INFORMATIVE ENERGY METRIC FOR SIMILARITY MEASURE IN REPRODUCING KERNEL HILBERT SPACES

Songhua LIU [1,2,*], Junying ZHANG [2], Caiying DING [1,3,4]

[1] *College of Science, Jiangxi University of Science and Technology, 341000 Ganzhou, P.R.China*
E-mail: sooh.liu@gmail.com

[2] *School of Computer Science and Technology, Xidian University, 710071 Xi'an, P.R.China*
E-mail: jyzhang@xidian.edu.cn

[3] *Center of Interdisciplinary, Lanzhou University, 730000 Lanzhou, P.R.China*

[4] *Institute of Physics, Chinese Academy of Science, 100190 Beijing, P.R.China*
E-mail: dcy06@lzu.edu.cn

Received 15 October 2010
Accepted 11 January 2012

Abstract

In this paper, information energy metric (IEM) is obtained by similarity computing for high-dimensional samples in a reproducing kernel Hilbert space (RKHS). Firstly, similar/dissimilar subsets and their corresponding informative energy functions are defined. Secondly, IEM is proposed for similarity measure of those subsets, which converts the non-metric distances into metric ones. Finally, applications of this metric is introduced, such as classification problems. Experimental results validate the effectiveness of the proposed method.

Keywords: Kernel methods; Similarity measure; Reproducing kernel Hilbert space; Non-metric distance

1. Introduction

Similarity measure in reproducing kernel Hilbert spaces (RKHS) has attracted much attention of researchers from diverse areas such as computer vision, machine learning and pattern recognition during the past few years [1,2]. In these applications, a notion of similarity is induced by computing kernel functions on arbitrary training sample pairs in input space. However, many similarity measures are developed based on metric distance under high-dimensional setting and samples in the RKHS often violate one or more metric axioms [3,4], this may impair the performance of machine learning algorithms. Therefore, how to choose a "good" similarity measure is one of the key concerns of these algorithms.

Previous studies on similarity measure in the RKHS can be divided into two modes, i.e., metric distance mode and non-metric distance mode.

In the metric distance mode, distance was developed based on metrics satisfying the metric axioms.

[*] Corresponding author. Address: College of Science, Jiangxi University of Science and Technology, 341000 Ganzhou, P.R.China (S.H.Liu).
E-mail:sooh.liu@gmail.com

Published by Atlantis Press

Euclidean distance is the most widely used similarity measure, such as inner products [2] in kernel-based machines. Edit distances, Hamming distance, and the sophisticated distances are introduced in Ref.[5]; information distance is studied in Ref.[6]. All these similarity measures are encoded in the so-called kernel matrix.

Although they have been widely used in many applications, they may violate one or more metric axioms under high-dimensional setting, which is called non-metric distance [3,4,7,8]. As indicated by Cover and Thomas in Ref.[9], the proposed information-theoretic based metrics, such as Kullback-Leibler (KL) divergence, can capture data structure beyond second order statistics. The KL divergence is often intuited as a distance metric, but it is not a true metric. Actually, it can be derived from Bregman divergence [10]. It is similar to a metric, but does not satisfy the triangle inequality or symmetry. Distance based on these similarity measures often deviate from the perceptual distance of human beings, and may impair the performance of machine learning algorithms [11,12,13].

Motivated and inspired by the above works, in this paper, we introduce informative energy metric (IEM) for similarity measure in the RKHS. Our analysis suggests that the IEM method can convert the non-metric problems into metric ones, and it holds both for Euclidean distance and manifold setting.

The rest of this paper is organized as follows: Section 2 gives a brief review of preliminaries. Section 3 presents the derivation of IEM, including proof of IEM and the objective function to update the output coefficient matrix. Performance evaluation of IEM is shown in Section 4 based on the benchmark problems in the area of projecting visualization and classification. Conclusions based on the study are highlighted in Section 5.

2. Preliminaries

In this section, we give all the necessary background material needed for the development of IEM in Section 3. We begin with a brief description of notations for similarity measure.

Given a set of training samples $\{x_i, c_i\}_{i=1}^{N}$, where $x_i \in X \subset R^{N\times d}$, d is dimensionality of the samples, X is input space, $c_i \in C = \{1, 2, \ldots, N_c\}$ is class label, C is class label set, and N_c is the number of classes. We now make use of a dual notation for the sample x in the input space, it is written with a single subscript x_i when its class is irrelevant, index $1 \leqslant i \leqslant N$. If the class is relevant, assume that we have J_p samples for pth class, we write x_{pj}, where the class index $1 \leqslant p \leqslant N_c$, and the index within class $1 \leqslant j \leqslant J_p$.

We recall also briefly the notations and lemmas which can also be found in Ref.[14].

Definition 1. (Reproducing Kernel Hilbert Space) Let X be a nonempty set and H a Hilbert space of functions: $f : X \to R$. Then H is called a reproducing kernel Hilbert space endowed with the inner product $\langle\cdot,\cdot\rangle$ if there exists a function $k : X \times X \to R$ with the following properties. (1) k has the reproducing property $\langle f, k(x,\cdot)\rangle = f(x)$, for all $f \in H$; in particular, $\langle k(x,\cdot), k(x',\cdot)\rangle = k(x,x')$. (2) k spans H.

Note that the RKHS uniquely determines k, we know the following definition using the Mercer's theorem in Ref.[14].

Definition 2. (Mercer Kernel Map) If k is a kernel satisfying the Mercer's theorem, we can construct a mapping Φ into a space where k acts as an inner product, $\langle\Phi(x), \Phi(x')\rangle = k(x,x')$, for almost all $x, x' \in X$. Moreover, give any $\varepsilon > 0$, there exist a map Φ_n into an N-dimensional inner product space such that $|k(x,x') - \langle\Phi_N(x), \Phi_N(x')\rangle| < \varepsilon$.

