Chapter – 4

A study material for the students of GLS University
Compiled by Dr. Krupa Mehta

Measures of Feature redundancy

- Correlation-based measures
- Distance-based measures, and
- Other coefficient-based measure

Correlation-based similarity measure

- measure of linear dependency between two random variables
- Pearson's product moment correlation coefficient
- Correlation value
- between +1 and

$$\alpha = \frac{cov(F_1, F_2)}{\sqrt{var(F_1).var(F_2)}}$$

$$cov(F_1, F_2) = \sum (F_{1_i} - \overline{F_1}).(F_{2_i} - \overline{F_2})$$

$$var(F_1) = \sum (F_{1_i} - \overline{F_1})^2, \text{where } \overline{F_1} = \frac{1}{n}.\sum F_{1_i}$$

$$var(F_2) = \sum (F_{2_i} - \overline{F_2})^2, \text{where } \overline{F_2} = \frac{1}{n}.\sum F_{2_i}$$

Distance-based similarity measure

- Euclidean distance

$$d(F_1, F_2) = \sqrt{\sum_{i=1}^{n} (F_{1_i} - F_{2_i})^2}$$

Aptitude (F_1)	Communication (F_2)	($F_1 - F_2$)	($F_1 - F_2$)^2
2	6	–4	16
3	5.5	–2.5	6.25
6	4	2	4
7	2.5	4.5	20.25
8	3	5	25
6	5.5	0.5	0.25
6	7	–1	1
7	6	1	1
8	6	2	4
9	7	2	4
			81.75

Distance-based similarity measure

- A more generalized form of the Euclidean distance is the **Minkowski distance**
- L_2 norm (when r= 2)

$$d(F_1, F_2) = \sqrt[r]{\sum_{i=1}^{n} (F_{1_i} - F_{2_i})^r}$$

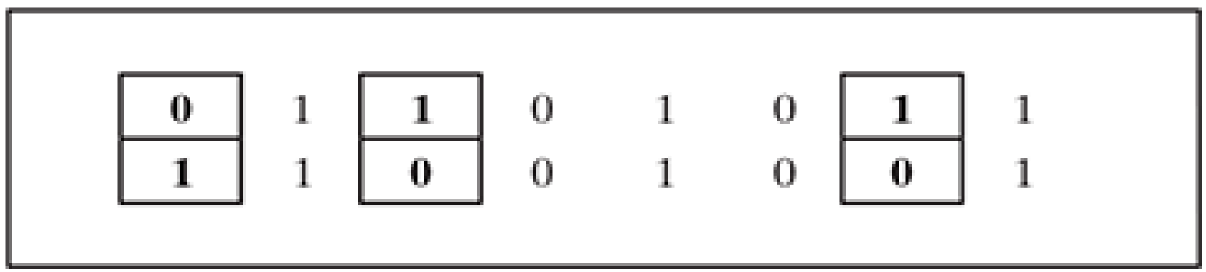
- Manhattan distance = L_1 norm=r·1

$$d(F_1, F_2) = \sum_{i=1}^{n} |F_{1_i} - F_{2_i}|$$

-

Distance-based similarity measure

- Hamming distance: calculate the distance between binary vectors

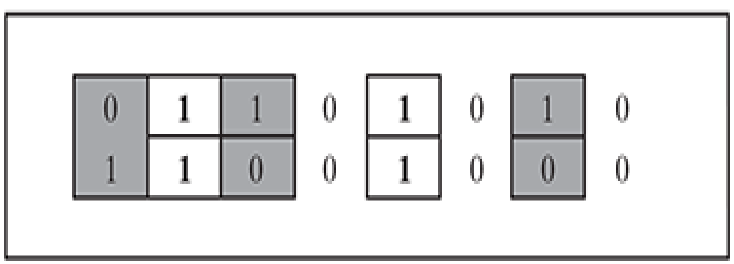


(a) Hamming distance measurement

Other similarity measures

$$J = \frac{n_{11}}{n_{01} + n_{10} + n_{11}}$$

- Jaccard index/coefficient is used as a measure of similarity between two features
  - n_{11} = number of cases where both the features have value 1
  - n_{01} = number of cases where the feature 1 has value 0 and feature 2 has value 1
  - n_{10} = number of cases where the feature 1 has value 1 and feature 2 has value 0



