Periodicals of Engineering and Natural Sciences ISSN 2303-4521
Vol. 8, No. 1, March 2020, pp.142-155

Using nonlinear dimensionality reduction techniques in big data analysis

Omar Adel abd Alwahab [1], Mohammed Sadiq Abd Alrazak[2]
[1] Statistic Department, College of Administration and Economics, University of Diyala
[2] Statistic Department, College of Administration and Economics, University of Baghdad

ABSTRACT

In recent years, the huge development in the measure of data has been noted. This becomes a first step of the big data. Big data can be defined as high volume, velocity and variety of data that require a new high-performance processing. The reduction of dimensions is one of the most important methods that are necessary in the field of big data analysis. Ideally, the reduced representation should have a dimensionality that corresponds to the intrinsic dimensionality of the data. There are two important procedures, the first one is dimensions reduction, and the second one is putting the data into its model and then estimating it. In this paper, we use two techniques of nonlinear dimensional reduction. The first one includes kernel principal components analysis (KernelPCA) along with modified kernel principal components analysis smooth (KernelPCAS) and the second one is neural network. The mean square error (MSE) was used to demonstrate the effectiveness of nonlinear reduction methods in the analysis of big data as a resourceful tool for this purpose.

Keywords: Nonlinear dimensions reduction, Kernel principal components analysis, ANN, big data

Corresponding Author:

Omar Adel Abd Alwahab
Teacher Assistant
Statistic Department ,College of Administration and Economics
University of Diyala, Iraq
E-mail: Omeradil@ecomang.uodiyala.edu.iq

1. Introduction

The progress in the technology and informatics in the present time has a significant impact in the development of medical, natural and human sciences. It was reflected clearly on the science of statistics and its close association with it, which led to this development to increase the variables that describe these cases [1-2]. Here, the role of statistics is to link between this development in information technology and the rest of science in terms of the study of phenomena and the collection of information as well as their related data. Then, they can be classified statistically and analyzed in a manner that fits the phenomenon in question and writes findings to make a decision. When the causal relationship between the random variables is known, the appropriate statistical analysis is called the parametric-analysis, where the parameters provide a brief summary of the observations that facilitate statistical inference, either in the case of the instability of these conditions or not knowing the relationship between random variables in terms of causal or behavioral relationship. The appropriate analysis is called non-parametric analysis. There is no information about the characteristics (or indications) of the phenomenon. The abundance of information and data can be obtained from the phenomenon under study in light of the development of data storage. These are in the form of databases are stored in many warehouses and data warehouses [3-5]. Here, it comes the idea of big data [6, 7, 11]. Big data can be defined as high volume, velocity and variety of data that require a new high-performance

processing, which in turn requires analysis. This classification is not sufficient if we want to analyze these data. Therefore, we will face a greater problem in the order of such data if the data are considered preliminary and therefore need to be classified. After the classification of this data, we can have a very large number of special variables. This phenomenon is under study and therefore we will face another problem in the analysis of these data because it contains reliability between the variables. It is also be difficult to determine the explanatory variables and the dependent variable. After determining above cases, the explanatory variables are linked to each other because of the large number of variables, which will generate a problem of curse of dimensionality. This causes a difficulty to find the relationship between them and the depend variable and thus get inaccurate parameters and then get incorrect estimates.The 'curse of dimensionality' is one of the difficult problems faced by researchers in data analysis, so it was necessary to work to solve the problem in order to obtain accurate results. One of the most important methods used to solve this problem is the reduction of dimensions. This method reduces the number of variables in the data under study without losing a number of variables, which is to maintain the content of information. After the reduction of high dimensions, it is worth mentioning the type of relationship between these variables as linear relationship or non-linear relationship. Here, the role of nonlinear models comes in the interpretation of the relationship between the dimensions of the reduced and depended variable. It is known that the linear models are a special case of non-linear models and are the best in the interpretation of the relationship between explanatory variables and the dependent variable. Machine learning is one of newest technique which can be defined as a field of computer science that often uses statistical techniques to give computers the ability to "learn" with data, without being explicitly programmed. Machine learning considers the modern technique that dealing with non-linear models to deal with difficult and complex models. The aim of this work is to reduce the high dimensions to low dimensions. Also, the paper intends to estimate that dimensions, and then finding the best nonlinear relationship between the explanatory variables and depending variable in high-dimensional problem through using some styles of nonlinear reduction methods and using of methods estimating of non-linear reduction.

