A Study on Distance Metric Learning using Distance Structure among Category Centroids

†Kenta Mikawa
Department of Information Science
Shonan Institute of Technology, Fujisawa, Japan
Tel: (+81) 466-30-0212, Email: mikawa@info.shonan-it.ac.jp

Manabu Kobayashi
Department of Information Science
Shonan Institute of Technology, Fujisawa, Japan
Tel: (+81) 466-30-0232, Email: kobayasi@info.shonan-it.ac.jp

Masayuki Goto
School of Creative Science and Engineering
Waseda University, Tokyo, Japan
Tel: (+81) 3-5286-3290, Email: masagoto@waseda.jp

Shigeichi Hirasawa
Research Institute for Science and Engineering
Waseda University, Tokyo, Japan
Tel: (+81) 3-5286-3290, Email: hira@waseda.jp

Abstract. The development in information technology has resulted in more diverse data characteristics and a larger data scale. Therefore, pattern recognition techniques have received significant interest in various fields including industrial management. In this study, we focus on a pattern recognition technique based on distance metric learning, which is known as the learning method in metric matrix under an arbitrary constraint from the training data. This method can acquire the distance structure, which takes account of the statistical characteristics of the training data. Most distance metric learning methods estimate the metric matrix from pairs of training data. However, the computational complexity of deriving the optimal metric matrix becomes high especially when the input data dimension becomes high. In this study, we focus on each category centroid to reduce the computational complexity of calculating the optimal metric matrix. To obtain an effective and robust result, we introduce the Alternating Direction Method for Multiplier (ADMM) and the regularization approach. To verify the effectiveness of our proposed method from the viewpoint of classification accuracy, simulation experiments using benchmark data sets are conducted.

Keywords: Distance metric learning, ADMM, Pattern recognition, Regularization

1. INTRODUCTION

The development in information technology highlighted the importance of knowledge discovery from enormous electronic data. Many techniques to obtain valid information have been proposed and widely used in business fields (Bishop, 2006). In this study, we focus on the vector space model, which is widely used in various fields. The basic methods of the vector space model are *k*-nearest neighbor, *k*-means, and template matching (Duda, 2000). However, the performance of these methods depends on the distance metric that is adopted. Generally, the Euclidean distance or cosine measure is often used because their computational cost is relatively small. These measures cannot consider the correlation between each element of input data; therefore, these distance measures sometimes do not improve the performance. On the other hand, the Mahalanobis distance can consider the correlation between each element of input data. However, in case of its adoption, it needs to estimate the suitable Mahalanobis matrix (metric matrix) that can express the correlation between each element of the training data.

The distance metric learning, which is the estimation of suitable distance metric under arbitrary constraints, is

one of the proposed methods in the field of machine learning (Yang et al., 2006). Generally, the distance metric learning adopts the Mahalanobis distance and can calculate the suitable metric matrix from input data information.

To obtain optimal metric matrix, most of the distance metric learning methods use iterative procedure. However, the computational complexity of these methods is relatively high especially when the dimension of input data becomes high. To reduce the computational complexity, the method of deriving optimal metric matrix using each category centroid with regularization is proposed (Mikawa et al., 2015). The computational complexity of this method is relatively small compared to other conventional methods because it does not use the iterative procedure. However, this method merely focuses on the relation between each category centroid and its training data, and it does not consider the distance structure among each category centroid.

From the above discussions, we propose the method to derive optimal metric matrix with consideration of the distance structure among each category centroid based on the method of Mikawa et al. The optimal metric matrix, which is derived from our proposed method, needs iterative procedure to obtain the optimal metric matrix. However, we show the effective way to derive the metric matrix using the alternating direction method of multiplier (Boyd et al., 20xx). To verify the effectiveness of our proposed method, simulation experiments are conducted.

