Local Discriminative Distance Metrics and Their Real World Applications

Yang Mu, Wei Ding
Department of Computer Science
University of Massachusetts Boston
Boston, Massachusetts 02125–3393
{yangmu, ding}@cs.umb.edu

***Abstract*—The ultimate goal of distance metric learning is to use discriminative information to keep data samples in the same class close, and those in different classes separate. Local distance metric methods can preserve discriminative information by considering neighborhood influence. We propose a discriminative distance metric approach by maximizing local pairwise constraints. Based on the local learning framework, we then extend this approach to a multiple metrics approach, local discriminative distance metrics (LDDM), by learning distance metrics on the local vicinity of each training sample. This extension avoids the global optimization for irrelevant pairwise constraints and can thus maximize the discriminative information in each local area. Theoretical analysis for the error bound of the proposed methods has been provided. In addition, we have studied three challenging real-world problems: crater detection, crime prediction, and accelerometer based activity recognition. We design and apply three local distance learning metrics to achieve the best performance for each particular task.**

I. Introduction

Real world applications usually possess various unique properties, and hence any single algorithm cannot apply to all real world problems. In this paper, we work on three different types of real world applications: a Mars crater detection project funded by National Aeronautics and Space Administration (NASA), a crime prediction project funded by Department of Justice (DOJ) and an accelerometer based activity prediction project funded by National Institutes of Health (NIH).

Many machine learning and data mining approaches can be used to analyze these real world tasks, e.g., K-means, K-Nearest Neighbors, kernel SVMs [16], [19], [21]. Among these metric-related approaches, distance metric learning plays a crucial role.

These distance learning tasks fall into two categories: unsupervised and supervised distance metric learning. In supervised distance metric learning [18], the ultimate goal is to use discriminative information in distance metric learning to keep all the data samples in the same class close and those from different classes separated. Zhang *et al.* [20] have shown that a distance metric incorporating discriminative information from labeled data usually outperforms the standard Euclidean distance in classification tasks.

Supervised distance metric learning can be further divided into global and local distance metric learning. The first step is to learn a global distance metric from training data to satisfy all pairwise constraints simultaneously [22], [17]. The most representative work is Xing's algorithm [17], which learns a distance metric on a global scale which minimizes the distance between data pairs according to equivalence constraints, while separating data pairs from each other according to inequivalence constraints. If data classes exhibit multimodal distributions, equivalence or inequivalence constraints from different data distributions may conflict with each other. Therefore, it is difficult to satisfy all the constraints on a global level. Local distance metric learning is introduced to cope with this problem by considering the locality of data distribution [13], [15], [5]. These local algorithms only consider neighboring pairwise constraints and avoid adopting conflicting constraints. By incorporating neighboring constraints, many approaches have achieved great successes in manifold learning [7], [11].

All aforementioned approaches try to learn a single metric on all data samples. The deficiencies of learning a single metric include: 1) a single metric is likely to be inappropriate for pairwise constraints from all training samples, and many pairwise constraints may be irrelevant to others; 2) a single local metric may be easily influenced by pairwise constraints from noisy samples; 3) a single global metric cannot deal with the multimodal distribution problem. It is recommended to learn multiple metrics to describe different localities of training samples [6], [15], [2], [3].

In this paper, we firstly propose a local distance metric approach and then show how it can be extended to a multiple local distance metrics technique. A key strategy for the distance learning approach is to design appropriate distance metrics based on the particular data properties of real world data. For example, global distance metrics can explore structure information and is suitable for datasets with a single mode distribution; a single local distance metric can be designed for a dataset with multimodal distribution but it cannot handle noisy data samples and irrelevant local constraints. On the other hand, multiple local distance metrics can cope with noisy samples as well as extensive irrelevant local constraints, which is often the case of classification with many classes such that the constraints from one class are irrelevant to the constraints from other classes. In addition, for multidimensional data, it is necessary to extend the vector based distance metric approaches to tensor form in order to take into account the geometric location of the data in spatial-temporal domain.

Based on the proposed solution, we have studied three challenging real-world problems: crater detection, crime prediction and accelerometer based activity recognition. We design and apply three local distance learning metrics to achieve the best performance for each particular task. Crater detection from remote sensing images is an important task in planetary science, since impact craters are topographic features on planetary surfaces resulting from impacts of meteoroids and crater counts are the only available tool for measuring remotely the relative ages of geologic formations on planets [10]. Crater detection is a task of binary classification (a crater or not a crater) and for each class it may obey multimodal distributions due to different crater formations, and so we apply a single local distance metric for this problem. Crime prediction is also a binary classification problem but the data involves spatial-temporal information. When using the vector based approach, the geometric structure will be broken [8]. To address this problem, we extend our single local distance metric approach to tensor form in order to directly deal with tensor inputs.

The rest of the paper is organized as follows: Section 2 gives the formal definition of the distance metric problem. Section 3 explains how to learn a single local discriminative distance metric. Section 4 discusses our multiple local discriminative distance metrics approach and Section 5 provides theoretical analysis. Section 6 gives our local distance metric learning applications to three real world applications. Section 7 concludes the paper.

II. Definition of Distance Metric

A. Problem Definition

Given a set of d-dimensional training samples $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]$, where $\{x_i\}_1^n \in \mathbb{R}^d$, and their associated labels $\mathbf{Y} = [y_1, y_2, \cdots, y_n]$, a generic distance metric to measure two samples $\mathbf{x}_i, \mathbf{x}_j$ is in the form of

$$d_A(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_A = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T A (\mathbf{x}_i - \mathbf{x}_j)}, \quad (1)$$

where A is positive semi-definite, and parameterizes a family of Mahalanobis distances [17]. Technically, it allows pseudo-metrics, such that $d_A(\mathbf{x}_i, \mathbf{x}_j) = 0$ does not imply $\mathbf{x}_i = \mathbf{x}_j$.

