Learning a Distance Metric by Balancing KL-Divergence for Imbalanced Datasets

Lin Feng, Huibing Wang, Bo Jin, Senior Member, IEEE, Haohao Li, Mingliang Xue, and Le Wang

Abstract—In many real-world domains, datasets with imbalanced class distributions occur frequently, which may confuse various machine learning tasks. Among all these tasks, learning classifiers from imbalanced datasets is an important topic. To perform this task well, it is crucial to train a distance metric which can accurately measure similarities between samples from imbalanced datasets. Unfortunately, existing distance metric methods, such as large margin nearest neighbor, information-theoretic metric learning, etc., care more about distances between samples and fail to take imbalanced class distributions into consideration. Traditional distance metrics have natural tendencies to favor the majority classes, which can more easily satisfy their objective function. Those important minority classes are always neglected during the construction process of distance metrics, which severely affects the decision system of most classifiers. Therefore, how to learn an appropriate distance metric which can deal with imbalanced datasets is of vital importance, but challenging. In order to solve this problem, this paper proposes a novel distance metric learning method named distance metric by balancing KL-divergence (DMBK). DMBK defines normalized divergences using KL-divergence to describe distinctions between different classes. Then it combines geometric mean with normalized divergences and separates samples from different classes simultaneously. This procedure separates all classes in a balanced way and avoids inaccurate similarities incurred by imbalanced class distributions. Various experiments on imbalanced datasets have verified the excellent performance of our novel method.

Index Terms—Distance metric by balancing KL-divergence (DMBK), distance metric learning (DML), geometric mean, imbalanced dataset.

I. INTRODUCTION

With the development of the information era, imbalanced datasets [1], [2] frequently occur in a variety of applications. Because negative samples heavily outnumber the positive ones, most machine learning methods [3] tend to favor the majority classes rather than the minority ones. Especially for those real-world applications which aim to detect a rare but important case, such as fraud detection [4], loan default prediction [5], and cancer detection, their performances drop significantly due to the bad influences caused by imbalanced class distributions. Take cancer detection as an example, patients with cancers are only a small part of all people in the real-world. Cancer detection aims to isolate patients with cancer, which is actually a classification task on imbalanced datasets. However, misclassifying a patient with cancer into healthy ones will have dire consequences to those patients, which will delay their treatments and endanger their life. By comparison, misclassifying a healthy person into a patient causes minor impact. Therefore, how to measure similarities between samples from imbalanced datasets is of vital importance.

During the last decade, distance metric learning (DML) [6]–[10] has received great attention in various data mining tasks due to its wide applications. Xing et al. [11] proposed a convex objective function for DML which aims to learn a global Mahalanobis metric. Two constraints are added to ensure a feasible solution. Large margin nearest neighbor (LMNN) [12], [13] takes label information into considerations and learns a distance metric with the goal that k-nearest neighbors always belong to the same class while samples from different classes are separated by a large margin. Neighborhood components analysis [14] constructs a nonparameter probabilistic optimization model which maximizes the probabilities of that samples and their stochastic neighbors are from the same class. Information-theoretic metric learning (ITML) [15] combines information theory with distance metric and converts the learning process to a Bregman optimization problem. It aims to find an optimal distance metric which can minimizes the differential relative entropy between two multivariate Gaussians. Maximally collapsing metric learning [16] constructs a convex optimization with the assumption that samples in the same class are simultaneously near each other and far from samples in the other classes. Both constraint-margin maximization (CMM) [17] and constrained metric learning (CML) [18] utilize large-margin criterion to learn distance metrics by considering pairwise constraints between samples. They aim to construct an optimal subspace.
in where the Euclidean metric can be employed to measure similarities. Similar to CMM and CML, traditional subspace embedding algorithms [19–24], such as principle component analysis (PCA) [19], linear discriminant analysis [20], and locality preserving projections [21], can also be regarded as DML methods.

However, the performances of most algorithms, including DML methods, will be severely affected due to the imbalanced class distributions. To address this issue, a number of approaches have been proposed for dealing with imbalanced datasets. All these approaches can be categorized into two main divisions: 1) sampling approaches [25–30] or 2) algorithmic approaches [31–37]. Sampling approaches include undersampling [26, 28] and oversampling [27]. Undersampling is a nonheuristic method which aims to balance class distributions by eliminating samples from the majority classes. Unfortunately, various potentially useful samples which may be important will be discarded. Oversampling aims to balance class distributions by replicating samples from the minority classes. However, due to the replications of the original samples, oversampling makes all algorithms overfitting easily. Because these sampling approaches above are always ineffective, some algorithmic approaches [35] are proposed to fill in this blank. Akbani et al. [31] proposed an algorithm which is based on a variant of the synthetic minority over-sampling technique [32] algorithm and applied support vector machines to imbalanced datasets successfully. Reference [36] is a novel ensemble approach based on switching with a new technique to select the switched examples based on nearest enemy distance, which can deal with imbalanced classification well. Even though there are various methods proposed to deal with imbalanced datasets from different views, few researchers focus on training an optimal distance metric just for imbalanced datasets. Because DML methods aim to better measure similarities between samples, they are not conflict with sampling methods or other methods to deal with imbalanced datasets. Meanwhile, accurate similarities between samples play a crucial part for many applications, such as classification, clustering, etc. And training a distance metric after sampling methods may achieve better performances in most situations.

Even though there are many DML methods proposed [30, 36–39], most DML methods fail to provide appropriate distance metrics to measure similarities between samples from imbalanced datasets. The main reason is that most DML methods care more about distances between samples while fail to take imbalanced class distributions into considerations. Take the DML method proposed by Xing et al. [11] as an example, it aims to maximize the objective function \(\sum_{(x_i, x_j) \in D} ||x_i - x_j||^2 \) which is the sum of distances between samples from different classes using distance metric matrix \(A\). However, faced with imbalanced datasets (as Fig. 1), this DML method tends to separate samples from majority classes (Class 1 and 2) which can maximize the objective function as much as possible. Samples from the minority classes (Class 3) can be neglected easily because they have less effects on the objective function than samples from the majority ones. However, distinctions between samples from Class 1 and 3 should be further emphasized because they may be misclassified easily, which may be neglected by most traditional DML methods. Therefore, for those real-world applications which aim to detect a rare but important case, such as fraud detection, loan default prediction, and cancer detection, the neglect of those minority classes will severely affect the performances of most DML methods.

