ECAI 2016
G.A. Kaminka et al. (Eds.)

doi:10.3233/978-1-61499-672-9-1553

Burg Matrix Divergence Based Multi-Metric Learning

Yan Wang and **Han-Xiong Li** [1]

Abstract. The basic idea of most distance metric learning methods is to find a space that can optimally classify data points belong to different categories. However, current methods only learn one Mahalanobis distance for each data set, which actually fails to perfectly classify different categories in most real world applications. To improve the classification accuracy of k-nearest-neighbour algorithm, a multi-metric learning method is proposed in this paper to completely classify different categories by sequentially learning sub-metrics. The proposed algorithm is based on minimizing the Burg matrix divergence between metrics. The experiments on five UCI data sets demonstrate the improved performance of Multi-Metric learning when comparing with the state-of-the-art methods.

1 Introduction

Learning a good distance metric in feature space is crucial in many learning algorithms, such as nearest neighbors classifier and K-means clustering [2]. Over the past decade, a large number of distance metric learning (DML) algorithms have been proposed to learn a Mahalanobis distance in feature space, and some of them have been successfully applied to real world applications. In order to learn a distance metric that can well classify the dis-similar data pairs, an earlier work [2] uses a semi-definite programming formulation under similarity and dissimilarity constraints. In [5], the Large Margin Nearest Neighbor (LMNN) is suggested to learn a Mahalanobis distance metric for kNN Classification.

Current DML methods are actually learning one metric space that properly classify different categories. Unfortunately, due to the complexity and uncertainty, a linear space that can perfectly classify different categories may not exist. In order to remedy the disadvantages, we propose a method to learn multi-metric spaces so that all the training data can be correctly classified in at least one metric space, as shown in figure 1. In section 2, the base-metric learning problem and sub-metric learning problem will be defined. In section 3, the optimization of these two problems are introduced. In case study, we conduct experiments on five public data sets to demonstrate the effectiveness of the proposed method.

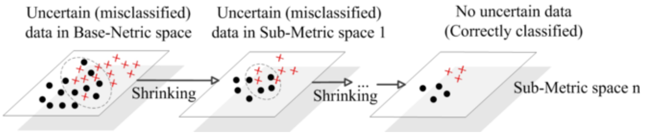
Uncertain (misclassified) data in Base-Netric space
Uncertain (misclassified) data in Sub-Metric space 1
No uncertain data (Correctly classified)
Shrinking
Shrinking
...
Sub-Metric space n

Figure 1. The Sequentially shrinking of Inseparable Set

2 Problems and Definitions

Given a data set x_i, where $x_i \in R^d, i = 1, 2, \ldots, n$, the Mahalanobis distance parameterized by positive semi-definite (PSD) matrix A is expressed as:

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

The uncertain data in this metric space can be defined as follows.

Definition 1 *Uncertain Data. For data x_i, if*

$$d_A(x_i, \mu_{i,k}) - d_A(x_i, \mu_{i,j}) < \rho \tag{2}$$

then x_i is uncertain in metric space A. In this formula, $\mu_{i,j}$ is the class center with the same label to x_i, and $\mu_{i,k}$ is the nearest class center with different class label to x_i, ρ is a hyper-parameter, which represents the desired margin between different classes

Definition 2 *Uncertain Set. If x_i is uncertain in all the existing metric spaces, then it will belong to uncertain set U. As shown in figure 1, the goal of proposed method is gradually shrinking the uncertain set to empty: $U \to \emptyset$.*

Definition 3 *Base-Metric Learning. The base-metric will be a Mahalanobis distance parameterized by a PSD matrix A^0. It is a global optimal metric learnt with the following problem:*

$$\begin{aligned} &\min_{A^0} D_\phi(A^0, M) \\ &s.t. 1) \sum_i^n (d_{A^0}(x_i, \mu_{i,j}) - u) \leq 0 \\ &\quad 2) \sum_i^n (d_{A^0}(x_i, \mu_{i,k}) - l) \geq 0; \end{aligned} \tag{3}$$

where $D_\phi()$ denotes the distance between matrix A^0 and M, M is a baseline matrix (we choose M equal to the covariance matrix of training set in this paper), u and l are upper limit and lower limit for distance, respectively.

Definition 4 *Sub-Metric Learning. The Sub-metrics is a group of Mahalanobis distances parameterized by PSD matrix $A^1, \cdots, A^k$. Under different distance constraints that force the data in uncertain set to be correctly classified, these distance metrics will be learnt with following problem:*

$$\begin{aligned} &\min_{A^{new}} D_\phi(A^{new}, A^0) \\ &s.t. \sum_i^{x^i \in U} (d_{A^{new}}(x_i, \mu_{i,k}) - d_{A^{new}}(x_i, \mu_{i,j})) \geq \rho * N_U \end{aligned} \tag{4}$$

where U denotes the uncertain set defined in definition 2, N_U is the size of U. The details of this problem will be analyzed in section 3.

