Non-Metric Label Propagation

Yin Zhang and **Zhi-Hua Zhou***
National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210093, China
{zhangyin, zhouzh}@lamda.nju.edu.cn

Abstract

In many applications non-metric distances are better than metric distances in reflecting the perceptual distances of human beings. Previous studies on non-metric distances mainly focused on supervised setting and did not consider the usefulness of unlabeled data. In this paper, we present probably the first study of label propagation on graphs induced from non-metric distances. The challenge here lies in the fact that the triangular inequality does not hold for non-metric distances and therefore, a direct application of existing label propagation methods will lead to inconsistency and conflict. We show that by applying spectrum transformation, non-metric distances can be converted into metric ones, and thus label propagation can be executed. Such methods, however, suffer from the change of original semantic relations. As a main result of this paper, we prove that any non-metric distance matrix can be decomposed into two metric distance matrices containing different information of the data. Based on this recognition, our proposed *NMLP* method derives two graphs from the original non-metric distance and performs a joint label propagation on the joint graph. Experiments validate the effectiveness of the proposed *NMLP* method.

1 Introduction

Distance plays an important role in many machine learning and pattern recognition techniques. Most distances were developed based on metrics satisfying the metric axioms, i.e., non-negativity, self-similarity, symmetry and triangular inequality. Such kind of distances are called *metric distances*. Although they have been applied widely and achieved great success, it has been found that in many applications metric distances deviate from the perceptual distances of human beings [Tversky, 1977; Santini and Jain, 1999]; in particular, the triangular inequality often violates human perceptual distances. For example, in the illustration shown in Figure 1,

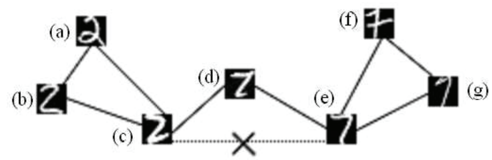


Figure 1: An illustration of non-metric distance that violates the metric axiom of triangular inequality.

the pattern (d) is similar to both the pattern (c) and the pattern (e), but the patterns (c) and (e) are dissimilar to each other! Distances which violate one or more metric axioms are called *non-metric distances*. The triangular inequality is the most often violated axiom, and it has been reported that by using non-metric distances which violate this axiom, it is possible to achieve better classification or recognition performances than using metric distances [Jacobs *et al.*, 2000; Tan *et al.*, 2006; 2009]. In this paper, if without clarification, we consider only this kind of non-metric distances.

Previous studies on non-metric distances all worked in supervised setting [Wu *et al.*, 2005; Tan *et al.*, 2006; 2009], neglecting the usefulness of unlabeled data. During the past decade, the use of unlabeled data has attracted much attention. A major paradigm is semi-supervised learning [Zhu, 2005; Chapelle *et al.*, 2006] which attempts to improve learning performance by automatically exploiting unlabeled training examples that are readily available. Many semi-supervised learning methods have been developed, among which graph-based methods [Zhu *et al.*, 2003; Zhou *et al.*, 2004] are particularly attractive due to their neat theoretical properties and good empirical performances.

The key of graph-based semi-supervise learning is label propagation, i.e., propagating labels of the labeled training examples to unlabeled instances based on graph structure, where the graphs are constructed by using the edges to reflect the distances/similarities between pairs of instances. At a first glance it might be thought that for tasks where non-metric distances are more suited than metric ones, we can realize label propagation simply by using non-metric distances to construct the graph and then applying existing label propagation methods on the graph. Unfortunately, this neglects the challenge posed by the violation of triangular inequality. As illustrated in Figure 1, we want to discriminate the hand-

*This research was supported by NSFC (60635030, 60721002), 863 Program (2007AA01Z169), JiangsuSF (BK2008018) and MSRA IST Program.

written digits ‘2’ and ‘7’. The pattern (d) is similar to both (c) and (e), violating the metric for $D_{cd} + D_{de} < D_{ce}$. If we apply label propagation on this graph, the label ‘2’ will be propagated to pattern (e), (f) and (g) via (d) and the label ‘7’ will be propagated to pattern (a), (b) and (c) via (d). Thus, inconsistency and conflict occur.

In this paper, we present probably the first study on label propagation based on non-metric distances. First, we show that it is possible to use spectrum transformation to convert the non-metric distances into metric ones, and thus label propagation can be executed on the metric-based graphs. Such methods, however, suffer from the change of semantics of the original similarity. Our main contribution is to propose the *NMLP* (Non-Metric Label Propagation) method, based on our proof that any non-metric distance matrix can be decomposed into two metric distance matrices. Our *NMLP* method works by decomposing the non-metric distance matrix at first, and then running a joint label propagation on the joint graph. The effectiveness of the *NMLP* method is validated in experiments. Note that we are studying how to deal with non-metric distances instead of proposing yet another semi-supervised learning approach, and it is possible to use our method to enable graph kernel approaches to handle non-metric distances. We are not concerning with metric learning, and we assume that the pairwise distances are given and reliable.

