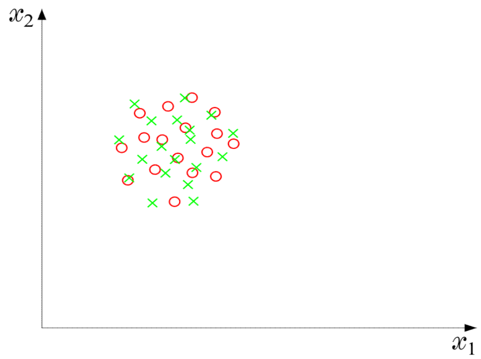
FEATURE SELECTION

- The goals:
  - Select the "optimum" number l of features
  - Select the "best" l features

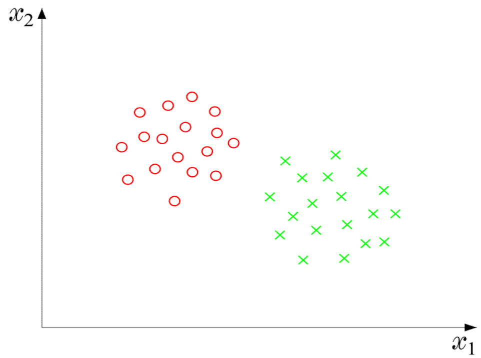
- Large l has a three-fold disadvantage:
  - High computational demands
  - Low generalization performance
  - Poor error estimates

- Given N
  - l must be large enough to learn
    - what makes classes different
    - what makes patterns in the same class similar
  - l must be small enough not to learn what makes patterns of the same class different.
  - In practice, $l < N/3$ has been reported to be a sensible choice for a number of cases.

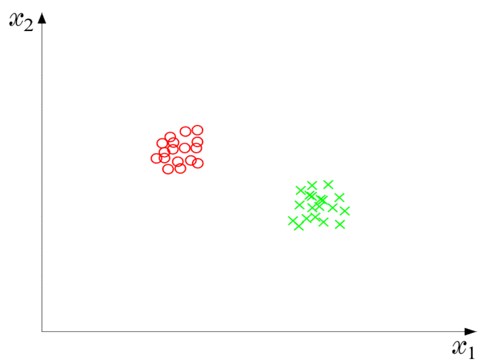
- Once l has been decided, choose the l most informative features
  - Best: **Large** between class distance, **Small** within class variance



Bad choice



Not bad choice



Good choice

- The basic philosophy
  - Discard individual features with poor information content
  - The remaining information rich features are examined jointly as vectors

- Feature Selection Based on Statistical Hypothesis Testing
  - **The Goal:** For each individual feature, find whether the values, which the feature takes for the different classes, differ significantly.

    That is, answer

$$\begin{cases} H_1: & \text{The values of the feature differ significantly} \\ H_0: & \text{The values of the feature do not differ significantly} \end{cases}$$

    If they do not differ significantly reject feature from subsequent stages.

- Hypothesis Testing Basics

- The steps:
  - N measurements $x_i, i = 1,2,\ldots,N$ are known
  - Define a function of them

