Non-Metric Locality-Sensitive Hashing

Yadong Mu, Shuicheng Yan
Department of Electrical and Computer Engineering, National University of Singapore, Singapore
{elemy, eleyans}@nus.edu.sg

Abstract

Non-metric distances are often more reasonable compared with metric ones in terms of consistency with human perceptions. However, existing *locality-sensitive hashing* (LSH) algorithms can only support data which are gauged with metrics. In this paper we propose a novel locality-sensitive hashing algorithm targeting such non-metric data. Data in original feature space are embedded into an implicit *reproducing kernel Kreĭn space* and then hashed to obtain binary bits. Here we utilize the norm-keeping property of p-stable functions to ensure that two data's collision probability reflects their non-metric distance in original feature space. We investigate various concrete examples to validate the proposed algorithm. Extensive empirical evaluations well illustrate its effectiveness in terms of accuracy and retrieval speedup.

Introduction

Over the last decade we have witnessed an explosive growth in the scale of image and video data. Billions of visual data are publicly available on the Web, part of which are accompanied with manual annotation. It brings both challenges and opportunities to traditional algorithms developed on small to median scale data sets. Particularly, *approximate nearest-neighbor* (ANN) search has become a key ingredient in many large-scale machine learning and computer vision tasks.

A well-defined distance is crucial in ANN. Most of the popular distances are subject to the metric axioms, i.e., non-negativity, symmetry and triangular inequality. Although these metric distances empirically prove successful, however, it is argued that in many real-world applications they are actually inconsistent with the perceptual distances of human beings (Laub et al. 2006). Such an example is presented in Figure 1. As can be seen, both the objects "man" and "horse" are perceptually similar to their composition, but the two obviously differ from each other. In the computer vision research, some authors reported superior performance of non-metric distances over traditional metric ones (Jacobs, Weinshall, and Gdalyahu 2000).

In this paper we devise a novel *locality-sensitive hashing* (LSH) scheme for non-metric distances, which extends traditional metric-based LSH algorithms (Andoni and Indyk 2008). Examples of non-metric distances include *Chamfer distance* between curves, the *Kullback- Leibler distance*, *the dynamic time warping* (DTW) distance for comparing time series, the *edit distance* for comparing strings and the *hyperbolic tangent kernel* $k(x, x') = \tanh\big(< x, x' > -1\big)$ of neural networks. Particularly, many non-metric distances such as KL can be regarded as special cases of Bregman divergence (Bregman 1967). We give its definition to illustrate non-metrics: let ϕ be a real-valued strictly convex function defined over a convex set $S \subseteq \mathbb{R}^m$. The ϕ-induced *Bregman divergence* is defined as:

$$D_\phi(p,\ q) = \phi(p) - \phi(q) - \big\langle \nabla\phi(q),\ p - q \big\rangle. \qquad (1)$$

Figure 2 gives an intuitive interpretation for this definition. Table 1 lists some widely used non-metric Bregman divergences and the corresponding ϕ's. Throughout the paper we constrain the distances to be symmetrized, e.g., for Bregman divergence, we use the symmetric form $\widetilde{D}_\phi(q,\ p) = D_\phi(p,\ q) + D_\phi(q,\ p)$.

In the rest of this paper, we first survey related works, and then elaborate on the proposed algorithm after introducing several related concepts. Finally, empirical evaluations on various benchmarks are presented.

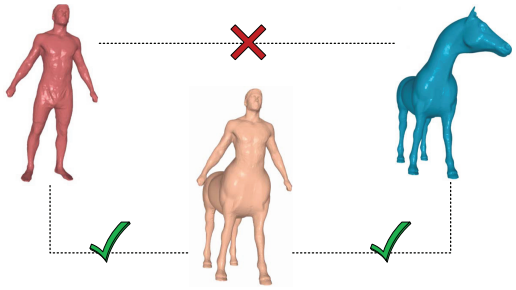

Figure 1: An example to illustrate non-metric distance that deviates from triangular inequality. The $\surd$ symbol indicates that two samples are similar while the $\times$ symbol means that these two samples are dissimilar.

Related Works

Those previous works tightly related to this research can be roughly casted into two categories:

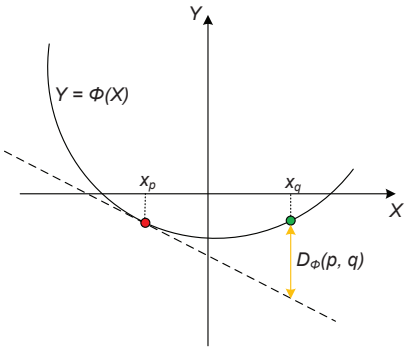

Figure 2: Illustration for Bregman divergence (1-D case).

