Engineering Applications of Artificial Intelligence 67 (2018) 173–186

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Sparse Bayesian similarity learning based on posterior distribution of data

Davood Zabihzadeh, Reza Monsefi *, Hadi Sadoghi Yazdi

Computer Department, Engineering Faculty, Ferdowsi University of Mashhad (FUM), Mashhad, Iran

ARTICLE INFO

Keywords:
Similarity learning
Metric learning
Latent space
Posterior distribution
Bayesian inference

ABSTRACT

A major challenge in similarity/distance learning is attaining a strong measure which is close to human notions of similarity. This paper shows why the consideration of data distribution can yield a more effective similarity measure. In addition, the current work both introduces a new scalable similarity measure based on the posterior distribution of data and develops a practical algorithm that learns the proposed measure from the data. To address scalability in this algorithm, the observed data are assumed to have originated from low dimensional latent variables that are close to several subspaces. Other advantages of the currently proposed method include: (1) Providing a principled way to combine metrics in computing the similarity between new instances, unlike local metric learning methods. (2) Automatically identifying the real dimension of latent subspaces, by defining appropriate priors over the parameters of the system via a Bayesian framework. (3) Finding a better projection to low dimensional subspaces, by learning the noise of the latent variables on these subspaces. The present method is evaluated on various real datasets obtained from applications, such as face verification, handwritten digit and spoken letter recognition, network intrusion detection, and image classification. The experimental results confirm that the proposed method significantly outperforms other state-of-the-art metric learning methods on both small and large-scale datasets.

1. Introduction

Similarity/Distance plays a key role in many machine learning and pattern recognition tasks, such as classification, clustering, Content-Based Information Retrieval (CBIR), recommender systems (Guo et al., 2016), visual tracking (Jiang et al., 2012; Li et al., 2012), image annotation (Guillaumin et al., 2009), web page archiving (Law et al., 2012), and cartoon synthesis (Yu et al., 2012) to name a few (Bellet et al., 2014). For many applications, standard measures, such as Euclidian distance or cosine similarity, are neither rich nor flexible enough to capture the human notion of similarity. Often, a better similarity measure can be learned from data. Metric learning algorithms aim to learn a distance function from data which brings semantically similar data items closer and keeps the conceptually dissimilar ones at a distance. Despite their success, metric learning algorithms are still restricted from the following aspects:

1- These algorithms assume that an optimal distance function takes the form of a Mahalanobis distance in input or the feature space induced by a kernel. However, this measure is not versatile enough to model human notions of similarity.

2- Mahalanobis methods require learning $O(d^2)$ parameters where d denotes the dimension of data. On the other hand, kernel methods require $O(n^2)$ parameters where n represents the number of training instances. Thus, these methods are infeasible on both high dimensional and large databases.

3- Local metric learning methods learn several metrics across the data manifold. However, these approaches do not provide a principled way to combine the learned metrics when computing the distance between new examples.

4- State-of-the-art approaches aim only to learn a low-rank linear or nonlinear projection and so ignore the noise of data on low dimensional latent spaces.

5- Hyperparameters of the system are often found by ad-hoc approaches, such as cross-validation, which required much time to learn from training data.

To deal with these issues, this paper initially presents a better notion of similarity by considering the structure and distribution of data. Then, a new similarity measure is introduced based on the posterior distribution of data. The present work also develops a practical algorithm that learns the proposed similarity measure from data. To address scalability

[*] Corresponding author.
E-mail addresses: d.zabihzadeh@mail.um.ac.ir (D. Zabihzadeh), monsefi@um.ac.ir (R. Monsefi), h-sadoghi@um.ac.ir (H.S. Yazdi).

https://doi.org/10.1016/j.engappai.2017.09.023
Received 11 January 2017; Received in revised form 7 September 2017; Accepted 25 September 2017
0952-1976/© 2017 Elsevier Ltd. All rights reserved.

