2014 IEEE 28-th Convention of Electrical and Electronics Engineers in Israel

Metric Learning Using Labeled and Unlabeled Data for Semi-Supervised/Domain Adaptation Classification

Hadas Benisty and Koby Crammer
Department of Electrical Engineering
Technion, Israel Institute of Technology
Technion City, Haifa 32000, Israel
{hadasbe@tx,koby@ee}.technion.ac.il

***Abstract*—Metric learning includes a wide range of algorithms aiming to improve the classification accuracy by capturing the spatial structure of the training set. The performance of those (supervised) methods greatly depends on the amount of labeled data available for training. In practice, however, it is usually not easy to obtain a large-scale labeled set, as opposed to an unlabeled one.**

In this paper we propose a new method for metric learning using a small-scale labeled set and a large-scale unlabeled set. This method can be applied in two setups - a Semi-Supervised (SS) classification setup and a Domain Adaptation (DA) setup. We used two sources of hand-written digits images to demonstrate the performance of our proposed method. We show that in both SS and DA setups, the proposed method leads to fewer classification errors compared to Euclidean distance and to Large Margin Nearest Neighbor (LMNN).

A. Introduction

The K-Nearest Neighbor (KNN) classifier is a simple, data-driven classifier, which assumes that the true labeling function is locally continuous, i.e., close vectors share the same label, whereas distant vectors may not. There is no training process; a test vector is classified according to the majority vote obtained using its K-nearest neighbors from among the training vectors. In many practical cases the continuity assumption is not sustained w.r.t the Euclidean metric, but it may be sustained w.r.t a different, unknown metric.

Metric learning algorithms aim to capture the spatial structure of the feature vectors and learn an alternative distance measure, leading to improved classification accuracy. Unsupervised methods such as Principal Component Analysis (PCA), and other dimensionality reduction algorithms can be viewed as methods for learning an equivalent metric using a mapping to a low dimensional space [1]. Some supervised methods use labeled training sets for casting the learning task as an optimization problem. For example, maximizing the sum of distances between dissimilar pairs and diminishing the sum of distances between pairs having similar labels [2]. Another recently proposed approach deals with triplets composed of a similar pair and a third differently labeled vector. The metric is learned such that the distance between the similar pair would be smaller than the distance to the dissimilar vector [3]. One of the best performing methods for metric learning is Large Margin Nearest Neighbor (LMNN) classification [4]. It is based on minimizing the distance between each sample and its KNN having the same label, and also sustaining a margin between the target neighbors and all other differently labeled vectors.

Successful training of supervised methods often requires having a large-scale training set. However, large-scale labeled data is not always available, whereas unlabeled data is usually easily obtained by simply sampling more data points. Moreover, sometimes the available labeled set is sampled from a different distribution than the test vectors, leading to discrepancies and biased results. In this paper we propose a new method for metric learning using a small labeled set and a large-scale unlabeled set. This method can be applied in two setups: 1) a Semi-Supervised (SS) setup - the training set (consisting labeled and unlabeled vectors) and the test set are sampled from the same distribution 2) a Domain Adaptation (DA) setup - the unlabeled training set and testing set are sampled from the same distribution, which is, however, different from the distribution of the labeled training vectors. We introduce the notion of *unlabeled neighbors* that is the set of all KNN of a certain labeled vector, from among the unlabeled set. We assume that the unlabeled set is more dense than the labeled one (since it is much larger) and thus require that the unlabeled neighbors of each labeled vector should stay close w.r.t the new metric. In addition we require that a margin between the K-unlabeled neighbors and all other differently labeled vectors is sustained. Our overall optimized objective is a weighted sum of two terms: one - considering the unlabeled vectors as described above, and the other - considering the labeled set similarly to LMNN.

We used two sources of hand-written digits images to demonstrate the performance of our proposed method (MNIST[1], and USPS[2]). We showed that in both SS and DA

[1] Available at http://yann.lecun.com/exdb/mnist/

[2] Available at http://www.cs.nyu.edu/ roweis/data.html

setups, our proposed method leads to a higher classification accuracy compared to Euclidean distance and to LMNN.

I. Problem Preliminaries

Let $\mathcal{X} \in \mathbb{R}^d$ and $\mathcal{Y} = (0, 1)$ be feature and label domains, correspondingly. Let $S^0 = \left(\mathbf{x}_1^0, y_1^0\right), ..., \left(\mathbf{x}_m^0, y_m^0\right)$ be a labeled training set of examples in $\mathcal{X} \times \mathcal{Y}$ sampled i.i.d. according to a distribution $\mathcal{D}^0$ over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{D}_x^0$ is the induced marginal distribution over $\mathcal{X}$.