2. Theoretical aspect

Linear regression is a powerful technique for analyzing data described by models, which are linear in the parameters. However, the researchers have an expression which connected with response variable to the predictor variables, and these models are sometimes nonlinear in the parameters. When linear regression techniques extended, appear to greater need for a regression model that provides a good explanation of the phenomenon under study. In spite of the wide use of linear models, they fail to explain the relationship between the explanatory variables and the dependent variable if the relationship is nonlinear. Therefore, there is a need to have models that have the necessary flexibility to find the shape of this relationship, where non-linear models are the most efficient in explaining the relationship between the explanatory variables and dependent variable. A model can be nonlinear in its parameters, nonlinear in its observed variables, or nonlinear in both its parameters and variables. Nonlinear in the parameters means that the mathematical relationship between the variables and parameters is not required to have a linear form. From another hand, a linear model is a special case of a nonlinear model [2, 3].

2.1 Nonlinear regression

The functional relationship between the explanatory variables and dependent variable can be expressed in through a general linear model as [6, 11]:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \ , i = 1,2 \dots, n \qquad (1)$$

Where,

$y_{i:}$ Dependent variable.

$x_{i1}, x_{i2}, \dots, x_{ip}$: Explanatory variables.

$\beta_0, \beta_1, \beta_2, \ldots, \beta_p$: Parameters of Linear Regression Model.

ε_i: Error term.

A nonlinear regression model can be written:

$$Y_i = f(X_n, \beta) + \varepsilon_i \quad , i = 1,2 \ldots, n \qquad \ldots \qquad (2)$$

Where, f is the expectation function and x_n is a vector of associated regresses variables or independent variables for the n^{th} case. This model is of exactly the same form as (2) except that the expected responses are nonlinear functions of the parameters. That is, for nonlinear models, at least one of the derivatives of the expectation function with respect to the parameters dependents on at least one of the parameters. There are many common nonlinear models use nonlinear regression model based on exponential regression model. When there is one predictor variable, one regression model with error is:

$$Yi = \beta_0 \exp(\beta_1 X_i) + \varepsilon_i \qquad \ldots \qquad (3)$$

Where,
β_0 and β_1 are parameters.
X_i are known constants.
ε_i are independent $N(0, \sigma^2)$
The response function for this model is:

$$f(X, \beta) = \beta_0 \exp(\beta_1 X) \qquad (4)$$

Note that this model is not linear in the parameters β_0 and β_1 .
Another important nonlinear regression model is the logistic regression model. This model with one explanatory variable and normal error terms is:

$$Yi = \frac{\beta 0}{1+\beta 1 \, exp(\gamma 2 Xi)} + \varepsilon_i \qquad (5)$$

The response function for this model is:

$$f(X, \beta) = \frac{\beta 0}{1+\beta 1 \, exp(\beta 2 Xi)} + \varepsilon_i \qquad (6)$$

2.2 Dimensionality reduction for Big Data

Direct reduction of dimensions is a very common technique. In addition to facilitating visualization, there is often a need to address data storage issues. Dimension reduction often involves the assumption of differentiation. Depending on the application techniques for reducing linear dimensions may need to be extended to non-linear techniques, if the data structure cannot be captured in a linear subspace. However, they are more difficult to calculate parameters. There is a new way to reduce the dimensions of large and complex data sets by using an illustrative tool to analyze data. Increased form information forces a low-dimensional model structure. In turn, makes inference more complex, the interpretation of models easier, and leads to greater durability against noise [2, 7].

Regularization techniques that enforce dispersion have been study widely for last years, including the lasso, adaptive lasso, the smoothly clipped absolute deviation penalty and cholesky decomposition modification. These newer methods and accessories are now essential tools in dealing with high-dimensional data. In high-dimensional data analysis, spaces between "points", or observations, are needed for many analytical procedures. For example, optimization methods usually move into a complex parameter space, and an idea of the step size for this movement is needed.

2.3 Methods of non-linear dimensionality reduction for big data

Basically, dimension reduction refers to the process of converting a set of data. That data needs to having vast dimensions into data with lesser dimensions. Also, it needs to ensure that it conveys similar information

concisely. Although, we use these techniques to solve curse of the dimensionality problems. Many natural phenomena behave in a nonlinear way meaning that the observed data describe a curve or curved subspace in the original data space. Identifying such nonlinear manifolds becomes more and more important in the field of molecular biology. In general, molecular data are of very high dimensionality because of thousands of molecules that are simultaneously measured at a time. Since the data are usually located within a low-dimensional subspace. They can be well described by a single or low number of components. Experimental time course data are usually located within a curved subspace which requires a nonlinear dimensionality reduction [4].

2.3.1 Kernel PCA Methods

In recent years, there has been an explosion of work on kernel methods. For supervised learning these include support vector machines, Scholkopf, Smola, and Muller (1998) used this trick to define kernel PCA ,and can be derived using the known fact that PCA can be carried out on the dot product matrix instead of the covariance matrix, Let $\{x_1, ..., x_N\}$ be a set of training data in the input space $\chi = R^d$ [1].

