Learning Compact and Effective Distance Metrics with Diversity Regularization

Pengtao Xie(✉)

Machine Learning Department, Carnegie Mellon University,
5000 Forbes Avenue, Pittsburgh, PA 15213, USA
pengtaox@cs.cmu.edu

Abstract. Learning a proper distance metric is of vital importance for many distance based applications. Distance metric learning aims to learn a set of latent factors based on which the distances between data points can be effectively measured. The number of latent factors incurs a trade-off: a small amount of factors are not powerful and expressive enough to measure distances while a large number of factors cause high computational overhead. In this paper, we aim to achieve two seemingly conflicting goals: keeping the number of latent factors to be small for the sake of computational efficiency, meanwhile making them as effective as a large set of factors. The approach we take is to impose a diversity regularizer over the latent factors to encourage them to be uncorrelated, such that each factor can capture some unique information that is hard to be captured by other factors. In this way, a small amount of latent factors can be sufficient to capture a large proportion of information, which retains computational efficiency while preserving the effectiveness in measuring distances. Experiments on retrieval, clustering and classification demonstrate that a small amount of factors learned with diversity regularization can achieve comparable or even better performance compared with a large factor set learned without regularization.

1 Introduction

In data mining and machine learning, learning a proper distance metric is of vital importance for many distance based tasks and applications, such as retrieval [22], clustering [18] and classification [16]. In retrieval, a better distance measure can help find data entries that are more relevant with the query. In k-means based clustering, data points can be grouped into more coherent clusters if the distance metric is properly defined. In k-nearest neighbor (k-NN) based classification, to find better nearest neighbors, the distances between data samples need to be appropriately measured. All these tasks rely heavily on a good distance measure. Distance metric learning (DML) [3,16,18] takes pairs of data points which are labeled either as similar or dissimilar and learns a distance metric such that similar data pairs will be placed close to each other while dissimilar pairs will be separated apart. While formulated in various ways, most DML approaches choose to learn a Mahalanobis distance $(x-y)^{\mathsf{T}}M(x-y)$, where x, y are d-dimensional feature vectors and $M \in \mathbb{R}^{d\times d}$ is a positive semidefinite matrix to be

A. Appice et al. (Eds.): ECML PKDD 2015, Part I, LNAI 9284, pp. 610–624, 2015.
DOI: 10.1007/978-3-319-23528-8_38

learned. DML can be interpreted as a latent space model. By factorizing M into $M = A^\mathsf{T}A$, the Mahalanobis distance can be written as $\|Ax - Ay\|^2$, which can be interpreted as first projecting the data from the original feature space to a latent space using the linear projection matrix $A \in \mathbb{R}^{k \times d}$, then measuring the squared Euclidean distance in the latent space. Each row of A corresponds to one latent factor (or one dimension of the latent space). Ax is the latent representation of x and can be used as input of downstream tasks. These latent factors are aimed at capturing the latent features of the observed data. The latent features usually carry high-level semantic meanings and reflect the inherent characteristics of data, thus measuring distance in the latent feature space could be more effective.

In choosing the number k of latent factors (or the dimension of the latent space), there is an inherent tradeoff between the effectiveness of the distance matrix A and computational efficiency. A larger k would bestow A more expressiveness and power in measuring distances. However, the resultant latent representations would be of high dimensionality, which incurs high computational complexity and inefficiency. This is especially true for retrieval where performing nearest neighbor search on high-dimensional representations is largely difficult. On the other hand, a smaller k can reduce the computational cost, but would render the distance matrix less effective.

In this paper, we aim to investigate whether it is possible to achieve the best of both worlds: given a sufficiently small k which facilitates computational efficiency, can the effectiveness of the distance matrix be comparable to that of a large k? In other words, the goal is to learn a compact (with small k) but effective distance matrix. Our way to approach this goal is motivated by Principal Component Analysis (PCA). Similar to DML, PCA also learns a linear projection matrix, where the row vectors are called components. Unlike DML which imposes no constraints on the row vectors (latent factors), PCA requires the components to be orthogonal to each other. Such an orthogonality constraint renders the components to be uncorrelated and each component captures information that cannot be captured by other components. As a result, a small number of components are able to capture a large proportion of information. This inspires us to place an orthogonality constraint over the row vectors of A in DML, with the hope that a small number of latent factors are sufficient to effectively measure distances. However, as verified in our experiments, requiring the latent factors to be strictly orthogonal may be too restrictive, which hurts the quality of distance measurement. Instead, we impose a *diversity* regularizer over the latent factors to encourage them to approach orthogonality, but not necessarily to be orthogonal. We perform experiments on retrieval, clustering and classification to demonstrate that with diversity regularization, a distance matrix with small k can achieve comparable or even better performance in comparison with an unregularized matrix with large k.