Table 1: Some convex functions and the corresponding Bregman divergences.

	$\phi(x)$	$D_\phi(x,\ x')$
Itakura-Saito	$-\sum_i \log x_i$	$\sum_i (\frac{x_i}{x'_i} - \log \frac{x_i}{x'_i} - 1)$
Kullback-Leibler	$\sum x_i \log x_i$	$\sum x_i \log \frac{x_i}{x'_i}$
Hinge	$\lvert x \rvert$	$\max(0, -2\text{sign}(x')x)$

Tree based methods: researchers in computational geometry and database management communities developed various tree-based structures for fast nearest neighbor retrieval, such as KD-tree (Bentley 1975) and VP-tree (Yianilos 1993). Most of these methods perform hierarchical space decomposition, attaching similar data to adjacent leaf nodes. These methods can be easily adapted to handle non-metric data, including Bregman divergences. For example, the so-called *Bregman ball tree* (BB-tree) was proposed in (Cayton 2008), where each tree node is pertained to a Bregman ball $B(\mu,\ R) = \{x \mid D_\phi(x,\ \mu) \leq R\}$ (μ and R are center and radius respectively). Given a query, the search for k-NN proceeds in a branch-and-bound way. A tree node is pruned if the distance between query datum and its projection onto the Bregman ball exceeds current upper bound. In (Zhang et al. 2009), similar tricks are developed to adapt for R-tree and VA-file. The major disadvantages of these tree-based methods lie in the tremendous requirement of memory to store tree node information (exponentially grows with respect to data number and feature dimensionality) and limited performance enhancement when handling high-dimensional data.

LSH based methods: For many machine learning tasks, an approximate nearest neighbor is almost as good as the exact one in existence of noises. The concept of *locality-sensitive hashing* (LSH) is supposed to well fit this appeal, especially for high-dimensional data, c.f. (Andoni and Indyk 2008) for a brief survey. Denote $\mathcal{H}$ as a family of hash functions mapping $\mathcal{R}^d$ to binary space. $\mathcal{H}$ is called "locality sensitive" if under any hash function $h \in \mathcal{H}$, the collision probability $P_{\mathcal{H}}\big[h(p) = h(q)\big] = \text{Sim}(p,\ q)$ (other formulations exist but are in spirit the same), where $\text{Sim}(\cdot, \cdot)$ is a function measuring pairwise similarity.

Existing LSH families rely on well-defined metrics or similarity functions in the original feature space, e.g., Jaccard coefficient (Broder et al. 1997), Hamming distance (Indyk and Motwani 1998), Arccos distance (Charikar 2002), and ℓ^p distance with $p \in [0, 2)$(Datar et al. 2004). However, none prior LSH work is devoted to the non-metric distances such as Bregman divergence (although some of its special cases are metrics, in general it is not).

Indefinite Kernels and Kreĭn Space

Given a data set $\mathcal{X} = \{x_i\},\ i = 1 \ldots n$, we can construct an $n \times n$ distance matrix $D = (D_{ij})$. As stated above, D is assumed to be symmetric, and zero only for the diagonal element. D is called *squared-Euclidean* if it is derived from the ℓ^2 metric. Let $K = -\frac{1}{2}QDQ$ where $Q = I - \frac{1}{n}ee^T$. Q is the projection matrix onto the orthogonal complement of $e = (1, 1, \ldots, 1)^T$. The transform results in a centralized kernel matrix $K = (K_{ij})$, with $K_{ij} = \big\langle \psi(x_i), \psi(x_j) \big\rangle$, where ψ is unknown mapping function. We have the following observation (Young and Householder 1938) (Laub and Müller 2004):

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

To plug K into kernel-based learning algorithms like support vector machine (SVM), one of the key requirements is the positive definiteness. Unfortunately, kernel matrix K induced from non-metric distance matrix D occasionally violates this condition and falls into the family of *indefinite kernels* (Pekalska and Haasdonk 2009). Figure 3 shows the spectrum of such a matrix, where negative eigenvalues are observed.

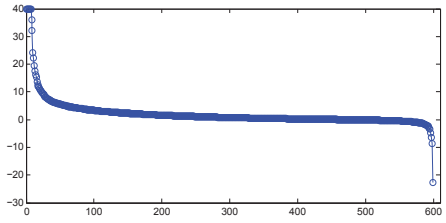

Figure 3: Typical spectrum of non-metric distance matrices.