Kernel PCA performs the traditional linear PCA in the feature space corresponding to the kernel k (.,.). Analogous to linear PCA, it involves the following Eigen decomposition [9]:

$$HKH = U\Sigma U^T \quad \ldots\ldots \qquad (7)$$

Where, K is the kernel matrix with entries $K_{ij} = k(x_i, x_j)$

U is orthogonal Eigen vector matrix

Σ covariance matrix

$$H = I - \frac{1}{N}11^T \qquad (8)$$

Where, H is the cantering matrix

I is the $N \times N$ identity matrix, $1 = [11...1]^T$ is an $N \times 1$ vector, $U = [a_1, ..., a_N]$ with $a_i = [a_{i1}, ..., a_{iN}]^T$ is the matrix containing the eigenvectors and $\Sigma = diag(\lambda_1, ..., \lambda_N)$ contains the corresponding eigenvalues. Define the "centered" map $\tilde{\varphi}$ as:

$\Phi : X \rightarrow F$

$$\varphi(x) = X^*X^T \qquad (9)$$

And

$$\tilde{\varphi}(x_i) = \varphi(x) - \bar{\varphi} \qquad (10)$$

Where,

$$\bar{\varphi} = \frac{1}{N}\sum_{i=1}^{N} \varphi(xi) \qquad (11)$$

The k^{th} orthogonal eigenvector of the covariance matrix in the feature space can then be shown

$$V_k = \sum_{i=1}^{N} \frac{a_{ki}}{\sqrt{\lambda_k}} \tilde{\varphi}(x_i) \qquad (12)$$

Denote the projection of the φ point x onto the k^{th} component by β_k.

Then

$$\beta_k = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{N} a_{ki}\, \tilde{K}(x,xi) \qquad (13)$$

And

$$k(xi, xj) = e^{-d^2(xi,xj)/2\sigma 2} \qquad (14)$$

Where, $d^2(x_i, x_j)$ is a distance measure in the input space.

Then projection of φ(x) onto the subspace spanned by the first n eigenvectors is given by

$$y = \sum_{k=1}^{N} B_k V_k + \bar{\varphi} \quad (15)$$

2.3.2 Kernel PCA smooth method (KPCAS)

A new develops of Kernel PCA method are modified. It relies on smooth of genuine dimension from Kernel PCA, and the following are the steps of algorithm modified method [9, 10]:

1- Decomposition metric *X* by using equation (7) in basic Kernel PCA.
2- Finding the k^{th} orthogonal eigenvector of the covariance matrix by using equation (12) in basic Kernel PCA.
3- Using equations (13), (14) and (15) to compute genuine dimension.
4- Estimating the genuine dimension X_g by smooth of local regression using weighted linear least squares and a 1st degree polynomial model.

2.3.3 Artificial Neural Networks (ANN)

Neural network (ANN) is an information processing system that has certain performance characteristics in a manner that simulates biological neural networks [8, 11]. In other words, artificial neural networks are a simulation of the way in which the human mind performs a particular task. Neural networks (NN) are used for systems that recognize an activity or systems that predict phenomena within a given framework, or control some software. Artificial neural networks (ANN) are learnable systems through examples, and artificial simple processing units, which store the practical information to make it available to the user by adjusting the weights. Therefore, artificial neural networks are installations of the parallel processing, which dependents mainly on the processing component capable of working as local memory with processing operations. It has a single output as it is dependent on the input values as well as the values stored in the local memory of these computational elements. The use of neural networks changes with the problem faced by the network, which gave the networks a high flexibility in the face of problems in various sciences. In 1959, by Bernard Fedro and Marcian Hoof, a fairly complex mathematical algorithm is presented capable of solving all numerical problems that are not subject to fixed mathematical laws. It is a way to reduce dimensions and work to overcome the problem of dimension, often consisting of neural network of a number of neurons internally linked between them. However, the quality of the linkage for internal neurons in addition to its nature are determined by the architecture and type of network and types of artificial neural networks: Feed forward networks, Back propagation networks and Recurrent networks

Back propagation is one of the methods of teaching neural networks that provide the transmission of information by the reverse propagation of the original direction of the arrival of information. This method is based on the Observer learning. In the training stage, you need special data to learn the network, when neural networks provide with input data with the desired output data, the network then feeds forward the input data to get the output value. The network then compares the calculated output with the desired output [5].

