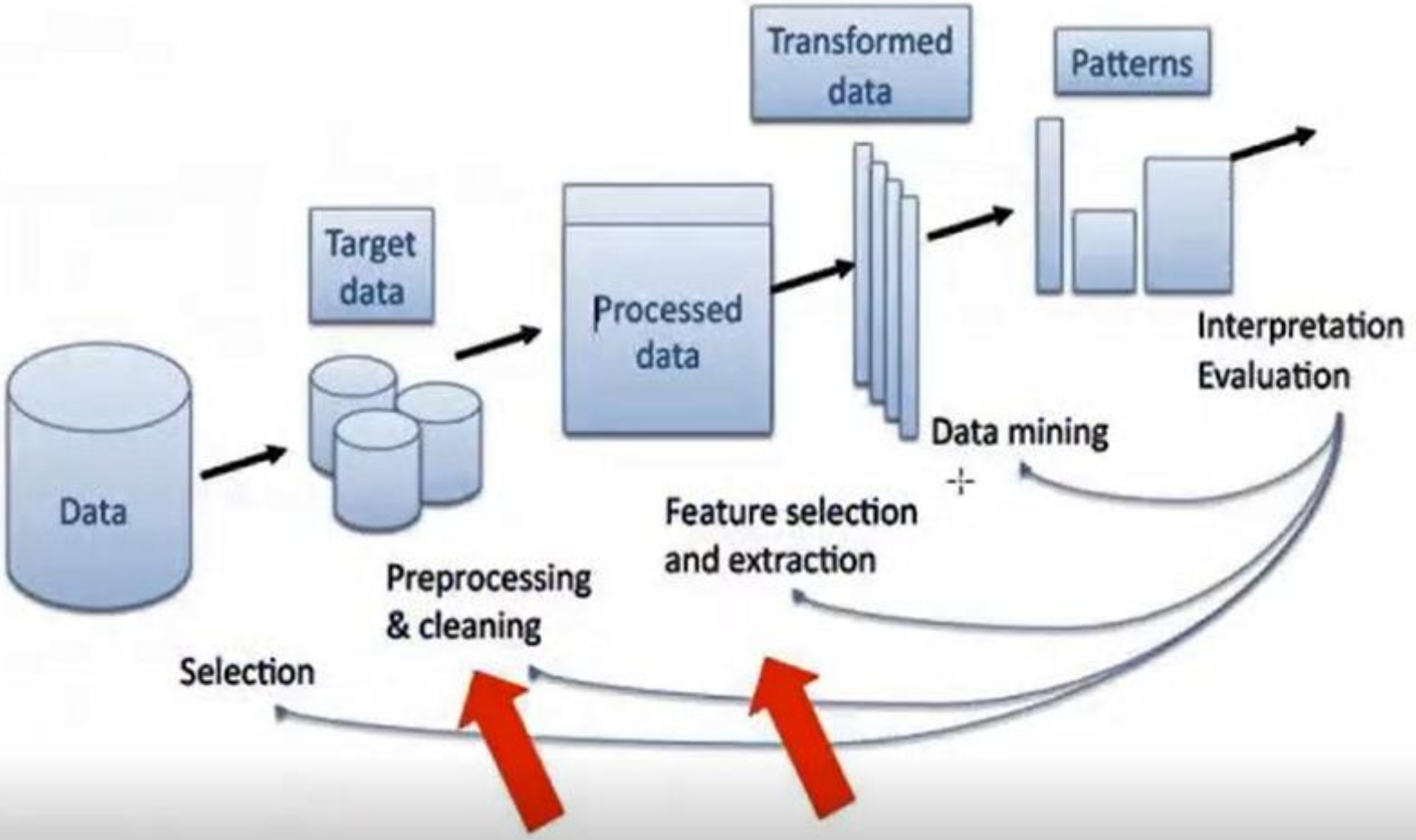
Overview

Introduction



Introduction

innovate achieve lead

Why we need FS:

- to improve performance (in terms of speed, predictive power, simplicity of the model).
- To visualize the data for model selection.
- To reduce dimensionality and remove noise.

Feature Selection is a process that chooses an optimal subset of features according to a certain criterion.

BITS Pilani
Hyderabad Campus

Feature selection methods

innovate achieve lead

Feature ranking techniques:

In Feature ranking technique, some decisive factors have been considered to rank each individual feature and then some features are selected that are suitable for a given project.

