Mach Learn (2016) 102:57–83
DOI 10.1007/s10994-015-5495-y

On some transformations of high dimension, low sample size data for nearest neighbor classification

Subhajit Dutta[1] · **Anil K. Ghosh**[2]

Received: 25 August 2013 / Accepted: 18 March 2015 / Published online: 3 April 2015

Abstract For data with more variables than the sample size, phenomena like concentration of pairwise distances, violation of cluster assumptions and presence of hubness often have adverse effects on the performance of the classic nearest neighbor classifier. To cope with such problems, some dimension reduction techniques like those based on random linear projections and principal component directions have been proposed in the literature. In this article, we construct nonlinear transformations of the data based on inter-point distances, which also lead to reduction in data dimension. More importantly, for such high dimension low sample size data, they enhance separability among the competing classes in the transformed space. When the classic nearest neighbor classifier is used on the transformed data, it usually yields lower misclassification rates. Under appropriate regularity conditions, we derive asymptotic results on misclassification probabilities of nearest neighbor classifiers based on the l_2 norm and the l_p norms (with $p \in (0, 1]$) in the transformed space, when the training sample size remains fixed and the dimension of the data grows to infinity. Strength of the proposed transformations in the classification context is demonstrated by analyzing several simulated and benchmark data sets.

Keywords Bayes risk · HDLSS data · High-dimensional geometry · Inter-point distances · Law of large numbers · Misclassification probability · Reproducing kernel Hilbert space · ρ-mixing sequences

Editor: Xiaoli Z. Fern.

✉ Subhajit Dutta
tijahbus@gmail.com

Anil K. Ghosh
akghosh@isical.ac.in

[1] Department of Mathematics and Statistics, Indian Institute of Technology, Kanpur 208016, U.P., India

[2] Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, 203, B. T. Road, Kolkata 700108, India

1 Introduction

Nearest neighbor classification (see, e.g., Cover and Hart 1967; Fix and Hodges 1989) is a simple and popular nonparametric method in the machine learning literature. For a fixed value of k, the k nearest neighbor classifier assigns an unlabeled observation $\mathbf{z}$ to the class having the maximum number of representatives in the set of k labeled observations closest to $\mathbf{z}$. When the training sample size n (i.e., the number of labeled observations) is large compared to the data dimension d, the k nearest neighbor classifier usually performs well. For an appropriate choice of k (which increases with n), its misclassification probability converges to the Bayes risk as n grows to infinity (see, e.g., Hall et al. 2008). However, like other nonparametric methods, the classic nearest neighbor (NN) classifier suffers from curse of dimensionality. So, it may yield poor performance for data in high dimensions. Radovanovic et al. (2010) discussed the presence of hubs and the violation of the cluster assumptions for such high-dimensional data. They further studied adverse effects of hubness on supervised, semi-supervised and unsupervised learning based on nearest neighbors.

For high-dimensional data, Francois et al. (2007) showed that pairwise distances between all observations in a class (after appropriate scaling) concentrate around a single value. Hall et al. (2005) also studied the geometry of a data cloud in high dimension, low sample size (HDLSS) situations and proved results related to the concentration of pairwise distances under appropriate conditions. Like Francois et al. (2007), their analysis showed that the observations in each class have a tendency to lie deterministically at the vertices of a regular simplex, and the randomness in the data appears only as a random rotation of that simplex. Further, Hall et al. (2005) used this high-dimensional geometry of the data to analyze the behavior of some popular classifiers including the classic NN classifier.

We now demonstrate adverse effects of concentration of pairwise distances on the performance of the classic NN classifier in high dimensions. Consider a classification problem between two d-dimensional normal distributions with mean vectors $\mathbf{0}_d = (0, \ldots, 0)^T$ and $\boldsymbol{\nu}_d = (\nu, \ldots, \nu)^T$, and dispersion matrices $\sigma_1^2\mathbf{I}_d$ and $\sigma_2^2\mathbf{I}_d$, respectively. Here, $\sigma_1^2 \neq \sigma_2^2$, and $\mathbf{I}_d$ denotes the $d \times d$ identity matrix. Now, if $\mathbf{X} = (X_1, \ldots, X_d)^T$ and $\mathbf{X}' = (X_1', \ldots, X_d')^T$ are two independent observations from class-1, $\|\mathbf{X} - \mathbf{X}'\|^2/2\sigma_1^2 \sim \chi_d^2$, the chi-square distribution with d degrees of freedom (df). Here, $\|\cdot\|$ denotes the usual Euclidean distance. Since $E(\|\mathbf{X} - \mathbf{X}'\|^2/d) = 2\sigma_1^2$ and $Var(\|\mathbf{X} - \mathbf{X}'\|^2/d) = 8\sigma_1^4/d \to 0$ as $d \to \infty$, we have $\|\mathbf{X} - \mathbf{X}'\|^2/d \xrightarrow{P} 2\sigma_1^2$ as $d \to \infty$. Similarly, if $\mathbf{Y}$ and $\mathbf{Y}'$ are two independent observations from class-2, we have $\|\mathbf{Y} - \mathbf{Y}'\|^2/d \xrightarrow{P} 2\sigma_2^2$ as $d \to \infty$. Now, if $\mathbf{X}$ is from class-1 and $\mathbf{Y}$ is from class-2, $\|\mathbf{X} - \mathbf{Y}\|^2/(\sigma_1^2 + \sigma_2^2) \sim \chi_d^2(\delta)$, the non-central chi-square distribution with d df and the non-centrality parameter $\delta = \nu^2/(\sigma_1^2 + \sigma_2^2)$. One can show that $\|\mathbf{X} - \mathbf{Y}\|^2/d \xrightarrow{P} \sigma_1^2 + \sigma_2^2 + \nu^2$ as $d \to \infty$. Suppose that we have two sets of labeled observations $\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}$ and $\mathbf{y}_1, \ldots, \mathbf{y}_{n_2}$ from class-1 and class-2, respectively. Now, for any future observation $\mathbf{z}$ from class-1, $\|\mathbf{z} - \mathbf{x}_i\|/\sqrt{d} \xrightarrow{P} \sigma_1\sqrt{2}$ for all $1 \leq i \leq n_1$, while $\|\mathbf{z} - \mathbf{y}_j\|/\sqrt{d} \xrightarrow{P} \sqrt{\sigma_1^2 + \sigma_2^2 + \nu^2}$ for all $1 \leq j \leq n_2$ as $d \to \infty$. So, $\mathbf{z}$ is correctly classified by the NN classifier if $\nu^2 > \sigma_1^2 - \sigma_2^2$. Similarly, a future observation from class-2 is correctly classified if $\nu^2 > \sigma_2^2 - \sigma_1^2$. Therefore, *the classic NN classifier correctly classifies all unlabeled observations if $\nu^2 > |\sigma_1^2 - \sigma_2^2|$ (also see Hall et al. (2005, p. 436)). Otherwise, irrespective of the choice of k, it classifies all unlabeled observations to a single class.* For instance, if two high-dimensional normal distributions differ only in their scales (i.e., $\nu^2 = 0$

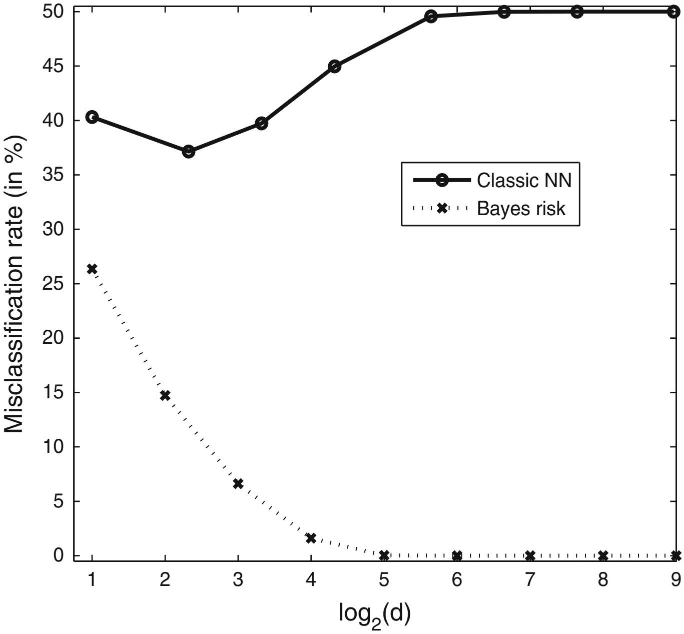


Fig. 1 Misclassification rates for $d = 2, 5, 10, 20, 50, 100, 200$ and 500

and $\sigma_1^2 \neq \sigma_2^2$), the classic NN classifier classifies all observations to the class having smaller spread.

To illustrate this, we considered a classification problem involving $N_d(\mathbf{0}_d, \mathbf{I}_d)$ and $N_d(\mathbf{0}_d, 1/4\mathbf{I}_d)$. We generated 10 observations from each class to form the training sample, and they were used to classify 200 unlabeled observations (100 from each class). This procedure was repeated 250 times to compute the average misclassification rate of the NN classifier. Figure 1 shows these misclassification rates and the corresponding Bayes risks for various choices of d. For large values of d, while the Bayes risk was close to zero, the classic NN classifier failed to discriminate between the two classes and classified all unlabeled observations to class-2.

Several attempts have been made in the literature to reduce dimension of the data, and use the NN classifier on the reduced subspace. The simplest method of dimension reduction is by projecting the data along some random directions (see, e.g., Fern and Brodley 2003). Another popular approach is to use projections based on principal component analysis (see, e.g., Deegalla and Bostrom 2006). Other approaches to NN classification for high-dimensional data include Goldberger et al. (2005), Weinberger et al. (2006), Tomasev et al. (2011). Chen and Hall (2009) proposed a robust version of the NN classifier for high-dimensional data, but it is applicable to a specific type of two-class location problem.

In this article, we propose some nonlinear transformations of the data that lead to substantial reduction in dimension for HDLSS data. These transformations are motivated from theoretical results on the high-dimensional geometry of a data cloud. They are based on interpoint distances, and enhance separability among the competing classes in the transformed space. As a result, when the NN classifier is used on the transformed data, it usually yields improved performance. We carry out a theoretical investigation on the misclassification probabilities of these classifiers, and show that concentration of pairwise distances can be used to develop a ‘perfect learning machine’ for HDLSS data.

2 Transformation based on averages of pairwise distances

Consider a two class problem with labeled observations $\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}$ from class-1 and $\mathbf{y}_1, \ldots, \mathbf{y}_{n_2}$ from class-2. If the component variables in each class are independent and identically distributed (i.i.d.) Gaussian random variables, for any of these $n = n_1 + n_2$ labeled observations, its distances from the observations in each class (after dividing by $\sqrt{d}$) converge to a constant. These two constants (one for each class) depend on the class label of the observation. So, if we transform all labeled observations based on average distances, we expect to have two distinct clusters in the transformed two-dimensional space, one for each class. For $n_1, n_2 \geq 2$, these transformed data points are given as follows

$$\mathbf{x}_i^* = \left(\frac{1}{n_1 - 1} \sum_{j=1, j \neq i}^{n_1} \frac{\|\mathbf{x}_i - \mathbf{x}_j\|}{\sqrt{d}}, \frac{1}{n_2} \sum_{j=1}^{n_2} \frac{\|\mathbf{x}_i - \mathbf{y}_j\|}{\sqrt{d}} \right)^T \quad \text{for} \quad 1 \leq i \leq n_1 \text{ and}$$

$$\mathbf{y}_j^* = \left(\frac{1}{n_1} \sum_{i=1}^{n_1} \frac{\|\mathbf{y}_j - \mathbf{x}_i\|}{\sqrt{d}}, \frac{1}{n_2 - 1} \sum_{i=1, i \neq j}^{n_2} \frac{\|\mathbf{y}_j - \mathbf{y}_i\|}{\sqrt{d}} \right)^T \quad \text{for} \quad 1 \leq j \leq n_2. \qquad (1)$$

Recall the two class classification problem involving the distributions $N_d(\mathbf{0}_d, \mathbf{I}_d)$ and $N_d(\mathbf{0}_d, 1/4\mathbf{I}_d)$ discussed in Sect. 1. In this example, for higher values of d, the classic NN classifier could not discriminate between the two classes and led to an average misclassification rate of almost 50 % (see Fig. 1). Figure 2 shows the scatter plots of transformed training sample observations for $d = 50$ and 500, where the black dots and the gray dots represent observations from class-1 and class-2, respectively. Clearly, the transformation based on average distances not only reduces the data dimension, but enhances separability between the two classes. This separability becomes more prominent as the dimension increases.

In a J class ($J > 2$) problem, we project an observation to a J-dimensional space, where the i-th co-ordinate denotes its average distance from the observations in the i-th class ($1 \leq i \leq J$). Here, we expect to have J distinct clusters, one for each class (assume that there

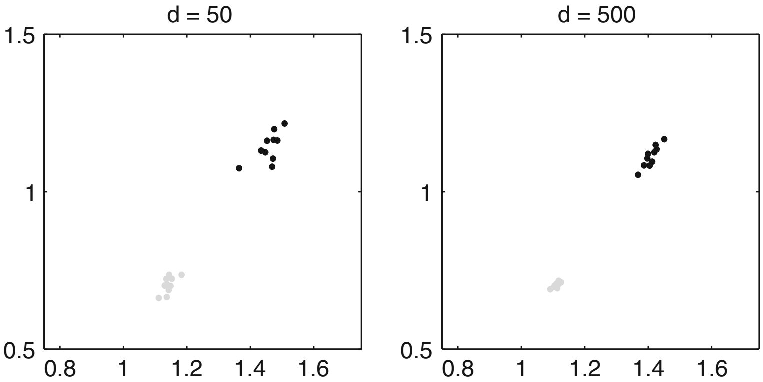


Fig. 2 Scatter plots of training data points after transformation based on average distances

are at least two labeled observations from each class). So, it is more meaningful to use the NN classifier on the transformed data. The NN classifier, which is used after this **TR**ansformation based on **A**verage **D**istances (henceforth referred to as NN-TRAD) possesses nice theoretical properties, and we state them in the following sub-sections.