In this paper, we propose a novel supervised DML method named distance metric by balancing KL-divergence (DMBK). KL-divergence can describe the difference between two probability distributions. Therefore, DMBK defines normalized divergence using KL-divergence to represent between-class divergence. Because DMBK should treat different classes equally and separate them balanced, DMBK combines geometric mean with normalized divergences to make all between-class divergences balanced. Then, two constraints are adopted by DMBK to achieve a feasible and excellent distance metric. In order to further treat all classes equally, Log function is utilized for geometric mean in DMBK. Our method is based on posing DML as an optimization problem, which is solved by gradient ascent and iterative projection.

This paper is organized as follows. In Section II, some related works are introduced. First, we introduce some basic knowledge of DML. Then, some typical DML methods are illustrated. In Section III, we illustrate the construction procedure of DMBK in detail. Meanwhile, the optimization procedure is explained. In Section IV, various experiments on imbalanced datasets are conducted to verify the excellent performance of DMBK. Section V concludes this paper.

II. RELATED WORKS

In this section, we introduce some related knowledge of DML. Meanwhile, four properties, which all distance metrics must satisfy, are listed. Then, we summarize the descriptions of important parameters which are utilized in this paper. Finally, we review two well-known DML methods which have aroused widespread attention.

A. Problem Definition

Assume we are given a set of \(N\) samples in a \(d\)-dimensional space \(X = \{x_1, x_2, \ldots, x_N\} \subseteq \mathbb{R}^d\), together with labels \(Y = \{y_1, y_2, \ldots, y_N\}\). Any \(y_i \in Y, 1 \leq i \leq N\) belongs to the label set \(C = \{1, 2, \ldots, c\}\). Table I shows the descriptions of some important parameters in this paper. For any \(x_i\) and \(x_j\), the Mahalanobis distance between these two samples can be calculated as follows:

\[
d_A(x_i, x_j) = \sqrt{(x_i - x_j)^T A (x_i - x_j)}
\]

where \(A \in \mathbb{R}^{d \times d}\) is the distance metric matrix. \(d_A(x_i, x_j)\) is the Mahalanobis distance between \(x_i\) and \(x_j\) using the metric matrix \(A\). The goal of DML is to develop a method to train a distance metric matrix \(A\) which can measure similarities between samples appropriately. For any distance metric matrix \(A\), these four properties must be satisfied as follows.

1) Non-Negativity: \(d_A(x_i, x_j) \geq 0\).
2) Symmetry: \(d_A(x_i, x_j) = d_A(x_j, x_i)\).
Some excellent DML methods have been proposed which can appropriately measure similarities between samples in various circumstances. DML methods can provide better tools to measure similarities between samples to improve the performances of most applications. And it is of vital importance for most classification systems. In this section, we introduce two well-known DML methods (ITML and LMNN) which have been widely utilized.

1) Information-Theoretic Metric Learning: ITML [15] is an information-theoretic approach which learns a Mahalanobis distance metric by minimizing the differential relative entropy between two multivariate Gaussians. Then, ITML formulates this problem as a particular Bregman optimization problem which minimizes the LogDet divergence subject to linear constraints. Constraints between pairs of distances are added to improve discriminative ability of ITML. And the objective function of ITML is formulated as follows:

$$
\min_{A \succeq 0} D_{\ell_d}(A, A_0) \\
\text{s.t. } \text{tr} \left( A (x_i - x_j)(x_i - x_j)^T \right) \leq u (i, j) \in S \\
\text{tr} \left( A (x_i - x_j)(x_i - x_j)^T \right) \geq l (i, j) \in D
$$

(2)

where tr(·) is the trace of a matrix and

$$
D_{\ell_d}(A, A_0) = \text{tr}(AA_0^{-1}) - \log \det \left( AA_0^{-1} \right) - n.
$$

(3)

It is well-known that $D_{\ell_d}(A, A_0)$ is the LogDet which is called Stein’s loss [40]. $A_0$ is an initial generalized Mahalanobis distance metric matrix. $S$ and $D$ are two sets which consist of samples from the same class or different classes, respectively. ITML aims to find an optimal metric matrix $A$ which can minimize $D_{\ell_d}(A, A_0)$ subject to two constraints. It can achieve excellent performance which generalizes well on unseen samples.

2) Large-Margin Nearest Neighbor: LMNN [12] trains a distance metric with the goal that the $k$-nearest neighbors always belong to the same class while samples from different classes are separated by a large margin. Based on the hinge loss, LMNN formulates a convex optimization using large margin criterion which is similar to SVMs [3] but requires

$$
\min_{A \succeq 0} D_{\ell_d}(A, A_0) \\
\text{s.t. } \text{tr} \left( A (x_i - x_j)(x_i - x_j)^T \right) \leq u (i, j) \in S \\
\text{tr} \left( A (x_i - x_j)(x_i - x_j)^T \right) \geq l (i, j) \in D
$$

(2)

where tr(·) is the trace of a matrix and

$$
D_{\ell_d}(A, A_0) = \text{tr}(AA_0^{-1}) - \log \det \left( AA_0^{-1} \right) - n.
$$

(3)

It is well-known that $D_{\ell_d}(A, A_0)$ is the LogDet which is called Stein’s loss [40]. $A_0$ is an initial generalized Mahalanobis distance metric matrix. $S$ and $D$ are two sets which consist of samples from the same class or different classes, respectively. ITML aims to find an optimal metric matrix $A$ which can minimize $D_{\ell_d}(A, A_0)$ subject to two constraints. It can achieve excellent performance which generalizes well on unseen samples.

B. Two Well-Known Works

During the last decades, DML has been a hot topic which attracts wide attentions from researchers all over the world. Fig. 1. Negative effect caused by imbalanced datasets. (a) Example of imbalanced datasets. (b) Tendencies of most algorithms.