$$J = \frac{n_{11}}{n_{01} + n_{10} + n_{11}}$$

Other similarity measures

$$SMC = \frac{n_{11} + n_{00}}{n_{00} + n_{01} + n_{10} + n_{11}}$$

- **Simple matching coefficient (SMC)**
  - n_{11} = number of cases where both the features have value 1
  - n_{01} = number of cases where the feature 1 has value 0 and feature 2 has value 1
  - n_{10} = number of cases where the feature 1 has value 1 and feature 2 has value 0
  - n_{00} = number of cases where both the features have value 0

$$\therefore \text{SMC of } F_1 \text{ and } \mathrm{F}_2 = \frac{n_{11} + n_{00}}{n_{00} + n_{01} + n_{10} + n_{11}} = \frac{2 + 3}{3 + 1 + 2 + 2} = \frac{1}{2} \text{ or } 0.5.$$

Other similarity measures

$$\|x\| = \sqrt{\sum_{i=1}^{n} x_i^2} \text{ and } \|y\| = \sqrt{\sum_{i=1}^{n} y_i^2}$$

- **Cosine Similarity**

$$cos(x, y) = \frac{x.y}{\|x\|.\|y\|}$$

  - **x = (2,4, 0, 0, 2, 1, 3, 0, 0)**
  - **y = (2, 1, 0, 0, 3, 2, 1, 0, 1)**
  - **x.y = 2*2 + 4*1 + 0*0 + 0*0 + 2*3 + 1*2 + 3*1 + 0*0 + 0*1 = 19**

$$\|x\| = \sqrt{2^2 + 4^2 + 0^2 + 0^2 + 2^2 + 1^2 + 3^2 + 0^2 + 0^2} = \sqrt{34} = 5.83$$

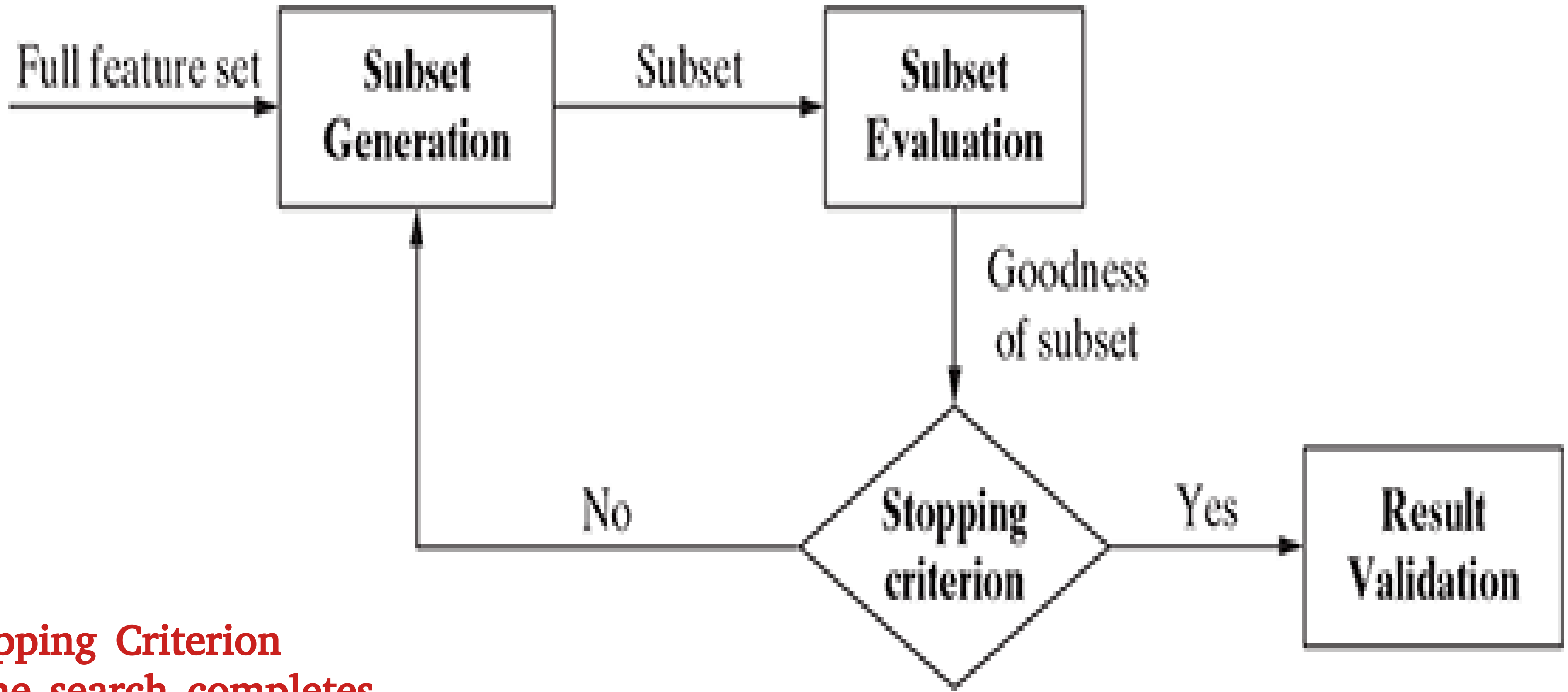
$$\|y\| = \sqrt{2^2 + 1^2 + 0^2 + 0^2 + 3^2 + 2^2 + 1^2 + 0^2 + 1^2} = \sqrt{20} = 4.47$$