2.1 Distributions with independent component variables

We have observed that if the components of $\mathbf{X}$ and $\mathbf{Y}$ are i.i.d. Gaussian variables, then the transformed labeled observations from the two classes converge to two points in $\mathbb{R}^2$ as the dimension increases. While the $\mathbf{x}_i^*$'s converge in probability to $\mathbf{a}_1 = (\sigma_1\sqrt{2}, \sqrt{\sigma_1^2+\sigma_2^2+\nu^2})^T$, the $\mathbf{y}_j^*$'s converge to $\mathbf{a}_2 = (\sqrt{\sigma_1^2+\sigma_2^2+\nu^2}, \sigma_2\sqrt{2})^T$ in probability. Here $\sigma_1^2 = Var(X_1), \sigma_2^2 = Var(Y_1)$ and $\nu^2 = E(X_1 - Y_1)^2$. This result continues to hold if the the components of $\mathbf{X}$ and $\mathbf{Y}$ are not necessarily Gaussian, but they are i.i.d. with finite second moments. In such cases, the distance convergence results follow from the weak law of large numbers [WLLN] (see, e.g., Feller 1968). For instance, $\|\mathbf{X}-\mathbf{Y}\|^2/d = \sum_{q=1}^d (X_q - Y_q)^2/d \xrightarrow{P} E(X_1 - Y_1)^2 = (\sigma_1^2+\sigma_2^2+\nu^2)$ as $d \to \infty$. Now, $\mathbf{a}_1$ and $\mathbf{a}_2$ are indistinguishable if and only if $\sigma_1^2 = \sigma_2^2$ and $\nu^2 = 0$. So, for high-dimensional data, unless we have $\nu^2 = 0$ and $\sigma_1^2 = \sigma_2^2$, the transformed observations $\mathbf{x}_1^*, \ldots, \mathbf{x}_{n_1}^*$ and $\mathbf{y}_1^*, \ldots, \mathbf{y}_{n_2}^*$ form two distinct clusters.

Using this transformation on an unlabeled observation $\mathbf{z}$, we get

$$\mathbf{z}^* = \left(\frac{1}{n_1}\sum_{i=1}^{n_1} \frac{\|\mathbf{x}_i - \mathbf{z}\|}{\sqrt{d}}, \frac{1}{n_2}\sum_{j=1}^{n_2} \frac{\|\mathbf{y}_j - \mathbf{z}\|}{\sqrt{d}} \right)^T. \tag{2}$$

For any $\mathbf{z}$ from class-1 (respectively, class-2), $\|\mathbf{z}-\mathbf{x}_i\|/\sqrt{d}$ converges to $\sigma_1\sqrt{2}$ (respectively, $\sqrt{\sigma_1^2+\sigma_2^2+\nu^2}$) for $1 \le i \le n_1$, and $\|\mathbf{z}-\mathbf{y}_j\|/\sqrt{d}$ converges to $\sqrt{\sigma_1^2+\sigma_2^2+\nu^2}$ (respectively, $\sigma_2\sqrt{2}$) for $1 \le j \le n_2$. So, $\mathbf{z}^*$ converges to $\mathbf{a}_1$ (respectively, $\mathbf{a}_2$) if $\mathbf{z}$ comes from class-1 (respectively, class-2). Therefore, $\mathbf{z}$ is correctly classified by the NN-TRAD classifier with probability tending to one as d grows to infinity.

We can observe this concentration of pairwise distances and hence optimality of the misclassification rate of the NN-TRAD classifier even when the components of $\mathbf{X}$ and $\mathbf{Y}$ are independent, but not identically distributed. In such cases, one needs stronger assumptions. For instance, we have the convergence of the pairwise distances if the fourth moments of the component variables are uniformly bounded (see $(A1)$ in Sect. 2.2). In this case, $\|\mathbf{X}-\mathbf{X}'\|^2/d$, $\|\mathbf{Y}-\mathbf{Y}'\|^2/d$ and $\|\mathbf{X}-\mathbf{Y}\|^2/d$ converge in probability to $2\sigma_1^2$, $2\sigma_2^2$ and $(\sigma_1^2+\sigma_2^2+\nu^2)$, respectively, where σ_1^2, σ_2^2 and ν^2 are defined as the limiting values (as $d \to \infty$) of $\sigma_{1,d}^2 = \sum_{q=1}^d Var(X_q)/d$, $\sigma_{2,d}^2 = \sum_{q=1}^d Var(Y_q)/d$ and $\nu_{12,d}^2 = \sum_{q=1}^d \{E(X_q) - E(Y_q)\}^2/d$ (also see $(A2)$ in Sect. 2.2), respectively.

2.2 Distributions with dependent component variables

Under appropriate conditions (see $(A1)$ and $(A2)$ stated below), the distance convergence holds for *uncorrelated* measurement variables as well (follows from Lemma 1 in the “Appendix”). However, Francois et al. (2007) observed that for high-dimensional data with highly correlated or dependent measurement variables, pairwise distances are less concentrated than if all variables are independent. They claimed that the concentration phenomenon depends on the intrinsic dimension of the data, instead of the dimension of the embedding space.

So, in order to have distance concentration in high dimensions, one needs high intrinsic dimensionality of the data or weak dependence among the measurement variables. Hall et al. (2005) assume such weak dependence (see $(A3)$ stated below) for investigating the distance concentration property in high dimensions. Motivated from Hall et al. (2005), we consider the following assumptions:

$(A1)$ *In each of the J competing classes, fourth moments of the component variables are uniformly bounded.*

$(A2)$ *Let $\boldsymbol{\mu}_{i,d}$ and $\boldsymbol{\Sigma}_{i,d}$ be the d-dimensional mean vector and the $d \times d$ dispersion matrix for the i-th class $(1 \leq i \leq J)$. There exists constants $\sigma_i^2 > 0$ for all $1 \leq i \leq J$ and ν_{ij} for all $i \neq j (1 \leq i, j \leq J)$, such that $(i) d^{-1} trace(\boldsymbol{\Sigma}_{i,d}) \to \sigma_i^2$ and $(ii) d^{-1} \|\boldsymbol{\mu}_{i,d} - \boldsymbol{\mu}_{j,d}\|^2 \to \nu_{ij}^2$, as $d \to \infty$.*

$(A3)$ *Let $\mathbf{U} = (U_1, U_2, \ldots)^T$ and $\mathbf{V} = (V_1, V_2, \ldots)^T$ be two independent observations either from the same class or from two different classes. Under some permutation of the components variables (which is same in all classes), the ρ- mixing property holds for the sequence $\{(U_q - V_q)^2, q \geq 1\}$, i.e.,*

$$sup_{1 \leq q < q' \leq \infty,\ |q - q'| > r} \left| Corr \left\{ (U_q - V_q)^2, (U_{q'} - V_{q'})^2 \right\} \right| \leq \rho(r),$$

where $\rho(r) \to 0$ as $r \to \infty$.

Note that Jung and Marron (2009) assumed almost similar conditions to prove high-dimensional consistency of estimated principal component directions. Biswas and Ghosh (2014) and Biswas et al. (2014) used similar conditions to derive consistency of their two-sample tests for HDLSS data.

Conditions $(A1)$–$(A3)$ are quite general. The first two are moment conditions that ensure some 'regularity' of the random variables. In classification problems, we usually get more information about class separation as the sample size increases. But, in HDLSS set up, we consider the sample size to be fixed, and under $(A2)$, we expect information about class separation to increase as d increases (unless $\sigma_1^2 = \sigma_2^2$ and $\nu^2 = 0$). Assumption $(A3)$ implies a form of weak dependence among the measurement variables so that WLLN holds for the sequence of dependent random variables as well (see Lemma 1 in the "Appendix"). For time series data, this indicates that the lag correlation shrinks to zero as the length of the lag increases. In particular, for data generated from discrete ARMA processes, all these conditions are satisfied. Importantly, stationarity of the time series is not required here. These assumptions also hold for m-dependent processes and Markov processes over finite state spaces. Recall that if the measurement variables are i.i.d., $(A2)$ and $(A3)$ hold automatically, and instead of $(A1)$, we only need existence of second moments for the weak convergence of pairwise distances.

Under $(A1)$ and $(A3)$, $\left|\sum_{q=1}^d (U_q - V_q)^2/d - \sum_{q=1}^d E(U_q - V_q)^2/d\right| \xrightarrow{P} 0$ as $d \to \infty$ (see Lemma 1). Now, depending on the choice of $(\mathbf{U}, \mathbf{V}) = (\mathbf{X}, \mathbf{X}')$, $(\mathbf{Y}, \mathbf{Y}')$ or $(\mathbf{X}, \mathbf{Y})$, under $(A2)$, $\sum_{q=1}^d E(U_q - V_q)^2/d$ converges to $2\sigma_1^2$, $2\sigma_2^2$ or $(\nu^2 + \sigma_1^2 + \sigma_2^2)$, respectively. Hence, we have

$$(i) \|\mathbf{X} - \mathbf{X}'\|/\sqrt{d} \xrightarrow{P} \sigma_1\sqrt{2},\ (ii) \|\mathbf{Y} - \mathbf{Y}'\|/\sqrt{d} \xrightarrow{P} \sigma_2\sqrt{2} \text{ and}$$
$$(iii)\ \|\mathbf{X} - \mathbf{Y}\|/\sqrt{d} \xrightarrow{P} \sqrt{\sigma_1^2 + \sigma_2^2 + \nu^2} \text{ as } d \to \infty. \quad (3)$$

So, depending on whether $\mathbf{z}$ comes from class-1 or class-2, $\mathbf{z}^*$ converges to $\mathbf{a}_1$ or $\mathbf{a}_2$ as before. For large values of d, $\mathbf{z}^*$ is expected to lie closer to the cluster formed by the transformed observations from the same class as $\mathbf{z}$. The same argument can be used for J class (with

$J > 2$) problems as well. The following theorem shows the high-dimensional optimality of the misclassification probability for the NN-TRAD classifier.

Theorem 1 *Suppose that the J competing classes satisfy assumptions (A1)–(A3). Also assume that $\sigma_i^2 \neq \sigma_j^2$ or $\nu_{ij}^2 > 0$ for every $1 \leq i < j \leq J$. Then the misclassification probability of NN-TRAD classifier converges to* 0 *as $d \rightarrow \infty$.*

The proof of this theorem is given in the "Appendix". Recall that under $(A1)$–$(A3)$, the classic NN classifier fails to achieve the optimal misclassification rate when $\nu_{ij}^2 < |\sigma_i^2 - \sigma_j^2|$ for some $1 \leq i < j \leq J$, but NN-TRAD works well even in such situations. Instead of $(A1)$–$(A3)$, Andrews (1988) and de Jong (1995) considered another set of assumptions to derive weak and strong laws of large numbers for mixingales. One may also use those assumptions to prove the distance convergence in high dimensions. Note that Francois et al. (2007) assumed stronger conditions for almost sure convergence of the distances.

3 A new transformation based on inter-point distances

The transformation based on average distances (TRAD) may fail to extract meaningful discriminating features from the data if one or more of the competing populations have some hidden sub-populations (e.g., the class distribution is a mixture of two or more unimodal distributions). In such cases, it may happen that $(A1)$–$(A3)$ do not hold for the whole class distribution, but they hold for each of the sub-class distributions. Consider an example where each class is an equal mixture of two d-dimensional (we used $d = 100$) normal distributions, each having the same dispersion matrix $\mathbf{I}_d$. For class-1, the location parameters of the two distributions were taken to be $\mathbf{0}_d$ and $(\mathbf{10}_2^T, \mathbf{0}_{d-2}^T)^T$, while in class-2 they were $(10, \mathbf{0}_{d-1}^T)^T$ and $(0, 10, \mathbf{0}_{d-2}^T)^T$. Taking an equal number of observations from these two classes, we generated a training set of size 20 and a test set of size 200. When TRAD was applied to this data, all transformed data points overlapped with each other (see Fig. 3). As a result, NN-TRAD misclassified almost half of the test set observations.

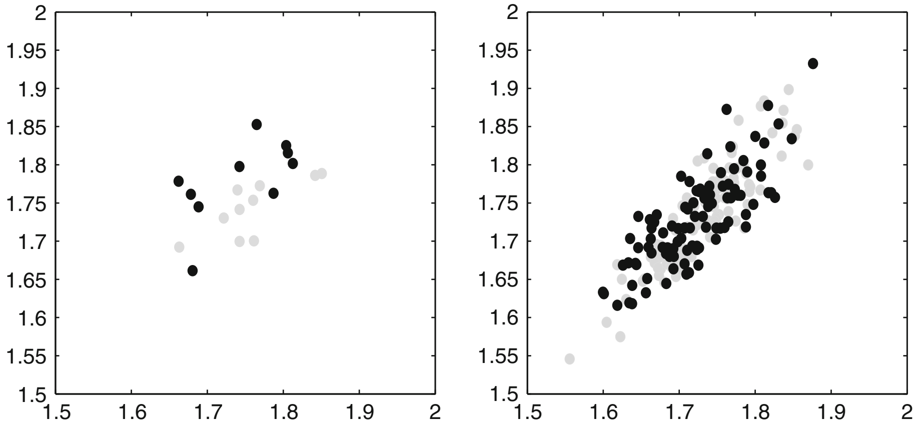


Fig. 3 Scatter plots of training (*left*) and test (*right*) data points after TRAD

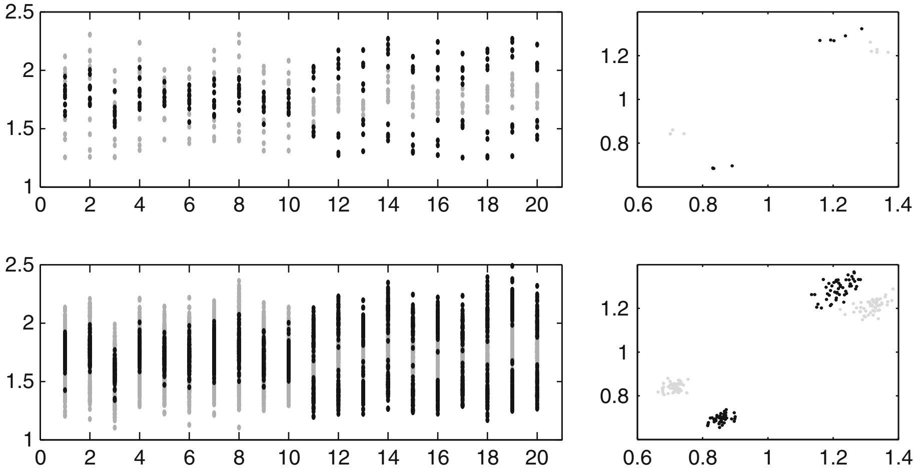


Fig. 4 The *left panel* shows different co-ordinates of the training (*top row*) and the test (*bottom row*) data points after TRIPD. The *right panel* shows the joint structure of the 10-th and the 11-th co-ordinate of the transformed observations

In order to overcome this limitation of TRAD and retain the discriminatory information contained in pairwise distances, we propose the following transformation of the training data:

$$\mathbf{x}_i^{**} = \left(\frac{\|\mathbf{x}_i - \mathbf{x}_1\|}{\sqrt{d}}, \ldots, \frac{\|\mathbf{x}_i - \mathbf{x}_{n_1}\|}{\sqrt{d}}, \frac{\|\mathbf{x}_i - \mathbf{y}_1\|}{\sqrt{d}}, \ldots, \frac{\|\mathbf{x}_i - \mathbf{y}_{n_2}\|}{\sqrt{d}}\right)^T \text{ and}$$

$$\mathbf{y}_j^{**} = \left(\frac{\|\mathbf{y}_j - \mathbf{x}_1\|}{\sqrt{d}}, \ldots, \frac{\|\mathbf{y}_j - \mathbf{x}_{n_1}\|}{\sqrt{d}}, \frac{\|\mathbf{y}_j - \mathbf{y}_1\|}{\sqrt{d}}, \ldots, \frac{\|\mathbf{y}_j - \mathbf{y}_{n_2}\|}{\sqrt{d}}\right)^T \tag{4}$$

for $1 \le i \le n_1$ and $1 \le j \le n_2$. Note that the i-th component in $\mathbf{x}_i^{**}$ is 0, while the (n_1+j)-th component in $\mathbf{y}_j^{**}$ is 0. For any new observation $\mathbf{z}$, using this transformation we get

$$\mathbf{z}^{**} = \left(\frac{\|\mathbf{z} - \mathbf{x}_1\|}{\sqrt{d}}, \ldots, \frac{\|\mathbf{z} - \mathbf{x}_{n_1}\|}{\sqrt{d}}, \frac{\|\mathbf{z} - \mathbf{y}_1\|}{\sqrt{d}}, \ldots, \frac{\|\mathbf{z} - \mathbf{y}_{n_2}\|}{\sqrt{d}}\right)^T. \tag{5}$$

Here, we get a $(n_1 + n_2)$-dimensional projection. In a J class problem, we consider a n-dimensional projection, where $n = n_1 + \cdots + n_J$. In the HDLSS setup (where d is larger than n), this transformation leads to substantial reduction in data dimension.