![Fig. 1. Negative effect caused by imbalanced datasets. (a) Example of imbalanced datasets. (b) Tendencies of most algorithms.](image)

**TABLE I**

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_i$</td>
<td>the $i$-th sample</td>
</tr>
<tr>
<td>$X$</td>
<td>the set of all labeled samples</td>
</tr>
<tr>
<td>$c$</td>
<td>the number of classes</td>
</tr>
<tr>
<td>$N$</td>
<td>the number of all labeled samples</td>
</tr>
<tr>
<td>$Y$</td>
<td>the set of labels</td>
</tr>
<tr>
<td>$A$</td>
<td>the distance metric matrix</td>
</tr>
<tr>
<td>$d_A(x_i, x_j)$</td>
<td>the distance between $x_i$ and $x_j$ using $A$</td>
</tr>
<tr>
<td>$K(p_i</td>
<td></td>
</tr>
<tr>
<td>$E_{A_{ij}}^A$</td>
<td>normalized divergence between class $i$ and $j$</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>tradeoff between two terms of the objective function</td>
</tr>
<tr>
<td>$L_A$</td>
<td>the value of the objective function</td>
</tr>
<tr>
<td>$\Lambda$</td>
<td>diagonal matrix consists of eigenvalues of $A$</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>$i$-th eigenvalue of $A$</td>
</tr>
<tr>
<td>$W$</td>
<td>the projection matrix and $A = WW^T$</td>
</tr>
</tbody>
</table>

3) Triangular Inequality: $d_A(x_i, x_j) + d_A(x_j, x_k) \geq d_A(x_i, x_k)$.
4) Distinguishability: $d_A(x_i, x_j) = 0 \iff x_i = x_j$. It is obvious that the metric matrix $A$ is positive semi-definite. Therefore, $A$ can be decomposed as $A = WW^T (W \in \mathbb{R}^{d \times l}, l \leq d$). It is equivalent for a distance metric to find a rescaling of samples which replaces each sample $x$ with $W^T x$. Then, the standard Euclidean metric can be utilized to measure similarities between these rescaled samples. And we summarize the descriptions of important parameters which are utilized in this paper in Table I.

For imbalanced datasets, most distance metrics tend to favor the majority classes while neglecting those minority ones. Therefore, it is crucial to train an appropriate distance metric which can treat samples from different classes equally. Meanwhile, distances between samples from different classes should be maximized as large as possible. In this paper, a distance metric which can deal with imbalanced datasets should be surveyed.

![Table I: Descriptions of Important Parameters](image)
no modification or extension for problems in multiway classification. And the objective function of LMNN is described as follows:

$$
\min_A \sum_{ij} \eta_{ij} (x_i - x_j)^T A (x_i - x_j) + \lambda \sum_{ij} \eta_{ij} (1 - \theta_{ij}) \xi_{ij} \\
\text{s.t. } (x_i - x_j)^T A (x_i - x_j) - (x_i - x_j)^T A (x_i - x_j) \geq 1 - \xi_{ij} \\
\xi_{ij} \geq 0 \\
A \succeq 0
$$

(4)

where $\eta_{ij}$ is 1 indicates that $x_i$ and $x_j$ are target neighbors, 0 otherwise. If $\eta_{ij} = 1$, the objective function of (4) aims to minimize the sum of distances between two samples from the neighborhood. $\theta_{ij} \in \{0, 1\}$ indicates whether $x_i$ and $x_j$ are from the same class. If $\theta_{ij} = 1$, it aims to minimize the distance between samples from the same class. In (4), $\lambda$ is the regularized parameter which determines the tradeoff between two terms of LMNN. $\xi_{ij}$ is a slack variable which can prevent LMNN from overfitting.

III. DISTANCE METRIC BY BALANCING KL-DIVERGENCE

In order to train an appropriate distance metric for imbalanced datasets, we propose a novel DML method called DMBK in this section. First, we expound the motivation of this method in Section III-A. Then, the construction procedure of DMBK is described in Section III-B. Finally, we introduce the optimization procedure of DMBK in Section III-C.

A. Motivation

Even though DML has been a hot topic during the last years, most DML methods fail to provide appropriate distance metrics to measure similarities between samples from imbalanced datasets. These methods care more about distances between samples while failing to take imbalanced class distributions into consideration. In order to solve this problem, this paper aims to construct a novel distance metric which can fully take imbalanced class distributions into considerations.

B. Construction Procedure of DMBK

In this section, the construction procedure of DMBK is illustrated in detail. For most traditional DML methods that have natural tendencies to favor the majority classes, they care more about the distinctions between majority classes with different labels. Distances between samples from those majority classes will be further emphasized while minority classes may be neglected easily. Therefore, it is of vital importance for DMBK to treat samples from different classes (both majority classes and minority classes) equally and separate them balanced. First of all, DMBK assumes that different classes are sampled from Gaussian densities with different expected values but one identical covariance. This assumption ensures that KL-divergence between different classes can better measure their similarities, which is of vital importance for DMBK. Then, it defines the normalized divergences using KL-divergence to describe the distinctions between different classes. Because it is crucial for DMBK to separate samples from different classes balanced, DMBK combines geometric mean with normalized divergences. Meanwhile, Log function is utilized for geometric mean to further balance all normalized divergences. Large margin criterion is adopted by DMBK to increase its discriminative ability. Finally, two constraints are added to ensure that DMBK can achieve a feasible and optimal distance metric.

For any two Gaussian distributions $p_i = N(x; u_i, \Sigma_i)$ and $p_j = N(x; u_j, \Sigma_j)$, where $u_i$ and $u_j$ are the mean vectors of samples from the corresponding classes, and $\Sigma_i$ and $\Sigma_j$ are two within-class covariance matrices of the $i$th and $j$th classes, the KL-divergence [24], [41] between these two Gaussian distributions can be described as follows:

$$
K(p_i||p_j) = \int dx N(x; u_i, \Sigma_i) \ln \frac{N(x; u_i, \Sigma_i)}{N(x; u_j, \Sigma_j)} \\
= \frac{1}{2} \left[ \ln |\Sigma_i| - \ln |\Sigma_j| + \text{tr} \left( \Sigma_i^{-1} \Sigma_j \right) + \text{tr} \left( \Sigma_j^{-1} D_{ij} \right) \right]
$$

(5)

where $\ln(\cdot)$ calculates the Napierian logarithm and $|\Sigma| = \text{det}(\Sigma)$. $D_{ij} = (u_i - u_j)(u_i - u_j)^T$ is a symmetrical matrix. $K(p_i||p_j)$ is the KL-divergence which can reflect the similarity between class $i$ and $j$. Because training a distance metric is equivalent to finding a rescaling of samples which replaces each sample $x$ with $W^T x$, the KL-divergence between two rescaled Gaussian distributions $p(W^T x|y = i)$ and $p(W^T x|y = j)$ is described as follows:

$$
K_A(p_i||p_j) = K(p(W^T x|y = i)||p(W^T x|y = j)) \\
= \frac{1}{2} \left[ \ln |W^T \Sigma_i W| - \ln |W^T \Sigma_j W| + \text{tr} \left( (W^T \Sigma_j W)^{-1} (W^T (\Sigma_i + D_{ij}) W) \right) \right]
$$