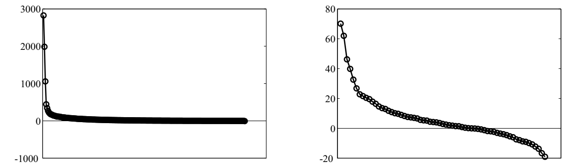
The rest of the paper is organized as follows. Section 2 reviews some related work. Section 3 presents the method of using spectrum transformation to make non-metrics become metrics. Section 4 presents our main contribution, the *NMLP* method. Section 5 reports on our experiments, and finally, Section 6 concludes the paper.

2 Related Work

Let $D = [D_{ij}]_{n\times n}$ be an $n \times n$ distance matrix where D_{ij} is the distance between the instances $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. Note that here $\boldsymbol{x}$ is not necessarily a vector. Denote $\mathcal{X}$ as the set of all $\boldsymbol{x}$. A distance matrix D will be called metric if there exists a metric function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $D_{ij} = d(\boldsymbol{x}_i, \boldsymbol{x}_j)$. In other words, D should satisfy that: a) $D_{ij} \geq 0$, $D_{ii} = 0$; b) $D_{ij} = D_{ji}$; and c) $D_{ij} \leq D_{ik} + D_{kj}$ where $1 \leq i, j, k \leq n$. Define the squared distance matrix $A = [A_{ij}]_{n\times n}$ where $A_{ij} = D_{ij}^2$. A is called squared-Euclidean if the corresponding metric function d is derived from the Euclidean norm. Let $K = -\frac{1}{2}\mathbf{H}A\mathbf{H}$ where $\mathbf{H} = I - \frac{1}{n}\boldsymbol{e}\boldsymbol{e}^T$, I is the identity matrix and $\boldsymbol{e} = [1, 1, \dots, 1]^T$. We have the following important theorem [Young and Householder, 1938; Laub and Müller, 2004]:

Theorem 1. *D is metric if and only if K is positive semi-definite.*

Decompose K as $K = V\Lambda V^T$ where $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)$ is the diagonal matrix of the eigenvalues (sorted in descending order) and $V = [\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_n]$ is the orthonormal matrix of corresponding eigenvectors. Since K is semi-definite, we have $\lambda_i \geq 0$ $(1 \leq i \leq n)$. Thus, we can define $\tilde{\boldsymbol{x}}_i \triangleq \Lambda^{\frac{1}{2}}\boldsymbol{v}_i$ and D_{ij} is thus the Euclidean distance between $\tilde{\boldsymbol{x}}_i$ and $\tilde{\boldsymbol{x}}_j$. If the distance matrix D is non-metric, K is no more positive semi-definite and λ_i can contain negative values. Through analyzing the spectrum of negative eigenvalues,



a) Spectrum of ‘Proteins’ b) Spectrum of ‘Cat-Cortex’

Figure 2: Examples of spectrums corresponding to two different kinds of non-metrics.

Laub and Müller [2004] attributed to two causes: 1) the distance is intrinsically metric but corrupted by noise, and 2) the distance is intrinsically non-metric. If the negative eigenvalues are close to zero, the non-metric is probably caused by the first reason and the second otherwise. Figure 2 shows some examples, where the two subfigures present the spectrums of the distance matrices of the data *Proteins* and *Cat-Cortex* (see Section 4 for details of these data), respectively. From Figure 2 we can see that the non-metric of *Proteins* is likely caused by noise while that of *Cat-Cortex* is likely caused by intrinsic non-metric. Laub and Müller [2004] argued that the negative eigenvalues of the first kind of non-metric is harmful while those of the second kind may contain useful information. Actually, non-metrics have been found useful in many applications [Jacobs *et al.*, 2000; Wu *et al.*, 2005; Tan *et al.*, 2006; 2009].

Semi-supervised learning attempts to exploit unlabeled data to improve the performance of learning with limited amount of labeled training data. Many effective methods have been developed, among which graph-based methods have attracted much attention. Such methods construct a graph by considering the pairwise relation between the instances, and then try to propagate the label information over the graph. Label propagation is the key to these methods. Many label propagation methods have been developed, such as the harmonic function method [Zhu *et al.*, 2003], the local and global consistency method [Zhou *et al.*, 2004], *etc.* Fujino *et al.* [2007] proposed a semi-supervised method for multi-component data based on a hybrid generative and discriminative approach. Zhou and Burges [2007] extended label propagation to multi-view data by generalizing normalized cut from a single view to multiple views, which forms a mixture of Markov chains defined on different views.

3 Spectrum Transformation

Different strategies can be used to construct a graph from a distance matrix D. The graph to be constructed can be represented by an $n \times n$ matrix W, where w_{ij} is the edge weight reflecting the similarity between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. If D is non-metric, as mentioned before, directly applying existing label propagation methods to such graphs may lead to inconsistency. One possible approach to avoid the problem is to convert the non-metric distance matrix to a metric one. There are several ways to achieve this goal, which mainly fall into the category of *spectrum transformation*.