2. PRELIMINARIES

2.1 Notation

Let the set of category C be $C = \{C_1, C_2 \cdots, C_N\}$, and W be the dimensional input vector, $\boldsymbol{x}_i \in R^W$ be $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{iW})^{\mathrm{T}}$ $(i = 1, 2, \cdots, D)$, where T is the transpose of a vector that belongs to category $C_n \in C$. The task of pattern recognition is to predict the correct category of a new input data $\boldsymbol{x}$, whose category is unknown, by using a classification rule that is acquired from the training data.

On the other hand, the distance measure of the vector space model, the Euclidean distance, and cosine measure are often used mainly from the viewpoint of computational cost. However, as mentioned previously, these measures cannot consider the statistical relationship between each element of input data. In contrast, the Mahalanobis distance is well-known and widely used as a distance metric that can consider the statistical relationship between each element of input data. The Mahalanobis distance can be defined using the metric matrix $M = [M_{ij}] \in R^{W \times W}$ whose elements can be estimated from the correlation between each element of input data. The Mahalanobis distance $d_M(\boldsymbol{x}_i, \boldsymbol{x}_j)$ between input data $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is denoted by

$$d_M(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T M (\boldsymbol{x}_i - \boldsymbol{x}_j)}. \quad (1)$$

In addition, we use the centroid of each category to obtain the optimal metric matrix in this study. Here, let the centroid of category C_n be $\boldsymbol{\mu}_n$, which is formulated by

$$\boldsymbol{\mu}_n = \frac{1}{|C_n|} \sum_{\boldsymbol{x}_i \in C_n} (x_{i1}, x_{i2}, \cdots, x_{iW})^{\mathrm{T}} . \quad (2)$$

Here, $|C_n|$ denotes the number of training data that belongs to category C_n.

2.2 Distance metric learning

Distance metric learning is the method of estimating an appropriate distance metric from the training data based on its information (Yang et al., 2006). Generally, the metric matrix M in equation (1) is calculated by solving the optimization problem under arbitrary constraint. To obtain the metric matrix, many methods have been proposed. In general, metric matrix is often used for pattern recognition tasks (*e.g.,* face recognition, clustering, classification, image retrieval among others).

One of the most well-known constraint widely used in the context of distance metric learning is the method of Xing et al. (2003). This method is called Mahalanobis Metric for Clustering (MMC), and defines the set of constraint (S or D) if the training data $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are similar (or dissimilar) in the following equations:

$$S\colon (\boldsymbol{x}_i, \boldsymbol{x}_j) \in S \quad \text{if} \quad \boldsymbol{x}_i and\ \boldsymbol{x}_j are\ similar,$$

$$D\colon (\boldsymbol{x}_i, \boldsymbol{x}_j) \in D \text{if} \quad \boldsymbol{x}_i and\ \boldsymbol{x}_j are\ dissimilar.$$

Xing et al. (2003) assumed that the information of similar or dissimilar has already been given beforehand and formulated the optimization problem that can estimate the optimal metric matrix $\widehat{M}$ as follows:

$$\widehat{M} = \mathrm{argmax}_M \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in D} d_M(\boldsymbol{x}_i, \boldsymbol{x}_j), \quad (3)$$

subject to

$$\sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in S} d_M^2(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq 1, \quad (4)$$

$$M \succeq 0.$$

Here, $\mathrm{M} \succeq 0$ denotes that the matrix M is positive semidefinite.

As mentioned previously, the MMC can derive the optimal metric matrix when the similarity or dissimilarity of the training data has already been given. On the other

hand, in this study, we focus on supervised learning, where the category information of input data has already been given. In summary, we can obtain more information of the input data in this setting. To utilize the category information and to obtain optimal metric matrix, we use the centroid of each category.

