Big Data Analytics

Workshop 2. Weka Decision Trees II

1. **Build the Decision Tree Model with “training set”, and test the model with “cross-validation”**

Open and Save the file churn2.csv

Build a decision tree classifier on this dataset using J48 with default settings

Classify tab

* classifiers (click Choose) => tree => J48
* click Object Editor area (the area on the right-hand-side of the choose button), set the following parameters
  + set unpruned to true
  + force moderately large leaves with minNumObj=5000
* Test options: select “cross-validation”
* click “start” to run the result

1. **Compare results with various sizes of minNumObj**

**Questions:**

* 1. Start from 5000 instances per node, and record the Results.

Number of nodes (exclude leaves and root node) = Size of tree - Number of leaves – 1

**% Incorrectly Classified Instances (Error Rate)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **minNumObj** | **Error Rate (%)**  **Using Training set** | **Error Rate (%) using Cross-validation** | **Size of tree** | **Number of leaves** | **Number of nodes** |
| **5000** |  |  |  |  |  |
| **1000** |  |  |  |  |  |
| **500** |  |  |  |  |  |
| **200** |  |  |  |  |  |
| **100** |  |  |  |  |  |
| **50** |  |  |  |  |  |
| **30** |  |  |  |  |  |
| **20** |  |  |  |  |  |
| **15** |  |  |  |  |  |

* 1. Please use the information from the table to plot the graph

Error Rate (%)

Number of Nodes

* 1. What do you observe from the above results?