$ is a bit different from (24). First, we define a vector $\mathbf{Z}=[y_1,y_2,\ldots,y_l,\mathbf{K}_{l+1,L}\boldsymbol{\theta},\mathbf{K}_{l+2,L}\boldsymbol{\theta},\ldots,\mathbf{K}_{n,L}\boldsymbol{\theta}]$ where $\mathbf{K}_{i,L}=\mathbf{K}(i,L)$ and $\mathbf{K}_{i,L}\boldsymbol{\theta}$ is equivalent to $\sum_{e\in L}\theta_e K(\mathbf{x}_i,\mathbf{x}_e^T)$, which is actually $f_i$. Then, the auxiliary variable $r'_{ij}$ is computed
$$r'_{ij}=\begin{cases}0, & \text{if } (i,j\in L) \text{ or } (i,j\in U)\\ -\exp\left(-\frac{(z_i-z_j)^2}{2\sigma^2}\right), & \text{else.}\end{cases}\quad i,j\in N \tag{30}$$
The auxiliary variable $r'$ remains symmetrical.
5) $\mathbf{y}_L=[y_1,y_2,\ldots,y_l]^T$ is the $l$-dimensional labels vector.

There are many existing solvers [49] for the standard quadratic programming in (26). Owing to the quadratic optimization in (25) being translated into the standard form in (26), the alternate two steps to maximize the objective function in (21) make corresponding subtle changes. First, use $\theta$ calculated at the last iteration to update the auxiliary variables $p$, $q$, and $r$ by (22), (23), and (30). Second, use the auxiliary variables $p$, $q$ and $r$ obtained in the first step to compute the coefficients of the standard quadratic programming in (26) by (27) and (28), and then solve the programming problems to obtain the updated $\theta$ .

The algorithm to acquire $\theta$ is summarized as Algorithm 1. According to Proposition 1, Algorithm 1 is convergent and Proposition 1 is proven in the Appendix.

*Proposition 1:* The augmented objective function sequence $E(\theta^t,p^t,q^t,r^t)$ generated by Algorithm 1 converges.

### C. Prediction

Temporarily, we consider transductive SSL. Transductive SSL only predicts the samples that are visible in the training process, i.e., $\{\mathbf{x}_i|i\in L\cup U\}$. For a binary classification task, we should first obtain auxiliary variable $\theta$ using Proposition 1. According to (17), we can restore the classification function values of unlabeled samples $\mathbf{f}_u=[f_{l+1},f_{l+2},\ldots,f_n]^T$ by

$$\mathbf{f}_u=\mathbf{K}_{U,L}\boldsymbol{\theta}. \tag{31}$$

**Algorithm 1**

**Input:** Input: feature set $X = \{\mathbf{x}_i | i \in L \cup U\}$ including labeled samples and unlabeled samples;
label set $y_L = \{y_1, y_2, \ldots, y_l\}$;
trade-off coefficients $\beta_1$ and $\beta_2$;
kernel parameter $\mu$ of weight matrix in eq. (4);
k-nearest neighbors number $k$ of weight matrix;
kernel parameter $\lambda = \frac{1}{2\sigma^2}$ in eq. (22), eq. (23), eq. (24) and eq. (30).

**Output:** $\theta$

- Initialize $f_i^0 = \begin{cases} y_i, i \in L \\ 0, i \in U \end{cases}$, then initialize $p^0$, $q^0$, $r'^0$ according to eq. (17), eq. (22), eq. (23), eq. (30);
- Calculate weight matrix $\mathbf{W}$ according to eq. (4), and enforce the sparsity on $\mathbf{W}$ by only computing the k-nearest neighbors.
- Calculate kernel matrix $\mathbf{K}(N, L)$ according to eq. (29).

**Rpeat**:

- Using $p^{t-1}$, $q^{t-1}$, $r'^{t-1}$, compute $\mathbf{H}^t$ and $\mathbf{c}^t$ according to eq. (27) and eq. (28);
- Obtain $\theta^t$ by solving the quadratic programming in eq. (26);
- Using $\theta^t$, update $p^t$, $q^t$, $r'^t$ according to eq. (22), eq. (23) and eq. (30).

**Until** convergence is achieved.

Then, the predicted labels of unlabeled training samples $\widehat{\boldsymbol{y}}_U = [\widehat{y}_{l+1}, \widehat{y}_{l+2}, \ldots, \widehat{y}_n]$ can be obtained

$$\widehat{\boldsymbol{y}}_U = \operatorname{sgn}(\mathbf{f}_U). \tag{32}$$

Next, the situation of a multiclassification task is discussed. Given a dataset including $c$ categories $\{(\upsilon_i)_{i=1}^c\}$, we decompose the multiclassification task into $c(c-1)/2$ binary classification tasks using a one-against-one strategy. $C_{ij}$ is a binary-classifier between class $\upsilon_i$ and class $\upsilon_j$, where class $\upsilon_i$ is viewed as positive class, and class $\upsilon_j$ is viewed as negative class. If $C_{ij}$ determines an unlabeled sample $\mathbf{x}_i$ is a positive sample, class $\upsilon_i$ gets a vote, otherwise, class $\upsilon_j$ gets a vote. Finally, $\mathbf{x}_i$ is predicted from the class label that gains the most votes.

Finally, consider the inductive SSL. For an unseen sample $\mathbf{x}$ to be forecast, we utilize the auxiliary variable $\theta$ obtained using Algorithm 1 to compute the classification function value of $\mathbf{x}$

$$f(\mathbf{x}) = \sum_{e \in L} \theta_e K(\mathbf{x}, \mathbf{x}_e). \tag{33}$$

The predicted label of $\mathbf{x}$ can be calculated by

$$\widehat{y} = \operatorname{sgn}(f(\mathbf{x})). \tag{34}$$

Similarly, we use a one-against-one strategy for multiclassification tasks.

## IV. Experiment

### A. Experimental Setup

To verify the effectiveness and robustness of the proposed SSL methods, some representative state-of-the-art methods are used for comparison, as follows.