2.3 Alternating Direction Method of Multiplier

The ADMM (Boyed et al., 2011) is an algorithm that can derive the optimal solution to optimize each variable. Here, let each variable be $\boldsymbol{y} \in R^l$, $\boldsymbol{z} \in R^m$, $A \in R^{p\times l}$, $B \in R^{p\times m}$, and $c \in R^p$, then the optimization problem is formulated as follows:

$$\text{minimize } f(\boldsymbol{y}) + g(\boldsymbol{z}), \tag{5}$$

$$\text{subject to } A\boldsymbol{y} + B\boldsymbol{z} = \boldsymbol{c}. \tag{6}$$

Here, it is assumed that $f(\boldsymbol{y})$ and $g(\boldsymbol{z})$ are (not strictly) convex. To solve the above optimization problem, the augmented Lagrangian L_ρ is defined by

$$L_\rho(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{u}) = f(\boldsymbol{y}) + g(\boldsymbol{z}) + \frac{\rho}{2}\|\boldsymbol{r} + \boldsymbol{u}\|_2^2 \tag{7}$$

In equation (7), the expression of the augmented Lagrangian can be simplified by scaled form (Boyed et al., 20xx). Here, $\boldsymbol{r}$ is $= A\boldsymbol{y} + B\boldsymbol{z} - \boldsymbol{c}$, and ρ is a tuning parameter satisfying $\rho > 0$.

The ADMM can derive optimal solutions by iterating the following equations to update each variable.

$$\boldsymbol{y}^{k+1} \coloneqq \text{argmin}_{\boldsymbol{y}}\ L_\rho(\boldsymbol{y}, \boldsymbol{z}^k, \boldsymbol{u}^k) \tag{8}$$

$$\boldsymbol{z}^{k+1} \coloneqq \text{argmin}_{\boldsymbol{z}}\ L_\rho(\boldsymbol{y}^{k+1}, \boldsymbol{z}, \boldsymbol{u}^k) \tag{9}$$

$$\boldsymbol{u}^{k+1} \coloneqq \text{argmin}_{\boldsymbol{u}}\ L_\rho(\boldsymbol{y}^{k+1}, \boldsymbol{z}^{k+1}, \boldsymbol{u}) \tag{10}$$

The optimal solution can be derived by iterating these equations under certain condition (Boyd et al., 2011).

3. CONVENTIONAL METHOD

As mentioned previously, many algorithms that use iterative procedures to obtain the optimal metric learning have been proposed. On the other hand, (Mikawa et al., 2015) proposed the method of deriving the optimal metric matrix without any iterative procedure under the supervised framework with regularization. This method minimizes the sum of Mahalanobis distances between each category centroid $\boldsymbol{\mu}_n$ and training data belonging to same category in the following:

$$\widehat{M} = \text{minimize}_M \sum_{n=1}^{N} \text{tr}(S_n M) + \eta\text{tr}(M), \tag{11}$$

$$\text{subject to } \log\det M = 0, \quad M \succeq 0. \tag{12}$$

Here, $\text{tr}(\cdot)$ denotes the trace of matrix and $S_n \in R^{W\times W}$ is denoted by

$$S_n = \sum_{\boldsymbol{x}_i \in C_n} (\boldsymbol{x}_i - \boldsymbol{\mu}_n)(\boldsymbol{x}_i - \boldsymbol{\mu}_n)^{\mathrm{T}}. \tag{13}$$

η is a regularization parameter that takes a positive value.

The above optimization problem can be solved analytically using the method of Lagrange multiplier. The optimal solution $\widehat{M}$ can be derived as

$$\widehat{M} = \det(V)^{\frac{1}{W}} V^{-1}\ , \tag{14}$$

where $V \in \mathrm{R}^{W\times W}$ satisfies the following equation:

$$V = \sum_{n=1}^{N} S_n + \eta I\ , \tag{15}$$

where I is the $W \times W$ identical matrix.

As mentioned previously, this method enables to derive the optimal solution analytically; therefore, it does not need to use an iterative procedure. This method is superior to other conventional methods in terms of computational complexity because the optimal metric matrix can be derived analytically.

4. PROPOSED METHOD

Most of the distance metric learning methods use similar or dissimilar information between training data. They also use iterative procedure to derive the optimal metric matrix. Therefore, if the number of training data increases, the computational cost also increases because the number of constraints becomes large. On the other hand, Mikawa et al. (2015) shows the derivation of optimal metric matrix by using the centroid of each category. However, from equations (11) and (12), this method does not use the relationship between categories.