Replacing A with $\mathbf{W}^T\mathbf{W}$ in Equation (1), where $\mathbf{W} = \mathbf{A}^{1/2}$, we get:

$$\begin{aligned} d_A(\mathbf{x}_i, \mathbf{x}_j) =& \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{W}\mathbf{W}^T (\mathbf{x}_i - \mathbf{x}_j)} \\ =& \left\|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\right\|. \end{aligned} \quad (2)$$

To solve the matrix $\mathbf{A}$ in Equation (1) or $\mathbf{W}$ in Equation (2), there are two basic approaches: 1) using the structural information based approach, e.g., PCA [4], LDA [4]; 2) using the pairwise constraints based approach. Structure information based approaches are mostly global approaches, while pairwise constraints based approaches can be global or local depending on whether to use the global or local pairwise constraints.

B. Pairwise constraints

The pairwise constraints based approaches have achieved great performance [15]. Pairwise constraints consist of two parts: similar pairs and dissimilar pairs. We can have certain pairs of them being "similar" and "dissimilar" based on their labels (supervised learning) or geo-location information (unsupervised learning). Under a desired distance metric, similar samples have a smaller distance while dissimilar samples have a larger distance. Distance metric approaches either minimize the similarity constraints as well as penalize the dissimilarity constraints or maximize the dissimilarity constraints together with constraining the similarity constraints. Based on this property, we can form the following objective function:

$$\begin{aligned} &\arg\min_{\mathbf{A}} \textstyle\sum_{(\mathbf{x}_i, \mathbf{x}_j) \in S} d^2_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_j), \\ &s.t. \textstyle\sum_{(\mathbf{x}_i, \mathbf{x}_j) \in D} d^2_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_j) \geq \theta \end{aligned} \quad (3)$$

where θ is a parameter, S and D are the similarity and dissimilarity sets, respectively. S contains pairs of samples that share the same class label or have closed geo-location distance, while D contains pairs of samples with different class labels or have large geo-location distance.

The global distance metric approaches construct S and D using the entire training dataset, e.g., Xing [17], while the local distance metric approaches only consider pairs of samples in a local vicinity [9]. The different forms of S and D distinguish the global and local approaches.

In addition, as shown in Equation (3), the main objectives of different algorithms are the same. The merits of differences are embodied in the constraint term θ, which optimizes the dissimilarity constraints. For example, typical global approach Xing's method [17] sets $\theta = 1$, while LMNN[15], a representative local approach, has

$$\theta = 1 + \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in S} d^2_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_j).$$

Comparing different definitions of θ, we can tell LMNN has a stronger constraint on the dissimilar pairs. A larger θ corresponds to a longer pairwise distance between dissimilar samples. From this aspect, LMNN is expected to incorporate more discriminative information, since dissimilar samples are forced to have longer distance and thus form a larger margin between data samples in two different classes.

III. Discriminative distance metric

In this section, we propose a more aggressive constraint on dissimilar samples by simply maximizing the θ and minimizing the similar constraints simultaneously.

Here we define two new objective functions to minimize/maximize the distance of two data samples in same/different class:

$$\arg\min_{\mathbf{A}} \textstyle\sum_{(\mathbf{x}_i, \mathbf{x}_j) \in S} d^2_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_j), \quad (4)$$

and

$$\arg\max_{\mathbf{A}} \textstyle\sum_{(\mathbf{x}_i, \mathbf{x}_j) \in D} d^2_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_j). \quad (5)$$

Using a parameter β to align two objectives in Equations (4) and (5), we have

$$\arg\min_{\mathbf{A}} \textstyle\sum_{(\mathbf{x}_i, \mathbf{x}_j) \in S} d^2_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_j) - \beta \textstyle\sum_{(\mathbf{x}_i, \mathbf{x}_j) \in D} d^2_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_j). \quad (6)$$

Substituting Equation (2) into Equation (6), we have

$$\begin{aligned}\arg\min_{\mathbf{W}} \quad & \textstyle\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in S} \left\|\mathbf{W}^T(\mathbf{x}_i-\mathbf{x}_j)\right\|_2^2 \\ & -\beta \textstyle\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in D} \left\|\mathbf{W}^T(\mathbf{x}_i-\mathbf{x}_j)\right\|_2^2 .\end{aligned} \tag{7}$$

To learn a local distance metric from Equation (7), we form the similar pairwise constraint set S and the dissimilar pairwise constraint set D by incorporating neighborhood information. Sets S_i and D_i are constructed for each training sample $\mathbf{x}_i$ as:

$$S_i = \{(\mathbf{x}_i, \mathbf{x}_j)|\mathbf{x}_j \in \mathbf{X}_i, y_i = y_j\},$$

and

$$D_i = \{(\mathbf{x}_i, \mathbf{x}_j)|\mathbf{x}_j \in \mathbf{X}_i, y_i \neq y_j\},$$

where $\mathbf{X}_i$ is the set of samples in $\mathbf{x}_i$'s local vicinity [9], also known as a local patch in [20], which contains $\mathbf{x}_i$ the k_1 samples with the same class label of $\mathbf{x}_i$ and k_2 samples with the different class label of $\mathbf{x}_i$. The projection matrix $\mathbf{W}$ to optimize the pairwise constraints in the local vicinity of $\mathbf{x}_i$ can be defined as

$$\begin{aligned}\arg\min_{\mathbf{W}} \quad & \textstyle\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in S_i} \left\|\mathbf{W}^T(\mathbf{x}_i-\mathbf{x}_j)\right\|_2^2 \\ & -\beta \textstyle\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in D_i} \left\|\mathbf{W}^T(\mathbf{x}_i-\mathbf{x}_j)\right\|_2^2 .\end{aligned} \tag{8}$$

Reorganizing Equation (8) in matrix form, we have the equivalent optimization problem:

$$\arg\min_{\mathbf{W}} \text{tr}(\mathbf{W}^T\mathbf{X}_i\mathbf{L}_i\mathbf{X}_i^T\mathbf{W}), \tag{9}$$

where $\mathbf{L}_i \in \mathbb{R}^{(k_1+k_2+1)\times(k_1+k_2+1)}$ is defined as

$$\mathbf{L}_i = \begin{bmatrix} \sum_{j=1}^{k_1+k_2}(w_i)_j & -w_i^T \\ -w_i & diag(w_i) \end{bmatrix}, \tag{10}$$

and the w_i is the coefficient vector defined as

$$w_i = \begin{bmatrix} \overbrace{1,\cdots,1}^{k_1} & \overbrace{-\beta,\cdots,-\beta}^{k_2} \\ \end{bmatrix} \tag{11}$$