In practice, we are given a finite amount of samples $x_1, \ldots, x_N$, we do not want to (or are unable to) analyze a given kernel k analytically, we can still compute a map Φ such that k corresponds to an inner product in the linear span of the $\Phi(x_i)$.

Lemma 1. *(Data-Dependent Kernel Map [14]) Suppose the data $x_1, \ldots, x_N$ and the kernel k are such that the kernel matrix $K_{ij} = k(x_i, x_j)$ is positive definite. Then it is possible to construct a map Φ into an N-dimensional feature space H such that $k(x_i, x_j) = \langle\Phi(x_i), \Phi(x_j)$. Conversely, given an arbitrary map Φ into some feature space H, the matrix $K_{ij} = \langle\Phi(x_i), \Phi(x_j)\rangle$ is positive definite.*

This is a kernel map defined from pairwise similarities in the RKHS. For the sake of measuring the similarity between classes in the feature space, a kernel-induced distance between data sets in the input space needs to be defined. It can be expressed in the entries of the kernel matrix[16],

$$\begin{aligned}
&\|\Phi(x_i)-\Phi(x_j)\|^2\\
&=(\Phi(x_i)-\Phi(x_j))^T(\Phi(x_i)-\Phi(x_j))\\
&=\Phi(x_i)^T\Phi(x_j)-2\Phi(x_i)^T\Phi(x_j)+\Phi(x_j)^T\Phi(x_j)\\
&=k(x_i,x_i)+k(x_j,x_j)-2k(x_i,x_j). \qquad (1)
\end{aligned}$$

As mentioned above, it implies that similarity in the RKHS can be learned from metric distance.

3. Derivation of IEM

As we have seen in Section 2, by using the kernel k, similarity measure can be carried out implicitly in the feature space that Φ maps into, which can have a very high (maybe infinite) dimensionality. However, the non-metric distances will lead to inconsistency and conflict, when the metric violate one or more metric axioms as in Refs.[3,4]. So it is necessary to change this metric distance. This section will give a novel metric named as IEM in the case that it meets the metric axioms.

3.1. Problem description

To facilitate understanding, we regard each data sample in the RKHS as a physical particle experiencing force acting on it imposed by an overall “information energy” of the data set. This idea is similar to the Ref.[17], the main difference is that we add a parameter to control the granularity of the similar/dissimilar subsets which can presents the force acting on the considering particle. If we were given a candidate particle $y_{ci}=\Phi(x_{ci})$, the neighboring particles can be divided into two subsets

$$\begin{aligned}
S&=\{i\mid y_{ci}=\Phi(x_{ci}),\ c\in C,\ i=1,\ldots,k_s\},\\
F&=\{l\mid y_{pl}=\Phi(x_{pl}),\ p\in C, p\neq c,\ l=1,\ldots,k_f\}, \qquad (2)
\end{aligned}$$

where S and F are called the similar subset and dissimilar subset, respectively. k_s and k_f are the granularity parameters, which represent the similar and dissimilar neighboring particles of the candidate particle. Note that by the above definitions, two particles can be regarded as similar if they belong to the same class. However, they may have relatively big Euclidean distance, which could impair the performance of the learning machines. For example, Fig. 1(a) shows particles in manifold setting.

In this Figure, we can observe that two particles are similar to each other with Euclidean distance, but dissimilar in manifold setting. So we address this problem with an informative energy criterion, making the similarity setting become automatic and adaptive in nature.

3.2. Method deduction

As mentioned before, we have modeled the neighborhoods of the candidate particle. In this section, we discuss how to induce the proposed metric holds both for the Euclidean and manifold setting.

For this purpose, we learn a similarity measure in the RKHS. Considering the candidate particle y_{ci}, the learned distance function tries to put k_s similar particles close together and k_f dissimilar particles far away from each other. Then, the interactions between pairs of similar or dissimilar particle can be obtained, which are computed using energy function proposed following. Actually, similarity between two praticles in high-dimensional RKHS can be quantified by their energy, this is so called informative energy. Finally, our metric IEM can be derived from following three stages.

In the first stage of IEM, we propose two informative energy functions. The main idea is that we quantify the amount of information between particles according to their graph energy [18]. Our goal is to transform the kernel space so that the distance in the transformed space correlated with the difference of the labels of particles. So, we need to define the informative energy.

The graph energy in Ref. [18] is defined as

$$E(\sigma)=\frac{1}{Z}\sum_{i=1}^{N}G(x_j-x_i)H(x_j,x_i), \qquad (3)$$

where $G(y)=\exp\left(\dfrac{-y^Ty}{2\sigma^2}\right)$ is Gaussian kernel function, σ is the kernel width parameter, $Z=$

$\sum_{j=1}^{N}\sum_{i=1}^{N} G(x_j - x_i)$ is a normalization variable. H is an indicator function, its value is one when the particles x_i and x_j are in the same class, otherwise zero when they are in the different class.

If we were given a candidate particle $y_{ci} = \Phi(x_{ci})$ in the kernel space, we define its informative energy function according to the graph energy model as Eq. (3). The main difference is that we consider each sample in the kernel space as a particle, and pull or push other particles in the transformed space. This means that the resultant effect of a particle is the sum of the effects between the particle pairs in the same or different classes. For each particle we defined two informative energy functions: similar and dissimilar energy. The first one is computed as follows

$$E_c(y_{ci}) = \frac{1}{N}\sum_{j=1}^{J_c} G(y_{cj} - y_{ci}). \tag{4}$$

Then the dissimilar energy function considering particles in set F is computed as

$$E_{p\neq c}(y_{ci}) = \frac{1}{N}\sum_{p=1}^{N_c}\sum_{l=1}^{J_p} G(y_{pl} - y_{ci}). \tag{5}$$

These two informative energy functions vary between zero and one. A high E_c indicates that two particles in the same class are quite similar. But a low $E_{p\neq c}$ indicates that two particles in the different class are quite different. We can use these two values to quantify the amount of information between any particles.

