Dimensionality Reduction of Data with Neighbourhood Components Analysis

Hannah Kariuki, Samuel Mwalili, Anthony Waititu

Department of Statistics and Actuarial Science, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya

Email address:

hkariuki059@gmail.com (H. Kariuki), samuel.mwalili@gmail.com (S. Mwalili), agwaititu@gmail.com (A. Waititu)

*Corresponding author

Abstract: In most research fields, the amount of data produced is growing very fast. Analysis of big data offers potentially unlimited opportunities for information discovery. However, due to the high dimensions and presence of outliers, there is a need for a suitable algorithm for dimensionality reduction. By performing dimensionality reduction, we can learn low dimensional embeddings which capture most of the variability in data. This study proposes a new approach, Neighbourhood Components Analysis (NCA) a nearest-neighbor-based non-parametric method for learning low-dimensional linear embeddings of labeled data. This means that the approach uses class labels to guide the dimensionality reduction (DR) process. Neighborhood Components Analysis learns a low-dimensional linear projection of the feature space to improve the performance of a nearest neighbour classifier in the projected space. The method avoids making parametric assumptions about the data and therefore, can work well with complex or multi-modal data, which is the case with most real-world data. We evaluated the efficiency of our method for dimensionality reduction of data by comparing the classification errors and class separability of the embedded data with that of Principal Component Analysis (PCA). The result shows a significant reduction in the dimensions of the data from 754 to 55 dimensions. Neighborhood Components Analysis outperformed Principal Components Analysis in classification error across a range of dimensions. Analysis conducted on real and simulated datasets showed that the proposed algorithm is generally insensitive to the increase in the number of outliers and irrelevant features and consistently outperformed the classical Principal Component Analysis method.

Keywords: Dimensionality Reduction, Neighbourhood Components Analysis (NCA), Principal Component Analysis (PCA), Outlier Detection

1. Introduction

Data from real-world settings such as signal processing, speech recognition, digital photographs, neuroinformatic, and bioinformatics usually has high dimensionality [1]. For instance, health care data on the status of patients with many recorded parameters from age, weight, blood analysis, nutrition, immune system status, genetic background, operations, treatments, diagnosed diseases, etc. can be high-dimensional. Each dimension corresponds to a specific parameter.

The large number of dimensions enhance data information content, but at the same time result in a greater possibility of noise and redundancy. Also, the number of features can exceed that of observations. It, therefore, becomes difficult to predict certain properties when more variables are added to a dataset because each variable added makes the predictive power to decrease exponentially.

Large datasets are often contaminated by highly deviating samples, faulty measurements, and noise often referred to as outliers. Although there is no widely agreed-upon concept of an outlier, D. M. Hawkins defined outliers as observations that are different or inconsistent with the remainder of a dataset, as to cause suspicion that it was as a result of unrelated mechanism [2]. Outliers can arise due to fraudulent behaviour, human error, mechanical faults, changes in system behaviour, instrument error, deviations in populations. Increasing dimensionality of data adds to the complexity of detecting such anomalies. By reducing the number of attributes in the data, it becomes easier to apply statistical techniques that can extract

useful information. This also addresses the issues brought about by the curse of dimensionality.

Dimensionality reduction is the transformation of data from high-dimensional space to lower-dimensional space without loss of meaningful information in the original data [3]. Dimension reduction aims to discard redundant features and undesired properties of high dimensional space. This can be achieved through feature extraction or feature selection.

Feature selection methods involve selecting a subset of variables from the original set of variables and discarding others to preserve crucial information. Feature selection methods preserve the original physical meaning of the features [4]. Feature extraction methods construct new reduced features from the original large number of features through linear or non-linear combinations, which preserves the class separability as much as possible in the transformed space. The extracted features do not preserve the meaning of the original features, but each of the original features may contribute to making the transformed features.

Linear dimensionality reduction techniques are popular because they are both fast and relatively immune to overfitting. Both Neighbourhood Components Analysis (NCA) and Principal Component Analysis (PCA) are linear dimensionality reduction methods that apply a linear operator to the original data to achieve a reduced representation. This study aims to compare PCA and NCA techniques for dimensionality reduction of high dimensional data in presence of outliers.

2. Literature Review

Wu et al. described big data as a large volume of complex, growing datasets with multiple independent sources [5]. According to Fan et al., high dimensionality combined with a large sample size cause problems such as noise accumulation, false correlations, incidental homogeneity, heavy computational cost, and algorithmic instability [6]. Shetta & Niranjan observed the problem with high-dimensional datasets is that when the number of variables increases the volume of the space increases in such a way that available samples are not adequate to get statistically significant results [7].

According to Roweis et al., it is important to reduce the dimensionality of input data, either for regularization of a subsequent learning algorithm or for computational savings [8]. Dimension reduction is also useful in exploratory analysis and machine learning because it allows for visualization of samples which can then be used to detect outliers and identify clusters. Machine learning models with fewer variables also generalize better during the fitting process.

Several linear and nonlinear approaches have been proposed in the literature to derive meaningful low-dimensional representations of high-dimensional data, ranging from unsupervised to supervised techniques. Real-world data are likely to form a highly nonlinear manifold. Although much recent effort has focused on non-linear methods, linear dimensionality reduction techniques are still popular, they are both fast and relatively immune to overfitting. Linear projections also preserve some essential

topology of the original data [8].

PCA is a classical multivariate technique and first choice for dimensionality reduction of data. The method was first introduced by Pearson [9], and later developed independently by Hotelling [10]. The central idea of PCA is to reduce the dimensionality of data in which there is a large number of correlated variables while retaining as much as possible of the variation present in the dataset [11].