The rest of the paper is organized as follows. In Section 2, we review related works. In Section 3, we present how to diversity DML. Section 4 gives experimental results and Section 5 concludes the paper.

2 Related Works

Many works [3,5,8,16,18,21] have been proposed for distance metric learning. Please see [15,19] for an extensive review. There are many problem settings regarding the form of distance supervision, the type of distance metric to be learned and the learning objective. Among them, the most common setting [3,5,18] is given data pairs labeled either as similar or dissimilar, learning a Mahalanobis distance metric, such that similar pairs will be placed close to each other and dissimilar pairs will be separated apart. As first formulated in [18], a Mahalanobis distance metric is learned under similarity and dissimilarity constraints by minimizing the distances of similar pairs while separating dissimilar pairs with a certain margin. Guillaumin *et al* [5] proposed to use logistic discriminant to learn a Mahalanobis metric from a set of labelled data pairs, with the goal that positive pairs have smaller distances than negative pairs. Kostinger *et al* [6] proposed to learn Mahalanobis metrics using likelihood test, which defines the Mahalanobis matrix to be the difference of covariance matrices of two Gaussian distributions used for modeling dissimilar pairs and similar pairs respectively. Ying and Li [21] developed an eigenvalue optimization framework for learning a Mahalanobis metric, which is shown to be equivalent to minimizing the maximal eigenvalue of a symmetric matrix.

Some works take other forms of distance supervision such as class labels [14], rankings [10], triple constraints [13], time series alignments [8] to learn distance metrics for specific tasks, such as k-nearest neighbor classification [16], ranking [10], time series aligning [8]. Globerson and Roweis [4] assumed the class label for each sample is available and proposed to learn a Mahalanobis matrix for classification by collapsing all samples in the same class to a single point and pushing samples in other classes infinitely far away. Weinberger *et al* [16] proposed to learn distance metrics for k-nearest neighbor classification with the goal that the k-nearest neighbors always belong to the same class while samples from different classes are separated by a large margin. This method also requires the presence of class labels of all samples. Trivedi *et al* [14] formulated the problem of metric learning for k nearest neighbor classification as a large margin structured prediction problem, with a latent variable representing the choice of neighbors and the task loss directly corresponding to classification error. In this paper, we focus on the pairwise similarity/dissimilarity constraints which are considered as the most common and natural supervision of distance metric learning. Other forms of distance supervision, together with the corresponding specific-purpose tasks, will be left for future study.

To avoid overfitting, various methods have been proposed to regularize distance metric learning. Davis *et al* [3] imposed a regularization that the Mahalanobis distance matrix should be close to a prior matrix and the Bregman divergence is utilized to measure how close two matrices are. Ying and Li [20] and Niu *et al* [11] utilized a mixed-norm regularization to encourage the sparsity of the projection matrix. Qi *et al* [12] used ℓ_1 regularization to learn sparse metrics for high dimensional problems with small sample size. Qian *et al* [13] applied dropout to regularize distance metric learning. In this paper,

Algorithm 1.. Algorithm for solving DDML.

Input: $\mathcal{S}$,$\mathcal{D}$,k,λ
repeat
  Fix $\widetilde{A}$, optimize g using subgradient method
  Fix g, optimize $\widetilde{A}$ using projected subgradient method
until converge

Table 1. Statistics of datasets

	Feature Dim.	#training data	#data pairs
20-News	5000	11.3K	200K
15-Scenes	1000	3.2K	200K
6-Activities	561	7.4K	200K

we investigate a different regularizer for DML: diversity regularization, which has not been studied by previous works.

The problem of diversifying the latent factors in latent variable models has been studied in [7,17], with the goal to reduce the redundancy of latent factors and improve the coverage of infrequent latent features/structures. Zou and Adams [23] used a Determinantal Point Process (DPP) prior to diversify the latent factors and Xie *et al* [17] defined a diversity measure based on pairwise angles between latent factors. In this paper, we study the diversification of distance metric learning, aiming to learn compact distance metrics without compromising their effectiveness.

3 Diversify Distance Metric Learning

In this section, we begin with reviewing the DML problem proposed in [18] and reformulate it as a latent space model using ideas introduced in [16]. Then we present how to diversify DML.

