Neural Processing Letters (2021) 53:2277–2304
https://doi.org/10.1007/s11063-021-10472-3

NRIC: A Noise Removal Approach for Nonlinear Isomap Method

Mahwish Yousaf[1] · Muhammad Saadat Shakoor Khan[2] · Tanzeel U. Rehman[1] · Shamsher Ullah[3] · Li Jing[1]

Accepted: 22 February 2021 / Published online: 8 March 2021

Abstract
Nonlinear manifold learning is a popular dimension reduction method that determines large and high dimensional datasets' structures. However, these nonlinear manifold learning methods, including isomap and locally linear embedding, are sensitive to noise. In this paper, we focus on the noisy nonlinear manifold learning method, such as Isomap. The main problem of the Isomap is sensitivity to noise. Our proposed new method noise removal isomap with a classification (NRIC), is based on the local tangent space alignment (LTSA) algorithm with classification techniques to remove noises and optimize neighborhood structure Isomap. The primary purpose of the NRIC is to increase efficiency, reduce noise, and improve the performance of the graph. Experiments on the real-world datasets have shown that the NRIC method outperforms efficiently and maintains an accurate low dimensional representation of the noisy nonlinear manifold learning data. The results show that LTSA with classification techniques provides high accuracy, mean-precision, mean-recall, and areas under the (ROC) curve (AUC) of the high dimensional datasets and optimizes the graphs.

✉ Li Jing
lj@ustc.edu.cn

Mahwish Yousaf
mahwish@mail.ustc.edu.cn

Muhammad Saadat Shakoor Khan
saadat@mail.ustc.edu.cn

Tanzeel U. Rehman
tanzeel@mail.ustc.edu.cn

Shamsher Ullah
shamsherullah@nwpu.edu.cn

1 School of Computer Science and Technology, University of Science and Technology of China, Hefei 230026, People's Republic of China

2 Key Laboratory of Strongly-Coupled Quantum Matter Physics Chinese Academy of Sciences Department of Physics and Hefei National Laboratory for Physical Sciences at the Microscale, University of Science and Technology of China, Hefei 230026, People's Republic of China

3 School of Software,Taicang Campus, Northwestern Polytechnical University, Jiangsu, People's Republic of China

Keywords Noise removal isomap with classification · Manifold learning · Isomap · Classification · Local tangent space alignment

1 Introduction

Nonlinear Manifold learning is an efficient approach for dimension reduction. Various methods and algorithms have been offered for analyzing the basic structure of the high and large-scale dimensional data, which attracted much more attention in the machine learning area [55]. The basic idea of manifold learning is to transform the high dimensional data into low dimensional space and retain the most important information [17]. In recent years has attracted much more attention to the great importance of studying manifold learning for nonlinear dimension reduction [51,58].

In 2000, the Isometric Mapping algorithm became a hot research topic in nonlinear manifold learning and information science [54]. Isomap is a nonlinear manifold learning algorithm that is widely used for nonlinear dimension reduction [14]. The basic idea of Isomap is a variant of the Multidimensional Scaling (MDS) metric, which preserves the global intrinsic structure of the data points. It maps the high dimensional data into low dimensional space [45]. Heeyoul and Seungjin [2,6] proposed the kernel Isomap algorithm to solve the noises and the outlier problem in topological stability. The limitation of the proposed schemes [2,6] is it destroyed the original Isomap algorithm. This approach preserves topological stability when dealing with outliers and noises. They also suggested the robust kernel Isomap method for topological stability, noises, and outliers problems [8]. They reduced the effect of outliers based on the topological structure with network flow help [7,8]. H. Chang and DY proposed [5] the robust LLE method for noises and outliers. This method can improve the robustness of LLE, eliminating the outliers and noises in data points. The main drawback of this method outliers and noises are still controlled in data points, and robustness is still reduced to some data points. Kouropteva et al. [23,24] and Shao et al. [42] proposed the selection of the optimal parameter values method for LLE, and Isomap and Saxena et al. [41] proposed the integrated approach for Isomap and LLE.

In addition, Bo Li et al. [28] proposed the expanded Isomap approach for improving the robust LLE process, and the robustness of the original Isomap was reduced. This method uses the weighted Principal Component Analysis (PCA) [5] to measure the noises and outliers in data points. So every weight point in the data set will be allocated through local robust PCA. To detect weighted noises and outliers, R. McGill et al. use the box statistic method [30]. After de-noising, this method will easily retain the topological structure [28].

In our motivation, we have focused on studying the nonlinear Isomap noise problem. The Isomap algorithm is also noise sensitive. Isomap algorithm is not suitable for real-world datasets, as the datasets are noiseless. We propose a novel approach called Noise Removal Isomap with Classification (NRIC) method for overcoming the Isomap noise problem.We have used the Local Tangent Space Alignment (LTSA) algorithm with classification techniques for the Isomap noise problem. To effectively eliminate the noises in data points, we used the concept of LTSA as a nonlinear manifold learning technique. We have used different classification techniques such as Support Vector Machine (SVM) [49,50], K Nearest Neighbor (KNN) [16,29], Naïve Bayes (NB) [20], and Random Forest (RF) [25]. Isomap algorithm can't easily map high-dimensional data to low-dimensional space by using classification techniques and real-world datasets because it's very noisy.

Our proposed method results show that the LTSA algorithm with classification techniques can significantly improve the original Isomap in a noisy environment. Also, our proposed method produces accurate results for large and high-dimensional datasets while reducing data point noise. In Sect. 3, we explain in detail the techniques and the algorithms.

Contributions In summary, the contribution of our proposed method is given below:

1. We propose the NRIC method for the Isomap noise problem. Our proposed method used an LTSA algorithm with well-known classification techniques. Our proposed NRIC method can easily embed the high dimensional data space into low dimensional space and optimize the neighborhood graph.
2. We conduct extensive experiments to analyze our NRIC method on five datasets empirically. We calculate the accuracy, mean-precision, mean-recall, and Area under the ROC (Receiving Operating Characteristics) Curve (AUC) for our proposed method and provide the effective noise removal results.
3. We improve the Isomap noise problem's performance by using the different neighborhood value of K. The experiment section shows the effectiveness of our proposed NRIC method according to K values.

The paper is organized as follows. In Sect. 2, we will give brief details of the Isomap and Machine learning classifier. We will provide details of the proposed method and LTSA, and classification techniques in Sect. 3. Section 4 describes the experimental results on five large scale datasets. Finally, the paper is concluded in Sect. 5.

2 Related Work

Classical Isomap is viewed as a variant of Multidimensional Scaling (MDS) metric to model nonlinear data using its geodesic distance. The primary purpose of Isomap preserves the geometry of data and gets the geodesic distance between all pairs of data points. The geodesic distance is divided into two parts, such as neighborhood data points and faraway data points. In neighborhood points, the Euclidean distances between neighboring points are provided approximated geodesic distance by input-space. In faraway points, the geodesic distances are calculated the approximated by the shortest paths in neighborhood points [28,37,38]. The main three steps of Isomap are given below:

Step-1: Build neighborhood graph G Firstly, build the K nearest neighbor (KNN) graph G of manifold learning based on the Euclidean distance d between two data points in the input space X_i and X_j, i.e., $d{=}X_i$, $X_j = \|X_i{-}X_j\|$ [28].

Step-2: Calculate the shortest distance When builds the neighborhood graph G, then calculate the geodesic distance matrix between sub-neighborhood faraway data points and computes the shortest path distance between any two data points X_i and X_j is the graph G by Dijkstra and Floyd's algorithm [18].

Steps-3: Build a d-dimensional embedding graph Isomap uses the MDS algorithm to compute the low d-dimensional embedding of the data points and make the geodesic distance dense matrix [43].

2.1 Machine Learning Classifier

Machine Learning (ML), which is used in different research fields, including Artificial intelligence, data classification, and Statistic concerned with the automatic acquisition of knowledge datasets. These techniques are capable of improving the performance of datasets

from experience [34]. The famous research field of ML is data classification. Data classification is provided various algorithms such as Support Vector Machine (SVM), K Nearest Neighbor (KNN), Random Forest (RF), Naïve Bayes (NB), Artificial Neural Network (ANN), Classification and Regression Tree (CART), Decision Tree, etc. [33]. The primary process of classification is to predict the label's data points from given datasets. The label data points are sometimes called classes, targets, and categories. Classification Predictive Modeling (CPM) is mapped the input data points through a mapping function and predicts the possible output data points. A detailed description of classification algorithms is given in Sect. 3.

3 Noise Removal Isomap with Classification (NRIC)

In this section, we propose a novel approach, which is called the Noise Removal Isomap with Classification (NRIC) method for the Isomap noise problem. The main idea of our NRIC method is to eliminate the noise quickly, map the high dimensional data into low dimensional space, and then easily optimize the neighborhood graph. We have using the LTSA algorithm with classification techniques in our proposed method. Our NRIC method provided high accuracy rather than Isomap and reduced the noise from data points. We have used classification techniques, such as SVM, KNN, NB, and RF, with different K values. The algorithm 1 of our NRIC method is given below:

Algorithm 1: Noise Removal Isomap with Classification (NRIC)

Input: Dataset ***X***, Noise Removal Output Dataset (NROD) ***Y***
Output: Graph of Accuracy ***A***, Mean-Precision ***P***, Mean-Recall ***R***, and Area under the (ROC) Curve ***(AUC)***

1: Perform the **LTSA** method on the input dataset **X** for noise removal and see Algorithm **(2)**
2: Noise Removal Output Dataset (NROD) **Y**
3: **Then**
4: We have performed Classification techniques on NROD **Y**, including SVM, KNN, NB, and RF. See the Algorithm **(3) (4) (5) (6)**
5: These classification techniques split the NROD **Y**
6: **Then**
7: Calculate the Accuracy **A**, Mean-Precision **P**, Mean-Recall **R**, and AUC
8: At the end build the Graphs of Accuracy, Mean-Precision, Mean-Recall, and ROC Curve

3.1 Local Tangent Space Alignment (LTSA) Algorithm

In 2004, Zhang and Zha introduced the nonlinear local tangent space alignment method for embedding [59]. This method can easily embed the high dimensional data into low dimensional space. LTSA can be used for noise problems and very efficiently eliminates noise. LTSA's central concept is LLE variants and employed the same geometric manifolds as LLE. LTSA uses a distinct method to the embedded manifold space compared with LLE. In LLE, every point of the datasets is locally linearly embedded into the manifold's linear plot then constructed the low dimensional datasets. So that preserved the locally linear relationships

of the original datasets. Moreover, LTSA has built a locally linear patch by using the PCA method on the neighbors, and then the patch can be evaluated as an approximation of local tangent space at the point [52]. The algorithm 2 of LTSA is given below:

Algorithm 2: Local Tangent Space Alignment (LTSA) Algorithm

Input: High dimensional dataset $\boldsymbol{X=\{X_i,...,X_k\}}$, $\boldsymbol{K}$ and ϵ neighborhood, $\boldsymbol{d}$ is the smallest eigenvectors, and I_i is the representation of $\boldsymbol{K}$ neighborhood.
Output: Low dimensional embedding dataset $\boldsymbol{Y=\{Y_i,....,\ Y_k\}}$
1: Search **K** and ϵ neighborhood
if *(**K** and ϵ neighborhood* $>$ ***d**)* **then**
  ϵ neighborhood depends on the dimensions of the data and **K** neighborhood has used the neighborhood of LTSA;
2: Computation of Local Coordinates5
  Centralized the data and Calculated the mean $\mathbf{X=\{}X_i\mathbf{-X...}X_k\mathbf{-X\}}$.
  Search the local coordinates of X_i by the PCA method, and local coordinates correspond to the 1st smallest eigenvectors **d**
3: Alignment of Local Coordinates
  Create the alignment matrix **A** and the initial value of matrix **A** is zero
  $\mathbf{A}(I_i, I_i) \leftarrow \mathbf{A}\ (I_i, I_i) + \mathbf{I}\text{-}G_i\ G_i^T$**, i=1....N**, and **I** is the **N** $\times$ **N** identity matrix
4: Calculating the Smallest Eigen decomposition vector
  The Eigen matrix **A** is the corresponding of **2nd - (d+1)** the smallest eigenvector and the global coordinates of $\mathbf{Y = [}u_1,...,u_{d+1}\mathbf{]\ T}$

3.2 Support Vector Machine (SVM) Classifier

SVM has attracted much more attention and used very actively in several research applications such as regression, learning classification, and ranking function. The basic idea of SVM is dependent on the Structural Risk Minimization (SRM) principle and statistical learning theory and identifying the position of decision space, also called hyperplane, that generates the optimal partition of classes [4,9,13,35]. SVM uses an isolating hyperplane to create an SVM event model classifier. The main issues of SVM cannot be isolated directly in the information space. This method provides a probability function to identify an answer by performing principle information space improvement in high dimensional space, where a perfect portioning hyperplane can be found [39]. In the experiment, we have used the linear kernel model for the SVM classifier. The algorithm 3 of SVM [47] is given below:

Algorithm 3: Support Vector Machine (SVM) Classifier

Input: Dataset ***Y***, ***A***: number of samples, ***B*** is labeled where $B_i \in$ ***(1,..., N)***, vector ***V=1***, and ***svm*** is the kernel classifier model
Output: Validation (Accuracy ***A***, Mean-Precision ***P***, Mean-Recall ***R***), and Area under the (ROC) Curve ***(AUC)***
1: **Train data** ← Split data (Dataset **Y**, size=**0.7**)
2: **Test data** ← Split data (Dataset **Y**, size=**1.0**)
3: **for** *(**n** in **(1,..., N))* **do**
    Built the vector **V**;
    if *(B_i=**K**)* **then**
        V=0;
    else
        Apply model **svm** to **A**, and vector **V** obtain a list of SVM classifiers;
4: Test the model using **"Test data"**
5: Calculate the Accuracy **A**, Mean-Precision **P**, Mean-Recall **R**, and AUC

3.3 K- Nearest Neighbors (KNN) Classifier

The KNN classifier is the simplest classification method and used in machine learning and data mining. It is beneficial and easy to implement. It does not require a fitting model for classifying various types of datasets and provide the best performance of the multiple types of datasets [21]. In contrast, the best performance of KNN depends on the distance metric for calculating the distance between data points of Euclidean. The KNN data points often use Euclidean distance for similarity [1]. The KNN makes the training samples by itself according to the laws of classification. The KNN algorithm is classifying the objects based on the nearest training samples in the attributes. KNN method is a kind of lazy learning and instance-based learning because the KNN function is locally approximated, and all execution of KNN is delayed until classification [36]. The KNN classifier can easily find the closest samples from training datasets. In the results section, K parameters are used as an optimal value [56,57]. The algorithm 4 of KNN [44] is given below:

Algorithm 4: K- Nearest Neighbors (KNN) Classifier

Input: Dataset ***Y*** is the train data, ***A*** is class labels of Y, ***x*** is an unknown sample, ***d*** is the Euclidean distance, and ***K*** is the knearest classifier model
Output: Validation (Accuracy ***A***, Mean-Precision ***P***, Mean-Recall ***R***), and Area under the (ROC) Curve **(AUC)**
1: **for** *(**I = 1** to **n**)* **do**
    Calculate distance **d**(Y_i**, x**);
2: Apply model **K** on a distance of knearest classifiers
3: **K**= Compute Knearestneighbors **d**(Y_i**, x**)
4: **return** $agrmax[A_i where$ **i =K**]
5: Calculate the Score
6: Accuracy **A**, Mean-Precision **P**, Mean-Recall **R**, and AUC

3.4 Naïve Bayes (NB) Classifiers

NB Classifiers are easy probability classifiers based on the Bayes Theorem [27], and the NB classifier is mostly used when input data dimensionality is high. This classifier is efficient for computing the available output data based on the input data. It adds new available raw input data at runtime and has an efficient probabilistic classifier [35]. Different types of NB classifier is accessible by the assumptions on the distribution of features; these are called event models of NB classifier [22], including Bernoulli or multinomial distributions, Gaussian distributions [32], and discrete features. We have used the Gaussian event model in our proposed method for calculating the accuracy. In our proposed method, we used the Gaussian event model to calculate accuracy. It's also slightly quicker and more efficient than SVM [53]. The NB [40] algorithm 5 is given below:

Algorithm 5: Naive Bayes (NB) Classifier

Input: Dataset ***Y*** is the training data, ***A*** is test data, where **A=**{a_1,a_2,....,a_n}, and ***GNB*** is the GussianNB classifier model
Output: Validation (Accuracy ***A***, Mean-Precision ***P***, Mean-Recall ***R***), and Area under the (ROC) Curve ***(AUC)***
1: Read the **Y** training data
2: Compute the standard deviation and mean of the prediction class
3: **for** ***N*** *Times* **do**
  Compute the probability of A_i using the **GNB** classifier model for each class. Until the probability of all prediction class {a_1,a_2,.....,a_n} has been computed.;
4: Calculate the Score
5: Accuracy **A**, Mean-Precision **P**, Mean-Recall **R**, and AUC

3.5 Random Forest (RF) Classifier

T. Kam Ho [15] was introduced RF in 1995, which uses the tree as parallel. RF is a collaborative learning classification algorithm (ensemble) combining the same and different types of more than one algorithm to classify the object. RF classifier is a randomly selected subset of the training dataset in the set of decision trees. It is a fast method to train the dataset rather than other techniques such as deep learning, although less slow to predict once trained datasets [3,25]. The algorithm 6 of RF [46] is given below:

Algorithm 6: Random Forest (RF) Classifier

Input: Dataset ***Y*** is the training data, ***A*** is input instance to be used for each tree, and ***RF*** is the Randomforest classifier model
Output: Validation (Accuracy ***A***, Mean-Precision ***P***, Mean-Recall ***R***), and Area under the (ROC) Curve ***(AUC)***
1: **for** ***(i=1 to B)*** **do**
  Y_i=BootstrapSample (**Y**);
  Z_i= Create **RF** Randomforest classifier model (Y_i, **A**);
  E=E ∪ { Z_i } ;
  Next **i**;
2: **return** E
3: Calculate the Score
4: Accuracy **A**, Mean-Precision **P**, Mean-Recall **R**, and AUC

Table 1 Experimental datasets

Datasets-name	Total datasets	Actual-dimensions	Classes	Use-dimensions
Iris	150	4	3	4
Wine	178	13	3	13
LFW people	13233	5828	5479	100
Breast cancer	569	30	2	30
Digits	1797	64	10	50

4 Results and Discussion

This section analyzes the effectiveness and efficiency of the NRIC, general experiments on high dimensional and large scale datasets. We have analyzed the proposed algorithms' performance with an LTSA algorithm and classification techniques such as SVM, KNN, NB, and RF. Moreover, we have also compared our proposed NRIC method with different neighborhood K values.

4.1 Datasets

The experiments were organized on a large scale and high dimensional datasets such as Iris [11], Wine [12], Labeled Faces in the Wild (LFW) people [19], Breast Cancer [31], and Digits [26]. The detailed information on the datasets is listed in Table 1.

4.2 Evaluation Methods

Various metrics are used to assess the proposed approach and its effectiveness. The classification algorithm provides multiple criteria for evaluating the resulting datasets, such as precision, accuracy, recall, and area under the ROC curve (AUC). Here, TP is the number of correct positive predictions. TN is the number of correct negative predictions, FP is the number of false-positive predictions, and FN is the number of false-negative predictions. The significance of these four parameters is based on the classification application. The division of correct predictions overall predictions is called accuracy. The ratio of correctly positive prediction overall positives predictions is called precision. The ratio of correctly negative predictions overall negative predictions is called recall [25]. The ROC curve area is well known for the classification evaluation method used in machine learning and data mining areas. The ROC curve takes True Positive Rate (TPR) and False Positive Rate (FPR) for a given classification algorithm [48]. ROC graphs are beneficial to organize a better visualize and classifier performance. Also, ROC graphs are present a better relationship between TPR and FPR. AUC can reasonably measure the model's prediction quality and display the classifier performance in a single value. If the AUC value is greater, the classifier performance is better, and otherwise, the performance is not well [10]. The formulas of accuracy Eq. (1), mean-precision Eq. (2), mean-recall Eq. (3), and area under the ROC (Receiving Operating Characteristics) curve (AUC) Eqs. (4) and (5) are defined below:

$$Accuracy = \left[\frac{(TP + TN)}{(TP + FP + FN + TN)} \right] \tag{1}$$

Table 2 Accuracy of iris dataset with SVM, KNN, NB, and RF

K	LTSA-SVM	ISO-SVM	LTSA-KNN	ISO-KNN	LTSA-NB	ISO-NB	LTSA-RF	ISO-RF
10	1.00	0.92	1.00	0.93	1.00	0.91	1.00	0.94
15	1.00	0.90	0.97	0.93	0.97	0.90	0.97	0.91
20	1.00	0.89	0.97	0.93	0.90	0.86	0.93	0.87
25	1.00	0.87	0.90	0.83	0.83	0.86	0.93	0.87
30	0.97	0.85	0.83	0.80	0.83	0.79	0.90	0.88
40	1.00	0.85	0.93	0.79	0.97	0.90	0.93	0.89

$$Mean - Precision = \left[\frac{(TP)}{(TP + FP)}\right] \tag{2}$$

$$Mean - Recall = \left[\frac{(TP)}{(TP + FN)}\right] \tag{3}$$

$$TPR = \left[\frac{(TP)}{(TP + FN)}\right] \tag{4}$$

$$FPR = \left[\frac{(FP)}{(FP + TN)}\right] \tag{5}$$

4.3 Accuracy

Tables 2, 3, 4, 5, 6 shows the calculated accuracy of Iris, Wine, LFW, Breast Cancer, and Digits Datasets from the Eq. (1). We have represented the calculated accuracy of five different high dimensional datasets in Figs. 1, 2, 3, 4, 5. Our proposed method graph consistently achieves high accuracy with different neighborhood K values. We used the different K values such as 10, 15, 20, 25, 30, and 40 for calculating the accuracy of the five datasets. We have compared LTSA with four classification techniques such as SVM, KNN, NB, and RF for five datasets. These classification techniques are work very well with LTSA and provide efficient and effective results for the Isomap noise problem. Therefore, the Isomap method cannot work well with classification techniques rather than our proposed method.

For the Iris dataset, we have achieved 100% accuracy for different values of K, such as 10, 15, 20, 25, 30, and 40. These classification techniques also have achieved 100% accuracy for different values of K. According to a comparison of classification techniques; the LTSA algorithm is provided 100% accuracy with SVM, KNN, NB, and RF in Fig. 1. In Table 2, LTSA-SVM is provided the 100% accuracy for values of K (10, 20, 25, and 40). LTSA-KNN is provided 100% accuracy for values of K=10. LTSA-NB and LTSA-RF are delivered with 100% accuracy only on K=10. Therefore, we have compared the Isomap with classification techniques. Still, ISO-SVM has only achieved 92% accuracy for the iris dataset on K=10, ISO-KNN has achieved 93%, ISO-NB has achieved 91%, and ISO-RF has achieved 94%. In Fig. 1, LTSA is performed well with SVM, KNN, NB, RF rather than Isomap with classification techniques.

For the Wine dataset, we have achieved 100% accuracy only on K=15, 30, and 40 for the LTSA-SVM. In Table 3, LTSA-KNN has gained 100% accuracy on K=15 and 20, LTSA-NB has increased 100% accuracy on K=25, and LTSA-RF has achieved 98% accuracy on K=15. Therefore, we have compared the Isomap with classification techniques. Still, ISO-SVM has only achieved 90% accuracy on K=10 and 20, ISO-KNN has achieved 92% on K=15,

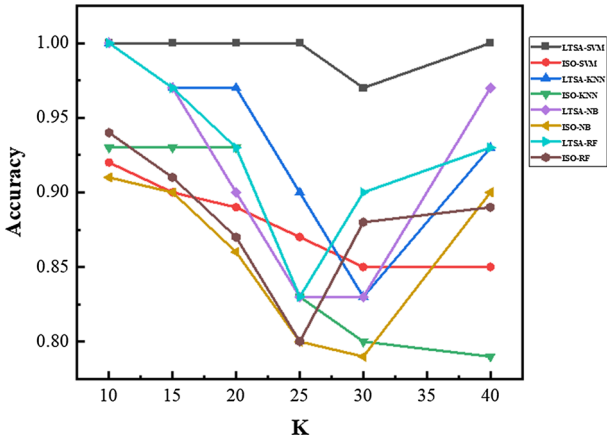

Fig. 1 Accuracy of Iris dataset with NRIC and isomap methods

Table 3 Accuracy of wine dataset with SVM, KNN, NB, and RF

K	LTSA-SVM	ISO-SVM	LTSA-KNN	ISO-KNN	LTSA-NB	ISO-NB	LTSA-RF	ISO-RF
10	0.94	0.90	0.86	0.82	0.97	0.93	0.94	0.90
15	1.00	0.90	1.00	0.92	0.98	0.91	0.98	0.86
20	0.94	0.89	1.00	0.91	0.94	0.90	0.92	0.89
25	0.97	0.87	0.97	0.89	1.00	0.89	0.94	0.82
30	1.00	0.85	0.92	0.82	0.97	0.87	0.83	0.79
40	1.00	0.82	0.97	0.83	0.94	0.84	0.89	0.77

Table 4 Accuracy of LFW people dataset with SVM, KNN, NB, and RF

K	LTSA-SVM	ISO-SVM	LTSA-KNN	ISO-KNN	LTSA-NB	ISO-NB	LTSA-RF	ISO-RF
10	0.90	0.87	0.81	0.80	0.62	0.77	0.68	0.63
15	0.90	0.82	0.85	0.79	0.75	0.72	0.68	0.65
20	0.93	0.88	0.81	0.78	0.75	0.70	0.68	0.65
25	0.92	0.87	0.84	0.76	0.70	0.65	0.68	0.65
30	0.92	0.85	0.88	0.75	0.79	0.62	0.68	0.64
40	0.92	0.88	0.85	0.74	0.81	0.75	0.68	0.62

ISO-NB has achieved 93%, and ISO-RF has achieved 90% both on k=10. In Fig. 2, we have shown the comparison of NRIC and Isomap method with classification techniques. However, LTSA works well with SVM rather than KNN, NB, and RF, and rather than Isomap with classification techniques in Fig. 2.

For LFW people dataset having 13233 data points and 5828 dimensions. We have used 100 dimensions for calculating the accuracy of the proposed algorithm. In Table 4, the LFW dataset is vast. Therefore, the accuracy of the LFW dataset is reduced very severely. We achieved 93% accuracy only on K=20 for the LTSA-SVM, 88% accuracy on K=30 for LTSA-KNN, and LTSA-NB has achieved 81% accuracy K=40. LTSA-RF has reached 68% accuracy for all K values. Therefore, ISO-SVM has only achieved 88% accuracy on K (20, 40). ISO-KNN has achieved 80% ISO-NB has reached 77% both on the value of K=10, and

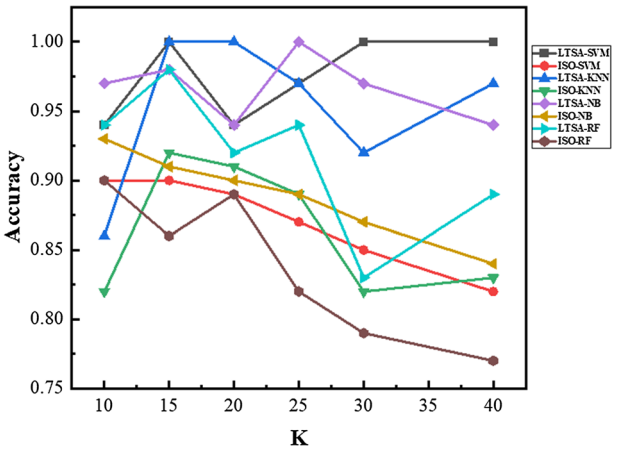

Fig. 2 Accuracy of wine dataset with NRIC and isomap methods

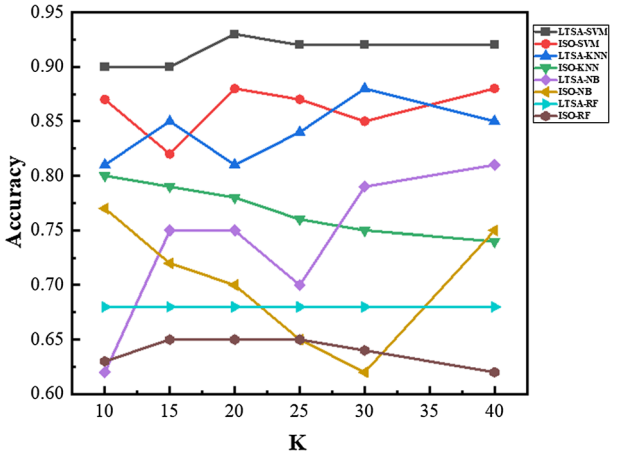

Fig. 3 Accuracy of LFW people dataset with NRIC and isomap methods

ISO-RF has reached 65% accuracy. In Fig. 3, we have shown the comparison between NRIC and Isomap method with classification techniques. However, LTSA is performed well with SVM, KNN, and NB rather than RF and Isomap.

For the Breast Cancer dataset, we have achieved 98% accuracy on K=40 for the LTSA-SVM, 94% accuracy for LTSA-KNN on K=40, 94% accuracy for LTSA-NB, and LTSA-RF on K=30, as shown in Table 5. Therefore, 94% accuracy for ISO-SVM on K (20, 40), ISO-KNN has achieved 90%, ISO-NB has gained 88% both on the value of K=10 ISO-RF has reached 90% accuracy on K=30. In Fig. 4, we have shown the comparison between NRIC and Isomap with classification techniques. However, LTSA works well with SVM, KNN, NB, and RF rather than Isomap.

