Improved Metric-Learning in Artificial Intelligence

Fan Gao

Abstract As artificial intelligence develops, it has become more and more concerned to measure the distance and similarity between different samples, especially in classification. It is known that AI meets the difficulty to properly describe the distance between different samples. Metric-learning is one field of machine learning performed on training samples in order to better describe the distribution of training samples, thereby improving the performance of classification. This paper proposes a new triplet constraint and uses the topology information combined with the unlabeled samples to design the regularizer of the model to ensure that the model meets the three basic assumptions of semi-supervised learning: smoothness, low density and manifold. This paper provides a new semi-supervised metric-learning model by using the new triplet constraint and regular terms. Then, using a gradient descent algorithm to solve the model, the optimal Mahalanobis matrix is obtained, which could ensure the Mahalanobis matrix is positive definite and symmetry during the iterative process. The Mahalanobis matrix solved by the model can be applied in clustering, classification and many other artificial intelligence fields. In this paper, the model is used for classification problem as an application. The experimental results are consistent with the idea of the model, reflecting the advantages of this model compared to other models, providing a new method to measure the sample distance in artificial intelligence and machine learning.

Keywords Semi-supervised metric-learning · Artificial intelligence · Machine learning · Mahalanobis distance · Triplet constraint

1 Introduction

High-dimensional data is inconvenient to process, so dimensionality reduction is aimed for a suitable lower-dimensional space where computation should be reduced while retaining sample information. The motivation for metric-learning is to find

F. Gao (✉)
Shanghai University, Jiading, China
e-mail: Gavin_FanGao@163.com

C.-T. Yang et al. (eds.), *Innovative Computing*,
Lecture Notes in Electrical Engineering 675,
https://doi.org/10.1007/978-981-15-5959-4_55

a suitable distance function, which has been applied extensively in fields as text classification and image-recognition. Traditionally, the Euclidean distance is used, however, Euclidean distance is evenly distributed where features' proportion and relationship are not considered, which affects the ability of characterizing sample structures.

Xing first proposed to use Mahalanobis distance [1] and solved it with a simple gradient algorithm. Schultz and Joachims proposed a model for determining the parameters of the metric matrix and the learning weight diagonal matrix [2]. Later, Weinberger and Saul designed the LMNN model, combining the idea of neighbor prediction [3]. The principal component analysis (PCA) is typical in dimensionality reduction algorithms [4], which can be regarded as a special metric-learning algorithm. Another classic dimension reduction method is the multi-dimensional dimensioning algorithm MDS [5], which converts a distance into the form of the inner product of the bit matrix. Roweis and Sau proposed local linear embedding by finding low-dimensional manifolds that maintain high-dimensional spatial neighbor structures [6]; Tenenbaum proposed ISOMAP algorithm by replacing Euclidean distances with local geodesic distances. [7]; Belkin and Niyogi proposed the Laplacian characteristic mapping [8]; Donoho and Grimes proposed the Hesse feature mapping algorithm [9].

In this paper, a new triplet constraint is designed, considering the previous research of the semi-supervised metric-learning model. The new triplet constraint combines the advantages of two types of previous triplet constraints, so that the learned matrix can measure the distance more effectively. In order to make the model meet the three premise assumptions of semi-supervised learning, a regular term is designed in this paper using unlabeled samples.

2 Semi-supervised Metric-Learning

2.1 Mahalanobis Distance

Euclidean metric is defined as

$$d(x_i, x_j) = \sqrt{(x_i - x_j)^T (x_i - x_j)} \tag{1}$$

In Euclidean metric, the proportion of different features of the sample is evenly distributed, so the coupling relationship between different features is not considered. Mahalanobis distance is used to improve this limitation and defined as follow.

$$d(x_i, x_j) = \sqrt{(x_i - x_j)^T A (x_i - x_j)} \tag{2}$$

where $A \in R^{n\times n}$ is the metric matrix. As a semi-symmetric positive definite matrix, A can be decomposed into $A = Q^T Q$, and Eq. (2) is rewritten as follow.

$$d_A(x_i, x_j) = \sqrt{(x_i - x_j)^T A(x_i - x_j)} = \sqrt{(x_i - x_j)^T Q^T Q(x_i - x_j)} \quad (3)$$

It is naturally required that the distance between similar samples is as small as possible, and the distance between heterogeneous samples is as large as possible.