A. Source Classification

Given a test point $\tilde{\mathbf{x}}^0$ drawn from the induced marginal distribution $\mathcal{D}_x^0$, its k-NN in S^0 are determined w.r.t. a distance metric $d_{\mathbf{M}}\left(\mathbf{x}_i - \mathbf{x}_j\right) = \left(\mathbf{x}_i - \mathbf{x}_j\right)^\top \mathbf{M} \left(\mathbf{x}_i - \mathbf{x}_j\right)$, parameterized by a matrix $\mathbf{M}$ is:

$$\widehat{\mathcal{N}}_k \left(\tilde{\mathbf{x}}^0; \mathbf{M}, S^0\right) = \{\mathbf{x}_1^{0*}, \ldots, \mathbf{x}_k^{0*}\}, \tag{1}$$

where,

$$\mathbf{x}_1^{0*} = \underset{\mathbf{x}^0 \in S^0}{\text{argmin}}\ d_{\mathbf{M}} \left(\tilde{\mathbf{x}}^0 - \mathbf{x}^0\right), \tag{2}$$

$$\mathbf{x}_2^{0*} = \underset{\mathbf{x}^0 \in S^0 \setminus \{\mathbf{x}_1^{0*}\}}{\text{argmin}}\ d_{\mathbf{M}} \left(\tilde{\mathbf{x}}^0 - \mathbf{x}^0\right), \tag{3}$$

and so on for $\mathbf{x}_3^{0*}, ..., \mathbf{x}_k^{0*}$. The label $\tilde{y}^0$ of the test point, $\tilde{\mathbf{x}}^0$, is estimated, using the k-NN rule, as the majority decision of the k-NN: $\hat{y}^0 = \text{MAJ}\left(y_1^{0*}, \ldots, y_k^{0*}\right)$.

In a supervised setup, a matrix $\mathbf{M}$ is learned by minimizing a cost function that uses the labeled training set, S^0, drawn from $\mathcal{D}^0$. In a semi-supervised (SS) setup, an unlabeled set $\{\mathbf{x}_i^{'0}\}_{i=1}^n$ (drawn from the induced marginal distribution $\mathcal{D}_x^0$), is also considered for learning the matrix $\mathbf{M}$.

B. Target Classification Using Domain Adaptation (DA)

Let $\tilde{\mathbf{x}}^1$ be a test point drawn from a target marginal distribution $\mathcal{D}_x^1$ (of $\mathcal{D}^1$ over $\mathcal{X} \times \mathcal{Y}$). In this setup, the goal is to learn a matrix $\mathbf{M}$ for classification of a target test point $\tilde{\mathbf{x}}^1$, given a labeled set S^0 (drawn from a source distribution) and an unlabeled set $\{\mathbf{x}_i^1\}_{i=1}^n$ drawn from the target marginal distribution $\mathcal{D}_x^1$. The k-NN of $\tilde{\mathbf{x}}^1$ from among the (labeled) training set S^0 are determined w.r.t. a metric $d_{\mathbf{M}}(\cdot, \cdot)$:

$$\widehat{\mathcal{N}}_k \left(\tilde{\mathbf{x}}^1; \mathbf{M}, S^0\right) = \{\mathbf{x}_1^{0*}, \ldots, \mathbf{x}_k^{0*}\} \tag{4}$$

where $\left\{\mathbf{x}_1^{0*}, ..., \mathbf{x}_k^{0*}\right\}$ have the same form as in eqns. (2) and (3), replacing $\tilde{\mathbf{x}}^0$ with $\tilde{\mathbf{x}}^1$. The label $\tilde{y}^1$ of the test point, $\tilde{\mathbf{x}}^1$, is estimated using the k-NN rule, as the majority decision of the k-NN: $\hat{y}^1 = \text{MAJ}\left(y_1^{0*}, \ldots, y_k^{0*}\right)$.

In a DA setup, the source and target distributions are assumed to share some of the features affecting the true labeling of the data. Therefore, the labeled (source) set is used for learning the relevance of the features to the true labeling function, and the unlabeled (target) set is used for adaptating to the spatial structure of the target.