in this algorithm, it is assumed that data are generated from latent variables that are close to one or more low dimensional manifolds. The current method is named "Sparse Similarity Learning based on Posterior distribution (SSLP)". When compared to similar works, its other advantages are:

1- SSLP learns the noise of data on latent subspaces so as to find more efficient projection matrices from an input space to low-dimensional latent subspaces.
2- Unlike available local metric learning methods, SSLP provides a principled way to combine metrics in computing the similarity between new observations.
3- The optimal values of most hyperparameters in the currently proposed model are automatically adjusted by the introduced algorithm, thus reducing training time and rendering the algorithm practical for many real applications.
4- SSLP finds an optimal sparse solution that increases the generalization of the proposed model and decreases its evaluation time on new data points.

The rest of the current paper is organized as follows: Section 2 reviews related works. In Section 3, the proposed similarity measure and its probabilistic model are presented. The learning algorithm is developed in Section 4 and, in Section 5, some theoretical results are established which guarantee generalization of the proposed method. A comparison with state-of-the-art methods and experimental results are reported in Section 6. Finally, Section 7 concludes with remarks and recommendations for future work.

2. Related works

Distance metric learning is an active research field in the machine learning community. Many studies in this area have focused on finding the optimal Mahalanobis metric, which is equivalent to learning an optimal linear projection. Seminal Mahalanobis works include MMC (Xing et al., 2003), LMNN (Weinberger and Saul, 2009), ITML (Davis et al., 2007), BoostMetric (Shen et al., 2012), SERAPH (Niu et al., 2014) and MSML (Qian et al., 2015a). In order to enforce the positive-semi-definite (p.s.d.) constraint, Mahalanobis metric learning algorithms need a full eigenvector decomposition at each stage with cost $O(d^3)$, where d represents the dimension of the input data. Hence, these methods are not scalable for high dimensional datasets. To address this problem, ITML uses the $log\ det$ regularization term that provides an automatic way to satisfy the p.s.d constraint. In MSML, *dual random projection* is proposed for high dimensional metric learning. This method first estimates dual variables on the low dimensional space of a random projection matrix and then employs these variables, along with input data, to construct a metric in the original space. In problems with a large number of constraints, MSML divides the original problem into multiple stages. At each stage, only a small subset of constraints is selected by sampling methods and the current solution is updated only by these constraints.

Some approaches directly learn the low-rank projection matrix $\boldsymbol{W} \in \mathbb{R}^{d\times p}$ (Goldberger et al., 2005; Xiang et al., 2008; Soleymani and Shouraki, 2010; Der and Saul, 2017; Wang et al., 2014; Perrot and Habrard, 2015). In the case of $p < d$, these methods reduce the dimensionality of the input data. In applications in which data lie close to a latent manifold with a dimensionality of $p \ll d$, these methods show better scalability with respect to the dimensionality of data. Xiang's method (Xiang et al., 2008) is a major work in this area which aims to find a projection matrix that both maximizes the sum of the square distances between dissimilar data items and also minimizes the sum of the square distances between similar data items. Here, the algorithm finds an optimal metric by utilizing an efficient binary search method. This method is extended by Wang et al. (2014) using the l_1 norm. However, because the introduced objective function is very sensitive to the initial value of the projection matrix, this method uses Xiang's method to initialize the solution.

LCA (Der and Saul, 2017) is a probabilistic model proposed for metric learning in latent space. Although this method learns the noise of data on latent subspaces, this noise is simply ignored when computing the distance at evaluation time. LCA is also based on the Maximum Likelihood that is prone to overfitting, especially for small and high dimensional training sets.