In the second stage of IEM, we derive the informative energy model as an objective function. As mentioned above, we have the simple idea that $E_c(y_{ci})$ should be as large as possible, and $E_{p\neq c}(y_{ci})$ should be as small as possible. This can ensure that separation between the different classes and aggregation within the same classes. Then, the total resultant effect can be computed as

$$E_{\text{total}(y)} = \frac{1}{N}\sum_{c=1}^{N_c}\sum_{i=1}^{J_c} E_c(y_{ci}) - E_{p\neq c}(y_{ci}). \tag{6}$$

The above equation implies that the optimal transformation of the kernel space is achieved only when the first term is maximized and the second term is minimized. From the definition of the informative energy function, we observe that $E_c(y_{ci})$ and $E_{p\neq c}(y_{ci})$ are controlled not only by the particles in the same class but also those in the different classes. So we add a constrain condition $0 \leqslant \alpha \leqslant 1$ into Eq. (6) and denote $k_{\text{total}} = k_s + k_f$ as the total number of those particles influence the candidate particle. Then, it can be computed as

$$\alpha = \left[\left(1 - \frac{J_c}{k_{\text{total}}+1}\right)^2 + \sum_{\substack{p=1\\ p\neq c}}^{N_c}\left(\frac{J_p}{k_{\text{total}}+1}\right)^2\right]. \tag{7}$$

where $\frac{J_c}{k_{\text{total}}+1}$ is a priori probability of particles (in cth class) in all neighboring particles, so the first term represents the total effects from other particles except those in cth class. Simultaneously, the second term means the sum of other classes' individual effects. In this way, k_{total} can be considered as the number of neighborhood of the candidate particle. Then we can modify the objective function as

$$E_{\text{total}(y)} = \frac{1}{N}\sum_{c=1}^{N_c}\sum_{i=1}^{J_c} \alpha E_c(y_{ci}) - (1-\alpha)E_{p\neq c}(y_{ci}). \tag{8}$$

In the third stage, we reformulate the objective function of Eq. (8) as an instance of semidefinite programming [19,20] in the following steps.

Step 1: Reformulate the candidate particle in the kernel space as another form

$$y_{ci} = \langle v, \Phi(x_{ci})\rangle = \textstyle\sum_j Lk(x_j, x_{ci}),$$

where v is a projection vector, L is a coefficient matrix needs to be determined.

Step 2: For simplicity, we denote $k_{h,l} = (\sum_h k(x_h,x_i) - \sum_l k(x_l,x_j))$, then, similarity measure for those particles

$$\begin{aligned}(y_i - y_j)^T(y_i - y_j) &= k_{h,l}^T L^T L k_{h,l}\\ &= k_{h,l}^T M k_{h,l}\\ &= D_M(k(x_h,x_i), k(x_l,x_j)).\end{aligned}$$

In this way, we work in term of a new variable $M = L^T L$. With this change of variable, we can transform the similarity measure to a coefficient matrix learning problem.

Step 3: Reformulate Eq. (8). The first term in Eq. (8) penalizes large distance between each input and its similar particles. In terms of the coefficient matrix, it can be given by

$$\xi_{\text{pull}}(M) = \frac{1}{N}\sum_{c=1}^{N_c}\sum_{i=1}^{J_c}\alpha E_c(y_{ci}).$$

The second term penalizes small distance between dissimilar particles and can be given by

$$\xi_{\text{push}}(M) = \frac{1}{N}\sum_{c=1}^{N_c}\sum_{i=1}^{J_c}(\alpha-1)E_{p\neq c}(y_{ci}).$$

Then, we combine the two terms $\xi_{\text{pull}}(M)$ and $\xi_{\text{push}}(M)$ into a single informative function for similarity measure learning. The two terms can have competing effects, a weighting parameter $\mu \in [0,1]$ balances these goals

$$\xi_{\text{total}}(M) = (1-\mu)\xi_{\text{pull}}(M) + \mu\xi_{\text{push}}(M). \tag{9}$$

Generally, the parameter μ can be tuned via cross validation, though in our experience, the results from maximizing the informative energy function in Eq. (9) did not depend sensitively on the value of μ. In practice, the value $\mu = 0.4$ worked well.

Finally, we introduce a margin E_0, it is used to measure the amount of informative energy when push away the dissimilar particles. For convenient, we compute the proportion of the number of dissimilar particles and the total number N. Then, E_0 can be computed as

$$E_0 = \frac{k_f}{N},$$

and the objective function Eq. (9) can be formulated as a semidefinite programming problem

$$\max_M (1-\mu)\xi_{\text{pull}}(M) + \mu\xi_{\text{push}}(M) + E_0,$$
$$\text{s.t.} \begin{cases} (1)\ \ \xi_{\text{pull}}(M) + E_0 \leqslant \xi_{\text{push}}(M), \\ (2)\ \ M \geqslant 0. \end{cases} \tag{10}$$

While semidefinite programming problem in this form can be solved by standard solver packages [20], we can obtain the coefficient matrix M.

3.3. Characteristic of IEM

We show that IEM has some advantages which are useful for similarity measure and obtain the following theorems.

Firstly, we prove IEM meets the metric axioms.

Theorem 2. *(Informative Energy Metric, IEM) Given a feature space H and the informative energy metric E, E satisfies the metric axioms, for all $x,y,z \in H$.*

Proof. (1) Non-negativity and symmetry.