Indefinite kernels fail to be embedded into the so-called *reproducing kernel Hilbert space* (RKHS) (Scholkopf and Smola 2001), but is fortunately interpreted by the notation of *reproducing kernel Kreĭn space* (RKKS) (Bognar 1974)(Pekalska and Haasdonk 2009). Here "Kreĭn space" refers to a vector space $\mathcal{K}$ equipped with an indefinite inner product $\big\langle \cdot, \cdot \big\rangle_{\mathcal{K}} : \mathcal{K} \times \mathcal{K} \to \mathbb{R}$ such that $\mathcal{K}$ admits an orthogonal decomposition as a direct sum $\mathcal{K} = \mathcal{K}_+ \oplus \mathcal{K}_-$, where $(\mathcal{K}_+, \kappa_+(\cdot, \cdot))$ and $(\mathcal{K}_-, \kappa_-(\cdot, \cdot))$ are separable Hilbert spaces with their corresponding positive definite inner products. The inner product of $\mathcal{K}$, however, is the difference of κ_+ and κ_-, i.e., for any $\xi_+, \xi'_+ \in \mathcal{K}_+$ and $\xi_-, \xi'_- \in \mathcal{K}_-$, we have

$$\big\langle \xi_+ + \xi_-, \xi'_+ + \xi'_- \big\rangle_{\mathcal{K}} = \kappa_+(\xi_+, \xi'_+) - \kappa_-(\xi_-, \xi'_-). \quad (2)$$

Strong relationship between Kreĭn space and indefinite matrix K exists. Perform singular value decomposition

$K = V^T \Lambda V$[1] and decompose the elements in the spectrum $\Lambda = \text{Diag}(\lambda_i)$ into two parts, i.e., $\Lambda_+ = \max(\Lambda, 0)$ and $\Lambda_- = -\min(\Lambda, 0)$. K can be transformed to be the difference of two positive semi-definite matrices:

$$K = K_+ - K_- = V^T \Lambda_+ V - V^T \Lambda_- V, \tag{3}$$

which implies that K can be embedded into a RKKS. Same to the kernel tricks in RKHS, we need not know the explicit mapping functions for $\mathcal{K}_+$ and $\mathcal{K}_-$.

The Proposed Hashing Algorithm

Step 1: ℓ^2-Keeping Projection in Hilbert Space

Based on the embedding into RKKS as in Equation 2, we devise a two-step LSH scheme. The goal of step 1 is to pursue two LSH families $\mathcal{H}_+$ and $\mathcal{H}_-$ such that $\forall h_+ \in \mathcal{H}_+$, $P_{\mathcal{H}_+}\big[h_+(p) = h_+(q)\big]$ monotonically decreases with respect to the ℓ^2 distance between p and q in $\mathcal{K}_+$, i.e., ℓ^2-LSH in $\mathcal{K}_+$. The situation in $\mathcal{K}_-$ is the same. Note that the operations in $\mathcal{K}_+$ and $\mathcal{K}_-$ are independent, thus in sequel we ignore the subscripts $+$, $-$ without confusion. Previous work in (Datar et al. 2004) discovers that ℓ^p-norm keeping hashing is feasible based on *p-stable distribution*, which is defined as below:

Definition 1. *(p-stable Distribution): a distribution Π over $\mathbb{R}$ is called p-stable, if $\exists p \geq 0$ such that for any n real numbers $v_1 \dots v_n$ and i.i.d. variables $X_1 \dots X_n$ from Π, $\sum_i v_i X_i$ shares the same distribution with $(\sum_i |v_i|^p)^{1/p} X$, where X is a random variable from Π.*

Such stable distributions exist for any $p \in (0, 2]$, e.g.,

- *Cauchy distribution* Π_C with density function $c(x) = \frac{1}{\pi} \frac{1}{1+x^2}$ is 1-stable,
- *Gaussian distribution* Π_G with density function $g(x) = \frac{1}{\sqrt{2\pi}} \exp\big(-\frac{x^2}{2}\big)$ is 2-stable.

Here we are interested in the case of $p = 2$ and thus capitalize on the above Gaussian distribution. Suppose the data lie in $\mathbb{R}^d$. Existing hashing family $\mathcal{H} : \mathbb{R}^d \to \mathbb{R}$, each $h \in \mathcal{H}$ has the form $h(x) = r^T x$ for any $x \in \mathbb{R}^d$. Here each entry of the hashing vector $r \in \mathbb{R}^d$ is randomly sampled from Π_G. It is provable that $\mathcal{H}$ is norm-keeping, i.e. for v_1 and v_2, the difference $|r^T v_1 - r^T v_2| = |r^T (v_1 - v_2)|$ is distributed as $\|v_1 - v_2\|_2 X$ where X is a random variable from Π_G.