Therefore, this method does not make use of the category information to gain optimal metric matrix. Generally, if the data belongs to a different category, then the statistical characteristic of that is also different. Consequently, there is a possibility to improve the performance of classification to use the category information relationship.

On the contrary, most of the distance metric learning methods solved the optimization problem under the

constraint of similarity or dissimilarity of training data to derive the optimal metric matrix. As mentioned previously, increasing the number of training data results in the drastic increase in computational complexity because the number of constraints becomes large. Consequently, many methods of distance metric learning need high computational complexity.

In this study, we assume that the centroid of each category can express the characteristics of its category, and show the way to derive the optimal metric matrix by using the distance between each category centroid as the constraint. From this assumption, our proposed method can reduce the number of constraints while considering the statistical difference of each category. In addition, we introduce the regularization similar to the method of GSML (Huang et al., 2011) to achieve the robust parameter estimation.

Let l be an arbitrary positive constant and $L \in R^{W\times W}$ be an arbitrary regularization matrix. The optimization problem in this study can be formulated by

$$\widehat{M} = \text{minimize}_M \sum_{n=1}^{N} tr(S_n M) - \log\det M + \eta \text{tr}(LM),$$

$$\text{subject to } d_M(\boldsymbol{\mu}_n, \boldsymbol{\mu}_{n'}) \geq l, \quad \forall n' : n < n', \tag{16}$$

$$M \succeq 0.$$

To solve the above optimization problem, we adopt the ADMM (Boyd et al., 20xx). Here, let the objective function of equation (16) be $f(M)$, and $r_{nn'}(M), I(v)$ is denoted by

$$r_{nn'}(M) = I(d_M(\boldsymbol{\mu}_n, \boldsymbol{\mu}_{n'}) - l), \tag{17}$$

$$I(v) = \begin{cases} 0 & (v \geq 0), \\ \infty & (v < 0). \end{cases} \tag{18}$$

To apply the ADMM, the above optimization problem is transformed as follows:

$$\text{minimize}_M = f(Z) + \sum_{n:n<n'} r_{nn'}(M_{nn'}), \tag{19}$$

$$\text{subject to} \quad Z = M_{nn'}, \tag{20}$$

where $Z \in R^{W\times W}, M_{nn'} \in R^{W\times W}$, respectively. Moreover, the augmented Lagrangian of equation (19) can be defined by

$$L_\rho(M_{nn'}, Z, U_{nn'}) = f(Z) + \sum_{n:n<n'} r_{nn'}(M_{nn'}) + \frac{\rho}{2} \sum_{n:n<n'} \|M_{nn'} - Z + U_{nn'}\|_F^2,$$

where $\|A\|_F = \sqrt{\text{tr}(A^{\text{T}}A)}$ denotes the Frobenius norm.

The ADMM form of the above optimization problem can be formulated as follows:

$$M_{nn'}^{k+1} := \text{argmin}_{M_{nn'}} L_\rho(M_{nn'}, Z^k, U_{nn'}^k), \tag{22}$$

$$Z^{k+1} := \text{argmin}_Z L_\rho(M_{nn'}^{k+1}, Z, U_{nn'}^k), \tag{23}$$

$$U_{nn'}^{k+1} := M_{nn'}^{k+1} - Z^{k+1} + U_{nn'}^k, \tag{24}$$

where k is the number of iteration, and $U_{nn'}$ is a matrix whose element is a multiplier. To obtain the optimal solution, the following equations need to be solved.

$$M_{nn'}^{k+1} := \text{argmin}_{M_{nn'}} \sum_{n:n<n'} r_{nn'}(M_{nn'}) + \frac{\rho}{2} \sum_{n:n<n'} \left\|M_{nn'} - Z^k + U_{nn'}^k\right\|_F^2, \tag{25}$$

$$Z^{k+1} := \text{argmin}_Z \; f(Z) + \frac{\rho}{2} \sum_{n:n<n'} \|M_{nn'} - Z + U_{nn'}\|_F^2. \tag{26}$$