Without loss of generality, suppose the training sample x belongs to the cth class, y belongs to the pth class. We can compute the total informative energy as $E_{\text{total}(x,y)} = E_{\text{total}(x)} + E_{\text{total}(y)}$ according to Eqs. (4,5). We can get

$$\begin{cases} E_{\text{total}(x)} = \alpha E_c(x) + (1-\alpha)E_{p\neq c}(x), \\ E_{\text{total}(y)} = \alpha' E_p(y) + (1-\alpha')E_{c\neq p}(y). \end{cases} \tag{11}$$

Because we only consider two particles x and y, according to Eqs. (4,5), we know that $E_c(x) = E_p(y) = 1$ and $\alpha = \alpha' = 1/2$, so we can rewrite Eq. (11) to

$$E_{\text{total}(x,y)} = 1 - \frac{E_{p\neq c}(x) + E_{c\neq p}(y)}{2}.$$

From the characteristics of the Gaussian kernel function in Ref. [16], we know

$$E_{p\neq c} = E_{c\neq p} = G(x-y),$$

here $0 \leqslant G(x-y) \leqslant 1$, and $G(x-y) = G(y-x)$. Then, we can obtain

$$E_{\text{total}(x,y)} = 1 - G(x-y) = E_{\text{total}(y,x)} \geqslant 0.$$

(2) Distinguishability.

Assume that x belongs to the cth class, we can compute its informative energy as

$$E_{\text{total}(x,x)} = E_{\text{total}(x)} + E_{\text{total}(x)},$$

where $E_{\text{total}(x)} = \alpha E_c(x) - (1-\alpha)E_p(x)$.

Here, we only consider the candidate particle x, it has 0 neighborhood in the same or different classes. Then, from Eq. (7), we know that $J_c = 1$, $J_p = 0$, $k_{\text{total}} = 0$ and $\alpha = 0$. So, $E_{\text{total}(x)} = \alpha E_c(x) - (1-\alpha)E_p(x)$ is reformulated as $E_{\text{total}(x)} = \alpha E_c(x)$ according to Eq. (8). From Eqs. (4,5), $E_c = 1$, $E_p = 0$, we can obtain

$$E_{\text{total}(x,x)} = 0.$$

(3) Triangle inequality.

Assume that $x,\ y$ belong to the cth class, z belongs to pth class. Because we only consider three samples here, so we found that $J_c = 2$, $J_p = 1$, $k = 2$. We can obtain

$$\begin{cases} E_{\text{total}(x,z)} = \dfrac{2G(x-y)}{9} - \dfrac{8G(x-z)}{9} - \dfrac{G(y-z)}{9} + \dfrac{11}{9}, \\ E_{\text{total}(x,y)} = \dfrac{4G(x-y)}{9} - \dfrac{7G(x-z)}{9} - \dfrac{7G(y-z)}{9} + \dfrac{4}{9}, \\ E_{\text{total}(y,z)} = \dfrac{7G(x-y)}{9} - \dfrac{G(x-z)}{9} - \dfrac{3G(y-z)}{9} + \dfrac{15}{9}. \end{cases}$$

then we know

$$E_{\text{total}(x,z)} \leqslant E_{\text{total}(x,y)} + E_{\text{total}(y,z)}.$$

In conclusion, IEM meets the metric axioms. □

Secondly, we found that different setting of parameter α will lead to special versions of IEM metric, which are highly related to the popular criterion mutual information (MI)[17]. We have the following theorem.

Theorem 3. *Let α in Eq. (7) a constant, when $k_{total} = N-1$, then our objective function in Eq. (8) is the same as MI criterion.*

Proof. In Ref. [17], MI is computed by Renyi entropy, according to definitions in our method, we can rewrite the MI computed in Ref.[17] as

$$MI = V_{IN} + V_{ALL} - 2V_{BTW}, \tag{12}$$

where the quantities appearing in Eq. (12) are as follows

$$V_{IN} = \frac{1}{N^2} \sum_{p=1}^{N_c} \sum_{k=1}^{J_p} \sum_{l=1}^{J_p} G(y_{pk} - y_{pl}), \tag{13}$$

$$V_{ALL} = \frac{1}{N^2} \left(\sum_{p=1}^{N_c} \left(\frac{J_p}{N} \right)^2 \right) \sum_{k=1}^{N} \sum_{l=1}^{N} G(y_k - y_l), \tag{14}$$

$$V_{BTW} = \frac{1}{N^2} \sum_{p=1}^{N_c} \frac{J_p}{N} \sum_{j=1}^{J_p} \sum_{k=1}^{N} G(y_{pj} - y_k). \tag{15}$$

Substituting Eqs. (12), (13) and (14) into Eq. (11), we can obtain the MI values.

In our IEM, the objective function in Eq. (8) can be modified by step one in the third stage as

$$E_{\text{total}(M)} = \frac{1}{N} \sum_{c=1}^{N_c} \sum_{i=1}^{J_c} \alpha E_c(y_{ci}) - (1-\alpha) E_{p \neq c}(y_{ci}), \tag{16}$$

where α is computed as Eq. (7).

When we consider all $N-1$ particles in the kernel space as neighborhoods of certain particle y_{ci}, this means $k_{\text{total}} = N-1$. Then, we can computed the parameter in Eq. (7) as

$$\alpha = 1 + \left(\frac{J_c}{N} \right)^2 + \sum_{\substack{p=1 \\ p \neq c}}^{N_c} \left(\frac{J_p}{N} \right)^2 - \frac{2J_c}{N}. \tag{17}$$

Substituting Eq. (17) into Eq. (16), we rewrite the objective function as

$$E_{\text{total}(M)} = \frac{1}{N^2} \left(E^1 + E^2 + E^3 \right),$$

where E^1, E^2, E^3 are computed as follows

$$E^1 = \sum_{c=1}^{N_c} \sum_{i=1}^{J_c} \sum_{j=1}^{J_c} G(y_{cj} - y_{ci}),$$

$$\begin{aligned} E^2 = {} & \sum_{c=1}^{N_c} \left(\frac{J_c}{N} \right)^2 \sum_{i=1}^{J_c} \left(\sum_{j=1}^{J_c} G(y_{cj} - y_{ci}) \right. \\ & \left. + \sum_{\substack{p=1 \\ p \neq c}}^{N_c} \sum_{l=1}^{J_p} G(y_{pl} - y_{ci}) \right) \\ & + \sum_{\substack{p=1 \\ p \neq c}}^{N_c} \left(\frac{J_p}{N} \right)^2 \sum_{c=1}^{N_c} \left(\sum_{i=1}^{J_c} G(y_{cj} - y_{ci}) \right. \\ & \left. + \sum_{l=1}^{J_p} G(y_{pl} - y_{ci}) \right), \end{aligned}$$

$$\begin{aligned} E^3 = -2 & \sum_{c=1}^{N_c} \frac{J_c}{N} \sum_{i=1}^{J_c} \left(\sum_{j=1}^{J_c} G(y_{cj} - y_{ci}) \right. \\ & \left. + \sum_{\substack{p=1 \\ p \neq c}}^{N_c} \sum_{l=1}^{J_p} G(y_{pl} - y_{ci}) \right). \end{aligned}$$