Feature subset selection techniques:

In feature subset selection, subset of features are searched which collectively have good predictive capability.

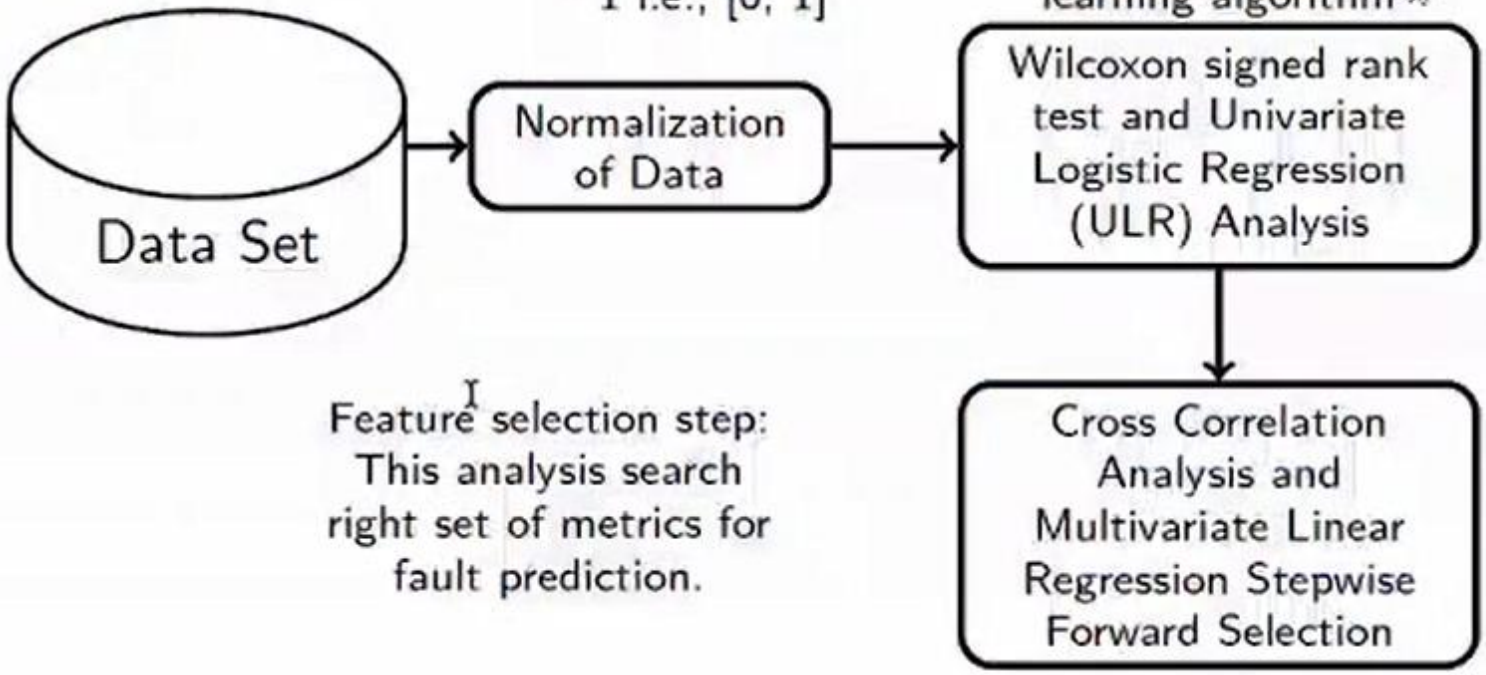
BITS Pilani
Hyderabad Campus

PROPOSED SOFTWARE METRICS VALIDATION METHOD

Data set containing software metrics and fault in software modules

Metrics are normalized over the range between 0 to 1 i.e., [0, 1]

pre-processing step: selection of metrics without involving learning algorithm



Feature selection step: This analysis search right set of metrics for fault prediction.

BITS Pilani
Hyderabad Campus

Python code

Python code

```
from scipy.stats import mannwhitneyu
w,p=mannwhitneyu(f0,f1)
```

Python code

```
from matplotlib import pyplot
pyplot.boxplot(x,labels=['Not-faulty','Faulty'])
pyplot.grid(True)
pyplot.xlabel('Metrics')
pyplot.ylabel('95%CI')
fna='C:/Users/lov/Documents/dsv/'+str(i)+".png"
pyplot.savefig(fna) pyplot.close()
```

Feature ranking methods

Feature ranking methods rank features independently without using any learning algorithm.