Unfortunately, since all we have is the kernel matrices K_+, K_-, such a linear representation is infeasible. Here we resort to a trick similar to the ones previously used in the kernel PCA (Schölkopf, Smola, and Müller 1998) and kernel LSH (Kulis and Grauman 2009). Our main observation is as below:

Theorem 2. *(ℓ^2-Keeping Projection in RKHS) Denote $\kappa(\cdot, \cdot)$ to be the inner product in Hilbert space $\mathcal{K}$. Given an m-cardinality data set $\mathcal{X}$ and corresponding Gram matrix G, the ℓ^2-metric keeping projection can be expressed as $p(x) = \sum_{i=1}^m \omega(i) \kappa(x, x_i)$, where $\omega(i)$ only relies on G.*

[1] In this paper we use K or G for the Gram matrices and $\mathcal{K}$ for Hilbert spaces.

Proof. Denote the implicit Hilbert mapping function as ψ. The geometric mean can be computed as $\mu_\psi = \frac{1}{m} \sum_{i=1}^m \psi(x_i)$. For a t-cardinality subset $\mathcal{S} \subset \{1 \dots n\}$, let $z = \frac{1}{t} \sum_{i \in S} \psi(x_i)$ and $\widetilde{z} = \sqrt{t}(z - \mu_\psi)$. According to the *central limit theorem*, $\widetilde{z}$ is distributed as Gaussian $\Phi(0, \Sigma)$, where Σ is the covariance matrix of $\mathcal{X}$. Further applying a whitening transform, we can obtain the desired hash vector in $\mathcal{K}$, i.e. $r = \Sigma^{1/2} \widetilde{z}$. For any datum x, $h(x) = \psi(x)^T \Sigma^{1/2} \widetilde{z}$.

Given Gram matrix $G = \Psi^T \Psi$, where each column of Ψ corresponds to a feature vector in data set $\mathcal{X}$. Similar to (Schölkopf, Smola, and Müller 1998), it is easily verified that $\widetilde{z}^T \Sigma^{1/2} \psi(x) = \widetilde{z}^T (\Psi Q)(QGQ)^{-\frac{1}{2}} (\Psi Q)^T \psi(x)$, where $Q = I - \frac{1}{m} e e^T$. Substituting $\widetilde{z} = \sqrt{t} \Psi (\frac{1}{t} \delta_S - \frac{1}{m} e)^T$, where δ_S is a binary indicator vector for subset $\mathcal{S}$. Finally we get

$$p(x) = \Big[\sqrt{t}(\frac{1}{t}\delta_S - \frac{1}{m}e) GQ(QGQ)^{-\frac{1}{2}} Q^T\Big] \Psi^T \psi(x) \tag{4}$$

Let $\omega \triangleq \Big[\sqrt{t}(\frac{1}{t}\delta_S - \frac{1}{m}e) GQ(QGQ)^{-\frac{1}{2}} Q^T\Big]$, thus the conclusion holds. □

Step 2: LSH in Kreĭn Space

The relationship between the Kreĭn Space $\mathcal{K}$ and the associated Hilber spaces $\mathcal{K}_+, \mathcal{K}_-$ can be summarized in Equation 2. For any $\xi, \xi' \in \mathcal{K}$, denote the pairwise ℓ^2 distance in $\mathcal{K}$ as $\|\xi - \xi'\|_{\mathcal{K}}$. Based on Equation 2 and the orthogonality of $\mathcal{K}_+, \mathcal{K}_-$, we have

$$\begin{aligned} \|\xi - \xi'\|_{\mathcal{K}}^2 &= \|\xi_+ - \xi'_+\|_{\mathcal{K}_+}^2 - \|\xi_- - \xi'_-\|_{\mathcal{K}_-}^2 \\ &= \big(\|\xi_+ - \xi'_+\|_{\mathcal{K}_+} - \|\xi_- - \xi'_-\|_{\mathcal{K}_-}\big) \times \\ &\quad \big(\|\xi_+ - \xi'_+\|_{\mathcal{K}_+} + \|\xi_- - \xi'_-\|_{\mathcal{K}_-}\big) \end{aligned} \tag{5}$$

Denote $D_-(\xi, \xi') \triangleq (|p_+(\xi) - p_+(\xi')| - |p_-(\xi) - p_-(\xi')|)$ and $D_+(\xi, \xi') \triangleq (|p_+(\xi) - p_+(\xi')| + |p_-(\xi) - p_-(\xi')|)$. It is easy to verify that the means of D_- and D_+ are proportional to the magnitudes of the two factors in Equation 5. We can thus make the following approximation:

$$\begin{aligned} \|\xi - \xi'\|_{\mathcal{K}}^2 \propto\ & \big(|p_+(\xi) - p_+(\xi')| - |p_-(\xi) - p_-(\xi')|\big) \times \\ & \big(|p_+(\xi) - p_+(\xi')| + |p_-(\xi) - p_-(\xi')|\big) \end{aligned} \tag{6}$$

