ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 5, Issue 7, July 2016

Application of Dimensionality Reduction techniques in Real time Dataset

M.Sabitha, M.Mayilvahanan

***Abstract* – Data pre-processing in data mining refers to transforming the raw data into understandable format for further analysis. The real time data is incomplete, robust and unorganised which should be cleaned and transformed to make it efficient for pre-processing. In this paper, we have discussed about three dimensionality reduction techniques namely Principal Component Analysis (PCA), Singular Value decomposition and Learning Vector Quantization applied to solar irradiance dataset. The dataset consists of temperature, solar irradiance, and humidity data for 25 years for selected eight south Indian cities. The dimensionality reduction was done by applying the above mentioned three algorithms and their efficiency was evaluated, to find the best suiting algorithm to apply for the dataset.**

***Index Terms* - Data Mining, Dimensionality Reduction, Learning Vector Quantization, Principal Component Analysis (PCA), Singular Value Decomposition.**

I. INTRODUCTION

Real time data are incomplete, noisy and inconsistent in nature that they need to be pre-processed for further analysis [1]. Dimensionality Reduction is done to extract or narrow down the data to facilitate analysis [2]. Data pre-processing consists of steps like cleaning, integration, transformation, reduction and discretization. In this paper, dimensionality reduction algorithms namely Principal Component Analysis (PCA), Singular Value Decomposition and Learning Vector Quantization are applied. The efficiency and performance of the algorithms are tested to state which algorithm suits best for the dataset.

II. DATASET

The dataset is a real time solar irradiance data of South Indian cities namely Chennai, Salem, Erode, Nilgiris, Bangalore, Vellore, Coimbatore and Madurai. The dataset consists of month-wise temperature, solar irradiance and relative humidity data for 25 years.

***M.Sabitha**, Department of Computer Science, PSG College of Arts and Science, Coimbatore, India.*

***M.Mayilvahanan,** Department of Computer Science, PSG College of Arts and Science, Coimbatore.*

III. DATA PREPROCESSING

In real time data are (i) incomplete which lacks attribute values, lacking certain attributes of interest or containing only aggregate data, (ii) noisy containing errors and (iii) inconsistent containing discrepancies in codes or names. The data need to be pre-processed to make it suitable for analysis. The steps involved in data pre-processing are cleaning, integration, transformation, reduction and discretization [3].

Data Cleaning is done by filling in missing values of attributes, identifying outliers and smooth out noisy data and correct inconsistent data .Data Transformation is normalizing the data, aggregating, generalizing and constructing the attribute. Data Reduction is reducing the number of attributes, reducing the number of attribute values, reducing number of tuples.

IV. DIMENSIONALITY REDUCTION

In Machine Learning and statistics, dimensionality reduction is termed as the process of reducing the number of random variables. It can be divided into feature selection and feature extraction [4]. Feature Selection approaches try to find a subset of the original variables. Feature extraction transforms the data in the high-dimensional space to low-dimensional space. The data transformation technique may be linear or non-linear. One such example of linear transformation technique is principal component analysis.

Dimensionality Reduction techniques taken for study here is Principal Component Analysis, Singular Value Decomposition and Learning Vector Quantization.

A. EIGENVAUES AND EIGENVECTORS

In linear algebra, an eigenvector[5] of a linear transformation T from a vector space V over a field F into itself is a non-zero vector that does not change its direction when that linear transformation is applied to it. If V is a vector that is not the zero vectors, then it is an eigenvector of a linear transformation T if T(V) is a scalar multiple of V. This condition can be written as the mapping

T:v→ **λv**

Where **λ** is a scalar in the field F, known as the eigenvalue associated with the eigenvector V.

B. PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations possibly correlated variables into a set of values of linearly uncorrelated variables called principal components [6]. The number of principal components is less than or equal to the original variables. The idea is to treat the set of tuples as a matrix M and find the eigenvectors for MM^T or M^TM. The matrix of these eigenvectors can be thought of as a rigid rotation in a high dimension space. When we apply this transformations to the original data, the axis corresponding to the principal eigenvector is the one along which points are most "spread out".

Each principal component in Principal Component Analysis is the linear combination of the variables and gives a maximized variance [7]. Let *X* be a matrix for *n* observations by *p* variables, and the covariance matrix is *S*. Then for a linear combination of the variables

$$z_1 = \sum_{i=1}^{p} a_{1i} x_i$$

where x_i is the *i*th variable , $a_{1i}\ i = 1, 2, \ldots, p$ are linear combination coefficients for z_1, they can be denoted by a column vector a_1, and normalized by $a_1^T a_1 = 1$.

The variance of z_1 will be $a_1^T S a_1$.

The vector a_1 is found by maximizing the variance. And z_1 is called the first principal component. The second principal component can be found in the same way by maximizing:

$a_2^T S a_2$ Subject to the

Constraints $a_2^T a_2 = 1$ and $a_2^T a_1 = 0$

It gives the second principal component that is orthogonal to the first one. Remaining principal components can be derived in a similar way. In fact coefficients $a_1, a_2, \ldots, a_p$ can be calculated from eigenvectors of the matrix *S*. Origin uses different methods according to the way of excluding missing values.