To solve these equations, each variable can be updated by

$$M_{nn'}^{k+1} := \begin{cases} Z^k - U_{nn'}^k & \alpha_{nn'} \geq l, \\ \Gamma_{nn'} & \alpha_{nn'} < l, \end{cases} \tag{27}$$

$$Z^{k+1} := Q^k \tilde{Z}^k Q^{k\text{T}}, \tag{28}$$

$$U_{nn'}^{k+1} := M_{nn'}^{k+1} - Z^{k+1} + U_{nn'}^k. \tag{29}$$

Here, $\Gamma_{nn'}, \alpha_{nn'}, \eta_{nn'}$ are in the following:

$$\Gamma_{nn'} = \frac{\eta_{nn'}}{\rho}(\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n'})(\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n'})^T + Z^k - U_{nn'}^k, \tag{30}$$

$$\alpha_{nn'} = (\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n'})(Z^k - U_{nn'}^k)(\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n'})^{\text{T}}, \tag{31}$$

$$\eta_{nn'} = -\frac{\rho}{\|\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n'}\|_2^4} \times \{l - (\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n'})^{\text{T}}(Z^k - U_{nn'}^k)(\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n'})\}, \tag{32}$$

In addition, $\tilde{Z}$ is a $\tilde{Z} = \text{diag}[\tilde{Z}_{11}, \tilde{Z}_{22}, \cdots, \tilde{Z}_{WW}]$, and $\tilde{Z}_{ii}$ satisfies

$$\tilde{Z}_{ii} = \frac{\lambda_i + \sqrt{(\lambda_i)^2 + 4\rho N}}{2\rho N}, \quad (33)$$

where λ_i is the ith eigenvalue of $-\sum_{n=1}^{N} S_n - \eta L + \rho N(\bar{M} + \bar{U})$. Here, $\bar{M}, \bar{U}$ is an average of $M_{nn'}, U_{nn'}$, respectively. Moreover, Q is the orthonormal matrix, which can be derived by the eigendecomposition of $-\sum_{n=1}^{N} S_n - \eta L + \rho N(\bar{M} + \bar{U})$.

On the other hand, when L equals to the identical matrix, the regularization term becomes $\mathrm{tr}(M)$, and in case of $L = M$, the regularization term becomes $\|M\|_F^2$. In addition, in case of $L = \sum_{n=1}^{N} S_n$, the regularization term equals to the sum of squared Mahalanobis distance between the category centroid and its training data.

5. SIMULATION EXPERIMENTS

5.1 Experimental Conditions

To verify the effectiveness of our proposed method, we conducted the classification experiment by using UCI machine learning repository (Asuncion, 2007). We used four data sets (iris, wine, balance, and ionosphere) and compared with the classification accuracy of other distance metric learning methods. We conducted the proposed method by using $L = I, M$, and $\sum_{n=1}^{N} S_n$, respectively. The parameters of η and ρ are decided in advance by the prior experiments. Classification accuracy is calculated as the average of ten observations. Basic information of each dataset is shown in Table 1.

Table 1: Basic information of datasets.

	Iris	Wine	Balance	Ionosphere
# of training data	135	161	563	316
# of test data	15	17	62	35
# of dimensions	3	13	4	34
# of categories	3	3	3	2

The classification accuracy is compared using the method of (Mikawa et al., 2015), Large Margin Nearest Neighbor (LMNN) (Weinberger, 2009), Information Theoretic Metric Learning (ITML) (Davis et al., 2007), and template matching by using Euclidean distance and cosine measure. To conduct LMNN and ITML, we use *k*-NN (Cover, 1969) $k = 3$.

5.2 Result of Experiments

The result of the experiments is shown in Table 2. The result shows that the proposed method is superior to the conventional methods without using the balance dataset. In addition, when the role of the regularization parameter changed, the classification accuracy is not drastically changed.