In feature ranking methods, ranking of features are based on the score of the features.

Further top $\lceil log_2 n \rceil$ features out of "n" number of features have been considered to develop a model.

Measures of Feature Impurity

innovate achieve lead

Gini Index

$$Gini_{Split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i) \quad (2)$$

$$GINI(t) = 1 - \sum_{i=1}^{k} (p(\frac{j}{t})^2) \quad (3)$$

Entropy

$$Entropy(t) = -\sum_{i=1}^{k} p(\frac{j}{t}) * log_2 p(\frac{j}{t}) \quad (4)$$

BITS Pilani
Hyderabad Campus

Measures of Feature Impurity

Information Gain

$$InformationGain = Entropy(P) - \sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \quad (5)$$

Classification error

$$Error(t) = 1 - max(p(\frac{j}{t})) \quad (6)$$

Measures of Feature Impurity

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Feature-subset selection method

Sequential Forward Generation (SFG):

It starts with an empty set of features S. As the search starts, features are added into S according to some criterion that distinguish the best feature from the others. S grows until it reaches a full set of original features. The stopping criteria can be a threshold for the number of relevant features m or simply the generation of all possible subsets in brute force mode.

Sequential Backward Generation (SBG):

It starts with a full set of features and, iteratively, they are removed one at a time. Here, the criterion must point out the worst or least important feature. By the end, the subset is only composed of a unique feature, which is considered to be the most informative of the whole set. As in the previous case, different stopping criteria can be used.

Principal Component Analysis (PCA)

Attribute reduction using Principal Component analysis (PCA) is achieved by transforming high dimension data space into lower dimension data space.

Takes a data matrix of n objects by p variables, which may be correlated, and summarizes it by uncorrelated axes (principal components or principal axes) that are linear combinations of the original p variables