    $q = f(x_1, x_2, \ldots, x_N)$: test statistic

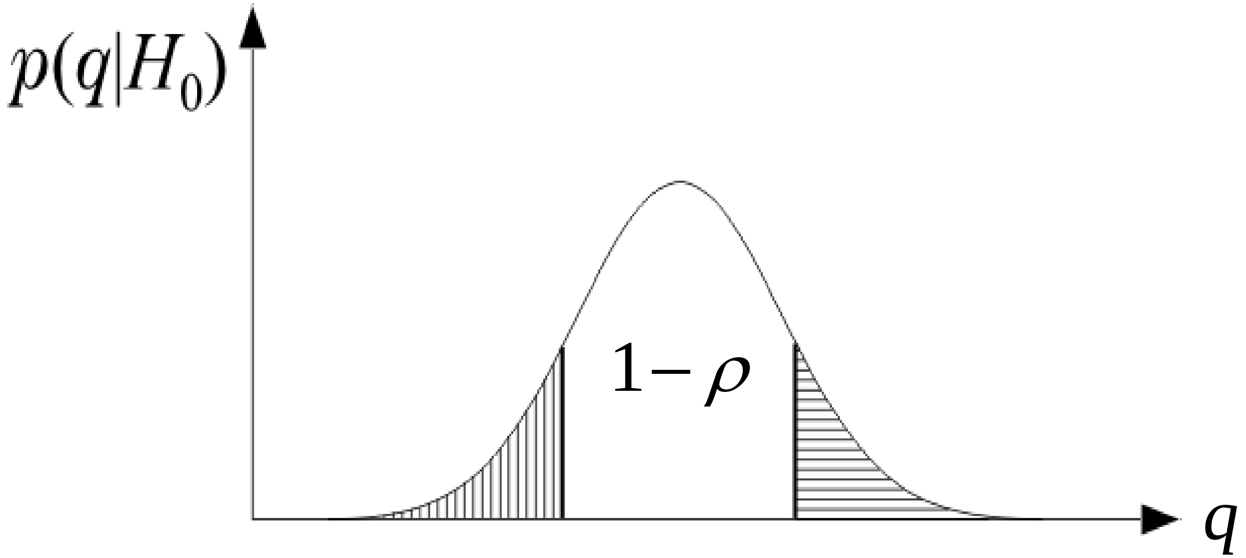
    so that $p_q(q;\theta)$ is easily parameterized in terms of θ.
  - Let D be an interval, where q has a high probability to lie under H_0, i.e., $p_q(q|\theta_0)$
  - Let $\bar{D}$ be the complement of D

    $D \longrightarrow$ Acceptance Interval

    $\bar{D} \longrightarrow$ Critical Interval
  - If q, resulting from $x_1, x_2, \ldots, x_N$, lies in D we accept H_0, otherwise we reject it.

➢ Probability of an error

$$p_q(q \in \overline{D} | H_0) = \rho$$



- ρ is preselected and it is known as the significance level.

❖ Application: The known variance case:

- Let x be a random variable and the experimental samples, $x_i = 1,2,\ldots,N$, are assumed mutually independent. Also let

$$E[x] = \mu$$

$$E[(x-\mu)^2] = \sigma^2$$

- Compute the sample mean

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- This is also a random variable with mean value

$$E[\bar{x}] = \frac{1}{N}\sum_{i=1}^{N} E[x_i] = \mu$$

That is, it is an Unbiased Estimator

- The variance $\sigma_{\bar{x}}^2$

$$E[(\bar{x}-\mu)^2] = E[(\frac{1}{N}\sum_{i=1}^{N} x_i - \mu)^2]$$

$$= \frac{1}{N^2}\sum_{i=1}^{N} E[(x_i-\mu)^2] + \frac{1}{N^2}\sum_i \sum_j E[(x_i-\mu)(x_j-\mu)]$$

Due to independence

$$\sigma_{\bar{x}}^2 = \frac{1}{N}\sigma_x^2$$

That is, it is Asymptotically Efficient

- Hypothesis test

$$H_1: \ E[x] \neq \hat{\mu}$$

$$H_0: \ E[x] = \hat{\mu}$$

- Test Statistic: Define the variable $q = \frac{\bar{x}-\hat{\mu}}{\sigma/\sqrt{N}}$

- Central limit theorem under H_0

$$p_{\bar{x}}(\bar{x}) = \frac{\sqrt{N}}{\sqrt{2\pi}\sigma} \exp\left(-\frac{N(\bar{x}-\hat{\mu})^2}{2\sigma^2} \right) \qquad \bar{x} \sim N\left(\hat{\mu}, \tfrac{\sigma^2}{N}\right)$$