II. Distance Metric Learning for LMNN

A metric $d_{\mathbf{M}}(\cdot, \cdot)$ can be written as a Euclidean distance in a transformed space, defined by a matrix $\mathbf{L} \in \mathbb{R}^{d \times d}$ such that $\mathbf{M} = \mathbf{L}^\top \mathbf{L}$:

$$\begin{aligned} d_{\mathbf{M}}\left(\mathbf{x}_i - \mathbf{x}_j\right) &= \left(\mathbf{x}_i - \mathbf{x}_j\right)^\top \mathbf{L}^\top \mathbf{L} \left(\mathbf{x}_i - \mathbf{x}_j\right) = \\ &= \left\|\mathbf{L}\left(\mathbf{x}_i - \mathbf{x}_j\right)\right\|^2 \end{aligned} \tag{5}$$

LMNN is a method for supervised learning of a linear transformation $\mathbf{L}$, used for k-NN (source) classification [4]. Given a labeled set S^0, a linear transformation $\mathbf{L} \in \mathbb{R}^{d \times d}$ is obtained by minimizing the following objective:

$$\begin{aligned} \epsilon_{\text{LMNN}}(\mathbf{L}) &= \omega_1 \sum_{i,j} \eta_{i,j} \left\|\mathbf{L}\left(\mathbf{x}_i^0 - \mathbf{x}_j^0\right)\right\|^2 + \\ &+ \left(1 - \omega_1\right) \sum_{i,j,l} \eta_{i,j} \left(1 - y_{i,l}\right) [1 + \\ &+ \left\|\mathbf{L}\left(\mathbf{x}_i^0 - \mathbf{x}_j^0\right)\right\|^2 - \left\|\mathbf{L}\left(\mathbf{x}_i^0 - \mathbf{x}_l^0\right)\right\|^2\Big]_+ \end{aligned} \tag{6}$$

where ω_1 is a parameter, and $y_{i,l} = 1$ if $y_i^0 = y_l^0$ and $y_{i,l} = 0$, otherwise. The parameter $\eta_{i,j} = \{0, 1\}$, indicates whether $\mathbf{x}_j^0$ is one of the *K-labeled neighbors* of $\mathbf{x}_i^0$, meaning that $\mathbf{x}_j^0$ is one of K closest neighbors of $\mathbf{x}_i^0$, from among all the samples in S^0 having the same label as $\mathbf{x}_i^0$.

By minimizing the cost in eqn. (6) the distance between each vector and its K-labeled neighbors is minimized, while any other vector having a dissimilar label is "pushed" further beyond a margin, as illustrated in Fig. 1. This method is very effective, since it targets the distance values (as opposed to their sum, for example, as suggested by Xing et al. [2]), and also, since it considers just the closest neighbors of each vector and not the entire training set.

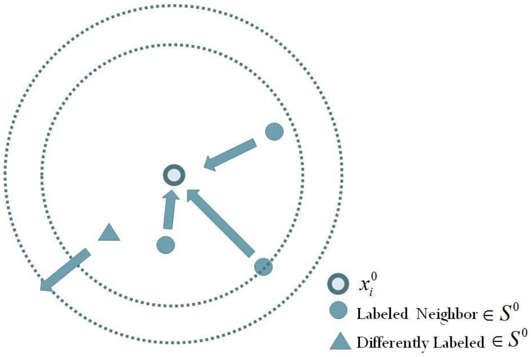

Fig. 1. LMNN - The metric is learned such that the K-labeled neighbors would be closer to $\mathbf{x}_i^0$ than any other dissimilar vectors [4].

III. Metric Learning Using Labeled and Unlabeled Data

In this section we propose a new approach for metric learning using labeled and unlabeled data. We assume that the unlabeled set is much larger than the labeled one and therefore much more dense. Our proposed approach uses the labeled set as used in LMNN. The dense unlabeled set is used for capturing the spatial structure of the data by requiring that the close unlabeled neighbors of each labeled vector remain close w.r.t. the learned metric, as illustrated in Fig. 2.

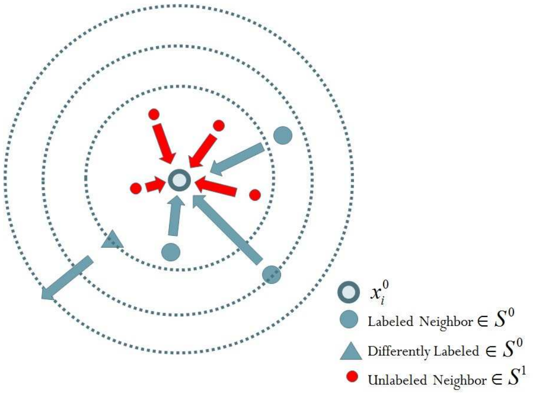
x_i^0
Labeled Neighbor $\in S^0$
Differently Labeled $\in S^0$
Unlabeled Neighbor $\in S^1$

Fig. 2. The proposed approach for metric learning using labeled and unlabeled sets.

