Subject

How to compare supervised learning algorithm with cross validation?

We want to compare two learning methods:

- K-NN (Nearest Neighbor), we use the HVDM distance metric (see references on the web site) which allows to mix discrete and continuous attributes;
- Decision tree algorithm (ID3).

Dataset

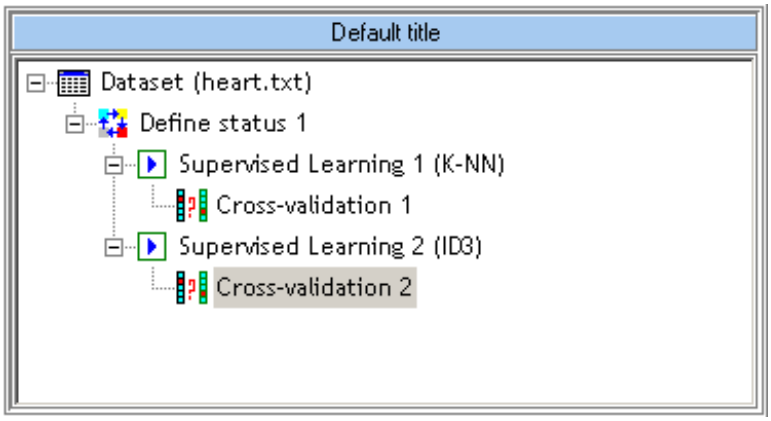
UCI's HEART DISEASES DIAGNOSTIC (Cleveland), THAL is omitted.

Experimentation steps

1. Load "dr_heart.bdm"
2. Insert « Define Status » and set attributes as:

Class attribute – TARGET	Cœur
Descriptors -- INPUT	Age, Sexe, Type_Douleur, Pression, Cholesterol, Sucre, Electro, Taux_Max, Angine, Depression, Pic, Vaisseau

3. Insert two supervised methods K-NN and ID3. For each of them, you must add in the first time a "Supervised learning" (from Meta-Spv Learning palette) in which you insert the learning algorithm (from Spv-Learning palette – K-NN and ID3). Set the following parameters:
   - K-NN: Number of neighbor = 5
   - ID3: Min Size For Split = 20, Min Size of Leaves = 5, Max depth of tree = 10, Min Entropy gain for splitting = 0.03
4. Insert in the diagram, after each learning method, a "cross-validation" component (from Spv Learning Assessment). Do not modify default parameters. You have the following diagram.



5. Then, you can execute each learning method on the whole dataset, resubstitution error rates for each method are very similar: 0.144 for 5-NN, and 0.152 for ID3. Are these methods having the same performances on this dataset?
6. To verify this, we use a cross-validation, we see that the "true" (less biased) error rate shows a better performance of 5-NN: **#0.19 for 5-NN** and **#0.26 for ID3**.
7. Conclusion: use always an unbiased error rate estimation to evaluate a learning algorithm (test set, cross-validation, bootstrap, etc.)