The following matlab code was implemented

```
[pc,score,latent,tsquare] = princomp(data2);
 red_dim = score(:,1:10);

X5 = bsxfun(@minus, data2, mean(data2,1))
covariancex =
(X5'*X5)./(size(X5,1)-1);

[V D] = eigs(covariancex, 10);
% reduce to 10 dimension

Xtest = bsxfun(@minus, data2, mean(data2,1));
pcatest = Xtest*V
```

C. SINGULAR VALUE DECOMPOSITION

Singular value decomposition (SVD) allows an exact representation of any matrix, and it is used in easy elimination of less important parts of the representation to produce an approximate representation with any desired number of dimensions [8].

The following matlab was implemented

```
while (sum(abs(A(~eye(m,n)))) > e)
% termination condition
      for i = 1:n
          for j = i+1:n
              [J1,J2] =
jacobi(A,m,n,i,j);
              A =
mtimes(J1,mtimes(A,J2));
              U = mtimes(U,J1');
              V = mtimes(J2',V);
          end
          for j = n+1:m
              J1 =
jacobi2(A,m,n,i,j);
              A = mtimes(J1,A);
              U = mtimes(U,J1');
          end
      end
  end
```

```
  S = A;
  % check if we need less than
three output arguments
  if (nargout < 3)
      Uout = diag(S);
  else
      Uout = U; Sout =
times(S,eye(m,n)); Vout = V;
  end
end
```

D. LEAST VECTOR QUANTIZATION

Vector Quantization is a quantization technique from signal processing that allows the modelling of probability density functions by the distribution of prototype vectors [9]. It works by dividing a large set of points (vectors) into groups having approximately the same number of points closest to them.

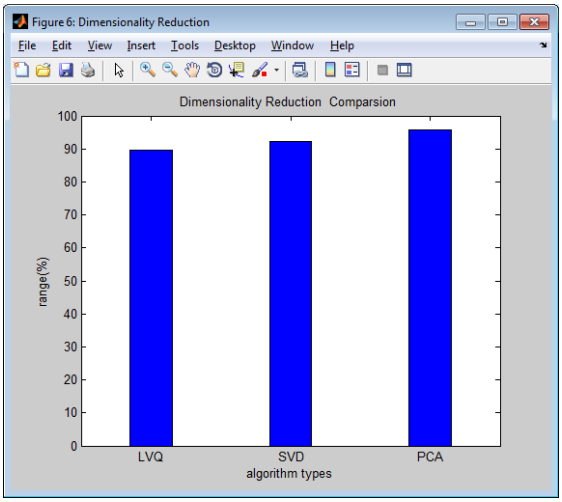
The Least Square Quantization technique works on the principle of finding a sample point and moving the nearest quantization vector centroid towards this sample point, by a small fraction of the distance.

A training algorithm for vector quantization

1. Pick a sample point at random
2. Move the nearest quantization vector centroid towards this sample point, by a small fraction of the distance
3. Repeat

V. COMPARING THE TECHNIQUES

The dimensionality reduction techniques are available as linear and non-linear based upon the transformations. Choosing the appropriate algorithm for the dataset will improve the efficiency of application and saving the runtime. Here in this paper , we compared three dimension reduction techniques namely Principal Component Analysis (PCA) , Singular Value Decomposition (SVD) and Least Square Quantization (LSQ) and found Principal Component Analysis performed best in the dataset as the dataset mainly consist of numerical value.



VI. CONCLUSION

Data Pre-processing is an important step in Data Mining, as it facilitate efficient analysis of data. Dimensionality reduction is a step in Data pre-processing in which the dimensions of the data are reduced so that the data to be analysed is narrowed down to perform well. In this paper, we applied three dimensionality reduction techniques namely Principal Component Analysis (PCA), Singular Value Decomposition (SVD) and Least Vector Quantization (LVQ) in the dataset. The dataset consists of temperature, solar irradiance and relative humidity data for 25 years of selected eight South Indian Cities. Since the data was mainly numeric, Principal Component Analysis (PCA) performed well compared to the other two techniques.

REFERENCES

[1] Rinnan, Åsmund, et al. "Data pre-processing." *Infrared spectroscopy for food quality analysis and control* (2009): 29-31.

[2]Wikipedia, https://en.wikipedia.org/wiki/ data_pre-processing

[3] Kuhn, Max, and Kjell Johnson. "Data pre-processing." *Applied Predictive Modeling*. Springer New York, 2013. 27-59.

[4] Dash, Manoranjan, Hua Liu, and Jun Yao. "Dimensionality reduction of unsupervised data." *Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on*. IEEE, 1997.

[5]Wikipedia, https://en.wikipedia.org/wiki/Eigenvalues_ and_eigenvectors

[6] Bro, Rasmus, and Age K. Smilde. "Principal component analysis." *Analytical Methods* 6.9 (2014): 2812-2831.

[7] Roweis, Sam T., and Lawrence K. Saul. "Nonlinear dimensionality reduction by locally linear embedding." *Science* 290.5500 (2000): 2323-2326.

[8] Golub, Gene H., and Christian Reinsch. "Singular value decomposition and least squares solutions." *Numerische mathematik* 14.5 (1970): 403-420.

[9] Gersho, Allen, and Robert M. Gray. *Vector quantization and signal compression*. Vol. 159. Springer Science & Business Media, 2012.