Procedures of PCA

innovate achieve lead

1st Step

Pre-treatment of Data Matrix→Scaling

2nd Step

Calculation of covariance matrix

3rd Step

Calculation of eigenvalues and eigenvectors of covariance matrix

4th Step

Calculation of scores

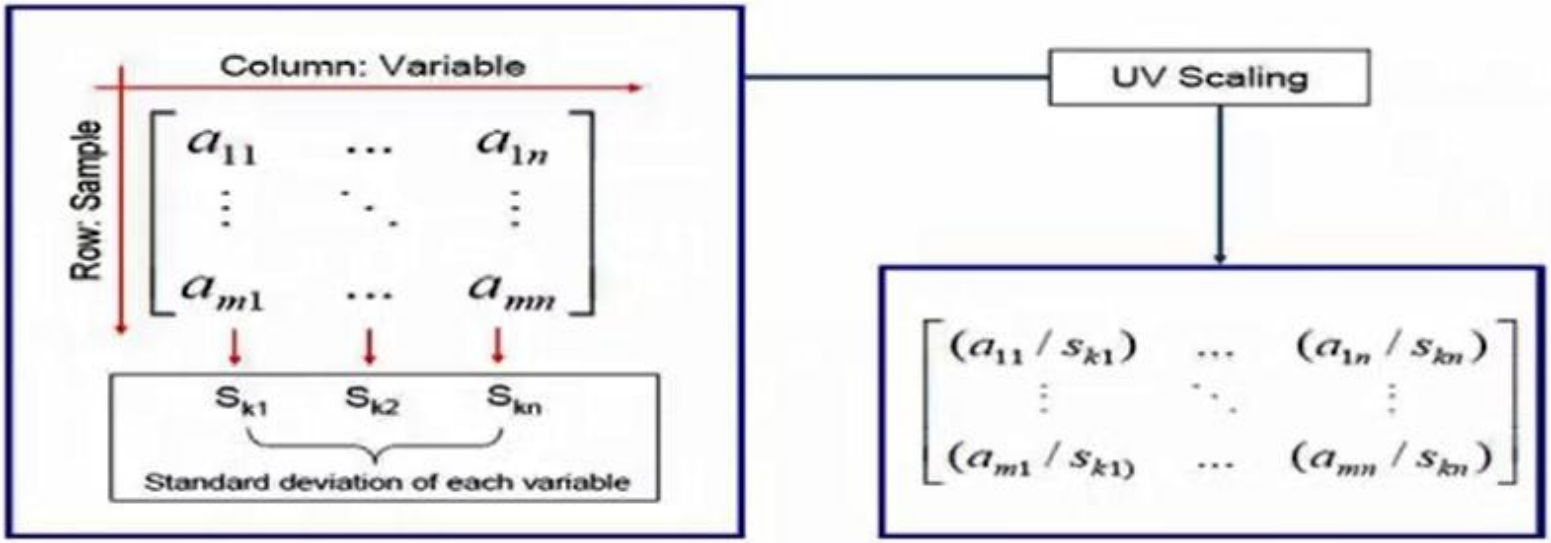
BITS Pilani
Hyderabad Campus

1st Step: Pre-treatment of Data Matrix

innovate achieve lead

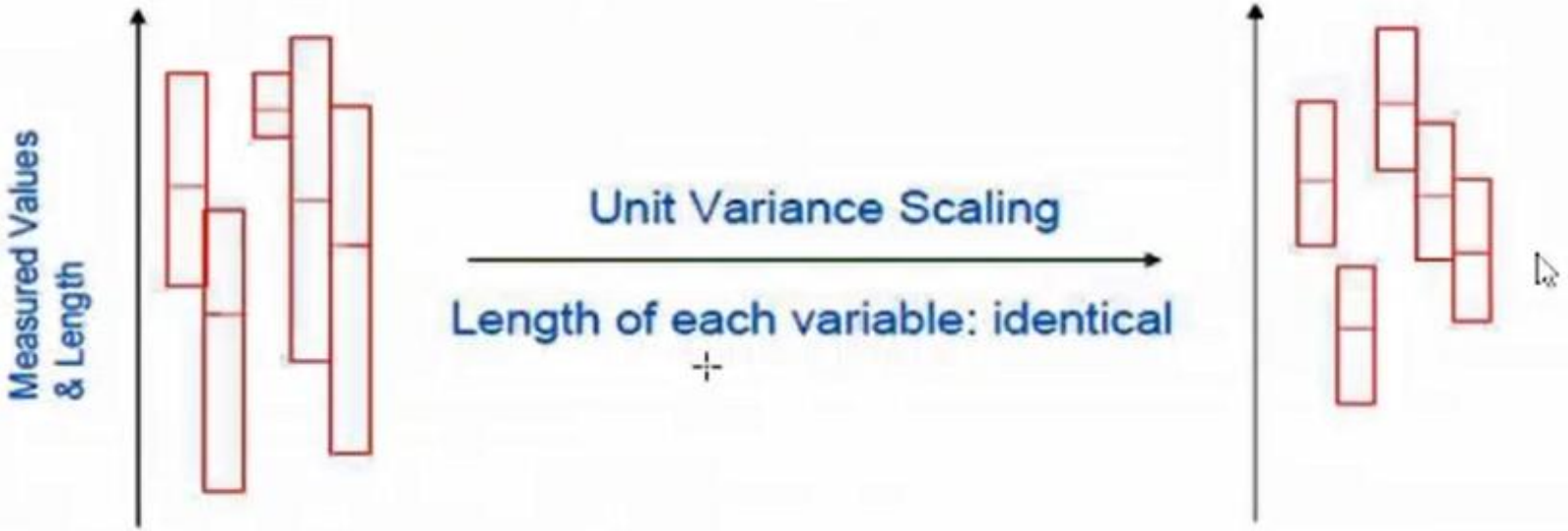
Pre-treatment of Data Matrix-Scaling

- Unless the data are normalized, a variable with a large variance will dominate
- Most common scaling technique - Unit variance (UV) scaling



