Analysis of dimensionality Reduction Using Principal Component Analysis (PCA) and Two Dimensional Haar Wavelets

Rajesh Kumar E

Department of CSE, Koneru Lakshmaiah Education Foundation,

Vaddeswaram, AP, India. rajthalopo@gmail.com,

Dr. Sivakumar Selvarasu

Department of Computer Applications, Faculty of Science and Humanities,

SRM Institute of Science and Technology, KTR Campus, Kattankulathur, Tamil Nadu - 603 203.

Abstract

Due to advances of digital technology in all sectors ranging from healthcare, production, web organization a huge data had been generated through these domains. Machine learning algorithms are used to uncover the hidden features of these data. The main curse of these data is that, it also generates huge volume of redundant data. So we need a data reduction technique to reduce the data and analyze the key features hidden in the data. In this work an attempt has been made to use Principal component Analysis (PCA) along with Two dimensional Haar wavelet(2DDHW) has been used to reduce the dimension of the data. To validate the above techniques we used Cardiotography (CTG) data set has been used . Our experimental result further confirms that by using two dimensional Haar wavelets along with the traditional PCA gives us better results.

INTRODUCTION

In the digital era, numerous number of data are generated daily in different field such as healthcare, business analytics, social media, banking sector, bioinformatics.[1] However, all the data that are generated through this sectors may not be useful but only a portion is useful for deciding the decision making task. Although the Machine Learning

Algorithms (MLAs) can be used to process this kind of big data [2-5]. MLAs are used in the different areas to predict and classify the test data to produce accurate results. In Recent days there is a constant usage of Maching Learning classfiers models in medical field. They have to be proven an indispensable tool in diagnosis of various medical datasets [6-7] .

Although MLAs can process the medical dataset in an effective manner that efficiency will decreases when the dimensionality increases [8]. When the number of attributes increases and as a result the learned model becomes more complicated to analyze and exhibit the exact prediction using the learned model. To overcome the above problems one need a less number of data this in turn gives the exact prediction of the data. This was done my applying Dimensionality Reduction Algorithm (DRA). DRA aims to solve the curse of the dimension of the data without the compromising the features of the data.

During the pregnancy, most of the women are affected with diabetes and high blood sugar. This will in turn the affect of the grown of the baby in the womb. Cardiotography (CTG) is a procreducre to diagnose the test under the formation of baby during pregnancy period [9]

CTG produces a recorded result of the mother uterine contraction signal and fetal heart rate [10]. Any irregularities of the fetal can be easily diagnosed through CTG. This is the standard procedure that is adopted to monitor the fetal growth across all the nations. [11]

For this type of clinical dataset prediction of the vital organs through out the pregnancy period plays a major role. In this connection, dimensionality plays a very crucial role in reducing the high dimensionality data to a low dimensional data.

In this work, one of the most popular dimensionlaityu reduction technique namely Principle component analysis are investigated along with Two Dimensional Discrete Haar wavelet (2DDHW) transform are used in MLAs such as Decision Tree, Navie Bayes, Random Forest, and Support Vector Machine using publicly available CTG dataset form UCI machine learning repository[12]

In the proposed work, first step is that, the CTG dataset is subjected by applying feature engineering to improve the quality of the dataset. In the next step the most important dimensionality reduction technique Principle Component Analysis (PCA) are applied individually on the CTG dataset that will extract the most important attrinubu where tes. The extracted features are put into trained in the MLAs where the confustion matrix is arrived. Later 2DDHW are applied to access the sensitivity, accuracy, and specificity.

1. Dimensionality Reduction Techniques

Dimensionality reduction is a one of the techniques which extract the important features thereby reducing the size of the dataset without affecting the characteristics of data. This technique provides better data visualization, data compression, improved classification accuracy, fast and efficient data retrieval. The main merits of the dimensionality reduction technique are reducing the dimension enhances the prediction accuracy of the classifier with good performance and also reduces the computational cost which was shown in Fig. 1