(6)

$A = WW^T (W \in \mathbb{R}^{d \times l}, \ l \leq d)$ and $W$ is the projection matrix. $K_A(p_i||p_j)$ is the KL-divergence between class $i$ and $j$ using the distance metric matrix $A$. Then DMBK normalizes all KL-divergence to describe the distinctions between classes as follows:

$$
E_{ij}^{A} = \frac{q_i q_j K_A(p_i||p_j)}{\sum_{1 \leq m < n \leq c} q_m q_n K_A(p_m||p_n)}
$$

(7)

where $q_i$ is the number of samples in class $i$. $E_{ij}^{A}$ is the normalized divergence between class $i$ and $j$, which can reflect the distinctions between these two classes. As we discussed above, it is important for DMBK to treat samples from different classes equally, which can prevent the learned distance metric from tending to favor the majority classes. Therefore, for imbalanced datasets, DMBK combines geometric mean with normalized divergences to separate samples from different classes balanced as follows:

$$
A^* = \arg \max_A \left( \prod_{1 \leq i < j \leq l} \frac{q_i q_j K_A(p_i||p_j)}{\sum_{1 \leq m < n \leq c} q_m q_n K_A(p_m||p_n)} \right)^{\frac{1}{c(c-1)}}
$$

$$
= \arg \max_A \left( \prod_{1 \leq i < j \leq l} E_{ij}^{A} \right)^{\frac{1}{c(c-1)}}
$$

(8)

where $A^*$ is the optimization solution of the maximization problem in (8). Because of the arithmetic-geometric average
inequality, the geometric mean of all normalized divergences in (8) achieve its maximum value if and only if the normalized divergences are equal to each other. Therefore, (8) tends to emphasize the effects of minority classes while decreasing the effects of those majority ones. For imbalanced datasets, the optimal metric matrix \( A \) can take imbalanced class distributions into considerations. Meanwhile, samples from both majority and minority classes are separated balanced, which ensures that DMBK can avoid inaccurate similarities incurred by imbalanced class distributions. In order to further emphasize the effects of minority classes, Log function is utilized for geometric mean of all normalized divergences as follows:

\[
A^* = \arg \max_A \log \left( \prod_{1 \leq i < j \leq c} \frac{q_i q_j K_A(p_i \| p_j)}{\sum_{1 \leq m < n \leq c} q_m q_n K_A(p_m \| p_n)} \right) \frac{1}{(c-1)}
\]

\[
= \arg \max_A \log \left( \prod_{1 \leq i < j \leq c} E_{ij}^A \right) \frac{1}{(c-1)}.
\]  

(9)

Even though the optimal \( A \) in (9) can treat samples from different classes (both majority and minority classes) equally, DMBK should lengthen distances between samples from different classes to improve its discriminative ability. With that in mind, we utilize large margin criterion and formulate a regularized framework for DMBK as follows:

\[
\max_A L_A = \log \left( \prod_{1 \leq i < j \leq c} E_{ij}^A \right) \frac{1}{(c-1)} + \lambda \sum_{1 \leq i < j \leq N} \theta_{ij} d_A(x_i, x_j)
\]

(10)

where \( \lambda \) is the regularization parameter which can adjust the tradeoff between the geometric mean of normalized divergences and large margin criterion. \( \theta_{ij} \in \{0, 1\} \) reflects that whether \( x_i \) and \( x_j \) are from the same class. If \( x_i \) and \( x_j \) belong to the same class, \( \theta_{ij} = 0 \), 1 otherwise. We can clearly find that the first term of (10) aims to treat samples from different classes equally, which can help DMBK to deal with imbalanced datasets. Maximizing the first term tends to balance the normalized divergences, which emphasizes the effects of minority classes while decreasing the majority ones. Meanwhile, the second-term utilizes the large margin criterion and increases the discriminative ability of DMBK. However, (10) cannot ensure that the optimal \( A \) is positive semidefinite. Therefore, there should be a constraint \( A \succeq 0 \) to ensure that a feasible solution can be obtained by DMBK. At the same time, DMBK constraints that the sum of distances between samples from the same class should be less than a certain constant. Then, the objective function of DMBK is formulated as follows:

\[
\max_A L_A = \log \left( \prod_{1 \leq i < j \leq c} E_{ij}^A \right) \frac{1}{(c-1)} + \lambda \sum_{1 \leq i < j \leq N} \theta_{ij} d_A(x_i, x_j)
\]

s.t. \( g(A) = \sum_{1 \leq i < j \leq N} (1 - \theta_{ij}) d_A^2(x_i, x_j) \leq 1 \)

(11)

where \( g(A) \) is a constraint which restricts that the sum of distances between samples from the same class should be less than a certain constant. DMBK sets the constant as 1 because changing this constant to any other positive constant \( t \) results only in \( A \) being replaced by \( t^2 A \), which cannot have any substantial impact on the metric matrix \( A \). Even though the objective function has considered the imbalanced issue of samples from different classes, the constraint \( g(A) \leq 1 \) does not consider imbalanced issue between samples from the same class. Therefore, according to the numbers of samples from different classes, the constant to restrict the sum of distances between samples from the same class should be determined automatically. And this operation will be considered in our future work.

C. Optimization Procedure of DMBK

In this section, the optimization procedure of DMBK is introduced in detail. Because it is a tough task for the optimization algorithms to find the optimal \( A \) of DMBK directly, DMBK replaces the optimization variable \( A \) with \( W \) due to \( A = W W^T \). And the optimal metric matrix \( A \) can be calculated as \( W W^T \). Gradient ascent is adopted by DMBK to update \( W \), which can find the maximum value of the objective function. Meanwhile, in order to meet these two constraints in (11), DMBK utilizes iterative projection which adjusts the metric matrix \( A \) to ensure a feasible solution.