BITS Pilani
Hyderabad Campus

1st Step: Pre-treatment of Data Matrix

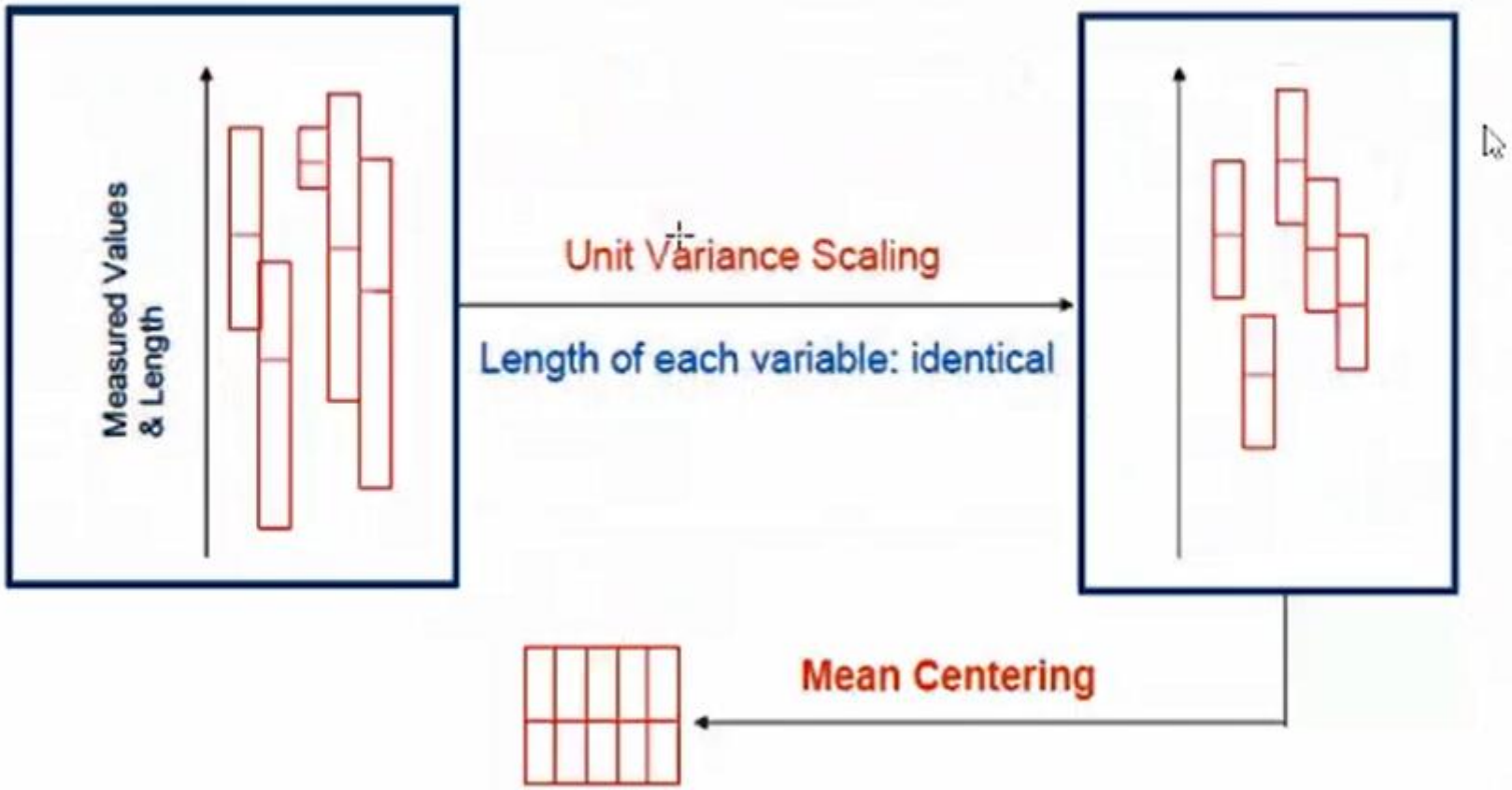


Pre-treatment of Data Matrix-Scaling

- Note: However, the mean values still remain different
- Therefore mean-centering as a second part of pre-data processing
  - Average value of each variable is calculated
  - Subtracted from the data

1st Step: Pre-treatment of Data Matrix

1) Pre-treatment of Data Matrix- Scaling (Continued)