If the results do not match, the network calculates the difference between them for each output layer neuron which represents the error value, the propagation stage and the network recalculates the error value in each of the hidden networks. Finally the stage of updating the value of weights and the network recalculates each weight is compensated by the calculated new values. In the back propagation, the activation functions used by the neurons must be derived, because in the modernization of the weights, the derived function of the activation function is used to calculate the new values, the learning stages on which the network dependents can be divided into two stages:

1- Front-feeding stage to train inputs

The foreground stage starts and receives as the unit (Xt) the input signal, and then this signal is transmitted to the unit or units of the hidden layer according to their number. The hidden layer adds its weighted input values and signal by weight according to the equation [8]:

$$Y = \sum_{i=1}^{n} X_{ij} Y_{ij} \quad \forall j \quad (16)$$

And then calculate the logistic function by the equation:

$$\boldsymbol{Y = \frac{1}{1+e^{-Y}}} \quad (17)$$

Which is used to convert the data into linearity, thus the value obtained from the previous equations is transferred to the output layer unit.

2- Spread the rear of the phase error

After moving the values to the output layer whose value is calculated by the previous steps, a comparison is made between the calculated values and the desired values (error calculation) by the difference between the values of those outputs by the following error equation [8]:

$$\boldsymbol{E = (d_i - Y_i)} \quad (18)$$

Where,

$\boldsymbol{d_i}$: The desired output represents the response variable.

$\boldsymbol{Y_i}$: The output value of the network represents the estimated response variable.

E: the error.

The weight is then corrected and adjusted through the learning process done on the network. The rear emission method is summarized as follows:

1-Start with weights and (Offsets), give the weights and node (Offsets) a few random values.

2-Initialize the input and describe the desired output, prepare the continuous values of the input vector $X_0,X_1,X_2, \dots , X_{n-1}$ (Customize) desired outputs $d_0,d_1,d_2, \dots , d_{m-1}$

3-To calculate the real output, the Sigmoid Function was used to calculate the output $Y_0,Y_1,Y_2, \dots , Y_{M-1}$.

4-Adjust the weights.

The algorithm begins by inserting in the output nodes and works backwards into a hidden layer and adjusts the weights by:

$$W_{ij}(t+1)=W_{ij}(t)+ \eta \delta_i \bar{x}_I \quad (19)$$

Where,

$W_{ij}(t)$: Weight in hidden node (i) (or input to node (j) in time (t)).

δ_i: Error term for node (j).

If j represents an output node:

$$\delta_i = Y_i (1 - Y_i) (d_i - Y_i) \quad (20)$$

Where,

d_i represents the desired output of node (j).

Y_i represents true output.

If it is an internal hidden node,

$$\delta_i = \overline{X}_i (1- \overline{X}_i) \sum_K \delta_k W_{ik} \quad (21)$$

Where,

k is related to all nodes in layers above node (j) [5].

The threshold value of the internal nodes is adjusted in a similar way by assuming that they are correlated weights and correlated from fixed-value entries. For fast convergence, the term Momentum is added and is denoted by α . It helps to change the weights regularly as follows:

$$W_{ij}(t+1) = W_{ij}(t) + \eta \delta_i \bar{x}_I + \alpha (W_{ij}(t) - W_{ij}(t-1)) \quad (22)$$

Where,

$$0 < \alpha < 1$$

And η is learning rate within $0 < \eta < 1$

In this paper, the neural network was used to reduce the nonlinear high dimensions by making the number of layers within the network and the number of nodes within each layer is equal to the number of explanatory variables in the neural network.

The reduction requirement is that the number of nodes within each hidden layer is less than the number of variables entering the network and thus the reduction is done, as the reduction is done within each node so that no information of the variables is lost.

3. Application

In this research, the theoretical aspect was applied to the Ministry of Planning data based on information from the Central Statistical Organization (CSO) from field statistical surveys to reveal the living reality of the individual and the Iraqi family. The sample of the survey included (25488) households in all Iraqi governorates. This survey aims to provide indicators at the national level and at the provincial and district levels and for both urban and rural environments, in terms of providing the necessary information to understand the causes of poverty among Iraqi families. This survey also provides comprehensive and up-to-date information on a number of standard indicators. The survey consists of a number of components, some of which are repeated periodically and some of which change depending on the country's need for data. Preparation of the statistical report for this survey has been in partnership with the Central Statistical Organization in Iraq and with the support of the United Nations organizations in Iraq. This preparation is to provide the relevant authorities and decision makers with full data to be able to study the progress made in Iraq through the results of the survey in this concern. The field researchers completed forms about (25488) Iraqi families from all governorates of Iraq, which represents the size of the sample studied for this survey. The data recorded in this form stands for the scientific material for this research. The survey forms consist of four parts based on manpower, consumer spending, basic services and food security.

3.1 Data Description

About 150 variables and 1236 observation form (IHSES-II) 2012 were used, which is the final form of this survey and approved by the Central Statistical Organization (CSO). The explanatory variable in the research is depicted in Appendix (A) and the dependent variable is based on total Expenditure paid prices.