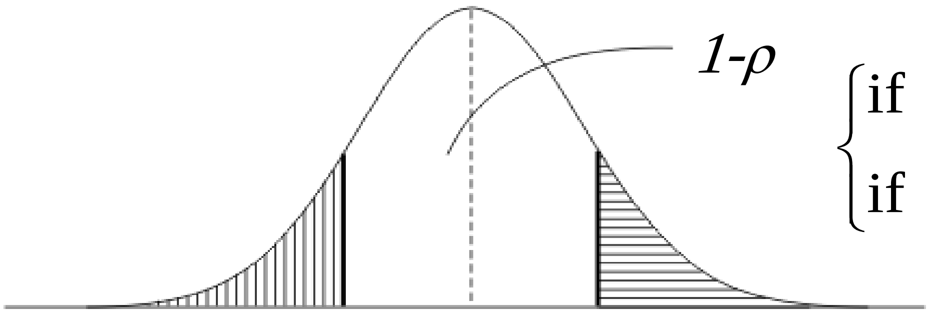
- Thus, under H_0

$$p_q(q) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{q^2}{2} \right) \quad q \sim N(0,1)$$

$$q = \frac{\bar{x}-\hat{\mu}}{\sigma / \sqrt{N}}$$

➢ The decision steps

- Compute q from x_i, $i = 1, 2, \ldots, N$
- Choose significance level ρ
- Compute from $N(0,1)$ tables $D = [-x_\rho, x_\rho]$



$$\begin{cases} \text{if} \quad q \in D & \text{accept } H_0 \\ \text{if} \quad q \in \overline{D} & \text{reject } H_0 \end{cases}$$

➢ An example: A random variable x has variance $\sigma^2 = (0.23)^2$. $N = 16$ measurements are obtained giving $\bar{x} = 1.35$. The significance level is $\rho = 0.05$.

Test the hypothesis $\begin{cases} H_0 : \mu = \hat{\mu} = 1.4 \\ H_1 : \mu \neq \hat{\mu} \end{cases}$

- Since σ^2 is known, $q = \dfrac{\bar{x} - \hat{\mu}}{\sigma / 4}$ is $N(0,1)$.

  From tables, we obtain the values with acceptance intervals $[-x_\rho, x_\rho]$ for normal $N(0,1)$

$1-\rho$	0.8	0.85	0.9	0.95	0.98	0.99	0.998	0.999
x_ρ	1.28	1.44	1.64	1.96	2.32	2.57	3.09	3.29

- Thus

$$\text{Prob}\left\{-1.967 < \frac{\bar{x} - \hat{\mu}}{0.23/4} < 1.967\right\} = 0.95$$

or

$$\text{Prob}\left\{-0.113 < \bar{x} - \hat{\mu} < 0.113\right\} = 0.95$$

or

$$\text{Prob}\left\{1.237 < \hat{\mu} < 1.463\right\} = 0.95$$

- Since $\hat{\mu} = 1.4$ lies within the above acceptance interval, we accept H_0, i.e.,

$$\mu = \hat{\mu} = 1.4$$

The interval [1.237, 1.463] is also known as **confidence interval** at the 1 - ρ = 0.95 level.

We say that: There is no evidence at the 5% level that the mean value is not equal to $\hat{\mu}$

❖ The Unknown Variance Case

- Estimate the variance. The estimate

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

is unbiased, i.e.,

$$E[\hat{\sigma}^2] = \sigma^2$$

- Define the test statistic

$$q = \frac{\bar{x} - \mu}{\hat{\sigma}/\sqrt{N}}$$

- This is no longer Gaussian. If x is Gaussian, then q follows a ***t*-distribution**, with N-1 degrees of freedom

$$q = \frac{\bar{x} - \mu}{\hat{\sigma} / \sqrt{N}}$$

- An example:

x is Gaussian, $N = 16$, obtained from measurements, $\bar{x} = 1.35$ and $\hat{\sigma}^2 = (0.23)^2$. Test the hypothesis

$H_0 : \ \mu = \hat{\mu} = 1.4$

at the significance level $\rho = 0.025$.

- Table of acceptance intervals for t-distribution

Degrees of Freedom	1-ρ	0.9	0.95	0.975	0.99
12		1.78	2.18	2.56	3.05
13		1.77	2.16	2.53	3.01
14		1.76	2.15	2.51	2.98
15		1.75	2.13	2.49	2.95
16		1.75	2.12	2.47	2.92
17		1.74	2.11	2.46	2.90
18		1.73	2.10	2.44	2.88