The plots in the left panel of Fig. 4 show co-ordinates of the transformed observations. The training and the test set cases from the two classes are indicated using gray and black dots, respectively. In each plot, we can observe two clusters of either black or gray dots along each co-ordinate. This gives us an indication that discriminative information is contained in almost all co-ordinates. It is more transparent from the scatter plots of the 10th and the 11th co-ordinates of the transformed observations shown in the right panel of Fig. 4. When the NN classifier was used after **TR**ansformation based on **I**nter-**P**oint **D**istances (henceforth referred to as NN-TRIPD), it correctly classified almost all test set observations.

Good performance of NN-TRIPD for classification among such high-dimensional mixture populations is asserted by Theorem 2(a) under assumption $(A4)$ stated below.

$(A4)$ *Suppose that the distribution of the* i-th $(1 \le i \le J)$ *class is a mixture of* R_i $(R_i \ge 1)$ *many sub-class distributions, where each of these sub-class distribu-*

tions satisfy $(A1)$–$(A3)$. *If* $\mathbf{X}$ *is from the* s*-th sub-class of the* i*-th class, and* $\mathbf{Y}$ *is from the* t*- th sub-class of the* j*- th class* $(1 \leq i \neq j \leq J, 1 \leq s \leq R_i, 1 \leq t \leq R_j)$ $\sum_{q=1}^{d} Var(X_q)/d$, $\sum_{q=1}^{d} Var(Y_q)/d$ *and* $\sum_{q=1}^{d}\{E(X_q) - E(Y_q)\}^2/d$ *converge to* $2\sigma_{i_s}^2, 2\sigma_{j_t}^2$ *and* $\nu_{i_s j_t}^2$, *respectively, as* $d \to \infty$.

If the competing classes satisfy $(A1)$–$(A3)$, $(A4)$ holds automatically. But, $(A4)$ holds in many other cases including the example involving mixture distributions discussed above, where $(A1)$–$(A3)$ fail to hold. The following theorem gives an idea about the asymptotic (as $d \to \infty$) behavior of the misclassification probability of the NN-TRIPD classifier under this assumption. Throughout this article, we will assume that there are at least two labeled observations from each of the sub-classes.

Theorem 2(a) *Suppose that the* J *competing classes satisfy assumption (A4). Further assume that for every* i, j, s *and* t *with* $1 \leq s \leq R_i, 1 \leq t \leq R_j, 1 \leq i \neq j \leq J$, *we either have* $\nu_{i_s j_t}^2 > |\sigma_{i_s}^2 - \sigma_{j_t}^2|$ *or* $0 < \nu_{i_s j_t}^2 < |\sigma_{i_s}^2 - \sigma_{j_t}^2| - 8(n_{i_s j_t} - 1)\max\{\sigma_{i_s}^2, \sigma_{j_t}^2\}/n_{i_s j_t}^2$, *where* $n_{i_s j_t}$ *is the total training sample size of these two sub-classes. Then the misclassification probability of NN-TRIPD classifier based on the* l_2 *norm converges to* 0 *as* $d \to \infty$.

The proof of the theorem is given in the "Appendix". Let us now consider the case when $J = 2$ and $R_1 = R_2 = 1$. Recall that NN-TRAD is optimal here in the sense that it only requires $\sigma_1^2 \neq \sigma_2^2$ or $\nu_{12}^2 > 0$ for the misclassification probability to go to zero. But, NN-TRIPD possess this asymptotic optimality if $0 < \nu_{12}^2 < |\sigma_1^2 - \sigma_2^2| - 8(n-1)\max\{\sigma_1^2, \sigma_2^2\}/n^2$ or $\nu_{12}^2 > |\sigma_1^2 - \sigma_2^2|$. For $n > 2$, we have a wide variety of examples (see, e.g., Example-1 in Sect. 5) where $\nu_{12}^2 < |\sigma_1^2 - \sigma_2^2| - 8(n-1)\max\{\sigma_1^2, \sigma_2^2\}/n^2 < |\sigma_1^2 - \sigma_2^2|$. In such cases, the classic NN classifier fails, but NN-TRIPD works well. NN-TRIPD has better theoretical properties if one uses the l_1 norm or the l_p norm with fractional $p(0 < p < 1)$ instead of the usual l_2 norm for NN classification in the transformed space. Like NN-TRAD, such a classifier requires only $\sigma_1^2 \neq \sigma_2^2$ or $\nu_{12}^2 > 0$ to achieve asymptotic optimality. One should note that both versions of NN-TRIPD usually outperform the NN-TRAD classifier if the competing populations are mixtures of several sub-populations.

Theorem 2(b) *Suppose that the* J *competing classes satisfy assumption (A4). Further assume that either* $\sigma_{i_s}^2 \neq \sigma_{j_t}^2$ *or* $\nu_{i_s j_t}^2 > 0$ *for every* $1 \leq s \leq R_i, 1 \leq t \leq R_j$ *and* $1 \leq i \neq j \leq J$. *Then, for any* $p \in (0, 1]$, *the misclassification probability of NN-TRIPD classifier based on the* l_p *norm converges to* 0 *as* $d \to \infty$.

The proof of the theorem is given in the "Appendix". Aggarwal et al. (2001) carried out an investigation on the performance of the l_p norms for high-dimensional data with varying choices of p. They observed that the l_p norm is more relevant for $p = 1$ and $p = 2$ than value of $p \geq 3$, while fractional values of $p(0 < p < 1)$ are quite effective in measuring proximity of data points. The relevance of the Euclidean distance has also been questioned in the past, and fractional norms were introduced to fight the concentration phenomenon (see, e.g., Francois et al. 2007).

4 Some discussions on NN-TRAD and NN-TRIPD

Although TRAD and TRIPD are motivated from theoretical results on concentration of pairwise distances, such transformations are not new in the machine learning literature (see, e.g., Cazzanti et al. 2008 for more details). For transformations based on principal component directions (see, e.g., Deegalla and Bostrom 2006) or multi-dimensional scaling (see, e.g., Young and Hoseholder 1938), one needs to compute pairwise distances among training sample observations. Like these methods, the proposed transformations also embed observations in a lower-dimensional subspace. Embedding of observations in Euclidean and pseudo-Euclidean space was also considered by Goldfarb (1985) and Pekalska et al. (2001).

In similarity based classification (see, e.g., Cazzanti et al. 2008; Chen et al. 2009), an index is defined to measure similarity of an observation with respect to training sample observations, and those similarity measures are used as features to develop a classifier. For instance, a nonlinear support vector machine (SVM) (see, e.g., Vapnik 1998) can be viewed as a similarity based classifier, where a kernel function is used to measure similarity/dissimilarity between two observations. Graepel et al. (1999) and Pekalska et al. (2001) used several standard learning techniques on these similarity based features. A discussion on similarity based classifiers including SVM, kernel Fisher discriminant analysis and those based on entropy can be found in Cazzanti et al. (2008). It has been observed in the literature that NN classifiers on similarity measures usually yield low misclassification rates (see, e.g., Cost and Salzberg 1993; Pekalska et al. 2001). One main goal of this paper is to provide a theoretical foundation for these two similarity based NN classifiers, namely, NN-TRAD and NN-TRIPD in the context of HDLSS data.

We now discuss the computational complexity of our methods. For transformation of n labeled observations, both NN-TRAD and NN-TRIPD require $O(n^2 d)$ computations to calculate all pairwise distances. For the classic NN classifier, one need not compute these distances unless a cross-validation type method is used to choose a value of k. However, this is an off-line calculation. Given a test case $\mathbf{z}$, all these methods need $O(nd)$ computations to calculate n distances. After the distance computation, the classic NN classifier with a fixed k requires $O(n)$ computations to find the k neighbors of $\mathbf{z}$ (see Aho et al. 1974). So, classification of $\mathbf{z}$ requires $O(nd)$ calculations. NN-TRAD performs NN classification in the transformed J-dimensional space. So, re-computation of n distances in that space and finding the k neighbors require $O(n)$ calculations. Therefore, its computational complexity for a test case is also $O(nd)$. In the case of NN-TRIPD, since the transformed space is n-dimensional, re-computation of distances and finding neighbors require $O(n^2)$ calculations. We deal with HDLSS data (where $d \gg n$) where $O(nd)$ dominates $O(n^2)$. In such situations, NN-TRIPD also requires $O(nd)$ computations to classify $\mathbf{z}$.

5 Results from the analysis of simulated data sets

We analyzed some high-dimensional simulated data sets to compare the performance of the classic NN, NN-TRAD and NN-TRIPD classifiers. For NN-TRIPD, we used l_p norms for several choices of p to compute the distance in the transformed space. The overall performance of NN-TRIPD classifiers for $p > 2$ was inferior compared with $p = 2$. The performance for fractional values of p was quite similar to $p = 1$. So, we have reported results for $p = 1$ and $p = 2$ only. These two classifiers are referred to as NN-TRIPD$_1$ and NN-TRIPD$_2$, respectively. In each example, we generated 10 observations from each of the two classes to form the training sample, while a test set of size 200 (100 from each class) was used. This procedure was repeated 250 times to compute the average test set misclassification rates of different classifiers. Average misclassification rates were computed for a set of increasing values of d (namely, 2, 5, 10, 20, 50, 100, 200 and 500).

Recall that the classic NN classifier needs the value of k to be specified. Existing theoretical results (see, e.g., Hall et al. 2008) give us an idea about the optimal order of k when the sample size n is large. But these results are not applicable to HDLSS data. In such situations (with low sample sizes), Chen and Hall (2009) suggested the use of $k = 1$. Hall et al. (2008) reported the optimal order of k to be $n^{4/(d+4)}$, which also tends to 1 as d grows to infinity. In practice, when we deal with a fixed sample size, we often use the cross-validation technique on the training data (see, e.g., Duda et al. 2000; Hastie et al. 2009) to choose the optimum value of k. However, there is high variability in the cross-validation estimate of the misclassification rate, and this method often fails to choose k appropriately (see, e.g., Hall et al. 2008; Ghosh and Hall 2008). In most of our experiments with simulated and real data sets, cross-validation led to inferior results compared to those obtained using $k = 1$. So, throughout this article, we used $k = 1$ for the classic NN classifier. To keep our comparisons fair, we used the same value of k for NN-TRAD, and for both versions of NN-TRIPD as well.

Let us begin with the following examples involving some normal distributions, and their mixtures.

Example-1 : Class-1: $N_d(\mathbf{0}_d, \mathbf{I}_d)$ and Class-2: $N_d((10, \mathbf{0}_{d-1}^T)^T, \mathbf{I}_d)$.

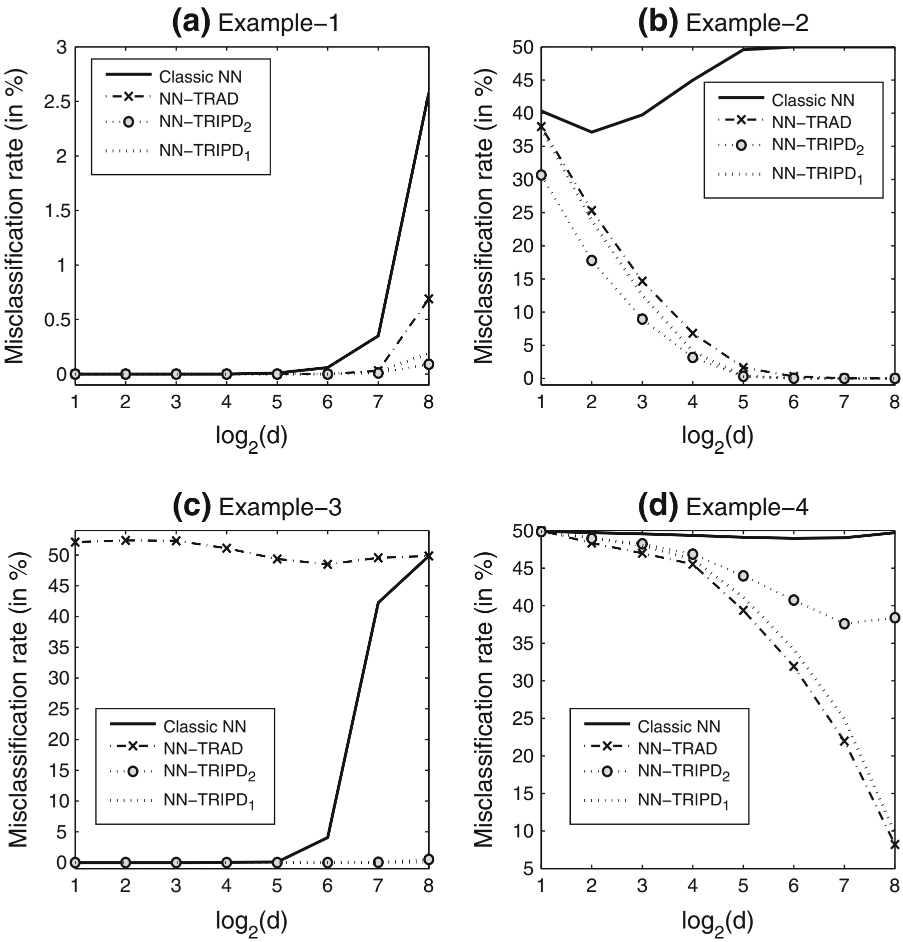
Example-2 : Class-1: $N_d(\mathbf{0}_d, \mathbf{I}_d)$ and Class-2: $N_d(\mathbf{0}_d, \frac{1}{4}\mathbf{I}_d)$.