For the Digits datasets having 1797 data points and 64 dimensions. We have used 50 dimensions for calculating the accuracy of the proposed method. Table 6 achieved 97% accuracy for LTSA-SVM on K=30, 95% LTSA-KNN, 94% LTSA-NB, and 84% LTSA-RF on the same value of K=10. Therefore, 95% accuracy for ISO-SVM on K (30), ISO-KNN has

Table 5 Accuracy of breast cancer dataset with SVM, KNN, NB, and RF

K	LTSA-SVM	ISO-SVM	LTSA-KNN	ISO-KNN	LTSA-NB	ISO-NB	LTSA-RF	ISO-RF
10	0.96	0.90	0.92	0.90	0.90	0.88	0.90	0.87
15	0.95	0.93	0.91	0.88	0.86	0.85	0.82	0.80
20	0.97	0.94	0.90	0.84	0.89	0.86	0.88	0.86
25	0.95	0.91	0.90	0.75	0.93	0.87	0.80	0.80
30	0.95	0.91	0.91	0.77	0.94	0.89	0.94	0.90
40	0.98	0.94	0.94	0.73	0.92	0.86	0.82	0.80

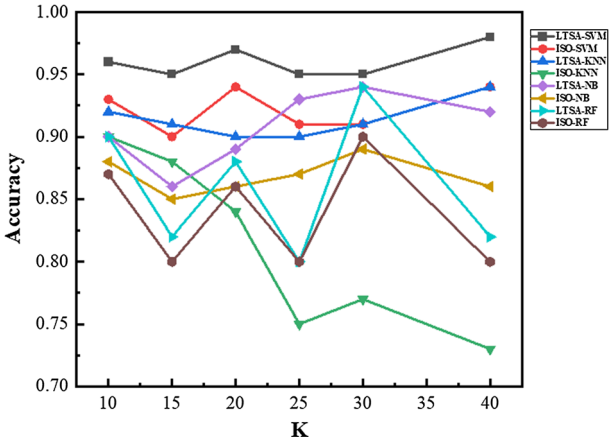

Fig. 4 Accuracy of breast cancer dataset with NRIC and isomap methods

Table 6 Accuracy of digits dataset with SVM, KNN, NB, and RF

K	LTSA-SVM	ISO-SVM	LTSA-KNN	ISO-KNN	LTSA-NB	ISO-NB	LTSA-RF	ISO-RF
10	0.95	0.93	0.95	0.93	0.94	0.92	0.84	0.80
15	0.96	0.94	0.94	0.91	0.92	0.90	0.79	0.75
20	0.96	0.94	0.95	0.90	0.91	0.90	0.76	0.75
25	0.96	0.94	0.93	0.87	0.89	0.87	0.68	0.63
30	0.97	0.95	0.93	0.88	0.88	0.86	0.67	0.62
40	0.95	0.94	0.92	0.81	0.91	0.88	0.75	0.69

achieved 93%, ISO-NB has reached 92%, and ISO-RF has reached 80% accuracy on K=10. In Fig. 5, we have shown the comparison between NRIC and Isomap with classification techniques. LTSA does work well with SVM, KNN, NB, and RF instead of Isomap.

In addition, the NRIC method's overall performance consistently achieves high accuracy for the five high dimensional datasets. Our NRIC method is much faster than Isomap and very effectively reduced the noise in the dataset. Then the NRIC method easily maps the high dimensional data into a low dimensional manifold. Sometimes, accuracy is high and down according to different K's values because sometimes the KNN are close to each other and discover quickly. Otherwise, the KNN is far away from each other and recognize hardly them.

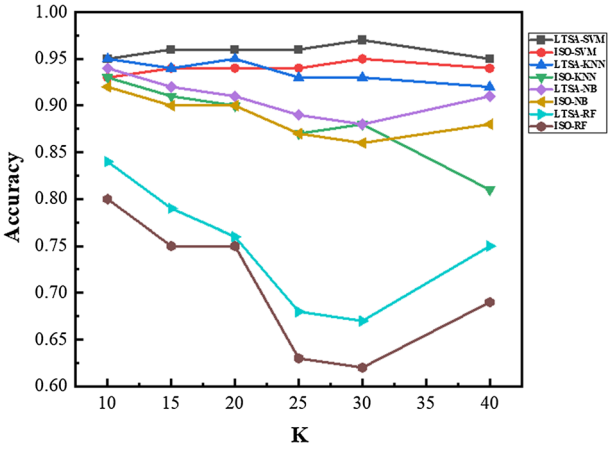

Fig. 5 Accuracy of digits dataset with NRIC and isomap methods

Table 7 Mean-precision of iris dataset with SVM, KNN, NB, and RF

K	LTSA-SVM	ISO-SVM	LTSA-KNN	ISO-KNN	LTSA-NB	ISO-NB	LTSA-RF	ISO-RF
10	1.00	0.96	1.00	0.95	1.00	0.94	1.00	0.93
15	1.00	0.95	0.97	0.93	0.97	0.92	0.97	0.91
20	1.00	0.93	0.97	0.91	0.90	0.87	0.93	0.90
25	1.00	0.91	0.90	0.83	0.83	0.80	0.85	0.80
30	0.97	0.90	0.85	0.80	0.83	0.78	0.91	0.88
40	1.00	0.88	0.96	0.90	0.97	0.88	0.93	0.86

Moreover, according to different neighborhood K values, some dataset accuracy performance is better, and several other dataset's accuracy performances are in "V" shape. The "V" shape shows that the K is optimal data-dependent.

4.4 Mean-Precision

Tables 7, 8, 9, 10, 11 shows the calculated mean-precision of Iris, Wine, LFW, Breast Cancer, and Digits Datasets from the Eq. (2). In Figs. 6, 7, 8, 9, 10, we have represented the calculated mean-precision of five different high dimensional datasets with NRIC and Isomap methods. The graph of our proposed NRIC method consistently achieves the high mean-precision with different neighborhood K values. We have compared NRIC and Isomap with four classification techniques such as SVM, KNN, NB, and RF for five datasets. The mean-precision performance of NRIC with classification techniques very well and provided effective results for the Isomap noise problem.

We have attained 100% mean-precision for the Iris dataset for different values of K (10, 15, 20, 25, and 40). Table 7 shows that LTSA-SVM is provided a 100% mean-precision for the values of K (10, 15, 20, 25, and 40); LTSA-KNN, LTSA-NB, and LTSA-RF are provided the 100% mean-precision on the value of K=10. The mean-precision of ISO-SVM is provided 96% on K=10, 95% mean-precision for ISO-KNN on K=10, 94% mean-precision for ISO-

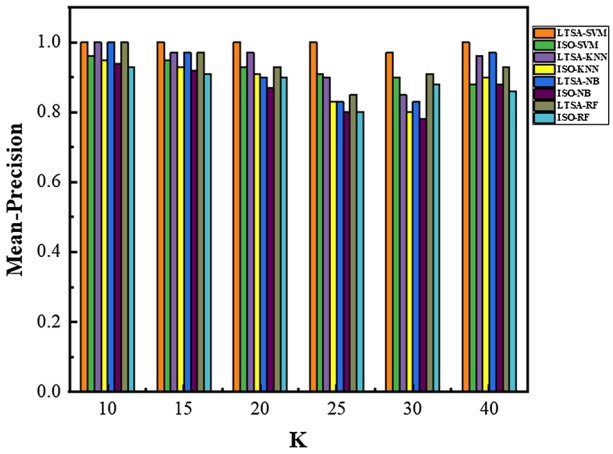

Fig. 6 Mean-precision of iris dataset with NRIC and isomap methods

Table 8 Mean-precision of wine dataset with SVM, KNN, NB, and RF

K	LTSA-SVM	ISO-SVM	LTSA-KNN	ISO-KNN	LTSA-NB	ISO-NB	LTSA-RF	ISO-RF
10	0.95	0.90	0.90	0.87	0.97	0.92	0.94	0.91
15	1.00	0.89	1.00	0.85	0.97	0.90	0.97	0.88
20	0.95	0.87	1.00	0.82	0.95	0.88	0.92	0.85
25	0.97	0.85	0.97	0.80	1.00	0.85	0.95	0.83
30	1.00	0.83	0.92	0.78	0.97	0.83	0.83	0.79
40	1.00	0.80	0.97	0.86	0.95	0.80	0.92	0.75

NB on K=10, and 93% for ISO-RF on K=10. In Fig. 6, the performance of the LTSA with SVM, KNN, NB, RF is better than Isomap.

For the Wine dataset, we have attained 100% mean-precision only on K (15, 30, and 40) for the LTSA-SVM. In Table 8, mean-precision of 100% LTSA-KNN (K=15 and 20), 100% of LTSA-NB on K=25, and 97% LTSA-RF (K=15). Therefore, we have compared Isomap with classification techniques such as ISO-SVM is presented with 90% mean-precision, 87% of ISO-KNN, 92% of ISO-NB, and 91% of ISO-RF are provided mean-precision on the same value of K=10. In Fig. 7, we have shown the mean-precision of the NRIC and Isomap methods with classification techniques. However, the mean-precision of LTSA with SVM, KNN, NB, and RF is better than the Isomap method.

For Labeled Faces in the Wild (LFW), we have attained 100% mean-precision on the value of (K=10,15,20,25,30,40) for LTSA-SVM and LTSA-RF, as shown in Table 9. The mean-precision performance of the 90% LTSA-KNN (K=20, 25, 40), 83% LTSA-NB (K=40). Therefore, mean-precision performance of the 89% ISO-SVM (K=20,40), 87% ISO-KNN (K=25), 80% ISO-NB (K=30,40), and 96% ISO-RF (K=10). In Fig. 8, we have shown the comparative performance of the mean-precision of the LTSA and Isomap method with classification techniques. However, LTSA is performed better with all classification techniques rather than Isomap.

For the Breast Cancer dataset, we have reached a 98% mean-precision for LTSA-SVM, 95% LTSA-KNN both on K=40, 95% LTSA-NB, and LTSA-RF (K=30), as shown in Table 10.

