Neurocomputing 129 (2014) 185–198

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Globality and locality incorporation in distance metric learning

Wei Wang [a,*], Bao-Gang Hu [b], Zeng-Fu Wang [a]

[a] Department of Automation, University of Science and Technology of China, Hefei 230026, PR China
[b] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, PR China

ARTICLE INFO

Article history:
Received 31 May 2013
Received in revised form
10 September 2013
Accepted 28 September 2013
Communicated by Ran He
Available online 26 October 2013

Keywords:
Mahalanobis distance
Metric learning
Convex optimization
Classification

ABSTRACT

Supervised distance metric learning plays a substantial role to the success of statistical classification and information retrieval. Although many related algorithms are proposed, it is still an open problem about incorporating both the geometric information (i.e., locality) and the label information (i.e., globality) in metric learning. In this paper, we propose a novel metric learning framework, called "Dependence Maximization based Metric Learning" (DMML), which can efficiently integrate these two sources of information into a unified structure as instances of convex programming without requiring balance weights. In DMML, the metric is trained by maximizing the dependence between data distributions in the reproducing kernel Hilbert spaces (RKHSs). Unlike learning in the existing information theoretic algorithms, however, DMML requires no estimation or assumption of data distributions. Under this proposed framework, we present two methods by employing different independence criteria respectively, i.e., Hilbert–Schmidt Independence Criterion and the generalized Distance Covariance. Comprehensive experimental results for classification, visualization and image retrieval demonstrate that DMML favorably outperforms state-of-the-art metric learning algorithms, meanwhile illustrate the respective advantages of these two proposed methods in the related applications.

1. Introduction

Distance functions are critical for many models and algorithms in machine learning and pattern recognition, such as k-nearest neighbor (kNN) classification and k-means clustering. The metric distances provide a measurement of dissimilarity between different points and significantly influence the performance of these algorithms. Due to limited prior knowledge, most algorithms use simple Euclidean distances. However, such distances cannot ensure satisfactory results in many types of applications where the intrinsic spaces of data are not Euclidean. Previous research results [1,4,7] have shown that empirically learnt distance metrics lead to substantial improvement to the Euclidean distances when the prior information is not available. In addition, metric learning has been successfully applied to a large portion of real-world problems, including visual object categorization [34], image retrieval [36] and cartoon synthesis [37].

Recently, many excellent algorithms have been developed for metric learning [2,3,5,6,21,33,35]. Among the related studies, most effort has been spent on learning a Mahalanobis distance from labeled training data. The Mahalanobis distances generalize standard Euclidean distances by scaling and rotating feature spaces. After gaining a deep insight into the popular Mahalanobis distance learning algorithms, we summarize a hierarchical diagram shown in Fig. 1 which classifies them into different categories. For better understanding of the significant differences between these approaches, we initially classify them in terms of the information they considered. One category attempts to learn distance metrics using class labels for classification, the other category considers both the label information (i.e., globality) and the geometric information (i.e., locality).

1.1. Globality metric learning

Research in the first category (i.e., globality metric learning) is driven by the need of keeping all the data points in the same class close together for *compactness* while ensuring those from different classes far apart for *separability*. To this end, a number of algorithms [4–6,12,22,24] have been proposed which can be further divided into the following subcategories as shown in Fig. 1.

(1) *Algorithms based on similarity/dissimilarity*: A natural intention in metric learning is to keep the points of the same label similar (i.e., the distance between them should be relative small) and others dissimilar (i.e., the distance should be larger). For instance, the method proposed by Xing et al. [4] formulates a convex metric learning program by reducing the averaged distance between similar instances under the constraint of separating dissimilar instances. Metric Learning by Collapsing Classes (MCML) [5] aims to find a distance metric that collapses all the points in

* Corresponding author.
E-mail address: weiyi889@mail.ustc.edu.cn (W. Wang).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.09.041

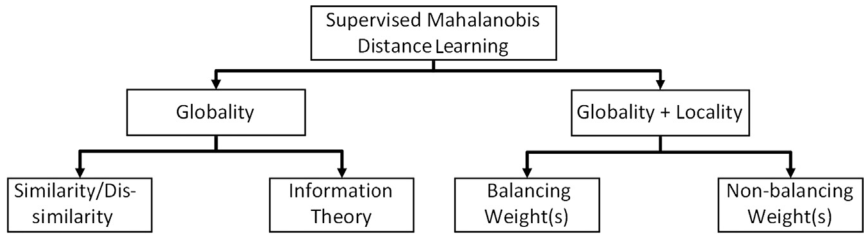

Fig. 1. Hierarchical diagram of classifying studies in supervised Mahalanobis distance learning.