TABLE I
Details of the Subsets of PASCAL VOC 2007, AWA, and ImageNet

| Data set | Categories | Samples |
|---|---|---|
| Subset1 of PASCAL VOC 2007 | 5 | 3461 |
| Subset2 of PASCAL VOC 2007 | 10 | 2159 |
| Subset of AWA | 11 | 5903 |
| Subset1 of ImageNet | 3 | 3570 |
| Subset2 of ImageNet | 3 | 3366 |
| Subset3 of ImageNet | 3 | 2826 |
| Subset4 of ImageNet | 5 | 4587 |
| Subset5 of ImageNet | 5 | 4890 |
| Subset6 of ImageNet | 5 | 4761 |

1) EAGR [18] is an SSL method with an efficient anchor graph regularization for large datasets.
2) SSC-SSR [20] is a semisupervised classification method based on subspace sparse representation.
3) PVM [19] is a graph-based SSL method using a set of sparse prototypes derived from the data. In this paper, PVM with square loss is used.
4) Support vector machine (SVM) [50] is a method for classical supervised statistical learning, which is now widely used and constantly improved [51].
5) SparseFME [35] is a discriminative sparse flexible manifold embedding method.
6) LNP [52] is a graph-based SSL approach based on a linear neighborhood model.
7) The proposed robust graph-based SSL classification method based on maximum correntropy criterion (RGSSL-MCC).

We carried out classification experiments on three image datasets. Because our focus is verifying the robustness to labeling noise, we have only extracted some subsets from the three image datasets to conduct experiences. This can help speed up the progress of the experiment. The PASCAL Visual Object Classes Challenge 2007 (PASCAL VOC 2007) [53] contains 20 visual object classes in realistic scenes. We have used a subset of the data including 9963 images to conduct multiclassification tasks. The images containing multiple objects are removed, and the remaining images are viewed as samples with single labels. Then, five-category datasets and ten-category datasets have been constructed from those samples for the multiclassification task. An 11-category dataset was extracted from animals with attributes (AWA) [54] for the multiclassification task in this paper. Three- and five-category datasets have also been constructed from ImageNet dataset [55] for the experiments. The details of the experimental datasets are shown in Table I.

Each dataset has been split into two parts: 1) 50% for training and 2) 50% for testing randomly. $P$% training samples have been randomly selected as labeled samples, and the remaining $(100 - P)$% training samples are regarded as unlabeled samples and have been predicted in the transductive learning. In this paper, the range of $P$ is set to {10, 30, 50}. The testing samples have been predicted in the inductive learning. In addition, to verify the robustness to labeling noise, we have

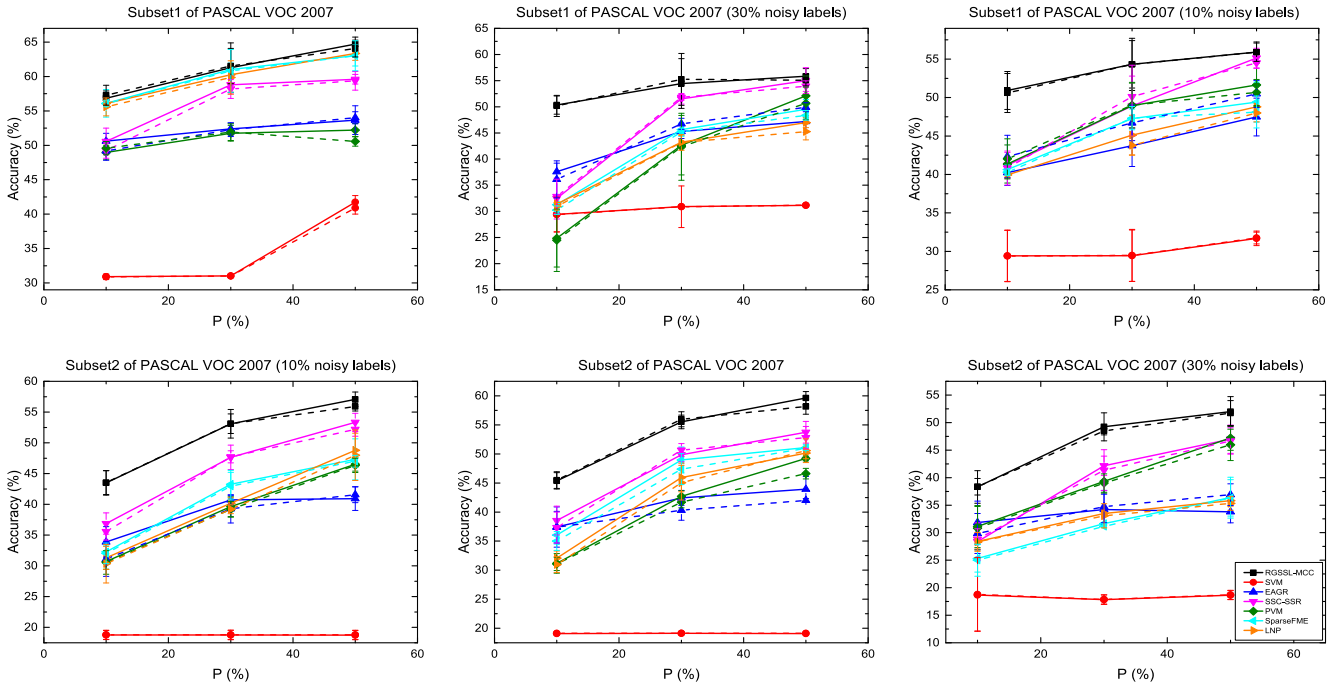



Fig. 2. Comparison of different SSL methods and classical supervised leaning method SVM on the two subsets of PASCAL VOC 2007. The solid line indicates the transductive learning results, and the dashed line indicates the inductive learning results. The first column of this figure shows the results with no noise; the second column shows the results with 10% noisy initial labels; and the third column shows the results with 30% noisy initial labels. Each broken line represents the average result of five independent runs.

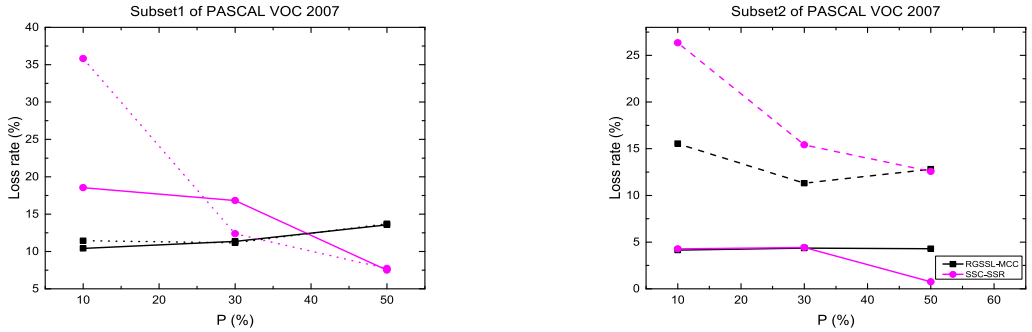



Fig. 3. Classification accurate decline rate when adding labeling noise on subsets of PASCAL VOC 2007. The horizontal axis shows the percentage of labeled samples contained in training set, and the vertical axis shows the classification accuracy decline rate. The solid line indicates the decline rate when adding 10% noisy labels and the dotted line indicates the decline rate when adding 30% noisy labels.

added 10% and 30% noisy initial labels. This can be consider as "class noise" which is mentioned in [56]. The noisy initial labels have been generated as follows: first, we have chosen 10% or 30% labeled training samples randomly; second, for each extracted training sample, we randomly select an incorrect label from the incorrect label set which does not include the true label of this sample and assign the incorrect label for this sample. Five independent experiments with random partition have been run to reduce the influences of randomness.