Since Equation (9) optimizes the pairwise constraints in sample $\mathbf{x}_i$'s vicinity, after summing over all the local vicinities, we have the equivalent optimization form for Equation (7):

$$\arg\min_{\mathbf{W}} \text{tr}(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}), \tag{12}$$

where $\mathbf{X}\mathbf{L}\mathbf{X}^T = \sum_i \mathbf{X}_i\mathbf{L}_i\mathbf{X}_i^T$

To make the projection matrix $\mathbf{W}$ linear and orthogonal, we impose $\mathbf{W}^T\mathbf{W} = \mathbf{I}_d$, where $\mathbf{I}_d$ is a $d \times d$ identity matrix. Equation (12) is then deformed to:

$$\min \ \text{tr}\left(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}\right) \ s.t. \mathbf{W}^T\mathbf{W} = \mathbf{I}_d. \tag{13}$$

Solutions of Equation (13) can be obtained with the standard eigen-decomposition:

$$\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{u} = \lambda\mathbf{u}. \tag{14}$$

Let the column vectors $\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_d$ be the solution of Equation (14), ordered according to eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_d$. The optimal projection matrix $\mathbf{W}$ is then given by: $\mathbf{W} = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}'_d]$, where $d' < d$. Once $\mathbf{W}$ is calculated, the local discriminative distance metric $\mathbf{A}$ can be calculated using Equation (2).

IV. Local Discriminative distance metrics

A. Multiple metrics

If only one distance metric is used to describe the whole training space, the tradeoff between the learning system and the number of samples may limit the learning performance [14]. The performance of this learning system can be measured as VC-dimension, which depicts the maximum number samples this learning system can shatter. From this aspect, it is well motivated to use multiple learning systems with each learning system only taking effect on a small portion of the data[14]. A typical classifier using multiple local learning systems is K-Nearest-Neighbor (KNN), which uses a small portion of the data to form a prediction on a local area.

A distance metric describes a distance space to be used for a learning system. Since there are multiple local learning systems, using multiple local distance metrics seems to be a natural extension. From the local learning theory, one single learning system cannot shatter all data samples. Similarly, one single distance metric cannot fulfill all pairwise constraints, which means some constraints have to be compromised during optimization.

To avoid pairwise constraints being compromised, we adopt a local optimization for each sample $\mathbf{x}_i$. In one optimization procedure, we only optimize those pairwise constraints containing sample $\mathbf{x}_i$. Other irrelevant constraints are excluded in the optimization so that the local vicinity of $\mathbf{x}_i$ can reflect our full expectation on this local area without any interference.

As shown in Equation (9), with the constraint $\mathbf{W}_i^T\mathbf{W}_i = I_d$, we can have a local discriminative distance metric $\mathbf{W}_i$ for the vicinity of $\mathbf{x}_i$ as

$$\arg\min_{\mathbf{W}_i} \text{tr}(\mathbf{W}_i^T\mathbf{X}_i\mathbf{L}_i\mathbf{X}_i^T\mathbf{W}_i), \ s.t. \mathbf{W}_i^T\mathbf{W}_i = \mathbf{I}_d. \tag{15}$$

B. A Probabilistic Approach for Ensemble Classifiers

Based on Equation (15), a distance metric $\mathbf{A}_i$ is defined on sample $\mathbf{x}_i$ as a *focal sample*.

Given an unknown test sample $\mathbf{x}_j$, let o be the class label of focal sample $\mathbf{x}_i$, the number of possible classes is N_o, the probability of $\mathbf{x}_j$ belonging to the class o, $Pr_i(o|\mathbf{x}_j)$, using the local distance metric $\mathbf{A}_i$ of the i^{th} focal sample $\mathbf{x}_i$ is

$$Pr_i(o|\mathbf{x}_j) = \begin{cases} \frac{\sum_{k=1}^{n}\{\theta(\mathbf{x}_k \in \mathbf{V}(\mathbf{x}_i))\theta(y_k = o)\}}{\sum_{k=1}^{n}\theta(\mathbf{x}_k \in \mathbf{V}(\mathbf{x}_i))} & \text{if } \mathbf{x}_j \in \mathbf{V}_K(\mathbf{x}_i) \\ \frac{1}{N_o} & \text{otherwise} \end{cases} \tag{16}$$

where $\mathbf{V}_K(\mathbf{x}_i)$ is the local vicinity of training sample $\mathbf{x}_i$ which contains K nearest neighbors of $\mathbf{x}_i$ with respect to the learned local distance metric $\mathbf{A}_i$. $\theta(\cdot)$ is an indicator function that returns 1 when the input argument is true, and 0 otherwise. $\theta(\mathbf{x}_j \in \mathbf{V}_K(\mathbf{x}_i)) = 1$ indicates $\mathbf{x}_j$ is among K nearest neighbors of $\mathbf{x}_i$ with respect to $\mathbf{A}_i$, which is calculated in Equation (14). Otherwise, the focal sample $\mathbf{x}_i$ has no influence on the unknown test sample $\mathbf{x}_j$. $\mathbf{V}(\mathbf{x}_i)$ defines a circular clique whose center is the focal sample $\mathbf{x}_i$. The radius r is the distance between the focal sample and the test sample $\mathbf{x}_j$ under the learned local distance metric $\mathbf{A}_i$. Probability $Pr_i(o|\mathbf{x}_j)$ is calculated as purity of circular clique $\mathbf{V}(\mathbf{x}_i)$. Please note that we propose a new prediction method in Equation (16) instead of the traditional KNN rules because of our objective function defined in Equation (15). We expect the vicinity of the focal sample to contain as many similar samples as possible. In this case, if a test sample is not in the K nearest neighbors of the focal sample, it is expected not to be similar to the focal sample. The metric is expected to *pull* the samples with the same/different label as the focal sample $\mathbf{x}_i$ closer to/away from $\mathbf{x}_i$. Note that if the test sample $\mathbf{x}_j$ is the closest sample to $\mathbf{x}_i$ in $\mathbf{V}_K(\mathbf{x}_i)$, the probability is 1 for the test sample $\mathbf{x}_j$ to be assigned as the same class label as $\mathbf{x}_i$.