From Theorem 1 we know that to make D metric is equivalent to make K positive semi-definite, or, to make K's eigenvalues non-negative. Spectrum transformation applies a mapping function $\hat{\lambda} = f(\lambda)$ on λ to make it non-negative. The induced matrix $\hat{K} = V\hat{\Lambda}V^T$ is thus a positive semi-definite matrix. Here we list four widely used spectrum transformation methods [Wu *et al.*, 2005] below:

1. *Denoise*: $f(\lambda) = \max(0, \lambda)$ [Graepel *et al.*, 1999; Pękalska *et al.*, 2001]. All negative eigenvalues are treated as noise and replaced by zeros.
2. *Flip*: $f(\lambda) = |\lambda|$ [Graepel *et al.*, 1999]. All negative eigenvalues are flipped on their signs.
3. *Diffusion*: $f(\lambda) = \exp(\beta\lambda)$ with $\beta > 0$ [Kondor and Lafferty, 2002]. It is derived from the diffusion kernel with the purpose to consider data distribution when computing pairwise similarity.
4. *Shift*: $f(\lambda) = \lambda + \eta$, where η is a constant [Roth *et al.*, 2003]. All eigenvalues are added by a constant to make them non-negative. It has been used in clustering and guarantees distortionless embedding of non-metric pairwise data with regard to cluster assignment [Roth *et al.*, 2003]. It can be proved that *Shift* is equivalent to adding a constant c to every off-diagonal elements in A, i.e., $\hat{A} = A + c(\boldsymbol{e}\boldsymbol{e}^T - I)$. If $\eta = -\lambda_n$ then c is the minimum constant to make D satisfy the triangular inequality [Roth *et al.*, 2003]. Experiments in [Wu *et al.*, 2005] show that *Shift* outperforms other methods for classification.

After spectrum transformation, $\hat{K}$ is positive semi-definite and the corresponding $\hat{D}$ is metric. Then, existing label propagation methods could be applied to $\hat{D}$.

However, although these four spectrum transformation methods can convert non-metric distances into metric ones, there are other problems preventing a successful non-metric label propagation. The *Denoise* method assumes that the non-metric is caused by noise, which enables it to filter out those noise when the assumption holds; but when non-metric is caused by intrinsically non-metric property, simply deleting those negative eigenvalues will miss useful information. For *Flip*, *Diffusion* and *Shift*, the semantic similarity may be changed. Take *Shift* for example. Assume that originally $D_{ij} > D_{st}$. The transformed distance satisfies that $\hat{D}_{ij} - \hat{D}_{st} = (D_{ij} - D_{st})\frac{D_{ij}+D_{st}}{\sqrt{D_{ij}^2+c}+\sqrt{D_{ij}^2+c}} < D_{ij} - D_{st}$ for any $c > 0$. Actually, the scheme of *Shift* to make the triangular inequality hold is to reduce the difference between the large and small distances. Such a change may not be what we want since samples belonging to different classes may become closer under the new distance.

4 The NMLP Method

In contrast to spectrum transformation methods, here we take a more straightforward solution which does not change the structure of the spectrum. First, we determine whether the non-metric is caused by noise. A simple method is to compare the ratio of the absolute value of the smallest eigenvalue and the largest one, i.e., $|\lambda_n/\lambda_1|$, with a threshold, θ. If $|\lambda_n/\lambda_1| \le \theta$, we perform *Denoise*. If $|\lambda_n/\lambda_1| > \theta$, applying *Denoise* would not be a good choice, and we will execute another routine.

NMLP(D, $\boldsymbol{y}_l$, θ, α)

Input:
D: $n \times n$ distance matrix
$\boldsymbol{y}_l$: $n_l \times 1$ label vector
θ : threshold
α : combination parameter

Process:
Construct A where $A_{ij} = D_{ij}^2$ and $K = -\frac{1}{2}\mathbf{H}A\mathbf{H}$.
Decompose $K = V\Lambda V^T$ and get $\lambda_1, \cdots, \lambda_n$.
if $|\lambda_n/\lambda_1| \le \theta$
  Denoise the spectrum as $\hat{\Lambda} = \max(\Lambda, 0)$ and $\hat{X} = \hat{\Lambda}^{\frac{1}{2}}V^T$. Based on $\hat{X}$, construct W and normalize $\bar{W} = \Delta^{-\frac{1}{2}}W\Delta^{-\frac{1}{2}}$
else
  Represent Λ as $\Lambda = \Lambda^+ - \Lambda^-$ and $X^+ = (\Lambda^+)^{\frac{1}{2}}V^T$, $X^- = (\Lambda^-)^{\frac{1}{2}}V^T$. Construct W^+ and W^- based on X^+ and X^-, respectively. Normalize W^+ and W^- and combine them to get $\bar{W} = \alpha\bar{W}^+ + (1-\alpha)\bar{W}^-$.
end if
Compute the normalized Laplacian matrix as $L = I - \bar{W}$ and the labels of unlabeled examples are predicted as $\tilde{\boldsymbol{y}}_u = -L_{uu}^{-1}L_{ul}\boldsymbol{y}_l$.