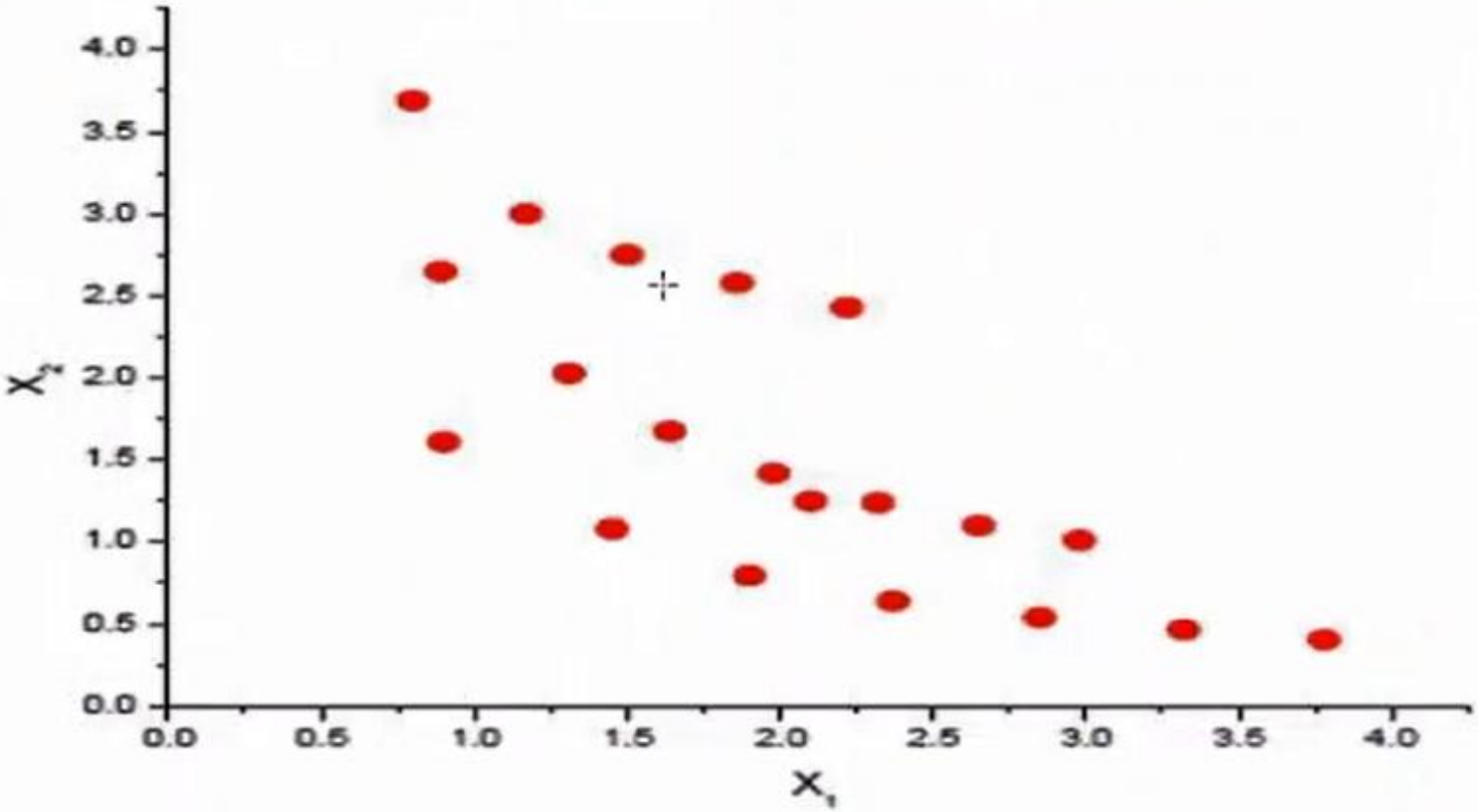
Example of Data Matrix: X

innovate achieve lead

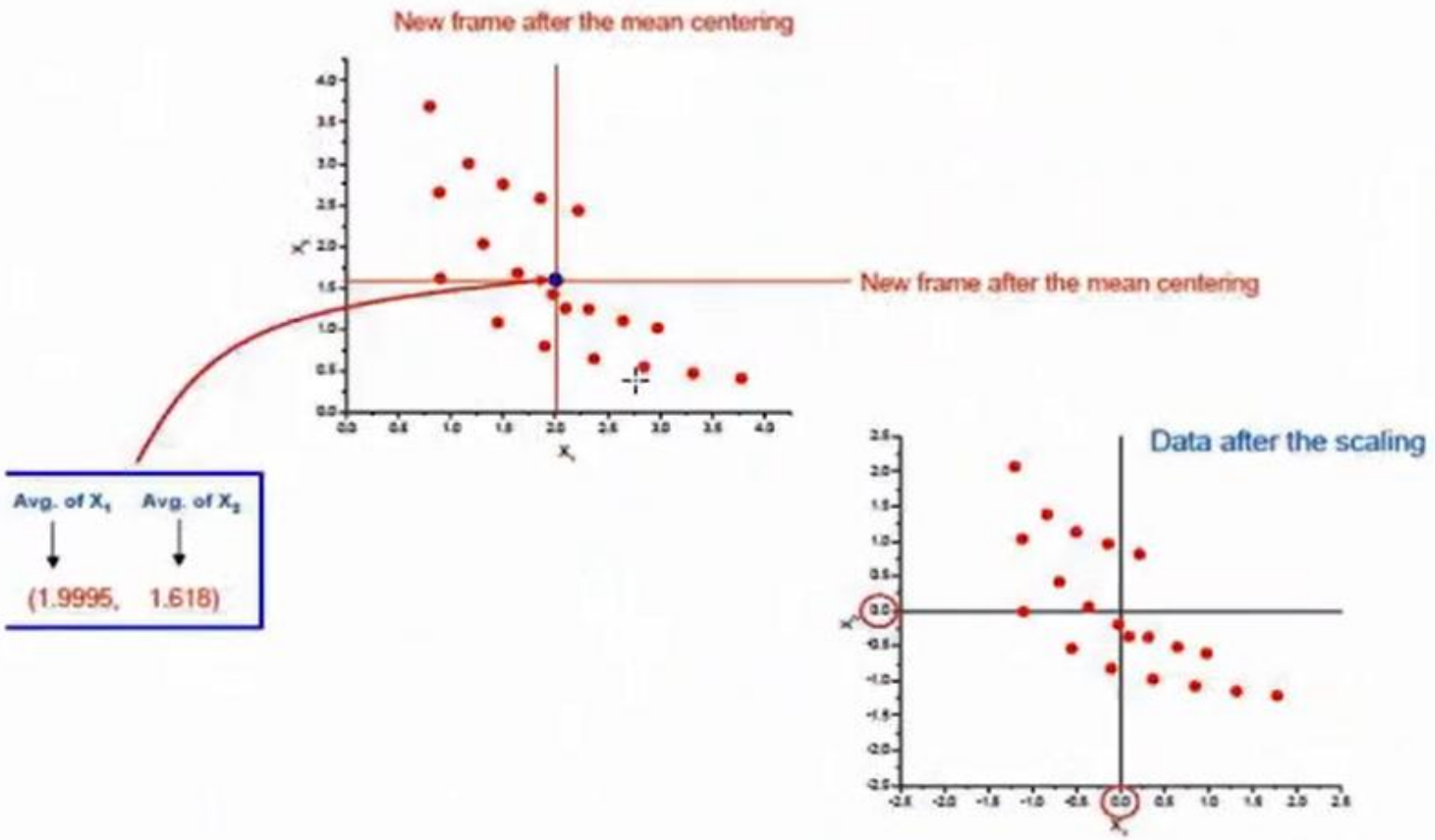
Sample NO.	Element	Martynov-Batsanov's Electronegativity (X_1)	Zunger's pseudopotenial core radii sum (X_2)	Sample NO.	Element	Martynov-Batsanov's Electronegativity (X_1)	Zunger's pseudopotential core radii sum (X_2)
1	H	2.1	1.25	11	Al	1.64	1.675
2	Li	0.9	1.61	12	Si	1.98	1.42
3	Be	1.45	1.08	13	P	2.32	1.24
4	B	1.9	0.795	14	S	2.65	1.1
5	C	2.37	0.64	15	Cl	2.98	1.01
6	N	2.85	0.54	16	K	0.8	3.69
7	O	3.32	0.485	17	Ca	1.17	3
8	F	3.78	0.405	18	Sc	1.5	2.75
9	Na	0.89	2.65	19	Ti	1.86	2.58
10	Mg	1.31	2.03	20	V	2.22	2.43

Example of Data Matrix: X

The scatter plot of X



Example of Data Matrix: X



2nd Step: Calculation of covariance mat

Pre-treatment of Data Matrix-Scaling

- Calculation of Covariance matrix(S) of Data Matrix(X)
  - Variance (1 dimensional concept): Measure of the spread of data in a given data set
  - Covariance (Multi-dimensional concept): Measure of the spread of data between dimensions (variables)

$$Cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})}{n-1} \quad (7)$$

3rd Step of PCA: Calculation of eigenva eigenvectors of covariance matrix

$$S = \begin{bmatrix} 0.6881 & -0.5929 \\ -0.5929 & 0.9026 \end{bmatrix} \rightarrow \begin{vmatrix} 0.6881-\lambda & -0.5929 \\ -0.5929 & 0.9026-\lambda \end{vmatrix} = 0$$