Astuti et al. used PCA for dimensionality reduction and SVM as a classifier for microarray data classification [12]. The results showed that the scheme performed well compared to previous research since the microarray data used was linearly separable. Reddy et al. investigated PCA and LDA dimensionality reduction techniques on four machine learning algorithms [13]. They found out that PCA outperformed LDA. Also, when the dimensionality of a dataset is high, PCA produces better results. Shetta & Niranjan observed that PCA is heavily affected by outliers present in the data [7]. Although PCA is widely considered as a basic model for dimensionality reduction, the method assumes that the data follows a specific model and requires a priori data knowledge, which is rare in real-world data.

Neighbourhood components analysis (NCA) was originally proposed by Roweis et al. for pattern recognition, classification, and dimensionality reduction [8]. This approach is well suited for dimensionality reduction since it does not lose any information during the process.

Singh-Miller et al. applied NCA method with a regularization term to acoustic modelling in a speech recognizer [14]. Singh-Miller applied NCA to the problem of acoustic modelling for speech recognition to perform dimensionality reduction on acoustic vectors [15]. Experiments showed that NCA performed competitively with heteroscedastic linear discriminant analysis (HLDA) a commonly employed dimensionality reduction technique in speech recognition.

Manit & Youngkong applied NCA to sEMG signal for gait phase pattern recognition and evaluated the efficiency of the method by comparing classification accuracy and class separability with that of PCA, linear discriminant analysis (LDA), and local preserving projection (LLP) [16]. They found that NCA outperformed the other methods in both classification accuracy and class separability.

Rizwan & Anderson performed dimensionality reduction to investigate computational and memory cost of speaker similarity score for phoneme classification using NCA method [17]. The results obtained showed a significant reduction in the dimensions of the TIMIT dataset from 50 dimensions 22 as well as 56% reduction in computational cost, and memory.

Ferdinando et al. explored how much NCA enhances emotion recognition using ECG-derived features [18]. The results showed that the method enhanced the performance and significantly improved the standard deviation for HRV-based features. Ferdinando & Alasaarela experiments showed that applying NCA enhanced the features such that new baselines were set by the performances in valence [19].

In order to address the limitations of PCA method, this study utilizes NCA technique which has proven to be a

suitable linear dimensionality reduction technique for real world data since the method assumes no parametric model for the class distributions or the boundaries between them. But instead relies on the strong regularization imposed by restricting a linear transformation of the original inputs.

3. Methodology

3.1. Principal Components Analysis Model

The PCA method calculates the covariance matrix of the given dataset and then finds the eigenvalues and eigenvectors of the covariance matrix [20]. A few eigenvectors whose eigenvalues have high variance are then selected to form the transformation matrix, which reduces the dimensions of the dataset. The main objective of PCA is to reduce the dimensions of a dataset from D to d and to project it onto a d-dimensional subspace where $(d \ll D)$ with high computational efficiency and retain most of the information.

$$x_i = \begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{Di} \end{pmatrix} \rightarrow \quad reduce \quad dimensionality \quad \rightarrow \begin{pmatrix} z_{1i} \\ z_{2i} \\ \vdots \\ z_{di} \end{pmatrix}$$

Given a pattern set x_i, where $x_i \in \mathbb{R}^D$, $i = 1, \dots, N$, the assumption made is that the data are centred around the origin. i.e $x_i \Leftrightarrow x_i - E(x_i)$.

The mean is given by;

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{1}$$

The covariance is given by;

$$C = \frac{1}{N} \sum_{i=1}^{N} x_i x_i^T \tag{2}$$

PCA formulates the eigenvalue problem.

$$\lambda v = Cv \tag{3}$$

where λ is eigenvalue, v is eigenvector, C is the corresponding covariance matrix over dataset x_i. Each eigenvector is a linear combination of original dimensions.

3.1.1. Eigenvalue Decomposition

Given that C is a square matrix, the solution for the eigenvalue problem consists of the following steps:

1) Solve all the λ which makes the matrix $(C - \lambda I)$ singular.

We determine a scalar λ which is the eigenvalue of matrix C such that the equation.

$$Cv = \lambda v, v \neq 0 \tag{4}$$

has a non-zero solution. v is the eigenvector associated with λ.

2) Given an eigenvalue λ, we solve for all non-zero vectors that meet $(C - \lambda I) = 0$.

To determine the non-zero vector v, we modify equation.

$$(C - \lambda I)v = 0 \tag{5}$$

For any vector v, the condition under which the equation $v \neq 0$ has a non-zero solution is, $\det(C - \lambda I) = 0$.

Spreading the left side of the determinant of the eigen matrix $(C - \lambda I)$ of C we obtain a polynomial equation.

$$\alpha_0 + \alpha_1 \lambda + \dots + \alpha_{n-1} \lambda^{n-1} + (-1)^n \lambda^n = 0 \tag{6}$$

The eigenvalues of C are denoted by $\lambda_1, \lambda_2, \dots, \lambda_n$. For each eigenvalue λ_i there is a corresponding eigenvector V which can be found by solving:

$$(C - \lambda_i I)V = 0 \tag{7}$$

Consider $\lambda_1, \lambda_2, \dots, \lambda_n$ eigenvalues of the covariance matrix C with $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ thus represents a proportion of the total variation.

$$Y_i = \frac{\lambda_i}{\sum_{i=1}^{n} \lambda_i} \tag{8}$$

The proportion of variation by the first d principal components should be large to avoid loss of information.

$$\frac{\lambda_1 + \lambda_2 + \dots + \lambda_n}{\lambda_1 + \lambda_2 + \dots + \lambda_n} \approx 1 \tag{9}$$

3.1.2. Low Dimensional Projection

Projecting down to d dimensions once all the principal components have been identified, the dimensionality of the dataset is reduced to d dimensions by projecting it onto the hyperplane defined by the first d PCs.