Output:
$\tilde{\boldsymbol{y}}_u$: the labels on unlabeled data

Figure 3: The *NMLP* method

For any non-metric distance matrix, we have:

Theorem 2. *Any squared non-metric distance matrix A can be decomposed into two squared metric distance matrix A^+ and A^- where $A = A^+ - A^-$.*

Proof. The proof is constructive. First we define $\Lambda^+ \triangleq \max(\Lambda, 0)$ and $\Lambda^- \triangleq \max(-\Lambda, 0)$. From those we can induce two feature spaces for the instance as $\boldsymbol{x}_i^+ = (\Lambda^+)^{\frac{1}{2}}\boldsymbol{v}_i$ and $\boldsymbol{x}_i^- = (\Lambda^-)^{\frac{1}{2}}\boldsymbol{v}_i$. Thus construct A^+ and A^- as the squared metric distance matrix for $\boldsymbol{x}^+$ and $\boldsymbol{x}^-$, i.e., $A_{ij}^+ \triangleq \|\boldsymbol{x}_i^+ - \boldsymbol{x}_j^+\|^2$ and $A_{ij}^- \triangleq \|\boldsymbol{x}_i^- - \boldsymbol{x}_j^-\|^2$. Now we prove $A = A^+ - A^-$. Define $K^+ \triangleq V\Lambda^+V^T$ and $K^- \triangleq V\Lambda^-V^T$, therefore, $K^+ = -\frac{1}{2}\mathbf{H}A^+\mathbf{H}$ and $K^- = -\frac{1}{2}\mathbf{H}A^-\mathbf{H}$. Since $\Lambda = \Lambda^+ - \Lambda^-$, we have $K = K^+ - K^-$ and finally $A = A^+ - A^-$. □

Based on Theorem 2 we can reconstruct two different kinds of features, $\boldsymbol{x}^+$ and $\boldsymbol{x}^-$, corresponding to the positive and negative eigenvalues, respectively. From experiments of [Laub and Müller, 2004; Laub *et al.*, 2007] it can be seen that these two kinds of features reflect different views of the data. However, how/why the two features represent different similarities remains an open problem [Laub and Müller, 2004; Laub *et al.*, 2007], and so we use both and treat them separately in order not to miss important information.

From $\boldsymbol{x}^+$ and $\boldsymbol{x}^-$ we can construct two graphs $\mathcal{G}^+ = (\mathcal{V}, \mathcal{E}^+, W^+)$ and $\mathcal{G}^- = (\mathcal{V}, \mathcal{E}^-, W^-)$ respectively, where $\mathcal{V}$ is the set of vertices, $\mathcal{E}^*$ is the set of edges and $W^* = [w_{ij}^*]_{n\times n}$ is the weight matrix, $* \in \{+, -\}$. The degree of a

Table 1: Databases used in experiments

Database	Distance description
Proteins	Evolutionary distances of 226 protein sequences in 4 classes of globins
Cat-Cortex	Connection strengths between 65 regions of the cat's celebral cortex with 4 functional classes
Kimia	Symmetric modified Hausdorff distances between 72 binary shape images of 6 classes [Pękalska *et al.*, 2001]
UNIPEN	Dynamic time-warping distance of 250 handwritten lower case characters of 5 classes [Bahlmann *et al.*, 2002]
USPS	Tangent-distance of 1000 USPS handwritten digits in 2 classes [Keysers *et al.*, 2004]
music-EMD	Earth mover's distance of music incipits corresponding to 2 identical sets of 2 classes [Typke *et al.*, 2003]
Music-PTD	Proportional transportation distance of music incipits corresponding to 2 identical sets of 2 classes [Typke *et al.*, 2003]

vertex u is $\delta^*(u) = \sum_{v\in\mathcal{V}} w^*(u,v)$ and the volume of $\mathcal{V}$ is $\text{vol}^*\mathcal{V} = \sum_{u\in\mathcal{V}} \delta^*(u)$. The transition probabilities are $p^*(u,v) = w^*(u,v)/\delta^*(u)$ and the stationary probabilities $\pi^*(u) = \delta^*(u)/\text{vol}^*\mathcal{V}$. Define

$$\gamma^+(u) = \frac{\alpha\delta^+(u)/\text{vol}^+\mathcal{V}}{\alpha\delta^+(u)/\text{vol}^+\mathcal{V} + (1-\alpha)\delta^-(u)/\text{vol}^-\mathcal{V}} \quad (1)$$

$$\gamma^-(u) = \frac{(1-\alpha)\delta^-(u)/\text{vol}^-\mathcal{V}}{\alpha\delta^+(u)/\text{vol}^+\mathcal{V} + (1-\alpha)\delta^-(u)/\text{vol}^-\mathcal{V}} \quad (2)$$

where $0 \le \alpha \le 1$ is the parameter controlling the combination. From the Markov mixture model, the transition probability of the combination of two graphs is

$$\begin{aligned} p(u,v) &= \gamma^+(u)p^+(u,v) + \gamma^-(u)p^-(u,v) \\ &= \frac{\alpha w^+(u,v)/\text{vol}^+\mathcal{V} + (1-\alpha)w^-(u,v)/\text{vol}^-\mathcal{V}}{\alpha\delta^+(u)/\text{vol}^+\mathcal{V} + (1-\alpha)\delta^-(u)/\text{vol}^-\mathcal{V}} \end{aligned} \quad (3)$$

and the stationary probability is

$$\pi(u) = \alpha\delta^+(u)/\text{vol}^+\mathcal{V} + (1-\alpha)\delta^-(u)/\text{vol}^-\mathcal{V}. \quad (4)$$

