Big Data

VICTORIA UNIVERSITY OF WELLINGTON
TE HERENGA WAKA

AIML427

Filter Feature Selection

Dr Bach Hoai Nguyen

Bach.Nguyen@vuw.ac.nz

Outline: This Week

- Single feature ranking
- Filter feature selection methods
- Embedded feature selection methods
- Feature selection applications

Single Feature Ranking

An easy (*naïve?*) way to do feature selection

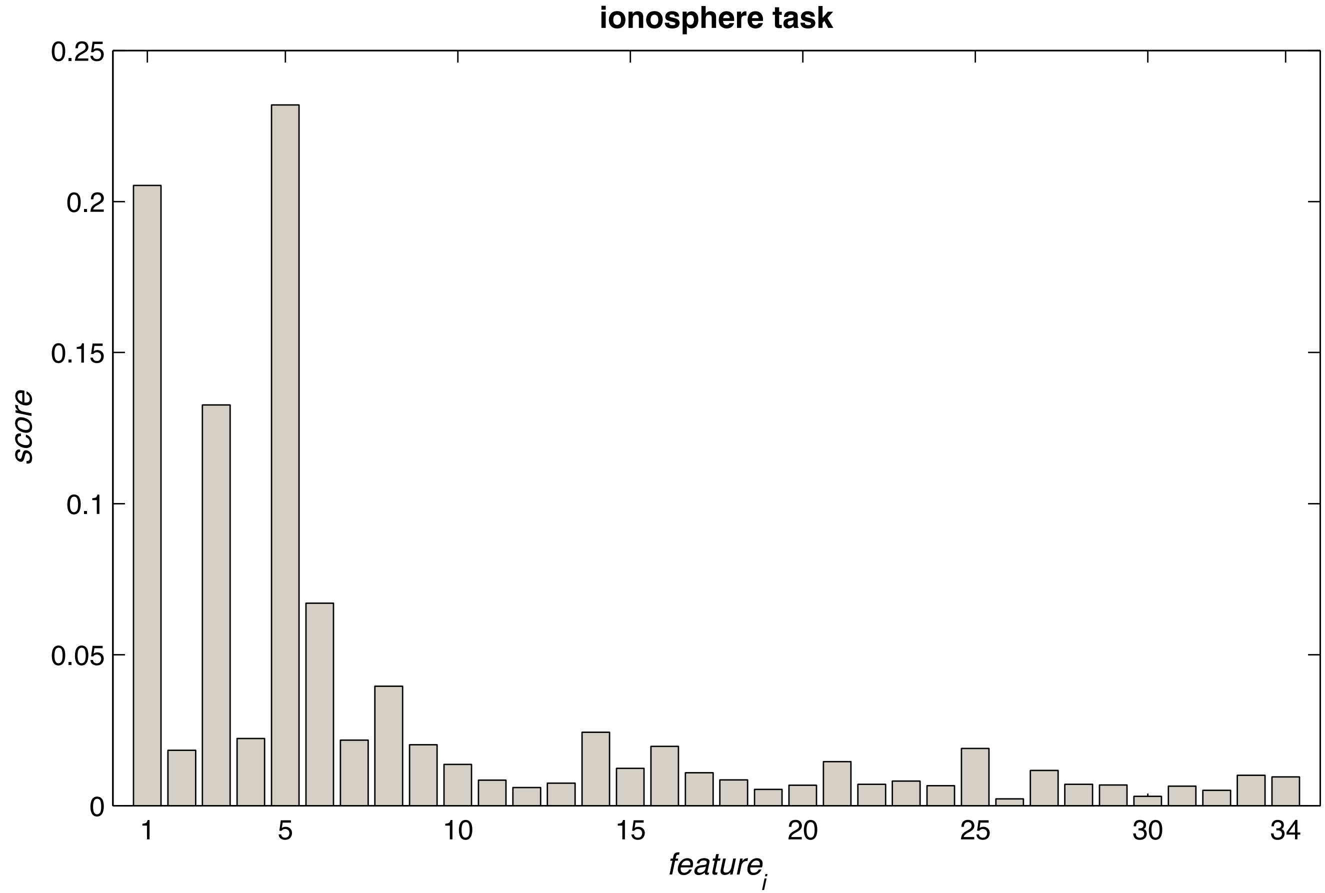
To select *m* features out of *n* original features:

1. Use an algorithm to measure the importance (goodness) of each feature individually
2. Sort (rank) all m features in the descending order of their importance
3. Choose m top (most important) features
4. The importance of a feature is determined depending on their "contribution" to the task, e.g. classification

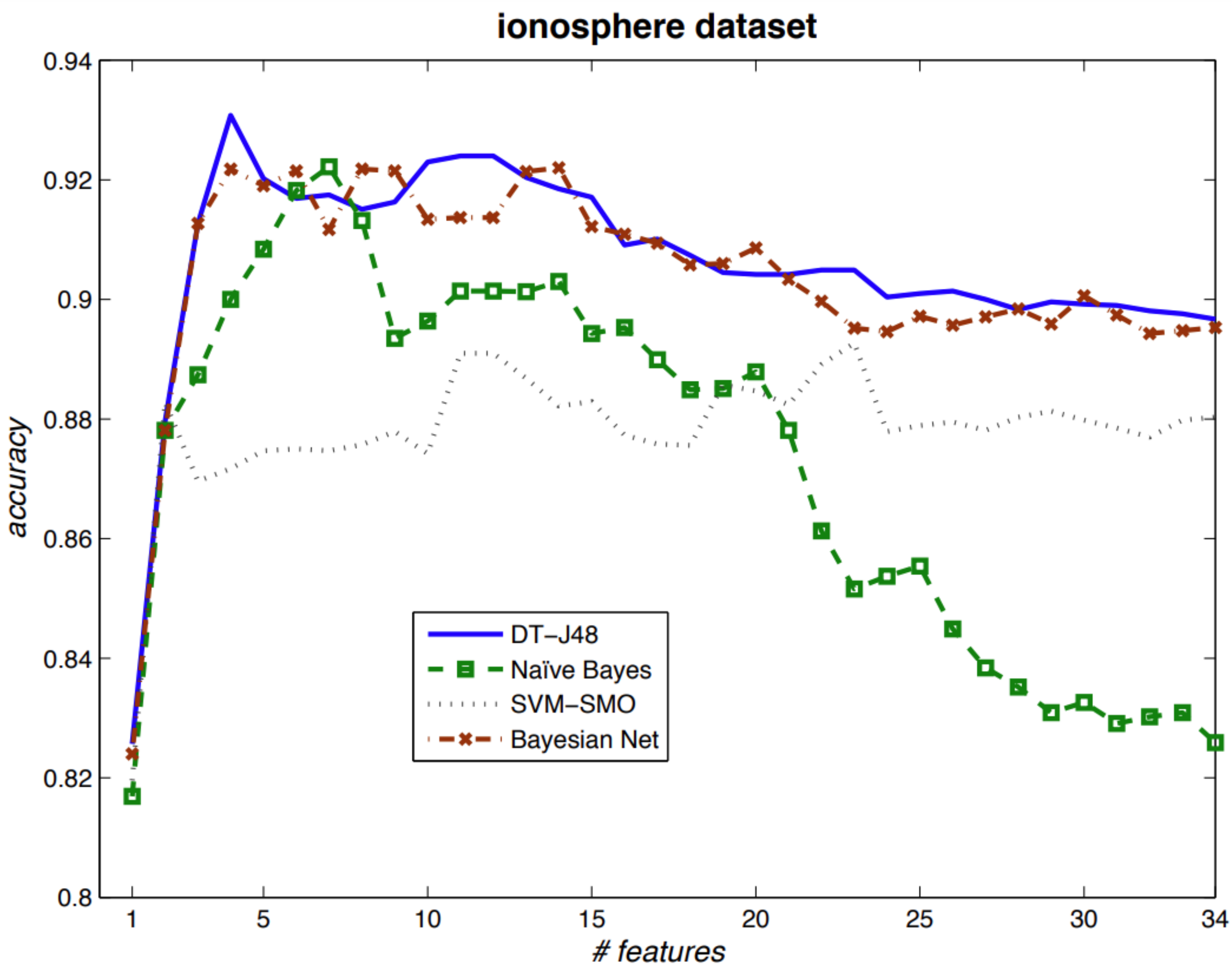
- Common measures of relevance/importance:
  - Pearson's correlation
  - Statistical testing (e.g. χ^2 test)
  - Information theory (e.g. Mutual Information, Information Gain)
  - Logistic Regression