$$\therefore cos(x, y) = \frac{19}{5.83*4.47} = 0.729$$

$\boxed{43.2°}$

  - **If cosine similarity has a value 1, the angle between x and y is 0° . 0 = 90°**

Overall feature selection process

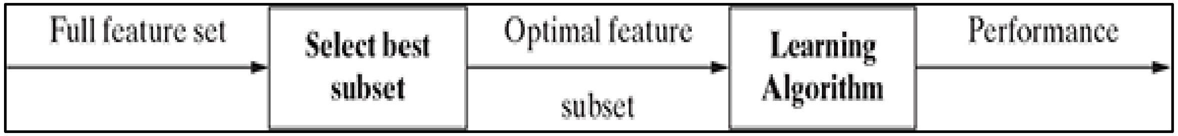


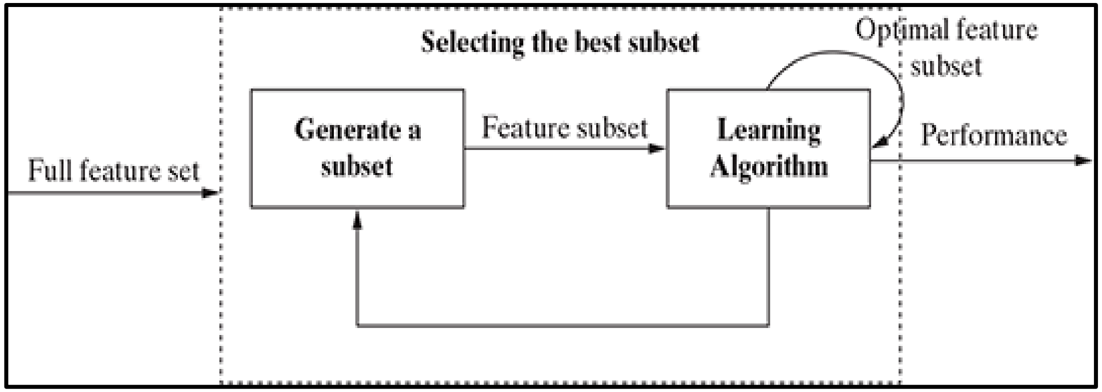
Stopping Criterion

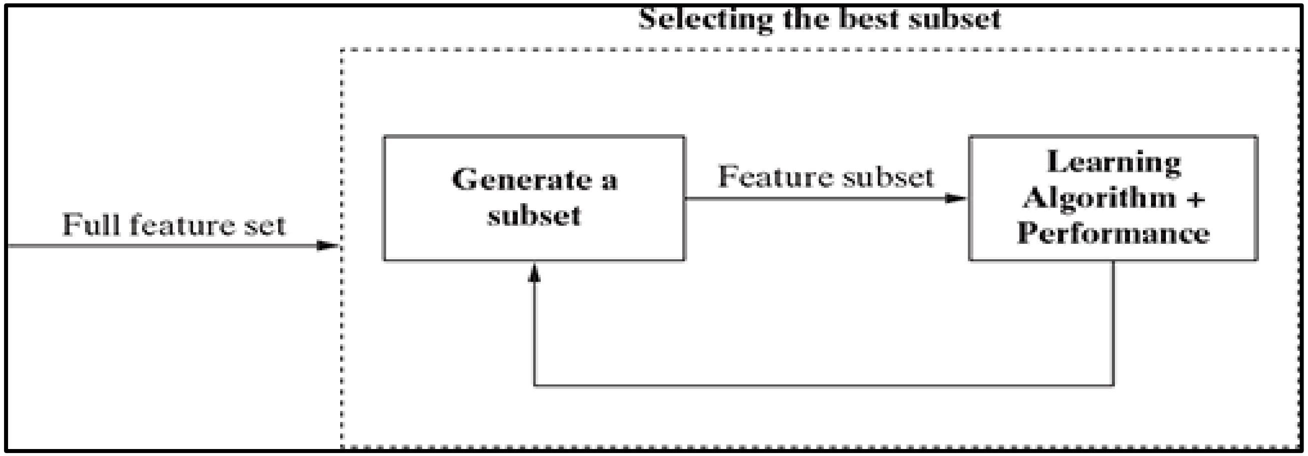
- the search completes
- some given bound (e.g. a specified number of iterations) is reached