For the comparative methods, the parameter selections refer to the original papers. For the proposed RGSSL-MCC methods, the tradeoff coefficients $\beta_1$ and $\beta_2$ are both fixed at 1 in all the datasets. The kernel parameter $\mu$ of the weight matrix in (4) is set to 0.005 in the subset of PASCAL VOC 2007 and AWA, and is set to 0.05 in the subset of ImageNet. The $k$-nearest neighbors number $k$ of weight matrix is fixed at 20 in all the datasets. The kernel parameter $\lambda = [(1)/(2\sigma^2)]$ in (22)–(24) and (30) is fixed at 0.1 in all the datasets.

### B. Experimental Results and Analysis

*1) Experimental Results of PASCAL VOC 2007 Subsets:* The experimental results for the subsets of PASCAL VOC 2007 are shown in Fig. 2. The horizontal axis shows the percentage of labeled samples contained in the training set and the vertical axis shows the accuracy of the classification experiments. The solid line indicates the experimental results in the transductive learning and the dashed line indicates the experimental results in the inductive learning. The first column of this figure shows the results when there is no noise; the second column shows the results when 10% noisy initial labels have been added; the third column shows the results when 30% noisy initial labels were added.

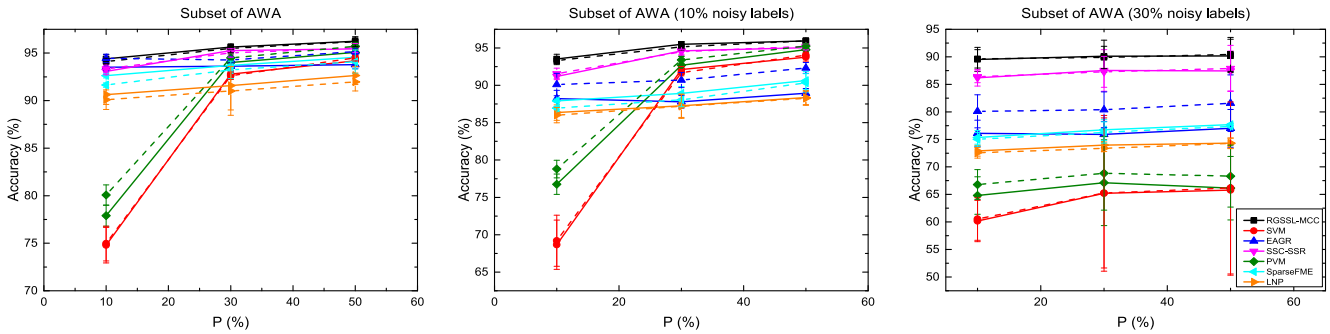



Fig. 4. Comparison of different SSL methods and classical supervised leaning method SVM on a subset of AWA. The solid line indicates the transductive learning results, and the dashed line indicates the inductive learning results. The first column of this figure shows the results with no noise; the second column shows the results with 10% noisy initial labels; and the third column shows the results with 30% noisy initial labels. Each broken line represents the average result of five independent runs.

Intuitively, the obtained results show that the proposed RGSSL-MCC method presents the best performance on both subsets 1 and 2 of PASCAL VOC 2007. When there is no noise, the proposed RGSSL-MCC method achieves observably higher performance than other comparative regularization-based SSL methods. This is due to the following two reasons. First, the maximum correntropy criterion could adjust the weight adaptively, resulting a more accurate result. Second, the third term in objective function makes the predicted values of unlabeled samples close to the labels of labeled samples which are similar to them in the feature space. When 10% or 30% noisy labels are added, the proposed RGSSL-MCC stays ahead of the competitors. This is because: when a label is incorrect with noises, the label may be much different from the predicted value, resulting in a very small correntropy. Thus, the terms with noisy labels would have a small impact on the objective function. The noisy labels cause a deterioration of the accuracy of all the SSL methods. We use the classification accuracy decline rate to express deterioration of the accuracy caused by noisy labels. The classification accuracy decline rate is defined as follows: the accuracy of the non-noise situation subtract the accuracy of the noisy labels situation, and is then divided by the accuracy of the non-noise situation. This yields a rate of decline.

From the perspective of avoiding wordy and redundant description, we have only analyzed the transductive learning classification results by comparing the accuracy decline rate of RGSSL-MCC with that of SSC-SSR. Because SSC-SSR have performed the best compared with other competitive methods for the subsets of PASCAL VOC 2007, we think this comparison is representative.

The classification decline rate when adding labeling noise on subsets of PASCAL VOC 2007 is shown in Fig. 3. The horizontal axis shows the percentage of labeled samples contained in the training set and the vertical axis shows the classification decline rate. The solid line indicates the decline rate when adding 10% noisy labels and the dotted line indicates the decline rate when adding 30% noisy labels.

On the whole, the accuracy decline rate of RGSSL-MCC is smaller than that of SSC-SSR when there are 10% and 30% labeled samples. In addition, the amount of labeling noise is greater, so this robust superiority of RGSSL-MCC is more

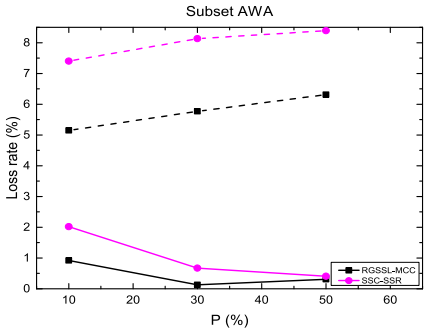



Fig. 5. Classification accuracy decline rate when adding labeling noise on subset of AWA . The horizontal axis shows the percentage of labeled samples contained in training set, and the vertical axis shows the classification accuracy decline rate. The solid line indicates the decline rate when adding 10% noisy labels and the dotted line indicates the decline rate when adding 30% noisy labels.

obvious. The reason may be that: when the number of labeled samples is fairly small, labeling noises may have a significant impact on label propagation; while the RGSSL-MCC can give a small weight to the samples with noisy labels through the maximum correntropy criterion, thus leading to an obvious advantage compared with other SSL methods; when the number of labeled samples is fairly large, the influences on label propagation caused by noisy labels are weaken. So the superiority compared with other SSL methods of RGSSL-MCC is weakened.