3.1 A Latent Space Modeling View of DML

Distance metric learning represents a family of models and has various formulations regarding the distance metric to learn, the form of distance supervision and how the objective function is defined. Among them, the most popular setting is: (1) distance metric: Mahalanobis distance $(x-y)^{\mathsf{T}}M(x-y)$, where x and y are feature vectors of two data instances and M is a symmetric and positive semidefinite matrix to be learned; (2) the form of distance supervision: pairs of data instances labeled either as similar or dissimilar; (3) learning objective: to learn a distance metric to place similar points as close as possible and separate dissimilar points apart. Given a set of pairs labeled as similar $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{|\mathcal{S}|}$ and a set of pairs labeled as dissimilar $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{|\mathcal{D}|}$, DML learns a Mahalanobis

Table 2. Retrieval average precision (%) on 20-News dataset

k	10	100	300	500	700	900
DML	72.4	74.0	74.9	75.4	75.8	76.2
DDML	**76.7**	**81.0**	**81.1**	**79.2**	**78.3**	**77.8**

Table 3. Retrieval average precision (%) on 15-Scenes dataset

k	10	50	100	150	200
DML	79.5	80.2	80.7	80.7	80.8
DDML	**82.4**	**83.6**	**83.3**	**83.1**	**82.8**

distance matrix M by optimizing the following problem

$$\begin{aligned} &\min_M \ \frac{1}{|\mathcal{S}|} \sum_{(x,y)\in\mathcal{S}} (x-y)^\mathsf{T} M (x-y) \\ &s.t. \quad (x-y)^\mathsf{T} M (x-y) \geq 1, \forall (x,y) \in \mathcal{D} \\ &\quad\ M \succeq 0 \end{aligned} \tag{1}$$

where $M \succeq 0$ denotes that M is required to be positive semidefinite. This optimization problem aims to minimize the Mahalanobis distances between pairs labeled as similar while separating dissimilar pairs with a margin 1. M is required to be positive semidefinite to ensure that the Mahalanobis distance is a valid distance metric.

By re-parametrizing M with $A^\mathsf{T}A$ [16], where A is a matrix of size $k \times d$ ($k \leq d$) and $A^\mathsf{T}A$ guarantees the positive semi-definiteness of M, the Mahalanobis distance can be written as $\|Ax - Ay\|^2$, which can be interpreted as first projecting the data from the original space to a latent space using the linear projection matrix A, then computing the squared Euclidean distance in the latent space. Each row of A corresponds to a latent factor. Accordingly, the problem defined in Eq.(1) can be written as

$$\begin{aligned} &\min_A \ \frac{1}{|\mathcal{S}|} \sum_{(x,y)\in\mathcal{S}} \|Ax - Ay\|^2 \\ &s.t. \quad \|Ax - Ay\|^2 \geq 1, \forall (x,y) \in \mathcal{D} \end{aligned} \tag{2}$$

To this end, we see that the DML problem can be interpreted as a latent space modeling problem. The goal is to seek a latent space where the squared Euclidean distances of similar data pairs are small and those of dissimilar pairs are large. The latent space is characterized by the projection matrix A.

3.2 Diversify DML

To diversify DML, we use the diversity measure proposed in [17] to regularize the latent factors to encourage them to approach orthogonality. Given k latent factors in $A \in R^{k\times d}$, one can compute the non-obtuse angle $\theta(a_i, a_j)$ between each pair of latent factors a_i and a_j, where a_i is the ith row of A. $\theta(a_i, a_j)$

Table 4. Retrieval average precision (%) on 6-Activities dataset

k	10	50	100	150	200
DML	93.2	94.3	94.5	94.5	94.5
DDML	**96.2**	**95.5**	**95.9**	**95.3**	**95.1**

Table 5. Retrieval average precision (%) on three datasets

	20-News	15-Scenes	6-Activities
EUC	62.8	65.3	85.0
DML [18]	76.2	80.8	94.5
LMNN [16]	67.0	70.3	71.5
ITML [3]	74.7	79.1	94.2
DML-eig [21]	71.2	71.3	86.7
Seraph [11]	75.8	82.0	89.2
DDML	**81.1**	**83.6**	**96.2**

is defined as $\arccos(\frac{|a_i \cdot a_j|}{\|a_i\|\|a_j\|})$. A larger $\theta(a_i, a_j)$ indicates that a_i and a_j are more different from each other. Given the pairwise angles, the diversity measure $\Omega(A)$ is defined as $\Omega(A) = \Psi(A) - \Pi(A)$, where $\Psi(A)$ and $\Pi(A)$ are the mean and variance of all pairwise angles. The mean $\Psi(A)$ measures how these factors are different from each other on the whole and the variance $\Pi(A)$ is intended to encourage these factors to be evenly different from each other. The larger $\Omega(A)$ is, the more diverse these latent factors are. And $\Omega(A)$ attains the global maximum when the factors are orthogonal to each other.

Using this diversity measure to regularize the latent factors, we define a Diversified DML (DDML) problem as:

$$\begin{aligned} &\min_A \frac{1}{|\mathcal{S}|} \sum_{(x,y)\in\mathcal{S}} \|Ax - Ay\|^2 - \lambda\Omega(A) \\ &s.t. \quad \|Ax - Ay\|^2 \geq 1, \forall (x, y) \in \mathcal{D} \end{aligned} \tag{3}$$

where $\lambda > 0$ is a tradeoff parameter between the distance loss and the diversity regularizer. The term $-\lambda\Omega(A)$[1] in the new objective function encourages the latent factors in A to be diverse. λ plays an important role in balancing the fitness of A to the distance loss $\sum_{(x,y)\in\mathcal{S}} \|Ax - Ay\|^2$ and its diversity. Under a small λ, A is learned to best minimize the distance loss and its diversity is ignored. Under a large λ, A is learned with high diversity, but may not be well fitted to the distance loss and hence lose the capability to properly measure distances. A proper λ needs to be chosen to achieve the optimal balance.