3.2 Applied methods

Nonlinear dimension reduction methods were applied to big data. MSE criterion was used to determine the effectiveness of these methods in analyzing big data.

Table 1. Kernel PCA Method

No. of component	MSE				
	1	2	3	4	5
KernelPCA	52.9487	52.9480	52.9466	52.9467	52.9480

In Table 1, we notice that the third component has the lowest mean quadratic error of five main components. However, this value is very high. Figure1 shows the results of KernelPCA method.

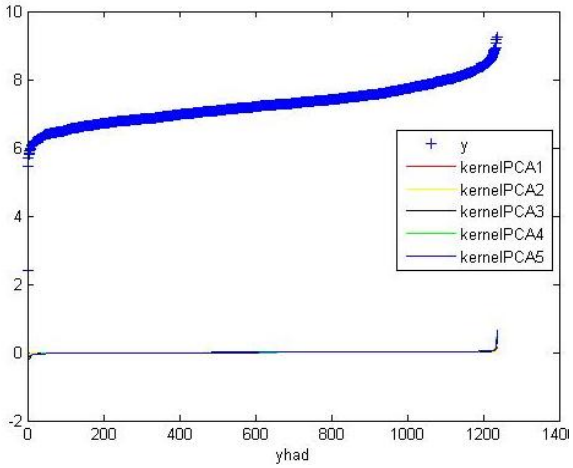

Figure 1. KPCA results

Table 2. Kernel PCA smooth method

Tuning parameter	MSE				
	0.01	**0.02**	**0.03**	**0.04**	**0.05**
KernelPCAS	0.2768	0.3018	0.3104	0.3141	0.3161

In Table 2, we observe that the proposed KPCAS with tuning parameter of the local regression using weighted linear least squares at 1st degree polynomial model has the lowest mean squared error in (0.01), however, all the tuning parameters are effective in reducing nonlinear high dimensions. Figure 2 shows the results of KernelPCAsmooth method.

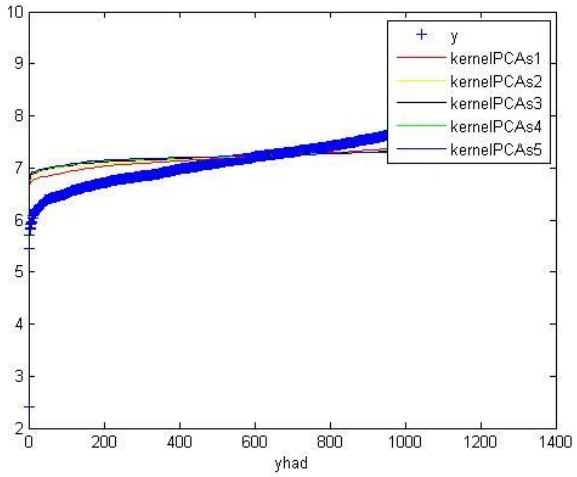

Figure 2. Kernel PCA smooth results

Table 3. ANN Method

No. of nods	MSE				
	5	**10**	**15**	**20**	**25**
NN	0.0365	0.0037	0.0121	0.0100	0.0145

In Table 3, we notice that the number of nodes in the ANN has an effect on the network output. The lowest mean square error was at the second node. However, all the nodes are considered effective in reducing nonlinear high dimensions. Figure 3 shows the results of ANN method.

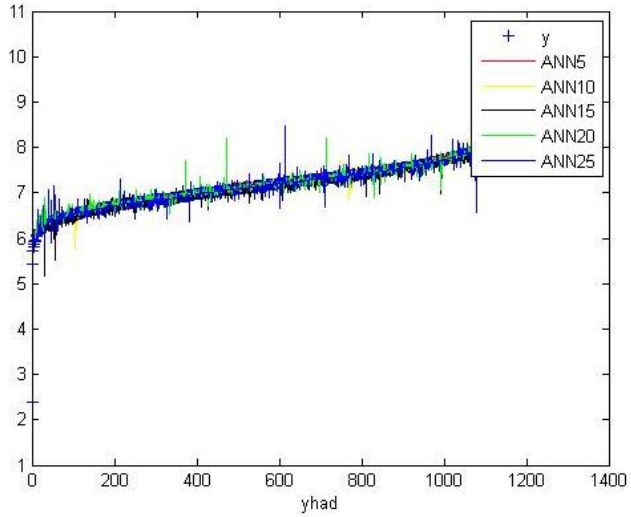

Figure 3. ANN results

Table 4. Methods comparison

	methods	MSE
1	KernelPCA	52.9487
2	KernelPCAsmooth	0.2768
3	ANN	0.0037

The results of the methods comparison in Table 4 indicate that the method (KernelPCA) has the largest (MSE) followed directly by the proposed method (KernelPCAsmooth), and the ANN has lowest (MSE). Figures 4-6 show the consequences of above methods.

