Feature Dimension Reduction

Presented by

Chowdhury Junayed Mahmud Abrar
ID – 15701059
Session 2014-2015
Department of Computer Science and
Engineering,
University of Chittagong

Supervised by

Nihad Karim Chowdhury
Associate Professor
Department of Computer Science and
Engineering
**University of Chittagong**

Contents

- Problem Statements

- Dimension Reduction

- Motivation

- Existing Methods

- How PCA works

- Features of the Proposed System

- Methodology

- Result

- Comparison with other methods

- Limitations

- Future works

Problem Statements

Problems for training dataset with high dimension

- Visualization problem

- Work on it

- Storage problem
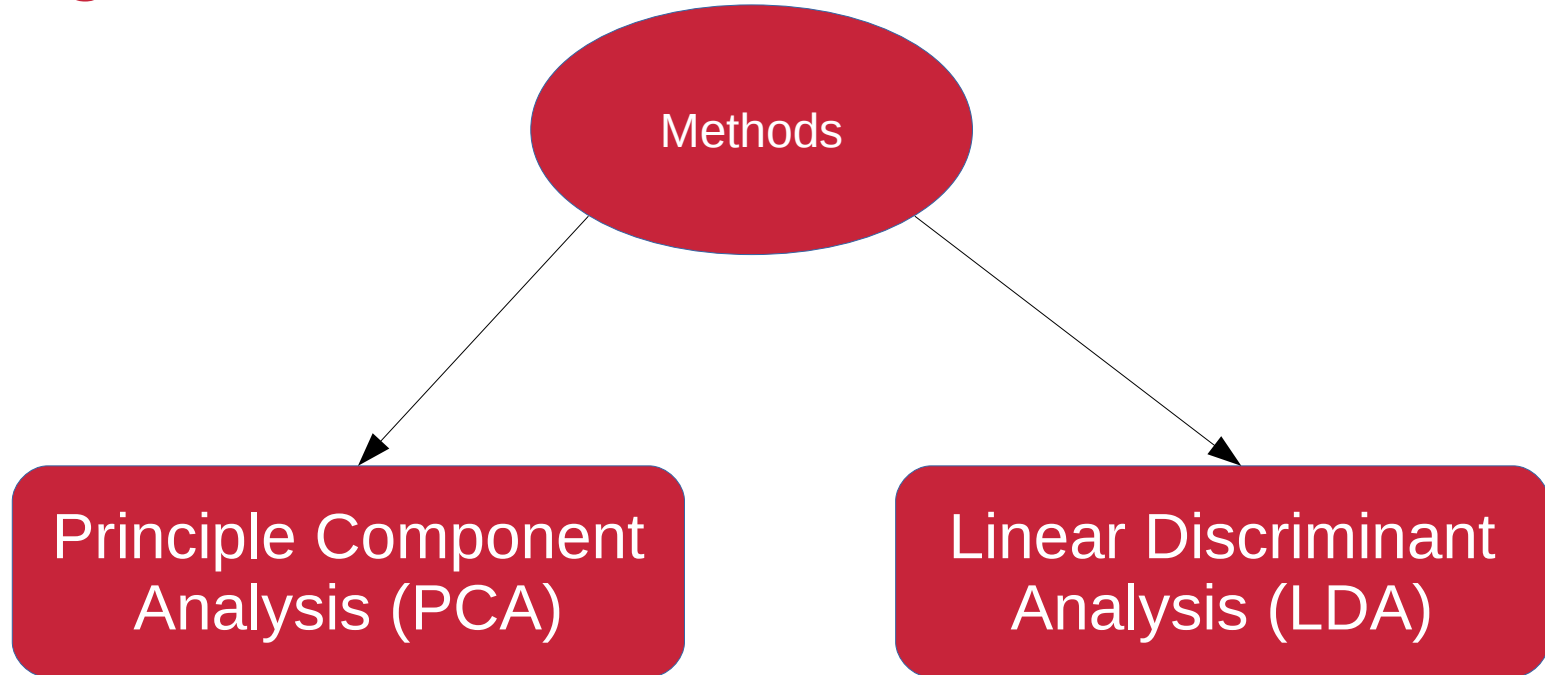
Dimension Reduction
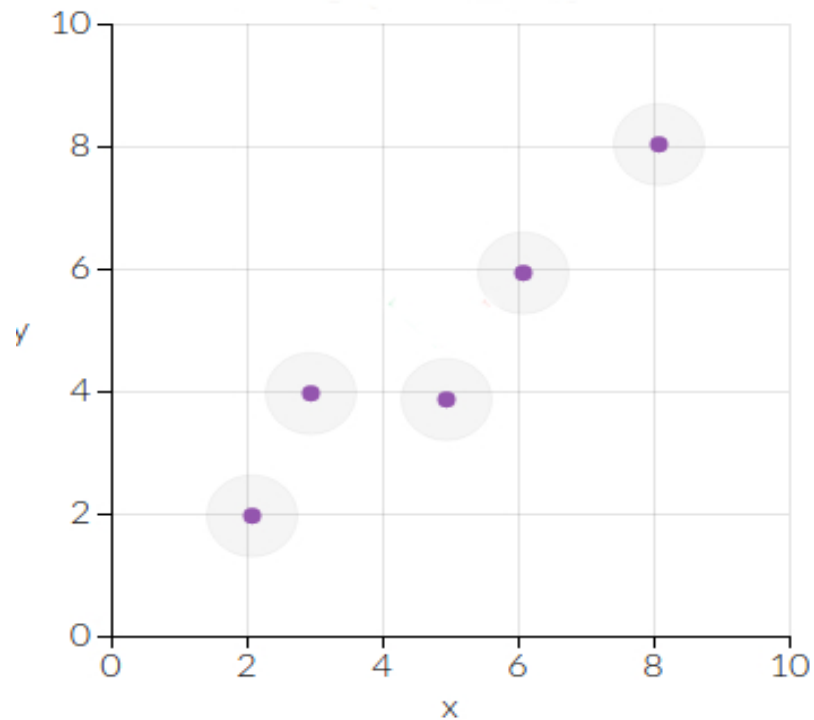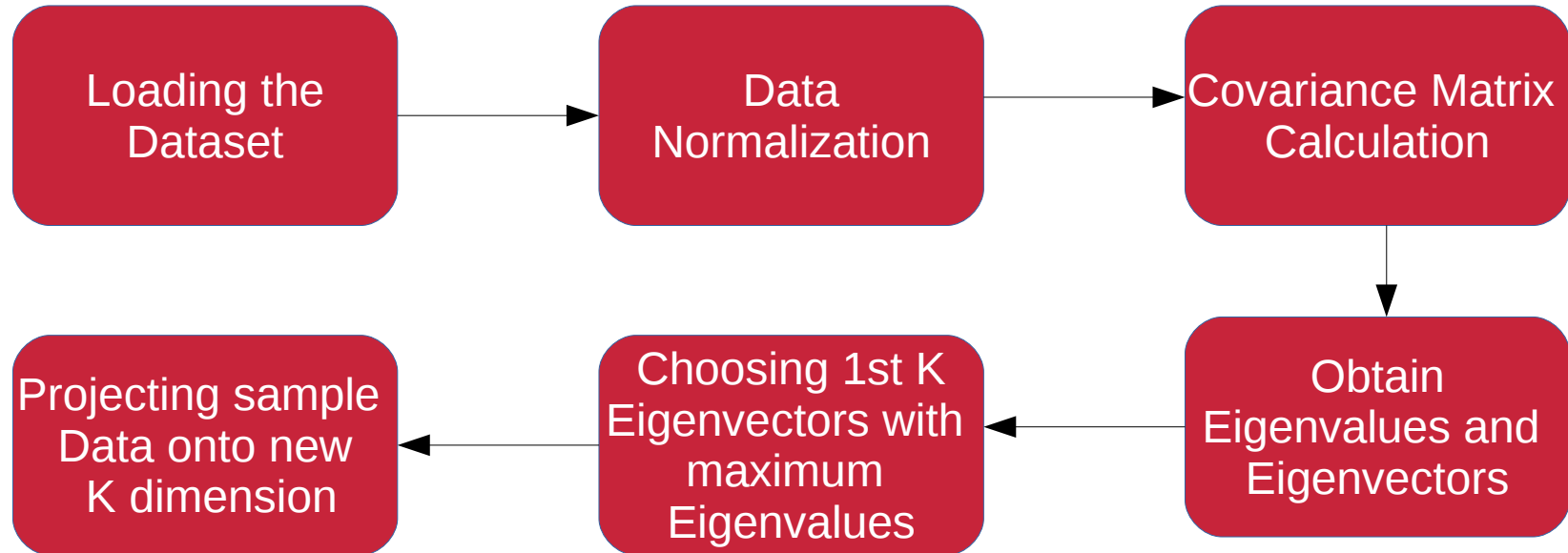