- $$\text{Prob}\left\{-2.49 < \frac{\bar{x}-\hat{\mu}}{\hat{\sigma}/4} < 2.49\right\}$$

$$1.207 < \hat{\mu} < 1.493$$

$$\text{Thus, } \hat{\mu} = 1.4 \text{ is accepted}$$

❖ Application in Feature Selection

- The goal here is to test against zero the difference μ_1-μ_2 of the respective means in ω_1, ω_2 of a single feature.
- Let x_i $i = 1, \ldots, N$, the values of a feature in ω_1
- Let y_i $i = 1, \ldots, N$, the values of the same feature in ω_2
- Assume in both classes $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (unknown or not)
- The test becomes $$\begin{cases} H_0: & \Delta\mu = \mu_1 - \mu_2 = 0 \\ H_1: & \Delta\mu \neq 0 \end{cases}$$

- Define

  $z = x - y$

- Obviously

  $E[z] = \mu_1 - \mu_2$

- Define the average

$$\bar{z} = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i) = \bar{x} - \bar{y}$$

- Known Variance Case: Define $q = \dfrac{(\bar{x} - \bar{y}) - (\hat{\mu}_1 - \hat{\mu}_2)}{\sigma\sqrt{\frac{2}{N}}}$

  - This is $N(0,1)$ and one follows the procedure as before.

- Unknown Variance Case:

Define the test statistic

$$q = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{S_z \sqrt{\frac{2}{N}}}$$

$$S_z^2 = \frac{1}{2N - 2} \left(\sum_{i=1}^{N} (x_i - \bar{x})^2 + \sum_{i=1}^{N} (y_i - \bar{y})^2 \right)$$

- q is t-distribution with $2N$-2 degrees of freedom,
- Then apply appropriate tables as before.

- Example: The values of a feature in two classes are:

ω_1: 3.5, 3.7, 3.9, 4.1, 3.4, 3.5, 4.1, 3.8, 3.6, 3.7

ω_2: 3.2, 3.6, 3.1, 3.4, 3.0, 3.4, 2.8, 3.1, 3.3, 3.6

Test if the mean values in the two classes differ significantly, at the significance level $\rho = 0.05$

- We have

$$\omega_1: \ \bar{x} = 3.73, \ \hat{\sigma}_1^2 = 0.0601$$
$$\omega_2: \ \bar{y} = 3.25, \ \hat{\sigma}_2^2 = 0.0672$$

For $N = 10$

$$S_z^2 = \frac{1}{2}(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)$$

$$q = \frac{(\bar{x} - \bar{y}) - 0}{S_z\sqrt{\frac{2}{10}}}$$

$$\boxed{q = 4.25}$$

- From the table of the t-distribution with 2N-2=18 degrees of freedom and $\rho = 0.05$, we obtain $D = [-2.10, 2.10]$ and since q=4.25 is outside D, H_1 is accepted and the feature is selected.

❖ Class Separability Measures

The emphasis so far was on individually considered features. However, such an approach cannot take into account **existing correlations among the features**. That is, two features may be rich in information, but if they are highly correlated we need not consider both of them. To this end, in order to search for possible correlations, we consider features jointly as elements of vectors. To this end:

- Discard poor in information features, by means of a statistical test.

- Choose the maximum number, ℓ, of features to be used. This is dictated by the specific problem (e.g., the number, N, of available training patterns and the type of the classifier to be adopted).

- Combine remaining features to search for the "best" combination. To this end:

  - Use different feature combinations to form the feature vector. Train the classifier, and choose the combination resulting in the best classifier performance.

    A major disadvantage of this approach is the high complexity. Also, local minima, may give misleading results.

  - Adopt a class separability measure and choose the best feature combination against this cost.

- Class separability measures: Let $\underline{x}$ be the current feature combination vector.
  - Divergence. To see the rationale behind this cost, consider the two-class case. Obviously, if on the average the value of $\ln \frac{p(\underline{x}|\omega_1)}{p(\underline{x}|\omega_2)}$ is close to zero, then $\underline{x}$ should be a poor feature combination. Define:

$$D_{12} = \int_{-\infty}^{+\infty} p(\underline{x}|\omega_1) \ln \frac{p(\underline{x}|\omega_1)}{p(\underline{x}|\omega_2)} d\underline{x}$$

$$D_{21} = \int_{-\infty}^{+\infty} p(\underline{x}|\omega_2) \ln \frac{p(\underline{x}|\omega_2)}{p(\underline{x}|\omega_1)} d\underline{x}$$

$$d_{12} = D_{12} + D_{21}$$

d_{12} is known as the **divergence** and can be used as a class separability measure.