3.3 Optimization

In this section, we present an algorithm to solve the problem defined in Eq.(3), which is summarized in Algorithm 1. First, we adopt a strategy similar to [16]

[1] Note that a negative sign is used here because the overall objective function is to be minimized but $\Omega(A)$ is intended to be maximized.

Table 6. Clustering accuracy (%) on 20-News dataset

k	10	100	300	500	700	900
DML	23.7	25.1	26.2	26.9	28.1	28.4
DDML	**33.4**	**42.7**	**44.6**	**39.5**	**40.6**	**41.3**

Table 7. Normalized mutual information (%) on 20-News dataset

k	10	100	300	500	700	900
DML	34.1	35.4	36.8	36.9	38.0	38.2
DDML	**42.5**	**49.7**	**51.1**	**47.2**	**47.8**	**48.1**

to remove the constraints. By introducing slack variables ξ to relax the hard constraints, we get

$$\begin{aligned} &\min_A \tfrac{1}{|\mathcal{S}|} \sum_{(x,y)\in\mathcal{S}} \|Ax - Ay\|^2 - \lambda\Omega(A) + \tfrac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} \xi_{x,y} \\ &s.t. \quad \|Ax - Ay\|^2 \geq 1 - \xi_{x,y}, \xi_{x,y} \geq 0, \forall (x,y) \in \mathcal{D} \end{aligned} \tag{4}$$

Using hinge loss, the constraint in Eq.(4) can be further eliminated

$$\begin{aligned} \min_A \tfrac{1}{|\mathcal{S}|} &\sum_{(x,y)\in\mathcal{S}} \|Ax - Ay\|^2 - \lambda\Omega(A) \\ &+ \tfrac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} \max(0, 1 - \|Ax - Ay\|^2) \end{aligned} \tag{5}$$

Since $\Omega(A)$ is non-smooth and non-convex, which is hard to optimize directly, Xie *et al* [17] instead optimized a low bound of $\Omega(A)$, and they proved that maximizing the lower bound can increase $\Omega(A)$. Factorizing A into $\mathrm{diag}(g)\widetilde{A}$, where g is a vector and g_i denotes the ℓ_2 norm of the ith row of A, then the ℓ_2 norm of each row vector in $\widetilde{A}$ is one. According to the definition of $\Omega(A)$, it is clear that $\Omega(A) = \Omega(\widetilde{A})$. The problem defined in Eq.(5) can be reformulated as

$$\begin{aligned} \min_{\widetilde{A},g} \tfrac{1}{|\mathcal{S}|} &\sum_{(x,y)\in\mathcal{S}} \|\mathrm{diag}(g)\widetilde{A}(x - y)\|^2 - \lambda\Omega(\widetilde{A}) \\ &+ \tfrac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} \max(0, 1 - \|\mathrm{diag}(g)\widetilde{A}(x - y)\|^2) \\ s.t. \quad &\|\widetilde{A}_i\| = 1, \forall i = 1, \cdots, k \end{aligned} \tag{6}$$

where $\widetilde{A}_i$ denotes the ith row of $\widetilde{A}$. This problem can be optimized by alternating between g and $\widetilde{A}$: optimizing g with $\widetilde{A}$ fixed and optimizing $\widetilde{A}$ with g fixed. With $\widetilde{A}$ fixed, the problem defined over g is

$$\min_g \tfrac{1}{|\mathcal{S}|} \sum_{(x,y)\in\mathcal{S}} \|\mathrm{diag}(g)\widetilde{A}(x - y)\|^2 + \tfrac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} \max(0, 1 - \|\mathrm{diag}(g)\widetilde{A}(x - y)\|^2) \tag{7}$$

Table 8. Clustering accuracy (%) on 15-Scenes dataset

k	10	50	100	150	200
DML	33.9	36.5	40.1	37.0	37.8
DDML	**46.9**	**51.3**	**46.2**	**46.5**	**49.6**

Table 9. Normalized mutual information (%) on 15-Scenes dataset

k	10	50	100	150	200
DML	41.4	41.0	42.0	41.4	41.6
DDML	**46.7**	**48.9**	**47.3**	**48.8**	**47.1**

which can be optimized with subgradient method. Fixing g, the problem defined over $\widetilde{A}$ is

$$\begin{aligned} \min_{\widetilde{A}} \ & \tfrac{1}{|\mathcal{S}|} \sum_{(x,y)\in\mathcal{S}} \|\mathrm{diag}(g)\widetilde{A}(x-y)\|^2 - \lambda\Omega(\widetilde{A}) \\ & + \tfrac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} \max(0, 1 - \|\mathrm{diag}(g)\widetilde{A}(x-y)\|^2) \\ s.t. \quad & \|\widetilde{A}_i\| = 1, \forall i = 1, \cdots, k \end{aligned} \tag{8}$$

