International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Analysis on PLSR in Contrast with PCR

Abhimanyu Kumar

Suresh Gyan Vihar University Jagatpura, Jaipur, Rajasthan, India

Abstract: ***In this paper, we have described a very efficient tool for data mining, i.e., predictive data mining techniques. Partial Least SquareRegression, predictive technique is briefly discussed and is compared with Principal ComponentRegression, another predictive technique. We have implemented Partial Least Square Regression (PLSR) technique on the Gasoline dataset. From other researchers, we have taken the result of Principal Component Regression (PCR) technique and compared our result of PLSR with the PCR's result, and it is observed that the result of PLSR is much more significant than the PCR, in respect to R-square error (R-sq) criteria for model comparison. Also we have seen that the 3 components of PLSR gives better result as compared to 4 components of PCR.***

Keywords: KDD- knowledge Discovery database, PDM – Predictive Data Mining, PCR – Principal Component Regression, PLSR – Partial Least Square Regression, PCA – Principal Component Analysis

1. Introduction

From the past few years, data-mining has become very efficient tool for the extraction and manipulation of the data for the establishment of patterns in order to produce valuable decision making information. For the extraction of the right information from the data set, the data mining techniques are not only dependent on techniques only but they also depend on the ability of the analyst.

Berry in year 1997, proposed that the human problems can be classified into six different data mining tasks: classification, estimation, prediction, affinity grouping, clustering, and description problems. This can be collectively called as "knowledge discovery". [1]

Weiss et al. in year 1998 classified DM into two parts: prediction and knowledge discovery. First part includes classification, regression, whereas, second part defines clustering, association rules, summarization, etc. [2]

Knowledge Discovery Database (KDD) has three stages.

- Data Preprocessing
- Data mining
- Data Post-processing

In the first stage, data preprocessing is done which results in data collection, data smoothing, data cleaning, data transformation and data reduction. The second stage, called Data Mining involves data classification commonly termed as prediction. The third stage is data post-processing, which shows the conclusion drawn from the analysis in stage two.

Predictive data mining (PDM) works in the same way as the analysis of small data set is done by any human, only the advantage of using PDM is that, it can be used for large data set and has fewer problems than any human analyst has. It learns from its past experience and does not repeat same mistake in the future situations.Nowadays, a predictive approach of data mining is its most developed part. It has the greatest potential pay-off and the description is very much precise.

2. Previous Work

This chapter gives the literature review of this research. It explains the predictive data mining techniques and the work done by other researchers on that technique.

According to Jolliffe, principal component analysis is just a beginning to the study of multivariate data. This type of analysis is a conventional multivariate techniques, it works on a subject of huge research ranging from algorithms approach to the neural network approach. It is versatile, using applications which are useful in many disciplines. It is very useful in the multivariate data analysis. [3]

In this paper, Naes presents formulae for prediction of error founded by the principal component regression (PCR). Difference between principal component regression and ordinary partial least square technique is discussed in context to this formula. [4].

Sun et al, also introduced principal component regression (PCR) as a multivariate calibration. It includes principal component in the regression model, based on the variance of the components. A principal component with small variance is not preferably used in the regression model. But some of the authors after him employed that low variance does not imply that corresponding component is unimportant. [5]

3. Methodology

This chapter introduces the partial least square regression, a predictive data mining technique for the knowledge of relationship between the variables and to study the nature or properties of the data set. The data set is first introduced, as well as preprocessing on the data set is done, in order to gain insight of the properties of the data set. By plotting the inputs over the output of the raw data set, relationship check is made. Preprocessing is done by the scaling or standardizing the data set, it is also known as data preparation. To know more about the relationship between the input variables and output variables, the correlation coefficients of each of the various dataset are computed.

Now the data sets are divided into two parts, setting the odd numbered data points as the "training set" and the even numbered data set as the "test validation data set". The train data set for each data set is used for the model building. A predictive data mining technique is used to build a model, and the various methods of that technique are engaged. The model accuracy is checked by the R-Square Error Method. In this paper, only the study of Gasoline data set has limited this study only to the model validation based on the data set only. Finally, all the methods of the PLSR technique is described and studied and compared with PCR's result and the best result is shown.