Table 2: Result of the experiments.

	Proposed ($L = I$)	Proposed ($L = M$)	Proposed ($L = \sum_{n=1}^{N} \mathrm{tr}(S_n)$)	Mikawa et al.	LMNN	ITML	Euclidean	Cosine
Iris	**98.00**	**98.00**	**98.00**	**98.00**	96.00	96.00	92.00	33.33
Wine	**99.44**	98.89	98.33	98.85	96.05	72.47	72.60	32.90
Balance	70.42	69.17	68.84	73.15	87.60	**88.96**	75.40	87.30
Ionosphere	**88.59**	87.47	87.46	88.33	87.20	86.89	71.00	75.50

6. DISCUSSIONS

In this study, we propose the method of deriving the optimal metric matrix by using each category centroid. Moreover, to change the arbitrary matrix L, the role of regularization changes. The result of the experiments shows that our proposed method is superior to the conventional methods in terms of classification accuracy. However, when using the balance dataset, the proposed method is not superior to the conventional methods. This is because our proposed method uses each category centroid when deriving the optimal metric matrix. However, because of the uneven distribution of balance dataset, the utilization of the centroid of each category is not suitable for improving the classification performance. In this case, it is suitable to use the conventional distance metric learning method with *k*-NN algorithms.

7. CONCLUSION AND FUTURE WORK

In this study, we proposed a method to estimate the optimal metric matrix using each category centroid. The result of the experiments showed that the proposed method is basically superior to the conventional method.

The proposed method uses iterative procedure; therefore, the computational cost is inferior to the method of Mikawa et al., especially when the number of dimensions becomes large, which will be addressed in our future work.

ACKNOWLEDGEMENT

This work was partly supported by Grant-in-Aid for Scientific Research for Young Scientists (C) 26750118.

APPENDICES

APPENDIX A. Derivation of equation (27)

The optimal solution can be obtained by solving the following equation.

$$M_{nn'}^{k+1} := \text{argmin}_{M_{nn'}} \sum_{n:n<n'} r_{nn'}(M_{nn'}) + \frac{\rho}{2} \sum_{n:n<n'} \left\| M_{nn'} - Z^k + U_{nn'}^k \right\|_F^2,$$

Here, to simplify the equation, let $A(M_{nn'})$ be

$$A(M_{nn'}) = \frac{\rho}{2} \sum_{n:n<n'} \left\| M_{nn'} - Z^k + U_{nn'}^k \right\|_F^2. \quad (34)$$

To obtain $M_{nn'}$ update, we consider the following two cases.

1. **In case of** $d_M(\boldsymbol{\mu}_n, \boldsymbol{\mu}_{n'}) \geq l$.

In this case, the first term of equation (25) becomes 0. Therefore, the optimal solution can be obtained to solve following function:

$$M_{nn'}^{k+1} := \text{argmin}_{M_{nn'}} A(M_{nn'}). \quad (35)$$

The gradient of equation (35) can be derived as follows:

$$\nabla_{M_{nn'}} A(M_{nn'}) = \rho(M_{nn'} - Z^k + U_{nn'}^k) \quad (36)$$

If it sets to zero, we have

$$M_{nn'} = Z^k - U_{nn'}^k. \quad (37)$$

To substitute equation (37) to equation (25), we have

$$I\left((\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n'})^T (Z^k - U_{nn'}^k)(\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n'}) - l\right). \quad (38)$$

From the above equations, if $(\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n'})^T (Z^k - U_{nn'}^k)(\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n'}) > l$ holds, $M_{nn'}^{k+1} = Z^k - U_{nn'}^k$ holds.

2. **In case of** $d_M(\boldsymbol{\mu}_n, \boldsymbol{\mu}_{n'}) < l$.