Since $\Omega(\widetilde{A})$ is non-smooth and non-convex, (sub)gradient method is not applicable. Xie *et al* [17] derived a smooth lower bound of $\Omega(\widetilde{A})$ and instead optimized the low bound with projected gradient descent. Please refer to [17] for details.

4 Experiments

In this section, on three tasks — retrieval, clustering and classification — we corroborate that through diversification it is possible to learn distance metrics that are both compact and effective.

4.1 Datasets

We used three datasets in the experiments: 20 Newsgroups[2] (20-News), 15-Scenes [9] and 6-Activities [1]. The 20-News dataset has 18846 documents from 20 categories, where 60% of the documents were for training and the rest were for testing. Documents were represented with *tfidf* vectors whose dimensionality is 5000. We randomly generated 100K similar pairs and 100K dissimilar pairs from the training set to learn distance metrics. Two documents were labeled as similar if they belong to the same group and dissimilar otherwise. The 15-Scenes dataset contains 4485 images belonging to 15 scene classes. 70% of the images were used for training and 30% were for testing. Images were represented with bag-of-words vectors whose dimensionality is 1000. Similar to 20-News, we generated 100K similar and 100K dissimilar data pairs for distance learning according to whether two images are from the same scene class or not. The 6-Activities

[2] http://qwone.com/~jason/20Newsgroups/

Table 10. Clustering accuracy (%) on 6-Activities dataset

k	10	50	100	150	200
DML	75.0	75.6	76.1	75.6	75.7
DDML	**94.9**	**96.3**	**96.6**	**95.1**	**95.7**

Table 11. Normalized mutual information (%) on 6-Activities dataset

k	10	50	100	150	200
DML	83.6	83.5	84.0	83.5	83.5
DDML	**90.3**	**91.9**	**91.3**	**91.4**	**91.1**

dataset is built from recordings of 30 subjects performing six activities of daily living while carrying a waist-mounted smart phone with embedded inertial sensors. The features are 561-dimensional sensory signals. There are 7352 training instances and 2947 testing instances. Similarly, 100K similar pairs and 100K dissimilar pairs were generated to learn distance metrics. Table 1 summarizes the statistics of these three datasets.

4.2 Experimental Settings

Our method DDML contains two key parameters — the number k of latent factors and the tradeoff parameter λ — both of which were tuned using 5-fold cross validation. We compared with 6 baseline methods, which were selected according to their popularity and the state of the art performance. They are: (1) Euclidean distance (EUC); (2) Distance Metric Learning (DML) [18]; (3) Large Margin Nearest Neighbor (LMNN) metric learning [16]; (4) Information Theoretical Metric Learning (ITML) [3]; (5) Distance Metric Learning with Eigenvalue Optimization (DML-eig) [21]; (6) Information-theoretic Semi-supervised Metric Learning via Entropy Regularization (Seraph) [11]. Parameters of the baseline methods were tuned using 5-fold cross validation. Some methods, such as ITML, achieve better performance on lower-dimensional representations which are obtained via Principal Component Analysis. The number of leading principal components were selected via 5-fold cross validation.

4.3 Retrieval

We first applied the learned distance metrics for retrieval. To evaluate the effectiveness of the learned metrics, we randomly sampled 100K similar pairs and 100K dissimilar pairs from 20-News test set, 50K similar pairs and 50K dissimilar pairs from 15-Scenes test set, 100K similar pairs and 100K dissimilar pairs from 6-Activities test set and used the learned metrics to judge whether these pairs were similar or dissimilar. If the distance was greater than some threshold t, the pair was regarded as similar. Otherwise, the pair was regarded as dissimilar. We used average precision (AP) to evaluate the retrieval performance.