[1] Department of Systems Engineering and Engineering Management, City University of Hong Kong

3 Optimization

3.1 Measure of Similarity Between Metrics

In problem (3) and (4), the objective is minimizing the difference $D_\phi()$ between target matrix A^0 and original matrix M. In this paper, The Burg matrix divergence is adopted to quantify this difference [4], which defines the $D_\phi()$ as:

$$\begin{aligned} D_\phi(A, M) &= KL(p(x, A)||p(x, M)) \\ &= tr(AM^{-1}) - \log|AM^{-1}| - d \end{aligned} \tag{5}$$

3.2 Base-Metric Learning

With formula (5), we can rewrite the base-metric learning problem in (3) as a Burg matrix optimization process:

$$\begin{aligned} &\min_{A^0} tr(A^0 M^{-1}) - \log|A^0 M^{-1}| - d \\ &s.t. 1) tr((A^0)^{-1} \sum_i^n (x_i - \mu_{i,j})(x_i - \mu_{i,j})^T) \le n * u \\ &\quad 2) tr((A^0)^{-1} \sum_i^n (x_i - \mu_{i,k})(x_i - \mu_{i,k})^T) \ge n * l; \end{aligned} \tag{6}$$

In above formulation, since the constraints here only have demands on the averaged distance to class centers, they are weaker than the commonly used pairwise constraints [4] or triplet constraints [3]. A colesd-form solution for this problem is derived in [1].

3.3 Sub-metric learning

The weaker constraints adopted in problem (6) may not be able to ensure a prefect metric that can found linear boundaries between different categories. To remedy the disadvantage of base-metric learning, we propose sub-metric learning problem shown in (4). The task of sub-metric learning is shrinking the uncertain set U so that each data instance can be correctly classified in at least one metric. Following formula (5), the sub-metric learning can be rewritten as:

$$\begin{aligned} &\min_{A^{new}} tr(A^{new} {A^0}^{-1}) - \log|A^{new} {A^0}^{-1}| - d \\ &s.t. \quad tr((A^{new})^{-1} \sum_i^{x_i \in U} ((x_i - \mu_{i,k})(x_i - \mu_{i,k})^T) \\ &\qquad - (x_i - \mu_{i,j})(x_i - \mu_{i,j})^T)) \le -\rho * N_U; \end{aligned} \tag{7}$$

where U denotes the uncertain set (defined in section 2), N_U is the size of U. A closed-form solution for this problem is proposed in [3], which is much more efficient than other common DML methods.

3.4 Classification

For input x_i, its confidence weight in metric space A^m is defined as:

$$\omega_i = -\log \frac{d_{A^m}(x_i, \mu_{1st})}{d_{A^m}(x_i, \mu_{2nd})} \tag{8}$$

where μ_{1st} denotes the nearest class center to x_i, and μ_{2nd} denotes the second nearest class center to x_i. Then, the classification result will be the class label with the maximum weight. For example, in $0-1$ classification, the probability of $y = 1$ will be:

$$p(y = 1|x, A^0, \ldots, A^m) = \frac{\sum_{i=0}^m \omega_i f_{A_i}(x*)}{\sum_{i=0}^m \omega_i} \tag{9}$$

4 Experiments

In this section we compare the proposed Multi-ML, method with a few methods: Euclidean distance, Mahalanobis distance, lda, ITML [4] and LMNN [5]. Experiments were run on 5 UCI data sets, that are: 1)Pima Indian Diabetes, 2)Breast Cancer Wisconsin Diagnostic, 3)Heart, 4)Liver Disorders and 5)Robot execution failures . All experimental results are obtained by averaging 50 runs. For each run, we randomly split the data sets 70% for training and 30% for testing.

Table 1. KNN (k=1) average classification accuracy of 50 random experiments via different metrics

	Diabetes	WDBC	Heart	Liver	Failures
Multi-ML	**0.693**	**0. 936**	**0.758**	0.607	**0.879**
LMNN	0.680	0.916	0.660	0.588	0.853
LDA	0.678	0.927	0.755	0.567	0.882
ITML	0.681	0.912	0.728	**0.608**	0.850
Euclidean	0.680	0.915	0.590	0.607	0.798
Mahalanobis	0.671	0.895	0.581	0.551	0.785

As we can see from Table 1, the proposed Multi-ML method outperforms other state-of-the-art methods on 4 of the 5 date sets. To understand the complexity of proposed method, the average computation time of different algorithms are listed in Table 2. We can find that Multi-ML is much faster than LMNN and ITML.

Table 2. Computation time (s)

	Diabetes	WDBC	Heart	Liver	Failures
LMNN	3.65	1.55	0.33	0.70	7.15
ITML	5.34	6.16	6.37	4.79	27.82
Multi-ML	0.10	0.17	0.062	0.071	1.01

5 Conclusion

Instead of current single-metric learning method, a multi-ML is proposed in this paper to improve the accuracy classification. By proposing base-ML and sub-ML problem as Burg matrix optimization problem, the proposed model enables us to derive an efficient close-form algorithm. The experiments on five UCI data sets prove the effectiveness of the proposed method.

REFERENCES

[1] M. Sustik B. Kulis and I. Dhillon, 'Learning low-rank kernel matrices', in *Proceedings of the 23rd international conference on Machine learning,(ICML 2007)*, pp. 505–512, (2006).

[2] S. Russell E. P. Xing, M. I. Jordan and A. Y. Ng, 'Distance metric learning with application to clustering with side-information', in *Advances in neural information processing systems (NIPS 2002)*, pp. 505–512, (2002).

[3] H. R. Karimi J. Mei, M. Liu and H. Gao, 'Logdet divergence based metric learning using triplet labels', in *Proceedings of the Workshop on Divergences and Divergence Learning (ICML 2013)*, ed., Springer, (2013).

[4] P. Jain S. Sra J. V. Davis, B. Kulis and I. S. Dhillon, 'Information-theoretic metric learning', in *Proceedings of the 24th international conference on Machine learning, (ICML 2007)*, ed., Springer, pp. 209–216, (2007).

[5] J. Blitzer K. Q. Weinberger and L. K. Saul, 'Distance metric learning for large margin nearest neighbor classification', in *Advances in neural information processing systems (NIPS 2005)*, pp. 1473–1480, (2005).