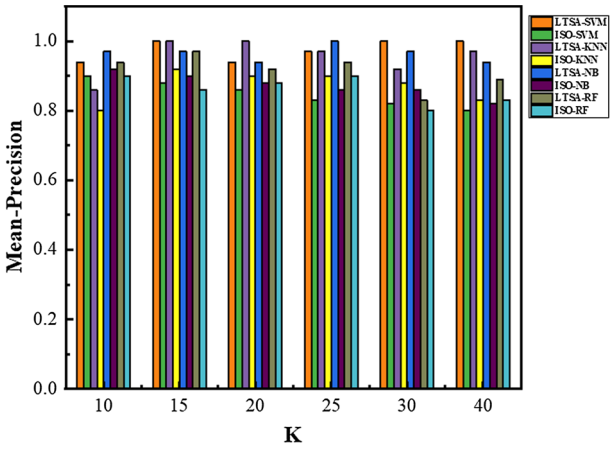

Fig. 7 Mean-precision of wine dataset with NRIC and isomap methods

Table 9 Mean-precision of LFW people dataset with SVM, KNN, NB, and RF

K	LTSA-SVM	ISO-SVM	LTSA-KNN	ISO-KNN	LTSA-NB	ISO-NB	LTSA-RF	ISO-RF
10	0.90	0.87	0.85	0.82	0.73	0.70	1.00	0.96
15	0.90	0.84	0.88	0.85	0.78	0.75	1.00	0.94
20	0.93	0.89	0.90	0.84	0.78	0.75	1.00	0.92
25	0.93	0.88	0.90	0.87	0.73	0.70	1.00	0.92
30	0.92	0.86	0.88	0.84	0.80	0.80	1.00	0.90
40	0.93	0.89	0.90	0.82	0.83	0.80	1.00	0.90

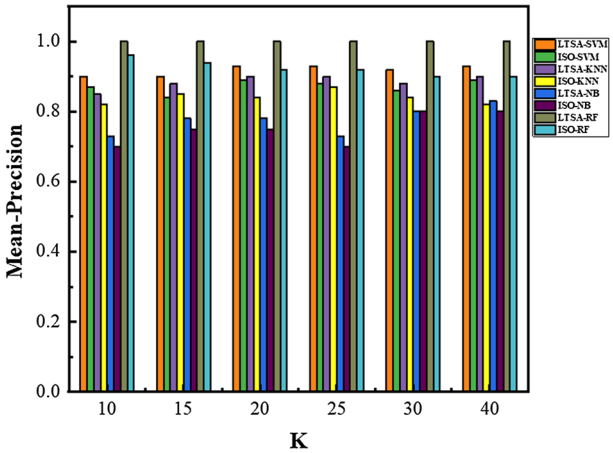

Fig. 8 Mean-precision of LWF people dataset with NRIC and isomap methods

Table 10 Mean-precision of breast cancer dataset with SVM, KNN, NB, and RF

K	LTSA-SVM	ISO-SVM	LTSA-KNN	ISO-KNN	LTSA-NB	ISO-NB	LTSA-RF	ISO-RF
10	0.96	0.93	0.93	0.90	0.91	0.90	0.91	0.88
15	0.95	0.92	0.92	0.90	0.89	0.87	0.89	0.87
20	0.96	0.92	0.92	0.90	0.90	0.87	0.91	0.85
25	0.95	0.90	0.93	0.90	0.94	0.87	0.90	0.88
30	0.95	0.90	0.93	0.90	0.95	0.90	0.95	0.93
40	0.98	0.95	0.95	0.91	0.94	0.87	0.89	0.86

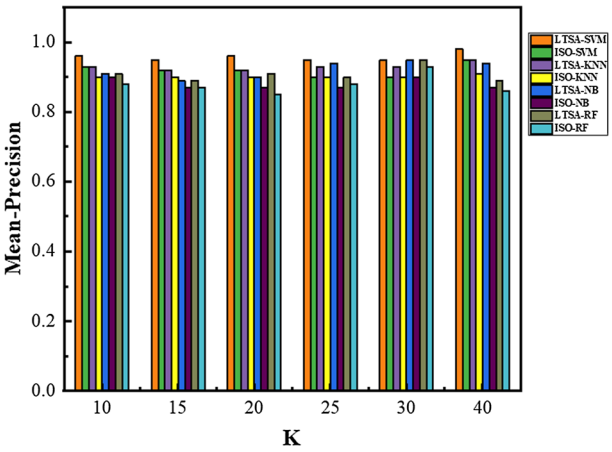

Fig. 9 Mean-precision of breast cancer dataset with NRIC and isomap methods

Table 11 Mean-precision of digits dataset with SVM, KNN, NB, and RF

K	LTSA-SVM	ISO-SVM	LTSA-KNN	ISO-KNN	LTSA-NB	ISO-NB	LTSA-RF	ISO-RF
10	0.95	0.92	0.95	0.93	0.94	0.90	0.87	0.84
15	0.96	0.93	0.94	0.91	0.93	0.90	0.84	0.80
20	0.96	0.93	0.95	0.90	0.91	0.89	0.79	0.76
25	0.97	0.93	0.94	0.89	0.89	0.86	0.78	0.76
30	0.97	0.94	0.93	0.89	0.89	0.85	0.79	0.75
40	0.96	0.94	0.93	0.84	0.92	0.85	0.78	0.72

Therefore, the mean-precision performance of the ISO-SVM is 95%, 91% ISO-KNN both on (K=40), ISO-NB is achieved 90%, and 93% ISO-RF on the same value of K = 30 in Table 10. The mean-precision performance of the NRIC is higher than the Isomap method. In Fig. 9, we have shown the comparative performance of the mean-precision of the LTSA and Isomap methods with classification techniques. However, LTSA is performed very well with SVM, KNN, NB, RF rather than the Isomap method.

For the Digits dataset, we have reached high mean-precision 97% on K=25, 30 for LTSA-SVM, as shown in Table 11. The mean-precision of the LTSA-KNN is attained 95% on (K=10, 20), 94% for LTSA-NB, and 87% LTSA-RF on the same value of K=10. Therefore,

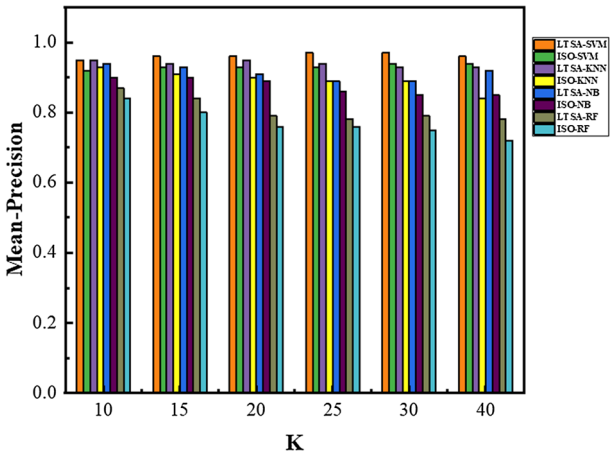

Fig. 10 Mean-precision of digits dataset with NRIC and isomap methods

ISO-SVM is achieved 94% on K=30 and 40, 93% ISO-KNN, 90% ISO-NB, and 84% ISO-RF on the same value of K=10. In Fig. 10, we have shown the comparative performance of the mean-precision of the LTSA and Isomap methods with classification techniques. However, the performance of the LTSA is well with SVM, KNN, NB, and RF rather than Isomap. Moreover, the overall mean-precision performance of the NRIC method consistently attains the high mean-precision for five high dimensional datasets. Sometimes, mean-precision performance is high and low according to the different values of K and classification techniques. Sometimes KNN values are close to each other and find easily. Otherwise, the KNN is far away from each other and find hardly.

4.5 Mean-Recall

Tables 12, 13, 14, 15, 16 shows the calculated mean-recall of Iris, Wine, LFW, Breast Cancer, and Digits Datasets from the Eq. (3). In Figs. 11, 12, 13, 14, 15, we have represented the calculated mean-recall of five different high dimensional datasets. The graph of our proposed NRIC method consistently reaches the high mean-recall of five datasets. We have used the different neighborhood values of K as same as accuracy and mean-precision. The mean-recall performance of LTSA with classification techniques very well and provided effective results for the proposed NRIC method.

We have attained 100% mean-recall for the Iris dataset for different K (10, 15, 20, 25, and 40) of LTSA-SVM. LTSA-KNN, LTSA-NB, and LTSA-RF are provided 100% mean-recall on the value of K=10 in Table 12. Therefore, 93% ISO-KNN mean-recall on K (10, 15, and 20) values, ISO-SVM is provided the 93% mean-recall, 92% ISO-NB, and 94% LTSA-RF, are on the same value of K=10. The performance of Isomap is lesser than the LTSA with classification techniques. In Fig. 11, the mean-recall performance of LTSA with SVM, KNN, NB, and RF is better than the Isomap method.

We have attained 100% mean-precision on (K=15,30, and 40) for the LTSA-SVM wine dataset. In Table 13, the mean-recall of the LTSA-KNN is 100% (K=15 and 20), LTSA-NB is 100% (K=25), and LTSA-RF is 97% (K=15). Therefore, the mean-recall of the ISO-SVM is 90% (K=10), ISO-KNN is 92% on (K=15), ISO-NB is 92% on (K=10), and ISO-RF is

Table 12 Mean-recall of iris dataset with SVM, KNN, NB, and RF

K	LTSA-SVM	ISO-SVM	LTSA-KNN	ISO-KNN	LTSA-NB	ISO-NB	LTSA-RF	ISO-RF
10	1.00	0.93	1.00	0.93	1.00	0.92	1.00	0.94
15	1.00	0.92	0.97	0.93	0.97	0.90	0.97	0.93
20	1.00	0.90	0.97	0.93	0.90	0.88	0.93	0.90
25	1.00	0.89	0.90	0.83	0.83	0.90	0.83	0.80
30	0.97	0.87	0.83	0.80	0.83	0.78	0.90	0.85
40	1.00	0.86	0.93	0.80	0.97	0.80	0.93	0.82

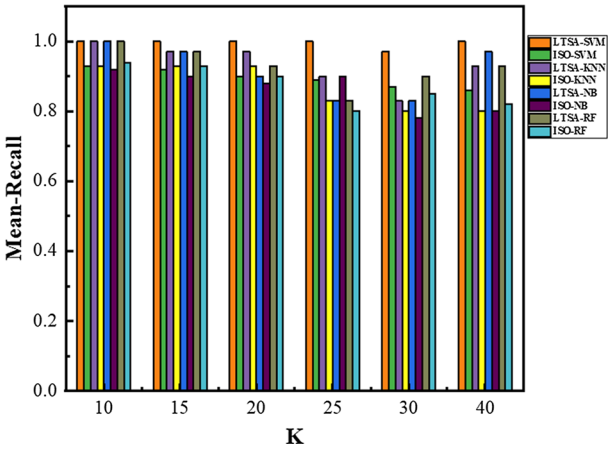

Fig. 11 Mean-recall of Iris dataset with NRIC and isomap methods

Table 13 Mean-recall of wine dataset with SVM, KNN, NB, and RF

K	LTSA-SVM	ISO-SVM	LTSA-KNN	ISO-KNN	LTSA-NB	ISO-NB	LTSA-RF	ISO-RF
10	0.94	0.90	0.86	0.80	0.97	0.92	0.94	0.90
15	1.00	0.88	1.00	0.92	0.97	0.90	0.97	0.86
20	0.94	0.86	1.00	0.90	0.94	0.88	0.92	0.88
25	0.97	0.83	0.97	0.90	1.00	0.86	0.94	0.90
30	1.00	0.82	0.92	0.88	0.97	0.86	0.83	0.80
40	1.00	0.80	0.97	0.83	0.94	0.82	0.89	0.83

90% on (k=10). In Fig. 12, we have shown the mean-recall of the LTSA algorithm with classification techniques and Isomap. However, the mean-recall performance of LTSA with NB, RF, KNN, and SVM is higher than the Isomap method.