Table 12. Clustering accuracy (%) on three datasets

	20-News	15-Scenes	6-Activities
EUC	36.5	29.0	61.6
DML [18]	28.4	40.1	76.1
LMNN [16]	32.9	33.6	56.9
ITML [3]	34.5	38.2	93.4
DML-eig [21]	27.3	26.6	63.3
Seraph [11]	**48.1**	48.2	74.8
DDML	44.6	**51.3**	**96.6**

Table 13. Normalized mutual information (%) on three datasets

	20-News	15-Scenes	6-Activities
EUC	37.9	28.7	59.9
DML [18]	38.2	42.0	83.6
LMNN [16]	33.3	34.3	58.2
ITML [3]	39.2	41.5	87.0
DML-eig [21]	34.0	31.8	58.6
Seraph [11]	49.7	47.5	71.1
DDML	**51.1**	**48.9**	**91.9**

Table 2, 3 and 4 show the average precision under different number k of latent factors on 20-News, 15-Scenes and 6-Activities dataset respectively. As shown in these tables, DDML with a small k can achieve retrieval precision that is comparable to DML with a large k. For example, on the 20-News dataset (Table 2), with 10 latent factors, DDML is able to achieve a precision of 76.7%, which cannot be achieved by DML with even 900 latent factors. As another example, on the 15-Scenes dataset (Table 3), the precision obtained by DDML with $k = 10$ is 82.4%, which is largely better than the 80.8% precision achieved by DML with $k = 200$. Similar behavior is observed on the 6-Activities dataset (Table 4). This demonstrates that, with diversification, DDML is able to learn a distance metric that is as effective as (if not more effective than) DML, but is much more compact than DML. Such a compact distance metric greatly facilitates retrieval efficiency. Performing retrieval on 10-dimensional latent representations is much easier than on representations with hundreds of dimensions. It is worth noting that the retrieval efficiency gain comes without sacrificing the precision, which allows one to perform fast and accurate retrieval. For DML, increasing k consistently increases the precision, which corroborates that a larger k would make the distance metric to be more expressive and powerful. However, k cannot be arbitrarily large, otherwise the distance matrix would have too many parameters that lead to overfitting. This is evidenced by how the precision of DDML varies as k increases.

Table 5 presents the comparison with the state of the art distance metric learning methods. As can be seen from this table, our method achieves the best performances across all three datasets. The Euclidean distance does not

Table 14. 3-NN accuracy (%) on 20-News dataset

k	10	100	300	500	700	900
DML	39.1	48.0	53.0	55.0	56.4	57.5
DDML	**51.3**	**64.1**	**64.5**	**63.3**	**62.9**	**61.4**

Table 15. 10-NN accuracy (%) on 20-News dataset

k	10	100	300	500	700	900
DML	39.4	49.4	54.3	56.2	57.9	58.6
DDML	**54.2**	**66.6**	**66.8**	**66.1**	**65.3**	**64.5**

Table 16. 3-NN accuracy (%) on 15-Scenes dataset

k	10	50	100	150	200
DML	47.7	47.7	50.8	51.7	51.1
DDML	**57.4**	**57.5**	**57.9**	**58.8**	**57.3**

Table 17. 10-NN accuracy (%) on 15-Scenes dataset

k	10	50	100	150	200
DML	51.6	51.7	54.0	54.4	54.9
DDML	**59.2**	**60.9**	**60.5**	**60.6**	**59.6**

incorporate distance supervision provided by human, thus its performance is inferior. DML-eig imposes no regularization over the distance metric, which is thus prone to overfitting. To avoid overfitting, ITML utilized a Bregman divergence regularizer and Seraph used a sparsity regularizer. But the performances of both regularizers are inferior to the diversity regularizer utilized by DDML. LMNN is specifically designed for k-NN classification, thus the learned distance metrics cannot guarantee to be effective in retrieval tasks.

4.4 Clustering

The second task we study is to apply the learned distance metrics for k-means clustering, where the number of clusters was set to the number of categories in each dataset and k-means was run 10 times with random initialization of the centroids. Following [2], we used two metrics to measure the clustering performance: accuracy (AC) and normalized mutual information (NMI). Please refer to [2] for their definitions.

Table 6,8 and 10 show the clustering accuracy on 20-News, 15-Scenes and 6-Activity test set respectively under various number of latent factors k. Table 7, 9 and 11 show the normalized mutual information on 20-News, 15-Scenes and 6-Activity test set respectively. These tables show that the clustering performance achieved by DDML under a small k is much better than DML under a much larger k. For instance, DDML can achieve 33.4% accuracy on the 20-News dataset (Table 6) with 10 latent factors, which is much better than the 28.4% accuracy

Table 18. 3-NN accuracy (%) on 6-Activities dataset

k	10	50	100	150	200
DML	94.9	94.8	94.6	95.1	95.0
DDML	**94.3**	**96.2**	**96.5**	**95.5**	**95.9**

Table 19. 10-NN accuracy (%) on 6-Activities dataset

	10	50	100	150	200
DML	95.3	95.0	95.2	95.2	95.3
DDML	**96.6**	**96.8**	**96.4**	**96.3**	**96.1**

Table 20. 3-NN accuracy (%) on three datasets

	20-News	15-Scenes	6-Activities
EUC	42.6	42.5	88.7
DML [18]	57.5	51.7	95.1
LMNN [16]	60.6	53.5	91.5
ITML [3]	50.9	51.9	93.5
DML-eig [21]	39.2	33.1	82.3
Seraph [11]	**67.9**	55.2	91.4
DDML	64.5	**58.8**	**96.5**