*2) Experimental Results of AWA Subset:* The experimental results on the AWA subsets are shown in Fig. 4. The meanings of these lines in Fig. 4 are identical with Fig. 2. We can observe from Fig. 4 that the proposed RGSSL-MCC method has achieved the highest accuracy among all the methods in all the cases. When there are no noises, the superiority of maximum correntropy criterion is not obvious. Here the advantage of RGSSL-MCC comes mainly from the introduction of supervised information to the third term in the objective function. As the number of noisy labels becomes larger, it is clear that the accuracy loss of RGSSL-MCC caused by labeling noise is smaller than other SSL methods, particularly when there are more noisy labels. This is because that: when the noisy labels arise, the correntropy between the noisy label and the corresponding predicted label is very small. Thus the term including

noisy label would make a tinny contribution to the objective function.

In order to quantify the superiority of the robustness of the RGSSL-MCC, we also used the classification accuracy decline rate defined in Section IV-B1. Succinctly, we chose the best competitive method SSC-SSR as a representative to be compared with the RGSSL-MCC. Because the results of transductive learning are similar to inductive learning, we only consider the results of transductive learning.

The classification loss rates when there is labeling noise on the subset of AWA are shown in Fig. 5. The settings of the line charts in Fig. 5 are identical with Fig. 3. The accuracy decline rate of RGSSL-MCC is smaller than that of SSC-SSR. This further shows that RGSSL is more robust to noisy labels.

*3) Experimental Results of ImageNet Subsets:* The experimental results on the subsets of ImageNet are shown in Fig. 6. The settings of the line charts in Fig. 6 are the same as those in Figs. 2 and 4.

Some observations from Fig. 6 are as follows. The proposed RGSSL-MCC method has the best classification accuracy of all the subsets. Under the non-noise condition, the good performance of the proposed RGSSL-MCC method demonstrates that introducing supervised information to the regularizer may reinforce label constraint on unlabeled instances effectively. With 10% and 30% noisy labels, the RGSSL-MCC maintains its leading place. This fully testifies that introducing the maximum correntropy criterion to the graph-based SSL framework can reduce the problem of noisy labels availably.

We consider the accuracy loss of each SSL method when adding noisy labels. Because the accuracy loss when adding noisy labels is unclear from Fig. 6, we compute the classification accuracy decline rate defined in Section IV-B1. Likewise, from this perspective of avoiding the wordy and redundant description, we only choose the most representative SSL methods to be compared with the RGSSL-MCC. For subset1, subset4, and subset6 of ImageNet, SSC-SSR is the best competitive method overall. Thus, we chose SSC-SSR as the comparison. On subset2 of ImageNet, SSC-SSR and EAGR have exhibited the best performance of all the comparison methods so we have chosen them for the comparison. For subset3 and subset5 of ImageNet, it is hard to judge which is the best among SSC-SSR, EAGR, and PVM so we have considered all of these methods. As the results of transductive learning and inductive learning have been very similar, we have only considered the results of transductive learning.

The classification accuracy decline rate when adding labeling noise on subsets of ImageNet is shown in Fig. 7. The settings of the line charts in Fig. 7 are identical with Figs. 3 and 5. Summarizing the law, the accuracy decline rate of RGSSL-MCC is smaller than that of SSC-SSR when there are 10% and 30% labeled samples on most of the subsets of ImageNet. Furthermore, on most of the subsets of ImageNet, the decline rate of RGSSL-MCC is much smaller than that of other SSL methods when adding 30% noisy labels. This, the proposed RGSSL-MCC method's robustness to noisy labels compared with the other SSL methods is relatively obvious when there are fewer labeled training samples and more noisy labels. This demonstrates that the proposed RGSSL-MCC method has some advantages in handling labeling noise over the other comparative SSL methods. The reason may be that introducing the maximum correntropy criterion in the graph-based SSL framework can focus on the data points in the local clustering centers instead of the noises, thus suppressing labeling noise effectively.

*4) Analysis of the Introduction of Supervised Information to the Regularizer and Maximum Correntropy Criterion:* To analyze the influences of the introduction of supervised information and maximum correntropy criterion, respectively, we have transformed the objective function in (18) to the following form, and take it as the comparison:

$$\max_{\theta} \sum_{i\in L} \exp\left(-\frac{\left\|\sum_{e\in L}\theta_e k(\mathbf{x}_i,\mathbf{x}_e)-y_i\right\|_2^2}{2\sigma^2}\right)+\beta\sum_{i\in N}\sum_{j\in N} \times \exp\left(-\frac{\left\|\sum_{e\in L}\theta_e k(\mathbf{x}_i,\mathbf{x}_e)-\sum_{e\in L}\theta_e k\left(\mathbf{x}_j,\mathbf{x}_e\right)\right\|_2^2}{2\sigma^2}\right)W_{ij}. \tag{35}$$

Equation (35) removes the supervised information of the regularizer from the objective function in (18), but retains the maximum correntropy criterion.

Besides, we takes the method described in (15) as the comparison. This method removes the maximum correntropy criterion from the objective function in (18), but retains the supervised information of the regularizer.

The experiments have been implemented in the subset2 of PASCAL VOC 2007. Some labeling noises have been added to the training set for robustness analysis. The results are shown in Fig. 8. We can observed that when there are no noise, the RGSSL-MCC performs slightly better than (15). This may be caused by that the maximum correntropy criterion could adjust the weight adaptively, resulting a more accurate result. The RGSSL-MCC and (15) performs better than (35). This shows the introduction of the supervised information to the regularizer can improve the performance. Besides, when there are labeling noises, the accuracy obtained by (15) decrease most greatly, while the RGSSL-MCC and (35) decrease slightly. The reason may be that the maximum correntropy criterion can focus on the data points which in the center of clustering instead of noisy points. The maximum correntropy criterion in the RGSSL-MCC and (35) makes the terms with error labels be in a tiny correntropy; however, the L2-norm of the difference between noisy labels and predicted value in (15) makes the errors amplify by square. These terms with noisy labels in the RGSSL-MCC and (35) have a smaller impact on the objective function than (15), and the influences from labeling noises are reduced as well. This shows the maximum correntropy criterion can repress the labeling noises.