$$\downarrow$$

$$(0.6881-\lambda)(0.9026-\lambda)-(0.5929)^2 = 0$$

$$\downarrow$$

$$\lambda^2 - 1.5907\lambda + 0.27 = 0$$

$$\downarrow$$

$$\lambda = \frac{1.5907 \pm \sqrt{(1.5907)^2 - 4\times 0.27}}{2}$$

$$\downarrow$$

$$\lambda_1 = 1.3978,\ \lambda_2 = 0.1928$$

Eigenvalues of covariance matrix S

BITS Pilani
Hyderabad Campus

Genetic algorithms

- Genetic algorithms are one of the few accepted techniques used for their ability to efficiently search large space about which little is known as inferable .
- Here two objective functions have been developed to evaluate the fitness value of each chromosome.
- The objective functions are based on the concept of minimization of number of attributes and maximization of accuracy.
- The fitness function may be formulated as:

$$F = minimize(\sqrt{(\frac{n}{m})^2 + (\frac{1}{Accuracy})^2}) \quad (8)$$

Genetic algorithms

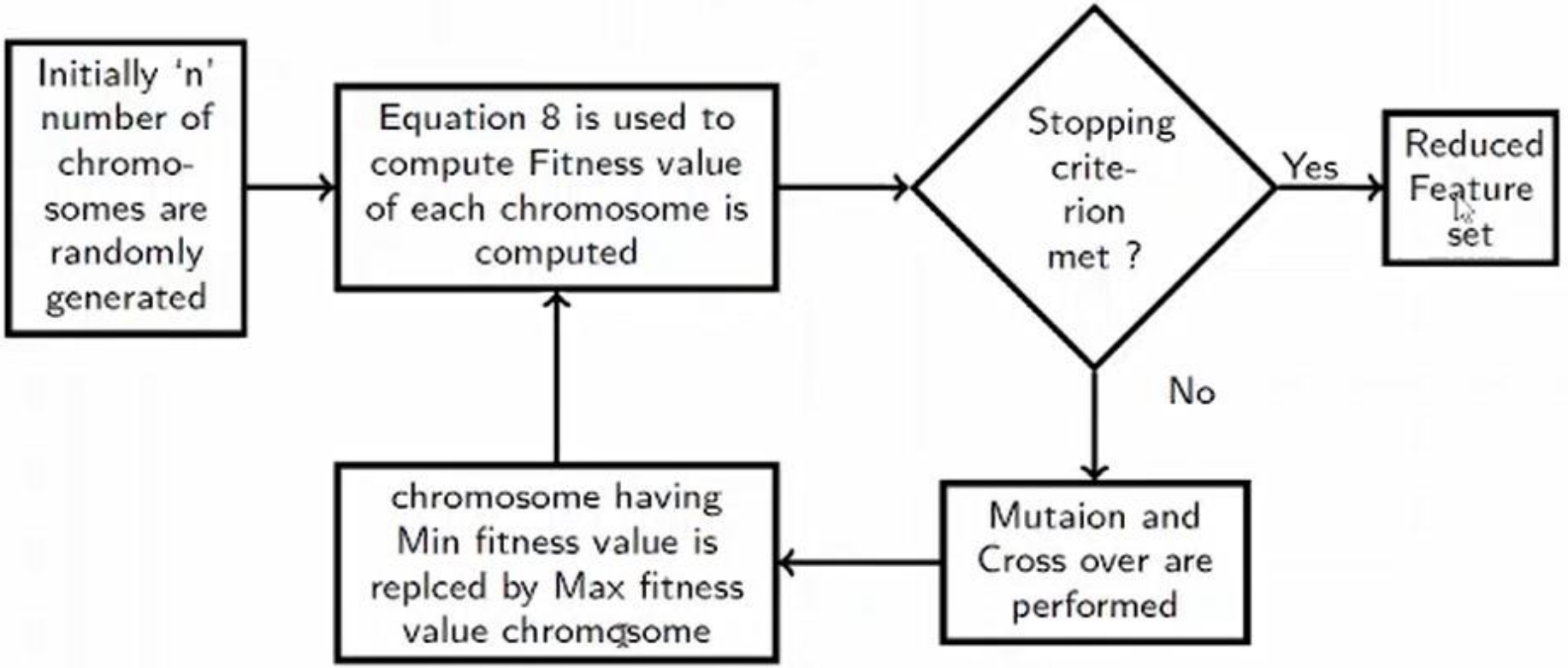


Figure: Flow chart representing GA execution

Any Question Please ?

BITS Pilani
Hyderabad Campus