Table 21. 10-NN accuracy (%) on three datasets

k	20-News	15-Scenes	6-Activities
EUC	41.7	44.9	90.2
DML [18]	58.6	54.9	95.3
LMNN [16]	62.7	56.2	91.5
ITML [3]	54.8	54.3	94.0
DML-eig [21]	43.8	34.0	82.8
Seraph [11]	**69.8**	60.3	92.5
DDML	66.8	**60.9**	**96.8**

obtained by DML with 900 latent factors. As another example, the NMI obtained by DDML on the 15-Scenes dataset (Table 9) with $k = 10$ is 46.7%, which is largely better than the 41.6% NMI achieved by DML with $k = 200$. This again corroborates that the diversity regularizer can enable DDML to learn compact and effective distance metrics, which significantly reduce computational complexity while preserving the clustering performance.

Table 12 and 13 present the comparison of DDML with the state of the art methods on clustering accuracy and normalized mutual information. As can be seen from these two tables, our method outperforms the baselines in most cases except that the accuracy on 20-News dataset is worse than the Seraph method. Seraph performs very well on 20-News and 15-Scenes dataset, but its performance is bad on the 6-Activities dataset. DDML achieves consistently good performances across all three datasets.

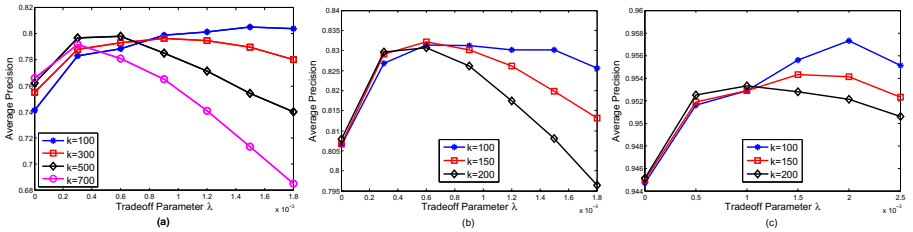

Fig. 1. Sensitivity of DDML to the tradeoff parameter λ on (a) 20-News dataset (b) 15-Scenes dataset (c) 6-Activities dataset

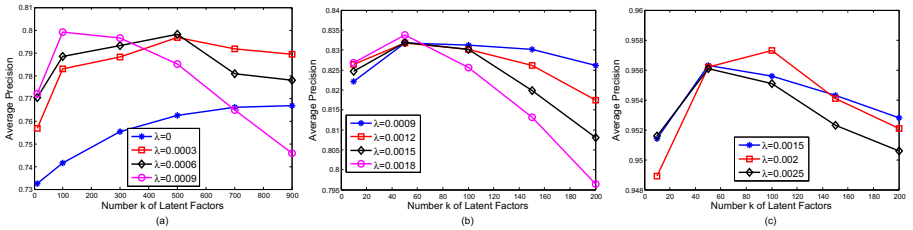

Fig. 2. Sensitivity of DDML to the number of latent factors k on (a) 20-News dataset (b) 15-Scenes dataset (c) 6-Activities dataset

4.5 Classification

We also apply the learned metrics for k-nearest neighbor classification, which is also an algorithm that largely depends on a good distance measure. For each testing sample, we find its k-nearest neighbors in the training set and use the class labels of the nearest neighbors to classify the test sample. Table 14, 16 and 18 show the 3-NN classification accuracy on the 20-News, 15-Scenes and 6-Activities dataset. Table 15, 17 and 19 show the 10-NN classification accuracy on the 20-News, 15-Scenes and 6-Activities dataset. Similar to retrieval and clustering, DDML with a small k can achieve classification accuracy that is comparable to or better than DML with a large k. Table 20 and 21 present the comparison of DDML with the state of the art methods on 3-NN and 10-NN classification accuracy. As can be seen from these two tables, our method outperforms the baselines in most cases except that the accuracy on 20-News dataset is worse than the Seraph method.

4.6 Sensitivity to Parameters

We study the sensitivity of DDML to the two key parameters: tradeoff parameter λ and the number of latent factors k. Figure 1 shows how the retrieval average precision (AP) varies as λ increases on the 20-News, 15-Scenes and 6-Activities dataset respectively. The curves correspond to different k. As can be seen from the figure, initially increasing λ improves AP. The reason is that a larger λ

encourages the latent factors to be more uncorrelated, thus different aspects of the information can be captured more comprehensively. However, continuing to increase λ degrades the precision. This is because if λ is too large, the diversify regularizer dominates the distance loss and the resultant distance metric is not tailored to the distance supervision and loses effectiveness in measuring distances.