The transformation based on the principal components is defined as

$$z_i = W_d^T x_i \tag{10}$$

Which is the dot product of the matrix x_i by the matrix W_d, which is the matrix containing the first d principal components. Where W is the matrix of the first d eigenvectors of the highest eigenvalues of the covariance matrix C. Selecting this hyperplane ensures that the projection will preserve as much variance as possible.

The attributes x_i from the initial D-dimensional space are therefore transformed into z_i low d dimensional space by selecting the first d eigenvectors associated with the highest eigenvalues. Which is a polynomial of degree n in λ.

3.2. Neighborhood Components Analysis Model

Given a labelled dataset comprising of n real-valued input vectors $x_1, x_2, \dots, x_n \in \mathbb{R}^D$ with corresponding class labels $y_1, y_2, \dots, y_n$, We find a quadratic distance metric that optimizes nearest neighbour classification performance [8]. For quadratic (Mahalanobis) distance metric we can write the expression as:

$$d_{ij} = (x_i - x_j)^T Q(x_i - x_j)$$

where Q is a positive semi-definite matrix that can be decomposed using eigen decomposition. Taking A to denote the transformation matrix to be learned, a metric $Q = A^T A$ is learned such that the distance between two points x and y is defined as,

$$d(x, y) = (x - y)^T Q(x - y) = (Ax - Ay)^T (Ax - Ay)$$

For NCA to perform linear dimensionality reduction, we restrict A to be a non-square matrix of size $d \times D$ where $d \ll D$ in our optimization procedure.

3.2.1. Stochastic Neighbor Selection

NCA selects a single neighbour stochastically. Each point i selects another point j as its neighbour among k points with some probability p_{ij} and inherits its class label from the point it selects [8]. The probability that a point i will select itself as a neighbour is zero.

$$p_{ij} = \frac{\exp\left(-\|Ax_i - Ax_j\|^2\right)}{\sum_{k \neq j} \exp(-\|Ax_i - Ax_k\|)}, P_{ii} = 0 \qquad (11)$$

The stochastic selection rule aids finding the probability p_i that a point i will be correctly classified.

$$p_i = \sum_{j \in Y_i} p_{ij} \text{ where } y_i = \{j | y_i = y_j\} \qquad (12)$$

3.2.2. Objective (Cost) Function

The objective of the neighbourhood components analysis is to maximize the expected number of points correctly classified. The objective (cost) function is given by,

$$f(A) = \sum_i \sum_{j \in Y_i} p_{ij} = \sum_i p_i \qquad (13)$$

Optimize the cost function (13) using gradient descent on a non-square matrix A. NCA maximizes the expected number of points correctly classified according to the objective function $f(A)$.

Differentiating f with respect to the transformation matrix A provides a gradient rule used to optimize A.

Denote $x_{ij} = x_i - x_j$

$$\frac{\partial f}{\partial A} = -2A \sum_i \sum_{j \in Y_i} p_{ij} \left(x_{ij} x_{ij}^T - \sum_k p_{ik} x_{ik} x_{ik}^T \right) \qquad (14)$$

$$\frac{\partial f}{\partial A} = -2A \sum_i \left(p_i \sum_k p_{ik} x_{ik} x_{ik}^T - \sum_{j \in Y_i} p_{ij} x_{ij} x_{ij}^T \right) \qquad (15)$$

The learned transformation matrix A will map the vector samples from D dimensional space to a low-dimensional space d.

$$Z = A.X \qquad (16)$$

Where Z represents the transformed features in d-dimensional space.

3.3. Performance Measures

To evaluate our method for dimensionality reduction and outlier detection, we consider classification error of the K-Nearest neighbour (KNN) on the transformed data.

$$Error = 1 - Accuracy$$

Where,

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

3.4. Data

In our experiments for this study, we used two different datasets, Parkinson's Disease Classification dataset from the UC Irvine repository and a simulated dataset.

The Parkinson's Disease Classification data used for this work was obtained from the Department of Neurology, in Cerrahpasa Faculty of Medicine, Istanbul University, and provided by UCI Machine Learning Repository [21]. The dataset consists of 755 variables and 756 instances. The target variable consists of two classes. The dataset contains various features for speech signal processing; including Time Frequency features, Mel Frequency Cepstral Coefficients (MFCCs), Wavelet Transform based Features, Vocal Fold Features, and tuneable Q-factor wavelet transform (TWQT) features which have been applied to the speech recordings of patients with Parkinson's Disease (PD) to extract clinically useful information for assessment of PD.

3.4.1. Data Simulation

The simulation has focused on a scenario where outliers are present in the high dimensional data. The dataset comprises a total of 1000 samples with 500 features, where 50 are relevant features and the rest redundant. The number of redundant features are generated as random linear combinations of the informative features. The inliers are drawn from $N(0,1)$. We added q variance outliers from a $N(q, \alpha \Sigma q)$ distribution, where $\alpha = 3$ and $\Sigma = 1$ [22].

3.4.2. Data Preparation

As in the previous studies [1, 22], the features in the raw data were normalized to [0, 1].

The id variable from the Parkinson's Disease Classification dataset was dropped because it was not useful in the analysis and the data type of the target variable was converted to categorical to represent the two classes.