Dimension
Reduction (2D to 1D)

Motivation

- Projecting high-dimensional data onto 2-3 dimensions

- Convert them to data points with minimum loss of information

- Ineffectiveness of most machine learning and data mining techniques on high-dimensional data
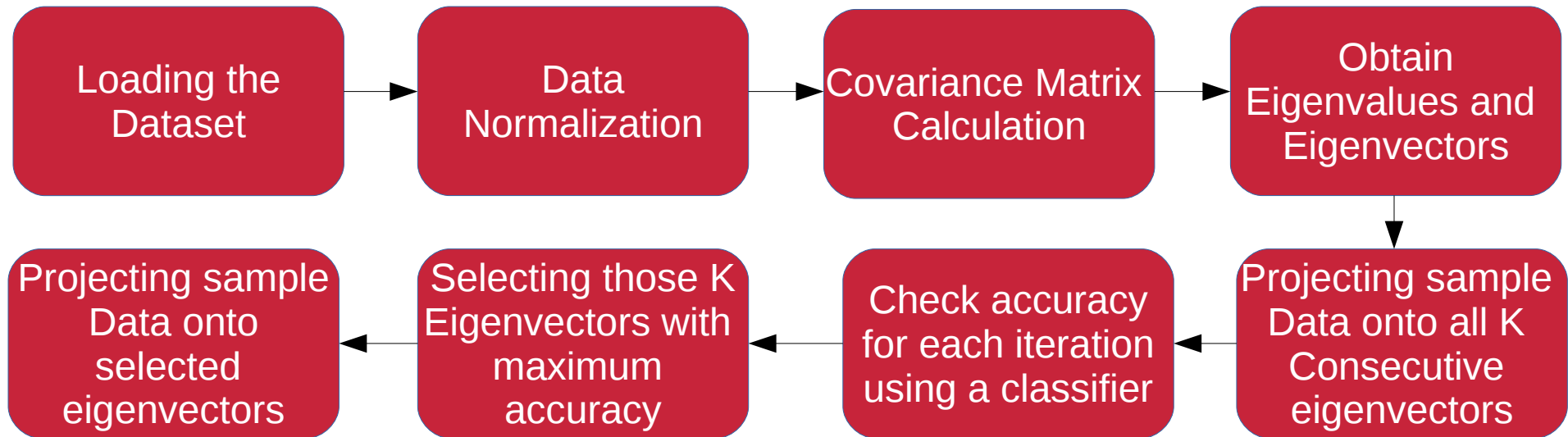
Existing Methods

How PCA works

Features of the Proposed System

- Modified from PCA

- Supervised Learning Method

- PCA takes always high variance dimensions. But, this method choose dimensions which is account for better accuracy

- When the training set is small, modified PCA has lower error rates than PCA

- Projected data can classify with better accuracy like LDA

Methodology

Loading the Dataset → Data Normalization → Covariance Matrix Calculation → Obtain Eigenvalues and Eigenvectors

↓

Projecting sample Data onto selected eigenvectors ← Selecting those K Eigenvectors with maximum accuracy ← Check accuracy for each iteration using a classifier ← Projecting sample Data onto all K Consecutive eigenvectors

Here, K = Number of Selected Dimensions

Results

Classification Accuracy Table for some datasets by Projecting Data onto 2 Dimensions:

Attribute	Instance	Classification Accuracy (Without Reduction)	Classification Accuracy (Applying PCA)	Classification Accuracy (Applying Modified PCA)	Classification Accuracy (Applying LDA)
5	80	62.5%	62.5%	75.0%	75.0%
4	748	73.3%	74.6%	77.3%	77.3%
23	1000	100.0%	92.0%	99.0%	100.0%
5	10000	99.2%	98.6%	98.6%	99.1%

Comparison with other methods
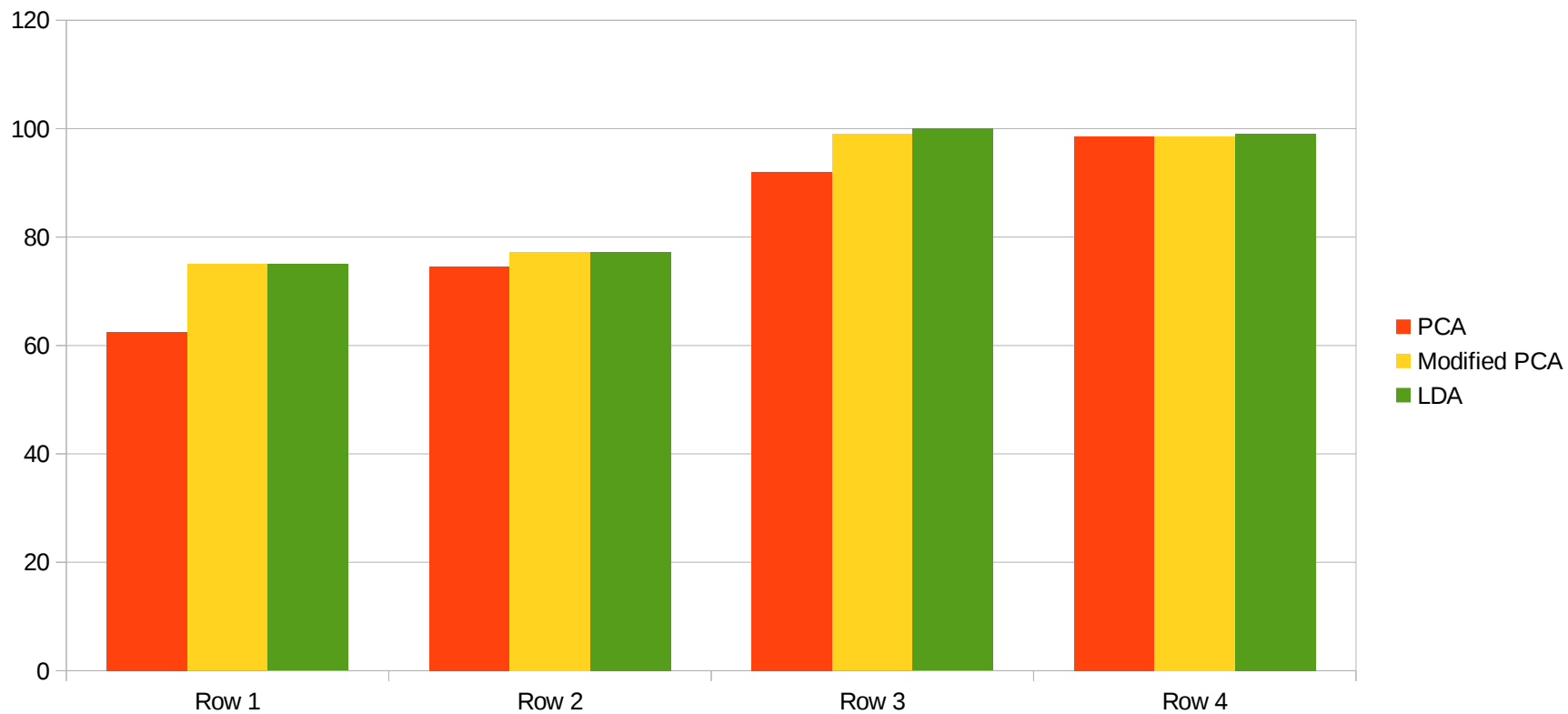
Limitations

- Not a good choice for large training datasets

- For large datasets, it will take much more time for computation than PCA

- Target class is necessary to implement it.

Future works

- Reduce computation for large datasets

- With more accuracy

Thank You