A. *Semi-Supervised Classification (SSC)*

Given a labeled set $S^0 = \left(\mathbf{x}_1^0, y_1^0\right), ..., \left(\mathbf{x}_m^0, y_m^0\right)$ drawn from a distribution $\mathcal{D}_x^0$, and an unlabeled set $\{\mathbf{x}_i^{'0}\}_{i=1}^n$ drawn from the marginal distribution $\mathcal{D}_x^0$, a linear transformation $\mathbf{L} \in \mathbb{R}^{d \times d}$ is obtained by minimizing the following objective:

$$\epsilon\left(\mathbf{L}\right) = \left(1 - \omega_3\right)\epsilon_{\text{LMNN}}\left(\mathbf{L}\right) + \omega_3 \epsilon_{\text{SSC}}\left(\mathbf{L}\right), \tag{7}$$

where ϵ_{LMNN} is defined above in eqn. (6) and:

$$\begin{aligned}\epsilon_{\text{SSC}}\left(\mathbf{L}\right) &= \omega_2 \sum_{i,j} \eta_{i,j}^0 \left\|\mathbf{L}\left(\mathbf{x}_i^0 - \mathbf{x}_j^{'0}\right)\right\|^2 + \\ &+ \left(1-\omega_2\right)\sum_{i,j,l}\eta_{i,j}^0\left(1 - y_{i,l}\right)\Big[1 + \\ &+ \left\|\mathbf{L}\left(\mathbf{x}_i^0 - \mathbf{x}_j^{'0}\right)\right\|^2 - \left\|\mathbf{L}\left(\mathbf{x}_i^0 - \mathbf{x}_l^0\right)\right\|^2\Big]_+ \end{aligned} \tag{8}$$

and where $y_{i,l}$ is defined the same as in LMNN (sec. II). The term $\eta_{i,j}^0 = \{0, 1\}$ indicates whether $\mathbf{x}_j^{'0}$ is an *unlabeled neighbor* of $\mathbf{x}_i^0$, meaning that $\mathbf{x}_j^{'0}$ is one of the k-NN of $\mathbf{x}_i^0$ from among the unlabeled set $\{\mathbf{x}_i^{'0}\}_{i=1}^n$, and that $\mathbf{x}_j^{'0}$ is closer to $\mathbf{x}_i^0$ than all $\mathbf{x}_l^0$ having a different label than y_i^0, i.e.

$$\eta_{i,j}^0 = \begin{cases} 1 & \begin{gathered}\mathbf{x}_j^{'0} \in \widehat{\mathcal{N}}_k\left(\mathbf{x}_i^0; \mathbf{L}^\top\mathbf{L}, \{\mathbf{x}_i^{'0}\}_{i=1}^n\right) \text{ and} \\ \left\|\mathbf{L}\left(\mathbf{x}_i^0 - \mathbf{x}_j^{'0}\right)\right\| < \left\|\mathbf{L}\left(\mathbf{x}_i^0 - \mathbf{x}_l^0\right)\right\|, \forall l : y_{i,l} = 0\end{gathered} \\ 0 & \text{otherwise}\end{cases} \tag{9}$$

By minimizing the cost function defined in eqn. (7) the learned metric considers labeled set as used in LMNN, while the close unlabeled neighborhood of each labeled vector remain close.

B. *Domain Adaptation (DA)*

Given a labeled set $S^0 = \left(\mathbf{x}_1^0, y_1^0\right), ..., \left(\mathbf{x}_m^0, y_m^0\right)$ drawn from a source distribution $\mathcal{D}_x^0$, and an unlabeled set,$\{\mathbf{x}_i^{'1}\}_{i=1}^n$, drawn from a target marginal distribution $\mathcal{D}_x^1$, a linear transformation $\mathbf{L} \in \mathbb{R}^{d \times d}$ is obtained by minimizing the following objective:

$$\epsilon\left(\mathbf{L}\right) = \left(1 - \omega_3\right)\epsilon_{\text{LMNN}}\left(\mathbf{L}\right) + \omega_3 \epsilon_{\text{DA}}\left(\mathbf{L}\right), \tag{10}$$