For E^1, it computes informative energy in the same class, this is the same as V_{IN} in Eq. (13) except the normalization factor $1/N^2$, so, $E^1 = V_{IN}/N^2$.

For E^2, we can find that $\sum_{j=1}^{J_c} G(y_{cj} - y_{ci})$ computes informative energy for class $c(c \neq p)$, the other term $\sum_{l=1}^{J_p} G(y_{pl} - y_{ci})$ computes informative energy for class $p(p \neq c)$. The sum of these two terms can be merged into one term without considering the class label. Then, E^2 can be modified as

$$E^2 = \left(\sum_{p=1}^{N_c} \left(\frac{J_p}{N} \right)^2 \right) \sum_{k=1}^{N} \sum_{l=1}^{N} G(y_k - y_l).$$

For E^3,

$$\sum_{j=1}^{J_c} G(y_{cj} - y_{ci}) + \sum_{\substack{p=1 \\ p \neq c}}^{N_c} \sum_{l=1}^{J_p} G(y_{pl} - y_{ci})$$

computes informative energies for class $c(c \neq p)$ and class $p(p \neq c)$, they can be merged into one term as $\sum_{i=1}^{J_c} \sum_{k=1}^{N} G(y_k - y_{ci})$. Then, E^3 can be modified as

$$E^3 = -2 \sum_{p=1}^{N_c} \frac{J_p}{N} \sum_{j=1}^{J_p} \sum_{k=1}^{N} G(y_{pj} - y_k),$$

substituting E^1, E^2 and E^3 into $E_{\text{total}(M)}$, we obtain

$$E_{\text{total}(M)} = V_{IN} + V_{ALL} - 2V_{BTW}.$$

Compare E^1, E^2 and E^3 with Eqs. (13), (14) and (15), we know that when $k = N - 1$, $E_{\text{total}(M)}$ is equal to the value of MI.

Then, the MI criterion is a special case of our objective function as Eq. (16). □

As mentioned above, we can get the same criterion as MI, the main difference is that we can change the parameter k_{total} to obtain the high performance of similarity measure, and it meets the metric axioms. This characteristic is especially desirable for kernel-based methods such as those yield very large kernel matrices for important feature extraction. Experimental results show that IEM appears promising in the contexts of projection and classification tasks.

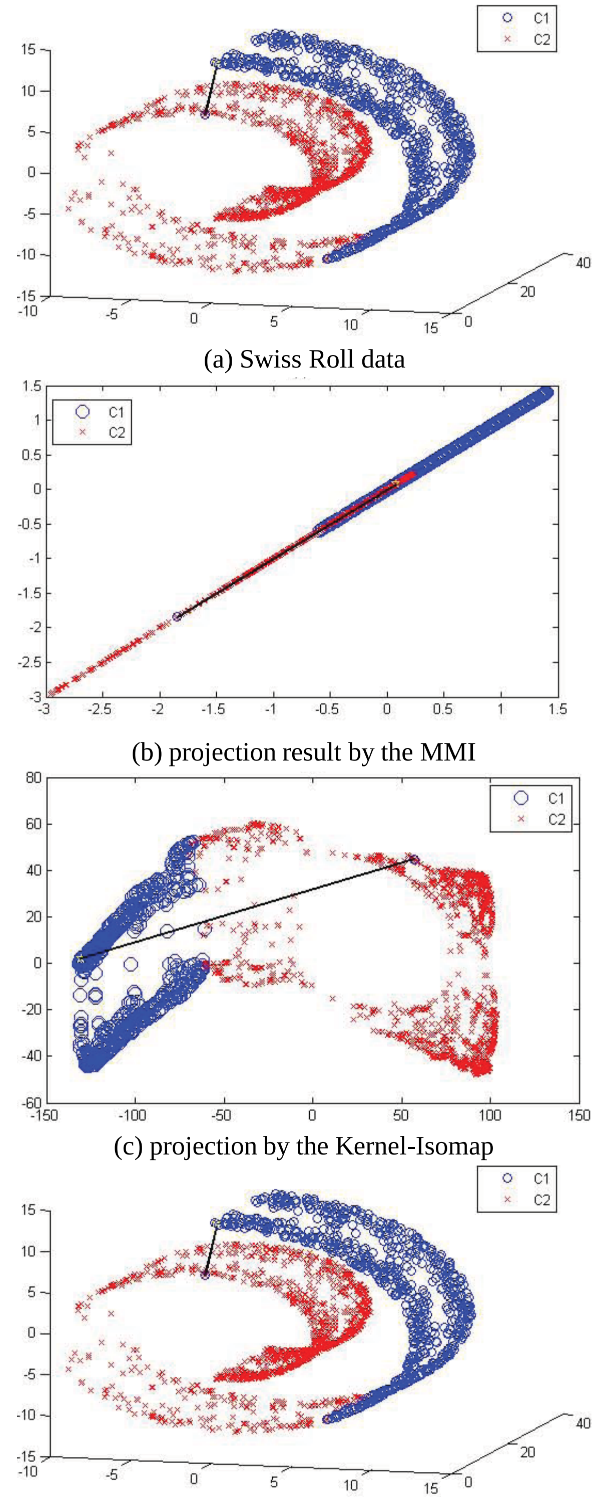
4. Applications of the main result

In this section, using the main result of Section 3, we demonstrate the validity of the proposed approach. In order to facilitate the comparison, we duplicate the main algorithm in Ref.[17] by ourselves, and denote their maximization mutual information as MMI. We will conduct two experiments on real benchmark data sets. The kernel width parameter σ is learned as the method described in Ref.[21]. Once the final coefficient matrix M is obtained, it can be used to low-dimensional visualization and classifier training. All experiments were run on the platform of Windows XP with 2.50GHz CPU and 2GB RAM using Matlab software.