As illustrated in Figure 1 from our published paper [9] , because the clique of the red circle $\mathbf{V}(\mathbf{x}_i)$ contains a focal sample, four red circles and one blue square, probability for the test sample belonging to the red circle class is $\frac{5}{6}$.

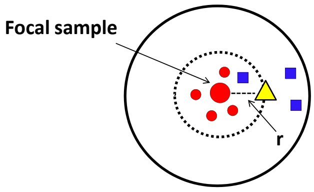

Fig. 1. Local distance metric prediction. Red circles and blue squares belong to two classes. The yellow triangle is an unknown test sample. The red circle in the center is the focal sample $\mathbf{x}_i$. Figure illustrates the local distance metric space $\mathbf{A_i}$ learned from the focal sample and its vicinity. The solid-line circle is $V_K(\mathbf{x}_i)$ and the dashed-line circle represents $V(\mathbf{x}_i)$. The probability for the yellow triangle belonging to the red circle class is the number of red circles in $V(\mathbf{x}_i)$ divided by total number of training samples in $V(\mathbf{x}_i)$.

We can obtain a set of locally learned classifiers described in a different data space, using the local classifier defined in Equation (16) under each local distance metric. This approach makes these local classifiers independent of each other to facilitate the alignment operation. Each obtained local distance metric best measures the vicinity of the focal sample and places the same class samples close to the focal sample and the different-class samples far away from the focal sample. To make the training model adjustable according to different test samples, we add a weight coefficient ϕ when combining n local prediction $Pr_i(o|\mathbf{x}_j)$ for a given test sample $\mathbf{x}_j$. Weight ϕ is decided by the distance between the test sample and focal sample. A final prediction is made by aligning n outputs in a probabilistic framework. The alignment process is formally defined as

$$Pr(o|\mathbf{x}_j) = \frac{1}{n}\sum_{i=1}^{n} \phi_i Pr_i(o|\mathbf{x}_j), \tag{17}$$

where n is the number of classifiers and $Pr_i(o|\mathbf{x}_j)$ is the probability of sample $\mathbf{x}_j$ belonging to class o predicted by the i^{th} local classifier. To simplify this process, we give all the training samples equal weights by letting $\phi_i = 1$. This makes the ensemble process behave like equal weight voting. The class label with the highest probability is the final label of test sample.

An overall summary of our local discriminative distance metrics (LDDM) method is described in Algorithm 1. In the training procedure, we need to calculate $\mathbf{W}_i$ by decomposing a $(k_1+k_2+1)\times(k_1+k_2+1)$ matrix $\mathbf{X}_i\mathbf{L}_i\mathbf{X}_i^T$ in Equation (14) for each focal sample $\mathbf{x}_i$ which has time complexity $O(n(k_1+k_2+1)^3)$. When testing an unknown sample, it is linear time $O(n)$ to the training set size, since all the local distance metrics were already obtained in the training phase. The test time complexity only depends on Equation (16) and Equation (17) which is just the ensemble of results of n training samples using pre-calculated local distance metrics. Note that the projection for all the training samples into the distance metric space can be conducted in the training phase. Despite the high training cost, we can parallelize the proposed model to make it scalable for large-scale problems. Local classifiers could also be learned offline in advance. For detailed performance and efficiency comparisons between LDDM and other distance metric approaches please refer to our paper published on Pattern Recognition [9].

Algorithm 1 LDDM: a multiple distance metrics approach for classification

```
Training procedure
for each training sample $\mathbf{x}_i$ do
    Get the focal vicinity $\mathbf{X}_i$ for $\mathbf{x}_i$
    Build the discriminative matrix $\mathbf{L}_i$ using Equation (10)

    solve the projection matrix $\mathbf{W}_i$ by Equation (14)
end for
Test procedure
for each test sample $\mathbf{x}_j$ do
    Calculate the probability of $\mathbf{x}_j$ belonging to class $o$
    when using the training sample $\mathbf{x}_i$ as the focal sample,
    $Pr_i(o|\mathbf{x}_j)$ by Equation (16)
    Ensemble all the predictions by different training samples according to Equation (17)
end for
```

V. Theoretical Analysis

We now theoretically prove the stability and efficiency of the proposed LDDM method by analyzing the convergence rate of the local discriminative distance metric and generalization bound of the local metrics and classifiers ensemble.

We assume that all the samples and their labels can be represented by an unknown distribution $F(\mathbf{x},\mathbf{y})$, defined by pairs $(\mathbf{x},\mathbf{y}) \in \mathbb{R}^d \times \mathbb{R}^1$. The pair $(\mathbf{x},\mathbf{y})$ is denoted as $\mathbf{z}$ for short. Model $\mathbf{x} \rightarrow f(\mathbf{x},\alpha)$ of the output $\mathbf{y}$ is controlled by a parameter $\alpha \in \Lambda$. $f(\mathbf{x},\alpha)$ refers to the local classifier defined in Equation (16) for LDDM. The $0-1$ loss function $Q(\mathbf{y}, f(\mathbf{x},\alpha))$ (or $Q(\mathbf{z},\alpha)$ for short) measures the quality of estimation by $f(\mathbf{x},\alpha)$ for output $\mathbf{y} \in \{-1,+1\}$. The global risk function is defined as

$$R(\alpha) = \int Q(\mathbf{z},\alpha)dF(\mathbf{z}) \tag{18}$$

over all functions $\{f(\mathbf{x},\alpha), \alpha \in \Lambda\}$, and samples $\{\mathbf{z}_i\}_{i=1}^n$ are independently drawn from the unknown distribution $F(\mathbf{z})$.
The empirical risk function with respect to the training samples