Example-3 : Class-1: $\frac{1}{2}[N_d(\mathbf{0}_d, \mathbf{I}_d) + N_d((10, 10, \mathbf{0}_{d-2}^T)^T, \frac{1}{4}\mathbf{I}_d)]$ and
Class-2: $\frac{1}{2}[N_d((10, \mathbf{0}_{d-1}^T)^T, \mathbf{I}_d) + N_d((0, 10, \mathbf{0}_{d-2}^T)^T, \frac{1}{4}\mathbf{I}_d)]$.

Example-4 : Class-1: $N_d(\mathbf{0}_d, \mathbf{I}_d)$ and Class-2: $N_d(\mathbf{0}_d, 0.81\mathbf{I}_d)$.

In Examples-1, 2, and 4, assumptions $(A1)$–$(A3)$ hold for each of the competing classes. In Example-3, all competing classes satisfy $(A4)$ (i.e., $(A1)$–$(A3)$ hold for each sub-class). Average misclassification rates of different classifiers for varying values of d are shown in Fig. 5.

For the location problem in Example-1, the two competing classes are widely separated, and the Bayes risk is almost zero for any value of d. In this example, we have information about class separability only in the first co-ordinate and accumulate noise as the value of d increases. Surprisingly, the presence of noise did not have any significant effect on the performance of any of these classifiers (see Fig. 5a). Except for $d = 500$, all the classifiers correctly classified almost all test set observations. But, the picture changed completely for the scale problem in Example-2. Since, all the co-ordinates have discriminatory information, one should expect the misclassification rates of all classifiers to converge to zero as d increases. However, the misclassification rate of the classic NN classifier dived a bit when we moved from $d = 2$ to $d = 5$, but thereafter it gradually increased with d. In fact, it performed as worse as a random classifier for values of d greater than 50. On the other hand, the misclassification rates of NN-TRAD and both versions of NN-TRIPD decreased steadily as d increased. For $d \geq 100$, almost all test set observations were classified correctly. Recall that for large d, the classic NN classifier correctly classifies all unlabeled observations if $\nu_{12,d}^2 > |\sigma_{1,d}^2 - \sigma_{2,d}^2|$. In Example-1, we have $\sigma_{1,d}^2 = \sigma_{2,d}^2 = 1$ and $\nu_{12,d}^2 = 100/d$ for all d, and hence $\nu_{12,d}^2 > |\sigma_{1,d}^2 - \sigma_{2,d}^2|$. So, the classic NN classifier worked well. For high values of d, since the difference was smaller, it misclassified some observations. In Example-2, we have $|\sigma_{1,d}^2 - \sigma_{2,d}^2| = 3/4$ but $\nu_{12,d}^2$ is 0 for all d. So, the classic NN classifier yielded almost 50 % misclassification rate even for moderately high values of d. In both these examples, NN-TRAD had a slight edge over NN-TRIPD as none of the competing classes had any further sub-classes.



(a) Example–1

(b) Example–2

(c) Example–3

(d) Example–4

Fig. 5 Misclassification rates of classic NN, NN-TRAD, NN-TRIPD$_1$ and NN-TRIPD$_2$ for different values of d

In the presence of sub-populations in Example-3, NN-TRAD yielded almost 50 % misclassification rate for all values of d, but NN-TRIPD led to substantial improvement (see Fig. 5c). In fact, it correctly classified almost all the test set observations for any value of d. The classic NN classifier performed perfectly up to $d = 50$, but its misclassification rate increased thereafter. The misclassification rate was 4.05 % for $d = 100$, but it increased sharply to 42.26 % for $d = 200$. To explain this behavior, let us consider the first sub-class in class-1 and the second sub-class in class-2. We have $\sigma^2_{1_1,d} = 1$, $\sigma^2_{2_2,d} = 1/4$ and $\nu^2_{1_1 2_2,d} = 100/d$ for all d. Note that $\nu^2_{1_1 2_2,d}$ is larger than $|\sigma^2_{1_1,d} - \sigma^2_{2_2,d}|$ for $d \leq 100$, but it is smaller than $|\sigma^2_{1_1,d} - \sigma^2_{2_2,d}|$ for $d \geq 200$. The same holds for the second sub-class of class-1 and the first sub-class of class-2. This led to sharp increase in the misclassification rate when we moved from $d = 100$ to $d = 200$. Example-4 shows the superiority of NN-TRIPD$_1$ over NN-TRIPD$_2$. Unlike Example-1, the condition given in Theorem 2(a) (with $R_1 = R_2 = 1$) fails to hold in this scale problem. NN-TRAD had good performance in this example because of unimodality of the class distributions.

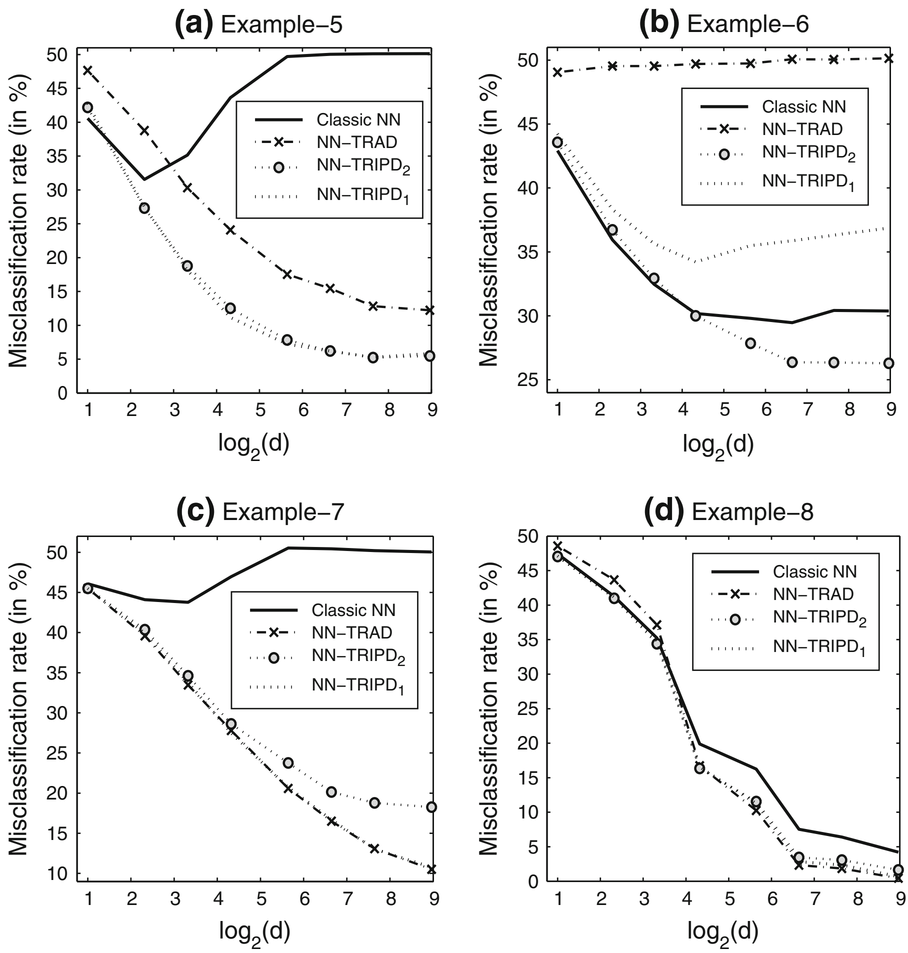


Fig. 6 Misclassification rates of classical NN, NN-TRAD, NN-TRIPD$_1$ and NN-TRIPD$_2$ for different values of d

We now consider some examples where the competing classes differ neither in their locations nor in their scales (i.e., $\sigma^2_{1,d} = \sigma^2_{2,d}$ and $\nu^2_{12,d} = 0$ for all d).

Example-5: Class-1: $N_d(\mathbf{0}_d, 0.9\mathbf{I}_d + 0.1\mathbf{J}_d)$ and Class-2: $N_d(\mathbf{0}_d, 0.1\mathbf{I}_d + 0.9\mathbf{J}_d)$, where $\mathbf{J}_d = \mathbf{1}_d\mathbf{1}_d^T$ and $\mathbf{1}_d = (1, \ldots, 1)^T$ is of length d.

Example-6: Class-1: $X_1 \sim N(0, \sigma^2)$ and $X_i = 0.5X_{i-1} + 0.5U_{i-1}$ for $i \geq 2$ and Class-2: $X_1 \sim N(0, \sigma^2)$ and $X_i = 0.5V_{i-1} - 0.5 X_{i-1}$ for $i \geq 2$, where U_i's and V_i's are i.i.d. $N(0, 1)$.

Example-7: Class-1: $N_d(\mathbf{0}_d, \mathbf{I}_d)$ and Class-2: $t_{3,d}(\mathbf{0}_d, \frac{1}{3}\mathbf{I}_d)$, where $t_{3,d}$ denotes the d-dimensional t-distribution with 3 d.f.

Figure 6 shows the performance of different classifiers for varying choices of d. In Examples-5 and 6, the two competing classes differ in their correlation structure. They have the same scatter matrix but differ in their shapes in Example-7. For Example-5, the classic NN classifier failed to discriminate between the two classes. For higher values of d, it misclassified almost half of the unlabeled observations. Performances of NN-TRIPD$_1$ and NN-TRIPD$_2$ were comparable, and both of them had lower misclassification rates than NN-TRAD. In Example-6, NN-TRAD performed very poorly, but the performance of NN-TRIPD was much better. In this example, NN-TRIPD$_2$ outperformed all its competitors. For $d = 500$, while NN-TRIPD$_1$ had an average misclassification rate of 36 %, NN-TRIPD$_2$ yielded an average misclassification rate close to 26 %. However, in Example-7 NN-TRIPD$_1$ had the best performance closely followed by NN-TRAD. For $d = 500$, the average misclassification rate of NN-TRIPD$_1$ was almost half of NN-TRIPD$_2$. We now consider an example from Chen and Hall (2009).

Example-8: Class-1: $X_1, X_2, \ldots, X_d$ are i.i.d. $N(0, 1)$ and Class-2: $X_i \overset{indep}{\sim} N(\nu_i, 1)$ for $1 \le i \le d$, where $\epsilon = d^{-\beta}$ proportion of ν_i's are $r \log d$ and the rest are 0, with $\beta = r = 0.75$.

In this example, the robust NN classifier of Chen and Hall (2009) failed to improve upon the performance of the classic NN classifier (see Chen and Hall (2009, p. 3201)), but NN-TRAD and both versions of NN-TRIPD outperformed the classic NN classifier for large values of d.

From the analysis of these simulated data sets, it is evident that both NN-TRIPD$_1$ and NN-TRIPD$_2$ have a clear edge over NN-TRAD for classifying HDLSS data. But, there is no clear winner among the first two. In practice, one needs to decide upon one of these two classifiers. We use the training sample to compute the leave-one-out cross-validation estimates of misclassification rates for both classifiers. The one with the lower misclassification rate is chosen, and it is used to classify all the tests cases. For further data analysis, this classifier will be referred to as the proposed classifier.

6 Comparison with other popular classifiers

We now compare the performance of our proposed classifier with some popular classifiers available in the literature. Here, we consider the examples studied in Sect. 5 for the case $d = 500$. As before, we use training sets and test sets of sizes 20 and 200, respectively, and each experiment is carried out 250 times. Table 1 shows the average test set misclassification rates of the classic NN classifier and our proposed NN classifier along with their corresponding standard errors reported within parentheses. Results are also reported for NN classifiers based on random projection (see, e.g., Fern and Brodley 2003) and principal component analysis [PCA] (see, e.g., Deegalla and Bostrom 2006). These two classifiers will be referred to as NN-RAND and NN-PCA, respectively. Misclassification rates are reported for linear and nonlinear (with radial basis function (RBF) kernel $K_\gamma(\mathbf{x}, \mathbf{y}) = \exp\{-\gamma\|\mathbf{x} - \mathbf{y}\|^2\}$) support vector machines [SVM] (see, e.g., Vapnik 1998) as well. In the case of SVM-RBF, the results are reported for the default value of the regularization parameter $\gamma = 1/d$ as used in http://www.csie.ntu.edu.tw/~cjlin/libsvm/. We also used the value of γ chosen by tenfold cross-validation method, but that did not yield any significant improvement in the performance of the resulting classifier. In fact, in more than half of the cases, $\gamma = 1/d$ led to lower misclassification rates than those obtained using the cross-validated choice of λ. In view of

Table 1 Misclassification rates (in %) of different classifiers on simulated data sets for $d = 500$ with the minimum indicated in bold

	Classic NN	NN-RAND	NN-PCA	RF	SVM-LIN	SVM-RBF	GLMNET	Proposed
Ex.- 1	2.58	20.76	26.50	24.84	43.01	43.69	**0.00**	0.43
	(0.12)	(0.33)	(0.34)	(0.33)	(0.23)	(0.23)	(0.00)	(0.04)
Ex.- 2	50.00	43.58	49.97	39.59	49.18	**0.00**	42.21	**0.00**
	(0.00)	(0.01)	(0.00)	(0.23)	(0.01)	(0.00)	(0.02)	(0.00)
Ex.- 3	49.85	5.86	0.23	24.23	40.72	41.84	**0.00**	0.30
	(0.02)	(0.17)	(0.07)	(0.39)	(0.21)	(0.20)	(0.00)	(0.03)
Ex.- 4	49.72	49.03	49.48	48.36	49.12	32.23	49.72	**9.74**
	(0.07)	(0.18)	(0.23)	(0.21)	(0.21)	(0.27)	(0.22)	(0.21)
Ex.- 5	50.13	50.07	40.82	50.35	50.02	9.27	42.05	**5.57**
	(0.06)	(0.06)	(0.23)	(0.28)	(0.07)	(0.50)	(0.18)	(0.28)
Ex.- 6	30.38	45.58	43.11	49.77	49.77	46.51	49.82	**29.27**
	(0.32)	(0.23)	(0.25)	(0.22)	(0.24)	(0.21)	(0.22)	(0.34)
Ex.- 7	50.06	50.56	46.04	44.83	49.29	19.82	46.32	**11.14**
	(0.01)	(0.04)	(0.17)	(0.21)	(0.11)	(0.52)	(0.19)	(0.33)
Ex.- 8	4.21	24.06	27.39	0.74	26.46	28.38	**0.03**	0.91
	(0.13)	(0.35)	(0.41)	(0.05)	(0.14)	(0.23)	(0.01)	(0.06)

high statistical instability of cross-validation estimates, this is expected in HDLSS situations, GLMNET, a method of logistic regression that uses convex combination of lasso and ridge penalties for dimension reduction (see, e.g., Simon et al. 2011), and a boosted version of classification tree known as random forest [RF] (see, e.g., Breiman 2001; Liaw and Wiener 2002) have also been used. For all these methods, we used available R codes with default tuning parameters.