4.1. Low-dimensional projection on manifold

We first evaluate whether our informative energy metric based feature space transformation method actually results in a new space that holds both for Euclidean distance and manifold setting. To this end, we evaluate whether our method improves visualization of low-dimensional projection task for Swiss Roll data.

The Swiss Roll data was used in Isomap and Kernel-Isomap, we used 2000 samples. In order to learning the informative energy metric for similarity measure, we defined “o” as the first class (C1) and “x” as the second class (C2). Two existing dimensional reduction methods are duplicated for comparison: maximization mutual information (MMI)[17] and Kernel-Isomap[22]. MMI finds a low-dimensional projection of the data points that best preserves their nonlinear features as measured in the high-dimensional input space. Kernel-Isomap is a manifold learning algorithm, which extends classical multidimensional scaling (MDS) by considering approximate geodesic distance instead of Euclidean distance. The parameter k_{total} is set to six, which is also used as neighborhood parameter in kernel-Isomap.



(a) Swiss Roll data

(b) projection result by the MMI

(c) projection by the Kernel-Isomap

(d) projection using IEM

Figure 1. Comparison of IEM with existing methods for the case of Swiss Roll data.

Fig. 1 shows the visualization results on Swiss Roll data. Fig. 1(a) shows the three-dimensional Swiss Roll data. We sampled two samples belong to different classes, which are connected by a line. These two samples can be considered as similar samples using Euclidean distance, but they are actually belong to the dissimilar set. Fig. 1(b) shows the projection result by the MMI, we can observe that two samples may be considered as in the same class, these two classes are highly overlapped. So we show the projection result by the kernel-Isomap in Fig. 1(c), which is a nonlinear dimensionality reduction method. From Fig. 1(c), we can observe that kernel-Isomap can find a smooth embedded manifold. Those two samples are separated as far as possible and can be considered belonging to the dissimilar set. Fig. 1(d) shows projection result by IEM, we can find that the samples of the first class are projected onto one line. Those two samples are moved away and can be useful for classification task.

As mentioned above, we can find that IEM can preserve the high-dimensional manifold features.

4.2. *Performance of classification*

In this experiment, we consider the classification task using the IEM in the RKHS. We use Statlog data for this problem. The Statlog data is the Statlog satellite image database cited from the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml), it has 4435 features for training, and 2000 for testing. Its dimensional is 36 and the number of classes is six, we randomly sample 1800 from the training data and use total testing data. Those six classes are according to the label in Fig. 2, where C1 represents red soil, C2 is cotton crop, C3 is grey soil, C4 is damp grey soil, C5 is soil with vegetation stubble, and C6 is very damp grey soil. This data is one of the many sources of information available for a scene. The interpretation of a scene by integrating spatial data of diverse types and resolutions including multispectral and radar data, is a difficult task.

In order to facilitate the visualization, the similarity measure method used here is MMI[17] and IEM. Once the coefficient matrix M is obtained, we can project test data using this coefficient matrix. Firstly, we show visualization of these test data in Figs. 2,3.

From the Fig. 2, we observed that MMI separates C2 and C5 but places other four classes almost on the top to each other, this is the same to the result in Ref.[17]. The criterion of IEM is a combination of representing each class as compactly as possible and as separated from each other as possible. Fig. 3 has achieved this: all classes are represented as quite compact clusters. But we should note that C1 is scattered and has small cross parts with other classes.

Fig. 3 shows the classification result of the IEM, we can find that C2 to C6 classes show better similarity and high aggregation except for C1.

In order to evaluate how those particles transform according to the IEM, we plot the gradient information of each particle. Fig. 3 shows the result of MMI, the direction is computed by partial differential of mutual information to the candidate particle. From this figure, we may find out why MMI cannot separate other four classes. The direction of each particle represents its transfer in the RKHS respect to the input space, it gives the trend of particle's aggregation. We can found that the directions of particles have no regular arrangement except C2, therefore, C1, C3 to C6 cannot be separated clearly.

For comparison, we computed partial differential of the energy function to the candidate particle. This can give the same result as in Fig. 4. Fig. 5 shows gradient information of the IEM, we observed that all classes can be separated clearly. From the direction of C1 in Fig. 4, we find out that particles in C1 have the tendency toward the within class center, and C1 has little overlapping area with other classes. This result is useful when the particle is difficult to classify. We can predict its potential class label according to its move direction in the RKHS.

Computed with the MMI, we can obtain the follow conclusions: (1) MMI can separate the cotton crop and vegetation stubble, but the other four soil features cannot be separated well. (2) IEM firstly separate the cotton crop from other soil features, the other four soil features are separated well. So our method can give better separation performance which is useful for classification task.