Therefore, we get the weight of the joint graph as

$$w(u,v) = p(u,v)\delta(u) = \alpha\frac{w^+(u,v)}{\text{vol}^+\mathcal{V}} + (1-\alpha)\frac{w^-(u,v)}{\text{vol}^-\mathcal{V}}. \quad (5)$$

Assume that the first n_l examples are with class labels $\boldsymbol{y} = (y_1, y_2, \cdots, y_{n_l})$ and the last n_u examples are unlabeled. Given the weight matrix W, the normalized weight matrix $\bar{W} = \Delta^{-\frac{1}{2}} W \Delta^{-\frac{1}{2}}$ where $\Delta = \text{diag}(\delta_1, \delta_2, \cdots, \delta_n)$ is a diagonal matrix with diagonal elements $\delta_i = \sum_j W_{ij}$. The optimal class assignment $\tilde{\boldsymbol{y}}$ is found by minimizing the following energy function on the joint graph:

$$\begin{cases} \min\limits_{\tilde{\boldsymbol{y}}} & \sum_{i,j=1}^n \bar{W}_{ij}(\tilde{y}_i - \tilde{y}_j) = \tilde{\boldsymbol{y}}^T L \tilde{\boldsymbol{y}} \\ \text{s.t.} & \tilde{\boldsymbol{y}}_l = \boldsymbol{y}_l \end{cases} \quad (6)$$

where $L = I - \bar{W}$ is the normalized Laplacian matrix and $\tilde{\boldsymbol{y}}_l$ stands for the first n_l elements of $\tilde{\boldsymbol{y}}$. The optimal class labels assigned to the unlabeled examples, denoted as $\tilde{\boldsymbol{y}}_u$, has the closed-form $\tilde{\boldsymbol{y}}_u = -L_{uu}^{-1} L_{ul} \boldsymbol{y}_l$ where the indices u and l stand for the parts of the Lapacian matrix that are related to the labeled and the unlabeled examples, respectively. The algorithm is summarized in Figure 3.

5 Experiments

The databases and non-metric distances used in our experiments are listed in Table 1. For *UNIPEN*, since it contains 5 classes, we merge class 1 and 2 into one class and the others into another class. Similarly, for *Kimia*, we reassign the data belonging to class 1 to 3 as the first class and other data as the second class. Each database has several data sets, say, four for *Proteins*, *Cat-Cortex* and *USPS* and two for the other databases. In total, we run experiments on 20 data sets.

Since previously there is no non-metric label propagation methods, here we evaluate the performances of *Denoise*, *Flip*, *Diffusion*, *Shift* and *NMLP*. In *Diffusion*, β is set to 0.1; in *Shift*, η is set to $-\lambda_n$, which is the minimum η to make D metric; in *NMLP*, the threshold θ is set to 1% and α is set to $\left(\sum_{\lambda_i>0}|\lambda_i|\right) / \left(\sum|\lambda_i|\right)$. Such a setting is based on the fact that the eigenvalue reflects the variance of the corresponding feature. How to get the optimal value for α is an open problem and will be studied in future. In addition, we also evaluate two degenerated versions of *NMLP*. The first uses only the positive eigenvalues, denoted as *Positive*; the second uses only the negative eigenvalues, denoted as *Negative*. Actually, *Positive* is equivalent to *Denoise* as well as *NMLP* when $\alpha = 1$, while *Negative* is equivalent to *NMLP* when $\alpha = 0$. Directly executing label propagation with the original non-metric distance is also evaluated, denoted as *Direct*. In the experiments the graph edges are defined by $w_{ij} = \exp(\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{-\delta^2})$ where $\delta = e^{-2}\bar{d}$ and $\bar{d}$ is the average pairwise distance among the data.

On each data set, we randomly select 5% and 10% samples to be used as labeled data and the remaining samples as unlabeled data. The results are shown in Tables 2 and 3. Note that *NMLP* judges in its first step that the non-metric of *Proteins* is caused by noise, while other data sets are intrinsically non-metric. The results show that the performance of *NMLP* on data sets with intrinsic non-metric is almost always the best, and its performance on data sets where the non-metric is caused by noise, i.e., *Proteins*, is also good. The performance of *Positive* is better than all other methods except *NMLP*; this suggests that the crucial information of non-metric chiefly lies in positive eigenvectors, as supported by the fact that the performances of *Negative* are quite poor. The inferiority of *Direct* compared with *NMLP* demonstrates the infeasibility of applying label propagation with original non-metric distance. The performances of the spectrum transformation methods are unstable and worse than *NMLP* on most data sets; this phenomenon is consistent with the analysis presented in Section 3.