4. Results

Dimensionality Reduction

The results under the PCA and NCA transformations are shown in this section. We applied our models to Parkinson's Disease Classification data to model dimensionality reduction of the high-dimensional features and the simulated data with High-dimensional features contaminated with outliers.

4.1. Simulated Results

PCA model and NCA model were fitted on the data. The dimensions of the simulated dataset were reduced from 500 dimensions (D=500) to 50 (d=50) dimensions.

To evaluate the effectiveness of the proposed method to filter irrelevant features and outliers, we added varied percentage of outliers for 5%, 10%, 15%, 20%, 25%, and 30% to the simulated dataset.

To evaluate the effectiveness of our methods for dimension reduction of high dimensional data contaminated with outliers, we evaluate NCA and PCA on the simulated data

set. To compare the methods based on the projected results, we applied K-nearest neighbor classification on the projected data using k = 3 and computed the classification error. The method with the least error preserves most of the important features than the other. Using the same projection learned at training, the training set and all future test points were projected into low-dimensional space and K-nearest neighbor classification was performed to assess the models. The results under the PCA and NCA transformations appear in Simulation results.

Figure 1. gives the classification errors of KNN on the simulated data set by PCA and NCA embeddings by varying the percentage of outlier contamination. It can be observed from the simulated data set, KNN on NCA embeddings achieved consistently lower classification errors using k = 3 compared to PCA. KNN on PCA embeddings achieved higher classification errors across a range of outlier contamination. It can be observed that PCA is highly affected by outliers. NCA is not heavily affected by the outliers since the outlying points contribute less to the labelling of the surrounding points.
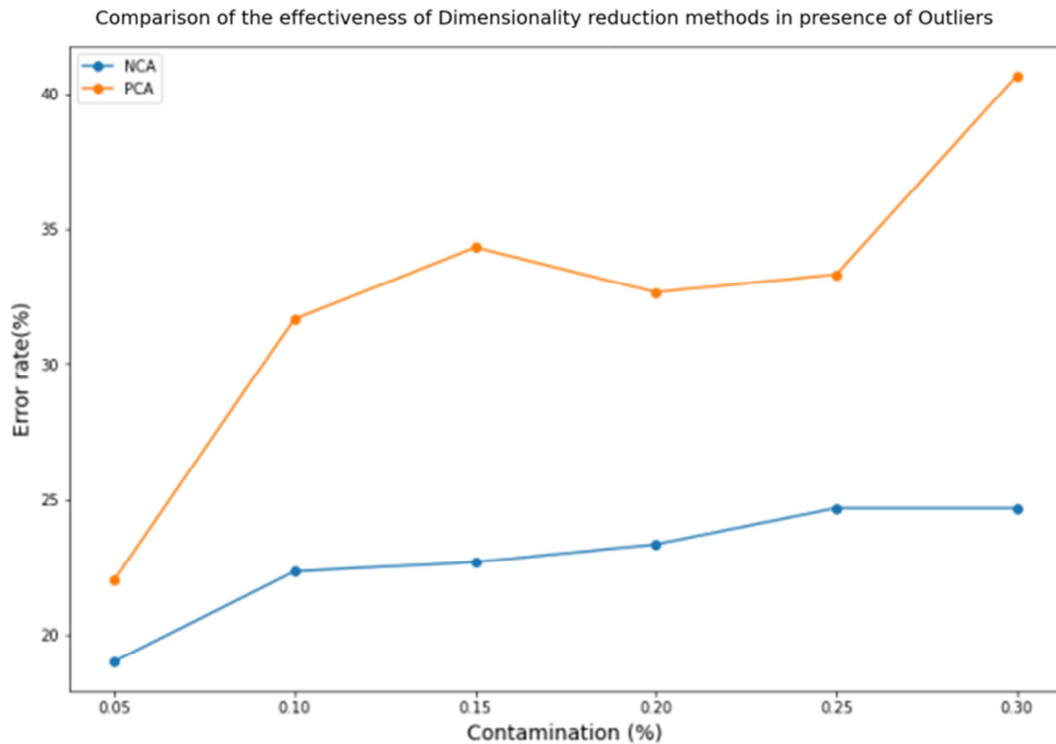

Figure 1. Classification Error % for various levels of outlier contamination.

4.2. Application on Real Data

The ability of PCA and NCA methods to perform dimensionality reduction was tested by applying the models to high dimensional dataset D=754.

Table 1. Proportion of variance explained and cumulative variance proportion for the principal components.

PCs	Proportion of Variance	Cumulated Proportion	PCs	Proportion of Variance	Cumulated Proportion
PC1	0.1295	0.130	PC36	0.0046	0.721
PC2	0.0939	0.223	PC37	0.0044	0.726
PC3	0.0825	0.306	PC38	0.0042	0.730
PC4	0.0429	0.349	PC39	0.0040	0.734
PC5	0.0357	0.385	PC40	0.0038	0.738
PC6	0.0300	0.415	PC41	0.0038	0.741
PC7	0.0251	0.440	PC42	0.0037	0.745
PC8	0.0221	0.462	PC43	0.0036	0.749
PC9	0.0204	0.482	PC44	0.0035	0.752
PC10	0.0182	0.500	PC45	0.0035	0.756
PC11	0.0177	0.518	PC46	0.0034	0.759
PC12	0.0150	0.533	PC47	0.0034	0.763
PC13	0.0131	0.545	PC48	0.0031	0.766
PC14	0.0127	0.559	PC49	0.0031	0.769
PC15	0.0127	0.572	PC50	0.0031	0.772
PC16	0.0117	0.583	PC51	0.0030	0.775
PC17	0.0107	0.594	PC52	0.0030	0.778