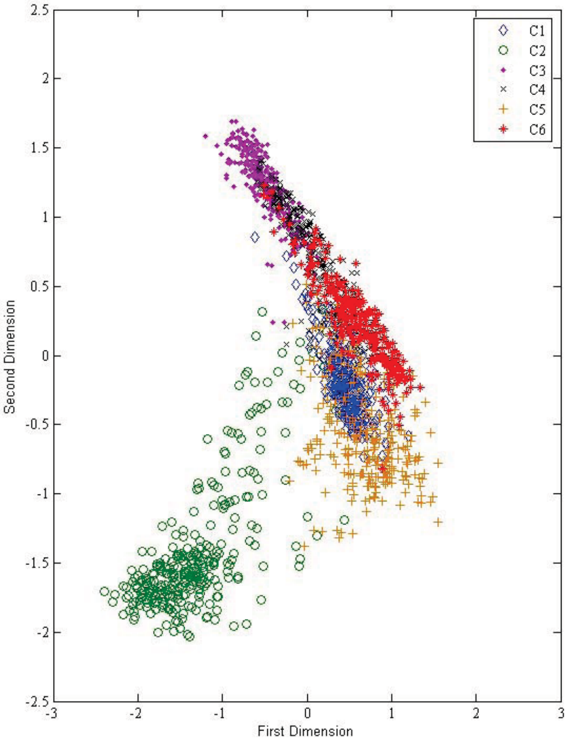


Figure 2. Classification result of the MMI

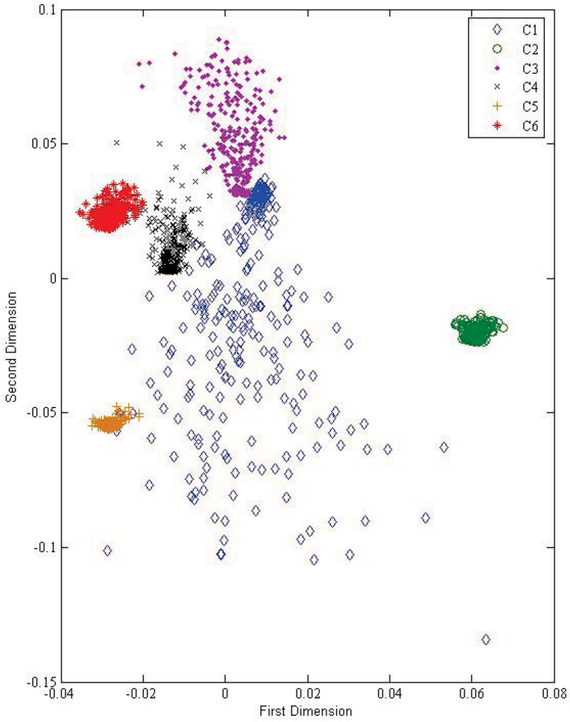


Figure 3. Classification result of the IEM

Published by Atlantis Press

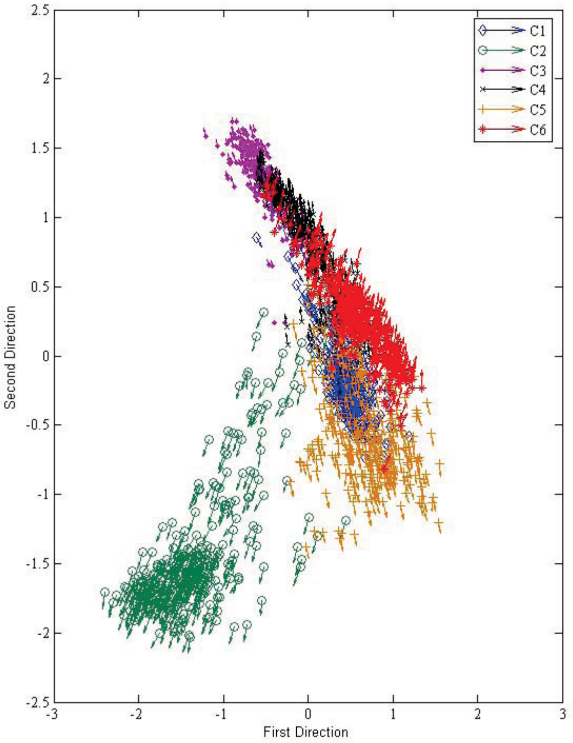


Figure 4. Gradient information of the MMI

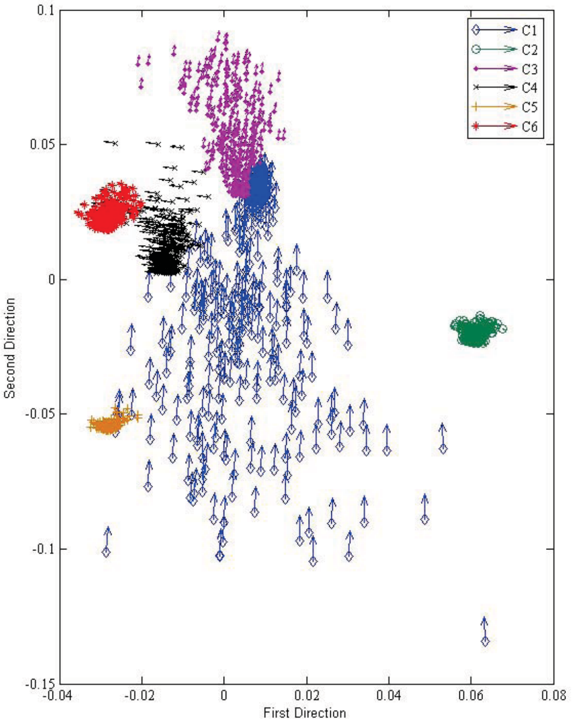


Figure 5. Gradient information of the IEM

Finally, we consider the influence of dimensionality to the classification performance. Here, we consider the Isolet data set, it is cited from the UCI Machine Learning Repository, which contains 6238 examples and 26 classes corresponding to letters of the alphabet, its dimensionality is 617. We compare our IEM with three existing methods: large margin nearest neighbor (LMNN)[23], MMI[17] and Multiclass SVM[24], kNN is incorporated in the classification stage in order to compute the final performance. Table 1 shows the test error rates on Isolet, we averaged over 10 runs with random sampling of 2000 training examples and its original testing samples.

Table 1. Benchmark test error rates on Isolet data

Dimensionality	17	51	85	119	172
LMNN	17.65	7.84	6.64	5.93	5.13
Multiclass SVM	-	-	-	-	3.40
MMI	23.25	15.22	12.13	6.45	5.84
IEM	14.00	6.25	5.56	4.83	3.96

From Table 1, we can observe that IEM is comparable to the Multiclass SVM. The result of Multiclass SVM is cited from Ref.[23]. We also note that IEM can obtain high classification accuracy under low-dimensional setting.