First, DMBK replaces \( A \) with \( W W^T \) to reconstruct the objective function. Then, DMBK takes a gradient step \( W = W + \alpha \nabla W L_A \) to update \( W \), which can maximize its objective function. \( \alpha \) is the step size and we set \( \alpha = 0.1 \) in our experiment. In order to calculate \( \nabla L_A \) easily, \( L_A \) can be transformed as follows:

\[
L_A = \frac{1}{c(c-1)} \sum_{1 \leq i < j \leq c} \log E_{ij}^A + \lambda \sum_{1 \leq i < j \leq N} \theta_{ij} d_A(x_i, x_j).
\]

(12)

We take \( A = W W^T \) into \( E_{ij}^A \) and \( d_A(x_i, x_j) \) and reorganize them as follows:

\[
E_{ij}^A = \frac{q_i q_j K_A(p_i \| p_j)}{\sum_{1 \leq m < n \leq c} q_m q_n K_A(p_m \| p_n)}
\]

\[
= \frac{q_i q_j K(p(W^T x | y = i) \| p(W^T x | y = j))}{\sum_{1 \leq m < n \leq c} q_m q_n K(p(W^T x | y = m) \| p(W^T x | y = n))}
\]

(13)

\[
d_A(x_i, x_j) = \sqrt{(x_i - x_j)^T W W^T (x_i - x_j)}.
\]

(14)

Then, \( \nabla W L_A \) can be expressed as follows:

\[
\nabla W L_A = \frac{1}{c(c-1)} \sum_{1 \leq i < j \leq c} \frac{\nabla W E_{ij}^A}{E_{ij}^A} + \lambda \sum_{1 \leq i < j \leq N} \theta_{ij} \nabla W d_A(x_i, x_j).
\]

(15)
Among (15), $\nabla_W E_{ij}^A$ and $\nabla_W d_A(x_i, x_j)$ can be calculated as follows:

$$\nabla_W E_{ij}^A = \frac{q_i q_j \nabla_W K_A(p_i \| p_j)}{\sum_{1 \leq m < n \leq c} q_m q_n K_A(p_m \| p_n)} q_i q_j K_A(p_i \| p_j) \left( \sum_{1 \leq m < n \leq c} q_m q_n \nabla_W K_A(p_m \| p_n) \right)$$

$$\nabla_W d_A(x_i, x_j) = \frac{(x_i - x_j) (x_i - x_j)^T W}{\sqrt{(x_i - x_j)^T W W^T (x_i - x_j)}}.$$

And in $\nabla_W E_{ij}^A$, the $\nabla_W K_A(p_i \| p_j)$ can be calculated as follows:

$$\nabla_W K_A(p_i \| p_j) = \left( \Sigma_j W(W^T \Sigma_j W)^{-1} - \Sigma_i W(W^T \Sigma_i W)^{-1} \right)$$

$$\left( + \left( \Sigma_i + D_{ij} \right) W(W^T \Sigma_i W)^{-1} - \Sigma_j W(W^T \Sigma_j W)^{-1} \right) W(W^T \Sigma_j W)^{-1}.$$

where $W(W^T \Sigma_i W)^{-1}$ may be rank-deficient, which cannot calculate the inverse matrix for it. Therefore, we always utilize generalized inverse matrix to obtain the approximate result. As we have discussed above, gradient ascent can be summarized as following.

$$A = I \in R^{d \times d}$$

**Optimization Procedure of DMBK**

1. **Iterate**
2. **Do**
3. $A = \left\{ \begin{array}{l} A' \mid \min_{A'} \| A' - A \|_F : A' \in S_1 \end{array} \right\}$
4. Decompose $A$ as $A = H^T \Lambda H$.
5. $A' = \text{diag}(\max \{ \gamma_1, 0 \}, \max \{ \gamma_2, 0 \}, \ldots, \max \{ \gamma_d, 0 \})$
6. Reconstruct $A$ as $H^T \Lambda' H$.
7. While $A$ converges
8. Decompose $A$ as $W W^T$ ($W \in R^{d \times l}$, $l \leq d$) using eigen decomposition.
9. Calculate $\nabla_W L_A$ using Eq.15 which is summarized in Table III.
10. Update $W$ as $W = W + \alpha \nabla_W L_A$.
11. Update $A$ as $W W^T$.
12. Until $A$ converges.

**Output**

The distance metric matrix $A$ of DMBK.

**TABLE III**

<table>
<thead>
<tr>
<th>Gradient Direction of $W$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A = I \in R^{d \times d}$ as a identity matrix.</td>
</tr>
<tr>
<td>The Optimization Procedure of DMBK:</td>
</tr>
<tr>
<td>1. <strong>Iterate</strong></td>
</tr>
<tr>
<td>2. <strong>Do</strong></td>
</tr>
<tr>
<td>3. $A = \left{ \begin{array}{l} A' \mid \min_{A'} | A' - A |_F : A' \in S_1 \end{array} \right}$</td>
</tr>
</tbody>
</table>
| 4. Decompose $A$ as $A = H^T \Lambda H$.
| 5. $A' = \text{diag}(\max \{ \gamma_1, 0 \}, \max \{ \gamma_2, 0 \}, \ldots, \max \{ \gamma_d, 0 \})$
| 6. Reconstruct $A$ as $H^T \Lambda' H$.
| 7. While $A$ converges
| 8. Decompose $A$ as $W W^T$ ($W \in R^{d \times l}$, $l \leq d$) using eigen decomposition.
| 9. Calculate $\nabla_W L_A$ using Eq.15 which is summarized in Table III.
| 10. Update $W$ as $W = W + \alpha \nabla_W L_A$.
| 11. Update $A$ as $W W^T$.
| 12. Until $A$ converges. |

**Output**

The distance metric matrix $A$ of DMBK.

**IV. Experiment**

In order to verify the performance of DMBK, we conducted several experiments on some benchmark datasets in this section. Six representative methods (including Xing et al. [11], CMM [17], ITML [15], LMNN [13], Euclidean, and Chebychev) are utilized as comparing distance metrics. Section IV-A describes all datasets which are utilized in our experiments. Meanwhile, the parameters of comparing metrics are introduced in detail. Then, in Section IV-B, we introduce several evaluation indices to describe the performances of classification tasks on imbalanced datasets. Finally, we show experiment results on these datasets in Section IV-C.
A. Datasets and Comparing Methods

There are eight datasets utilized in these experiments, including three UCI datasets (Wine, Glass, and ISOLET) and five image datasets (Yale, ORL, AR, Caltech 101, and Corel-10k). Table IV summarizes the attributes of these datasets in detail. Meanwhile, Fig. 2 shows some images from the image datasets.

For all datasets utilized in our experiment, we summarized the details of them in Table IV. Imbalanced ratio represents the ratio of the number of samples from majority class to the number of samples from the minority ones. For Wine, Glass, and Caltech 101, the numbers of samples from different classes are different with each other. Therefore, we listed the imbalanced ratios. For the other datasets, we will show how we select samples from them to make the training sets imbalanced.