*5) Discussion:* Taking all the experiments together, we can draw some conclusions. The proposed RGSSL-MCC method has performed significantly better in the classification task

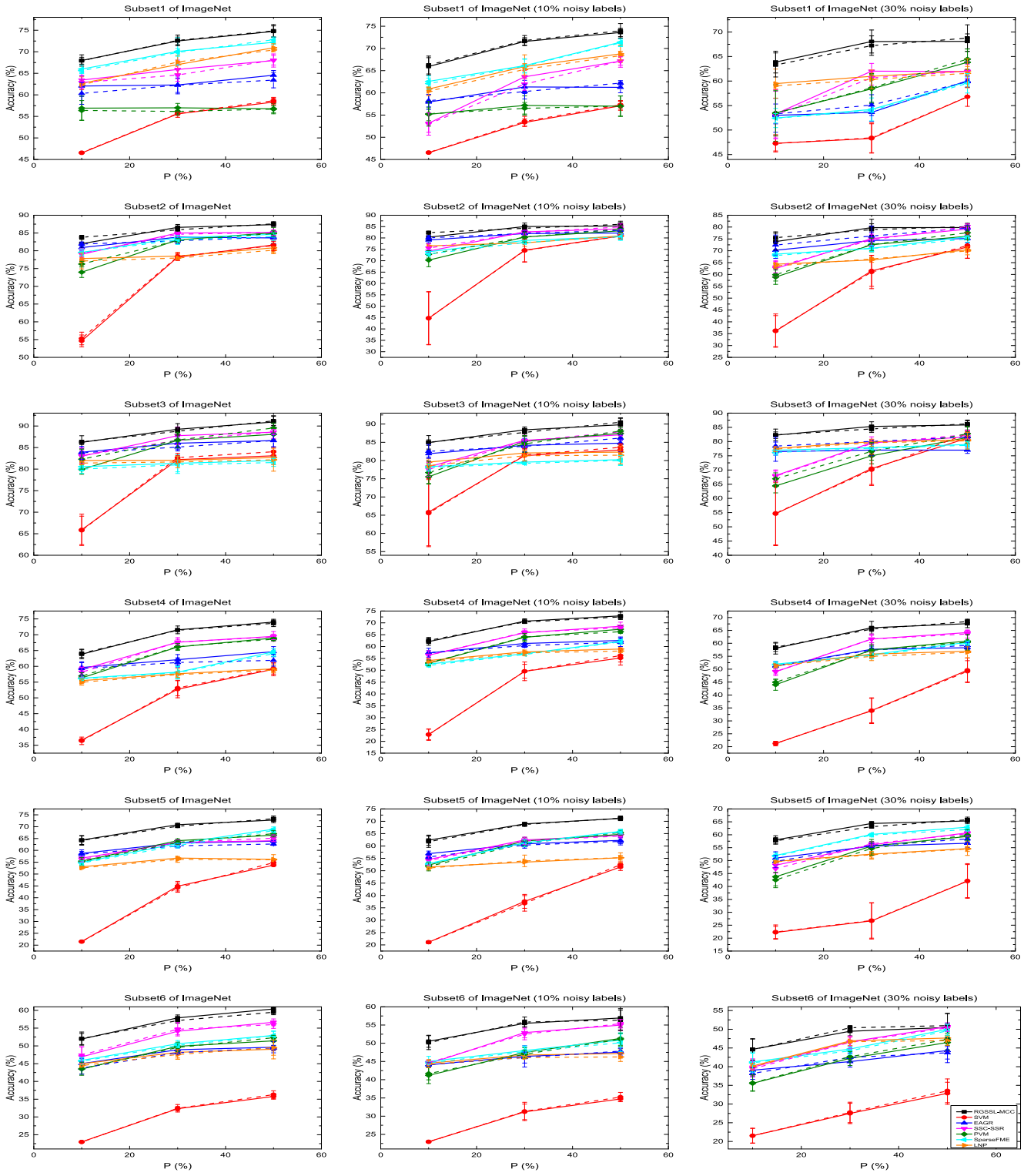



Fig. 6. Comparison of different SSL methods and classical supervised leaning method SVM on the subsets of ImageNet. The solid line indicates the transductive learning results, and the dashed line indicates the inductive learning results. The first column of this figure shows the results with no noise; the second column shows the results with 10% noisy initial labels; and the third column shows the results with 30% noisy initial labels. Each broken line represents the average result of five independent runs.

on all the datasets. When there are no noises, the superiority of RGSSL-MCC comes mainly from the introduction of the supervised information to the regularizer. This is because the third term in the objective function forces the unlabeled samples close to the labeled samples which are near to them in the feature space. When there are noisy labels, the accuracy losses of the RGSSL-MCC method are smaller than those of the other SSL methods, especially under the circumstance