Example: Single-Feature Ranking

- Decision Trees/Genetic Programming
- The frequency of features in good performing trees can be used to measure the importance of individual features.



Example: Single-Feature Ranking



Issues: Single-Feature Ranking for Selection

There are potential risks in using single-feature ranking methods for feature selection:

- Ignore *interactions between features*
- These methods cannot recognise the true worth of a group of features that seem to be individually weakly relevant
- High-ranked (top important) features might be redundant

Feature ranking

vs

Feature subset selection

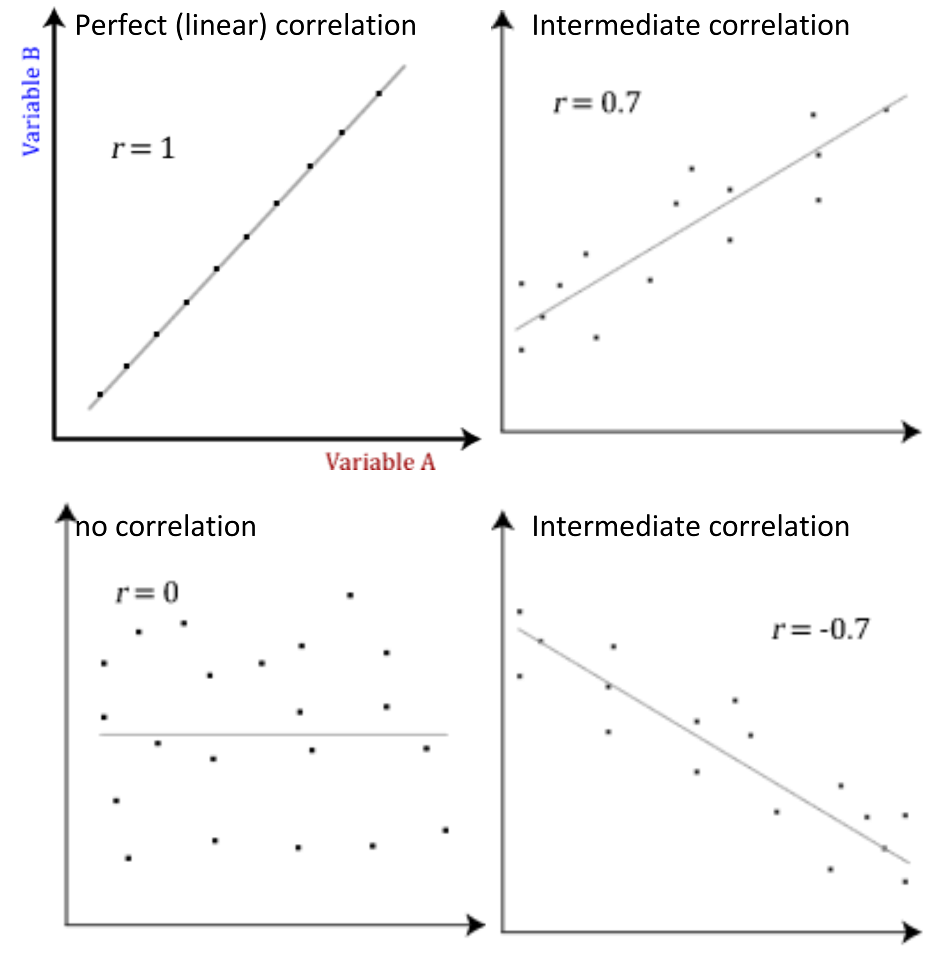
FILTER FEATURE SELECTION

Filter Approach

- Filter FS: does not involve any learning algorithm during the feature selection process
- Covers many feature selection algorithms:
  - Those that use a search strategy and a surrogate classifier
  - Those that use single-feature ranking for feature selection
  - Many other algorithms (e.g. reliefF, ...)

Pearson's correlation