We also study the influence of the setting of α on the performance of *NMLP*. We evaluate the performance of *NMLP* with α varying from 0.3 to 1 with 0.05 as interval. We use the first data set of each database and the lower labeled data rate (5%) setting. Other settings are as same as that used before. The results are shown in Figure 4. We can see that in *Pro-*

Table 2: Error rates (mean±std.) when 5% data are labeled. The best performance on each data set is bolded, and its comparable performances are underlined (statistical significance examined via paired t-tests at 95% significance level).

Database	*Flip*	*Diffusion*	*Shift*	*Positive*	*Negative*	*Direct*	*NMLP*
Proteins-1	0.039±0.004	0.403±0.093	**0.035±0.005**	**0.035±0.005**	0.429±0.083	**0.035±0.005**	**0.035±0.005**
Proteins-2	0.155±0.078	0.349±0.032	0.155±0.078	0.155±0.080	0.375±0.084	**0.151±0.077**	0.155±0.080
Proteins-3	0.015±0.020	0.015±0.004	**0.014±0.018**	0.015±0.020	0.318±0.144	**0.014±0.018**	0.015±0.020
Proteins-4	0.069±0.021	**0.019±0.000**	0.072±0.022	0.069±0.021	0.365±0.155	0.071±0.021	0.069±0.021
Cat-Cortex-1	0.097±0.078	0.198±0.201	0.201±0.121	0.079±0.109	0.295±0.127	0.361±0.200	**0.065±0.094**
Cat-Cortex-2	0.170±0.127	0.180±0.209	0.092±0.062	0.104±0.089	0.419±0.248	0.143±0.089	**0.078±0.063**
Cat-Cortex-3	0.211±0.188	0.115±0.221	0.112±0.045	0.102±0.111	0.389±0.189	0.152±0.111	**0.092±0.109**
Cat-Cortex-4	0.233±0.212	0.220±0.133	0.212±0.053	0.228±0.149	0.405±0.185	0.394±0.173	**0.185±0.109**
Kimia-1	0.304±0.139	0.433±0.116	0.347±0.189	**0.276±0.127**	0.494±0.064	0.338±0.175	**0.276±0.127**
Kimia-2	0.268±0.154	0.464±0.101	0.276±0.143	**0.237±0.166**	0.461±0.072	0.259±0.176	**0.237±0.166**
UNIPEN-1	0.382±0.063	0.231±0.094	0.231±0.085	0.230±0.082	0.466±0.080	0.232±0.090	**0.228±0.081**
UNIPEN-2	0.326±0.087	0.247±0.068	0.209±0.094	**0.200±0.096**	0.444±0.065	0.210±0.049	**0.200±0.095**
USPS-1	0.280±0.084	0.422±0.006	0.258±0.069	0.251±0.076	0.480±0.040	0.265±0.091	**0.250±0.076**
USPS-2	0.289±0.083	0.515±0.041	0.295±0.091	**0.275±0.083**	0.489±0.040	0.280±0.097	**0.275±0.083**
USPS-3	0.253±0.079	0.363±0.115	0.206±0.091	0.201±0.076	0.455±0.091	0.217±0.067	**0.200±0.077**
USPS-4	0.253±0.057	0.385±0.099	0.252±0.075	0.228±0.062	0.456±0.067	0.264±0.130	**0.227±0.061**
music-EMD-1	0.465±0.082	0.491±0.049	0.486±0.058	**0.463±0.064**	0.480±0.077	0.501±0.059	**0.463±0.059**
music-EMD-2	0.488±0.055	0.478±0.041	0.491±0.077	0.471±0.065	0.478±0.111	0.485±0.079	**0.467±0.064**
music-PTD-1	0.463±0.060	0.505±0.055	0.495±0.071	0.465±0.060	0.501±0.082	0.467±0.085	**0.455±0.059**
music-PTD-2	0.475±0.049	0.489±0.033	0.473±0.053	**0.467±0.041**	0.471±0.031	0.478±0.045	**0.467±0.041**