The optimal solution of $M_{nn'}$ can be derived by solving following optimization problem:

$$\text{minimize}_{M_{nn'}} A(M_{nn'}), \quad (39)$$

$$\text{subject to } d_M(\boldsymbol{\mu}_n, \boldsymbol{\mu}_{n'}) < l. \quad (40)$$

Here, letting the Lagrangian of above optimization problem be $B(M_{nn'}, \lambda)$, $B(M_{nn'}, \lambda)$ becomes

$$B(M_{nn'}, \lambda) = \frac{\rho}{2} \left\| M_{nn'} - Z^k + U_{nn'}^k \right\|_F^2 + \lambda \left\{ d_{M_{nn'}}(\boldsymbol{\mu}_n, \boldsymbol{\mu}_{n'}) < l \right\}. \quad (41)$$

The gradient of equation (41) becomes

$$\nabla_{M_{nn'}} B(M_{nn'}, \lambda) = \rho\left(M_{nn'} - Z^k + U_{nn'}^k\right) + \lambda(\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n'})(\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n'})^{\mathrm{T}}, \quad (42)$$

$$\frac{\partial B(M_{nn'}, \lambda)}{\partial \lambda} = (\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n'})^T (M_{nn'})(\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n'}) = l. \quad (43)$$

Therefore, equation (43) becomes

$$M_{nn'} = -\frac{\lambda}{\rho}(\boldsymbol{\mu}_n, \boldsymbol{\mu}_{n'})(\boldsymbol{\mu}_n, \boldsymbol{\mu}_{n'})^{\mathrm{T}} + Z^k + U_{nn'}^k. \quad (44)$$

To substitute equation (44) into equation (43), we have

$$(\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n'})^{\mathrm{T}} \left\{ -\frac{\lambda}{\rho}(\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n'})(\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n'})^{\mathrm{T}} + Z^k + U_{nn'}^k \right\} (\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n'}) = l. \quad (45)$$

Therefore, we have

$$(\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n'})^{\mathrm{T}} \left\{ -\frac{\lambda}{\rho}(\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n'})(\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n'})^{\mathrm{T}} \right\} \times (\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n'}) + (\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n'})^{\mathrm{T}} \left\{ Z^k + U_{nn'}^k \right\} (\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n'}) = l.$$

Consequently,

$$-\frac{\lambda}{\rho} \|\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n'}\|_2^4 = l - (\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n'})^{\mathrm{T}} \left\{ Z^k + U_{nn'}^k \right\} (\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n'}), \quad (47)$$

$$\lambda = -\frac{\rho}{\|\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n'}\|_2^4} \times \left\{ l - (\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n'})^{\mathrm{T}} (Z^k + U_{nn'}^k)(\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n'}) \right\},$$

is derived. From the above equations, $M_{nn'}$ update can be carried out as follows:

$M_{nn'}$

$$:= \begin{cases} Z^k - U^k_{nn'} & (d_M(\boldsymbol{\mu}_n\ , \boldsymbol{\mu}_{n'}) \geq l), \\ -\frac{\lambda}{\rho}(\boldsymbol{\mu}_n\ , \boldsymbol{\mu}_{n'})(\boldsymbol{\mu}_n\ , \boldsymbol{\mu}_{n'})^{\mathrm{T}} + Z^k + U^k_{nn'} & \\ & (d_M(\boldsymbol{\mu}_n\ , \boldsymbol{\mu}_{n'}) < l). \end{cases} \quad (48)$$

APPENDIX B. Derivation of equation (28)

The optimal solution can be obtained by solving the following equation.

$$Z^{k+1} \coloneqq \mathrm{argmin}_Z\ f(Z) + \frac{\rho}{2}\sum_{n:n<n'} \|M_{nn'} - Z + U_{nn'}\|_F^2. \quad (49)$$

The gradient of above equation becomes

$$\sum_{n=1}^{N} S_n - Z^{-1} + \eta L + \rho \sum_{n:n<n'} M_{nn'} - Z + U_{nn'}. \quad (50)$$