– For the multi-class case, define d_{ij} for every pair of classes ω_i , ω_j ; and the average divergence is defined as

$$d = \sum_{i=1}^{M} \sum_{j=1}^{M} P(\omega_i) P(\omega_j) d_{ij}$$

– Some properties:

$$d_{ij} \geq 0$$

$$d_{ij} = 0, \text{if } i = j$$

$$d_{ij} = d_{ji}$$

– Large values of d are indicative of good feature combination.

- **Scatter Matrices.** These are used as a measure of the way data are scattered in the respective feature space.
  - Within-class scatter matrix

$$S_w = \sum_{i=1}^{M} P_i S_i$$

where

$$S_i = E\left[\left(\underline{x} - \underline{\mu}_i\right)\left(\underline{x} - \underline{\mu}_i\right)^T\right]$$

and

$$P_i \equiv P(\omega_i) \approx \frac{n_i}{N}$$

n_i the number of training samples in ω_i.

Trace $\{S_w\}$ is a measure of the average variance of the features.

- Between-class scatter matrix

$$S_b = \sum_{i=1}^{M} P_i \left(\underline{\mu}_i - \underline{\mu}_0\right)\left(\underline{\mu}_i - \underline{\mu}_0\right)^T$$

$$\underline{\mu}_0 = \sum_{i=1}^{M} P_i \underline{\mu}_i$$

Trace $\{S_b\}$ is a measure of the average distance of the mean of each class from the respective global one.

- Mixture scatter matrix

$$S_m = E\left[\left(\underline{x} - \underline{\mu}_0\right)\left(\underline{x} - \underline{\mu}_0\right)^{\mathrm{T}}\right]$$

It turns out that:

$$S_m = S_w + S_b$$

- Measures based on Scatter Matrices.

  - $$J_1 = \frac{\text{trace}\{S_m\}}{\text{trace}\{S_w\}}$$
  - $$J_2 = \frac{|S_m|}{|S_w|} = |S_w^{-1} S_m|$$
  - $$J_3 = \text{trace}\{S_w^{-1} S_m\}$$
  - Other criteria are also possible, by using various combinations of S_m, S_b, S_w.

The above J_1, J_2, J_3 criteria take high values for the cases where:

- Data are clustered together within each class.
- The means of the various classes are far.

