Analysis of Dimensionality Reduction Techniques on Big Data

CSCE 5310: METHODS IN EMPERICAL ANALYSIS

Project Increment 1

# Group:6

## Team Members:

Bhavana Girni

Manasa Veena

Senthil Kumar Rajendran

# Motivation

The businesses of machine learning the classifier model in the health industry is constantly expanding. It has shown to be quite useful in the diagnosis of a different types of clinical and medical datasets. Women are most impacted by high blood pressure, jaundice during pregnancy time. This will have an impact on the unborn child's health and growth. Cardiotocography (CTG) is a diagnostic tool primarily used in pregnancy-related testing for women.

The fetal heart rate and values of signals from the mother's uterine contractions are recorded by the CTG scan. This data has high number of attributes/dimensions and to be able to train a ML model to classify the data to predict a fetus as ‘Normal’, ‘Suspect’ or ‘Pathologic’ requires the data to mapped to fewer dimensions to help train the a classifier model to be accurate and fast.

This is where dimensionality reduction techniques like PCA and LDA come in.

The goal of this work is to study and review the impact of the above mentioned DR methods on the performance of Machine Learning algorithms.

# Objectives and goals

• Outwriting a complex and methodical analysis of the backend information and literature on dimensionality reduction approaches

• To determine effectiveness of PCA and LDA on ML algorithms in relation to various metrics (Accuracy, Specificity, Recall, F1-score, Sensitivity, Precision)

• That the dimensionality reduction approaches do not impair the performance of ML algorithms

• When we process the data to be fed to machine learning algorithms, use feature engineering methods to ease out any ML performance bottlenecks

• To determine how the approach is working on the datasets with different dimensionalities

# Significance

In digital world generates lots of amount of the data from eventually every industry. To determine imaginary patterns from the data that is to be utilized to inform managing and commercial purpose decisions, machine learning algorithms are being created instead of regular methods. By significantly decreasing the time complexity of the machine learning algorithms' production phase, dimensionality reduction approaches might lighten their computing power to that of the machine learning algorithms

# Features

Listed below are some of the important features of the CTG dataset

* FHR Bottomline for LB (beats per every minute).
* Accelerations per second (AC)
* Light decelerations per every second, or DL
* DS stands for Decelerations per every Second
* Long-term accelerations per every second
* Percentage of time with anomalous short-term variability in heart (ASTV)
* Mean value of the short-term variability (MSTV)
* Amount of time with abnormal long-term variability, or ALTV
* Mean value of long-term variability, or MLTV

# Related Work

Machine Learning has revolutionized the way we look and use data. Digitization has resulted in a lot of data being generated across various areas and this enormous gold mine of data that had remained untouched or considered unusable is now allowing ML to learn from it and make better predictions and opening up for a plethora of applications in various fields.

This section presents a survey of numerous publications on machine learning algorithms and dimensionality reduction methods.

Advanced Machine Learning algorithms like Artificial Neural Networks, Convolutional Neural Networks, Recurrent Neural networks have been used for predictions on UCI Laboratory data for various types of applications. ANN with 18 neuron hidden layer was used to measure the performance of DT, SVM and Naive Bayes models in discovering catalyst materials.

CNN was used to predict heart diseases based on ECG signals and is also known to be used as the basis for one of the best fetal cardiac signal classifiers.

New methods to interpret CTG with PCA, using DSP for processing CTG signals better

ANN and SVM were integrated in an approach to diagnose Breast Cancer data from the Wisconsin Breast Cancer Dataset and the accuracy was measured to be around 99.71% . Integration of SVM and Simulated annealing algorithm resulted in a hybrid model that has 97% accuracy in diagnosing Hepatitis. The test analysis was performed using real time UCI dataset.

Dimensionality reduction techniques like PCA has been used in conjunction with K-means to predict diabetes and brain tumor. The PCA is first applied on the data received from blood samples and MRI images to reduce the high number of dimensions and K-means is applied to cluster the data. This hybridization proved to have high accuracy and speed

PCA firefly in combination with xgboost and deep learning neural network models have been used for classification of intrusion detection and diabetic retinopathy datasets respectively to good yield on classification results.

# Dataset

Cardiotocography is a technique used for monitoring and recording Fetal heart rate of a baby in the mother's womb and also to record the mother's uterine contraction signals. Its a standard practice followed in many places around the world , hence plays a vital role in diagnosing the fetal health of the baby during pregnancy.

The cardiotocography dataset is available in the UCI Machine Learning repository and has 2126 fetal cardiotocograms that were automatically processed and the respective diagnostic features measured. The CTGs were also classified by three expert obstetricians and a consensus classification label assigned to each of them.

