Feature selection

Уменьшение размерности

Feature selection (variable selection, attribute selection, variable subset selection)

Feature extraction

Цели Feature selection

1. Избежание переобучения и повышение качества классификации (в т.ч. для улучшения кластеризации).
2. Ускорение работы классифицирующих моделей.
3. Дополнительное понимание об изучаемых объектах.

Виды отбираемых аттрибутов

- Избыточные (Redundant) аттрибуты - не несут никакой дополнительной информации
- Иррелевантные (Irrelevant) аттрибуты - не несут вообще какой-либо информации

Оценка методов feature selection

- На различных датасетах.
- С различными классификаторами (если возможно).
- Добавляют в исходные датасеты векторы шумов и таргет вектор.

Виды Fearure Selection

1. Filter methods
    a. Univariate
    b. Multivariate
2. Wrapper methods
    a. Deterministic
    b. Randomized
3. Embedded methods

Filter methods

Оценивают качество тех или иных аттрибутов и, как правило, отсеивают худшие из них.

+ Вычислительно простые, легко масштабируются
- Игнорируют связи между аттрибутами, особенности используемого классификатора

Примеры Filter methods

Univariate:

- Euclidian distance
- Information gain
- Spearman corellation coefficient

Multivariate:

- CFS
- MBF

Spearman corellation coefficient

$$\rho = \frac{\sum_{ij}(x_{ij}-\bar{x}_j)(y_i-\bar{y})}{\sqrt{\sum_{ij}(x_{ij}-\bar{x}_j)^2 \sum_i (y_i-\bar{y})^2}}$$

$$\rho \in [-1;1] \qquad \rho \to 0$$

Wrapper methods

Получают некоторым способом подмножество аттрибутов из исходного.

+ Имеют более высокую точность чем Filtering, могут учитывать связи между аттрибутами, напрямую взаимодействуют с используемым классификатором.
- Долгое время работы, высокая вероятность переобучения.

Примеры Wrapper methods

Deterministic:

- SFS
- SBE
- SVM-RFE

Randomized:

- Randomized Hill Climbing
- Genetic Algorithms

SVM-RFE

1. Обучаем SVM на тренировочной выборке
2. Ранжируем признаки по полученным весам
3. Выкидываем последние признаки
4. Повторяем, пока не останется заданное количество признаков

Embedded

Учитывают особенности классификатора, в отличие от wrapper’ов, для которых классификатор - черная коробка.

Для каждого классификатора приходится использовать индивидуальный метод.

Random Forest

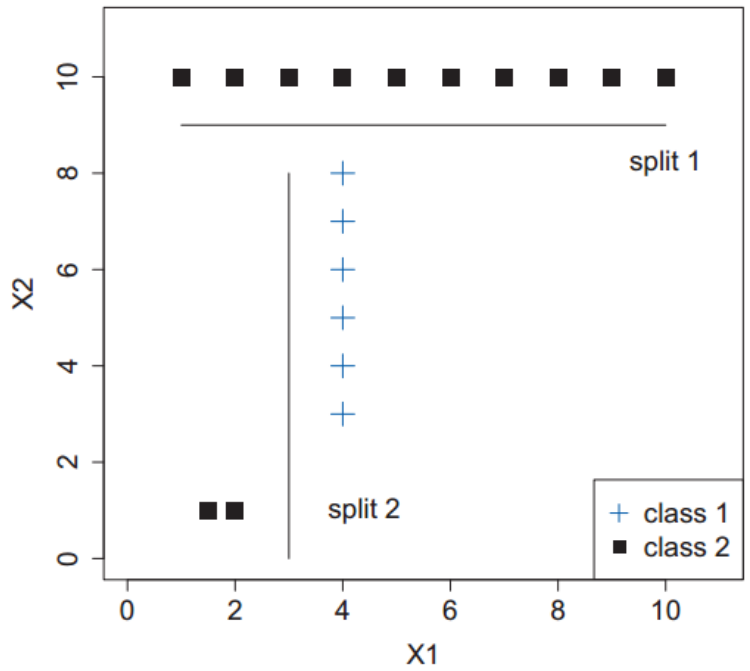
1. Для каждого дерева выбирается подвыборка размера N с повторениями
2. Строится решающее дерево. При выборе очередного признака для разбиения рассматриваются $m \approx \sqrt{M}$ признаков.
3. Выбирается наилучший по заданному критерию.

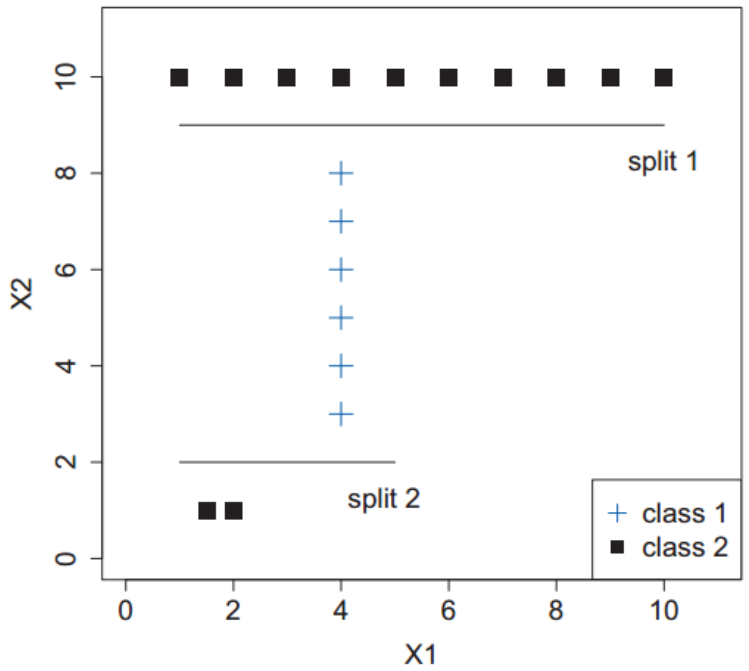
IG and IG

$$gini(T) = 1 - \sum_{i=1}^{k} \big(p(c_i)\big)^2 - \sum_{i=1}^{n} p(t_i) \sum_{j=1}^{k} p(c_j|t_i)\big(1 - p(c_j|t_i)\big).$$