If $S = \{(x_i, x_j): x_i, x_j \in R^n, \text{and of the same class}\}$ and $D = \{(x_i, x_j): x_i, x_j \in R^n, \text{and of different class}\}$, then a simple metric-learning model is defined

$$\min_A \sum_{(x_i,x_j)\in S} d^2(x_i, x_j), \quad \text{s.t.} \sum_{(x_i,x_j)\in D} d^2(x_i, x_j) \geq C \quad (4)$$

where $d(x_i, x_j) = \sqrt{(x_i - x_j)^T A(x_i - x_j)}$, A is a semi-positive symmetric matrix and C is constant. Many of the later models were derived from this basic model.

2.2 Improved Triplet Constraint

In this paper, note that $L = \{x_1, \ldots, x_l\}$ is the labled sample, $U = \{x_{l+1}, \ldots, x_n\}$ is the unlabeled sample, $D = L \cup U = \{x_1, \ldots, x_n\}$ for all sample sets, where $x_i \in R^m$. $X = [x_1, x_2, \ldots, x_n]$ is the data matrix containing the whole input samples.

Since it is desirable to minimize the distance between similar samples while maximizing it between heterogeneous ones, under this assumption, the triplet constraints can be defined as

$$\mathcal{T} = \{(x_i, x_j, x_k): D_{ij}^2 < D_{ik}^2\} \quad (5)$$

where $D_{ij}^2 = (x_i - x_j)^T A(x_i - x_j)$, x_i and x_j are in the same group and x_i and x_k do not belong to the same group. According to the definition of triplet constraints, if given the samples x_i , x_j, and x_k, the purpose is to find a semi-positive symmetric matrix A, and then use the matrix A to calculate the distance between these samples, maximizing $D_{ik}^2 - D_{ij}^2$ is the largest, that is equivalent to minimizing $D_{ij}^2 - D_{ik}^2$. Based on this requirement, an initial model can be derived:

$$\min_A \sum_T (D_{ij}^2 - D_{ik}^2), \quad \text{s.t. } A \geq 0 \quad (6)$$

In the regularized semi-supervised metric-learning (RSSML) [10], in order to introduce the sample distribution boundary, the RSSML model scales the triplet constraints and constructs the following form:

$$\mathcal{T}_2 = \left\{\left(x_i, x_j, x_k\right): D_{ij}^2 + 1 < D_{ik}^2\right\} \tag{7}$$

Then, the model can be converted to

$$\min_A \sum_{\mathcal{T}} \left[1 + D_{ij}^2 - D_{ik}^2\right]_+, \quad \text{s.t. } A \geq 0 \tag{8}$$

where $[z]_+ := \max\{z, 0\}$.If D_{ij}^2 plus a unit distance is still smaller than D_{ik}^2, a hinge loss function [11] is triggered.

The ability of this triplet constraint is influenced by the distribution of the sample points. When the distance between the same sample and the distance of the heterogeneous sample is very close, a judgment error may occur [12, 13].

Therefore, the parameter γ is introduced to avoid such problems, and thus the triplet constraint is rewritten as follow

$$\mathcal{T}_3 = \left\{\left(x_i, x_j, x_k\right): D_{ij}^2 < \gamma D_{ik}^2\right\} \tag{9}$$

where $\gamma (0 < \gamma \leq 1)$ is used to balance the impact of similar data and different types of data. $\gamma = 1/\left(1 + \bar{D}^{-1}\right)$, where $\bar{D}$ is the average distance of the data set D. When the difference between the homogeneous sample and the heterogeneous sample is very large, D is also very large, in which case γ tends to 1 and $\mathcal{T}_3$ is equivalent to $\mathcal{T}$. The revised model is:

$$\min_A \sum_{\mathcal{T}_3} \left(D_{ij}^2 - \gamma D_{ik}^2\right), \quad \text{s.t. } A \geq 0 \tag{10}$$

