TDT4300 Datavarehus og datagruvedrift
Assignment 3: Classification

Magnus L Kirø

March 11, 2013

1 Decision Trees

- 1: tot gini= 0.930801388889
- 2: id : gini= 0.95
- 3: age : gini= 0.665
- 4: student : gini= 0.5
- 5: creditworthiness : gini= 0.5
- 6: The best gini value is the total for the whole set. Besides that age is the best gini value. The id(UserID) is a unique value and would in many cases be useless when searchinf for information.

The two last gini indexes are "income : gini= 0.645" and "pc : gini= 0.48"

2.2 datasets

- Iris dataset

  The dataset contains three (3) variables.

  150 instances are available.

  The variables describe, petal length, petal width and class.

  I assume the petal width is the most difficult to classify.

- Diabetes dataset

  Has nine attributes.

  768 instances are present in the dataset.

  The data describes the condition and stats of diabetes patients.

  I would presume that the *insu* variable is the hardes to classify.

- Spambase dataset

  58 variables can be found in the dataset.

  4601 instances ar present in the dataset.

  The content of the dataset describes content of spam email.

  Class as a variable here would be the most difficult one to classiy.

2.3 Classification

List the algorithems, with corresponding findings.

- J48

  Iris; The petal width is the most important variable. The accuracy is 96.08%

  Diabetes; *Plas* is the most important variable here. The accuracy is 76.25%

  spambase; *word_freq_remove* is here the most important variable. Accuracy=92.20%

- k-NN

  Iris; Using the neares neighbour the petal width variable is the most significant one. accuracy=96.08%

  Diabetes: the plas variable would still be the most significant one. accuracy=72.80%

  spambase; The most deciding variable is *word_freq_remove* and the accuracy is 89.0%

- Support vector machines

  Iris; petal width has the most weight here. accuracy=96.08%

Diabetes; Plas has the most significant value here. accuracy=79.31%

spambase; *word_freq_remove* is still the most significant variable accuracy=90.54%

2.4 Evaluation

Cross-validation is the process of splitting the dataset into subsets, execute calculations, then comparing all the calculation of the subsets to create a better total estimate of the calculation.

With the k-NN algorithm the cross-validation gives the biggest difference in accuracy.

Cross-validation will be a better choice with bigger datasets and more complex datasets. Splitting the calculations into sub calculations can save time and complexity, thus reducing resource and time usage. And the over all result becomes a bit more accurat.

In the case where the data is sorted the percentage split won't help much. Then the subsets would be to different to give good results. Having four fruits, two apples and two oranges, there would be little point in dividing them into two sets where the same fruits are in the same sets. Splitting like this, calculating and then combine would only result in bigger resource usage and time delay.

2.5 Best Classifiers

J48 showed the best performance with all datasets. The REPTree algorithm showed better results then any of the others.