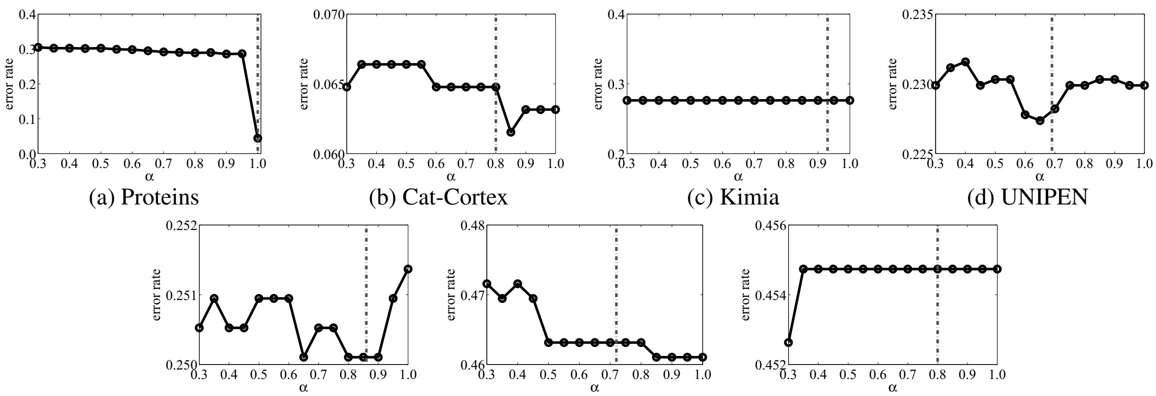
Table 3: Error rates (mean±std.) when 10% data are labeled. The best performance on each data set is bolded, and its comparable performances are underlined (statistical significance examined via paired t-tests at 95% significance level).

Database	*Flip*	*Diffusion*	*Shift*	*Positive*	*Negative*	*Direct*	*NMLP*
Proteins-1	0.034±0.007	0.336±0.066	**0.032±0.008**	**0.032±0.008**	0.438±0.087	**0.032±0.008**	**0.032±0.008**
Proteins-2	0.109±0.065	0.345±0.035	0.112±0.075	**0.107±0.066**	0.386±0.047	**0.107±0.066**	**0.107±0.066**
Proteins-3	0.004±0.012	0.020±0.011	**0.003±0.011**	0.004±0.012	0.282±0.111	**0.003±0.011**	0.004±0.012
Proteins-4	0.067±0.032	**0.019±0.002**	0.067±0.032	0.067±0.032	0.474±0.206	0.067±0.032	0.067±0.032
Cat-Cortex-1	0.067±0.067	0.046±0.016	0.128±0.069	**0.022±0.012**	0.226±0.058	0.147±0.082	0.024±0.017
Cat-Cortex-2	0.162±0.142	0.149±0.158	0.106±0.080	0.131±0.133	0.468±0.276	0.125±0.110	**0.092±0.070**
Cat-Cortex-3	0.074±0.055	0.044±0.012	0.113±0.074	0.067±0.073	0.339±0.113	0.152±0.089	**0.039±0.016**
Cat-Cortex-4	0.067±0.051	0.106±0.104	0.162±0.063	0.080±0.064	0.333±0.094	0.308±0.115	**0.060±0.051**
Kimia-1	0.094±0.034	0.437±0.113	0.099±0.032	0.057±0.037	0.392±0.045	0.085±0.036	**0.057±0.037**
Kimia-2	0.099±0.039	0.455±0.091	0.151±0.068	0.090±0.040	0.381±0.037	0.096±0.044	**0.090±0.040**
UNIPEN-1	0.280±0.117	0.152±0.092	0.201±0.078	0.128±0.100	0.446±0.086	0.173±0.078	**0.125±0.098**
UNIPEN-2	0.244±0.067	0.195±0.052	0.164±0.060	0.136±0.082	0.422±0.052	0.164±0.060	**0.135±0.080**
USPS-1	0.158±0.032	0.376±0.088	0.162±0.054	0.156±0.038	0.447±0.041	**0.155±0.043**	**0.155±0.038**
USPS-2	**0.183±0.063**	0.462±0.079	0.205±0.071	0.185±0.067	0.469±0.039	0.185±0.067	0.184±0.067
USPS-3	0.147±0.068	0.328±0.022	0.156±0.059	**0.142±0.077**	0.389±0.046	0.160±0.066	**0.142±0.077**
USPS-4	0.168±0.059	0.367±0.008	0.156±0.050	**0.148±0.047**	0.405±0.046	0.171±0.050	**0.148±0.046**
music-EMD-1	0.476±0.104	0.507±0.102	0.496±0.071	0.464±0.089	0.473±0.090	0.491±0.067	**0.462±0.089**
music-EMD-2	0.462±0.060	0.503±0.039	0.482±0.058	**0.431±0.079**	0.460±0.075	0.489±0.074	**0.431±0.079**
music-PTD-1	0.489±0.053	0.484±0.074	0.489±0.096	0.484±0.038	0.529±0.113	0.496±0.079	**0.478±0.037**
music-PTD-2	0.478±0.070	0.520±0.040	0.482±0.040	0.476±0.066	0.483±0.061	0.511±0.057	**0.474±0.065**

teins, where *NMLP* finds the non-metric is caused by noise, $\alpha = 1$ is the best; on other databases, where *NMLP* finds the non-metric is intrinsic, $\alpha = 1$ is not the best or there are other α values which can do as well as $\alpha = 1$. This shows that the judgement on the cause of non-metric is reasonable. Moreover, it can be seen that the estimated α is close to the optimal α value on most databases.