$$gain(T) = -\sum_{i=1}^{k} p(c_i) \log_2\big(p(c_i)\big) + \sum_{i=1}^{n} p(t_i) \sum_{j=1}^{k} p(c_j|t_i) \log_2\big(p(c_j|t_i)\big).$$

Избыточность

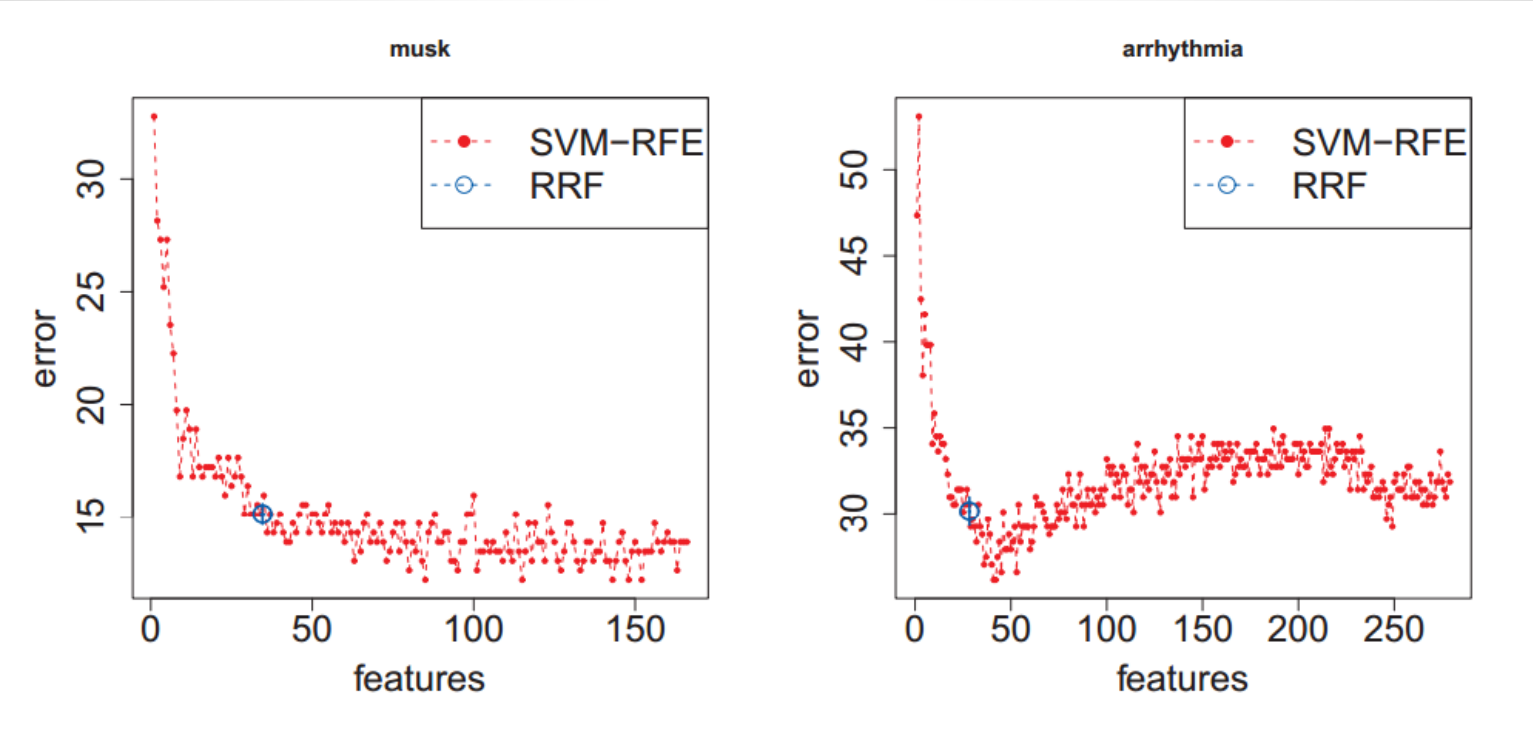




Regularization

$$gain_R(X_j) = \begin{cases} \lambda \cdot gain(X_j) & X_j \notin F \\ gain(X_i) & X_j \in F \end{cases}$$

SVM-RFE vs RRF



Но RRF работает на порядок быстрее.

Feature importance

$$Imp_i = \frac{1}{ntree} \sum_{v \in S_{X_i}} Gain(X_i, v)$$

$$Imp_i' = \frac{Imp_i}{max_{j=1}^{P} Imp_j}$$

GRRF

$$Gain_R(X_i, v) = \begin{cases} \lambda_i Gain(X_i, v) & X_i \notin F \\ Gain(X_i, v) & X_i \in F \end{cases}$$

$$\lambda_i = (1 - \gamma)\lambda_0 + \gamma Imp_i'$$

Заключение