- The Pearson correlation coefficient, r:
  - r in [-1, 1]
  - r = 0 indicates **no association** between the two variables
  - r > 0 indicates a positive association
  - r < 0 indicates a negative association



- r is calculated according to:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\,\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

Pearson's correlation

- Can measure the relevance between a feature & class label
- Binary classification: can use Pearson correlation directly
- Multi-class classification (>2 class values):
  - {Red, Green, Blue} – nominal -> no obvious distance
  - *k* classes, convert to *k* binary variables (one-hot encode)

Y	$\mathbf{Y_1}$	$\mathbf{Y_2}$	$\mathbf{Y_3}$
Red	1	0	0
Green	0	1	0
Blue	0	0	1

  - Calculate correlation based on these *k* binary variables $\mathbf{Y_1}$, $\mathbf{Y_2}$, $\mathbf{Y_3}$ with each feature.

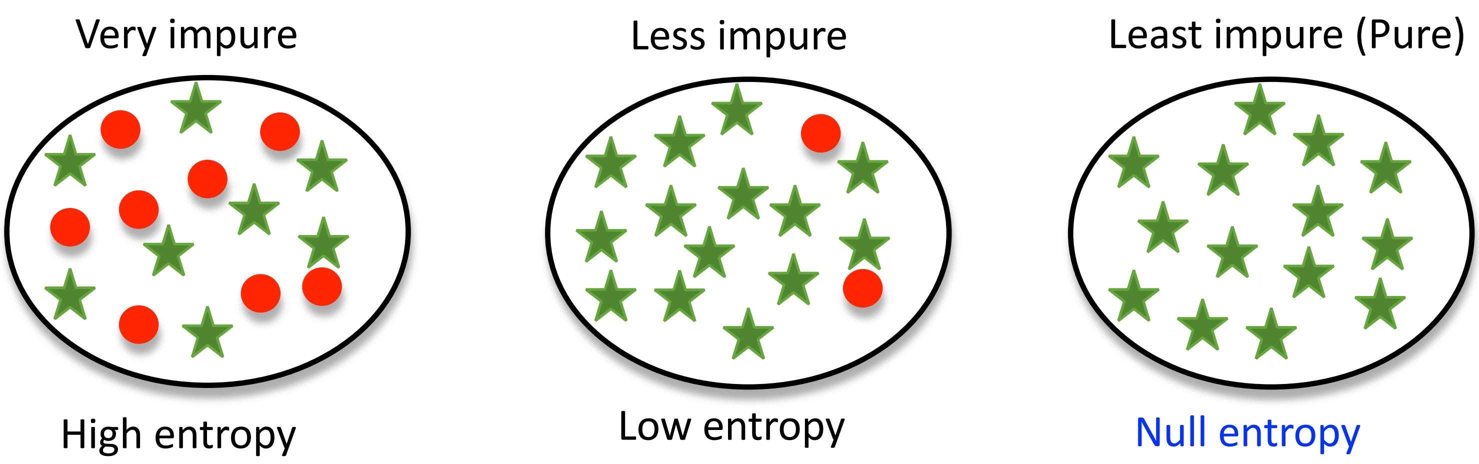
Information Theory: Entropy

- Entropy measures the impurity or uncertainty in a group of examples.