For complex datasets, where the discriminatory power of input features varies locally, Mahalanobis and linear projection methods are not flexible enough to fit the true distance function in different regions of input space. Local metric learning addresses this issue by learning one metric for each region of the input data (Weinberger and Saul, 2009; Verma et al., 2012; Wang et al., 2012) to the extreme of determining one metric for each training example (Noh et al., 2010; Mu et al., 2013; Fetaya and Ullman, 2015). For example, MM_LMNN (Weinberger and Saul, 2009) learns one metric per class. To ensure consistency when computing distances in different regions, this method simultaneously learns all the metrics. As the objective function in MM_LMNN does not have any regularization term, 30% of training data is considered as a validation set so as to avoid overfitting.

In PLML (Wang et al., 2012), for each training example, a local metric is learned as a linear combination of basis metrics $\boldsymbol{M}_{\boldsymbol{b1}}, \boldsymbol{M}_{\boldsymbol{b2}}, \dots, \boldsymbol{M}_{\boldsymbol{bm}}$, where $\boldsymbol{M}_{\boldsymbol{b_i}}$ is a metric associated to the anchor point $\boldsymbol{u_i}$ obtained by the k-means clustering algorithm. The proposed weight learning algorithm tries to learn a metric function that varies smoothly along the data manifold. PLML has numerous hyperparameters to be adjusted. It also requires a full eigenvector decomposition at each stage, making it inappropriate for high dimensional metric learning. To address these issues, SCML (Sparse Compositional Metric Learning) (Shi et al., 2014) uses ran$k$-1 p.s.d matrices as a basis. In other words, metric $\boldsymbol{M_x}$, associated with training example $\boldsymbol{x}$, is defined as:

$$\boldsymbol{M_x} = \sum_{i=1}^{m} w_i \boldsymbol{b}_i \boldsymbol{b}_i^t. \tag{1}$$

The basis set $\boldsymbol{B} = \{\boldsymbol{b}_1, \boldsymbol{b}_2, \dots, \boldsymbol{b}_m\}$ is generated using the Fisher discriminant analysis at several local regions. This approach automatically satisfies the p.s.d constraint. However, to achieve a reasonable performance, the number of bases should be large enough, so that the time of the weight learning algorithm is high for large scale problems. The optimal solution is also limited to a linear combination of the basis elements which are fixed during optimization.

The main limitation in these studied local metric learning methods is the inability of these methods to combine learned metrics in a principled way so as to compute a distance between two new data points. Also, for training points $\boldsymbol{x}, \boldsymbol{x}'$, the distance function is asymmetric and depends on the local metric used in computing the distance.

3. The probabilistic model

As stated, consideration of the structure and distribution of data often yields a better notion of similarity. For example, in Fig. 1, data items $\boldsymbol{x}$ and $\boldsymbol{z}$ are in the same subspace while $\boldsymbol{y}$ belongs to the other subspace. Although, by Euclidean measure, $\boldsymbol{x}$ is closer to $\boldsymbol{y}$ than $\boldsymbol{z}$, $\boldsymbol{x}$ and $\boldsymbol{z}$ are more similar considering the structure and distribution of data. The proposed method aims to define its similarity measure based on this observation. For this purpose, some notations are introduced.

In many real applications, data are generated from latent variables that are near to several low-dimensional subspaces. Let K and p denote the number and maximum dimension of these subspaces respectively. These subspaces can be coded by the k-dimensional hidden variable $\boldsymbol{q}$ via 1-of-K coding. In the current work, the posterior distribution of data is represented by $p(\boldsymbol{q}|\boldsymbol{x})$. For example, assume input data $\boldsymbol{x}$ belongs to latent space $j = 2$ and $K = 5$. The hidden variable $\boldsymbol{q}$ for $\boldsymbol{x}$ can then be represented as $\boldsymbol{q} = [0, 1, 0, 0, 0]^t$ with $p(\boldsymbol{q} = [\boldsymbol{0}, \boldsymbol{1}, \boldsymbol{0}, \boldsymbol{0}, \boldsymbol{0}]|\boldsymbol{x})$ being the probability of latent subspace $j = 2$ given $\boldsymbol{x}$. The low dimensional latent