PCs	Proportion of Variance	Cumulated Proportion	PCs	Proportion of Variance	Cumulated Proportion
PC18	0.0106	0.604	PC53	0.0029	0.781
PC19	0.0095	0.614	PC54	0.0028	0.784
PC20	0.0094	0.623	PC55	0.0028	0.786
PC21	0.0087	0.632	PC56	0.0028	0.789
PC22	0.0081	0.640	PC57	0.0027	0.791
PC23	0.0076	0.648	PC58	0.0027	0.795
PC24	0.0074	0.655	PC59	0.0026	0.797
PC25	0.0067	0.662	PC60	0.0026	0.800
PC26	0.0063	0.668	PC61	0.0026	0.800
PC27	0.0062	0.674	PC62	0.0026	0.800
PC28	0.0061	0.680	PC63	0.0026	0.800
PC29	0.0059	0.686	PC64	0.0026	0.800
PC30	0.0056	0.692	PC65	0.0026	0.800
PC31	0.0053	0.697	PC66	0.0026	0.800
PC32	0.0052	0.702	PC67	0.0026	0.800
PC33	0.0050	0.707	PC68	0.0026	0.800
PC34	0.0048	0.712	PC69	0.0026	0.800
PC35	0.0047	0.717	PC70	0.0022	0.823

Table 1 represents the proportion of variance explained and cumulative variance proportion for the principal components. The second column of the table represents the proportion of variance which is the ratio of each Eigenvalue to the total. As it can be seen from the third column which represents the cumulated proportion on of variance, the first principal component explained 13% of the total data variance and the second PC which contains 0.9% of the variance in the data explains the second largest variance and so on. The first 60 PCs in total accounted for 80% of the total variance. In this study the first 60 to 69 PCs are considered sufficient for data representation since they represent 80% of the variance in the data.

For the Neighbourhood Components Analysis algorithm, the dimensionality of the reduced representation which is the number of rows in the matrix A has to be set by the user which is a deficiency of the approach.

4.2.1. Visualizing the Projected Distribution

The algorithms were applied to the Parkinson's Disease Classification dataset and projected the data into a 2-dimensional space for visualization.

Visualization results of PCA applied on the dataset and projected into 2-dimensional embedding.
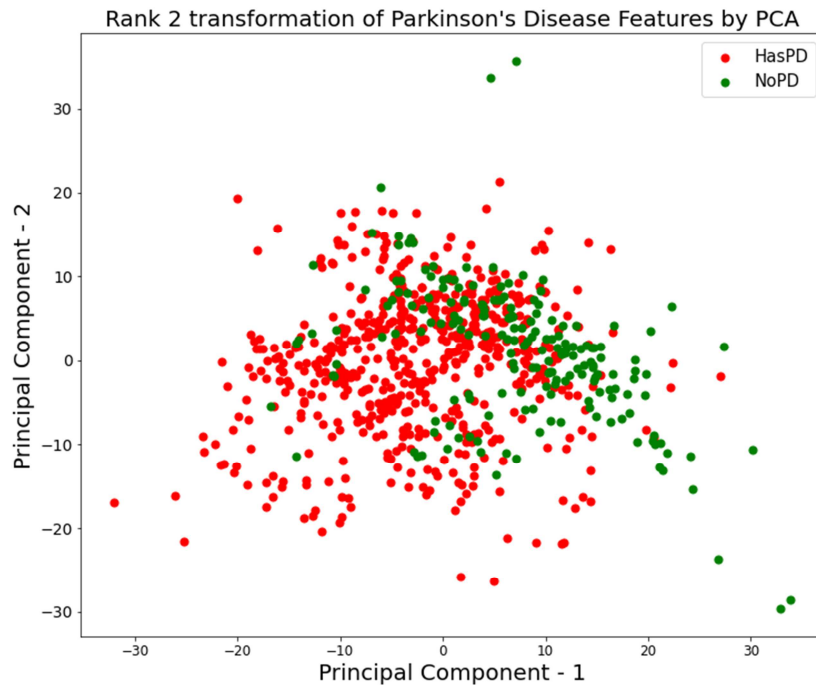

Figure 2. Dataset visualization results of PCA embedding. The data was reduced from its original dimensions D=754 and projected to d=2.

The data was successfully projected to a 2-dimensional feature space by PCA method. It can be observed from Figure 2. that the subjects in the first two principal components space overlap on each other and have no clear clustering for each group.

The NCA model is trained with a projection down to d = 2 dimensions to allow for visualization of the projected data.
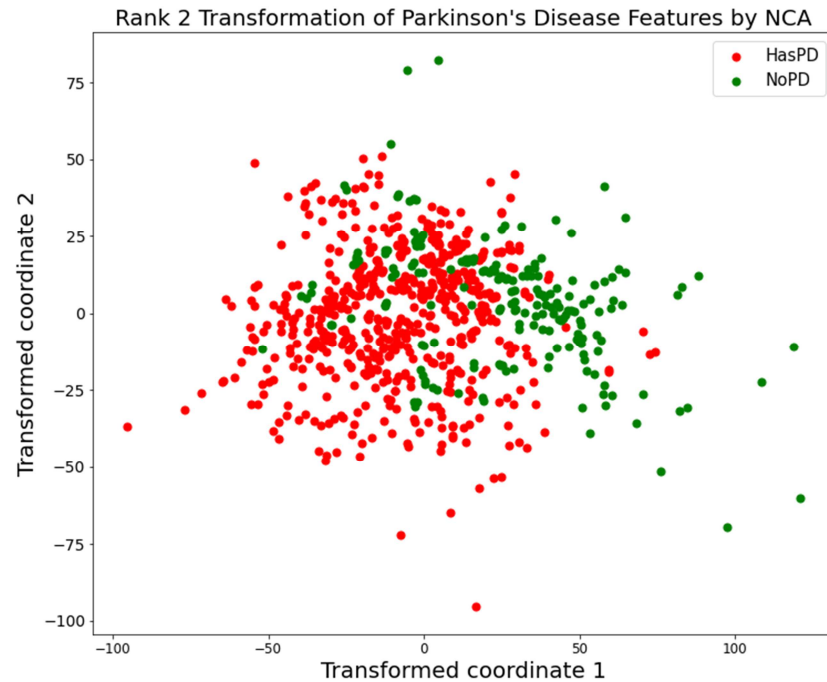
Figure 3. Dataset visualization results of NCA embedding. The data was reduced from its original dimensions D=754 to d=2.