A. Data Acquisition

Data set used for the research is Gasoline Dataset. For this work, MATLAB software is used for all the analyses. All the analyses (PCA, CCA) have been made before the software is used for the dataset. Before the technique is used on the dataset, an introductory analysis is done on the data set to gain knowledge of dataset.

B. Data Description and Preprocessing

The plot between each variable in the data set and the indices is made to see the dispersion between variables, which are different. The plot between the input variables and output variables is made to know the relationship between them. The correlation coefficient of the data matrix is calculated, to measure the correlation between the variables. To check for nonlinear relationship between the variables, score vectors were plotted against each other.

C. Partial Least Square Regression

PLS is a method of modeling input variables to predict a response variable. In this process the input data is transformed to a new variable or score (t) and the output data (y) is also transformed to a new score (u) making them uncorrelated factors and removing collinearity between the input and output variables. Between the score vectors **t** and **u**, a linear mapping **b** is performed. The score vectors are the values of the data on the loading vectors p and q. Analysis just like principle component is done on the new scores to create loading vectors (p and q).

D. Principal Component Regression

PCR is second technique of predictive data mining, in which PC analysis is done. PCA is an unsupervised parametric method in which a higher percentage of variance in data called principal component is selected and by help of those PC's it reduces and classifies the number of variables without significant loss of information.

E. Procedure

The data set is first introduced, as well as pre-processing on the data set is done, in order to gain insight of the properties of the data set. By plotting the inputs over the output of the raw data set, relationship check is made. To reduce the level of dispersion between the variables in the data set, the data is pre-processed. Pre-processing is done by the scaling or standardizing the data set, it is also known as data preparation.A predictive data mining technique is used to build a model, and the various methods of that technique are engaged.The model accuracy is checked by the R-Square Method. In this study, only the study of data set has limited this study only to the model validation based on the data set only. Finally, all the methods of the PLS & PCR technique is described and compared and the best model is checked.

F. Data Introduction

Spectral and octane data of gasoline is used as dataset in our thesis. NIR spectra and octane numbers of 60 gasoline samples are described here. NIR spectra are measured in 2nm intervals, i.e., ranging from 900 nm to 1700 nm. Prediction of the gasoline octane numbers with the near infrared rays in discussed here. This dataset has 700 independent variables and one dependent variable to predict the response variables.

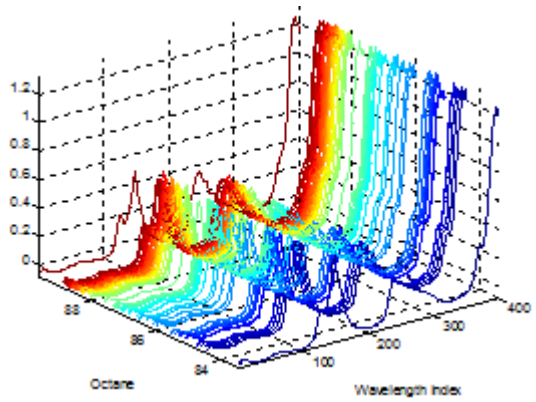

Figure 1: Analysis on data set to determine wavelength

G. Comparison Criteria

Mean Square Error, a criterion for model comparison is used in my research.MSE is the most significant criteria for determining and comparing different data mining techniques. MSE measures the difference between actual test outputs and the prediction test output. Smaller MSE is better. Large MSE values show poor prediction. The MSE of the predictions is the mean of the squared difference between the observed and the predicted values.

R-Square, also termed as R-Sq or R^2. It is used to measure the percentage variability in any of the data matrix, which is accounted for by the built model. The value closer to 1 of R-Sq, shows a better prediction.

4. Result and Comparison

In this chapter, analysis on the dataset is done. Firstly, PLSR, predictive data mining technique is applied on the dataset and then PCR, another predictive data mining technique is applied and then both are compared with each other, in order to check the effectiveness of the models. In both the techniques, the resulting prediction is compared with the actual output variables and the difference between them is measured using a statistical methods.

In the partial least square regression, the train data set is scaled and the means and standard deviation is defined to scale the test data set. Selection of the significant PLS component is done by the partial least square analysis.

The reduced singular value decomposition (SVD) and the resultant Eigen vectors acknowledge the variable that played significant role in each PLS component.