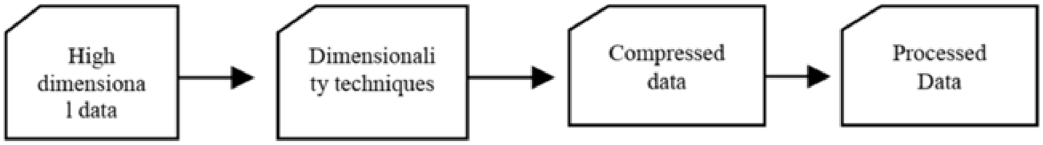


Fig. 1. Dimensionality Reduction Process

In this work some classifiers in ML like decision Tree, Naïve Baye, Random Forest and SVM are used to classify the data and arrives a confusion matrix,

1.1 Principal Component Analysis (PCA)

It is considered one of the standard dimensionality procedures to reduce the data size without much affecting the main attributes of the data. It works on the principle of orthogonal transformation. PCA converts a group of correlated variables to a group of uncorrelated variables [13] and which in turn expolore data analysis. It is also used for examine the connection amoung a group of variable. Hence it is considered as a vital tool for dimensionality reduction.

Algorithm of PCA

- Standardize the dataset.
- Calculate the covariance matrix for the features in the dataset.
- Calculate the eigenvalues and eigenvectors for the covariance matrix.
- Sort eigenvalues and their corresponding eigenvectors.
- Pick k eigenvalues and form a matrix of eigenvectors.
- Transform the original matrix.

In this way the raw data with n diemnsionlaity is reduced to an new m dimensional data.

Two Dimensional Wavelets

Wavelets are considered one of the best tool for signal processing, image processing. There are several wavelet are available in the literature like Haar wavelet, Daubechies wavelets, Coifman wavelet, Meyer wavelets.[14-16,]. Depends upon the applications of the problem different wavelets are used. One of the simplest wavelet is the Haar wavelet which was derived from the Haar scaling function and defined as

$$\varphi(x) = \{^{1 \quad 0 < x < 1}$$

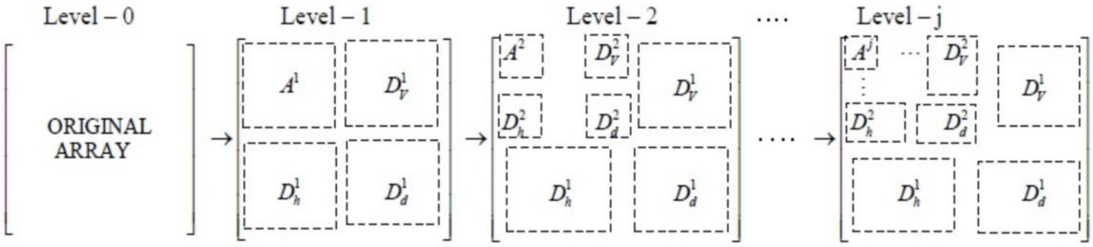
$$0 \ otherwise$$

Similarly one can define the two dimensional wavelets whose scaling function and vertical, horizontal, and diagonal scaling function is defined as follows

Using the above coefficients we can apply two dimensional discrete Haar wavelet transform. When the consider the data which is of the following matrix can be represent by

$$S = \left[\begin{array}{cc|cc|c|cc}
s_{11} & s_{12} & s_{13} & s_{14} & \cdots & s_{1(n-1)} & s_{1n} \\
s_{21} & s_{22} & s_{23} & s_{24} & \cdots & s_{2(n-1)} & s_{2n} \\
\hline
s_{31} & s_{32} & s_{33} & s_{34} & \cdots & s_{3(n-1)} & s_{3n} \\
s_{41} & s_{42} & s_{43} & s_{44} & \cdots & s_{4(n-1)} & s_{4n} \\
\hline
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
\hline
s_{(n-1)1} & s_{(n-1)2} & s_{(n-1)3} & s_{(n-1)4} & \cdots & s_{(n-1)(n-1)} & s_{(n-1)n} \\
s_{n1} & s_{n2} & s_{n3} & s_{n4} & \cdots & s_{n(n-1)} & s_{nn}
\end{array}\right]$$

Where S is the square matrix of order n x n. Where ($n = 2^j$) j represent the level of decomposition of the matrix. The input matrix in this case denotes the feature vector can be decomposed by applying wavelet coefficient as into a submatrix as shown in the figure.