In order to verify the performance of DMBK, six distance metrics are utilized as comparing methods. Among all these seven methods, there are some parameters need to be set for ITML, LMNN, and DMBK. For ITML, the initial given metric matrix is set as an identity matrix whose diagonal elements are 1, 0 others. In order to establish the lower and upper bounds which can improve its performance, we utilize (respectively) the 5th and 95th percentiles of the observed distribution of distances between pairs of samples within the dataset. For LMNN, the size of target neighbors is crucial for the performance and we set its number as 3. For DMBK, only the tradeoff of the objective function is need to be set and we set it as $\lambda = 0.05$. For the other four methods, they are nonparameter methods which have no parameter to be set.

B. Evaluation Indices

In this section, we introduce several evaluation indices which can describe the performances of classification tasks on imbalanced datasets. The results of classifications can be divided into four categories which we have summarized in Table V. And these four categories are known as true positive (TP), false negative (FN), false positive (FP), and true negative (TN). Meanwhile, we define the numbers of samples which belong to each categories as $N_{TP}$, $N_{FN}$, $N_{FP}$, and $N_{TN}$.

It is well-known that the overall accuracy (OA) is an important index which can reflect the performance of most classification tasks. And it is defined as follows:

$$OA = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{FP} + N_{FN} + N_{TN}}\quad (19)$$

where OA is the overall accuracy. Even though OA can reflect the performances of most classification tasks to some extent, it is not fair enough just to utilize this index for imbalanced dataset. The reason is that the number of negative samples in imbalanced dataset is much larger than that of the positive ones, which causes that the OA relies heavily on the classification accuracy on negative samples. Meanwhile, for some specific tasks, the cost of misclassifying a sample from minority classes is far greater than misclassifying a sample from majority ones. For these reasons, we calculate the two indices named Precision and Recall which are more suitable for classification tasks on imbalanced datasets, and these two indices are defined in [42] and [43]. Precision is a percentage which reflects the ratio between the truly classified positive samples and all predicted positive ones as (20). Recall is a percentage which reflects the ratio between the truly classified positive samples and all actual positive ones as (21). Both Precision and Recall can reflect the performances of all distance metrics on positive classes.

$$\text{Precision} = \frac{N_{TP}}{N_{TP} + N_{FP}}\quad (20)$$
$$\text{Recall} = \frac{N_{TP}}{N_{TP} + N_{FN}}.\quad (21)$$

Combined Precision and Recall into a single value, $F$-measure [42], [43] is another important index for imbalanced datasets. And $F$-measure is defined as

$$F\text{-measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}.\quad (22)$$

C. Classification Experiment on Imbalanced Datasets

In contrast with those traditional distance metrics, we conduct several experiments to verify the performance of DMBK in this section. In order to describe the imbalanced level for those imbalanced datasets, the imbalanced ratio is defined as the ratio of the number of samples from majority classes to the number of samples from minority ones. In this section, we train these seven distance metrics using training samples. For most DML methods, they aim to improve the performance of kNN classification and the construction framework is based on leaning a Mahalanobis distance metric, which is suitable for kNN classification. Therefore, 1NN is utilized to classify all the testing samples into different classes. Based on the classification result, we calculate four indexes (OA, Precision, Recall, 

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Size</th>
<th>Classes</th>
<th>Dimensions</th>
<th>Imbalance ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wine</td>
<td>178</td>
<td>3</td>
<td>13</td>
<td>1.06761</td>
</tr>
<tr>
<td>Glass</td>
<td>214</td>
<td>6</td>
<td>9</td>
<td>1.1184</td>
</tr>
<tr>
<td>ISOLET</td>
<td>7797</td>
<td>26</td>
<td>617</td>
<td>--</td>
</tr>
<tr>
<td>Yale</td>
<td>165</td>
<td>15</td>
<td>1024</td>
<td>--</td>
</tr>
<tr>
<td>ORL</td>
<td>400</td>
<td>40</td>
<td>1024</td>
<td>--</td>
</tr>
<tr>
<td>AR</td>
<td>1680</td>
<td>120</td>
<td>2000</td>
<td>--</td>
</tr>
<tr>
<td>Caltech 101</td>
<td>9144</td>
<td>102</td>
<td>--</td>
<td>1.0388</td>
</tr>
<tr>
<td>Corel-10k</td>
<td>10000</td>
<td>100</td>
<td>--</td>
<td>1:1</td>
</tr>
</tbody>
</table>

**Table IV: Attributes of the Datasets**

<table>
<thead>
<tr>
<th></th>
<th>Predicted positive</th>
<th>Predicted negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual positive</td>
<td>TP</td>
<td>FN</td>
</tr>
<tr>
<td>Actual negative</td>
<td>FP</td>
<td>TN</td>
</tr>
</tbody>
</table>

**Table V: Confusion Matrix**
Fig. 2. Some images from (a) ORL, (b) AR, (c) Yale, and (d) Caltech 101 image datasets used in our experiments.

Fig. 3. Prediction and Recall of seven distance metrics on Wine dataset.

and $F$-measure) to show the performance of these seven DML methods on imbalanced datasets.

First, we conduct classification experiments on Wine and Glass datasets. For Wine dataset, there are three classes which consist of 59, 71, and 48 samples. The imbalanced ratio of Wine dataset is 1:0.6761. The first and third classes are assigned as positive classes. For the Glass dataset, there are six classes which consist of 70, 76, 17, 13, 9, and 29 samples. And the imbalanced ratio of Glass dataset is 1:0.1184. The third to sixth classes are assigned as positive classes. We conducted fivefold cross-validation on these two datasets. And these seven distance metrics learning methods are utilized to train distance metrics. After 1NN classification using different distance metrics, the Precision and Recall are calculated and we show them in Figs. 3 and 4. All results are calculated ten times and we show the mean and max ones.

We can clearly find that DMBK outperforms the other six distance metrics on Wine and Glass datasets. Because DMBK can take imbalanced class distributions into consideration, it treats samples from different classes equally and separates them balanced. And it is obvious that DMBK is a better DML method to deal with imbalanced datasets.