Based on the project functions p_+, p_- obtained in step 1, we introduce two auxiliary functions a_1, a_2, whose definitions are as below:

$$a_1(\xi) = p_+(\xi) - p_-(\xi) \tag{7}$$

$$a_2(\xi) = p_+(\xi) + p_-(\xi) \tag{8}$$

Without loss of accuracy, both $a_1(\xi)$ and $a_2(\xi)$ are normalized to $[0, 1]$. Denote as $\tilde{a}_1(\xi)$ and $\tilde{a}_2(\xi)$ respectively. The adopted hash function $h : \mathbb{R}^2 \to \{0, 1\}^2$ casts 2-D vector $(a_1(\cdot)\ \ a_2(\cdot))^T$ into two binary bits $(h_1(\cdot)\ \ h_2(\cdot))^T$. The scheme is as below (here k denotes 1 or 2):

$$h_k(\xi) = \begin{cases} 1, & \tilde{a}_k(\xi) > \theta \\ 0, & \tilde{a}_k(\xi) \leq \theta \end{cases} \tag{9}$$

where θ is a real number randomly sampled from $[0, 1]$. In this way, for any $\xi, \xi' \in \mathcal{K}$, the Hamming distance between their hashing bits shall be one number of $\{0, 1, 2\}$. It is interesting to investigate how the Hamming distance is related to the distance in original feature space, i.e., $\|\xi - \xi'\|^2_{\mathcal{K}}$. In fact, we have the following observation:

Theorem 3. ***(Collision Probability)*** *For any $\xi, \xi' \in \mathcal{K}$, denote $|\tilde{a}_1(\xi) - \tilde{a}_1(\xi')| = p_1$ and $|\tilde{a}_2(\xi) - \tilde{a}_2(\xi')| = p_2$. Let D_{ham} be the Hamming distance between $h(\xi)$ and $h(\xi')$. Under the hash scheme in Equation 9, D_{ham} attains values 0, 1, 2 with probabilities $(1 - p_1)(1 - p_2)$, $p_1(1 - p_2) + (1 - p_1)p_2$, and $p_1 p_2$ respectively.*

Proof. The two terms $D_+(\xi, \xi'), D_-(\xi, \xi')$ in Equation 6 can be determined as below:

Case 1: $(p_+(\xi) - p_+(\xi')) \times (p_-(\xi) - p_-(\xi')) \geq 0$:

$$\begin{aligned} D_-(\xi, \xi') &= |(p_+(\xi) - p_-(\xi)) - (p_+(\xi') - p_+(\xi'))| \\ &= |\tilde{a}_1(\xi) - \tilde{a}_1(\xi')| \quad (10) \\ D_+(\xi, \xi') &= |\tilde{a}_2(\xi) - \tilde{a}_2(\xi')| \quad (11) \end{aligned}$$

Case 2: $(p_+(\xi) - p_+(\xi')) \times (p_-(\xi) - p_-(\xi')) < 0$:

$$\begin{aligned} D_-(\xi, \xi') &= |\tilde{a}_2(\xi) - \tilde{a}_2(\xi')| \quad (12) \\ D_+(\xi, \xi') &= |\tilde{a}_1(\xi) - \tilde{a}_1(\xi')| \quad (13) \end{aligned}$$

In either case, D_{ham} can be determined by p_1 and p_2, thus the conclusion can be easily verified. □

The Retrieval Algorithm

After the construction of hash tables, another challenge is to encode out-of-sample data and retrieve its approximate nearest neighbors. Given a query x, we calculate its non-metric distance to all elements in $\mathcal{X}$ to get $D_{\Psi,\psi(x)}$. Let $V_{diag}(K) \in \mathbb{R}^n$ be the vector formed by the diagonal elements of K. Recall that $K \in \mathbb{R}^{n \times n}$ contains the inner products of centralized data, it can be verified that $\kappa(x, x) = \frac{1}{n} e\big(D_{\Psi,\psi(x)} - V_{diag}(K)\big)$, and further obtain the inner product between $\mathcal{X}$ and x:

$$\Psi^T \psi(x) = -\frac{1}{2} Q\big(D_{\Psi,\psi(x)} - V_{diag}(K)\big) \quad (14)$$