For Labeled Faces in the Wild (LFW), we have attained a 93% mean-recall on the value of K=20 for the LTSA-SVM. The mean-recall of the LTSA-KNN is 88% on K=30, 81% LTSA-NB on (K=40), and 68% LTSA-RF have the same values for all K in Table 14. Therefore, the mean-recall of the ISO-SVM is 89% on K=40, 80% ISO-KNN on (K=10), 80% ISO-NB on (K=40) 65% ISO-RF have the same values for all K in Table 14. We have shown the mean-recall of the LTSA algorithm's comparative performance with classification techniques and Isomap in Fig. 13. However, LTSA performs better with SVM, KNN, and NB rather than the RF and Isomap.

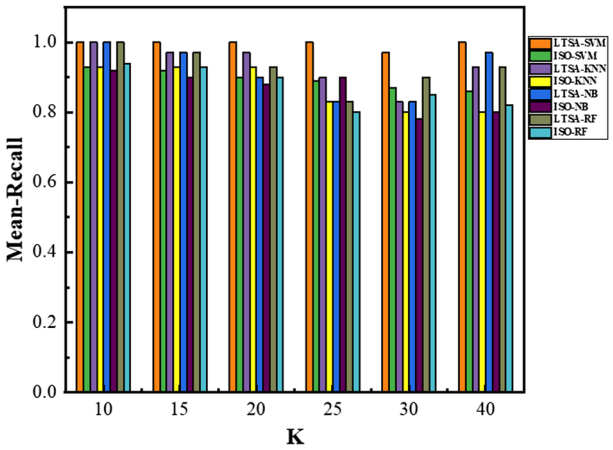

Fig. 12 Mean-recall of wine dataset with NRIC and isomap methods

Table 14 Mean-recall of LFW people dataset with SVM, KNN, NB, and RF

K	LTSA-SVM	ISO-SVM	LTSA-KNN	ISO-KNN	LTSA-NB	ISO-NB	LTSA-RF	ISO-RF
10	0.90	0.87	0.81	0.80	0.63	0.68	0.68	0.65
15	0.90	0.82	0.85	0.79	0.75	0.72	0.68	0.65
20	0.93	0.88	0.81	0.80	0.75	0.72	0.68	0.65
25	0.92	0.88	0.84	0.76	0.70	0.70	0.68	0.65
30	0.92	0.85	0.88	0.75	0.79	0.77	0.68	0.65
40	0.92	0.89	0.85	0.75	0.81	0.80	0.68	0.65

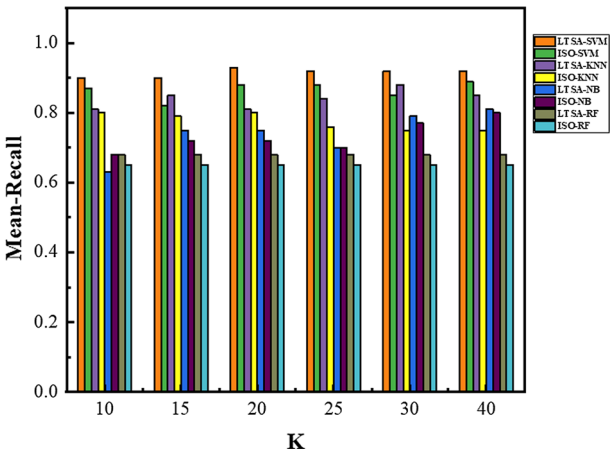

Fig. 13 Mean-recall of LWF people dataset with NRIC and isomap methods

Table 15 Mean-recall of breast cancer dataset with SVM, KNN, NB, and RF

K	LTSA-SVM	ISO-SVM	LTSA-KNN	ISO-KNN	LTSA-NB	ISO-NB	LTSA-RF	ISO-RF
10	0.96	0.93	0.92	0.91	0.90	0.88	0.90	0.91
15	0.95	0.92	0.91	0.88	0.86	0.87	0.82	0.82
20	0.96	0.98	0.90	0.84	0.89	0.89	0.88	0.86
25	0.95	0.92	0.90	0.75	0.93	0.87	0.80	0.80
30	0.95	0.92	0.91	0.77	0.94	0.89	0.94	0.91
40	0.98	0.94	0.94	0.73	0.92	0.86	0.82	0.80

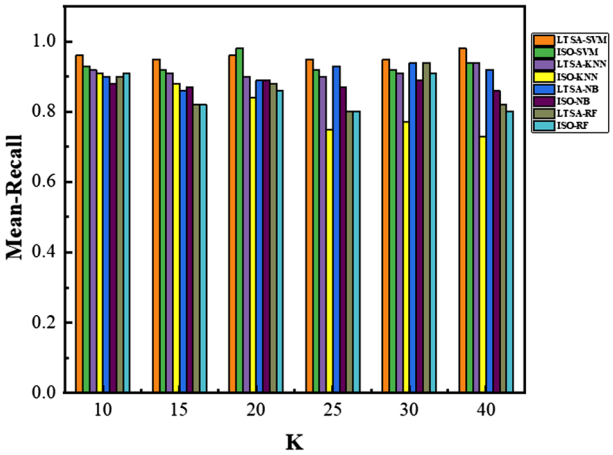

Fig. 14 Mean-recall of breast cancer dataset with NRIC and isomap methods

For the Breast Cancer dataset, we have reached a 98% mean-recall value of K=40 for the LTSA-SVM, 94% LTSA-KNN on the same value of K=10, 94% LTSA-NB, and LTSA-RF on the same value of K=30, as shown in Table 15. In Table 15, therefore, the mean-recall of the ISO-SVM is 98% on (K=20), 88% ISO-KNN on (K=15), 89% ISO-NB on (K=20, 30), and ISO-RF is 91% on the value of K = 10, 30. The ISO-SVM performance is higher than LTSA-SVM on K= 20, and ISO-RF is higher than LTSA-RF on K=10. In Fig. 14, we have shown the mean-recall of the LTSA algorithm's comparative performance with classification techniques and Isomap.

For the Digits dataset, we have reached a high mean-recall 97% on K=30 for the LTSA-SVM, as shown in Table 16. The mean-recall of the LTSA-KNN is attained 95% on K= (10, 20), 94% LTSA-NB, and 84% for LTSA-RF on the same value of K=10. Therefore, mean-recall of the ISO-SVM is attained 92% on K= (10, 15, 20, and 25), 93% ISO-KNN on K=10, and 90% for ISO-NB on (K=10, 15, 20, 25, and 40), and 80% ISO-RF on the value of (K=10, 20). In Fig. 15, we have shown the mean-recall of the LTSA algorithm's comparative performance with classification techniques and Isomap. However, the performance of the LTSA is better with KNN, NB, and RF rather than SVM and Isomap. Moreover, the overall mean-recall performance of the NRIC method consistently attains the high results of the five high dimensional datasets. The mean-recall performance is better and down according to the different neighborhood values of K and classification techniques.

Table 16 Mean-recall of digits dataset with SVM, KNN, NB, and RF

K	LTSA-SVM	ISO-SVM	LTSA-KNN	ISO-KNN	LTSA-NB	ISO-NB	LTSA-RF	ISO-RF
10	0.95	0.92	0.95	0.93	0.94	0.90	0.84	0.80
15	0.96	0.92	0.94	0.91	0.92	0.90	0.79	0.80
20	0.96	0.92	0.95	0.90	0.91	0.90	0.68	0.62
25	0.96	0.92	0.93	0.88	0.89	0.90	0.75	0.73
30	0.97	0.91	0.94	0.88	0.88	0.87	0.69	0.62
40	0.95	0.91	0.92	0.81	0.91	0.90	0.75	0.69

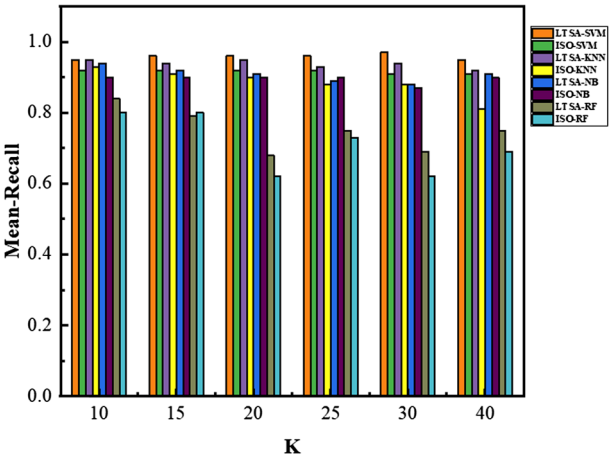

Fig. 15 Mean-recall of digits dataset with NRIC and isomap methods

4.6 Area Under the ROC Curve (AUC)

Tables 17, 18, 19, 20, 21 shows the calculated area under the ROC (Receiving Operating Characteristics) curve (AUC) of Iris, Wine, LFW, Breast Cancer, and Digits Datasets from Eqs. (4) and (5). In Figs. 16, 17, 18, 19, 20, we have represented the calculated AUC values of five different high dimensional datasets. The graph of our proposed NRIC method consistently reaches the high AUC values of five datasets. We have used the different neighborhood values of K as same as accuracy, mean-precision, and mean-recall. The area's performance under the ROC (Receiving Operating Characteristics) curve (AUC) of LTSA with classification techniques very well and provided effective results for the proposed NRIC method.

We have computed AUC values for the Iris dataset for different K (10, 15, 20, 25, and 40) values in Table 17. The AUC values of LT-SVM-A, LT-KNN-A, LT-NB-A, and LT-RF-A are 0.734, 0.736, 0.776, and 0.779, respectively, on the different values of K in Table 17. Therefore, the AUC values of IS-SVM-A (0.697), IS-KNN-A (0.693), IS-NB-A (0.761), and IS-RF-A (0.682) on the different values of K. According to Fig. 16, the performance of the ROC curve of LTSA with RF and NB is significant rather than SVM and KNN. In contrast, the high value of AUC is RF and NB. It significantly shows that RF and NB models are improved the performance of our proposed NRIC method. However, the ROC curve performance of the Isomap method with classification techniques is not better than our proposed method.