Decomposing of original matrix into various levels.

The A^1, D_v^{1}, D_h^{1}, D_d^{1} are obtained by applying scaling approximating , vertical wavelet coefficient , horizontal wavelet coefficient, and diagonal wavelet coefficient.

Dataset Description:

During the trimester period many pregnant women feels uneasy. During this course of time, a feutus heart rate also has some problems with respect to the oxygen supply. CTG is used to observe the fetal heart and contractions of the uterus. CTG dataset has 2126 instances and 23 attributes. The major attribute that used for constraction of the uterus and the fetal heart rate are UC (uterine contraction per second) and FM (fetal movements per second)

Data Set Characteristics	Multivariate	**Number of instances**	2126	**Area**	Life
Attribute Characteristic	Real	**Number of Attributes**	23	**Date Download**	2010/09/07
Associated Tasks	Classification	**Missing Values**	N/A	**Number of Web Hits**	209179

Major Attribute in the dataset

LB	FHR baseline
MC	Acceleration per second
DL	Light Deceleration per second
DS	Severe deceleration per second
DP	Prolonged Deceleration per second
ASTV	Percentage of time with abnormal short term variability
MSTV	Mean value of short term variability

MLTV	Mean value of long term variability

PROPOSED METHODOLOGY

This paper explores the effect of feature extraction and dimensionality reduction techniques on the performance of ML algorithms on CTG data set. The various step used in this work are discussed below.

Step1:

Feature extraction technique, normalization and conversion of categorical data to numerical data is applied CTG dataset. To normalize the input dataset, standard scalar normalization method is used.

Step 2:

The normalised dataset is tested using ML algorithms, Decision Tree, Naïve Bayes, Random Forest and Support Vector Machine (SVM). The performance of these classifiers is then evaluated on the metrics, Precision, Accuracy, Sensitivity and Specificity

Step 3:

PCA is applied on the normalized dataset to extract the most important features. The resultant dataset is then tested using the ML algorithms.

Step 4.

After obtained the confusion matrix two dimensional discrete Haar wavelet (2DDDHW)s are obtained for calculate the performance of the various measures are evaluated.

Metrics for Evaluation of the Model

It is the percentage of correct predictions that a classifier has made when comparted to the actual value of the label in testing phase.

The metric used in the analysis are Accuracy, Sensitivity, and Specificity Accuracy can be calculated using the following formula

Accuracy = (TN +TP)/(TN + TP + FN +FP) where, TP is true positives, TN represents true negatives, FP is false positives, FN is false negatives

Sensitivity

It is the percentage of true positives which are truly identified by the classifier during testing. It can be derived using the give formula TP/(TP + FN)

Specificity

It can calculate the percentage of true negatives that are correctly identified by the classifier during testing and is derived using the following TN/(TN + FP)

Performance Evaluation of Classifiers with PCA and 2DDHW

First the raw data that was obtained by the UCI repository are subjected with the following MLAs such as Decision Tree, Naïve Bayes, Random Forest and SVM. The confusion matrix shows that SVM and Random Forest algorithms perform slightly better than Decision Tree and Naïve Baye in terms of Precision, and F1Score.

32330Decision Tree Confusion Matrix: [4 54 0]
0 0 42

Naïve Bayes Confusion Matrix: [336 2 2
4 53 0]
0 6 24
325 1 0

Random forest Matrix: [5 53 0]
0 0 42
332 1 0

SVM Confusion Matrix: [4 54 0]
0 1 34

After obtained the confusion matrix we applied PCA as well as 2DDHW transform and the results are shown below

Summary of the results of CTG dataset

	Precision (%)	F1-Score (%)	Recall (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Number of Features

Decision Tree (DT)	98	98	98	98.52	99.06	92.1	36
Naïve Bayes (NB)	96	96	96	96.52	95.85.	92.84	36
Random Forest(RF)	99	99	99	98.56	99.6	92.5	36
Support Vector Machine (SVM)	99	99	99	98.5	96.5	95.25	36
DT +PCA	98	98	95	98	98.2	98.3	26
NB + PCA	95	95	95	95	100	80.3	26
RF + PCA	97	97	97	97.3	9.6	91.3	26
SVM+PCA	98	98	98	98.5	99	92	26
DT + 2DDHW	97	97	97	97.4	99.3	84.5	1
NB + 2DDHW	98	98	98	98.4	98.3	65	1
RF + 2DDHW	98	97	97	97.4	99.3	84.5	1
SVM + 2DDHW	98	98	98	98.5	98	86.2	1