$\{\mathbf{z}_i\}_{i=1}^n$ is

$$R_{emp}(\alpha) = \frac{1}{n}\sum_{i=1}^{n} Q(\mathbf{z}_i, \alpha). \tag{19}$$

In local algorithms, the local risk function $R(\alpha, \mathbf{x}_0)$ depends on the focal sample $\mathbf{x}_0$ and the vicinity of $\mathbf{x}_0$. The nonnegative locality function $D(\mathbf{x}, \mathbf{x}_0, A)$, which embodies the vicinity information of the focal sample, is defined as

$$D(\mathbf{x}, \mathbf{x}_0, A) = \begin{cases} 1 & \text{if } \|\mathbf{x} - \mathbf{x}_0\|_A \leq r \\ 0 & \text{otherwise,} \end{cases} \tag{20}$$

where A is the distance metric obtained by letting $\mathbf{x}_0$ be the focal sample and r is the soft threshold of the locality function, which is defined by the distance between the focal sample and the test sample, and illustrated in Figure 1 for LDDM, where K is number of neighbors to be considered in the vicinity. The norm of the locality function is defined as

$$\|D(\mathbf{x}_0, A)\| = \int D(\mathbf{x}, \mathbf{x}_0, A) dF(\mathbf{z}). \tag{21}$$

Based on the definition of the locality function, samples and labels can be represented by a new distribution $F(\mathbf{z}, A)$ corresponding to local distance metric A. The distribution is defined as

$$\int_A dF(\mathbf{z}, A) = \int_A \frac{D(\mathbf{x}, \mathbf{x}_0, A)}{\|D(\mathbf{x}_0, A)\|} dF(\mathbf{z}). \tag{22}$$

The local distance metric-based unnormalized local risk function is defined as:

$$\mathcal{R}(\alpha, A, \mathbf{x}_0) = \int Q(\mathbf{z}, \alpha) D(\mathbf{x}, \mathbf{x}_0, A) dF(\mathbf{z}), \tag{23}$$

and the local empirical risk function is based on the summation over all focal samples, which is defined as:

$$\mathcal{R}_{emp}(\alpha, A, \mathbf{x}_0) = \frac{1}{n}\sum_{i=1}^{n} Q(\mathbf{z}_i, \alpha) D(\mathbf{x}_i, \mathbf{x}_0, A). \tag{24}$$

Next, we give the bound on the convergence rate of a local classifier, risk bound of one local classifier and risk bound of the ensemble of a set of local classifiers.

A. Convergence Rate of Local Classifier

In this paper, we define the concept of local domain-based VC-dimension, which is a VC-dimension of a set of functions under a local vicinity. Convergence rate bound of the global risk function only depends on the number of training samples and the VC-dimension that measures the complexity and the expressive power of the set of loss functions $\{Q(\mathbf{z}, \alpha), \alpha \in \Lambda\}$.

In the existing distance metric learning methods, all the VC-dimension and loss functions are under the same distance metric. Thus these distance metric methods obey the bound in the following theorem [14].

Theorem 5.1: Let $\{Q(\mathbf{z}, \alpha), \alpha \in \Lambda\}$ be a set of nonnegative real functions with VC-dimension h. Then the following bound holds

$$P\left\{\sup_{\alpha\in\Lambda} \frac{R(\alpha) - R_{emp}(\alpha)}{\sqrt{\int Q^2(\mathbf{z}, \alpha) dF(\mathbf{z})}} > \epsilon a(\epsilon)\right\} \tag{25}$$
$$< 12\left(\frac{2ne}{h}\right)^h \exp\left\{-\frac{\epsilon^2 n}{4}\right\},$$

where

$$a(\epsilon) = \sqrt{1 - \frac{1}{2}\ln\epsilon}.$$

Theorem 5.1 shows the bound for the test error $\mathcal{R}_{emp}(\alpha)$. The left part is a probability corresponding to the difference between training error $\mathcal{R}(\alpha)$ and test error $\mathcal{R}_{emp}(\alpha)$. The probability approaches 0 when the test error and the training error have an acceptable difference This probability has been proved to be converged to 0 when there are enough training samples [14]. For our local discriminative distance metrics algorithm, the loss functions are different according to the focal samples since they obey their own local distance metrics obtained from the focal samples. To obtain the convergence rate of a local classifier, we assume that the loss function with the local distance metric satisfy the following mild condition:

$$\sup_{\alpha, A} \frac{\sqrt{\int Q^2(\mathbf{z}, \alpha) dF(\mathbf{z}, A)}}{\int Q(\mathbf{z}, \alpha) dF(\mathbf{z}, A)} < \tau. \tag{26}$$

This means that the probability that $\sup_\alpha Q(\mathbf{z}, \alpha)$ exceeds some value will decrease quickly with the value increasing. Value τ determines how fast it decreases. We can get the following theorem for convergence rate of local risk function which is bounded in the term of local domain-based VC-dimension h^*.

Theorem 5.2: Let the vicinity of x_0 be under the local distance metric A and the set of loss functions $\{Q(\mathbf{z}, \alpha) D(\mathbf{x}, \mathbf{x}_0, A), \alpha \in \Lambda\}$ have the local domain based VC-dimension h^*. Then the following bound holds:

$$P\left\{\sup_{\alpha\in\Lambda} \frac{\mathcal{R}(\alpha, A, \mathbf{x}_0) - \mathcal{R}_{emp}(\alpha, A, \mathbf{x}_0)}{\mathcal{R}(\alpha, A, \mathbf{x}_0)} > \frac{\tau\epsilon a(\epsilon)}{\sqrt{\|D(\mathbf{x}_0, A)\|}}\right\} \tag{27}$$
$$< 12\left(\frac{2Ke}{h^*}\right)^{h^*} \exp\left\{-\frac{\epsilon^2 K}{4}\right\}$$

where

$$a(\epsilon) = \sqrt{1 - \frac{1}{2}\ln\epsilon}.$$

Proof is shown in section Appendix A.