Figure 2 shows how AP varies as k increases on the 20-News, 15-Scenes and 6-Activities dataset respectively. The curves correspond to different λ. When k is small, the average precision is low. This is because a small amount of latent factors are insufficient to capture the inherent complex pattern behind data, hence lacking the capability to effectively measure distances. As k increases, the model capacity increases and the AP increases accordingly. However, further increasing k causes performance to drop. This is because a larger k incurs higher risk of overfitting to training data.

5 Conclusions

In this paper, we study the problem of diversifying distance metric learning, with the purpose to learn compact distance metrics without losing their effectiveness in measuring distances. Diversification encourages the latent factors in the distance metric to be different from each other, thus each latent factor is able to capture some unique information that is hard to be captured by other factors. Accordingly, the number of latent factors required to capture the total information can be greatly reduced. Experiments on retrieval, clustering and classification corroborate the effectiveness of the diversity regularizer in learning compact and effective distance metrics.

References

1. Anguita, D., Ghio, A., Oneto, L., Parra, X., Reyes-Ortiz, J.L.: Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine. In: Ambient Assisted Living and Home Care, pp. 216–223. Springer (2012)
2. Cai, D., He, X., Han, J.: Locally consistent concept factorization for document clustering. IEEE Transactions on Knowledge and Data Engineering **23**(6), 902–913 (2011)
3. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.: Information-theoretic metric learning. In: Proceedings of the 24th International Conference on Machine Learning, pp. 209–216. ACM (2007)
4. Globerson, A., Roweis, S.T.: Metric learning by collapsing classes. In: Advances in Neural Information Processing Systems, pp. 451–458 (2005)
5. Guillaumin, M., Verbeek, J., Schmid, C.: Is that you? metric learning approaches for face identification. In: IEEE International Conference on Computer Vision, pp. 498–505. IEEE (2009)
6. Kostinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H.: Large scale metric learning from equivalence constraints. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2288–2295. IEEE (2012)

7. Kwok, J.T., Adams, R.P.: Priors for diversity in generative latent variable models. In: Advances in Neural Information Processing Systems, pp. 2996–3004 (2012)
8. Lajugie, R., Garreau, D., Bach, F., Arlot, S.: Metric learning for temporal sequence alignment. In: Advances in Neural Information Processing Systems, pp. 1817–1825 (2014)
9. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 2169–2178. IEEE (2006)
10. Lim, D., Lanckriet, G., McFee, B.: Robust structural metric learning. In: Proceedings of The 30th International Conference on Machine Learning, pp. 615–623 (2013)
11. Niu, G., Dai, B., Yamada, M., Sugiyama, M.: Information-theoretic semisupervised metric learning via entropy regularization. Neural Computation, 1–46 (2012)
12. Qi, G.-J., Tang, J., Zha, Z.-J., Chua, T.-S., Zhang, H.-J.: An efficient sparse metric learning in high-dimensional space via l1-penalized log-determinant regularization. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 841–848. ACM (2009)
13. Qian, Q., Hu, J., Jin, R., Pei, J., Zhu, S.: Distance metric learning using dropout: a structured regularization approach. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 323–332. ACM (2014)
14. Trivedi, S., Mcallester, D., Shakhnarovich, G.: Discriminative metric learning by neighborhood gerrymandering. In: Advances in Neural Information Processing Systems, pp. 3392–3400 (2014)
15. Wang, F, Sun, J.: Survey on distance metric learning and dimensionality reduction in data mining. Data Mining and Knowledge Discovery, 1–31 (2014)
16. Weinberger, KQ., Blitzer, J., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. In: Advances in Neural Information Processing Systems, pp. 1473–1480 (2005)
17. Xie, P., Deng, Y., Xing, E.P.: Diversifying restricted boltzmann machine for document modeling. In: ACM SIGKDD Conference on Knowledge Discovery and Data Mining (2015)
18. Xing, E.P., Jordan, M.I., Russell, S., Ng, A.Y.: Distance metric learning with application to clustering with side-information. In: Advances in Neural Information Processing Systems, pp. 505–512 (2002)
19. Liu, Y., Rong, J.: Distance metric learning: A comprehensive survey. Michigan State Universiy, vol. 2 (2006)
20. Ying, Y., Huang, K., Campbell, C.: Sparse metric learning via smooth optimization. In: Advances in Neural Information Processing Systems, pp. 2214–2222 (2009)
21. Ying, Y., Li, P.: Distance metric learning with eigenvalue optimization. The Journal of Machine Learning Research **13**(1), 1–26 (2012)
22. Zhang, P., Zhang, W., Li, W.-J., Guo, M: Supervised hashing with latent factor models. In: SIGIR (2014)
23. Zou, J.Y., Adams, R.P.: Priors for diversity in generative latent variable models. In: Advances in Neural Information Processing Systems, pp. 2996–3004 (2012)