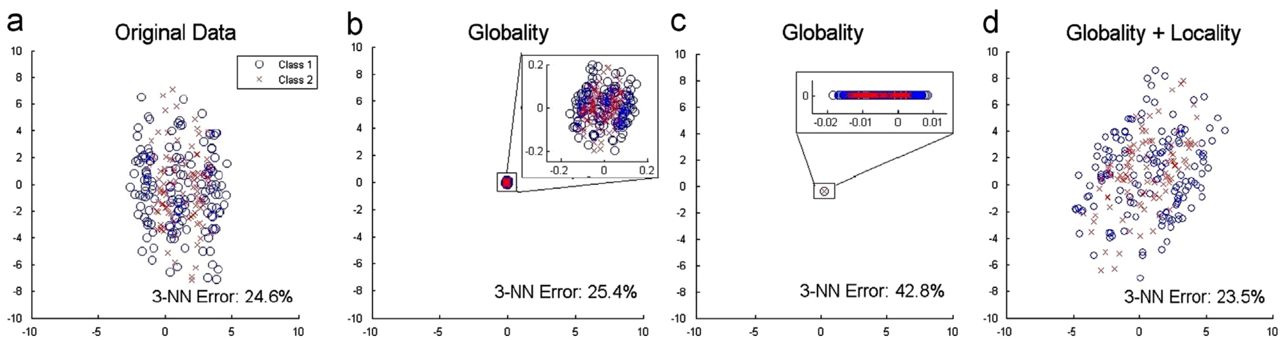

Fig. 2. The multimodal data distributions (a) without projection, adjusted by (b) globality method [4], (c) globality method [5] and (d) globality+locality method [16].

the same class while maintaining the separation between different classes.

(2) *Algorithms based on information theory*: To introduce information theory into metric learning, these algorithms define two Gaussian distributions. The first Gaussian is based on the Mahalanobis distance to be learnt and the second is determined heuristically. In this term, the distance metric can be learnt by minimizing the relative entropy between these two distributions. In particular, Information-Theoretic Metric Learning (ITML) [6] defines the second Gaussian from prior knowledge and searches the optimal metric by minimizing Kullback–Leibler (K–L) divergences between them, subject to a set of similarity and dissimilarity constraints. The information geometry algorithm [23] uses an ideal kernel given from the class labels [24] to construct the second Gaussian and also minimizes K–L divergences between two distributions.

1.2. *Globality + locality metric learning*

The previously discussed globality metric learning algorithms have been successfully applied to various fields. In these algorithms, data points are generally assumed to have unimodal distributions. Based on this assumption, they attempt to minimize distances between all pairs in the same class. For this reason, however, the globality algorithms are not appropriate for multimodal data distributions, since their goals (i.e., compactness and separability) conflict and can be hardly achieved simultaneously [17]. To alleviate this problem, it is expected to incorporate the geometric information (i.e., locality) with the label information (i.e., globality) in metric learning. This issue is of particular importance and has got significant progress recently [15–20]. As an empirical validation, we compare two globality methods (i.e., Xing et al. [4] and Globerson and Roweis [5]) with a globality + locality method (i.e., Weinberger et al. [16]) on a synthetic 2-D data set. The 2-D data set has two classes and each class has distinct modes. From their results shown in Fig. 2, it is clear that the globality methods [4,5] actually increase the 3-NN classification error, since they focus on minimizing all the pairwise distances between similarly labeled data points. Even worse, [5] collapses the entire data set into a straight line. By contrast, [16] reduces the error rate, due to the adaption to local structures. Though this evaluation is conducted on a synthetic data set, it shows in general the problems posed by multimodal data distributions and illustrates the advantages of considering the global and the local information simultaneously.

In the globality+locality category, the fundamental challenge is how to combine locality and globality. As concluded in Fig. 1, typical algorithms [16,19,20] address this challenge by directly generating balancing weight(s). These weights are involved through two strategies: (1) combining the term of local manifold into the objective function of globality metric learning; (2) formulating the globality learning objective function subject to a set of *triplet-based* constraints which state that a point should be similar to another point but dissimilar to a third point in the learnt metric. Particularly, Large Margin Nearest Neighbor (LMNN) [16] learns a distance metric from the local neighborhood and is solved by a semidefinite programming (SDP) [38] incorporating a great number of triplet-based constraints. Hoi et al. [19] provide an SDP to consider the topological structure of data along with similarity and dissimilarity constraints. Zhong et al. [20] propose a parametric manifold regularizer to the metric learning model based on a large set of triplet-based constraints. Although the algorithms discussed above have been extensively investigated in the literature, they present a limitation that many different balancing weights need to be tuned or optimized, such as slack variables, the amount of constraints and the weight in the objective function. The values of these balancing weights have great influences on metric learning performance. Even worse, the computational complexity may be increased rapidly for large-scale and high-dimensional data, since many optimizations, e.g., the general-purpose solver of SDP in LMNN, involve iterative procedures and have to solve massive numbers of constraints in each iteration. Therefore, it is promising to design an efficient metric learning algorithm to address the challenges originated from balancing weight(s). [15,17] are the researches most related to this paper, in which a globality+locality distance metric is learnt without optimizing balancing weight(s) and iteratively satisfying constraints. Despite similar goals, our approach differs significantly in the essential conception and optimization. Goldberger et al. [15] propose a I prefer "Neighbourhood" since this is the title