Figure 3 shows 2 dimensional projections of the Parkinson's Disease Classification dataset obtained by NCA embedding. The projected representation of the data in 2 dimensions is well represented by NCA in terms of visualization compared to PCA. The classes are consistently much better separated by the NCA transformation. It can be observed that the two classes "haspd" (Has Parkinson's disease) and "nopd" (Has no Parkinson's disease) which is the control group, when projected to a 2-dimensional space, can be clearly separated. NCA appears to project data in such a way as to keep points from the same class close together. The clustering enforced by NCA is visually meaningful despite the large reduction in dimension. Other observations can be that the "haspd" class is spread out as compared to the control class.

4.2.2. Model Comparison on the Parkinson's Disease Classification Dataset

The dataset was trained with 605 samples (80%) and tested with 151 samples (20%) and the embeddings were evaluated on the test set using KNN with k=3.

The data was first projected into a 2-dimensional space by PCA and NCA methods for visualization.
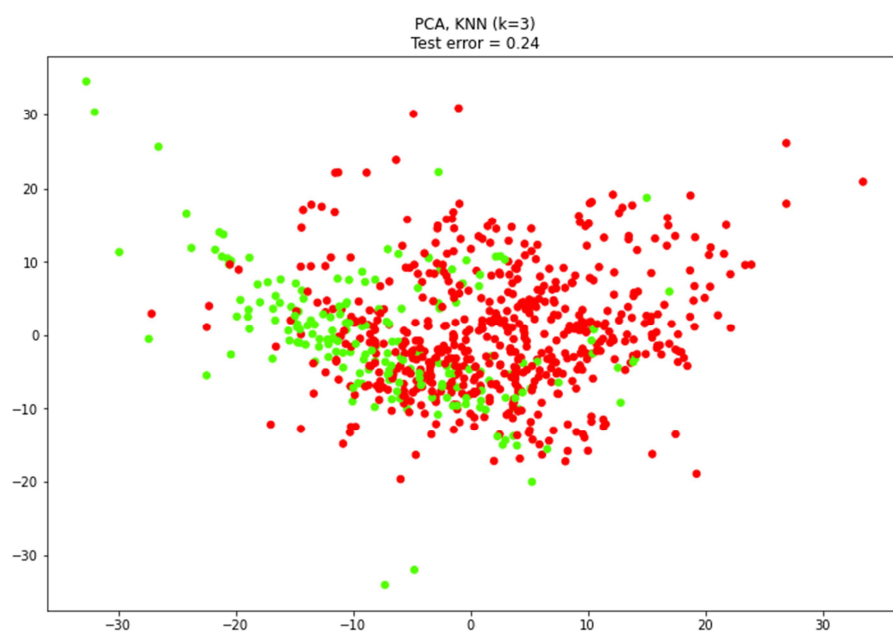

Figure 4. 2-dimensional PCA embedding on the test data.

It can be observed that from Figure 4. KNN on 2-dimensional PCA embeddings achieved an error of 0.24 which is slightly higher compared to that of NCA. There is no class separability acieved by PCA embedding on the test data. Rather the test points from the two classes overlap on each other.
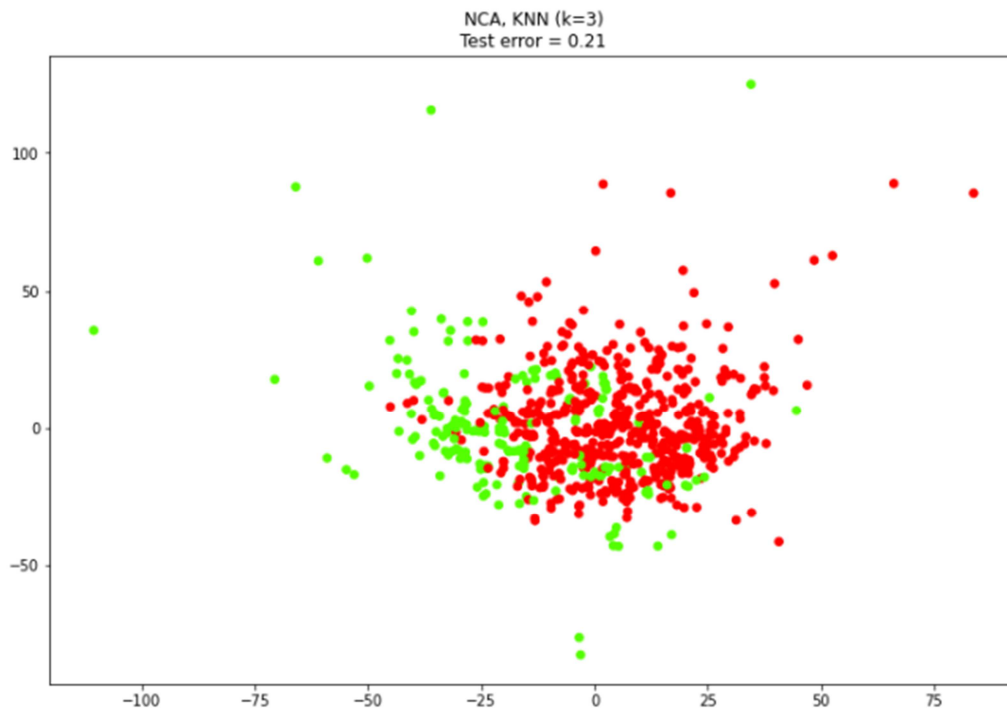