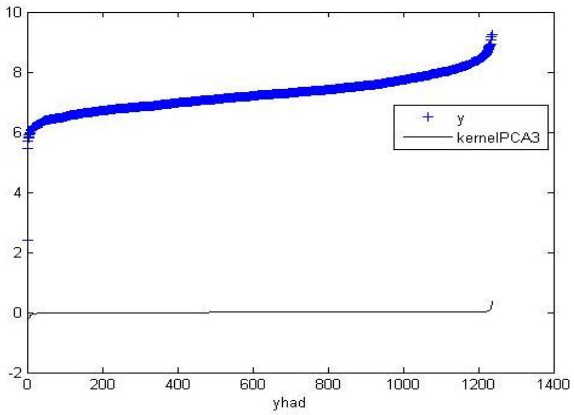

Figure 4. KPCA3 results

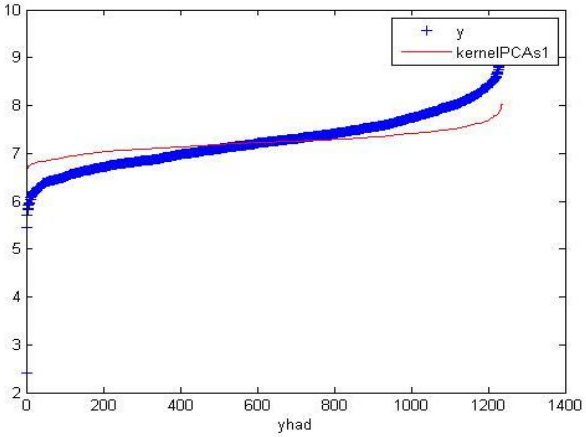

Figure 5. KPCAS1 results

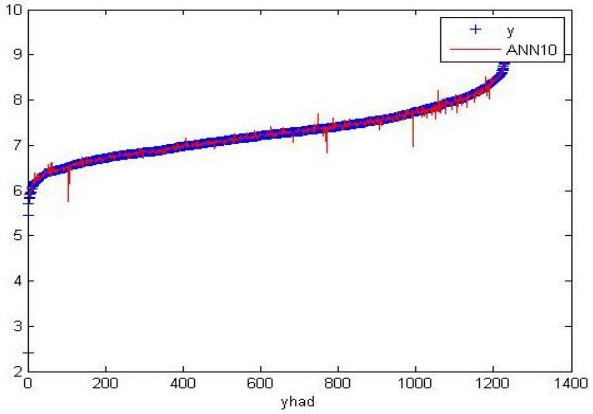

Figure 6. ANN10 results

Through the results of above comparison criterion, the most effective methods are shown in the analysis of big data through the dependent variable and explanatory variable. We note that the preference of the proposed method and the method of neural networks are over the traditional method through the reduction parameters of the explanatory variables and the formation of a new single axis with the dependent variable. The boot property used in the proposed method is for dealing with big data proved its effectiveness through the criterion (MSE), which depends on the booting of the new axis with the axis of the dependent variable. This boot property is used to study the effect of the new variable vector on the dependent variable.

4. Conclusions

1- From the above results, it is clear that the classic method of nonlinear reduction does not give good results when dealing with big data, because of data problems and the problem of curse high dimensions.
2-Through the result of the modified method kernelPCA smooth, we note its superiority over the classical method kernelPCA in nonlinear reduction. It gives it the advantage and ability to deal with big data when analyzing.
3- It is clear from the result of neural networks, the number of nodes within the neural network is equal to the number of explanatory variables outweigh the modified kernelPCA smooth method when dealing with big data. However, there is a problem of long time required for big data analyzing because of the property of rear emission error correction within the network.

5. Recommendations

Once analyzing data characterized by big data, characteristics should be avoided from the use of traditional methods of dimensional reduction because it cannot deal with the amount of large data and diversity in variables and processing speed. Consequently, nonlinear reduction methods should be used as kernrelPCAsmooth method. The results of the method of neural networks is worthy. In a kernelPCAsmooth manner, if time is important in the data analysis process, we recommend that you do not use it.