There are five subsets contained in ISOLET in total and each subset contains 26 classes. For each class, it contains about 60 samples. Ten classes are randomly assigned as positive classes while the other classes are assigned as negative ones. We randomly select ten samples from positive classes and 20 samples from negative classes as training samples. Therefore, the imbalanced ratio of training samples is 0.5. And there are 420 training samples utilized for each subset. Among all these training samples, only 100 samples are positive ones and the other training samples are negative ones. Except all training samples, the rest of the samples are utilized as testing ones. Because most DML methods utilized in our experiments are time consuming, such as Xing, LMNN, ITML, etc. We utilized PCA to project highly dimensional datasets into a 30-D subspace. PCA preprocess is utilized to all comparing methods and our proposed one, which will not influence the experiment results. All seven distance metrics are utilized using training samples and 1NN classification is performed to classify all testing samples. Then, Precision and Recall are calculated according to the classification results and we summarize them in Fig. 5.

It is obvious that DMBK outperforms the other six distance metrics in most situations. Meanwhile, LMNN and ITML can also achieve good performances. Therefore, faced with imbalanced dataset, DMBK can fully take imbalanced class distributions into considerations.

Then, we conduct experiments on Yale, ORL, and AR face datasets to show the performances of these distance metrics. In these experiments, we randomly select several
classes as positive classes while other classes are selected as negative ones. For original Yale, ORL, and AR, numbers of samples from different classes are all the same with each other. Therefore, the imbalanced ratio equals 1:1. In our experiments, we randomly select 5, 10, and 30 classes as positive classes, respectively, while the other classes are assigned as negative ones. Samples from both positive classes and negative classes are randomly selected as training samples. In order to show the performances of these distances on imbalanced datasets, the number of negative samples is bigger than the number of positive ones. Because NNS (NPS) represents the number of samples from the negative (positive) class. Therefore, imbalanced ratios for training sets can be calculated as NNS/NPS. PCA is utilized to preprocess all samples and project them into a 30-D subspace. 1NN classifier is utilized after the implements of all distance metrics.

Four indices (OA, Precision, Recall, F-measure) are calculated in Tables VI–VIII. All results are calculated ten times and we show the mean values in these three tables. We can clearly find that DMBK can achieve excellent performances in most situations. Meanwhile, LMNN is another good distance metric compared with other distance metrics. Through results

| TABLE VI | CLASSIFICATION ON YALE DATASETS |
| --- | --- | --- | --- | --- | --- |
| Distance Metric | NPS | NNS | OA | Precision | Recall | F-measure |
| DMBK | 25 | 50 | 73.33 | 75.66 | 76.67 | 76.16 |
| Xing | 25 | 50 | 69.81 | 68.30 | 72.08 | 70.14 |
| CMM | 25 | 50 | 67.22 | 70.86 | 68.14 | 69.48 |
| ITML | 25 | 50 | 70.74 | 69.20 | 69.14 | 69.18 |
| LMNN | 25 | 50 | 75.56 | 72.52 | 73.58 | 73.05 |
| Euclidean | 25 | 50 | 73.80 | 85.04 | 69.67 | 76.59 |
| Chebychev | 25 | 50 | 58.52 | 64.97 | 61.67 | 63.28 |
|  | 25 | 50 | 57.40 | 75.10 | 51.33 | 60.98 |
|  | 25 | 50 | 53.33 | 56.23 | 55.42 | 55.82 |
|  | 25 | 50 | 52.80 | 70.77 | 45.00 | 55.02 |

1 NPS represents the number of positive samples
2 NNS represents the number of negative samples

| TABLE VII | CLASSIFICATION ON ORL DATASETS |
| --- | --- | --- | --- | --- | --- |
| Distance Metric | NPS | NNS | OA | Precision | Recall | F-measure |
| DMBK | 50 | 150 | 93.95 | 95.39 | 93.60 | 94.48 |
| Xing | 50 | 150 | 85.20 | 85.02 | 86.57 | 85.79 |
| CMM | 50 | 150 | 93.80 | 93.04 | 93.60 | 94.32 |
| ITML | 50 | 150 | 91.05 | 89.56 | 89.30 | 91.06 |
| LMNN | 50 | 150 | 91.55 | 92.65 | 90.40 | 91.51 |
| Euclidean | 50 | 150 | 85.65 | 84.29 | 83.80 | 85.03 |
| Chebychev | 50 | 150 | 76.15 | 76.46 | 73.40 | 74.90 |
|  | 50 | 150 | 80.51 | 85.78 | 75.43 | 80.27 |

1 NPS represents the number of positive samples
2 NNS represents the number of negative samples

| TABLE VIII | CLASSIFICATION ON AR DATASETS |
| --- | --- | --- | --- | --- | --- |
| Distance Metric | NPS | NNS | OA | Precision | Recall | F-measure |
| DMBK | 150 | 450 | 95.19 | 94.94 | 94.81 | 94.88 |
| Xing | 150 | 700 | 96.67 | 97.54 | 94.22 | 95.85 |
| CMM | 150 | 450 | 58.81 | 59.29 | 59.92 | 59.40 |
| ITML | 150 | 700 | 56.89 | 65.54 | 46.96 | 54.72 |
| LMNN | 150 | 450 | 94.80 | 94.16 | 94.44 | 94.30 |
| Euclidean | 150 | 700 | 96.00 | 97.89 | 93.04 | 95.40 |
| Chebychev | 150 | 450 | 91.13 | 90.58 | 90.52 | 90.55 |
|  | 150 | 700 | 92.17 | 94.90 | 86.96 | 90.76 |
|  | 150 | 450 | 95.15 | 94.52 | 94.30 | 94.41 |
|  | 150 | 700 | 96.20 | 97.71 | 93.70 | 95.66 |
|  | 150 | 450 | 73.60 | 74.41 | 73.52 | 73.96 |
|  | 150 | 700 | 78.12 | 83.64 | 69.33 | 75.82 |
|  | 150 | 450 | 58.91 | 58.84 | 58.81 | 59.32 |
|  | 150 | 700 | 64.64 | 73.93 | 54.22 | 62.56 |

1 NPS represents the number of positive samples
2 NNS represents the number of negative samples
from these three tables, recall decreases as the numbers of negative samples increase. However, it has no significant effect on the recall of DMBK. This shows that DMBK can deal with imbalanced datasets well.