# Detail Design of Features

## Raw Data format

|  |  |
| --- | --- |
| Column Name | Description |
| **FileName** | of CTG examination |
| **Date** | of the examination |
| **b** | start instant |
| **e** | end instant |
| **LBE** | baseline value (medical expert) |
| **LB** | baseline value (SisPorto) |
| **AC** | accelerations (SisPorto) |
| **FM** | foetal movement (SisPorto) |
| **UC** | uterine contractions (SisPorto) |
| **ASTV** | percentage of time with abnormal short term variability (SisPorto) |
| **mSTV** | mean value of short term variability (SisPorto) |
| **ALTV** | percentage of time with abnormal long term variability (SisPorto) |
| **mLTV** | mean value of long term variability (SisPorto) |
| **DL** | light decelerations |
| **DS** | severe decelerations |
| **DP** | prolongued decelerations |
| **DR** | repetitive decelerations |
| **Width** | histogram width |
| **Min** | low freq. of the histogram |
| **Max** | high freq. of the histogram |
| **Nmax** | number of histogram peaks |
| **Nzeros** | number of histogram zeros |
| **Mode** | histogram mode |
| **Mean** | histogram mean |
| **Median** | histogram median |
| **Variance** | histogram variance |
| **Tendency** | histogram tendency: -1=left assymetric; 0=symmetric; 1=right assymetric |
| **A** | calm sleep |
| **B** | REM sleep |
| **C** | calm vigilance |
| **D** | active vigilance |
| **SH** | shift pattern (A or Susp with shifts) |
| **AD** | accelerative/decelerative pattern (stress situation) |
| **DE** | decelerative pattern (vagal stimulation) |
| **LD** | largely decelerative pattern |
| **FS** | flat-sinusoidal pattern (pathological state) |
| **SUSP** | suspect pattern |
| **CLASS** | Class code (1 to 10) for classes A to SUSP |
| **NSP** | Normal=1; Suspect=2; Pathologic=3 |

### Features/ Dimensions considered for Dimensionality reduction

|  |
| --- |
| **Features** |
| **LB** |
| **AC** |
| **FM** |
| **UC** |
| **DL** |
| **DS** |
| **DP** |
| **ASTV** |
| **MSTV** |
| **ALTV** |
| **MLTV** |
| **Width** |
| **Min** |
| **Max** |
| **Nmax** |
| **Nzeros** |
| **Mode** |
| **Mean** |
| **Median** |
| **Variance** |
| **Tendency** |

### Target Label

NSP

# Analysis

To be able to predict diseases in infants before they are born or in the womb CTG dataset provides a lot of information. This dataset needs to be processed initially before feeding into a ML classifier model to be able to do accurate predictions. Since the dataset contains a wide range of attributes/dimensions it needs to preprocessed and this is where dimensionality reduction techniques come in to map the high dimensionality correlated attributes to lower dimensionality unit variance low correlated dimension.

Feature engineering also plays a vital role in extracting the transformed dimension/feature from raw data to actual input to the ML model.

Principal Component Analysis (PCA) and Linear Discernment Analysis (LDA) are the two popular dimensionality reduction techniques that are analyzed on some of the well known ML algorithms like Decision tree, Naïve Bayes, Random Forest and Support Vector Machine (SVM) using the CTG dataset from the UCI Machine Learning repository. The experiment is then repeated on datasets of varied dimensions like Diabetic Retinopathy and Intrusion Detection datasets

The performance of the ML algorithms on these various datasets before and after applying dimensionality reduction is measured and particular focus placed on the performance of PCA vs LDA against several established metrics like F1-score, Accuracy, Sensitivity Precision, Recall & Specificity

The end goal of this analysis is to identify and prove that dimensionality reduction techniques improve the speed and performance of ML algorithms

# Implementation

Following steps are followed to implement Dimensionality reduction and feed the transformed data in to the ML model

* Cleanse the dataset
* Standardize the dataset to have unit variance and zero mean
* Calculate the co-variance matrix
* Calculate the eigen vector and eigen value of the co-variance matrix
* Sort and pick the top k eigen vectors and do a dot product with the raw data, essentially mapping the high dimensional raw data on to the k-dimensional transformed data producing the dimensionally reduced dataset
* Train the ML models with the new transformed reduced dataset and test them with test data
* Calculate the various metrics and evaluate the ML models
* Perform step 6 & 7 without dimensionality reduction and compare
* Perform all the above steps for the Diabetic Retinopathy and Intrusion Detection Datasets to analyze the performance of PCA and LDA

# Preliminary Results

For this increment we have completed performing LDA, PCA for the CTG dataset and run the 4 ML classifiers on both the reduced datasets. For comparison purposes we have also run the ML models on full non-reduced dataset. Preliminary results are attached below

## Accuracy of ML models on a LDA Reduced dataset

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

## Accuracy of ML models on a PCA Reduced dataset

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

## Accuracy of ML models on a full dimensional dataset

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

# Project Management

## Implementation status report

Work Completed

Description

We have completed PCA and LDA coding part and a complex and methodical analysis of the backend information and literature on dimensionality reduction approaches for the CTG dataset

Contribution and Task

For this part of the project, we have done with the coding part of LDA and PCA and clear documentation part with the goals and the objective of the project and the significance and the features Here we have also described the major features we have used by all these things we have also have clear description for the expected output.

Bhavana:

Completed LDA python notebook

Documentation of motivation, goals and objectives and the features used.

Contributed 35% of overall effort.

Senthil:

Completed PCA python notebook

Documentation of Related work, Dataset, Detail Design of Features and Implementation and video voice over

Contributed 35% of overall effort.

Manasa:

Completed Non-DR python notebook

Documentation part of preliminary results, Project Management

Contributed 30% of overall effort.

Work to be Completed

Description

Evaluation of the following outstanding metrics for the CTG dataset

Specificity, Recall, F1-score, Sensitivity, Precision

PCA and LDA techniques to be implemented on 2 more datasets and metric evaluation

Diabetic retinopathy dataset

Intrusion Detection dataset

Contribution and Task

Bhavana:

Complete PCA analysis and coding

Documentation of visualization of results.

Contribute 33% to overall effort.

Senthil:

Complete Non DR analysis and coding

Documentation of visualization of results and document cleanup activities

Contribute 34% to overall effort.

Manasa:

Complete PCA analysis and coding

Documentation of visualization of results.

Contribute 33% to overall effort.

# References/Bibliography

## Github - https://github.com/senthilrajendran1/UNT.git