5. Conclusions

In this paper, we address the non-metric distance problem under manifold setting, a new similarity metric in the RKHS is presented. Based on this new metric, we show how it can be applied for manifold setting tasks. Experimental results show its validity.

Other advantages of IEM are

- IEM can extract high order statistics and nonlinear statistics from data sets. This method holds both for Euclidean and manifold setting,
- IEM makes a crucial addition to the MMI in Ref.[17]. The main difference is that we add a constrain condition α, and k_{total} is used to set the value of the constrain condition. Experimental results show that IEM can give better performance on manifold setting and Statlog satellite image database,

- IEM can be used as a quantification of sample similarity. It can improve the visual ability of projection, and can be used for classification task. We expect that this metric can be used in bioinformatics in future work.

In future work we intend to apply the proposed method to bioinformatics. We also want to modify our method to parallel implementations, which can alleviate the high computational complexity of semidefinite programming. In addition, how to choose the parameters used in this method is an interesting topic.

Acknowledgments

The authors are grateful to the Editor and the referees for their helpful comments and suggestions. This work was supported by the Doctoral Foundation of Jiangxi University of Science and Technology, Jiangxi Educational Committee Foundation for Youths, National Natural Science Foundation of China (61070137,60702063).

References

1. S. Santini, R. Jain, “Similarity measures,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **21** (9),871–883 (1999).
2. S. Kevin Zhou, R. Chellappa, “From sample similarity to ensemble similarity: probabilistic distance measure in reproducing kernel Hilbert space,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28** (6),917–929 (2006).
3. M. R. Ackermann, J. Bolmer, C. Sohler, “Clustering for metric and non-metric distance measures,” *in: Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms*, San Francisco, California, 799–808 (2008).
4. Y. Zhang, Z. H. Zhou, “Non-metric label propagation, ” *in: Proceedings of the 21th International Joint Conference on Artificial Intelligence (IJCAI’09)*, Pasadena, CA, 1357–1362 (2009).
5. S. Muthukrishnan, S. C. Sahinalp, “Approximate nearest neighbors and sequence comparison with block operation, ” *in: Proceedings of 32nd ACM Symposium on Theory of Computing*, Portland, OR, 416–424 (2000).
6. C. H. Bennett, P. Gǻcs, M. Li, P. M. B. Vitányi, W. Zurek, “Information distance, ” *IEEE Transactions on Information Theory*, **44** 1407–1423 (1998).
7. X. Tan, S. Chen, Z. H. Zhou, J. Liu, “Learning non-metric partial similarity based on maximal margin criterion, ” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, **2** 138–145 (2006).
8. S. Agarwal, J. Wills, L. Cayton, G. Lanckriet, D.Kriegman, S. Belongie, “Generalized non-metric multidimensional scaling,” *in: 11st International Conference on Artificial Intelligence and Statistics*, San Juan, Puerto Rico, (2007).
9. T. M. Cover, J. A. Thomas, “Elements of information theory, ” *John Wiley* & *Sons*, (1985).
10. B. A. Frigyik, S. Srivastava, M. R. Gupta, “Functional Bregman divergences and bayesian estimation of distributions, ” *IEEE Transactions on Information Theory*, **54** 5130–5139 (2008).
11. V. Athitsos, J. Alon, S. Sclaroff, “Efficient nearest neighbor classification using a cascade of approximate similarity measures,” *in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, Washington, DC, USA, 486–493 (2005).
12. H. A. Boubacar, S. Lecoeuche, “A new kernel-based algorithm for online clustering, ” *in: International Conference on Artificial Neural Networks (ICANN’05)*, Warsaw, Poland, 583–588 (2005).
13. S. Ilhan, N. Duru, E. Adali, “Improved fuzzy art method for initializing k-means,” *International Journal of Computational Intelligence Systems*, **3** (3) 274–279 (2010).
14. B. Schölkopf, A. J. Smola, “Learning with kernels, ” *MIT Press*, Cambridge, MA, (2002).
15. J. Mercer, “Function of positive and negative type and their connection with the theory of integral equations,” *Philosophical Transactions of the Royal Society*, London, **A 209** 415–446 (1909).
16. M. S. Baghshah, S. B. Shouraki, “Non-linear metric learning using pairwise similarity and dissimilarity constraints and the geometrical structure of data, ” *Pattern Recognition*, **43** (8) 2982–2992 (2010).
17. K. Torkkola, “Feature extraction by non-parametric mutual information maximization, ” *Journal of Machine Learning Research*, **3** 1415–1438 (2003).
18. Y. Lin, C. Lin, Y. Tsai, T. Ku, Y. Huang, C. Hsu, “A spectral graph theoretic approach to quantification and calibration of collective morphological differences in cell images,” *Bioinformatics*, **26** i29–i37 (2010).
19. B. Stephen, V. Lieven, “Convex optimization, ” *Cambridge Unversity Press*, (2004).
20. L. Vandenberghe, S. Boyd, “Semidefinite programming,” *SIAM Review*, **38** (1) 49–95 (1996).
21. B. Schölkopf, A. J. Smola, K.-R. Müller, “Nonlinear component analysis as a kernel eigenvalue problem,” *Neural Computation*, **10** 1299–1319 (1998).
22. H. Choi, S. Choi, “Kernel Isomap,” *Electronics Letters*, **40** (25) 1612–1613 (2004).

23. K. Q. Weinberger, L. K. Saul, “Distance metric learning for large margin nearest neighbor classification,” *Journal of Machine Learning Research*, **10** 207–244 (2009).

24. K. Crammer, Y. Singe, “On the algorithmic implementation of multiclass kernel-based vector machines,” *Journal of Machine Learning Research*, **2** 265–292 (2001).