However, when the distance between the similar sample and the heterogeneous sample is very small, the model imposes too strict restrictions on D_{ij}^2, which causes some similar samples to be excluded from the model. In this paper, another parameter is added to the model. $\omega = 1 - \gamma$, the triplet constraint is as follow

$$\mathcal{T}_4 = \left\{\left(x_i, x_j, x_k\right): D_{ij}^2 < \gamma D_{ik}^2 + \omega\right\} \tag{11}$$

Obviously, when γ is 1, $\mathcal{T}_4$ is equivalent to $\mathcal{T}$. When $\mathcal{T}$ is very small due to the small $\bar{D}$, ω acts to relax the constraint, but overall, $\mathcal{T}_4$ is still more strict than $\mathcal{T}_2$.

The model is revised to:

$$\min_A \sum_{\mathcal{T}_4} \left(D_{ij}^2 - \gamma D_{ik}^2 - \omega\right), \quad \text{s.t. } A \geq 0 \tag{12}$$

In the LMNN algorithm [6], the distances are calculated in pairs to minimize because KNN classifier does not need all samples of one kind clustered into one family. Only the k-nearest neighbor of each sample is needed.

Define two indicator functions here:

$$\eta_{ij} = \begin{cases} 1, \ x_i \text{ and } x_j \text{ are neighbours} \\ 0, \ \text{other} \end{cases} \tag{13}$$

$$y_{ij} = \begin{cases} 1, \ x_i \text{ and } x_j \text{ belong to the same class} \\ 0, \qquad\qquad \text{other} \end{cases} \tag{14}$$

The model can be rewritten as:

$$\min_{A} \sum_{i,j,k}^{l} \eta_{ij} \left(1 - y_{ij}\right) \left(D_{ij}^2 - \gamma D_{ik}^2 - \omega\right), \quad \text{s.t. } A \geq 0 \tag{15}$$

And l is the number of labeled samples.

2.3 Regularization Term

Based on neighbor relationships, the following regular term is used in LRML algorithm to combine unlabeled data.

$$rg(A) = \frac{1}{2} \sum_{i,j=1}^{n} S_{ij} D_{ij}^2 \tag{16}$$

where

$$S_{ij} = \begin{cases} 1, \ fi \in N(j) \text{ or } j \in N(i) i \neq j \\ 0, \qquad\qquad \text{other} \end{cases}$$

$N(i) = \left\{x_j | x_j \text{ and } x_i \text{ are neighbour}\right\}$, that is, $N(i)$ represents the neighbor of x_i measured by the Euclidean distance.

According to the smoothing hypothesis, if the samples are neighbors, their distance in the new space is minimized: $\sum_{i=1}^{n} \sum_{j \in N(i)} D_{ij}^2$. Introduce the similarity $\left[S_{ij}\right]$, and according to the manifold hypothesis: $\sum_{i=1}^{n} \sum_{j \in N(i)} S_{ij} D_{ij}^2$.

According to the clustering hypothesis, for the samples in the high-density region, their distance is minimized by the parameter $\beta_i = f(p(x_i)) \in R^+$, where $p(x_i)$ is x_i's density and $f\colon R \to R$ is a non-negative monotone increasing function, which gives: $\sum_{i=1}^{n} \beta_i \sum_{j \in N(i)} S_{ij} D_{ij}^2$.

Finally, rewrite the regular term as:

$$rg(A) = \frac{1}{4} \sum_{i=1}^{n} \beta_i \sum_{j \in N(i)} S_{ij} D_{ij}^2 \tag{17}$$

Combined with the previous model, a semi-supervised metric-learning model was obtained:

$$\min_{A} \frac{1}{4} c \sum_{i=1}^{n} \beta_i \sum_{j \in N(i)} S_{ij} D_{ij}^2 + (1-c) \sum_{i,j,k}^{l} \eta_{ij} \left(1 - y_{ij}\right) \left(D_{ij}^2 - \gamma D_{ik}^2 - \omega\right) \quad (18)$$

$$\text{s.t. } A \geq 0$$

And $0 \leq c \leq 1$ is a balance parameter used to control the weight between regular term and loss term.

