Information Sciences 282 (2014) 321–331

Contents lists available at ScienceDirect

Information Sciences

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

Multi-granularity distance metric learning via neighborhood granule margin maximization

Pengfei Zhu [a,b], Qinghua Hu [a,*], Wangmeng Zuo [c], Meng Yang [d]

[a] School of Computer Science and Technology, Tianjin Univerity, Tianjin 300073, China
[b] Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China
[c] School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
[d] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518000, China

ARTICLE INFO

Article history:
Received 9 July 2013
Received in revised form 6 April 2014
Accepted 8 June 2014
Available online 18 June 2014

Keywords:
Neighborhood granular margin
Metric learning
Neighborhood rough set
Multiple granularity

ABSTRACT

Learning a distance metric from training samples is often a crucial step in machine learning and pattern recognition. Locality, compactness and consistency are considered as the key principles in distance metric learning. However, the existing metric learning methods just consider one or two of them. In this paper, we develop a multi-granularity distance learning technique. First, a new index, neighborhood granule margin, which simultaneously considers locality, compactness and consistency of neighborhood, is introduced to evaluate a distance metric. By maximizing neighborhood granule margin, we formulate the distance metric learning problem as a sample pair classification problem, which can be solved by standard support vector machine solvers. Then a set of distance metrics are learned in different granular spaces. The weights of the granular spaces are learned through optimizing the margin distribution. Finally, the decisions from different granular spaces are combined with weighted voting. Experiments on UCI datasets, gender classification and object categorization tasks show that the proposed method is superior to the state-of-the-art distance metric learning algorithms.

1. Introduction

How to construct or learn a proper distance or similarity measure is a key problem in clustering and classification such as *k*-means, and *k* nearest neighbor searching [10,51,54]. Whereas, the optimal distance metric may be problem-specific and up to the underlying data structure and distributions. To this end, there have been increasing efforts made to learn a distance metric in recent years [6,10,37,39,41,50]. Metric learning methods can be categorized into unsupervised [50], semi-supervised [5] and supervised ones [1,11,25,27,34,53], according to the availability of the labels of training samples. Metric learning has been proved to successfully improve the clustering and recognition performance in information retrieval [29,30], bioinformatics [47] and computer vision tasks [6,10,12,16,39,37].

Generally speaking, metric learning aims to learn an effective distance metric, measured by which the samples from the positive sample pair (i.e., samples with the same class label or similar samples) could be as close as possible, while the samples from the negative sample pair (i.e., samples with the different class labels or dissimilar samples) could be as far as possible. In most cases, a metric learning model has three key components: sample pairs; objective function; regularization. In supervised learning, sample pairs can be generated from k nearest neighbors, e.g., large margin nearest neighbor (LMNN) [6]

* Corresponding author. Tel.: +86 22 27401839.
E-mail address: huqinghua@tju.edu.cn (Q. Hu).

http://dx.doi.org/10.1016/j.ins.2014.06.017
0020-0255/

and neighborhood component analysis (NCA) [15]. In verification tasks, samples pairs could be randomly generated by putting two similarly labeled samples into positive pairs and two differently labeled samples into negative pairs [16,34,37,39,53]. In weakly supervised learning, side information is provided and similar/dissimilar sample pairs are given [50,18]. The objective function is often established by minimizing the distance between two samples in positive pairs and maximizing the distance between two samples in negative pairs [6,16,37]. Besides, Bar et al. proposed to maximize mutual information between the original data and embedded data [3]. To get a stable solution and the expected property for the learned metric, regularizations such as trace of matrix [53], log-determinant regularization [42], sparse regularization [23,32] and nuclear norm [36] are imposed on the learned parameters in different applications [38].

In distance based classification, the performance of local classifiers, e.g., nearest neighbor classifier and neighborhood classifier, is greatly affected by the local distribution of the training samples. Many metric learning methods aim to learn a distance metric to get expected local data structure. The local data structure, i.e., neighborhood, can be evaluated from locality, compactness and consistency. Locality means the neighborhood relationship in the original space, which should be kept in the learning process. Locality preserving is widely applied in dimension reduction [17,40,43], spectral analysis [7,55] and sparse coding [48]. Compactness measures the closeness of samples in the neighborhood, which is the main principle in many clustering algorithms [24]. Consistency is used to measure ratio of the samples that can be recognized with the Bayes rule [9]. A good metric should be capable to preserve locality, lead to compact local data structure and high consistency.