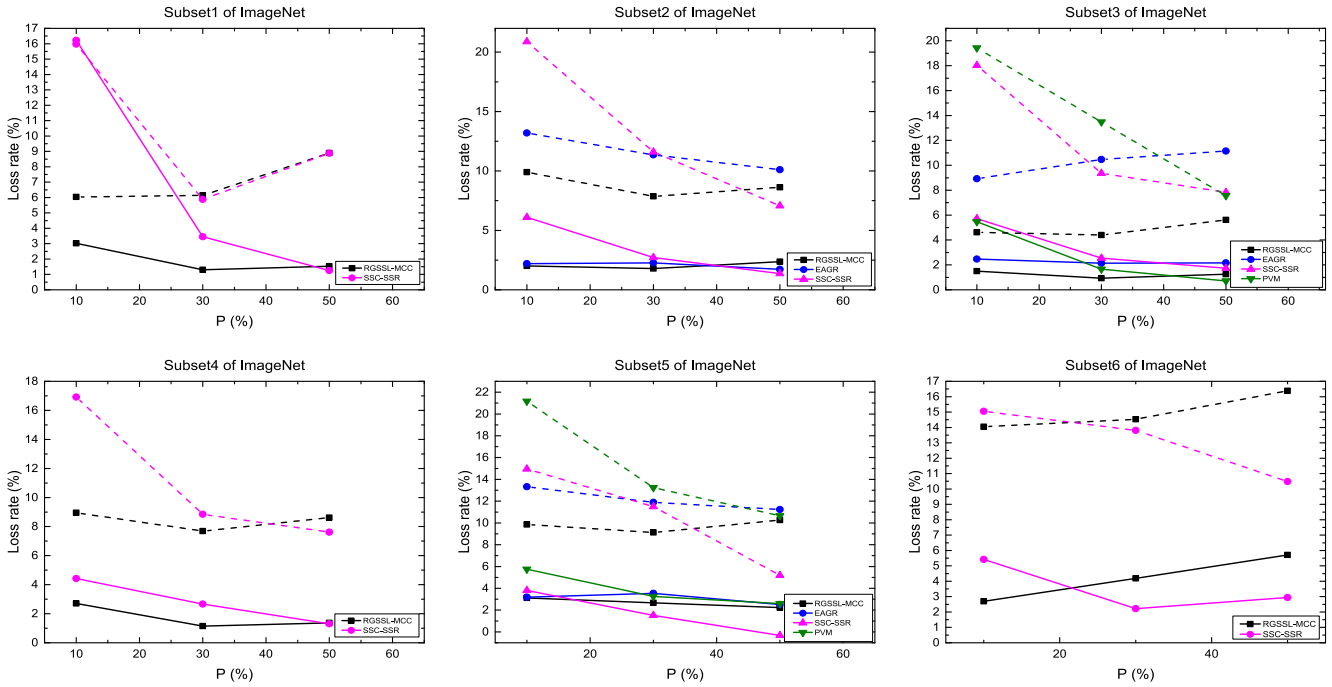



Fig. 7. Classification accuracy decline rate when adding labeling noise on subsets of ImageNet. The horizontal axis shows the percentage of labeled samples contained in training set, and the vertical axis shows the classification accuracy decline rate. The solid line indicates the decline rate when adding 10% noisy labels and the dotted line indicates the decline rate when adding 30% noisy labels.

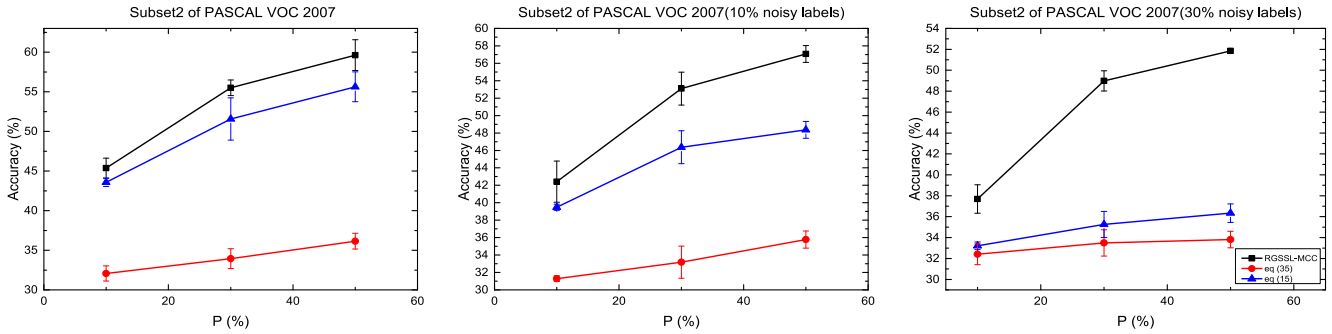



Fig. 8. Comparison of the proposed RGSSL-MCC methods with (15) and (35) on the subset2 of PASCAL VOC 2007. The first column of this figure shows the results with no noise; the second column shows the results with 10% noisy initial labels; and the third column shows the results with 30% noisy initial labels. Each line represents the average result of five independent runs.

of some stronger labeling noise but only a small number of labeled training samples. This is because when there are less samples and more labeling noise, the impact of the noisy labels on label propagation would be large in conventional SSL methods. However, the maximum correntropy criterion makes the correntropy between noisy labels and the predicted values become very small, thus reducing the impact of noisy label on the objective function.

All the SSL methods perform better than the supervised SVM. This shows the effectiveness of using the data distribution information. In addition, the accuracy of the RGSSL-MCC method in transductive learning and inductive learning is consistent. This proves that the inductive learning is effective.

## V. Conclusion

Noisy labels are very common in supervised learning scenarios and may cause a bias in the supervised information. This paper proposes a robust SSL method based on maximum correntropy criterion to solve this problem. The proposed graph-based SSL framework adds supervised information into the regularizer to reinforce the constraint on the labels, guaranteeing that the predicted labels of each cluster are close to the round truth. The introduction of the maximum correntropy criterion into the proposed graph-based SSL framework makes the model focus on the data points in the local cluster center and ignore the noise. A series of image classification experiments in situations of non-noise and label noise compared with some state-of-the-art SSL methods and the classical supervised learning method SVM has been carried out; the experimental results prove the generalization and robustness of the proposed SSL method. In the future, we aim to improve the proposed SSL method, especially for large-scale and high-dimensional data by using anchor graph [57] and feature selection and extraction [58].

## APPENDIX
## PROOF OF PROPOSITION 1

According to (19) and (21), we have

$$E(\theta^t, p^t, q^t, r^t) \leq E\left(\theta^t, p^{t+1}, q^{t+1}, r^{t+1}\right) \leq E\left(\theta^{t+1}, p^{t+1}, q^{t+1}, r^{t+1}\right).$$

The augmented objective function $E$ is increased at each alternating maximization step. Therefore, the augmented objective function sequence $E(\theta^t, p^t, q^t, r^t)$ is nondecreasing. According to the boundedness of correntropy [25], the augmented objective function $E(\theta^t, p^t, q^t, r^t)$ is bounded. Because the augmented objective function sequence $E(\theta^t, p^t, q^t, r^t)$ is nondecreasing and bounded, the sequence $E(\theta^t, p^t, q^t, r^t)$ is convergent. ■

## REFERENCES


[1] G. Camps-Valls, T. V. B. Marsheva, and D. Zhou, "Semi-supervised graph-based hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3044–3054, Oct. 2007.

[2] K. Lu, J. Zhao, and D. Cai, "An algorithm for semi-supervised learning in image retrieval," *Pattern Recognit.*, vol. 39, no. 4, pp. 717–720, 2005.

[3] X. You, Q. Peng, Y. Yuan, Y.-M. Cheung, and J. Lei, "Segmentation of retinal blood vessels using the radial projection and semi-supervised approach," *Pattern Recognit.*, vol. 44, nos. 10–11, pp. 2314–2324, 2011.

[4] T. Zhang, S. Liu, C. Xu, and H. Lu, "Boosted multi-class semi-supervised learning for human action recognition," *Pattern Recognit.*, vol. 44, nos. 10–11, pp. 2334–2342, 2011.