Our proposed classifier had the best overall performance among the classification methods considered here. It yielded the best performance in *five out of these eight* data sets. In other cases, its misclassification rates were quite close to the minimum. The NN classifiers based on dimension reduction techniques (i.e., NN-RAND and NN-PCA) had substantially higher misclassification rates than the classic NN classifier in Examples-1, 6 and 8. Only in Example-3, they yielded much lower misclassification rates compared to the classic NN classifier. Among other competitors, GLMNET yielded the best misclassification rate in three data sets, but it had very poor performance in five other data. In Examples-1, 3 and 8, we have discriminatory information only in a few components. GLMNET is a classification method developed specifically for this type of sparse data, and hence it performed well in these examples. The linear SVM classifier is expected to perform well when the population distributions differ in their locations. However, in the presence of small training samples, it failed to extract sufficient discriminating information. It yielded high misclassification rates even for the location problem in Example-1. Its nonlinear version, SVM-RBF had better performance. It led to the lowest misclassification rate in Example-2. RF had competitive performance in Example-8, but in all other examples, its performance was not comparable to our method.

6.1 Comparison based on the analysis of benchmark data sets

We further analyzed twenty benchmark data sets for assessment of our proposed method. The first fourteen data sets listed in the UCR Time Series Classification/Clustering Page (http://www.cs.ucr.edu/~eamonn/time_series_data/) are considered. The Madelon data set is from the UCI machine learning repository (http://archive.ics.uci.edu/ml/datasets.html). We also considered the first five data sets listed in the Kent Ridge Bio-medical Data Set Repository (http://levis.tongji.edu.cn/gzali/data/mirror-ketridge.html). All these data sets have specific training and test sets. However, instead of using those single training and test samples, we used random partitioning of the whole data to form 250 training and test sets. The sizes of the training and the test sets in each partition are reported in Table 2. Average misclassification rates of different classifiers were computed over these 250 partitions, and they are reported in Table 3 along with their corresponding standard errors inside parentheses.

The overall performance of the proposed method was fairly competitive. In *six out of these twenty* data sets, our proposed classifier yielded the lowest misclassification rate. It was either in second or in third position in six other data sets. One should also notice that in twelve out of these twenty data sets, the proposed method outperformed the classic NN classifier. In majority of the cases, it had lower misclassification rates than NN-RAND and NN-PCA as well. Among other classifiers, the overall performance of RF turned out to be very competitive, and it outperformed our proposed classifier in seven out of sixteen data sets. However, the R code for RF was computationally infeasible on four data sets with dimension greater than 7000. NN-PCA had similar problems with memory as the data dimension was high. For such high-dimensional data sets, GLMNET and linear SVM had competitive performance. In many of these high-dimensional benchmark data sets, the measurement variables were highly correlated. The intrinsic dimension of the data was low, and the pairwise distances failed to concentrate. As a consequence, although our proposed method had competitive performance, its superiority over other classifiers was not as prominent as it was in the simulated examples.

Table 2 Brief description of benchmark data sets

Data set	Dimension	Class	Train size	Test size	Data set	Dimension	Class	Train size	Test size
Syn. contorl	60	6	60	540	Face (four)	350	4	40	72
Gun-point	150	2	20	180	Lightning-2	637	2	60	61
CBF	128	3	30	900	Lightning-7	319	7	70	73
Face (all)	131	14	250	2000	ECG	96	2	30	171
OSU leaf	427	6	120	322	Madelon	500	2	600	2000
Swedish leaf	128	15	150	975	ALL-AML luke	7129	2	36	36
50 Wors	270	50	450	455	Breast cancer	24,481	2	49	48
Trace	275	4	60	140	Cent. nerv. sys.	7129	2	30	30
Two patterns	128	4	1000	4000	Colon tumor	2000	2	30	32
Wafer	152	2	700	6474	MLL lukemia	12,582	3	36	36

Table 3 Misclassification rates (in %) of different classifiers on benchmark data sets with the minimum indicated in bold

Data set	Classic NN	NN-RAND	NN-PCA	RF	SVM-LIN	SVM-RBF	GLMNET	Proposed
Synthetic control	18.45	21.11	21.38	14.14	14.93	10.11	24.04	**6.74**
	(0.15)	(0.19)	(0.21)	(0.14)	(0.13)	(0.20)	(0.17)	(0.14)
Gun-point	19.65	20.01	20.48	**15. 37**	19.97	24.28	20.56	21.21
	(0.32)	(0.34)	(0.37)	(0.36)	(0.35)	(0.30)	(0.36)	(0.30)
CBF	12.33	15.13	19.84	8.88	9.76	11.54	17.02	**7.67**
	(0.24)	(0.24)	(0.26)	(0.15)	(0.17)	(0.22)	(0.21)	(0.20)
Face (all)	20.09	22.94	23.17	**17.84**	22.00	20.20	28.69	25.54
	(0.07)	(0.07)	(0.08)	(0.08)	(0.08)	(0.08)	(0.09)	(0.09)
OSU leaf	**44.06**	45.45	51.55	47.70	60.82	45.83	59.20	44.78
	(0.18)	(0.17)	(0.20)	(0.21)	(0.19)	(0.20)	(0.20)	(0.18)
Swedish leaf	33.60	34.23	35.74	23.98	24.99	**23.87**	33.39	31.16
	(0.11)	(0.12)	(0.14)	(0.11)	(0.11)	(0.11)	(0.13)	(0.12)
50 Words	33.51	34.71	37.11	35.43	35.05	39.05	46.28	**31.53**
	(0.10)	(0.10)	(0.11)	(0.11)	(0.10)	(0.11)	(0.15)	(0.10)
Trace	29.26	29.24	33.94	19.29	**18.48**	27.00	23.50	28.66
	(0.22)	(0.22)	(0.25)	(0.27)	(0.21)	(0.19)	(0.26)	(0.24)
Two patterns	9.61	12.73	10.16	15.76	19.89	10.78	20.08	**9.59**
	(0.04)	(0.06)	(0.03)	(0.08)	(0.06)	(0.05)	(0.05)	(0.04)
Wafer	28.93	34.18	29.70	**22.98**	27.07	26.00	28.32	29.48
	(0.03)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)
Face (four)	14.75	16.27	18.02	**5.77**	7.96	8.74	12.18	18.89
	(0.27)	(0.28)	(0.30)	(0.17)	(0.19)	(0.18)	(0.26)	(0.28)
Lightning 2	28.11	28.60	28.16	22.08	35.64	27.83	33.72	**21.97**
	(0.31)	(0.31)	(0.32)	(0.32)	(0.35)	(0.32)	(0.37)	(0.28)
Lightning 7	37.33	39.12	39.75	**27.19**	34.59	37.37	47.30	32.86
	(0.25)	(0.30)	(0.31)	(0.28)	(0.29)	(0.24)	(0.31)	(0.26)
ECG	**16.50**	16.86	21.70	20.71	23.34	22.77	23.63	18.35
	(0.22)	(0.24)	(0.42)	(0.14)	(0.40)	(0.40)	(0.39)	(0.23)
Madelon	35.87	43.01	37.72	36.65	47.45	42.88	47.41	**29.71**
	(0.08)	(0.10)	(0.09)	(0.10)	(0.10)	(0.07)	(0.08)	(0.06)
ALL-AML lukemia	10.23	11.12	–	–	**5.03**	25.02	7.14	14.68
	(0.26)	(0.29)	–	–	(0.30)	(0.30)	(0.27)	(0.32)
Breast cancer	43.36	45.38	–	–	38.32	48.57	**36.33**	41.92
	(0.36)	(0.40)	–	–	(0.40)	(0.32)	(0.41)	(0.42)

Table 3 continued

Data set	Classic NN	NN-RAND	NN-PCA	RF	SVM-LIN	SVM-RBF	GLMNET	Proposed
Cental nervous system	38.70	42.09	–	–	36.83	**35.72**	41.32	41.69
	(0.42)	(0.48)	–	–	(0.48)	(0.44)	(0.46)	(0.45)
Colon tumor	24.06	24.23	25.27	19.12	**17.19**	28.82	19.06	19.03
	(0.36)	(0.34)	(0.35)	(0.30)	(0.28)	(0.33)	(0.30)	(0.33)
MLL leukemia	9.93	10.47	–	–	7.26	17.54	**6.88**	9.72
	(0.26)	(0.28)	–	–	(0.29)	(0.35)	(0.25)	(0.26)

7 Kernelized versions of NN-TRAD and NN-TRIPD

Observe that NN-TRAD and NN-TRIPD transform the data based on pairwise distances, and use the NN classifier on the transformed data. Classifiers like nonlinear SVM (see, e.g., Vapnik 1998) also adopt a similar idea. Using a function $\Phi : \mathbb{R}^d \to \mathcal{H}$, it projects multivariate observations to the reproducing kernel Hilbert space $\mathcal{H}$, and then constructs a linear classifier on the transformed data. Inner product between any two observations $\Phi(\mathbf{x})$ and $\Phi(\mathbf{y})$ in $\mathcal{H}$ is given by $\langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle = K_\gamma(\mathbf{x}, \mathbf{y})$, where K_γ is a positive definite (reproducing) kernel, and γ is the associated regularization parameter. Kernel Fisher discriminant analysis (see, e.g., Hofmann et al. 2008) is also based on this idea. The performance of these classifiers depends on the choice of K_γ and γ. Although RBF kernel is quite popular in the literature, it works well if γ is chosen appropriately. Unlike these methods, NN-TRAD and NN-TRIPD does not involve any tuning parameter.

Now, let us investigate the performance of the classic NN classifier on the transformed data in $\mathcal{H}$. Assume the kernel K_γ to be isotropic, i.e., $K_\gamma(\mathbf{x}, \mathbf{y}) = g(\gamma^{1/2}\|\mathbf{x}-\mathbf{y}\|)$, where $g(t) > 0$ for $t \neq 0$ (see, e.g., Genton 2001). The squared distance between $\Phi(\mathbf{x})$ and $\Phi(\mathbf{y})$ in $\mathcal{H}$ is given by $\|\Phi(\mathbf{x}) - \Phi(\mathbf{y})\|^2 = K_\gamma(\mathbf{x}, \mathbf{x}) + K_\gamma(\mathbf{y}, \mathbf{y}) - 2K_\gamma(\mathbf{x}, \mathbf{y}) = 2g(0) - 2g(\gamma^{1/2}\|\mathbf{x}-\mathbf{y}\|)$. Clearly, if g is monotonically decreasing in $[0, \infty)$ (for the RBF kernel, we have $g(t) = e^{-t^2}$), the ordering of pairwise distances in $\mathcal{H}$ remains the same. So, the NN classifier in $\mathcal{H}$ will inherit the same problems as the classic NN classifier in $\mathbb{R}^d$.

To understand this better, assume $(A1)$–$(A3)$ and let g be continuous with $g(t) \to 0$ as $t \to \infty$. Suppose $\mathbf{X}, \mathbf{X}'$ are two independent observations from class-1 and $\mathbf{Y}, \mathbf{Y}'$ are two independent observations from class-2. If γ remains fixed as d increases (or, γ decreases slowly such that $\gamma d \to \infty$ as $d \to \infty$), then $g(\gamma^{1/2}\|\mathbf{X}-\mathbf{Y}\|) = g((\gamma d)^{1/2} d^{-1/2}\|\mathbf{X}-\mathbf{Y}\|) \xrightarrow{P} 0$ as $d \to \infty$. So, separability among the competing classes decreases in high dimensions. Similarly, if $\gamma d \to 0$ as $d \to \infty$, $g(\gamma^{1/2}\|\mathbf{X}-\mathbf{X}'\|)$, $g(\gamma^{1/2}\|\mathbf{Y}-\mathbf{Y}'\|)$ and $g(\gamma^{1/2}\|\mathbf{X}-\mathbf{Y}\|)$ all converge in probability to $g(0)$. The kernel transformation becomes non-discriminative in both cases. Therefore, $\gamma = O(1/d)$ seems to be the optimal choice, and this justifies the use of $\gamma = 1/d$ as a default value in http://www.csie.ntu.edu.tw/~cjlin/libsvm/. Henceforth, we will assume that $\gamma d \to a_0$ as $d \to \infty$. Using (3), it can now be shown that

$$(i)\ \|\Phi(\mathbf{X}) - \Phi(\mathbf{X}')\|^2 \xrightarrow{P} 2g(0) - 2g\left(\sigma_1\sqrt{2a_0}\right),$$

$$(ii)\ \|\Phi(\mathbf{Y}) - \Phi(\mathbf{Y}')\|^2 \xrightarrow{P} 2g(0) - 2g\left(\sigma_2\sqrt{2a_0}\right) \text{ and}$$

$$(iii)\ \|\Phi(\mathbf{X}) - \Phi(\mathbf{Y})\|^2 \xrightarrow{P} 2g(0) - 2g\left(\sqrt{a_0(\sigma_1^2 + \sigma_2^2 + \nu_{12}^2)}\right) \tag{6}$$

Table 4 Misclassification rates (in %) of usual and kernelized version of proposed classifiers on simulated data sets with the minimum indicated in bold

	Ex-1	Ex-2	Ex-3	Ex-4	Ex-5	Ex-6	Ex-7	Ex-8
Usual	**0.43**	**0.00**	**0.30**	9.74	5.57	**29.27**	11.14	**0.91**
	(0.04)	(0.00)	(0.03)	(0.21)	(0.28)	(0.34)	(0.33)	(0.06)
Kernel	0.72	**0.00**	0.42	**9.68**	**4.51**	29.92	**10.73**	1.18
	(0.06)	(0.00)	(0.04)	(0.21)	(0.24)	(0.34)	(0.34)	(0.08)

as $d \to \infty$. Since g is monotone, the NN classifier on the transformed observations in $\mathcal{H}$ classifies all observations to a single class if $\nu_{12}^2 < |\sigma_1^2 - \sigma_2^2|$ (like the classic NN classifier in $\mathbb{R}^d$).