Recall that K can be decomposed as $V \Lambda V^T$, let $\widetilde{K}$ be the new kernel matrix with x plugged in. As an approximation, we assume $\widetilde{K}$ can be expressed in the following form:

$$\widetilde{K} = \begin{pmatrix} K \\ u^T \end{pmatrix} \approx \begin{pmatrix} V \\ u^T \end{pmatrix} \Lambda_+ V^T - \begin{pmatrix} V \\ u^T \end{pmatrix} \Lambda_- V^T, \quad (15)$$

where the vector $u \in \mathbb{R}^n$ is the parameter to estimate. The optimal u^* can be pursued by solving the following least-squared problem:

$$\begin{aligned} u^* &= \arg\min_u \left\| \big(V\Lambda_+ u - V\Lambda_- u\big) - \Psi^T \psi(x) \right\|_2^2 \\ &= \arg\min_u \left\| V\Lambda u - \Psi^T \psi(x) \right\|_2^2 \quad (16) \end{aligned}$$

whose closed-form solution is available, i.e., $u^* = ((V\Lambda)^T (V\Lambda))^{-1} (V\Lambda)^T \Psi^T \psi(x)$. After that, the inner products in $\mathcal{K}_+$, $\mathcal{K}_-$ can be simply determined according to $\kappa_+(x_k, x) = v_k \Lambda_+ u^*$ and $\kappa_-(x_k, x) = v_k \Lambda_- u^*$ respectively, where v_k is the k-th row of V.

Table 2: Data sets description and the corresponding non-metric distances (or similarities).

DATA SET	SAMPLE NUMBER	DISTANCE/SIMILARITY
INTERNETADS	2359	TVERSKY
LOCAL-PATCH	300K	ITAKURA-SAITO
CIFAR-10	60K	KL
USPS-DIGIT	10K	CHAMFER

Experiments

Dataset Description

In this section we provide quantitative study for the proposed non-metric LSH algorithm. We adopt four benchmarks, whose information and the corresponding applied non-metric distances are listed as below:

InternetAds represents a set of possible advertisements on Internet pages. Most features are binary, conveying webpage's information including the image's URL and alt text, the anchor text, and words occurring near the anchor text. On this data set, we adopt the *Tversky linear contrast similarity* (Tversky 1977) to measure the pairwise similarity.

CIFAR-10 is a labeled subset of the well-know 80 million "tiny image" data set constructed at MIT, consisting of 60K 32×32 highly-smoothed images in 10 categories. The dataset is constructed to learn meaningful recognition-related image filters whose responses resemble the behavior of human visual cortex. For each image, we extract 387-d GIST feature, and apply the Itakura-Saito divergence.

Local-Patch is a large-scale data set used to compare image matching algorithms in computer vision. It contains roughly 300K 32×32 image patches from photos of Trevi Fountain (Rome), Notre Dame (Paris) and Half Dome (Yosemite). For each image patch, we compute a 128-d SIFT vector as the holistic descriptor. These SIFT vectors are then ℓ^1-normalized and measured by KL-distance.

USPS-digit contains grayscale handwritten digit images scanned from envelopes by the U.S. Postal Service. It includes roughly 10K samples for digits $0 \sim 9$. For each image, we first convert it into binary image via thresholding, and then adopt Chamfer distance (Barrow et al. 1977) as distance measure.

Table 2 summaries the information of these benchmarks, and in Figure 4, we present several images sampled from the above-mentioned data sets.

Figure 4: Example images from benchmarks used in this paper.

Evaluation Methodology

In our evaluation, each data set is divided into two part: one is used to construct the hash functions (the sample numbers

vary among different data set, depending on the whole data size and intrinsic complexity lying in the data), and the rest is kept for evaluation. Once hash functions are obtained from the former subset, all elements in the evaluation subset are projected to the hash buckets accordingly. To gauge the performance, we adopt the *leave-one-out cross validation* (LOOCV) method. In practice, we randomly sample $500 \sim 1000$ data as queries and the final performance is averaged over all samples.

Specifically, the performance of a hash algorithm is estimated according to the *Good Neighbor Ratio* (GNR) criterion. Here "good neighbor" indicates the samples which are adjacent to the query. Given the non-metric distance definition, it is possible to calculate any sample's proximity rank relative to a query. Typically the top 5% nearest samples to the query are regarded as "good neighbor". To evaluate, we can simply count the proportion of good neighbors in the retrieved data set below a pre-defined Hamming distance (e.g., 1 or 2).

Experimental Results

To illustrate the indefinite property of non-metric distances (or equivalently similarities), we calculate the Gram matrices between 500 random samples for all four benchmarks, and then perform spectral analysis, as seen in Figure 5. The positive singular values are plotted in red, while the negative ones are colored in blue. It can be seen that in most cases the spectral energy from $\mathcal{K}_-$ cannot be ignored.