In [21], a neighborhood rough set model is proposed based on neighborhood granulation. The samples in the neighborhood of each sample form a neighborhood granule. Then, a family of neighborhood granules forms an elemental granule system that covers the universe. By computing the consistency of neighborhood granules, the universe is divided into decision positive regions and decision boundary regions. The percentage of samples in the decision positive regions is defined as neighborhood dependency [20,21]. Neighborhood dependency only counts the pure neighborhood granules and does not reflect the real consistency. Then the decision boundary regions are further grouped into recognizable and misclassified subsets based on the class probabilities in the neighborhood. The percentage of misclassified samples is defined as neighborhood decision error rate [19]. Neighborhood decision error reflects the consistency of neighborhood structures. However, it does not consider the locality and compactness. In this work we design a new evaluation index which simultaneously considers the locality, compactness and consistency.

When we learn the distance metric with neighborhood information, a problem appears, i.e., how to set the size of neighborhoods. It is suggested that multi-granularity data analysis may lead to performance improvement. Multiple granularity leads to diverse viewpoints of the world. In different granular spaces, we may view an object differently or get different decisions. This observation has been widely used in feature extraction, feature learning and classifier design. For example, in feature extraction, Gabor feature extracts features in different scales, that is, different down-sampling rate [33]. Additionally, spatial pyramid model in matching uses pooling technique to combine the feature extracted in different patch sizes [52,28]. In feature learning and representation, deep learning, is actually a multi-granularity method. Deep learning learns low-level, middle-level and high-level features, and each level can be interpreted as a granularity [4]. For classifier design, a multi-scale face recognition method is proposed by combining the decision of different scales [57]. In [56], an adaptive neighborhood granularity selection and combination method is proposed to solve the granularity-sensitive problem in neighborhood granular models. Hence, we can learn multiple distance metrics under different granularity and then combine the decisions made from the learned metrics.

In this paper, we propose a multi-granularity neighborhood distance metric learning (MGML) method. Firstly, we propose neighborhood granule margin to evaluate a distance metric. Neighborhood granule margin is defined by maximum log-likelihood of Bayes error. Then we formulate the metric learning problem as a support vector machines (SVM) model, which can be effectively solved by standard SVM solvers. Hence, it is quite efficient and has good scalability. As the optimal neighborhood size may be task-specific, we propose a multi-granularity method to combine the decisions of different granularity. For each neighborhood size, we can learn a distance metric, and then a decision is got. By margin distribution optimization, the granularity weights are learned. Finally, the decisions of different granularity are combined using the learned weights. Experiments on UCI datasets, gender classification, object categorization show that the proposed metric learning method is competent with the state-of-the-art metric learning methods.

The rest of this paper is organized as follows: Section 2 introduces neighborhood granule margin; Section 3 gives the metric learning model by maximizing neighborhood granule margin; Section 4 proposes the multi-granularity distance metric learning method; experimental analysis is described in Section 5, and Conclusions are given in Section 6.

2. Neighborhood granule margin

Given an information system$\langle \boldsymbol{U}, \boldsymbol{A}, \boldsymbol{D} \rangle$, $\boldsymbol{U} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ is a non-empty set of objects, $\boldsymbol{A} = \{a_1, \ldots, a_m\}$ is a set of attributes which describe samples, and $\boldsymbol{D}$ is the decision variable.

Definition 1. [21] Given $\boldsymbol{x}_i \in \boldsymbol{U}$, the neighborhood $\delta(\boldsymbol{x}_i)$ of $\boldsymbol{x}_i$ is defined as

$$\delta(\boldsymbol{x}_i) = \{\boldsymbol{x}_j | \boldsymbol{x}_j \in \boldsymbol{U}, \Delta(\boldsymbol{x}_i, \boldsymbol{x}_j) \leqslant \delta\}, \tag{1}$$

where Δ is a distance function defined in feature spaces and δ is the neighborhood size.