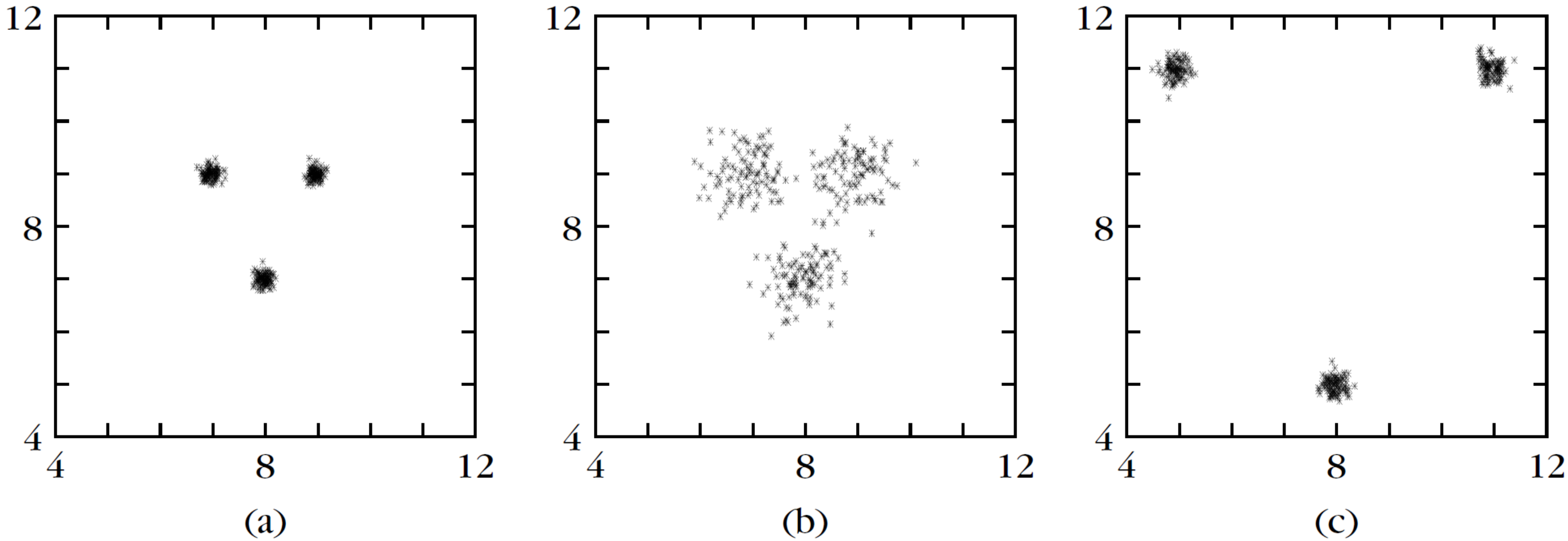


FIGURE 5.5

Classes with (a) small within-class variance and small between-class distances, (b) large within-class variance and small between-class distances, and (c) small within-class variance and large between-class distances.

- Fisher's discriminant ratio. In one dimension and for two equiprobable classes the determinants become:

$$|S_w| \propto \sigma_1^2 + \sigma_2^2$$

$$|S_b| \propto (\mu_1 - \mu_2)^2$$

and

$$\frac{|S_b|}{|S_w|} = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$

known as Fischer's ratio.

❖ Ways to combine features:

Trying to form all possible combinations of ℓ features from an original set of m selected features is a computationally hard task. Thus, a number of suboptimal searching techniques have been derived.

➢ Sequential forward selection. Let x_1, x_2, x_3, x_4 the available features (m=4). The procedure consists of the following steps:

- **Adopt a class separability criterion** (could also be the error rate of the respective classifier). Compute its value for ALL features considered jointly $[x_1, x_2, x_3, x_4]^T$.

- **Eliminate one feature and for each of the possible resulting combinations**, that is $[x_1, x_2, x_3]^T$, $[x_1, x_2, x_4]^T$, $[x_1, x_3, x_4]^T$, $[x_2, x_3, x_4]^T$, compute the class reparability criterion value C. Select the best combination, say $[x_1, x_2, x_3]^T$.

- From the above selected feature vector eliminate one feature and for each of the resulting combinations, $[x_1, x_2]^T, [x_2, x_3]^T, [x_1, x_3]^T$ compute C and select the best combination.

The above selection procedure shows how one can start from m features and end up with the "best" ℓ ones. Obviously, the choice is suboptimal. The number of required calculations is:

$$1+\frac{1}{2}\big((m+1)m-\ell(\ell+1)\big)$$

In contrast, a full search requires:

$$\binom{m}{\ell}=\frac{m!}{\ell!(m-\ell)!}$$

operations.

- Sequential backward selection.

  Here the reverse procedure is followed.

  - **Compute C for each feature**. Select the "best" one, say x_1
  - **For all possible 2D combinations** of x_1, i.e., $[x_1, x_2]$, $[x_1, x_3]$, $[x_1, x_4]$ compute C and choose the best, say $[x_1, x_3]$.
  - **For all possible 3D combinations** of $[x_1, x_3]$, e.g., $[x_1, x_3, x_2]$, etc., compute C and choose the best one.

The above procedure is repeated till the "best" vector with ℓ features has been formed. This is also a suboptimal technique, requiring:

$$\ell m - \frac{\ell(\ell-1)}{2}$$

operations.

- Floating Search Methods

The above two procedures suffer from the nesting effect. Once a bad choice has been done, there is no way to reconsider it in the following steps.

In the floating search methods one is given the opportunity in reconsidering a previously discarded feature or to discard a feature that was previously chosen.

The method is still suboptimal, however it leads to improved performance, at the expense of complexity.