References

[1] K.I.C. Williams, “On a Connection between Kernel PCA and Metric Multidimensional Scaling", Machine Learning, vol.46, pp.11–19, 2002.
[2] P. K. Chandrakar and A. K. Shrivas, “An Analysis of Big Data Dimensionality Reduction Technique", Review of Business and Technology Research, vol. 14, no. 2, ISSN 1941-9414,2017.
[3] D. Erdman and M. Little, " Nonlinear Regression Analysis and Nonlinear Simulation Models", *Survey of SAS System Features, SAS Institute Inc., Cary, NC* , 1998.
[4] J. Fan, Q. Sun, W.-X. Zhou, “Principal component analysis for big data", arXiv:1801.01602, 2018.
[5] A. K. Ghosh and S. Bose, "Backfitting Neural Network", Computational Statistic , no.19, 2004.
[6] M. Kutner, e al., “Applied Linear Regression Models”, McGraw-hill Book Company. ISBN: 9780073014661.Publisher. Volume: Edition: 4,2004.
[7] J. A. Lee and M. Verleysen, “Nonlinear Dimensionality Reduction", Springer Verlag, New York, ,2007.
[8] O. S. Qasim and I. R. Mohammed, "Artificial neural networks in the suitability of the model for medical diagnosis", Journal of Rafidain for Computer Science and Mathematics, vol.10, Issue 1,2013 .
[9] Y. Rathi, S. Dambreville, and A. Tannenbaum, “Statistical Shape Analysis using Kernel PCA", Proceedings of SPIE, The International Society for Optical Engineering 6064,2006 ·
[10] L. Wasserman, "All of Nonparametric Statistics (Springer Texts in Statistics) ",Book, Springer-Verlag Berlin, Heidelberg, 2006.
[11] Y.-R. Yeh, S.-Y. Huang and Y.-J. Lee, "Nonlinear Dimension Reduction with Kernel Sliced Inverse Regression", IEEE Transactions on Knowledge and Data Engineering, vol.21, no.11, pp. 1590 - 1603 , 2007 .

Appendix (A)
150 variables names form (IHSES-II) 2012

No.	Variables name	Type of data
1	Size of the Household	Quantitative
2	Day visit 1	Quantitative
3	Month visit 1	Quantitative
4	Year visit 1	Quantitative
5	Field Staff visit 2	Quantitative
6	Day visit 2	Quantitative
7	Month visit 2	Quantitative
8	Year visit 2	Quantitative
9	Field Staff visit 2	Quantitative
10	Day visit 3	Quantitative
11	Month visit 3	Quantitative
12	Year visit 3	Quantitative
13	Field Staff visit 3	Quantitative
14	Number of times having	{1, NEVER}...
15	Compare economic situation	{1, MUCH WORSE}...
16	Number of: Electric washing	Quantitative
17	Number of: Electric generator	Quantitative
18	Number of: TV	Quantitative
19	Number of: Cooker (gas,	Quantitative
20	Number of: Water heater	Quantitative
21	Number of: Heater (kerosene,	Quantitative
22	Number of: Electric fan	Quantitative
23	Number of: Air condition	Quantitative
24	Number of: Frozen	Quantitative
25	Number of: Water cooler	Quantitative
26	Number of: Dishwasher	Quantitative
27	Number of: Electric vacuum	Quantitative
28	Number of: Personal computer	Quantitative
29	Number of: Play station	Quantitative
30	Number of: Satellite	Quantitative
31	Number of: IPAD	Quantitative
32	Number of: TOTAL	Quantitative
33	Version of Data Entry	Quantitative
34	Number of people in the ration	Quantitative
35	Type of household unit	{1, HOUSE}...
36	Total area of dwelling	Quantitative
37	Built area occupied by the	Quantitative
38	Total area of land	Quantitative
39	The water is sufficient	{1, YES}...
40	Type of toilet	{1, FLUSHED TOILET}...
41	Days/Week from Public	Quantitative
42	Days/Week from Community	Quantitative
43	Use electricity from	{1, yes}...
44	Ownership status of the	{1, OWNED BY THE HOUSEHOLD}...