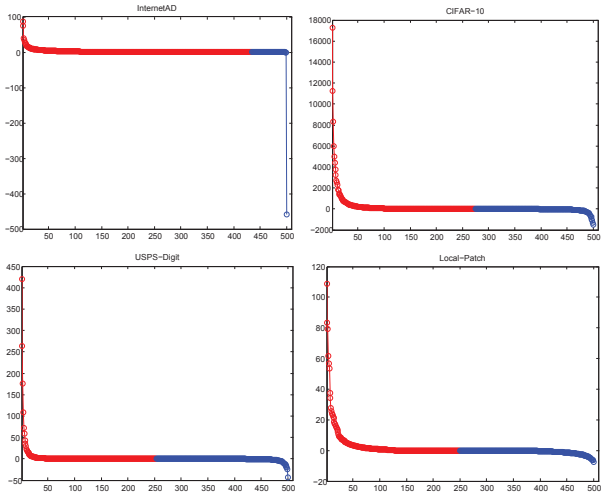

Figure 5: Spectrums corresponding to four different non-metric distances or similarities. For better viewing, please see original color pdf file.

Figure 7 presents the experimental results based on the GNR criterion. On each data set, we run 10 independent rounds to reduce randomness. The hashing methods involved in the figures are as below:

- Our proposed non-metric LSH algorithm
- Kernelized LSH (Kulis and Grauman 2009) based on K_+
- Kernelized LSH based on K_-
- First perform hashing based on K_+ or K_- independently, and then simply concatenate them (in other words, the hamming distance between two data are the summation of their distances calculated from K_+ and K_-).

Note that Kernelized LSH based on K is infeasible due to K's indefinite property. Our aim is to check whether the two Hilbert spaces pertained to K_+ and K_- contain complementary information to each other. In Figure 7, it is observed that in most cases the GNR values of the proposed LSH algorithm is consistently superior to the results based on either K_+ or K_-, or their direct concatenation, which serves as a strong sign that K_- carries useful information for data transform. Traditional learning algorithms from indefinite kernels tend to ignore negative singular values as noises. However, here we argue that such a treatment possibly bring information loss, and the proposed method provides a simple solution to overcome this issue.

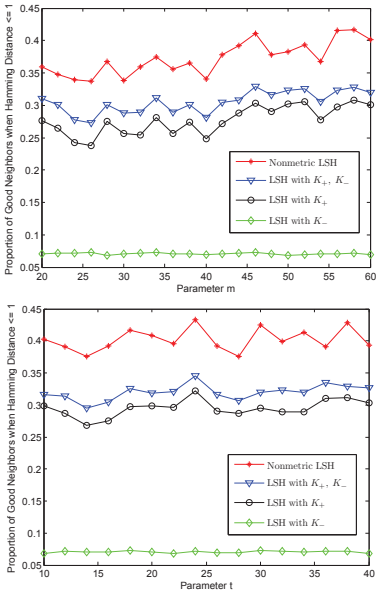

Figure 6: Parameter stability testing results. See text for description. For better viewing, please see original color pdf file.

We also study the influence of parameters m and t (see Theorem 2) used in hash function construction. We conduct two additional experiments on the InternetAD data set. In the first experiment, we vary m and fix $t = \frac{1}{4}m$. Figure 6 plots the performance evolutionary curve. While in the second experiment, we fix $m = 50$ an let t vary from 10 to 40, as shown in Figure 6. In both cases, the proposed method shows relatively stable performance.

Conclusions

Traditional LSH methods focus on well-known metrics such as Euclidean distance and Cosine similarity. In this paper we investigate the practicability of hashing methods given non-metric distance measure. We show that symmetric non-metric distances can be elegantly interpreted by Kreĭn space theory, and derive a two-step locality-sensitive hash-