Table 17 AUC of Iris dataset with SVM, KNN, NB, and RF

K	LT-SVM-A	IS-SVM-A	LT-KNN-A	IS-KNN-A	LT-NB-A	IS-NB-A	LT-RF-A	IS-RF-A
10	0.734	0.697	0.620	0.595	0.745	0.725	0.690	0.655
15	0.722	0.689	0.622	0.600	0.776	0.761	0.718	0.667
20	0.661	0.615	0.618	0.581	0.667	0.651	0.754	0.682
25	0.670	0.633	0.736	0.697	0.582	0.555	0.736	0.664
30	0.708	0.659	0.728	0.693	0.602	0.579	0.761	0.663
40	0.733	0.693	0.719	0.689	0.726	0.694	0.779	0.657

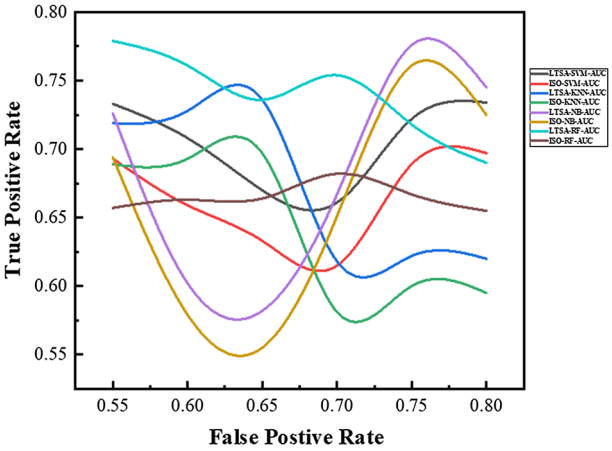

Fig. 16 ROC curve of Iris dataset with NRIC and isomap methods

Table 18 AUC of wine dataset with SVM, KNN, NB, and RF

K	LT-SVM-A	IS-SVM-A	LT-KNN-A	IS-KNN-A	LT-NB-A	IS-NB-AUC	LT-RF-A	IS-RF-A
10	0.542	0.540	0.521	0.428	0.538	0.507	0.507	0.490
15	0.501	0.497	0.584	0.409	0.485	0.429	0.518	0.500
20	0.544	0.534	0.455	0.430	0.512	0.500	0.452	0.420
25	0.578	0.511	0.463	0.406	0.596	0.488	0.491	0.430
30	0.561	0.462	0.461	0.396	0.567	0.505	0.448	0.425
40	0.655	0.402	0.471	0.451	0.465	0.434	0.453	0.422

For the Wine dataset, the AUC values of LT-SVM-A, LT-KNN-A, LT-NB-A, and LT-RF-A are 0.655, 0.584, 0.596, and 0.518, respectively, on the different values of K in Table 18. According to Fig. 17, the ROC curve of LTSA with SVM is better than KNN, NB, and RF because the high value of AUC is SVM. It significantly shows that the SVM model is significantly improved the performance of our proposed NRIC method. Therefore, the AUC values of IS-SVM-A (0.540), IS-KNN-A (0.451), IS-NB-A (0.507), and IS-RF-A (0.500) on the different values of K. However, the ROC curve performances of the Isomap method with classification techniques are not better than our proposed NRIC method.

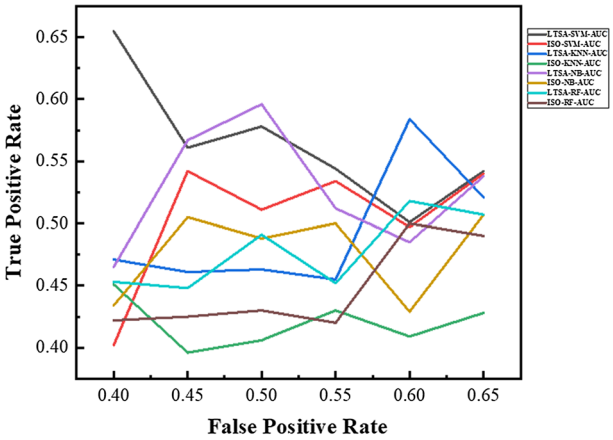

Fig. 17 ROC curve of wine dataset with NRIC and isomap methods

Table 19 AUC of LWF people dataset with SVM, KNN, NB, and RF

K	LT-SVM-A	IS-SVM-A	LT-KNN-A	IS-KNN-A	LT-NB-A	IS-NB-A	LT-RF-A	IS-RF-A
10	0.311	0.300	0.500	0.428	0.576	0.507	0.737	0.573
15	0.574	0.554	0.499	0.480	0.451	0.500	0.462	0.455
20	0.498	0.347	0.486	0.475	0.423	0.412	0.762	0.424
25	0.548	0.449	0.519	0.443	0.310	0.404	0.509	0.401
30	0.514	0.501	0.541	0.478	0.507	0.470	0.542	0.336
40	0.527	0.489	0.496	0.471	0.451	0.433	0.483	0.580

For the LWF People dataset, the AUC values of LT-SVM-A (0.574), LT-KNN-A (0.541), LT-NB-A (0.576), and LT-RF-A (0.737) on the different values of K in Table 19. In Fig. 18, the performance of the ROC curve of LTSA with RF is well rather than SVM, KNN, and NB. It significantly shows that the RF model is significantly improved the performance of our proposed NRIC method. Therefore, the AUC values of IS-SVM-A (0.554), IS-KNN-A (0.480), IS-NB-A (0.507), and IS-RF-A (0.580) on the different values of K. Moreover, the Isomap method works well with the RF model. However, the ROC curve performance of the Isomap method with classification techniques is not better than our proposed NRIC method.

For the Breast Cancer dataset, the AUC values of LT-SVM-A (0.744), LT-KNN-A (0.491), LT-NB-A (0.621), and LT-RF-A (0.521) on the different values of K in Table 20. In Fig. 19, the ROC curve of LTSA with SVM is better than KNN, NB, and RF. It significantly shows that the SVM model has improved our proposed NRIC method performance rather than other classification models. Therefore, the AUC values of IS-SVM-A (0.425, 0.547), IS-KNN-A (0.471), IS-NB-A (0.560), and IS-RF-A (0.626) on the different values of K. Moreover, the Isomap method works well with the RF model, but the IS-SVM-A value is higher than the LT-SVM-A on the value of K=15 in Table 20. However, the ROC curve performance of the Isomap method with classification techniques is not better than our proposed NRIC method.

For the Digits dataset, AUC values of LT-SVM-A (0.714 on K=40), LT-KNN-A (0.528 on K=20), LT-NB-A (0.528 on K=20), and LT-RF-A (0.530 on K=25 and 30) in Table 21. In Fig. 20, the ROC curve of LTSA with SVM is better than KNN, NB, and RF. It significantly

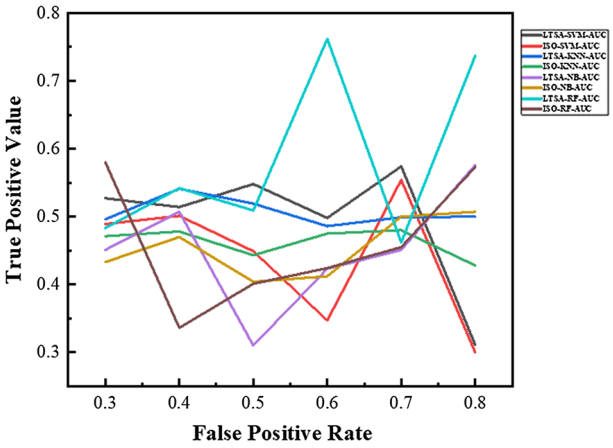

Fig. 18 ROC curve of LWF people dataset with NRIC and isomap methods

Table 20 AUC of breast cancer dataset with SVM, KNN, NB, and RF

K	LT-SVM-A	IS-SVM-A	LT-KNN-A	IS-KNN-A	LT-NB-A	IS-NB-A	LT-RF-A	IS-RF-A
10	0.524	0.520	0.491	0.463	0.594	0.424	0.432	0.626
15	0.424	0.425	0.486	0.471	0.472	0.406	0.377	0.342
20	0.506	0.451	0.474	0.465	0.544	0.471	0.409	0.399
25	0.379	0.342	0.490	0.443	0.518	0.414	0.521	0.513
30	0.744	0.465	0.477	0.471	0.621	0.560	0.520	0.466
40	0.548	0.547	0.471	0.411	0.437	0.464	0.460	0.452

Table 21 AUC of digits dataset with SVM, KNN, NB, and RF

K	LT-SVM-A	IS-SVM-A	LT-KNN-A	IS-KNN-A	LT-NB-A	IS-NB-A	LT-RF-A	IS-RF-A
10	0.511	0.500	0.485	0.451	0.524	0.524	0.445	0.447
15	0.544	0.532	0.488	0.427	0.421	0.418	0.441	0.437
20	0.464	0.416	0.528	0.514	0.528	0.528	0.446	0.424
25	0.444	0.439	0.516	0.493	0.379	0.361	0.530	0.387
30	0.638	0.582	0.393	0.380	0.433	0.426	0.530	0.517
40	0.714	0.506	0.483	0.484	0.373	0.325	0.433	0.276

shows that the SVM model has improved our proposed NRIC method's performance rather than other classification models. Therefore, the AUC values of IS-SVM-A (0.582 on K=30), IS-KNN-A (0.484 on K=40), IS-NB-A (0.524, 0.528 on K=10 and 20), and IS-RF-A (0.447 on K=10). Moreover, the Isomap method works well with the SVM model, but the IS-RF-A value is higher than LT-RF-A on the value of K=10 in Table 21. However, the overall ROC curve performance of the Isomap method with classification techniques is not better than our proposed NRIC method.

