Neurocomputing 156 (2015) 33–41

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Risk-based adaptive metric learning for nearest neighbour classification

Yanan Miao [a,b], Xiaoming Tao [a,b,*], Yipeng Sun [a,b], Yang Li [a,b], Jianhua Lu [a,b]

[a] Department of Electronic Engineering, Tsinghua University, Beijing 10084, PR China
[b] Tsinghua National Laboratory for Information Science and Technology (TNList), Beijing 10084, PR China

ARTICLE INFO

Article history:
Received 23 May 2014
Received in revised form
25 September 2014
Accepted 5 January 2015
Communicated by B. Apolloni
Available online 21 January 2015

Keywords:
Nearest neighbour
Metric learning
LOOCV risk
Non-parametric
Classification

ABSTRACT

The performance of *k*-nearest neighbour classification highly depends on the appropriateness of distance metric designation. Optimal performance can be obtained when the distance metric is matched to the characteristics of data. Existing works on distance-metric learning typically learn a global linear transform from training samples, and the effectiveness is limited to data, which are well-separated by linear decision boundaries. To address this problem, we propose a locally adaptive weighted distance-metric learning method to deal with the non-linearity of the data. The metric are learned based on local leave-one-out cross-validation (LOOCV) risks in each dimension, so that the local variations in feature component discriminability are taken into account. Experiments on both public datasets and hyper-spectral imagery classification demonstrate that the classification accuracy of the proposed method shows about 2–10% improvements over other competitive methods.

1. Introduction

k-Nearest neighbour (NN) classification and its variations are considered as simple and efficient tools in pattern recognition applications [1–6]. The performance of *k*-NN classification depends critically on the choice of distance metric [7,8]. In the absence of prior knowledge, Euclidean distance is most frequently used due to its simplicity. Unfortunately, the effectiveness of it is limited in many practical applications [9–11]. To address the limitation, many distance metric learning algorithms [2,9,12,10,13,5] have been proposed for various pattern recognition tasks. One kind of them can adaptively select the neighborhood shape by learning a local metric [2]. It can perform effectively under non-linear decision boundaries in training samples, yet needs some restrictive assumptions and suffers from high computational complexity. Some algorithms [12,14] learn Mahalanobis metric which can be viewed as a global linear transform of the input space. While it is straightforward to apply existing methods within local regions of the input, the estimation for distance metric can be unreliable because the number of training samples is small. Another kind of methods attempts to develop metric learning algorithm by minimize the empirical risk directly [15,16] under bounded loss function. Yet it relies on strong assumptions on the distribution of the examples.

In this paper, we propose to utilize empirical risk to learn a locally adaptive metric to model the non-linear spatial distribution of data. It is both learned and acting locally without many assumptions. We describe it as *risk-based weighted nearest neighbour* (RBWNN) classification approach. Specifically, we assign a different weight to each dimension to compute the distance. Instead of minimizing the empirical risk directly, each weight is computed based on its local empirical risk in each dimension, as measured by leave-one-out cross-validation (LOOCV) errors in a local search region of the input. Components with larger errors are less informative dimensions, thus can be suppressed while computing distances in our algorithm. Experiments show that our proposed metric is robust for different tasks and can provide flexibility to new kinds of data. We also note that our method needs only 1–3 iterations to achieve convergence without many manually tuned parameters.

The remainder of this paper is organised as follows. Some related works are introduced in Section 2. Section 3 describes the proposed method in detail and makes some discussions about our method. Section 4 shows the experiment results compared to other methods on several datasets. The last section concludes the paper.

2. Related works

As a simple and effective method, NN classification has been widely used and improved in many literatures. Ref. [17] decides the query point's label by considering the distance from it to the median point of a line through two same class samples. A local mean vector based method and its improved version are proposed

* Corresponding author at: Department of Electronic Engineering, Tsinghua University, Beijing 10084, PR China.
E-mail addresses: miaoyn12@mails.tsinghua.edu.cn (Y. Miao), taoxm@mail.tsinghua.edu.cn (X. Tao), sunyp10@mails.tsinghua.edu.cn (Y. Sun), liy-11@mails.tsinghua.edu.cn (Y. Li), lhh-dee@mail.tsinghua.edu.cn (J. Lu).

http://dx.doi.org/10.1016/j.neucom.2015.01.009
0925-2312/

by Mitani et al. [18] and Zeng et al. [19] respectively, to perform k-NN classification more efficiently. However, these methods have not fully utilized the local statistics by only considering the mean or the second-order moment. In our method, the local spatial distribution of data will be considered by using the empirical risk in a local search region of the input.

There are also kinds of algorithms pursuiting to learning a metric to compute the distances for improving NN classification. Hastie et al. [2] discovered the relation between discriminant analysis and the metric. Their algorithm acquired adaptive metric by computing intra/inter-class matrix under Gaussian data assumption. Another local metric [9] is developed through considering the most relevant features to adjust the weights imposed on each dimension using a Chi-square distance formulation. In [13], the distances between samples are simply scaled by defined costs in a local region so that the metric is locally adaptive. Although these methods improve the original k-NN rule, they cannot capture enough local statistics in data, and the computational complexity of such improvements is high. In our methods, we utilize the spatial distribution of data by associating empirical risks with the metric weights. More recently, a series of methods have been proposed to learn Mahalanobis distance metric directly from different perspectives. Neighbour component analysis (NCA) [12] optimizes the expected leave-one-out error, using a stochastic neighbour selection rule. Weinberger et al. [20,7] formulated the metric learning process a semi-definite programming with large margin constrains, and solved it using a combination of sub-gradient descent and alternating projections. Information-theoretic metric learning (ITML) [14,5] learns the metric by minimizing relative entropy (Kullback–Leibler divergence) between the under-estimated matrix and a prior one under Gaussian distribution assumption. We note that these metric learning algorithms are normally task-driven and have been applicated to object tracking [21–23], face recognition [24] and other applications [25]. All these algorithms are parametric methods and need pre-training under labelled samples to learn a global metric. Our method instead is non-parametric and learns in a local search region so that the metric is locally adaptive to the spatial distribution of data.