3 Model Simplification and Algorithm

Consider the convenient for calculation, simplify the model as follow

$$\min ctr\left(\mathrm{XLX^T}A\right) + (1-c)\left(tr(\mathrm{MA}) - \sum_{i,j}^{l} C_{ij}^1 \omega\right), \quad \text{s.t. } A \geq 0 \quad (19)$$

Where, $M = X_l(D_c - C)X_L^T$,$C = C^1 - C^0C^1 = \mathrm{diag}\left(\left(ee^T - Y\right)e\right)H$,, $C^0 = \gamma \mathrm{diag}(He)\left(ee^T - Y\right)$, $D_c = \mathrm{diag}(Ce)L = \sum_{i=1}^{n} L^{(i)}$, $L^{(i)} = D_w^{(i)} - W^{(i)}$ is a Laplacian matrix,$D_w^{(i)}$ is a diagonal matrix which consists of $D_w^{(i)}(k,k) = \sum_j W_{kj}^{(i)}$.

The optimal solution of the model can be obtained by using the steepest descent algorithm for the positive definite symmetric matrix (Table 1).

Table 1 Algorithm steps

Gradient descent algorithm for definite symmetric matrix
Input: L, U, D,
Output: Mahalanobis matrix A
1. Initialize the Mahalanobis matrix A, k (the norm for the nearest neighbor of any sample), α (step size per learning) and the maximum iterations T
2. For $t = 1$: T (1) Calculating the gradient (2) $G(t) = [A(t)]^{-\frac{1}{2}} \mathrm{Sym}[\nabla f(A(t))][A(t)]^{-\frac{1}{2}}$ (3) Update metric matrix A, $A(t+1) = A(t)^{\frac{1}{2}} \exp(-\alpha(t)G(t))A(t)^{\frac{1}{2}}$

4 Numerical Experiment

The experiment selected four different data sets (iris, wine, balance, breast cancer) in the UCI database for numerical experiments. Table 2 gives a brief description of the four types of data. $|L|$ represents the number of labeled samples, $|U|$ represents the total number of samples, and $|L|/|U|$ represents the labeling rate of the labeled samples.

During the experiment, the whole data is randomly divided into two parts: a labeled set L and an unlabeled set U. Consider each of its 4 neighbors for each sample and select 10 tag samples for each class. Randomly select 10 samples of each class as training samples and randomly select 5 samples of each class as a test set. Experiment was carried out 20 times for each data set and takes the average of 20 experiments as final.

As shown in Fig. 1, according to the average results of 20 experiments, the classification of these models is much better than the European distance classification. The

Table 2 Descriptive statistics

Data set	Sample nNumber	Feature number	Label number	$\|L\|$	$\|U\|$	$\|L\|/\|D\|(\%)$
Iris	150	4	3	15	135	10
Wine	178	13	3	15	163	8.43
Balance	625	4	3	15	610	2.4
Dermatology	358	34	6	30	328	8.38