6 Conclusion

Non-metric distances are better than metric distances in many applications since they reflect human perception better in some cases, however, previously it is unknown how to execute label propagation on graphs constructed based on non-metric distances. This paper presents probably the first study on non-metric label propagation. The key challenge here lies in the fact that the violation of triangular inequality axiom makes the direct application of existing label propagation methods suffer from the problems of inconsistency and conflict. We show that by applying spectrum transformation, non-metric distances can be converted into metric ones, and thus label propagation can be used. However, such methods are inappropriate since they may change the original semantic relations. We prove that any non-metric distance can be



(a) Proteins (b) Cat-Cortex (c) Kimia (d) UNIPEN

(e) USPS (f) music-EMD (g) music-PTD

Figure 4: Error rates under different α values. The dashed vertical line indicates the α value estimated by *NMLP*.

decomposed into two metric distances, based on which we propose the *NMLP* (Non-Metric Label Propagation) method and its effectiveness is validated in experiments.

An interesting future work is to apply the *NMLP* method to real tasks where non-metric distances have been found better than metric ones. Another future work is to design other kinds of non-metric label propagation methods.

Acknowledgements: We want to thank Yu-Yin Sun and Xin-Pan Xiao for their comments on a preliminary draft.

References

[Bahlmann *et al.*, 2002] C. Bahlmann, B. Haasdonk, and H. Burkhardt. On-line handwriting recognition with support vector machines - a kernel approach. In *IWFHR*, pages 49–54, 2002.

[Chapelle *et al.*, 2006] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[Fujino *et al.*, 2007] A. Fujino, N. Ueda, and K. Saito. Semi-supervised learning for multi-component data classification. In *IJCAI*, pages 2754–2759, 2007.

[Graepel *et al.*, 1999] T. Graepel, R. Herbrich, and K. Obermayer. Classification on pairwise proximity data. In *NIPS 11*, pages 438–444. 1999.

[Jacobs *et al.*, 2000] D. W. Jacobs, D. Weinshall, and Y. Gdalyahu. Classification with non-metric distances: Image retrieval and class representation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(6):583–600, 2000.

[Keysers *et al.*, 2004] D. Keysers, W. Macherey, H. Ney, and J. Dahmen. Adaptation in statistical pattern recognition using tangent vectors. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 26(2):269–274, 2004.

[Kondor and Lafferty, 2002] R. I. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete input spaces. In *ICML*, pages 315–322, 2002.

[Laub and Müller, 2004] J. Laub and K.-R. Müller. Feature discovery in non-metric pairwise data. *Journal of Machine Learning Research*, 5:801–818, 2004.

[Laub *et al.*, 2007] J. Laub, J. Macke, K.-R. Müller, and F. A. Wichmann. Inducing metric violations in human similarity judgements. In *NIPS 19*, pages 777–784. 2007.

[Pękalska *et al.*, 2001] E. Pękalska, P. Paclík, and R. P. W. Duin. A generalized kernel approach to dissimilarity-based classification. *Journal of Machine Learning Research*, 2:175–211, 2001.

[Roth *et al.*, 2003] V. Roth, J. Laub, M. Kawanabe, and J. M. Buhmann. Optimal cluster preserving embedding of nonmetric proximity data. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(12):1540–1551, 2003.

[Santini and Jain, 1999] S. Santini and R. Jain. Similarity measures. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 21(9):871–883, 1999.

[Tan *et al.*, 2006] X. Tan, S. Chen, Z.-H. Zhou, and J. Liu. Learning non-metric partial similarity based on maximal margin criterion. In *CVPR*, pages 138–145, 2006.

[Tan *et al.*, 2009] X. Tan, S. Chen, Z.-H. Zhou, and J. Liu. Face recognition under occlusions and variant expressions with partial similarity. *IEEE Trans. Information Forensics and Security*, 2009.

[Tversky, 1977] A. Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977.

[Typke *et al.*, 2003] R. Typke, P. Giannopoulos, R. C. Veltkamp, F. Wiering, and R. van Oostrum. Using transportation distances for measuring melodic similarity. In *ISMIR*, pages 107–114, 2003.

[Wu *et al.*, 2005] G. Wu, E. Y. Chang, and Z. Zhang. Learning with non-metric proximity matrices. In *ACM Multimedia*, pages 411–414, 2005.

[Young and Householder, 1938] G. Young and A. S. Householder. Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 3:19–22, 1938.

[Zhou and Burges, 2007] D. Zhou and C. J. C. Burges. Spectral clustering and transductive learning with multiple views. In *ICML*, pages 1159–1166, 2007.

[Zhou *et al.*, 2004] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *NIPS 16*, pages 321–328. 2004.

[Zhu *et al.*, 2003] X. Zhu, Z. Ghahramani, and J. Lafferty. Semisupervised learning using gaussian fields and harmonic functions. In *ICML*, pages 912–919, 2003.

[Zhu, 2005] X. Zhu. Semi-supervised learning literature survey. Technical Report 1530, Department of Computer Sciences, University of Wisconsin at Madison, Madison, WI, 2005.