Theorem 5.1 gives the convergence rate for the approaches based on a single distance metric in Equation 2. Theorem 5.2 gives the convergence rate using the local domain based VC-dimension for a single distance metric $\mathbf{A}_i$. In the following theorem, we show the risk bound of a local classifier according to Theorem 5.2.

B. Bound of Local Classifiers

For local classifiers learned on local distance metrics according to Equation 16, we have the following theorem.

Theorem 5.3: Let the distance metric of the vicinity of $\mathbf{x}_0$ be A. The set of loss functions $\{Q(\mathbf{z},\alpha)D(\mathbf{x},\mathbf{x}_0,A),\alpha\in\Lambda\}$ have the local domain-based VC-dimension h^*. The following inequality holds for all $\alpha\in\Lambda$ with probability $1-\eta$:

$$R(\alpha,A,\mathbf{x}_0)\le\frac{1}{\|D(\mathbf{x}_0,A)\|}\cdot \qquad (28)$$
$$\left[\mathcal{R}_{emp}(\alpha,A,\mathbf{x}_0)+\nu\left(1+\sqrt{1+\frac{4}{\nu}\mathcal{R}_{emp}(\alpha,A,\mathbf{x}_0)}\right)\right]$$

where

$$\nu=2\frac{(h^*)\left\{\ln[2K/(h^*)]+1\right\}-\ln\frac{\eta}{24}}{K}$$

Proof is shown in Appendix B.

C. Bound of Classifiers Ensemble

We now further explain the generalization bound of the classifier ensemble method discussed in Section 4. Since every training sample will be treated as a focal sample in turn, n samples drawn from the unknown distribution $F(\mathbf{x},\mathbf{y})$ can generate n local distance metrics. For each unknown test sample $\mathbf{x}$, the base classifier $f_i(\mathbf{x},A_i)\in\mathcal{H}$ can be obtained by Equation (16), where A_i is the local distance metric learned by focal sample $(\mathbf{x}_i,y_i)$, which embodies local discriminative information and the size of $\mathcal{H}$ is n. According to the alignment procedure in Equation (17), we define the final classifier after ensemble as

$$g(\mathbf{x})=\mathrm{sign}(\sum_{i=1}^{n}f_i(\mathbf{x},A_i)). \qquad (29)$$

In Equation (29), $g(\cdot)$ gives a wrong prediction on the sample $(\mathbf{x},\mathbf{y})$ only if $\mathbf{y}g(\mathbf{x})\le 0$. $f_i(\mathbf{x},A_i)$ is the i^{th} element of $f(x,A)$. The margin function is given by $\mathbf{y}g(\mathbf{x})$. Equation (29) is fundamentally a majority vote on all base classifiers. [12] has shown a bound which applies to all majority-vote classifiers. Inspired by this, we show the following theorem which states that the generalization error of the ensembled classifier can be bounded in terms of the number of training samples with the margin below a threshold θ and in the capacity of base classifier space $\mathcal{H}$.

Theorem 5.4: Let S be a set of n samples independently drawn from the distribution $F(\mathbf{x},\mathbf{y})$ over $X\times\{-1,+1\}$. Assume that the base-classifier space $\mathcal{H}$ is finite, and let $\sigma>0$. Then with probability at least $1-\sigma$ over the random choice of the training set S, every weighted average function $g(\cdot)$ satisfies the following bound for all $\theta>0$:

$$P_F(\mathbf{y}g(\mathbf{x})\le 0)\le P_S(\mathbf{y}g(\mathbf{x})\le\theta) \qquad (30)$$
$$+\,O\left(\frac{1}{\sqrt{n}}\left(\frac{\log n\log|\mathcal{H}|}{\theta^2}+\log(1/\sigma)\right)^{1/2}\right).$$

For detailed proof please refer to Theorem 1 in [12].

VI. Experiments

A. Crater Detection

Crater detection from panchromatic images faces unique challenges when compared to traditional object detection tasks. Craters are numerous, have a large range of sizes and textures, and continuously merge into image backgrounds. There are various reasons for the formation of the Mars impact craters. Therefore, crater images with different ages, shapes and textures yield a multi-modal distribution dataset. In Fig. 2, crater images may have shadows in different areas and rims are not always clear; non-crater images are different to crater images in their own ways.

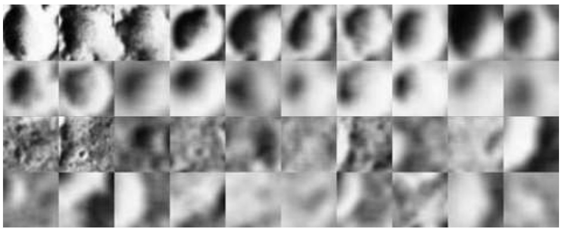

Fig. 2. Sample crater images. The top two rows correspond to real crater images. The bottom two rows correspond to non-crater images.

There are extensive studies showing that local distance metric approaches, which consider the neighborhood information, perform better than global approaches in the multi-model distribution case [9]. Therefore, we apply our local distance metric approach described in Equation (13) to tackle this problem.

A scale, location, and rotation invariant feature, Biologically Inspired Haar Feature [10], is extracted to represent the crater images. The left panel of Fig. 3 show the 2-dimensional PCA space of the original crater images, which roughly demonstrates the distribution and the complexity of the dataset. The right panel of Fig. 3 shows the proposed 2-dimensional discriminative distance metric space under the new feature representation, from which we can tell even though the original data may have a complicated distribution, under a good feature representation and discriminative method, craters and non-craters are well separated. Therefore a single distance metric is good enough for this dataset. A simplified version of LDDM method, denoted as Discriminative Locality Alignment (DLA) [10], [20], has been used in this experiment . We have 2,085 true craters and 2,085 non-craters. Under a 10-fold cross validation, we can achieve 0.93 F1 score, which is higher than the state-of-the-art method, transfer learning based boosting [1], with 0.90 F1 score, shown in Fig. 4.