[5] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.

[6] X. Zhu, "Semi-supervised learning literature survey," *Comput. Sci.*, vol. 37, no. 1, pp. 63–77, 2008.

[7] S. Baluja, "Probabilistic modeling for face orientation discrimination: Learning from labeled and unlabeled data," in *Proc. Conf. Adv. Neural Inf. Process. Syst. II*, 1999, pp. 854–860.

[8] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling, "Semi-supervised learning with deep generative models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 4. Montreal, QC, Canada, 2014, pp. 3581–3589.

[9] G. Druck and A. McCallum, "High-performance semi-supervised learning using discriminatively constrained generative models," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 319–326.

[10] E. J. Gradi, E. Govekar, and I. Grabec, "Avaya: Sentiment analysis on twitter with self-training and polarity lexicon expansion," *J. Sound Vib.*, vol. 252, no. 3, pp. 563–572, 2013.

[11] I. Triguero, J. A. Sáez, J. Luengo, S. García, and F. Herrera, "On the characterization of noise filters for self-training semi-supervised in nearest neighbor classification," *Neurocomputing*, vol. 132, no. 13, pp. 30–41, 2014.

[12] P. Kang, D. Kim, and S. Cho, "Semi-supervised support vector regression based on self-training with label uncertainty: An application to virtual metrology in semiconductor manufacturing," *Expert Syst. Appl.*, vol. 51, pp. 85–106, Jun. 2016.

[13] A. Appice, P. Guccione, and D. Malerba, "A novel spectral-spatial co-training algorithm for the transductive classification of hyperspectral imagery data," *Pattern Recognit.*, vol. 63, pp. 229–245, Mar. 2017.

[14] D. Roqueiro *et al.*, "In silico phenotyping via co-training for improved phenotype prediction from genotype," *Bioinformatics*, vol. 31, no. 12, pp. i303–i310, 2015.

[15] J. Slivka, G. Sladic, B. Milosavljevic, and A. Kovacevic, "RSSalg software: A tool for flexible experimenting with co-training based semi-supervised algorithms," *Knowl.-Based Syst.*, vol. 121, pp. 4–6, Apr. 2017.

[16] F. Dornaika and Y. E. Traboulsi, "Matrix exponential based semi-supervised discriminant embedding for image classification," *Pattern Recognit.*, vol. 61, pp. 92–103, Jan. 2017.

[17] Z. Lu and L. Wang, "Noise-robust semi-supervised learning via fast sparse coding," *Pattern Recognit.*, vol. 48, no. 2, pp. 605–612, 2015.

[18] M. Wang, W. Fu, S. Hao, D. Tao, and X. Wu, "Scalable semi-supervised learning by efficient anchor graph regularization," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 7, pp. 1864–1877, Jul. 2016.

[19] K. Zhang, L. Lan, J. T. Kwok, S. Vucetic, and B. Parvin, "Scaling up graph-based semisupervised learning via prototype vector machines," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 3, pp. 444–457, Mar. 2015.

[20] G. Yu, G. Zhang, Z. Zhang, Z. Yu, and L. Deng, "Semi-supervised classification based on subspace sparse representation," *Knowl. Inf. Syst.*, vol. 43, no. 1, pp. 81–101, 2015.

[21] O. Chapelle and J. Weston, "Cluster kernels for semi-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 15, 2003, p. 15.

[22] M. Belkin and P. Niyogi, "Semi-supervised learning on Riemannian manifolds," *Mach. Learn.*, vol. 56, no. 1, pp. 209–239, 2004.

[23] X. Lu, Y. Wang, and Y. Yuan, "Graph-regularized low-rank representation for destriping of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 7, pp. 4009–4018, Jul. 2013.

[24] W. Liu, P. P. Pokharel, and J. C. Principe, "Correntropy: Properties and applications in non-Gaussian signal processing," *IEEE Trans. Signal Process.*, vol. 55, no. 11, pp. 5286–5298, Nov. 2007.

[25] M. Nikolova and M. K. Ng, "Analysis of half-quadratic minimization methods for signal and image recovery," *SIAM J. Sci. Comput.*, vol. 27, no. 3, pp. 937–966, 2005.

[26] A. Blum and S. Chawla, "Learning from labeled and unlabeled data using graph mincuts," in *Proc. 18th Int. Conf. Mach. Learn.*, 2001, pp. 19–26.

[27] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," in *Proc. 20th Int. Conf. Mach. Learn.*, 2003, pp. 912–919.

[28] Y. Yuan, H. Lv, and X. Lu, "Semi-supervised change detection method for multi-temporal hyperspectral images," *Neurocomputing*, vol. 148, pp. 363–375, Jan. 2015.

[29] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2003, pp. 321–328.

[30] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[31] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 849–856.

[32] Z. Zhang, M. Zhao, and T. W. S. Chow, "Graph based constrained semi-supervised learning framework via label propagation over adaptive neighborhood," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 9, pp. 2362–2376, Sep. 2015.

[33] F. Nie, S. Xiang, Y. Liu, and C. Zhang, "A general graph-based semi-supervised learning with novel class discovery," *Neural Comput. Appl.*, vol. 19, no. 4, pp. 549–555, 2010.

[34] F. Nie, D. Xu, I. W.-H. Tsang, and C. Zhang, "Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction," *IEEE Trans. Image Process.*, vol. 19, no. 7, pp. 1921–1932, Jul. 2010.

[35] Z. Zhang *et al.*, "Discriminative sparse flexible manifold embedding with novel graph for robust visual representation and label propagation," *Pattern Recognit.*, vol. 61, pp. 492–510, Jan. 2017.

[36] Z. Zhang *et al.*, "Robust adaptive embedded label propagation with weight learning for inductive classification," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2017.2727526.

[37] O. Zoidi, A. Tefas, N. Nikolaidis, and I. Pitas, "Positive and negative label propagation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 2, pp. 342–355, Feb. 2018.

[38] Z. Zhang *et al.*, "Semi-supervised image classification by nonnegative sparse neighborhood propagation," in *Proc. 5th ACM Int. Conf. Multimedia Retrieval*, vol. 4. Shanghai, China, 2015, pp. 139–146.

[39] R. He, W.-S. Zheng, B.-G. Hu, and X.-W. Kong, "Nonnegative sparse coding for discriminative semi-supervised learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 2849–2856.

[40] K.-H. Jeong, W. Liu, S. Han, E. Hasanbelliu, and J. C. Principe, "The correntropy MACE filter," *Pattern Recognit.*, vol. 42, no. 5, pp. 871–885, 2009.