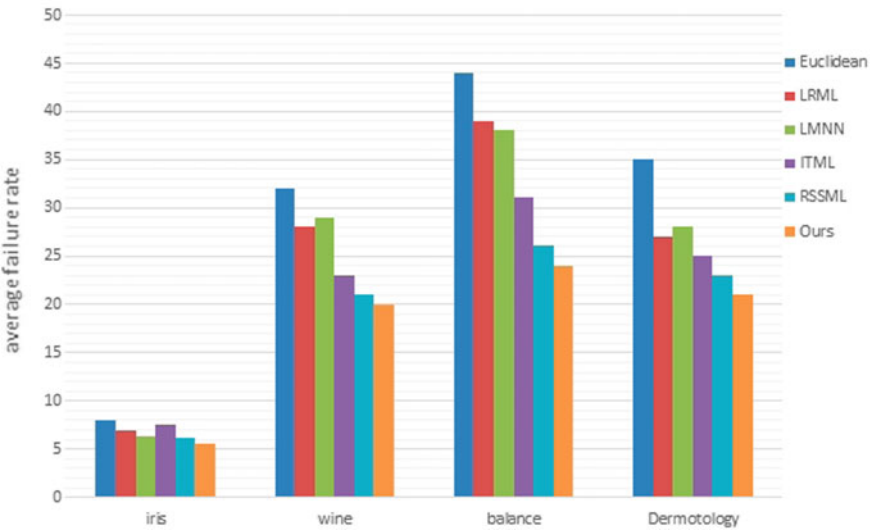

Fig. 1 Clustering results

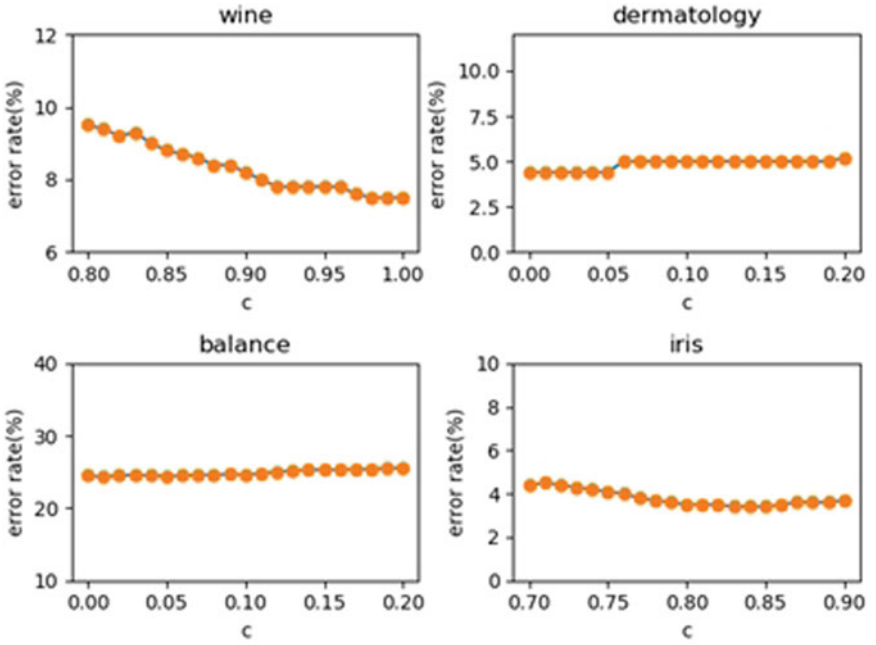

Fig. 2 Sensitivity for c

results presented in this paper are better than those tested on each data set. Second, the average error rate of the proposed method on each data set is the smallest.

When testing the model sensitivity of two parameters, c and v, the values of other parameters are fixed, and the change step of c is 0.01. For the dermatology and balance data sets, the range of c is selected as [0, 0.2]; for the wine data set, the range of c is set to be [0.8, 1]; for the iris data set, the range of c is set to be [0.7, 0.9]. Take the change step of v is 1, the range of change is [1, 15]. Figures 2 and 3 show the sensitivity for the two parameters, and the error rate of the experiment fluctuated slightly with the change of c and v.

5 Conclusions

This paper proposes a new semi-supervised metric-learning model, which uses topology information to construct regular terms to satisfy the three important assumptions of semi-supervised learning. For solving the model, a gradient descent method is used to get the metric matrix A. Then, we apply A to the KNN classifier and replace the commonly used Euclidean distance by the Mahalanobis distance calculated by the matrix to classify some data in the UCI database as a test set. Finally, compare the results of our model with several existing semi-supervised metric-learning methods.