3. Risk-based adaptive metric

To find meaningful 'nearest' neighbours, the metric should be adaptive to the local boundary of data [2,26]. We propose to learn a metric from the risk induced by different components of input. The risk is used to update the weight with an exponential kernel weighting scheme. In this section, the motivation and the basic principle of the proposed method will be described in detail.

3.1. Problem description

The main goal of supervised classification is to train a classifier identifying a new coming query sample $\mathbf{x}_0 \in \mathbb{R}^p$ to one of the predefined classes $y_0 = 1, 2, \ldots, J, J \geq 2$, by means of the labelled training samples $\mathcal{D} = \{(y_i, \mathbf{x}_i)\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^p$, and the tth component of $\mathbf{x}_i$ is $x_i(t)$. In RBWNN, the labels of the under-classified inputs are assigned by which the most frequently presented samples belonging to. At first, we revisit the final decision step in the NN-based method from a probability perspective. The predicted label is determined by considering the posterior probability $p(j|\mathbf{x}_0)$ where $j = 1, 2, \ldots, J$. Then according to *maximum a posterior* (MAP) estimation principle, the query's class y_0 is

$$y_0 = \arg\max_j p(j|\mathbf{x}_0), \tag{1}$$

where the posterior probability can be approximated by

$$p(j|\mathbf{x}_0) = \frac{1}{k}\sum_i \mathbf{1}\{y_i = = j\} \wedge \mathbf{x}_i \in N_k(\mathbf{x}_0). \tag{2}$$

Note that function $\mathbf{1}\{\cdot\}$ equals to 1 when the event in $\{\}$ is true and 0 otherwise. $N_k(\mathbf{x}_0)$ denotes the set of k nearest neighbours of $\mathbf{x}_0$. The discrete conditional probability is

$$p(j|\mathbf{x}_0) \sim \left[\frac{1}{k}, \frac{2}{k}, \ldots, \frac{k}{k}\right]. \tag{3}$$

Extending the neighbour number from one to k, the error upper bound can converge to the Bayes error rate when k goes to infinity with $k/n \to 0$ provided sufficient training samples [27]. When k is finite, the distribution in Eq. (3) is just an approximation to the theoretical continuous distribution. Therefore, from the perspective of sampling, the chosen nearest neighbours play an important role in local density estimation. If there are more relevant samples in the local search region of the query among the k nearest neighbours, the estimated bias can be cut down [1]. In the literature, neighbour samples are obtained in standard k-NN (the standard k-NN is summarized in Algorithm 1) by computing the Euclidean norm (ℓ^2-norm) as follows:

$$\mathbf{D}(\mathbf{x}_i, \mathbf{x}_0) = \| \mathbf{x}_i - \mathbf{x}_0 \|_2, \tag{4}$$

In practice, Euclidean measurement usually cannot match well with the observation prior. This distance metric measures dissimilarities in each input dimension equally, in spite of their unequal effects on computing distances. Therefore, Eq. (4) cannot find the most informative dimension to search the neighbours. As the posterior probability $p(j|\mathbf{x}_0)$ is estimated from the class distribution of neighbour samples, it is desired if the distances are small from the query $\mathbf{x}_0$ to the neighbours with the same label. Therefore, we seek adaptive weighted metric to fit the nonlinear boundary of data.

Algorithm 1. Standard k-NN.

Input: Training set $\mathcal{D} = \{(y_i, \mathbf{x}_i)\}_{i=1}^N$, the query sample $\mathbf{x}_0$, and the number of neighbour k
Output: y_0: the label of $\mathbf{x}_0$
1: **for all** $\mathbf{x}_i$ **do**
2: Get the distance to $\mathbf{x}_0$ from Eq. (4)
3: **end for**
4: Find the k nearest neighbour $N_k(\mathbf{x}_0)$
5: Get posterior probability $p(j|\mathbf{x}_0)$ according to Eq. (2)
6: $y_0 = \arg\ \max_j p(j|\mathbf{x}_0)$

3.2. Adaptive metric learning

In order to use more informative components of the input to find the neighbours, we propose to learn an adaptive metric which can impose unequal emphasis on each dimension. The metric can change adaptively as the location of the query varies among the samples. The distance metric is defined as follows:

$$\mathbf{D}(\mathbf{x}_i, \mathbf{x}_0; w_t) = \left(\sum_{t=1}^{p} w_t (x_i(t) - x_0(t))^2\right)^{1/2}, \tag{5}$$

$\mathbf{x}_0$ is a p-dimensional query sample. w_t can be used to weight the relevancy of dimension t with regard to the neighbourhood relationship of the input space. We define w_t according to a kernel weighting scheme

$$w_t = \frac{\phi_t}{\sum_{i=1}^{p} \phi_i}, \tag{6}$$