B. Crime Prediction

Police agencies have been collecting an increasing amount of information to better understand patterns in criminal activity. Recently there is a new trend in using the data collected to predict where and when crime will occur. Crime prediction is greatly beneficial because if it is done accurately, the police administrator would be able to allocate resources to the geographic areas most at risk for criminal activity and ultimately make communities safer. In this paper, we discuss a new four-order tensor representation for crime data. The tensor encodes

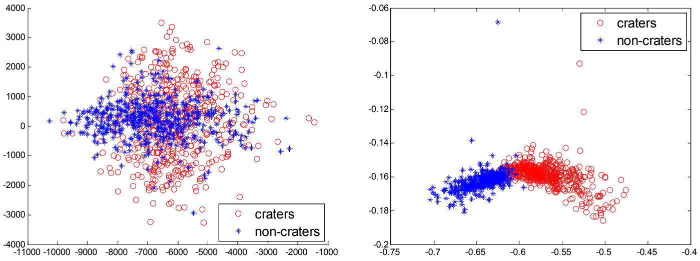

Fig. 3. Impact dataset visualization under 2-dimensional PCA space (left) and 2-dimensional discriminative distance metric space (right).

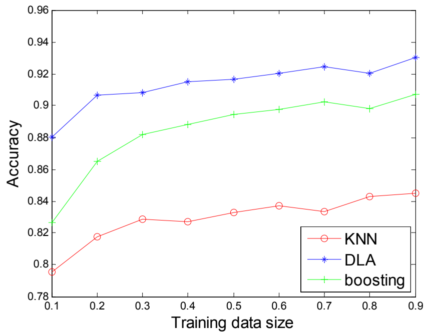

Fig. 4. The comparison of DLA and KNN, and boosting algorithm on crater detection.

the longitude, latitude, time, and other relevant incidents. Using the tensor data structure, we propose the Empirical Discriminative Tensor Analysis (EDTA) algorithm to obtain sufficient discriminative information while minimizing empirical risk simultaneously, which is a natural extension of the distance measurement in vector space. Detailed implementation can be found in [8].

We aim to predict residential burglaries. Each residential burglary is encoded by spatial and temporal information, including longitude, latitude, and time. We also collect other relevant geo-coded events selected by domain scientists that are believed to be associated with criminal activity, including construction permits, foreclosure, mayor hotline inputs, motor vehicle larceny, social events, and offender data. Crime data is rasterized into small grid cells because it is infeasible to make precise longitude and latitude coordinate predication. Fig. 5 further explains the internal structure of the third-order tensor using residential burglary as an example. The number of residential burglaries for a specific grid cell is the summation of all the crimes happened inside this grid cell. We aggregate the data by month and perform monthly prediction because daily crime data is too few and cannot provide sufficient features from the crime data collected in this northeastern city. Therefore, the ultimate objective of this crime forecasting task is to predict whether a grid cell will have high residential burglaries for a given month. Experimental results, comparing to other tensor based approaches, on predicting residential burglary prediction in 2007 using historical data are shown in Fig. 6.

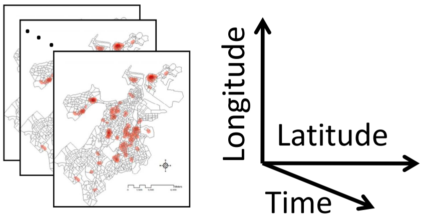

Fig. 5. The residential burglary third-order tensor example. Each map refers to a residential burglary map in different time. The combination of these maps by time forms a three order tensor.

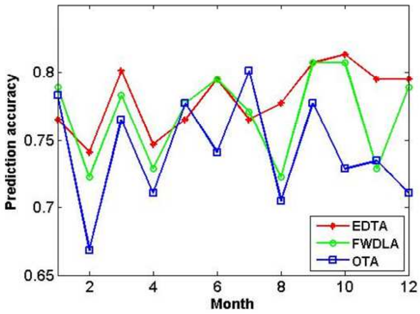

Fig. 6. Methods comparison on residential burglary prediction for 12 months in 2007.

C. Accelerometor data

The negative health consequences and rampant growth of childhood obesity causes it to be a major public health concern. Exercise is known to best combat childhood obesity. One popular method to measure the amount of exercise done by children is to estimate energy expenditure from attached accelerometers as they perform various activities. However, it remains an open research problem to accurately translate accelerometer output to estimates of energy expenditure due to the complexity of motion sensor data. Recent studies have shown that classifying physical activities can significantly improve estimation accuracy of energy expenditure. Totally there are 19 activities in 5 categories: Sedentary, Chores, Locomotion, Interactive Video Games, Exercise and Sports with 5487 samples. 5 percentile features and 9 autocorrelations features are used in this experiment. The high number of classes greatly raise the optimization constraints as shown in Equation 3. Therefore, LDDM is most appropriate for this complex dataset. The results can be found in Table 1 and the experimental result is reported in Leave One Person Out (LOPO).

VII. Conclusion and Future Work

In this paper, we have studied the properties of the local/global distance metric and single/multiple distance metric. We further present the single local discriminative distance metric and its multiple distance metrics form LDDM. We theoretically prove the convergence rate bound and the risk

	SVM	**NaiveBayes**	**NeuralNetwork**	**kNN**	**LDDM**
LOPO	74.02	61.14	78.59	74.93	**80.21**

TABLE I. Accuracy (%) of several classifiers in predicting activity category labels.

bound for local classifiers by introducing a new concept of local domain based VC-dimension. We also prove the risk bound of final classifiers ensemble. For different types of real world applications, we select different distance metric approaches. In the future, we plan to explore more properties for multiple distance metrics learning and extend it to more real world applications.

References

[1] W. Ding, T. F. Stepinski, Y. Mu, L. P. C. Bandeira, R. Vilalta, Y. Wu, Z. Lu, T. Cao, and X. Wu. Subkilometer crater discovery with boosting and transfer learning. *ACM TIST*, 2(4):39, 2011.

[2] C. Domeniconi. Adaptive nearest neighbor classification using support vector machines. *Proc. NIPS*, 2002.

[3] C. Domeniconi and D. Gunopulos. Locally adaptive metric nearest-neighbor classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1281–1285, Sept. 2002.

[4] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 2 edition, 1990.

[5] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighbourhood components analysis. *Proc. NIPS*, 2005.