where:

$$\begin{aligned}\epsilon_{\text{DA}}\left(\mathbf{L}\right) &= \omega_2 \sum_{i,j} \eta_{i,j}^1 \left\|\mathbf{L}\left(\mathbf{x}_i^0 - \mathbf{x}_j^{'1}\right)\right\|^2 + \\ &+ \left(1-\omega_2\right)\sum_{i,j,l}\eta_{i,j}^1\left(1 - y_{i,l}\right)\Big[1 + \\ &+ \left\|\mathbf{L}\left(\mathbf{x}_i^0 - \mathbf{x}_j^{'1}\right)\right\|^2 - \left\|\mathbf{L}\left(\mathbf{x}_i^0 - \mathbf{x}_l^0\right)\right\|^2\Big]_+ \end{aligned} \tag{11}$$

and where the term ϵ_{LMNN} and $y_{i,l}$ are defined the same as in LMNN (sec. II).

The term $\eta_{i,j}^1$, defined similarly to $\eta_{i,j}^0$ (see eqn. 9), indicates whether $\mathbf{x}_j^{'1}$ is an unlabeled neighbor of $\mathbf{x}_i^0$, from among the unlabeled set $\{\mathbf{x}_i^{'1}\}_{i=1}^n$:

$$\eta_{i,j}^1 = \begin{cases} 1 & \begin{gathered}\mathbf{x}_j^{'1} \in \widehat{\mathcal{N}}_k\left(\mathbf{x}_i^0; \mathbf{L}^\top\mathbf{L}, \{\mathbf{x}_i^{'1}\}_{i=1}^n\right) \text{ and} \\ \left\|\mathbf{L}\left(\mathbf{x}_i^0 - \mathbf{x}_j^{'1}\right)\right\| < \left\|\mathbf{L}\left(\mathbf{x}_i^0 - \mathbf{x}_l^0\right)\right\|, \forall l : y_{i,l} = 0\end{gathered} \\ 0 & \text{otherwise}\end{cases} \tag{12}$$

By minimizing the cost defined in eqn. (10), the labeled set is used for learning a metric reflecting the true labeling of the data, while the unlabeled set is used for capturing and adjustment to the target's domain.

IV. Experimental Results

A. *Synthetic Data*

All points were drawn using a Gaussian distribution with a mean vector $\mu = (0, 0, 0)^\top$ and two different covariance matrices:

$$\begin{aligned}\Sigma^0 &= \begin{pmatrix} 0.35 & 10^{-4} & 0 \\ -0.35 & 10^{-4} & 0 \\ 0 & 0 & 0 \end{pmatrix} \\ \Sigma^1 &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 10^{-4} & 0.35 \\ 0 & 10^{-4} & 0.35 \end{pmatrix}\end{aligned} \tag{13}$$

The labels were set for both source and target vectors, using a single decision rule: $y = \text{sign}\left(x(2)\right)$, where $x(2)$ is the second element of the vector $\mathbf{x}$. Using these definitions, the source and target marginal distributions share one non-zero element ($x(2)$), which in both cases captures the knowledge regarding the label.

All experiments were held using a training set comprised of 20 labeled samples drawn from the source distribution. Three linear transformations were learned using the following training setups:

1) Supervised classification – LMNN [4], learned using 20 labeled source samples.

2) SS classification – as described in Section III-A, learned using 20 labeled and 2000 unlabeled source samples.
3) DA – as described in Section III-B, learned using 20 labeled source samples and 2000 unlabeled target samples.

The source and target classification tasks were examined using two test-sets, each consisting of 3,000 samples, drawn from the source and target distributions, correspondingly. Three labeled neighbors were considered during the learning stage for all learned transformations. For SS and DA tasks, three unlabeled neighbors were considered. Several values for ω_i, $i = 1, 2, 3$ were examined: $\omega_i = (0.1, .3, .5, .7, .9, .95)$, $i = 1, 2$ and $\omega_3 = \left(10^{-2}, 0.1, .3, .5, .7, .9, .95\right)$. The values of ω_i, $i = 1, 2, 3$ were separately tuned for LMNN, SS and for DA, to achieve a minimal testing error, averaged over 5 different training and testing sets. The error rates were measured using 1, 3, 5, 7 and 9 -NN. Table I presents the minimal testing error rates achieved by the tuned parameters, averaged over another 50 different sets. In case of source classification, LMNN produces the highest error rate (even compared to Euclidean distance), probably due to the small amount of training vectors, while our proposed approach using unlabeled data reduces this error. In case of target classification LMNN is much worse (18.8%) than Euclidean (2.2%) since it does not consider the spatial structure of the target domain. Our SS approach reduces this error to 4.2% since much more data is used, even if sampled from the source distribution. A further reduction, leading to the lowest error rate is achieved using our proposed DA approach for the target classification task (0.8%).