Here, let the average of $M_{nn'}, U_{nn'}$ be $\bar{M}, \bar{U}$ respectively (*i.e.*, $\bar{M} = 1/N \sum_{n:n<n'} M_{nn'}$, $\bar{U} = 1/N \sum_{n:n<n'} U_{nn'}$). To transform equation (50), it becomes

$$\sum_{n=1}^{N} S_n - Z^{-1} + \eta L + \frac{\rho}{N}(\bar{M} + \bar{U}) + \rho N \cdot Z. \quad (51)$$

Therefore, we have

$$-Z^{-1} + \rho N Z = -\sum_{n=1}^{N} S_n - \eta L + \rho N(\bar{M} + \bar{U}). \quad (52)$$

The right hand side of equation (52) is a symmetric matrix, then it can be decomposed as $Q\Lambda Q^{\mathrm{T}}$, where Q is a orthonormal matrix, *i.e.* $QQ^{\mathrm{T}} = I$. In addition, Λ satisfies $\Lambda = \mathrm{diag}[\lambda_1, \lambda_2\ , \cdots, \lambda_W\ ]$ and λ_w denotes w the eigenvalue of the right hand side of equation (52). Therefore, equation (52) can be transformed

$$-Z^{-1} + \rho N Z = Q\Lambda Q^{\mathrm{T}}. \quad (53)$$

Moreover, let $\tilde{Z}$ be $\tilde{Z} = QZQ^{\mathrm{T}}$, then equation (53) becomes

$$-\tilde{Z}^{-1} + \rho N \tilde{Z} = \Lambda. \quad (54)$$

Here, $\tilde{Z} = \mathrm{diag}[\widetilde{M}_{11}, \widetilde{M}_{22}, \cdots, \widetilde{M}_{WW}\ ]$ holds; therefore,

$$\rho N \tilde{Z}_{ii} - \frac{1}{\tilde{Z}_{ii}} = \lambda_{ii}, \quad (55)$$

is derived. Accordingly, we have

$$\tilde{Z}_{ii} = \lambda_i + \frac{\sqrt{\lambda_i^2 + 4\rho N}}{2\rho N}\ . \quad (56)$$

To multiply Q to equation (56), the Z update can be carried out.

REFERENCES

C. M. Bishop. (2006) *Pattern Recognition and Machine Learning*. Springer–Verlag.

R. O. Duda, P. E. Heart, D. G. Stork. (2000) *Pattern Classification*. Wiley-Interscience.

L. Yang and R. Jin. (2006) Distance Metric Learning: A Comprehensive Survey. *Department of Computer Science and Engineering, Michigan State University, Technical Report*.

K. Mikawa and M. Goto. (2015) Regularized Distance Metric Learning for Document Classification and Its Application. *J. Japan Industrial and Management Association*, Vol. 66, No. 2E, 1–14.

E. P. Xing, A. Y. Ng, M. I Jordan, and S. J. Russell. (2003) Distance Metric Learning, with Application to Clustering with Side-Information. *Advances in Neural Inform. Processing Syst. 15*, 521–528.

S. Boyd N. Parikh, E. Chu B. Peleato and J. Eckstein. (2011) Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multiplier, *Foundation and Trends in Mach. Learn.,* Vol. 3, No. 1, 1–122.

K. Huang, Y. Ying and C. Campbell. (2011) Generalized Sparse Metric Learning with Relative Comparisons. *Knowledge and Information Systems*, Vol. 28, No. 1, 25–45.

A. Asuncion, D. J. Newman. (2007) UCI Machine Learning Repository. *http://www.ics.uci.edu/~mlearn/MLRepository.html*

J. V. Davis, B. kulis, P. Jain, S. Sra and I. S. Dhillon. (2007) Information-Theoritic Metric Learning. *Proc. the 24th Int. Conf. on Mach. Learn*, 209–216.

K. Q. Weinberger and L. K. Saul. (2009) Distance Metric Learning for Large Margin Nearest Neighbor Classification. *J. Mach. Learn. Res.*, Vol. 10, 207–244.

T. Cover and P. Hart. (1967) Nearest Neighbor Pattern Classification. *IEEE Trans. Inf. Theory,* Vol. 13, No. 1, 21–27.