- S is the (training) set, with $C_1, \ldots, C_N$ classes

$$H(S) = -\sum_{c=1}^{N} p_c * log_2(p_c)$$

- **H(S)** measures the Entropy of S
- $\boldsymbol{p_c}$ is the **proportion** of class $\boldsymbol{C_c}$ in S



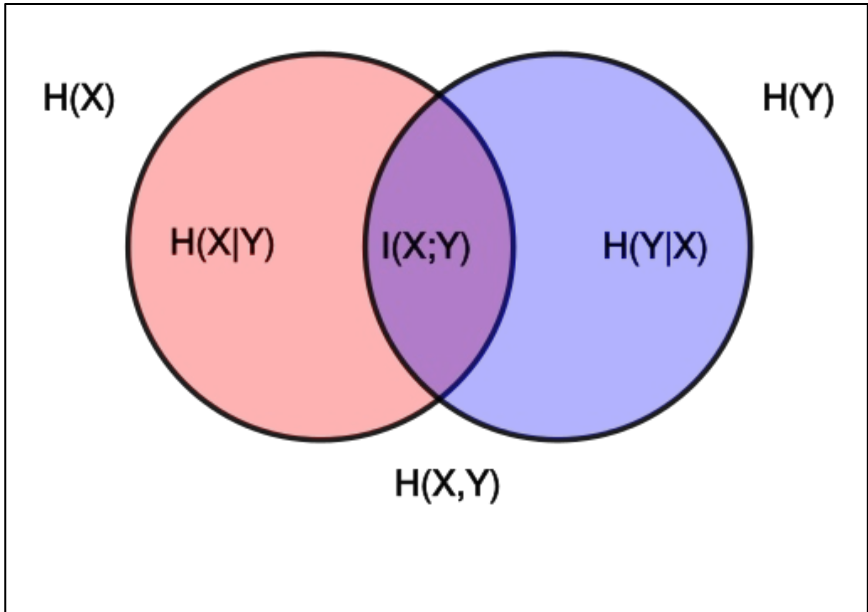
Conditional Entropy

- Entropy
  - $\mathrm{H(X)} = -\sum_{x\epsilon X} p(x) log_2 \, p(x)$
  - $p(x) = P(X = x)$ is the probability density function of X

- Conditional entropy:

$$\mathrm{H(X|Y)} = -\sum_{x\epsilon X, y\epsilon Y} p(x,y) \, log_2 \, p(x|y)$$

  - Entropy of X given Y
  - How much information needed to describe X given Y
  - $H(C|X_1) < H(C|X_2)$: which one is better, X_1 or X_2?



Mutual Information

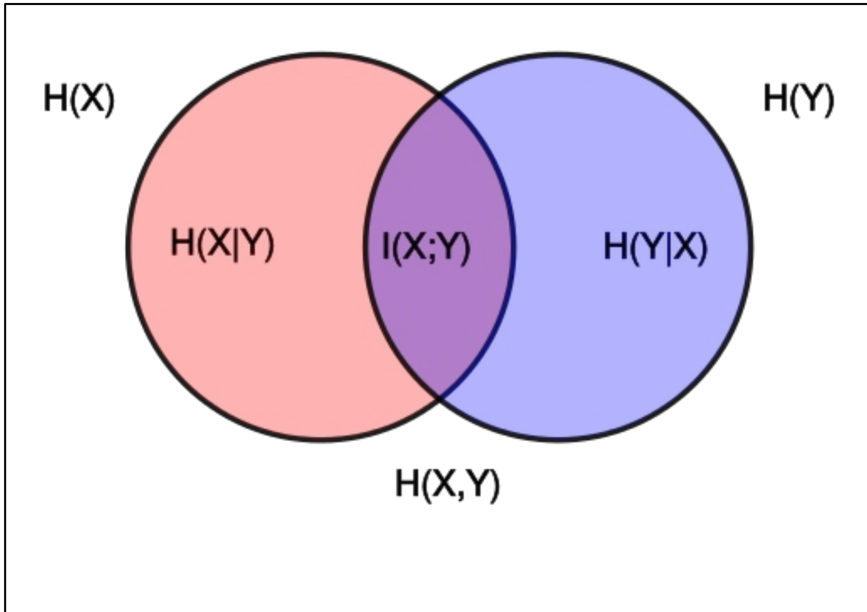
Mutual information of two random variables is a measure of the mutual dependence between the two variables