Fitting the Data

A plot between the number of components and the percentage variance explained in the response variable is shown in the figure. Use of ten PLS components and a response variable is shown. Use of ten components is more than what we actually needed to adequately fit the data. But this type of diagnosis is necessary in order to make a choice of a simple model with lesser components.

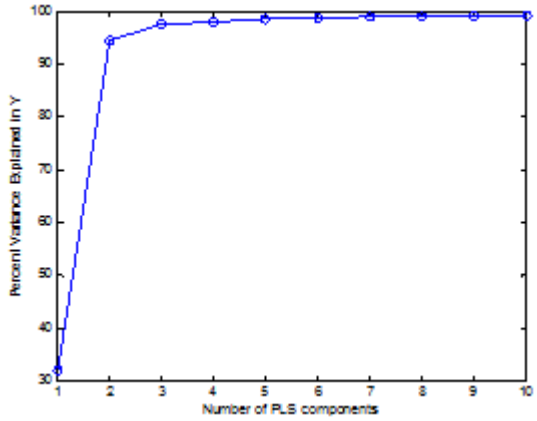

Figure 2: Plot between PLS components and the variance

Fitting Data on PLSR

We can check the accuracy of the model firstly by taking data on PLSR. Although, this is not a valid model accuracy measure criterion, it is very close to the accurate model analysis criteria. Plotting of the response variable against the two predictors is shown in the figure below.

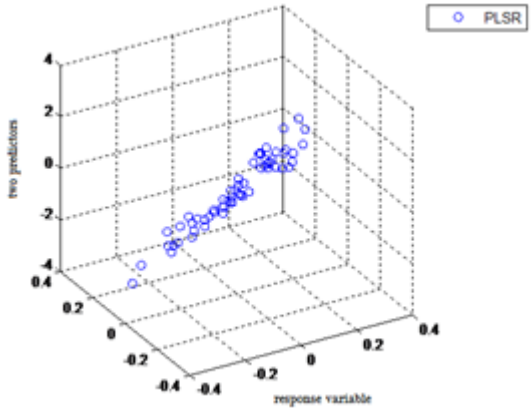

Figure 3: Plotting of response variable with two predictors

Model Parsimony / Comparison

PLSR with only 3 components gives the best prediction as compared to the PCR with 4 components. In the PLS, the original variables that define the PLS components, also defines the PLS weights. PLS weights define the strength of each component of the PLSR and in what direction.

R-square of PLSR = 0.9466
R-square of PCR = 0.1962

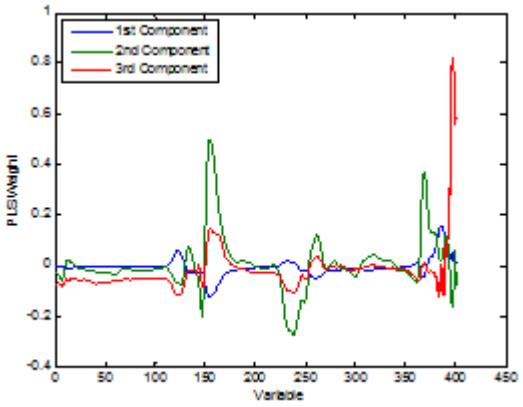

Figure 4: Plot showing PLSR with 3 components

As the PLSR model with 3 components gives a better prediction than PCR model with 4 components. Comparison of model is somehow depends on the R-square, model comparison criteria also. As we have also seen that the R-square result of PLSR is more accurate than the PCR's R-square result. The components that predict more about the data is chosen by its weight, and helps in choosing the data from the original set to a reduced subset with more accuracy.

Acknowledgment

I would like to appreciate my family and friends who always have motivated me to complete this research.

References

[1] Berry, Michael J. A. et al., Data-Mining Techniques for Marketing, Sales andCustomer Support. U.. S A: John Wiley and Sons (1997).
[2] Weiss, Sholom M. et al., Predictive Data-Mining: A Practical Guide. San Francisco, Morgan Kaufmann (1998).
[3] Jolliffe, I.T., Principal Component Analysis, New York: Springer-Verlag (1986).
[4] Naes, T., and H. Martens, "Principal Component Regression in NIR Analysis: Viewpoints, Background Details and Selection of Components," J.Chemom. 2 (1988).
[5] Sun, J., "A correlation Principal Component Regression Analysis of NIR." J.Chemom9 (1995)