Results and Discussions

The following points has to observed from the above table

1. DT. NB RFM and SVM perform the same without dimensionality reduction.
2. When the dataset is subjected to dimensionality reduction using PCA DT, SVM, RF classifiers performs better with respect to the measures. Performance of NB is reduced in terms of accuracy and specificity when the dimension is reduced.
3. While applying the 2DDHW on the dataset the specificity of NB, RFM performance better than the other classifiers.

Conclusions

In this work, a small initiative has been taken to CTG dataset to analyse the performance of MLAs using PCA and 2DDHW transforms. The dimensionality of data is well reduced on CTG dataset while applying PCA and features of the data is unchanged because of the orthogonal transformation. Whereas 2DDHW transform depends on the average and difference means of the data. Thus 2DDHW acts as one more filter for the CTG data set.

References

[1] F. Anowar, S. Sadoui, "Incremental neural-network learning for big fraud data", in 2000, IEEE international conference on system, Man, and Cybernetics (SMC), IEEE, 2002, pp 3551 – 3557.

[2] F. Anowar, S. Sadoui "International framework for real-world fraud detection environment". Comut. Intell. (2002) 1 -22.

[3] V. Spruyt "The Curse of dimensionality in classification", Comput., Vision Dummies, 21 (2002) 35 – 40.

[4] L.Van Der Maaten, E. Postma, J. Van den Herik, "Dimensionality Reduction A Comparative Review, J. Mach. Learn. Res. (2009), 66 – 71

[5] P. Jindal, D. Kumar, A review on dimensionality reduction techniques, Int.,J.Comput.Appl, (2017), 42 – 46.

[6] C. T. Rasmussen, "Gaussian Processing in Machine Learning", in Summer School on Machine Learing, Berlin, Germany, Springer, 2003, pp 63 -71

[7] T. G. Dietterich "Ensemble methods in machine learning", in Proc., Int., Workshop

Multiple Classifier Syst. Berlin, Germany, Springer, (2000) 1 -15.

[8] D.M.Hawkins, The problem of overfitting, J.Chem. Inf. Comput.Sci (2004) 1 – 12.

[9] R.M.Grivell, S.Alfirevic, G.M Gyte and D.Devane, "Antenatal cardiotocography for fetal assessment", Cochrane Database Systematic Rev., (2015), 1- 48.

[10] Z. Alfirevic, G.M. Gyte, A.Cuthbert and D.Devane" Continuous Cardiotocography (CTG) as a form of electronic fetal monitoring for fetal assessment during labour", Cochrane Database Systematic Rev., (2017) 1 – 108.

[11] D.Ayres – de-Campus, C.Y. Spong "FIGO consensum guidelines intrapartum fetal monitoring CTG",Int. J.Gynecol. Obsterics (2015), 13 -24.

[12] D.Ayres-de-Campos, J.Bernardes, A, Garrido and L.Pereira- Leite, "Sisporto 2.0 A program for automated analysis of CTG", J.Maternal-Fetal Med, (2000), 311 – 318.

[13] J.Leskovec,A.Rajaraman, J.Ullman, Dimensionlality Reduction , in Mining of Masive Datasets, 2014, pp. 415 – 447

[14] G.Hariaharan, Wavelet Analysis- An overview, Wavelet Solutions for Reaction–Diffusion Problems in Science and Engineering, 2000, 15 – 31

[15] V.Sumathy, S.Hemalatha, B.Sripathy, Comparitive Study of Two dimensional Legendre and Chebyhev Extended case, 2019, 13 – 17

[16] P Vijayaraju, B Sripathy, D Arivudainambi, S Balaji "Hybrid memetic algorithm with two-dimensional discrete Haar wavelet transform for optimal sensor placement", 2017 2267 – 2278.