[6] T. Hastie and R. Tibshirani. Discriminant adaptive nearest neighbor classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(6):607–616, June 1996.

[7] X. He and P. Niyogi. Locality preserving projections. In *In Advances in Neural Information Processing Systems 16*. MIT Press, 2003.

[8] Y. Mu, W. Ding, M. Morabito, and D. Tao. Empirical discriminative tensor analysis for crime forecasting. *Knowledge Science, Engineering and Management*, pages 293–304, 2011.

[9] Y. Mu, W. Ding, and D. Tao. Local discriminative distance metrics ensemble learning. *Pattern Recognition*, 46(8):2337–2349, 2013.

[10] Y. Mu, W. Ding, D. Tao, and T. F. Stepinski. Biologically inspired model for crater detection. In *International Joint Conference on Neural Networks*, pages 2487–2494, 2011.

[11] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *SCIENCE*, 290:2323–2326, 2000.

[12] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, Oct. 1998.

[13] M. Sugiyama. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *The Journal of Machine Learning Research*, 8:1027–1061, 2007.

[14] V. Vapnik and L. Bottou. Local algorithms for pattern recognition and dependencies estimation. *Neural Computation*, 5(6):893–909, 1993.

[15] K. Q. Weinberger and L. K. Saul. Distance Metric Learning for Large Margin Nearest Neighbor Classification. *Journal of Machine Learning Research*, 10:207–244, 2009.

[16] S. Xiang, F. Nie, and C. Zhang. Learning a mahalanobis distance metric for data clustering and classification. *Pattern Recogn.*, 41:3600–3612, December 2008.

[17] E. Xing, A. Ng, M. Jordan, and S. Russell. Distance metric learning with application to clustering with side-information. *Proc. NIPS*, 2003.

[18] L. Yang and R. Jin. Distance metric learning: A comprehensive survey. *Michigan State Universiy*, pages 1–51, 2006.

[19] Y. Yu, J. Jiang, and L. Zhang. Distance metric learning by minimal distance maximization. *Pattern Recogn.*, 44:639–649, March 2011.

[20] T. Zhang, D. Tao, X. Li, and J. Yang. Patch Alignment for Dimensionality Reduction. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1299–1313, Sept. 2009.

[21] W. Zhang, X. Xue, Z. Sun, H. Lu, and Y.-F. Guo. Metric learning by discriminant neighborhood embedding. *Pattern Recogn.*, 41:2086–2096, June 2008.

[22] Z. Zhang, J. Kwok, and D. Yeung. Parametric distance metric learning with label information. In *Proc. IJCAI*, 2003.

Appendix

A. Proof of Theorem 5.2

Proof: Theorem 5.1 implies the following inequality:

$$P\left\{\sup_{\alpha\in\Lambda}\frac{\mathcal{R}(\alpha,A,\mathbf{x}_0)-\mathcal{R}_{emp}(\alpha,A,\mathbf{x}_0)}{\sqrt{\int Q^2(\mathbf{z},\alpha)D^2(\mathbf{x},\mathbf{x}_0,A)dF(\mathbf{z})}}>\epsilon a(\epsilon)\right\} \tag{31}$$

$$<12\left(\frac{2Ke}{h^*}\right)^{h^*}\exp\left\{-\frac{\epsilon^2K}{4}\right\},$$

where K is the size of the focal vicinity defined in Equation (16). According to Equation (20) and Equation (22), we have

$$\begin{aligned}&\sqrt{\int Q^2(\mathbf{z},\alpha)D^2(\mathbf{x},\mathbf{x}_0,A)dF(\mathbf{z})}\\ \leq&\sqrt{\int Q^2(\mathbf{z},\alpha)\left\|D(\mathbf{x}_0,A)\right\|dF(\mathbf{z},A)}.\end{aligned} \tag{32}$$

According to Equation (22), (23) and (26), we have

$$\sqrt{\int Q^2(\mathbf{z},\alpha)dF(\mathbf{z},A)}<\tau\int Q(\mathbf{z},\alpha)dF(\mathbf{z},A)=\tau\frac{\mathcal{R}(\alpha,A,\mathbf{x}_0)}{\|D(\mathbf{x}_0,A)\|}. \tag{33}$$

According to Equation (32) and (33), we have

$$\sqrt{\int Q^2(\mathbf{z},\alpha)D^2(\mathbf{z},\mathbf{x}_0,A)dF(\mathbf{z})}<\tau\frac{\mathcal{R}(\alpha,A,\mathbf{x}_0)}{\sqrt{\|D(\mathbf{x}_0,A)\|}}. \tag{34}$$

The inequality Equation (27) can be obtained from Equations (31) and (34) immediately.
This completes the proof. ■

B. Proof of Theorem 5.3

Proof: In Equation (27), let $\eta/2$ denote the right-side. By solving the equation

$$12\left(\frac{2Ke}{h^*}\right)^{h^*}\exp\{-\frac{\epsilon^2K}{4}\}=\eta/2 \tag{35}$$

and replacing the result into Equation (27), we obtain the following inequality with probability $1-\eta/2$.

$$\mathcal{R}(\alpha,A,\mathbf{x}_0)\leq\mathcal{R}_{emp}(\alpha,A,\mathbf{x}_0)+\nu\left(1+\sqrt{1+\frac{4}{\nu}\mathcal{R}_{emp}(\alpha,A,\mathbf{x}_0)}\right) \tag{36}$$

where

$$\nu=2\frac{(h^*)\left\{\ln[2n/(h^*)]+1\right\}-\ln\frac{\eta}{24}}{n}.$$

By defining the normalized empirical risk for the vicinity of $\mathbf{x}_0$

$$R(\alpha,A,\mathbf{x}_0)=\int Q(\mathbf{z},\alpha)\frac{D(\mathbf{x},\mathbf{x}_0,A)}{\|D(\mathbf{x}_0,A)\|}dF(\mathbf{z}),$$

we can get Equation (28) by dividing both sides of inequality Equation (36) by $\|D(\mathbf{x}_0,A)\|$.
This completes the proof. ■