The nonlinear SVM classifier constructs a linear classifier in the transformed space $\mathcal{H}$. Under $(A1)$–$(A3)$, Hall et al. (2005, p. 434) showed that in a high-dimensional two class problem, the linear SVM classifier classifies all observations to a single class if $|\sigma_1^2/n_1 - \sigma_2^2/n_2|$ exceeds ν_{12}^2. Following their argument and replacing σ_1^2, σ_2^2 and ν_{12}^2 by $g(0) - g(\sigma_1\sqrt{2a_0})$, $g(0) - g(\sigma_2\sqrt{2a_0})$ and $g(\sigma_1\sqrt{2a_0}) + g(\sigma_2\sqrt{2a_0}) - 2g(\sqrt{a_0(\sigma_1^2 + \sigma_2^2 + \nu_{12}^2)})$, respectively (compare Eq. (3) from p. 7 and Eq. (6)), one can derive a similar condition when the nonlinear SVM classifier based on RBF classifies all observations to a single class. However, NN-TRAD and NN-TRIPD can work well on the transformed observations, and one can derive results analogous to Theorems 1, 2(a) and 2(b). The results are summarized below.

Theorem 3 *Assume that the reproducing kernel of the Hilbert space $\mathcal{H}$ is of the form $K_\gamma = g(\gamma^{1/2}\|\mathbf{x}-\mathbf{y}\|)$, where (i) $g : [0, \infty) \to (0, \infty)$ is continuous and monotonically decreasing, and (ii) $\gamma d \to a_0 (> 0)$ as $d \to \infty$.*

(a) Under the conditions of Theorem 1, the misclassification probability of the kernelized version of NN-TRAD classifier converges to 0 *as $d \to \infty$.*
(b) Suppose that the J competing classes satisfy (A4). Also assume the inequality in Theorem 2(a) with $\sigma_{i_s}^2, \sigma_{j_t}^2$ and $\nu_{i_s j_t}^2$ replaced by $g(0) - g(\sigma_{i_s}\sqrt{2a_0})$, $g(0) - g(\sigma_{j_t}\sqrt{2a_0})$ and $g(\sigma_{i_s}\sqrt{2a_0}) + g(\sigma_{j_t}\sqrt{2a_0}) - 2g(\sqrt{a_0(\sigma_{i_s}^2 + \sigma_{j_t}^2 + \nu_{i_s j_t}^2)})$, respectively. Then, the misclassification probability of the kernelized version of NN-TRIPD classifier based on the l_2 norm converges to 0 *as $d \to \infty$.*
(c) Under the conditions of Theorem 2(b), for any $p \in (0, 1]$, the misclassification probability of the kernelized version of NN-TRIPD classifier based on the l_p norm converges to 0 *as $d \to \infty$.*

We have used kernelized versions of NN-TRAD and NN-TRIPD classifiers on all the simulated and benchmark datasets from Sect. 6. The overall performance of the latter turned out to be better. Tables 4 and 5 present average misclassification rates of the kernelized NN-TRIPD classifier along with their corresponding standard errors. Misclassification rates of the usual NN-TRIPD classifier are also shown alongside to facilitate comparison.

For the kernelized version, we have used $\gamma = 1/d$ for the first fourteen data sets. In the other six cases, we used $\gamma = 10^{-t}/d$ (i.e., $a_0 = 10^{-t}$), where the non-negative integer t was chosen based on a small pilot survey. The overall performance of the kernelized version was fairly competitive. In three out of eight simulated data sets, it had lower misclassification rates than the usual version. The usual version had better performance in *four out of eight*

Table 5 Misclassification rates (in %) of usual and kernelized version of proposed classifiers on benchmark data sets with the minimum indicated in bold

Data	Usual	Kernel	Data	Usual	Kernel	Data	Usual	Kernel	Data	Usual	Kernel
Synth. cont.	6.74	**6.47**	Swed. leaf	31.16	**30.63**	Face (Four)	18.89	**17.36**	ALL-AML lukemia	14.68	**14.43**
	(0.14)	(0.14)		(0.12)	(0.12)		(0.28)	(0.29)		(0.32)	(0.33)
Gun point	**21.21**	21.76	50 Words	**31.53**	33.07	Lightning 2	**21.97**	22.80	Breast cancer	41.92	**41.50**
	(0.30)	(0.29)		(0.10)	(0.10)		(0.28)	(0.28)		(0.42)	(0.41)
CBF	**7.67**	7.82	Trace	28.66	**28.23**	Lightning 7	**32.86**	33.89	Central ner. sys.	**41.69**	41.97
	(0.20)	(0.21)		(0.24)	(0.25)		(0.26)	(0.28)		(0.45)	(0.47)
Face (all)	25.54	**23.01**	Two patterns	9.59	**8.41**	ECG	18.35	**17.88**	Colon tumor	**19.03**	19.32
	(0.09)	(0.09)		(0.04)	(0.04)		(0.23)	(0.22)		(0.33)	(0.34)
OSU leaf	44.78	**43.74**	Wafer	**29.45**	29.83	Madelon	29.71	**29.67**	MLL lukemia	9.72	**9.43**
	(0.18)	(0.19)		(0.04)	(0.04)		(0.06)	(0.06)		(0.26)	(0.27)

examples. In Example-2, both versions correctly classified all test set observations. The kernelized version performed better than the usual version in *twelve out of twenty* benchmark data sets as well.

8 Concluding remarks

In this article, we have proposed some nonlinear transformations of the data for nearest neighbor classification in the HDLSS setup. While the classic NN classifier suffers due to distance concentration in high dimensions, these transformations use this property to their advantage and enhance class separability in the transformed space. When the NN classifier is used on the transformed data, the resulting classifiers usually lead to improved performance. Using several simulated and real data sets, we have amply demonstrated this. We have derived asymptotic optimality of the misclassification probabilities for the resulting classifiers in the HDLSS asymptotic regime, where the sample size remains fixed and the dimension of the data grows to infinity. Similar optimality results have been derived for kernelized versions of these classifiers as well. As future work, it would be interesting to study the behavior of these classifiers in situations where the sample size increases simultaneously with the data dimension.

Throughout this article, we have used $k = 1$ for all nearest neighbor classifiers. However, NN-TRAD and NN-TRIPD classifiers with other values of k had better performance in some data sets. Due to high stochastic variation, the cross-validation method often failed to select those values of k. Other re-sampling techniques could be helpful in such cases. Similarly, other resampling methods can be used to choose between $p = 1$ and $p = 2$ in our proposed classifier. Recall that NN-TRAD often performs poorly in the presence of sub-classes. So, if we can identify these hidden sub-classes using an appropriate clustering algorithm, the performance of NN-TRAD can be improved. Similarity based clustering methods (see, e.g., Ding et al. 2005; Arora et al. 2013) can be used for this purpose. In this article, a theoretical investigation has been carried out on the good properties of the proposed transformations in the case of the nearest neighbor classification. A study for other well-known classifiers remains to be investigated.

Acknowledgments The authors would like to thank the Associate Editor and three anonymous referees for their careful reading of earlier versions of the manuscript. Their helpful comments and suggestions led to a substantial improvement of the manuscript. The first author is also thankful to Dr. Raphaël Huser for some useful comments.

Conflict of interest The authors declare that they have no conflict of interest.

Appendix: Proofs and mathematical details

Lemma 1 *If a sequence of random variables $\{W_q,\ q \geq 1\}$ has uniformly bounded second moments and $sup_{1\leq q,q'<\infty,|q-q'|>r}|Corr(W_q, W_{q'})| < \rho(r)$, where $\rho(r) \to 0$ as $r \to \infty$, then WLLN holds for the sequence $\{W_q,\ q \geq 1\}$.*

Proof of Lemma 1 Since the sequence of random variables $\{W_q,\ q \geq 1\}$ has uniformly bounded second moments, we have $\sup_{1\leq q<\infty} V(W_q) < C$ for some constant $C > 0$. For any fixed d, we have

$$V\left[\frac{1}{d}\sum_{q=1}^{d} W_q\right] \leq \frac{C}{d} + \frac{C}{d^2}\left[\sum_{q\neq q'} Corr(W_q, W_{q'})\right].$$

For every $\epsilon > 0$, one can choose an integer R_ϵ such that for every $r > R_\epsilon$, $|\rho(r)| < \epsilon/2C$. If we take $d > 6CR_\epsilon/\epsilon$, then

$$V\left[\frac{1}{d}\sum_{q=1}^{d} W_q\right] \leq \frac{C}{d} + \frac{C}{d^2}\left[\sum_{(q,q^{'}):1\leq|q-q^{'}|\leq R_\epsilon} Corr(W_q, W_{q^{'}})\right] + \frac{C}{d^2}\left[\sum_{(q,q^{'}):|q-q^{'}|>R_\epsilon} Corr(W_q, W_{q^{'}})\right] \leq \frac{C}{d} + \frac{2CR_\epsilon}{d} + \frac{\epsilon}{2} < \epsilon. \quad (7)$$

Equation (7) now implies $E\left[\frac{1}{d}\sum_{q=1}^{d} W_q - \frac{1}{d}\sum_{q=1}^{d} E[W_q]\right]^2 \to 0$ as $d \to \infty$, which in turn proves that $\left|\frac{1}{d}\sum_{q=1}^{d} W_q - \frac{1}{d}\sum_{q=1}^{d} E[W_q]\right| \xrightarrow{P} 0$ as $d \to \infty$. □

Proof of Theorem 1 We give the proof for $J = 2$. Suppose that $\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}$ and $\mathbf{y}_1, \ldots, \mathbf{y}_{n_2}$ are observations from two classes. Assuming $(A1) - (A3)$, recall the following from Eq. (3)

$$\mathbf{x}_i^* \xrightarrow{P} \mathbf{a}_1 = \left(\sigma_1\sqrt{2},\ \sqrt{\sigma_1^2 + \sigma_2^2 + \nu_{12}^2}\right) \text{ for } 1 \leq i \leq n_1, \text{ and}$$

$$\mathbf{y}_j^* \xrightarrow{P} \mathbf{a}_2 = \left(\sqrt{\sigma_1^2 + \sigma_2^2 + \nu_{12}^2},\ \sigma_2\sqrt{2}\right) \text{ for } 1 \leq j \leq n_2 \text{ as } d \to \infty. \quad (8)$$

Now, if a future observation $\mathbf{z}$ comes from the first class, we have $\mathbf{z}^* \xrightarrow{P} \mathbf{a}_1$ as $d \to \infty$. So, $\|\mathbf{z}^* - \mathbf{x}_i^*\|^2 \xrightarrow{P} 0$ for all $1 \leq i \leq n_1$, but $\|\mathbf{z}^* - \mathbf{y}_j^*\|^2 \xrightarrow{P} \|\mathbf{a}_1 - \mathbf{a}_2\|^2$ for all $1 \leq j \leq n_2$ as $d \to \infty$. Further, $\|\mathbf{a}_1 - \mathbf{a}_2\|^2 = 0$ if and only if $\sigma_1^2 = \sigma_2^2$ and $\nu_{12}^2 = 0$. Under the condition of the theorem, we have $\|\mathbf{a}_1 - \mathbf{a}_2\|^2 > 0$. Therefore, NN-TRAD correctly classifies $\mathbf{z}$ with probability tending to 1 as $d \to \infty$. Again, if $\mathbf{z}$ comes from the second class, we have $\mathbf{z}^* \xrightarrow{P} \mathbf{a}_2$ as $d \to \infty$. While $\|\mathbf{z}^* - \mathbf{x}_i^*\|^2 \xrightarrow{P} \|\mathbf{a}_1 - \mathbf{a}_2\|^2$ for all $1 \leq i \leq n_1$, $\|\mathbf{z}^* - \mathbf{y}_j^*\|^2 \xrightarrow{P} 0$ for all $1 \leq j \leq n_2$. Therefore, NN-TRAD correctly classifies $\mathbf{z}$ with probability tending to 1 as $d \to \infty$. Combining these above facts, now the proof follows from an application of the Dominated Convergence Theorem. For $J > 2$, this result can be proved using similar arguments. □

Proof of Theorem 2 (a) For the sake of simplicity, here we prove the result for the case when $J = 2$ and $R_1 = R_2 = 2$. Let us assume that there are n_{i_s} observations from the s-th sub-class in the i-th class, denoted by $P_{i_s}(i = 1, 2;\ s = 1, 2)$. Suppose that $\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}$ are observations from a the first class and $\mathbf{y}_1, \ldots, \mathbf{y}_{n_2}$ are observations from the second class. Without loss of generality, also assume that $\mathbf{x}_1, \ldots, \mathbf{x}_{n_{1_1}}$ are from P_{1_1}, while $\mathbf{x}_{n_{1_1}+1}, \ldots, \mathbf{x}_{n_{1_1}+n_{1_2}}$ are from $P_{1_2}(n_{1_1} + n_{1_2} = n_1)$. Similarly, assume that $\mathbf{y}_1, \ldots, \mathbf{y}_{n_{2_1}}$ are from P_{2_1}, and $\mathbf{y}_{n_{2_1}+1}, \ldots, \mathbf{y}_{n_{2_1}+n_{2_2}}$ are from $P_{2_2}(n_{2_1} + n_{2_2} = n_2)$. Since these sub-classes satisfy $(A1) - (A3)$ (recall $(A4)$), for any observation $\mathbf{x}_i$ (respectively, $\mathbf{y}_j$), $\mathbf{x}_i^{**} \xrightarrow{P} \mathbf{a}_{\mathbf{x}_i}$ (respectively, $\mathbf{y}_j^{**} \xrightarrow{P} \mathbf{a}_{\mathbf{y}_j}$) as $d \to \infty$, where $\mathbf{a}_{\mathbf{x}_i}$ (respectively, $\mathbf{a}_{\mathbf{y}_j}$) is a point in the $(n_1 + n_2)$-dimensional space.