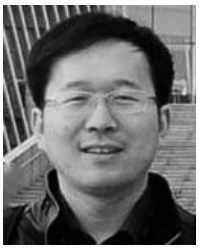
[41] X.-T. Yuan and B.-G. Hu, "Robust feature extraction via information theoretic learning," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 1193–1200.

[42] D. Erdogmus and J. C. Principe, *Information Theoretic Learning*. New York, NY, USA: Springer, 2010.

[43] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 1995, pp. 988–999.

[44] P. J. Huber, *Robust Statistics*. New York, NY, USA: Wiley, 2011.

[45] R. He, W.-S. Zheng, and B.-G. Hu, "Maximum correntropy criterion for robust face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1561–1576, Aug. 2011.
[46] N.-H. Yang, M.-M. Huang, R. He, and X.-K. Wang, "Robust semi-supervised learning algorithm based on maximum correntropy criterion," *J. Softw.*, vol. 23, no. 2, pp. 279–288, 2012.
[47] B. Du, Z. Wang, L. Zhang, L. Zhang, and D. Tao, "Robust and discriminative labeling for multi-label active learning based on maximum correntropy criterion," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1694–1707, Apr. 2017.
[48] R. T. Rockafellar, *Convex Analysis*. Princeton, NJ, USA: Princeton Univ. Press, 1970.
[49] M. Grant and S. Boyd. (Mar. 2014). *CVX: MATLAB Software for Disciplined Convex Programming, Version 2.1*. [Online]. Available: http://cvxr.com/cvx
[50] C. J. C. Burges, *A Tutorial on Support Vector Machines for Pattern Recognition*. Kluwer Acad., 1998.
[51] J. Han, G. Cheng, Z. Li, and D. Zhang, "A unified metric learning-based framework for co-saliency detection," *IEEE Trans. Circuits Syst. Video Technol.*, to be published, doi: 10.1109/TCSVT.2017.2706264.
[52] F. Wang and C. Zhang, "Label propagation through linear neighborhoods," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 1, pp. 55–67, Jan. 2008.
[53] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. (2007). *The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results*. [Online]. Available: http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html
[54] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 453–465, Mar. 2014.
[55] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
[56] X. Zhu and X. Wu, "Class noise vs. attribute noise: A quantitative study," *Artif. Intell. Rev.*, vol. 22, no. 3, pp. 177–210, 2004.
[57] Q. Zhang, J. Sun, G. Zhong, and J. Dong, "Random multi-graphs: A semi-supervised learning framework for classification of high dimensional data," *Image Vis. Comput.*, vol. 60, pp. 30–37, Apr. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885616301305
[58] L. Zhang *et al.*, "Simultaneous spectral-spatial feature selection and extraction for hyperspectral images," *IEEE Trans. Cybern.*, vol. 48, no. 1, pp. 16–28, Jan. 2018.



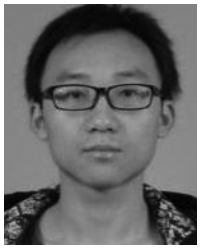
**Bo Du** (M'10–SM'15) received the B.S. and Ph.D. degrees in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, in 2005 and 2010, respectively.

He is currently a Professor with the School of Computer, Wuhan University. He has over 40 research papers published in the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE JOURNAL OF SELECTED TOPICS IN EARTH OBSERVATIONS AND APPLIED REMOTE SENSING, and the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS with 1796 Google Scholar citations and an H-index of 22. His current research interests include pattern recognition, hyperspectral image processing, and signal processing.

Dr. Du was a recipient of the Best Reviewer Awards from the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS for his service to the IEEE JOURNAL OF SELECTED TOPICS IN EARTH OBSERVATIONS AND APPLIED REMOTE SENSING in 2011, the ACM Rising Star Award for his academic progress in 2015, and the ESI Hot Papers or Highly Cited Papers for five of his papers. He was the Session Chair of both International Geoscience and Remote Sensing Symposium 2016 and the 4th IEEE GRSS Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing. He also serves as a Reviewer of 20 Science Citation Index magazines, including the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE JOURNAL OF SELECTED TOPICS IN EARTH OBSERVATIONS AND APPLIED REMOTE SENSING, and the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS.

**Xinyao Tang** received the B.E. degree from the Computer Science and Technology Department, Ocean University of China, Qingdao, China, in 2015. She is currently pursuing the master's degree with the School of Computer, Wuhan University, Wuhan, China.

Her current research interest includes pattern recognition.



**Zengmao Wang** received the B.S. degree in engineering of surveying and mapping from Central South University, Changsha, China, in 2013, and the M.S. degree from the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China, in 2015, where he is currently pursuing the Ph.D. degree with the School of Computer.

His current research interests include data mining and machine learning.

**Lefei Zhang** (S'11–M'14) received the B.S. and Ph.D. degrees from Wuhan University, Wuhan, China, in 2008 and 2013, respectively.

He is currently an Associate Professor with the School of Computer, Wuhan University. His current research interests include pattern recognition, image processing, and remote sensing.

Dr. Zhang is a Reviewer of over 30 international journals, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON IMAGE PROCESSING, and the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS.

**Dacheng Tao** (F'15) is a Professor of Computer Science and an ARC Laureate Fellow with the School of Information Technologies and the Faculty of Engineering and Information Technologies, and the Inaugural Director of the UBTECH Sydney Artificial Intelligence Centre, University of Sydney, Darlington, NSW, Australia. He mainly applies statistics and mathematics to artificial intelligence and data science. His research results have expounded in one monograph and over 500 publications at prestigious journals and prominent conferences, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON IMAGE PROCESSING, JMLR, IJCV, NIPS, ICML, CVPR, ICCV, ECCV, ICDM, and ACM SIGKDD. His current research interests include computer vision, data science, image processing, machine learning, and video surveillance.

Mr. Tao was a recipient of several best paper awards, such as the Best Theory/Algorithm Paper Runner Up Award in the IEEE ICDM'07, the Best Student Paper Award in the IEEE ICDM'13, the Distinguished Student Paper Award in IJCAI'17, the 2014 ICDM 10-Year Highest-Impact Paper Award, the 2017 IEEE Signal Processing Society Best Paper Award, the 2015 Australian Scopus-Eureka Prize, the 2015 ACS Gold Disruptor Award, and the 2015 UTS Vice-Chancellors Medal for Exceptional Research. He is a fellow of AAAS, OSA, IAPR, and SPIE.