No.	Variables name	Type of data
45	Problems of transportation	{1, YES}...
46	Number of household in	Quantitative
47	Number of other members	Quantitative
48	Years staying in this house	Quantitative
49	Type of household unit	{1, HOUSE}...
50	Principal material of the walls	{1, BRICK}...
51	Principal material of ceiling	{1, REINFORCED CONCRETE
52	Total area of land	Quantitative
53	Bedrooms exclusively use by	Quantitative
54	Hall exclusively use by this	Quantitative
55	Guest's room exclusively use	Quantitative
56	Dining room exclusively use	Quantitative
57	Other rooms exclusively use	Quantitative
58	Kitchen exclusively use by	Quantitative
59	Bathroom exclusively use by	Quantitative
60	Bathroom with utilities	Quantitative
61	Bedrooms jointly use with	Quantitative
62	Hall jointly use with	Quantitative
63	Frequency of the garbage	{1, DAILY}...
64	Main source of water	{1, PUBLIC NETWORK: HOUSING
65	Use electricity from Public	{1, yes}...
66	Use electricity from	{1, yes}...
67	Use electricity from Private	{1, yes}...
68	Days/Week from Public	Quantitative
69	Days/Week from Community	Quantitative
70	Hours/Day from Public	Quantitative
71	Hours/Day from Community	Quantitative
72	Hours/day subscription from	Quantitative
73	Amperes in the subscription	Quantitative
74	First source of energy use for	{1, ELECTRICITY FORM PUBLIC
75	Second source of energy use	{1, ELECTRICITY FORM PUBLIC
76	First source of energy use for	{1, ELECTRICITY FORM PUBLIC
77	Second source of energy use	{1, ELECTRICITY FORM PUBLIC
78	First source of energy use for	{1, ELECTRICITY FORM PUBLIC
79	Second source of energy use	{1, ELECTRICITY FORM PUBLIC
80	First source of energy use for	{1, ELECTRICITY FORM PUBLIC
81	Second source of energy use	{1, ELECTRICITY FORM PUBLIC
82	First source of energy use for	{1, ELECTRICITY FORM PUBLIC
83	Second source of energy use	{1, ELECTRICITY FORM PUBLIC
84	Ownership status of the	{1, OWNED BY THE HOUSEHOLD}...
85	Estimated rental monthly	Quantitative
86	Distance to Elementary	Quantitative
87	Distance to Mid, basic or high	Quantitative
88	Distance to Public hospital	Quantitative
89	Distance to Private clinic	Quantitative
90	Distance to Public medical	Quantitative
91	Distance to Pharmacy (km)	Quantitative

No.	Variables name	Type of data
92	Distance to Police station	Quantitative
93	Distance to Post office (km)	Quantitative
94	Distance to Place of worship	Quantitative
95	Distance to Youth center	Quantitative
96	Distance to Bank (km)	Quantitative
97	Distance to Fire station (km)	Quantitative
98	Distance to Municipal council	Quantitative
99	Distance to Private bus/taxi	Quantitative
100	Distance to Markets (km)	Quantitative
101	Distance to Paved road (km)	Quantitative
102	Distance to Ration agent (km)	Quantitative
103	Minutes to nearest	Quantitative
104	Minutes to nearest Mid, basic	Quantitative
105	Minutes to nearest Public	Quantitative
106	Minutes to nearest Private	Quantitative
107	Minutes to nearest Public	Quantitative
108	Minutes to nearest Pharmacy	Quantitative
109	Minutes to nearest Police	Quantitative
110	Minutes to nearest Post office	Quantitative
111	Minutes to nearest Place of	Quantitative
112	Minutes to nearest Youth	Quantitative
113	Minutes to nearest Bank	Quantitative
114	Minutes to nearest Fire	Quantitative
115	Minutes to nearest Municipal	Quantitative
116	Minutes to nearest Private	Quantitative
117	Minutes to nearest Markets	Quantitative
118	Minutes to nearest Paved road	Quantitative
119	Minutes to nearest Ration	Quantitative
120	Problems of transportation	{1, YES}...
121	Type of main road leads to the	{1, PAVED ROAD, NO PAVEMENT}...
122	amount of last payment for	Quantitative
123	amount of last payment for	Quantitative
124	amount of last payment for	Quantitative
125	days covered by last payment	Quantitative
126	days covered by last payment	Quantitative
127	estimated proper used for non-	Quantitative
128	estimated proper used for non-	Quantitative
129	Affected by smoke and gases	{1, LARGELY AFFECTED}...
130	Affected by dust	{1, LARGELY AFFECTED}...
131	Affected by bad odor	{1, LARGELY AFFECTED}...
132	Affected by noise	{1, LARGELY AFFECTED}...
133	Affected by insects, rodents,	{1, LARGELY AFFECTED}...
134	Affected by garbage near	{1, LARGELY AFFECTED}...
135	Affected by rain and stagnant	{1, LARGELY AFFECTED}...
136	Affected by outlets of sanitary	{1, LARGELY AFFECTED}...
137	Affected by humidity	{1, LARGELY AFFECTED}...
138	Affected by insufficient	{1, LARGELY AFFECTED}...

No.	Variables name	Type of data
139	Affected by security risks	{1, LARGELY AFFECTED}...
140	Affected by insufficient	{1, LARGELY AFFECTED}...
141	Number of rooms are	{1, EXTREMELY INADEQUATE}...
142	Minimum monthly income	Quantitative
143	Has any member of the	{1, YES}...
144	Agricultural holdings over	{1, YES}...
145	Animal production activity?	{1, YES}...
146	per capita expenditure market	Quantitative
147	quintile Per capita	Quantitative
148	Per capita expenditure group	{1.00, less than 60}...
149	Household expenditure group	{1.00, less than 400}...
150	Total Expenditure paid prices	Quantitative
Y	Total Expenditure market	Quantitative