Consider an observation $\mathbf{x}_1$ from P_{1_1}. The third column in Table 6 shows the elements of $\mathbf{a}_{\mathbf{x}_1}$. The first element of $\mathbf{a}_{\mathbf{x}_1}$ is $\|\mathbf{x}_1 - \mathbf{x}_1\| = 0$. The next $(n_{1_1} - 1)$ elements are $\sigma_{1_1}\sqrt{2}$, and they are the limiting values of scaled distances of $\mathbf{x}_1$ from other observations in P_{1_1}. Then we have n_{1_2} scaled distances of $\mathbf{x}_1$ from the observations in P_{1_2} which are $\sqrt{\sigma_{1_1}^2 + \sigma_{1_2}^2 + \nu_{1_1 1_2}^2}$. Similarly, the next n_{2_1} elements are $\sqrt{\sigma_{1_1}^2 + \sigma_{2_1}^2 + \nu_{1_1 2_1}^2}$, and they are followed by n_{2_2} elements all equal to $\sqrt{\sigma_{1_1}^2 + \sigma_{2_2}^2 + \nu_{1_1 2_2}^2}$. For any future observation $\mathbf{z}$ from P_{1_1}, $\mathbf{a}_{\mathbf{z}}$ differs from

Table 6 Elements of $\mathbf{a}_{\mathbf{x}_1}$, $\mathbf{a}_{\mathbf{y}_1}$ and $\mathbf{a}_{\mathbf{z}}$, where $\mathbf{z}$ comes from P_{1_1}

No. of elements	Limiting distance from (after re-scaling by $d^{-1/2}$)	Elements of $\mathbf{a}_{\mathbf{x}_1}$, $\mathbf{a}_{\mathbf{y}_1}$ and $\mathbf{a}_{\mathbf{z}}$		
		$\mathbf{x}_1$	$\mathbf{y}_1$	$\mathbf{z}$ from P_{1_1}
1	$\mathbf{x}_1$	0	$\tau(1_1, 2_1)$	$\sigma_{1_1}\sqrt{2}$
$n_{1_1} - 1$	$\mathbf{x}_i$ for $i = 2, \ldots, n_{1_1}$	$\sigma_{1_1}\sqrt{2}$	$\tau(1_1, 2_1)$	$\sigma_{1_1}\sqrt{2}$
n_{1_2}	$\mathbf{x}_i$ for $i = n_{1_1} + 1, \ldots, n_1$	$\tau(1_1, 1_2)$	$\tau(1_2, 2_1)$	$\tau(1_1, 1_2)$
1	$\mathbf{y}_1$	$\tau(1_1, 2_1)$	0	$\tau(1_1, 2_1)$
$n_{2_1} - 1$	$\mathbf{y}_i$ for $i = 2, \ldots, n_{2_1}$	$\tau(1_1, 2_1)$	$\sigma_{2_1}\sqrt{2}$	$\tau(1_1, 2_1)$
n_{2_2}	$\mathbf{y}_i$ for $i = n_{2_1} + 1, \ldots, n_2$	$\tau(1_1, 2_2)$	$\tau(2_1, 2_2)$	$\tau(1_1, 2_2)$

Here $\tau(i_s, j_t) = \sqrt{\sigma^2_{i_s} + \sigma^2_{j_t} + \nu^2_{i_s j_t}}$ for $i, j = 1, 2$, $s = 1, 2$ and $t = 1, 2$

$\mathbf{a}_{\mathbf{x}_1}$ only in one place (see the fifth column in Table 6). The limiting value of its scaled distance from $\mathbf{x}_1$ is not 0, but $\sigma_{1_1}\sqrt{2}$. So, we have $\|\mathbf{a}_{\mathbf{x}_1} - \mathbf{a}_{\mathbf{z}}\|^2 = 2\sigma^2_{1_1}$. For the other $\mathbf{x}_i$s ($2 \le i \le n_{1_1}$) from P_{1_1}, the first n_{1_1} elements of $\mathbf{a}_{\mathbf{x}_i}$ has one 0 (with 0 at the i-th place) and $(n_{1_1} - 1)$ many $\sigma_{1_1}\sqrt{2}$s. The rest of its elements are same as $\mathbf{a}_{\mathbf{x}_1}$. So, $\mathbf{a}_{\mathbf{z}}$ differs from $\mathbf{a}_{\mathbf{x}_i}$ only at the i-th place, and $\|\mathbf{a}_{\mathbf{z}} - \mathbf{a}_{\mathbf{x}_i}\|^2 = 2\sigma^2_{1_1}$ as before. Therefore, we get

$$\|\mathbf{z}^{**} - \mathbf{x}_i^{**}\|^2 \xrightarrow{P} \|\mathbf{a}_{\mathbf{z}} - \mathbf{a}_{\mathbf{x}_i}\|^2 = 2\sigma^2_{1_1} \text{ for all } 1 \le i \le n_{1_1} \text{ as } d \to \infty. \tag{9}$$

We now compute the distances of $\mathbf{z}^{**}$ from the transformed observations of the second class. Let us begin with $\mathbf{y}_1$ belonging to the sub-class P_{2_1}. As $d \to \infty$, we get $\mathbf{y}_1^{**} \xrightarrow{P} \mathbf{a}_{\mathbf{y}_1}$ (this is shown in the fourth column in Table 6). It contains the first n_{1_1} elements equal to $\sqrt{\sigma^2_{1_1} + \sigma^2_{2_1} + \nu^2_{1_1 2_1}}$ [limiting scaled distances from observations in P_{1_1}], next n_{1_2} elements equal to $\sqrt{\sigma^2_{1_2} + \sigma^2_{2_1} + \nu^2_{1_2 2_1}}$ [distances from observations in P_{1_2}], followed by one 0 [distance from itself], $(n_{2_1} - 1)$ elements equal to $\sigma_{2_1}\sqrt{2}$ [distances from other observations in P_{2_1}] and n_{2_2} elements equal to $\sqrt{\sigma^2_{2_1} + \sigma^2_{2_2} + \nu^2_{2_1 2_2}}$ [distances from observations in P_{2_2}]. For other values of $2 \le i \le n_{1_1}$, $\mathbf{a}_{\mathbf{y}_i}$ can be obtained by interchanging the $(n_1 + 1)$-th and the $(n_1 + i)$-th elements of $\mathbf{a}_{\mathbf{y}_1}$.

So, for all $1 \le j \le n_{2_1}$,

$$\begin{aligned}
&\|\mathbf{z}^{**} - \mathbf{y}_j^{**}\|^2 \xrightarrow{P} \|\mathbf{a}_{\mathbf{z}} - \mathbf{a}_{\mathbf{y}_j}\|^2 \\
&= n_{1_1}\left[\sigma_{1_1}\sqrt{2} - \sqrt{\sigma^2_{1_1} + \sigma^2_{2_1} + \nu^2_{1_1 2_1}}\right]^2 \\
&\quad + n_{1_2}\left[\sqrt{\sigma^2_{1_1} + \sigma^2_{1_2} + \nu^2_{1_1 1_2}} - \sqrt{\sigma^2_{1_2} + \sigma^2_{2_1} + \nu^2_{2_1 1_2}}\right]^2 \\
&\quad + \left[\sigma^2_{1_1} + \sigma^2_{2_1} + \nu^2_{1_1 2_1}\right] + (n_{2_1} - 1) \\
&\quad \times \left[\sqrt{\sigma^2_{1_1} + \sigma^2_{2_1} + \nu^2_{1_1 2_1}} - \sigma_{2_1}\sqrt{2}\right]^2 \\
&\quad + n_{2_2}\left[\sqrt{\sigma^2_{1_1} + \sigma^2_{2_2} + \nu^2_{1_1 2_2}} - \sqrt{\sigma^2_{2_1} + \sigma^2_{2_2} + \nu^2_{2_1 2_2}}\right]^2
\end{aligned} \tag{10}$$

as $d \to \infty$. Now, consider this as a two class problem, where we want to classify $\mathbf{z}$ either to P_{1_1} or to P_{2_1}. Clearly, $\mathbf{z}$ will be correctly classified to P_{1_1} if $\|\mathbf{a_z} - \mathbf{a_{y_1}}\|^2$ exceeds $\|\mathbf{a_z} - \mathbf{a_{x_1}}\|^2$. From Eqs. (9) and (11), we get

$$\|\mathbf{a_z} - \mathbf{a_{y_1}}\|^2 - \|\mathbf{a_z} - \mathbf{a_{x_1}}\|^2 \geq \left[\sigma_{1_1}\sqrt{2} - \sqrt{\sigma_{1_1}^2 + \sigma_{2_1}^2 + \nu_{1_1 2_1}^2}\right]^2 + \left[\sigma_{2_1}^2 + \nu_{1_1 2_1}^2 - \sigma_{1_1}^2\right] + (n_{2_1} - 1)\left[\sqrt{\sigma_{1_1}^2 + \sigma_{2_1}^2 + \nu_{1_1 2_1}^2} - \sigma_{2_1}\sqrt{2}\right]^2. \quad (11)$$

To obtain (12), we drop some non-negative terms from the expression of $\|\mathbf{a_z} - \mathbf{a_{y_1}}\|^2$ in (11). Let us consider two separate cases.

Case 1 If $\sigma_{2_1}^2 + \nu_{1_1 2_1}^2 \geq \sigma_{1_1}^2$, from (12), we get $\|\mathbf{a_z} - \mathbf{a_{y_1}}\|^2 - \|\mathbf{a_z} - \mathbf{a_{x_1}}\|^2 \geq \left[\sigma_{1_1}\sqrt{2} - \sqrt{\sigma_{1_1}^2 + \sigma_{2_1}^2 + \nu_{1_1 2_1}^2}\right]^2 + (n_{2_1} - 1)\left[\sqrt{\sigma_{1_1}^2 + \sigma_{2_1}^2 + \nu_{1_1 2_1}^2} - \sigma_{2_1}\sqrt{2}\right]^2$.

Here, the right hand side is 0 if and only if $\sigma_{1_1}^2 = \sigma_{2_1}^2$ and $\nu_{1_1 2_1}^2 = 0$, So, under the condition of the theorem, we have $\|\mathbf{a_z} - \mathbf{a_{y_1}}\|^2 - \|\mathbf{a_z} - \mathbf{a_{x_1}}\|^2 > 0$.

Case 2 If $\sigma_{2_1}^2 + \nu_{1_1 2_1}^2 < \sigma_{1_1}^2$, it is easy to check that

$$\left|\sigma_{1_1}\sqrt{2} - \sqrt{\sigma_{1_1}^2 + \sigma_{2_1}^2 + \nu_{1_1 2_1}^2}\right| < \left|\sigma_{2_1}\sqrt{2} - \sqrt{\sigma_{1_1}^2 + \sigma_{2_1}^2 + \nu_{1_1 2_1}^2}\right|. \quad (12)$$

So, from (12), we have

$$\|\mathbf{a_z} - \mathbf{a_{y_1}}\|^2 - \|\mathbf{a_z} - \mathbf{a_{x_1}}\|^2 \geq (n_{1_1} + n_{2_1} - 1)\left[\sqrt{\sigma_{1_1}^2 + \sigma_{2_1}^2 + \nu_{1_1 2_1}^2} - \sigma_{1_1}\sqrt{2}\right]^2 + \left[\sigma_{2_1}^2 + \nu_{1_1 2_1}^2 - \sigma_{1_1}^2\right]. \quad (13)$$

If we define $u_1 = \sqrt{\sigma_{1_1}^2 + \sigma_{2_1}^2 + \nu_{1_1 2_1}^2}$, $u_2 = \sigma_{1_1}\sqrt{2}$ and $m = (n_{1_1} + n_{2_1} - 1) = n_{1_1 2_2} - 1$, the right hand side of (14) can be expressed as

$$m(u_1 - u_2)^2 + (u_1^2 - u_2^2) = (u_1 - u_2)[m(u_1 - u_2) + (u_1 + u_2)].$$

Here, we have $u_2 > u_1 > 0$. So, the right hand side of (14) is positive if

$$\{m(u_1 - u_2) + (u_1 + u_2)\} < 0 \Leftrightarrow m > \frac{(u_2 + u_1)}{(u_2 - u_1)} \Leftrightarrow u_1 < \left(1 - \frac{2}{m+1}\right)u_2. \quad (14)$$

Squaring both sides and substituting u_1, u_2 and m with their original expressions, we get

$$u_1{}^2 < \left(1 - \frac{2}{m+1}\right)^2 u_2^2 \Leftrightarrow \sigma_{2_1}^2 + \nu_{1_1 2_1}^2 + \sigma_{1_1}^2 < 2\sigma_{1_1}^2\left(1 + \frac{4}{n_{1_1 2_1}^2} - \frac{4}{n_{1_1 2_1}}\right)$$
$$\Leftrightarrow \nu_{1_1 2_1}^2 < (\sigma_{1_1}^2 - \sigma_{2_1}^2) - (n_{1_1 2_1} - 1)8\sigma_1^2/n_{1_1 2_1}^2. \quad (15)$$

We can argue for the other sub-classes in a similar way. More general cases can be handled using these arguments but with extra notations. □

Proof of Theorem 2 (b) For any $\mathbf{x} = (x_1, \ldots, x_d)^T$ and $p > 0$, the l_p norm of $\mathbf{x}$ is defined as $\|\mathbf{x}\|_p = (|x_1|^p + \cdots + |x_d|^p)^{1/p}$. Like the proof of Theorem 2(a), here also we consider the simpler case when $J = 2$ with $R_1 = R_2 = 2$, and we continue to use the same notations.

From the expressions of $\mathbf{a}_{\mathbf{x}_1}$, $\mathbf{a}_{\mathbf{y}_1}$ and $\mathbf{a}_{\mathbf{z}}$ given in Table 6, we get

$$\|\mathbf{a}_{\mathbf{z}} - \mathbf{a}_{\mathbf{x}_i}\|_p^p = (\sigma_{1_1}\sqrt{2})^p \text{ for all } 1 \le i \le n_{1_1} \text{ and} \tag{16}$$

$$\|\mathbf{a}_{\mathbf{z}} - \mathbf{a}_{\mathbf{y}_j}\|_p^p \ge n_{1_1}\left|\sigma_{1_1}\sqrt{2} - \sqrt{\sigma_{1_1}^2 + \sigma_{2_1}^2 + \nu_{1_1 2_1}^2}\right|^p + \left|\sqrt{\sigma_{1_1}^2 + \sigma_{2_1}^2 + \nu_{1_1 2_1}^2}\right|^p$$
$$+ (n_{2_1} - 1)\left|\sigma_{2_1}\sqrt{2} - \sqrt{\sigma_{1_1}^2 + \sigma_{2_1}^2 + \nu_{1_1 2_1}^2}\right|^p \text{ for all } 1 \le j \le n_{2_1}. \tag{17}$$

Note that (18) is analogous to (12) used in the proof of Theorem 2(a).