- subsequent addition (or deletion) of the feature is not producing a better subset
- a sufficiently good subset (e.g. a subset having better classification accuracy than the existing benchmark) is selected

Feature selection approaches

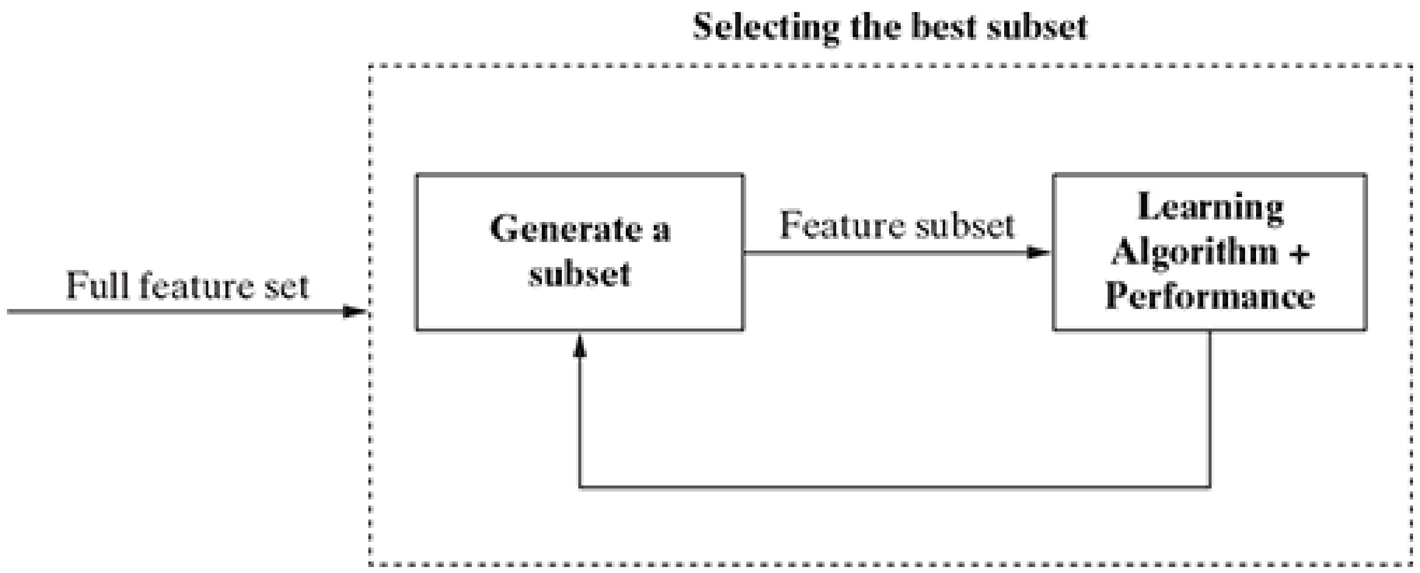
- **Filter approach**
- **Wrapper approach**
- **Hybrid approach**
- **Embedded approach**







- **Embedded approach**



Reference

- Machine Learning by Saikat Dutt, Subramanian Chandramouli, Amit Kumar Das published by Pearson