Furthermore, we conduct a classification experiment on Caltech 101 image datasets. Because different objects in Caltech 101 have different numbers of images, Caltech 101 is an imbalanced dataset. And the imbalanced ratio of this dataset equals $\frac{31}{800} = 0.03875$. Locality-constrained linear coding [44] is utilized to extract features from all images to represent them. Then we perform PCA to preprocess all features into a 30-D subspace. In this experiment, we randomly select 2000 as training samples. The other samples are assigned as testing ones. All distance metric methods are employed to measure the distances between different samples. And 1NN is utilized to classify all testing samples. These experiments are conducted ten times and the mean results of OA, Precision, Recall, and F-measure are described as Table IX. It is clear that DMBK can achieve a better performance on the Caltech 101 dataset. LMNN and Euclidean are another two good metrics for this dataset.

In order to show the performance of DMBK on balanced datasets, we conducted another experiment on Corel-10k which contains 10,000 images from 100 categories. Similar with the experiments above, we randomly selected 5000 samples as training ones while the other 5000 ones are assigned as testing ones. We utilized AlexNet [45] to train the training samples and extracted the second fully connected layers as features. PCA is utilized to preprocess all features into a 30-D subspace. All distance metric methods are employed to measure the distances between different samples. And 3NN is utilized to classify all testing samples. These experiments are conducted ten times to calculate the mean results of OA, Precision, Recall, and F-measure, which has been summarized in Table X.

This section conducts several experiments on some benchmark datasets to verify the good performance of DMBK. We can clearly find that DMBK can achieve good performances in most cases.

### D. Training Time and Parameter Selection

In this section, we discuss the training time of our proposed method in detail. Therefore, we compared DMBK with the other four methods in our experiments. Because Xing, ITML, LMNN, and DMBK are all solved using iterative optimization procedures, we recorded their training time instantly once their iterations converged. Tables XI and XII summarize the training time on Caltech 101 dataset. The training time was tested on PC with a dual-Core i5-2300 CPU (2.80 GHz) and 6 GB memory. Table XI shows the training time with different numbers of training samples from Caltech 101 (the input dimension is 30). Table XII shows the training time with different numbers of input dimensions (2000 training samples are utilized) which are preprocessed by PCA.

Because CMM is optimized by eigen decomposition which does not need to be trained by iterative procedure, they consume less time to train their distance metric. From these two tables above, we can clearly find that DMBK consumes less time than Xing and LMNN while it can achieve excellent performance in most situations. Meanwhile, it can be found in Table XI that LMNN is more sensitive to the numbers of training samples. And Table XII shows DMBK is more sensitive to the numbers of input dimensions because the training times increase greatly with the rise of input dimensions. It is because that the most time-consuming procedures to optimize DMBK are to calculate the gradient as (16)–(18), which are mainly determined by the dimensions of matrices. And the dimensions of these matrices are directly influenced by input dimensions of samples. Therefore, it will consume less time facing with samples with lower dimensions.

Meanwhile, we show why we select $\lambda$ as 0.05 for DMBK on Yale face dataset. We randomly select 25 samples as positive samples while 90 samples as negative samples, which is corresponding the experiment above in Table VI. We show the OA accuracies with different selections of $\lambda$ in Table XIII. This experiment is conducted ten times and Table XIII shows the mean results.
It can be found that $\lambda = 0.05$ is a better selection for DMBK. Therefore, we set $\lambda$ as 0.05 in all experiments of this paper. Because the performance of DMBK varies with different selections of $\lambda$, it indicates that the regularization of DMBK is indeed helpful.

### V. Conclusion

DML is a traditional topic which has been carefully researched during the last two decades. DML can be utilized before $k$NN to improve its classification performance. Meanwhile, it can also be utilized to better measure similarities between samples, which can be adopted by many applications, such as object tracking and semantic segmentation. Therefore, researches on DML are valuable and need to be carried on continuously.

In this paper, we proposed a novel DML method named DMBK. In contrast with some traditional DML methods, DMBK is proposed to deal with imbalanced datasets, which can fully take imbalanced class distributions into considerations. First, DMBK utilizes KL-divergences to define normalized divergences which are combined with geometric mean to separate samples from different classes balanced. Then, large-margin criterion is adopted to improve DMBK’s discriminative ability. Two constraints are added for DMBK to ensure that a feasible and optimal distance metric can be obtained successfully.

It is obvious that, faced with imbalanced datasets, DMBK outperforms the other six distance metrics in most situations. The construction process of DMBK ensures that both majority and minority classes can be treated equally, which can avoid inaccuracy correlations incurred by imbalanced datasets. Various experiments show the excellent performance of DMBK. Meanwhile, imbalanced datasets will be employed in various fields, which needs more attention. DMBK provides a novel idea to measure distances between samples from imbalanced datasets. However, DMBK ignores the imbalanced problems in calculating the distances of samples from the same class, which influences the performances of DMBK to some extent. Therefore, according to the numbers of samples from different classes, the constant to restrict the sum of distances between samples from the same class should be determined automatically. And in our future work, we will study one method to solve this problem to further improve its performance.

### ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to significantly improve the quality of this paper.

### REFERENCES


Huibing Wang is currently pursuing the Ph.D. degree in mechanical design automation from Dalian University of Technology, Dalian, China, in 2001, and the Ph.D. degree in computer application technology from Dalian University of Technology, Dalian, China, in 2009.

Bo Jin (M’12–SM’16) received the B.S. degree in mechanical design automation from Dalian University, Dalian, China, in 2008 and 2011, respectively, and the Ph.D. degree in computer application technology from Dalian University of Technology, Dalian, China, in 2013.

He is currently an Associate Professor with the School of Innovation and Entrepreneurship, Dalian University of Technology. His current research interests include information retrieval, data mining, and intelligent computing.

Dr. Jin is a TPC Member in some top data mining conferences, such as SIGKDD Conference on Knowledge Discovery and Data Mining 2015, IEEE International Conference on Data Mining 2014–2015, and SIAM International Conference on Data Mining 2015. He is a Senior Member of ACM and CCF.

Mingliang Xue received the B.S. and M.S. degrees from the Department of Automation, Dalian University of Technology, Dalian, China, in 2008 and 2011, respectively, and the Ph.D. degree from the Department of Computing, Curtin University, Perth, WA, Australia, in 2015.

He is currently a Lecturer with the College of Computer Science and Engineering, Dalian Minzu University, Dalian. His current research interests include image processing and pattern recognition.

Le Wang received the Ph.D. degree in computer application technology from the School of Computer Science and Technology, Dalian University of Technology, Dalian, China, in 2013.

She is currently an Associate Professor with Ningbo Dahongying University, Ningbo, China. Her current research interests include image processing and data mining.