Define $w_1 = \sqrt{\sigma_{1_1}^2 + \sigma_{2_1}^2 + \nu_{1_1 2_1}^2}$, $w_2 = \sigma_{1_1}\sqrt{2} - \sqrt{\sigma_{1_1}^2 + \sigma_{2_1}^2 + \nu_{1_1 2_1}^2}$ and $w_3 = \sigma_{2_1}\sqrt{2} - \sqrt{\sigma_{1_1}^2 + \sigma_{2_1}^2 + \nu_{1_1 2_1}^2}$. Using (17) and (18), we can show that

$$\|\mathbf{a}_{\mathbf{z}} - \mathbf{a}_{\mathbf{y}_1}\|_p^p - \|\mathbf{a}_{\mathbf{z}} - \mathbf{a}_{\mathbf{x}_1}\|_p^p \ge (|w_1|^p + |w_2|^p - |w_1 + w_2|^p)$$
$$+ (n_{1_1} - 1)|w_2|^p + (n_{2_1} - 1)|w_3|^p. \tag{18}$$

Note that $|w_1|^p + |w_2|^p \ge |w_1 + w_2|^p$ for all $p \in (0, 1]$, and $w_2 = w_3 = 0$ if and only if $\sigma_{1_1}^2 = \sigma_{2_1}^2$ and $\nu_{1_1 2_1}^2 = 0$. Under the conditions assumed in the theorem, we always have $\|\mathbf{a}_{\mathbf{z}} - \mathbf{a}_{\mathbf{y}_1}\|_p^p > \|\mathbf{a}_{\mathbf{z}} - \mathbf{a}_{\mathbf{x}_1}\|_p^p$, or $\|\mathbf{a}_{\mathbf{z}} - \mathbf{a}_{\mathbf{y}_1}\|_p > \|\mathbf{a}_{\mathbf{z}} - \mathbf{a}_{\mathbf{x}_1}\|_p$. In fact, we have $\|\mathbf{a}_{\mathbf{z}} - \mathbf{a}_{\mathbf{y}_j}\|_p > \|\mathbf{a}_{\mathbf{z}} - \mathbf{a}_{\mathbf{x}_i}\|_p$ for all $1 \le i \le n_{1_1}$ and $1 \le j \le n_{2_1}$. So, the NN-TRIPD classifier based on the l_p norm with $p \in (0, 1]$ correctly classifies any $\mathbf{z}$ from P_{1_1} with probability tending to 1. We can prove this result for observations from other sub-classes as well. □

Proof of Theorem 3 (a) Under $(A1)$–$(A3)$, and the conditions on K_γ and γ as stated in the theorem, we have convergence of the pairwise distances for the transformed observations. Define $\beta_{11} = 2g(0) - 2g(\sigma_1\sqrt{2a_0})$, $\beta_{22} = 2g(0) - 2g(\sigma_2\sqrt{2a_0})$ and $\beta_{12} = 2g(0) - 2g(\sqrt{a_0(\sigma_1^2 + \sigma_2^2 + \nu_{12}^2)})$.

If $\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}$ are independent observations from class-1, and $\mathbf{y}_1, \ldots, \mathbf{y}_{n_2}$ are independent observations from class-2, using (6) we have

$$\Phi^*(\mathbf{x}_i) = \left(\frac{1}{n_1 - 1}\sum_{j=1, j\ne i}^{n_1}\|\Phi(\mathbf{x}_i) - \Phi(\mathbf{x}_j)\|, \frac{1}{n_2}\sum_{j=1}^{n_2}\|\Phi(\mathbf{x}_i) - \Phi(\mathbf{y}_j)\|\right)^T$$
$$\xrightarrow{P} \boldsymbol{\beta}_1 = (\sqrt{\beta_{11}}, \sqrt{\beta_{12}})^T \text{ for all } i = 1, \ldots, n_1 \text{ and}$$
$$\Phi^*(\mathbf{y}_j) = \left(\frac{1}{n_1}\sum_{i=1}^{n_1}\|\Phi(\mathbf{y}_j) - \Phi(\mathbf{x}_i)\|, \frac{1}{n_2 - 1}\sum_{i=1, i\ne j}^{n_2}\|\Phi(\mathbf{y}_j) - \Phi(\mathbf{y}_i)\|\right)^T$$
$$\xrightarrow{P} \boldsymbol{\beta}_2 = (\sqrt{\beta_{12}}, \sqrt{\beta_{22}})^T \text{ for all } i = 1, \ldots, n_2, \text{ as } d \to \infty. \tag{19}$$

For a new observation $\mathbf{z}$ from the first class, we have

$$\Phi^*(\mathbf{z}) = \left(\frac{1}{n_1}\sum_{i=1}^{n_1}\|\Phi(\mathbf{x}_i) - \Phi(\mathbf{z})\|, \frac{1}{n_2}\sum_{j=1}^{n_2}\|\Phi(\mathbf{y}_j) - \Phi(\mathbf{z})\|\right)^T \xrightarrow{P} \boldsymbol{\beta}_1, \tag{20}$$

as $d \to \infty$. While $\|\Phi^*(\mathbf{z}) - \Phi^*(\mathbf{x}_i)\| \xrightarrow{P} 0$ for $1 \le i \le n_1$, $\|\Phi^*(\mathbf{z}) - \Phi^*(\mathbf{y}_j)\| \xrightarrow{P} \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|$ for $1 \le j \le n_2$ as $d \to \infty$. Since g is strictly monotone, $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\| = 0$ if and only if $\sigma_1^2 = \sigma_2^2$ and $\nu_{12}^2 = 0$. Therefore, the kernelized version of NN-TRAD correctly classifies

$\mathbf{z}$ with probability tending to 1 as $d \to \infty$. A similar argument can be obtained for any observation from class-2. □

Proof of Theorem 3 (b) We prove the result for $J = 2$ and $R_1 = R_2 = 2$ as in the proof of Theorem 2(a). Instead of $\mathbf{x}_i^{**}(1 \leq i \leq n_1)$, $\mathbf{y}_j^{**}(1 \leq j \leq n_2)$ and $\mathbf{z}^{**}$ (see Eqs. (4) and (5) in Sect. 3), we now use TRIPD on $\Phi^{**}(\mathbf{x}_i)$, $\Phi^{**}(\mathbf{y}_j)$ and $\Phi^{**}(\mathbf{z})$. Here Φ^{**} is defined as follows:

$$\Phi^{**}(\mathbf{z}) = \Big(\|\Phi(\mathbf{z}) - \Phi(\mathbf{x}_1)\|, \ldots, \|\Phi(\mathbf{z}) - \Phi(\mathbf{x}_{n_1})\|,$$
$$\|\Phi(\mathbf{z}) - \Phi(\mathbf{y}_1)\|, \ldots, \|\Phi(\mathbf{z}) - \Phi(\mathbf{y}_{n_2})\|\Big)^T.$$

We continue to use the same notations as in the proof of Theorem 2(a). Since all sub-classes satisfy $(A1)$–$(A3)$, if $\mathbf{x}$ and $\mathbf{y}$ both are from $P_{i_s}(i = 1, 2,\ s = 1, 2)$, $\|\Phi(\mathbf{x}) - \Phi(\mathbf{y})\|^2 \xrightarrow{P} 2g(0) - 2g(\sigma_{i_s}\sqrt{2a_0})$ as $d \to \infty$. Again, if $\mathbf{x}$ is from P_{i_s} and $\mathbf{y}$ is from P_{j_t} $(P_{i_s} \neq P_{j_t})$ for $i, j = 1, 2, s = 1, 2, t = 1, 2$, $\|\Phi(\mathbf{x}) - \Phi(\mathbf{y})\|^2 \xrightarrow{P} 2g(0) - 2g(\sqrt{a_0(\sigma_{i_s}^2 + \sigma_{j_t}^2 + \nu_{i_s j_t}^2)})$ now, repeating the arguments used in the proof of Theorem 2(a) with $\sigma_{i_s}^2$, $\sigma_{j_t}^2$ and $\nu_{i_s j_t}^2$ replaced by $g(0) - g(\sigma_{i_s}\sqrt{2a_0})$, $g(0) - g(\sigma_{j_t}\sqrt{2a_0})$ and $g(\sigma_{i_s}\sqrt{2a_0}) + g(\sigma_{j_t}\sqrt{2a_0}) - 2g(\sqrt{a_0(\sigma_{i_s}^2 + \sigma_{j_t}^2 + \nu_{i_s j_t}^2)})$, respectively, we have the proof. □

Proof of Theorem 3 (c) Consider $\mathbf{x}$ to be from P_{i_s}, and $\mathbf{y}$ to be from $P_{j_t}(i, j = 1, 2$ and $s, t = 1, 2)$. If a future observation $\mathbf{z}$ comes from P_{i_s}, following the arguments used in Eq. (19) of Theorem 2(b), one can show that $\|\Phi^{**}(\mathbf{z}) - \Phi^{**}(\mathbf{x})\|_p^p - \|\Phi^{**}(\mathbf{z}) - \Phi^{**}(\mathbf{y})\|_p^p$ converges to a quantity greater than or equal to

$$(|\tilde{w}_1|^p + |\tilde{w}_2|^p - |\tilde{w}_1 + \tilde{w}_2|^p) + (n_{1_1} - 1)|\tilde{w}_2|^p + (n_{2_1} - 1)|\tilde{w}_3|^p,$$

where $\tilde{w}_1 = [2g(0) - 2g(\sqrt{a_0(\sigma_{i_s}^2 + \sigma_{j_t}^2 + \nu_{i_s j_t}^2)})]^{1/2}$, $\tilde{w}_2 = [2g(0) - 2g(\sigma_{i_s}\sqrt{2a_0})]^{1/2} - \tilde{w}_1$ and $\tilde{w}_3 = [2g(0) - 2g(\sigma_{j_t}\sqrt{2a_0})]^{1/2} - \tilde{w}_1$ (analogous to Eq. (19)). Note that $a_0 > 0$, and g is a positive, decreasing function. The result can be proved using the same argument as in Theorem 2(b). □

References

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high-dimensional space. *Database theory—ICDT*, pp. 420–434.

Aho, A. V., Hopcroft, J. D., & Ullman, J. E. (1974). *Design and analysis of computer algorithms*. London: Addison-Wesley.

Andrews, D. W. K. (1988). Laws of large numbers for dependent non-identically distributed random variables. *Econometric Theory*, *4*, 458–467.

Arora, R., Gupta, M., Kapila, A., & Fazel, M. (2013). Similarity based clustering by left stochastic matrix factorization. *Journal of Machine Learning Research*, *14*, 1715–1746.

Biswas, M., & Ghosh, A. K. (2014). A nonparametric two-sample test applicable to high-dimensional data. *Journal of Multivariate Analysis*, *123*, 160–171.

Biswas, M., Mukhopadhyay, M., & Ghosh, A. K. (2014). An exact distribution-free two-sample run test applicable to high dimensional data. *Biometrika*, *101*, 913–926.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Cazzanti, L., Gupta, M., & Koppal, A. (2008). Generative models for similarity-based classification. *Pattern Recognition*, *41*, 2289–2297.

Chen, Y., Garcia, E. K., Gupta, M., Cazzanti, L., & Rahimi, A. (2009). Similarity-based classification: Concepts and algorithms. *Journal of Machine Learning Research*, *10*, 747–776.

Chen, Y.-B., & Hall, P. (2009). Robust nearest neighbor methods for classifying high-dimensional data. *Annals of Statistics*, *37*, 3186–3203.

Cost, S., & Salzberg, S. (1993). A weighted nearest neighbor algorithm for learning with symbolic features. *Machine Learning*, *10*, 57–78.

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions Information Theory*, *13*, 21–27.

Deegalla, S., & Bostrom, H. (2006). Reducing high-dimensional data by principal component analysis vs. random projection for nearest neighbor classification. In *5th International conference on machine learning and applications*, pp. 245–250.

de Jong, R. M. (1995). Laws of large numbers for dependent heterogeneous processes. *Econometric Theory*, *11*, 347–358.

Ding, C., He, X., & Simon, H. D. (2005). On the equivalence of nonnegative matrix factorization and spectral clustering. *Proceedings SIAM conference data mining*, pp. 606–610.

Duda, R., Hart, P., & Stork, D. G. (2000). *Pattern Classification*. New York: Wiley.

Feller, W. (1968). *An introduction to probability theory and its applications*. New York: Wiley.

Fern, X. Z., & Brodley, C. E. (2003). Random projections for high-dimensional data clustering. *Proceedings of 20th international conference machine learning*, pp. 186–193.

Fix, E., & Hodges, J. L, Jr. (1989). Discriminatory analysis: Nonparametric discrimination: Consistency properties. *International Statistical Review*, *57*, 238–247.

Francois, D., Wertz, V., & Verleysen, M. (2007). The concentration of fractional distances. *IEEE Transaction on Knowledge and Data Engineering*, *17*, 873–886.

Genton, M. G. (2001). Classes of kernels for machine learning: A statistics perspective. *Journal of Machine Learning Research*, *2*, 299–312.

Ghosh, A. K., & Hall, P. (2008). On error-rate estimation in nonparametric classification. *Statistica Sinica*, *18*, 1081–1100.

Goldberger, J., Roweis, S., Hinton, G., & Salakhutdinov, R. (2005). Neighbourhood component analysis. *Advance Neural Information Processing Systems 18 (NIPS)*, *17*, 513–520.

Goldfarb, L. (1985). A new approach to pattern recognition. *Program Pattern Recognition*, *2*, 241–402.

Graepel, T., Herbrich, R., & Obermayer, K. (1999). Classification on pairwise proximity data. *Advance Neural Information Processing System*, *11*, 438–444.

Hall, P., Marron, J., & Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society Series B*, *67*, 427–444.

Hall, P., Park, B. U., & Samworth, R. (2008). Choice of neighbor order in nearest neighbor classification. *Annals of Statistics*, *36*, 2135–2152.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference and prediction*. New York: Springer.

Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). Kernel methods in machine learning. *Annals of Statistics*, *36*, 1171–1220.

Jung, S., & Marron, J. S. (2009). PCA consistency in high dimension, low sample size context. *Annals of Statistics*, *37*, 4104–4130.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*, 18–22.

Pekalska, E., Paclíc, P., & Duin, R. P. W. (2001). A generalized kernel approach to dissimilarity-based classification. *Journal of Machince Learning Research*, *2*, 175–211.

Radovanovic, M., Nanopoulos, A., & Ivanovic, M. (2010). Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machince Learning Research*, *11*, 2487–2531.

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, *39*, 1–13.

Tomasev, N., Radovanovic, M., Mladenic, D., & Ivanovic, M. (2011). Hubness-based fuzzy measures for high-dimensional k-nearest neighbor classification. *Machine Learning Data Mining in Pattern Recogniton, LNCS*, *6871*, 16–30.

Vapnik, V. (1998). *Statistical Learning Theory*. New York: Wiley.

Weinberger, K., Blitzer, J., & Saul, L. (2006). Distance metric learning for large margin nearest neighbor classification. *Advances in Neural Information Processing Systems*, *18*, 1473–1480.

Young, G., & Hoseholder, A. S. (1938). Discussion of a set of points in terms of their mutual distances. *Psychometrika*, *3*, 19–22.