Figure 5. *2-dimensional NCA embedding on the test data.*

From Figure 5, KNN on 2-dimensional NCA embeddings achieved an error of 0.21. NCA embedding on test data achieves more class separability compared to the PCA embedding. The class separability can reveal how easy the dataset can be separated. This enables KNN classifier obtain less classification error on the embedded data.

The algorithms were used to reduce the dataset from its original dimensions 754 to d = 2, d = 5, d = 10, d = 15, d = 20, d = 25, d = 30, d = 35, d = 40, 45, d=50, d=55, d= 60, d=65 and d = 70 respectively in order to obtain optimum dimensions that the dataset can be reduced to without loss of information.

Table 1 represents the classification errors (%) across a range of projected dimensions on the Parkinson's Disease Classification dataset. The number of principal components considered for this analysis ranged from 2 PCs to 70 PCs.
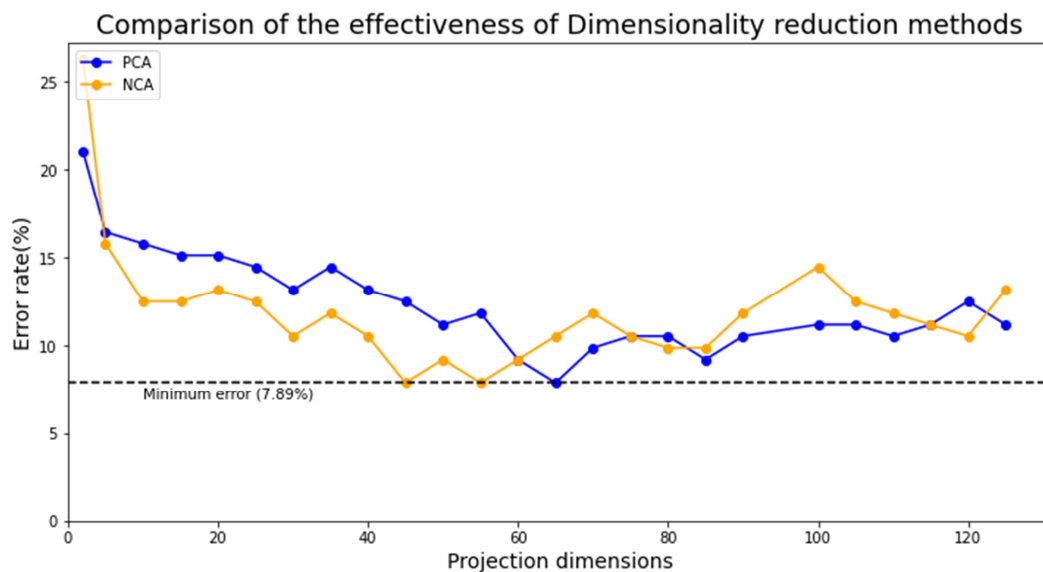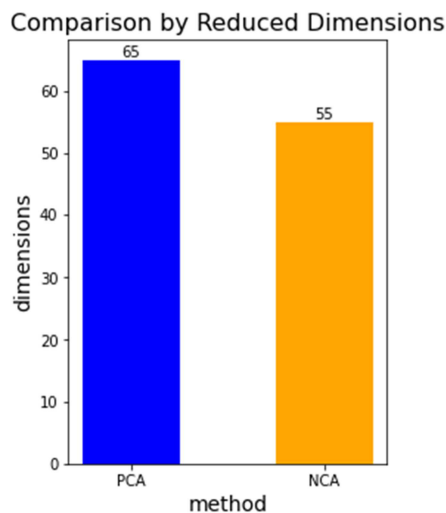

Figure 6. *KNN classification error on UCI Parkinson's Disease Classification dataset, across a range of projected dimensions.*

Table 1. *Classification Error (%) on various reduced dimensions with k=3.*

Projected dimension (d)	Classification Error (%) on PCA Embeddings	Classification Error (%) on NCA Embeddings
2	21.05	26.32
5	16.45	15.79
10	15.79	12.50
15	15.13	12.50
20	15.13	13.16
25	14.47	12.50
30	13.16	10.53
35	14.47	11.84
40	13.16	10.53
45	12.50	7.89
50	11.18	9.21
55	11.84	7.89
60	9.21	9.21
65	7.89	10.53
70	9.87	11.84

It can be observed from Table 2 that KNN on NCA embeddings attained the lowest classification error across the range of projected dimensions and consistently outperformed PCA.

Figure 7. *Comparison by reduced dimensions.*

From Figure 6. The optimal classification error obtained by the methods was 7.89%. However, the methods achieved the optimum error at different dimensions. The trend was as the number of dimensions increased, the classification error reduced up to an optimum number of dimensions and then as the number of dimensions increased, the classifier performed poorly (classification error increased). The optimum classification error shows the number of dimensions the data can be reduced to for a classifier to perform well.