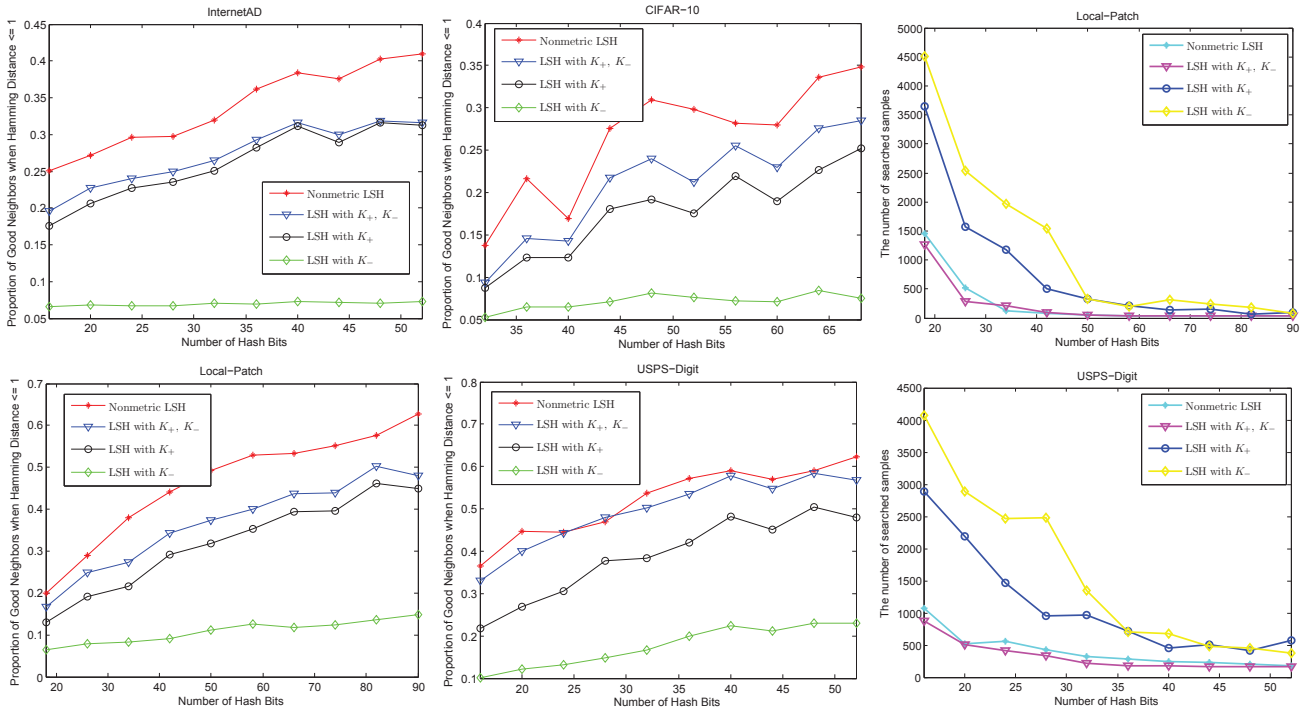

Figure 7: Experimental results. The left and middle columns presents GNR values of the four benchmarks, and the right column shows the decreasing tendency of retrieved samples when the hash bits increase. For better viewing, please see original color pdf file.

ing method, which captures information contained in negative singular values, rather than simply abandoning them.

Acknowledgment: the work in this paper is supported by NRF/IDM Program at Singapore, under research Grant NRF2008IDMIDM004-029.

References

Andoni, A., and Indyk, P. 2008. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Commun. ACM* 51(1):117–122.

Barrow, H. G.; Tenenbaum, J. M.; Bolles, R. C.; and Wolf, H. C. 1977. Parametric correspondence and chamfer matching: Two new techniques for image matching. In *IJCAI*, 659–663.

Bentley, J. 1975. Multidimensional binary search trees used for associative searching. *Commun. ACM* 18(9):509–517.

Bognar, J. 1974. *Indefinite inner product spaces*. Springer-Verlag.

Bregman, L. M. 1967. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics* 7:200–217.

Broder, A. Z.; Glassman, S. C.; Manasse, M. S.; and Zweig, G. 1997. Syntactic clustering of the web. *Computer Networks* 29(8-13):1157–1166.

Cayton, L. 2008. Fast nearest neighbor retrieval for bregman divergences. In *ICML*, 112–119.

Charikar, M. 2002. Similarity estimation techniques from rounding algorithms. In *STOC*, 380–388.

Datar, M.; Immorlica, N.; Indyk, P.; and Mirrokni, V. 2004. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the twentieth annual symposium on Computational geometry*, 253–262.

Indyk, P., and Motwani, R. 1998. Approximate nearest neighbors: towards removing the curse of dimensionality. In *STOC*.

Jacobs, D.; Weinshall, D.; and Gdalyahu, Y. 2000. Classification with nonmetric distances: Image retrieval and class representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 22(6):583–600.

Kulis, B., and Grauman, K. 2009. Kernelized locality-sensitive hashing for scalable image search. In *ICCV*.

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

Laub, J.; Macke, J.; Müller, K.-R.; and Wichmann, F. A. 2006. Inducing metric violations in human similarity judgements. In *NIPS*, 777–784.

Pekalska, E., and Haasdonk, B. 2009. Kernel discriminant analysis for positive definite and indefinite kernels. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 31(6):1017–1032.

Scholkopf, B., and Smola, A. J. 2001. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press.

Schölkopf, B.; Smola, A. J.; and Müller, K.-R. 1998. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation* 10(5):1299–1319.

Tversky, A. 1977. Features of similarity. *Psychological Review* 84:327–352.

Yianilos, P. N. 1993. Data structures and algorithms for nearest neighbor search in general metric spaces. In *SODA*.

Young, G., and Householder, A. 1938. Discussion of a set of points in terms of their mutual distances. *Psychometrika* 3(1):19–22.

Zhang, Z.; Ooi, B. C.; Parthasarathy, S.; and Tung, A. K. H. 2009. Similarity search on bregman divergence: Towards non-metric indexing. *PVLDB* 2(1):13–24.