TABLE I
Synthetic Data - Classification Error Rates [%].

	Source Classification Err. [%]	Target Classification Err. [%]
Euclidean Distance	1.2	2.2
LMNN [4]	1.4	18.8
SS	**1.3**	**4.2**
DA	**1.1**	**0.8**

B. Hand Written Digits

Two data sets of hand written digits were used: MNIST, and USPS. MNIST consists of $60,000$ training examples and $10,000$ testing examples. Each example is a 28×28 8-bit gray-scale image. USPS consists of 4600 training examples and 4600 testing examples, each is a 16×16 8-bit,gray-scale image. In order to facilitate a Euclidean metric between these two data-sets, the MNIST examples were pre-processed to match the size of USPS: a frame of 4 pixels was trimmed of every picture (this frame was artificially added to this data set) yielding a 20×20 picture, and then re-sampled by 4/5, producing a 16×16 picture. The weights ω_i, $i = 1, 2, 3$ and amount of neighbors were set for every task and setup using separate training and development sets (taken from the training examples) so a minimal error would be achieved.

Tables II and III present the classification error rates achieved for USPS and MNIST, respectively, using the learned matrices. Since the labeled set is relatively small, LMNN does not improve the classification compared to Euclidean distance; it leads to some increase of the error rate in the source classification task and to a substantial increase in the target classification tasks. Our proposed approach considerably reduces the error rate, compared to LMNN. Compared to Euclidean, our method leads to a slight improvement for USPS classification as it reduces the error rate by 1.2% for SS and by 2.2% for DA. For MNIST classification, our method improves the error rate for SS task by 1.1% and by 0.1% for DA .

TABLE II
Classification of USPS Hand Written Digits.

Training set	Learned Metric	Test Err. [%]
200 labeled USPS imgs.	Euclidean Distance LMNN [4] learned for USPS	12.0 13.3
200 labeled and 4000 unlabeled USPS imgs.	SS	**10.8**
200 labeled MNIST imgs.	Euclidean Distance LMNN [4] learned for MNIST	30.0 42.6
200 labeled MNIST imgs. and 4000 unlabeled USPS imgs.	DA	**27.8**

TABLE III
Classification of MNIST Hand Written Digits.

Training Set	Learned Metric	Test Err. [%]
200 labeled MNIST imgs.	Euclidean Distance LMNN [4] learned for MNIST	27.3 34.0
200 labeled and 4,000 unlabeled MNIST imgs.	SS	**26.2**
200 labeled USPS imgs.	Euclidean Distance LMNN [4] learned for USPS	41.1 66.3
200 labeled USPS imgs. and 4,000 unlabeled MNIST imgs.	DA	41.0

To conclude, LMNN deteriorates if a small labeled training set is used (source classification), and becomes even worse when it is sampled from a different distribution (target classification), providing higher error rates than a simple Euclidean distance. The proposed approach improves the learning process as it utilizes an additional unlabeled set, leading to the lowest classification error rates in all the examined cases.

V. Conclusion

Many metric learning approaches are supervised methods, meaning they rely on labeled training examples for successful classification. LMNN is one of the most effective methods for metric learning but its performance deteriorates when the available training set is small and becomes even worse if the training set is sampled from a different distribution than the test vectors.

In this paper we have proposed a new method for utilizing labeled and unlabeled data (semi-supervised or domain adaptation setups) for metric learning. The (small) data set is used as in LMNN, where the unlabeled set, assumed to be large-scale and dense, is used for capturing the spatial structure of the data. In all the cases that we examined, LMNN leads to the highest error rates (even compared to Euclidean distance), due to the very small labeled training sets. Our approach, considering an additional unlabeled set, reduces this error, leading to the lowest error rates.

References

[1] I. Jolliffe, *Principal component analysis*, Wiley Online Library, 2005.

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

[3] M. Schultz and T. Joachims, "Learning a distance metric from relative comparisons," *Advances in neural information processing systems (NIPS)*, p. 41, 2004.

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