- How much information does one variable give about another variable?

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$
$$= \sum_{x \in X, y \in Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$

- $I(X_1; C) > I(X_2; C)$:
  which one is better, X_1 or X_2?

- $I(X_1; X_2) = 0.8$,
  $I(X_2; X_3) = 0.4$,
  $I(X_1; X_3) = 0.5$:
  remove which feature?



Mutual Information

- Mutual information evaluates the information shared between each pair of features/variables

- Relevance:
  - Classification performance
  - The relevance (MI) between each selected feature and the class labels

- Redundancy:
  - Number of features
  - The redundancy (MI) between the selected features

Ranking using Information Theory Measures

- Categorical (nominal) data:
  - If it is a numeric feature it must first be *discretised*
- Mutual information estimation method can used
- Mutual information between a feature and the class labels
  - Rank features
  - Select top ranked features

Filter Method

Objective Function:

$$Rel = \sum_{x_i \in X} I(x_i; C)$$

$$Red = \sum_{\substack{x_i, x_j \in X, \\ and\ i \neq j}} I(x_i; x_j)$$

- X is the selected feature subset
- x_i, x_j : feature in X
- C is the class lables
- *Rel*: relevance between X and c
- *Red*: redundancy within X

$$\mathrm{I(X;Y)} = \mathrm{H(X)} - \mathrm{H(X|Y)} = \mathrm{H(Y)} - \mathrm{H(Y|X)}$$
$$= \sum_{x \epsilon X, y \epsilon Y} p(x,y) log_2 \frac{p(x,y)}{p(x)p(y)}$$

Minimum Redundancy-Maximum Relevance (mRMR)

- S is the feature subset, Ω is the pool of all candidate features, the minimum redundancy condition is:

$$\min_{s \subset \Omega} \frac{1}{|S|^2} \sum_{i,j \in S} I(f_i, f_j)$$

where |S| is the number of features in S.

- For classes $c=(c_i,....c_k)$ the maximum relevance condition maximises the total relevance of all features in *S:*

$$\max_{s \subset \Omega} \frac{1}{|S|} \sum_{i \in S} I(c, f_i)$$

H.C. Peng, F.H. Long, and C. Ding, Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, 2005, pp. 1226–1238.

Minimum Redundancy-Maximum Relevance (mRMR)

- The mRMR feature set optimises these two conditions simultaneously, either in quotient form:

$$\max_{s\subset\Omega}\{\frac{\sum_i I(c,f_i)}{\frac{1}{|S|}\sum_{i,j\in S} I(f_i,f_j)}\}$$

or in difference form:

$$\max_{s\subset\Omega}\{\sum_i I(c,f_i)\ -\frac{1}{|S|}\sum_{i,j\in S} I(f_i,f_j)\}$$

H.C. Peng, F.H. Long, and C. Ding, Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, 2005, pp. 1226–1238.

Filter Feature Selection

- Information theory-based approach:
  - max-relevance, and min-redundancy
- Rough set theory for feature selection
- Fast correlation based filter feature selection
- Evolutionary computation for filter feature selection
- …
- Issues:
  - Most filter approaches do not evaluate subsets of features