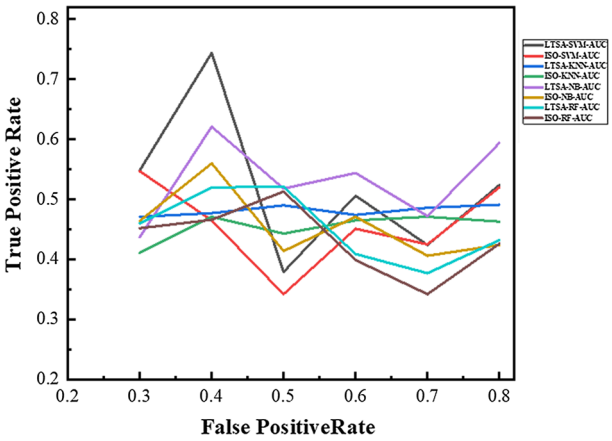

Fig. 19 ROC curve of breast cancer dataset with NRIC and isomap methods

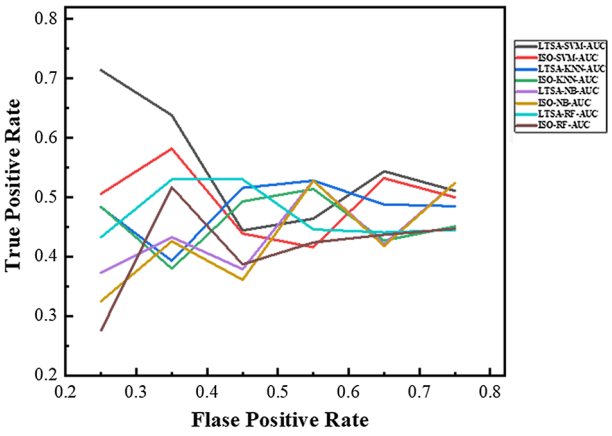

Fig. 20 ROC curve of digits dataset with NRIC and isomap methods

5 Conclusion

This paper proposed a Noise Removal Isomap with a Classification (NRIC) method for the Isomap noise problem. This research paper aims to focus on the noisy nonlinear manifold learning method, such as Isomap. The main problem of the Isomap is sensitivity to noise and cannot easily generate the data after de-noising. Our proposed method can easily map the high dimensional data into low dimensional space. Our NRIC method can identify the noise from data points and eliminate the noise in datasets. In this paper, we compared four classification techniques, including SVM, KNN, NB, and RF, with the LTSA algorithm to improve the accuracy of the noise of Isomap and improve the accuracy of the datasets. We compared four classification techniques on five different datasets to gain insight into what technique is suitable for Isomap noise. We calculated accuracy, mean-precision, mean-recall, and area under the (ROC) curve (AUC) for the NRIC method. Our experiment result shows that our proposed method is much efficient than Isomap and provides highly accurate results of high dimensional datasets, and can easily optimize the graph. In future work, we will analyze

the proposed method's performance on other datasets and calculated the other classification metrics and time complexity analysis. We will be used the same idea of other manifold learning techniques.

Funding This work is supported by the Strategic Priority Research Program of Chinese Academy of Sciences (A Class) NO. XDA19020102.

Declaration

Human and animal rights This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Al-Shalabi R, Kanaan G, Gharaibeh M (2006) Arabic text categorization using knn algorithm. In: Proceedings of The 4th international multiconference on computer science and information technology, vol 4, pp 5–7
2. Balasubramanian M, Schwartz EL, Tenenbaum JB, de Silva V, Langford JC (2002) The isomap algorithm and topological stability. Science 295(5552):7–7
3. Bansal H, Shrivastava G, Nguyen GN, Stanciu LM (2018) Social network analytics for contemporary business organizations. IGI Global, Singapore
4. Burges CJ (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Disc 2(2):121–167
5. Chang H, Yeung DY (2006) Robust locally linear embedding. Pattern Recogn 39(6):1053–1065
6. Choi H, Choi S (2004) Kernel isomap. Electron Lett 40(25):1612–1613
7. Choi H, Choi S (2005) Kernel isomap on noisy manifold. In: Proceedings. The 4th international conference on development and learning, 2005, IEEE, pp 208–213
8. Choi H, Choi S (2007) Robust kernel isomap. Pattern Recogn 40(3):853–862
9. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
10. Fawcett T (2006) An introduction to roc analysis. Pattern Recogn Lett 27(8):861–874
11. Fisher RA (1950) Contributions to mathematical statistics. American Psychological Association, Washington
12. Forina M (1991) Uci machine learning repository wine dataset. Institute of Pharmaceutical and Food Analysis and Technologies
13. Han J, Pei J, Kamber M (2011) Data mining: concepts and techniques. Elsevier, Amsterdam
14. Han Y, Cheng Q, Hou Y (2018) Fault detection method based on improved isomap and svm in noise-containing nonlinear process. In: 2018 international conference on control. automation and information sciences (ICCAIS), IEEE, pp 461–466
15. Ho TK (1995) Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition, IEEE, vol 1, pp 278–282
16. Ho TK (1998) Nearest neighbors in random subspaces. In: Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), Springer, Berlin, pp 640–648
17. Hong C, Yu J, Zhang J, Jin X, Lee KH (2018) Multimodal face-pose estimation with multitask manifold deep learning. IEEE Trans Industr Inf 15(7):3952–3961
18. Hougardy S (2010) The floyd-warshall algorithm on graphs with negative cycles. Inf Process Lett 110(8–9):279–281
19. Huang GB, Mattar M, Berg T, Learned-Miller E (2008) Labeled faces in the wild: a database forstudying face recognition in unconstrained environments. In: Workshop on Faces in 'Real-Life' Images: detection, alignment, and recognition. https://hal.inria.fr/inria-00321923
20. Jain A, Mandowara J (2016) Text classification by combining text classifiers to improve the efficiency of classification. Int J Comput Appl (2250-1797) 6(2)
21. Jiang S, Pang G, Wu M, Kuang L (2012) An improved k-nearest-neighbor algorithm for text categorization. Expert Syst Appl 39(1):1503–1509
22. John GH, Langley P (2013) Estimating continuous distributions in bayesian classifiers. arXiv preprint arXiv:1302.4964

23. Kouropteva O, Okun O, Pietikäinen M (2002) Selection of the optimal parameter value for the locally linear embedding algorithm. FSKD 2:359–363
24. Kouropteva O, Okun O, Pietikäinen M (2005) Incremental locally linear embedding. Pattern Recogn 38(10):1764–1767
25. Kowsari K, Jafari Meimandi K, Heidarysafa M, Mendu S, Barnes L, Brown D (2019) Text classification algorithms: a survey. Information 10(4):150
26. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324
27. Lewis DD (1998) Naive (bayes) at forty: the independence assumption in information retrieval. In: European conference on machine learning, Springer, pp 4–15
28. Li B, Huang DS, Wang C (2008) Improving the robustness of isomap by de-noising. In: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), IEEE, pp 266–270
29. Lowe DG (1995) Similarity metric learning for a variable-kernel classifier. Neural Comput 7(1):72–85
30. McGill R, Tukey JW, Larsen WA (1978) Variations of box plots. Am Stat 32(1):12–16
31. McMahan B, Ramage D (2017) Federated learning: collaborative machine learning without centralized training data. Google Research Blog 3
32. Metsis V, Androutsopoulos I, Paliouras G (2006) Spam filtering with naive bayes-which naive bayes? CEAS, Mountain View, CA 17:28–69
33. Miranda AL, Garcia LPF, Carvalho AC, Lorena AC (2009) Use of classification algorithms in noise detection and elimination. In: International conference on hybrid artificial intelligence systems, Springer, pp 417–424
34. Mitchell T (1997) Machine learning. Mcgraw-hill higher education, New York
35. Nikam SS (2015) A comparative study of classification techniques in data mining algorithms. Oriental J Comput Sci Technol 8(1):13–19
36. Nikhath AK, Subrahmanyam K, Vasavi R (2016) Building a k-nearest neighbor classifier for text categorization. Int J Comput Sci Inf Technol 7(1):254–256
37. Qu T, Cai Z (2015) A fast isomap algorithm based on fibonacci heap. In: international conference in swarm intelligence, Springer, pp 225–231
38. Qu T, Cai Z (2017) An improved isomap method for manifold learning. Int J Intell Comput Cybernet
39. Rana S, Singh A (2016) Comparative analysis of sentiment orientation using svm and naive bayes techniques. In: 2016 2nd international conference on next generation computing technologies (NGCT), IEEE, pp 106–111
40. Saputra MFA, Widiyaningtyas T, Wibawa AP (2018) Illiteracy classification using k means-naïve bayes algorithm. JOIV Int J Inform Vis 2(3):153–158
41. Saxena A, Gupta A, Mukerjee A (2004) Non-linear dimensionality reduction by locally linear isomaps. In: International conference on neural information processing, Springer, pp 1038–1043
42. Shao C, Huang H (2005) Selection of the optimal parameter value for the isomap algorithm. In: Mexican international conference on artificial intelligence, Springer, pp 396–404
43. Sumithra V, Surendran S (2015) A review of various linear and non linear dimensionality reduction techniques. Int J Comput Sci Inf Technol 6:2354–2360
44. Tay B, Hyun JK, Oh S (2014) A machine learning approach for specification of spinal cord injuries using fractional anisotropy values obtained from diffusion tensor images. Comput Math Methods Medicine 2014
45. Tenenbaum JB, De Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290(5500):2319–2323
46. Thongkam J, Xu G, Zhang Y (2008) Adaboost algorithm with random forests for predicting breast cancer survivability. In: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), IEEE, pp 3062–3069
47. Tzacheva A, Ranganathan J, Mylavarapu SY (2019) Actionable pattern discovery for tweet emotions. In: International conference on applied human factors and ergonomics, Springer, pp 46–57
48. Ullah S, Jeong M, Lee W (2018) Nondestructive inspection of reinforced concrete utility poles with isomap and random forest. Sensors 18(10):3463
49. Vapnik V, Vapnik V (1998) Statistical learning theory 1:624
50. Vapnik VN (1995) The nature of statistical learning theory. Springer, Berlin
51. Verma R, Khurd P, Davatzikos C (2007) On analyzing diffusion tensor images by identifying manifold structure using isomaps. IEEE Trans Med Imaging 26(6):772–778
52. Wang J (2012) Geometric structure of high-dimensional data and dimensionality reduction. Springer, Berlin
53. Xu S (2018) Bayesian naïve bayes classifiers to text classification. J Inf Sci 44(1):48–59

54. Ye DH, Desjardins B, Hamm J, Litt H, Pohl KM (2014) Regional manifold learning for disease classification. IEEE Trans Med Imaging 33(6):1236–1247
55. Yin J, Hu D, Zhou Z (2008) Noisy manifold learning using neighborhood smoothing embedding. Pattern Recogn Lett 29(11):1613–1620
56. Yu J, Tao D, Wang M (2012) Adaptive hypergraph learning and its application in image classification. IEEE Trans Image Process 21(7):3262–3272
57. Yu J, Rui Y, Tao D (2014) Click prediction for web image reranking using multimodal sparse coding. IEEE Trans Image Process 23(5):2019–2032
58. Zhang B (2011) Multiple features facial image retrieval by spectral regression and fuzzy aggregation approach. Int J Intell Comput Cybernet
59. Zhang Z, Zha H (2004) Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. SIAM J Sci Comput 26(1):313–338