Based on Figure 7, it can be observed that there was a significant reduction in the dimensions of the dataset from 754 to 55 by the NCA technique. The PCA method reduced the dataset to 65 dimensions retaining 80% of the information in the data. This shows that the high dimensional dataset can be represented in 55 dimensions. This is the number of dimensions the KNN classifier obtained optimum results on the embedded dataset.

5. Conclusion

Linear dimensionality reduction methods are widely used because they are not prone to overfitting and preserve the topology of data. We have investigated Neighborhood Components Analysis, a novel approach for dimensionality reduction of data which is based on the nearest-neighbor model and compared it with Principal Components Analysis which is a well-known standard approach. Neighbourhood Component Analysis (NCA) algorithm has been applied as a dimensionality reduction criterion for selecting relevant features in high-dimensional data. For a researcher, it is important to consider the most relevant features for either exploratory analysis or machine learning in their analysis to achieve high accuracy results. This can be achieved by employing a dimensionality reduction procedure on the data.

As per the results from this study, the Neighborhood Components Analysis model gives better performance in dimensionality reduction even in presence of outliers and in terms of visualization of class separation in low dimensional space compared to PCA. The classification error of the transformed features by NCA is lower than that of PCA. This signifies that the variance captured by the PCs is not certainly an important indicator of classification performance. Our experiments on Parkinson's Disease Classification dataset shows that by using reduced dimensions of the data, learning can be made easier. NCA is not affected as much by the outlier points since they contribute less to the labelling of the surrounding points during learning of low dimensional projection. PCA was highly affected by the outliers. It therefore can be concluded from the above experimental results that NCA is able to achieve a better dimensionality reduction than PCA. Future work should consider robust variants of NCA to improve its performance in presence of outliers.

References

[1] W. Yang, K. Wang and W. Zuo, "Neighborhood component feature selection for high-dimensional data.," *JCP,* vol. 7, p. 161–168, 2012.

[2] D. M. Hawkins, Identification of outliers, vol. 11, Springer, 1980.

[3] C. Qin, S. Song and G. Huang, "Non-linear neighborhood component analysis based on constructive neural networks," in *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2014.

[4] A. Datta, S. Ghosh and A. Ghosh, "PCA, kernel PCA and dimensionality reduction in hyperspectral images," in *Advances in Principal Component Analysis*, Springer, 2018, p. 19–46.

[5] X. Wu, X. Zhu, G.-Q. Wu and W. Ding, "Data mining with big data," *IEEE transactions on knowledge and data engineering,* vol. 26, p. 97–107, 2013.

[6] J. Fan, F. Han and H. Liu, "Challenges of big data analysis," *National science review,* vol. 1, p. 293–314, 2014.

[7] O. Shetta and M. Niranjan, "Robust subspace methods for outlier detection in genomic data circumvents the curse of dimensionality," *Royal Society open science,* vol. 7, p. 190714, 2020.

[8] S. Roweis, G. Hinton and R. Salakhutdinov, "Neighbourhood component analysis," *Adv. Neural Inf. Process. Syst.(NIPS),* vol. 17, p. 513–520, 2004.

[9] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science,* vol. 2, p. 559–572, 1901.

[10] H. Hotelling, "Analysis of a complex of statistical variables into principal components.," *Journal of educational psychology,* vol. 24, p. 417, 1933.

[11] I. T. Jolliffe, "Principal components in regression analysis," *Principal component analysis,* p. 167–198, 2002.

[12] W. Astuti and others, "Support vector machine and principal component analysis for microarray data classification," in *Journal of Physics: Conference Series*, 2018.

[13] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna, R. Kaluri, D. S. Rajput, G. Srivastava and T. Baker, "Analysis of dimensionality reduction techniques on big data," *IEEE Access,* vol. 8, p. 54776–54788, 2020.

[14] N. Singh-Miller, M. Collins and T. J. Hazen, "Dimensionality reduction for speech recognition using neighborhood components analysis," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[15] N. Singh-Miller, "Neighborhood analysis methods in acoustic modeling for automatic speech recognition," 2010.

[16] J. Manit and P. Youngkong, "Neighborhood components analysis in sEMG signal dimensionality reduction for gait phase pattern recognition," in *7th International Conference on Broadband Communications and Biomedical Applications*, 2011.

[17] M. Rizwan and D. V. Anderson, "Speaker similarity score based fast phoneme classification by using neighborhood components analysis," in *2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2016.

[18] H. Ferdinando, T. Seppänen and E. Alasaarela, "Emotion recognition using neighborhood components analysis and ECG/HRV-based features," in *International Conference on Pattern Recognition Applications and Methods*, 2017.

[19] H. Ferdinando and E. Alasaarela, "Enhancement of emotion recogniton using feature fusion and the neighborhood components analysis," 2018.

[20] G. R. Naik, Advances in Principal Component Analysis: Research and Development, Springer, 2017.

[21] C. O. Sakar, G. Serbes, A. Gunduz, H. C. Tunc, H. Nizam, B. E. Sakar, M. Tutuncu, T. Aydin, M. E. Isenkul and H. Apaydin, "A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform," *Applied Soft Computing,* vol. 74, p. 255–263, 2019.

[22] V. Fritsch, G. Varoquaux, B. Thyreau, J.-B. Poline and B. Thirion, "Detecting outlying subjects in high-dimensional neuroimaging datasets with regularized minimum covariance determinant," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2011.