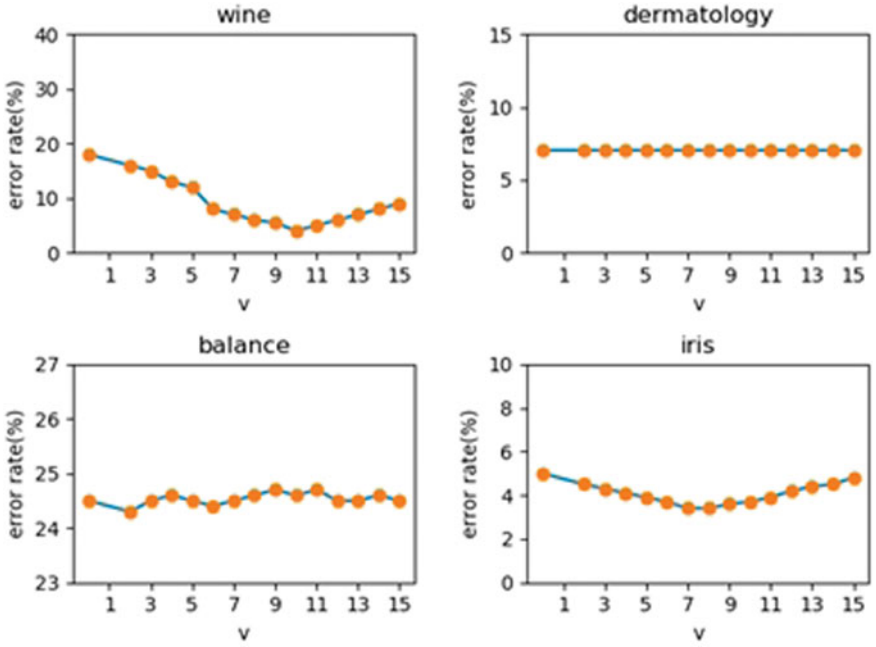

Fig. 3 Sensitivity for v

The numerical results illustrate that the classification effect of our model is better and meets the original design goals.

References

1. Xing, E.P., A.Y. Ng, M.I. Jordan, et al. 2002. Distance metric learning, with application to clustering with side-information. In *International Conference on Neural Information Processing Systems*, 521–528. Cambridge: MIT Press.
2. Schultz, M, and T. Joachims. 2003. Learning a distance metric from relative comparisons. In *International Conference on Neural Information Processing Systems*, 41–48. Cambridge: MIT Press.
3. Weinberger, K.Q., and L.K. Saul. 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research* 10 (1): 207–244.
4. Yaniv, Z. 2006. Principle component analysis.
5. Shang, Y, and W. Ruml. 2004. Improved MDS-based localization. In *Joint Conference of the IEEE Computer and Communications Societies*, vol. 4, 2640–2651. IEEE.
6. Lan, Y.D., H. Deng, and T. Chen. 2012. Dimensionality reduction based on neighborhood preserving and marginal discriminant embedding. *Procedia Engineering* 29 (4): 494–498.
7. Silva, V.D., and J.B. Tenenbaum. 2003. Unsupervised learning of curved manifolds. In *Nonlinear estimation and classification*, 453–465. New York: Springer.
8. Belkin, M., and P. Niyogi. 2003. *Laplacian Eigenmaps for dimensionality reduction and data representation*. Cambridgr: MIT Press.

9. Donoho, D.L., and C. Grimes. 2003. Hessian eigenmaps: locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences of the United States of America* 100 (10): 5591.
10. Wang, Q., P.C. Yuen, and G. Feng. 2013. *Semi-supervised metric learning via topology preserving multiple semi-supervised assumptions.* Amsterdam: Elsevier Science Inc.
11. Wu, Y., and Y. Liu. 2007. Robust truncated hinge loss support vector machines. *Publications of the American Statistical Association* 102 (479): 974–983.
12. Hoi, S.C.H., W. Liu, and S.F. Chang. 2010. Semi-supervised distance metric learning for collaborative image retrieval and clustering. *ACM Transactions on Multimedia Computing Communications & Applications* 6 (3): 1–26.
13. Der, M., and L.K. Saul. 2012. Latent coincidence analysis: a hidden variable model for distance metric learning. In *International Conference on Neural Information Processing Systems*, pp. 3230–3238. Curran Associates Inc. 2012.