CS760 – ML – Warm up Project

Exploring Weka and ARFF file format

Anmol Mohanty

Data Set:-

A dataset of CPU manufactures and their characteristics has been created. The ultimate classification is whether the CPU sub-system hardware is capable of handling graphics intensive games.

There are 20 examples with 10 attributes in each. The attributes in order are:-

- Vendor Name
- MYCT (Machine Cycle time, Nanoseconds)
- MMIN (Min frequency, Khz)
- MMAX (Max frequency, Khz)
- CACH (Size of Cache, MB)
- CHMIN (Mem channel min frequency, Mhz)
- CHMAX (Mem channel min frequency, Mhz)
- GRAPHICS(Presence of graphics card)
- MEMORY(Physical RAM, MB)
- games_class with nominal data type ({0,1}) representing if games can be played on this system[1] or not [0].

A sample example:-

Intel	100	8000	32000	64	8	32	1	2048	1

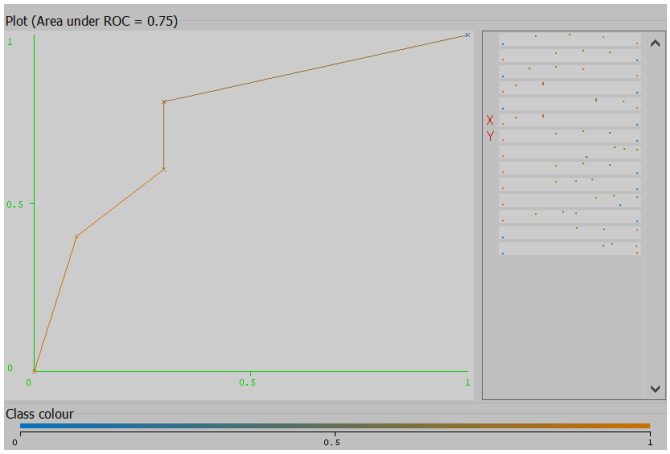
This shows that Intel is the vendor of this chip, with a cache size of 64 MB and top CPU frequency of 32 Mhz. It has a graphics card and is capable enough of playing games.

The data file corresponding to this is attached as “cpu.with.vendor.final.ARFF”

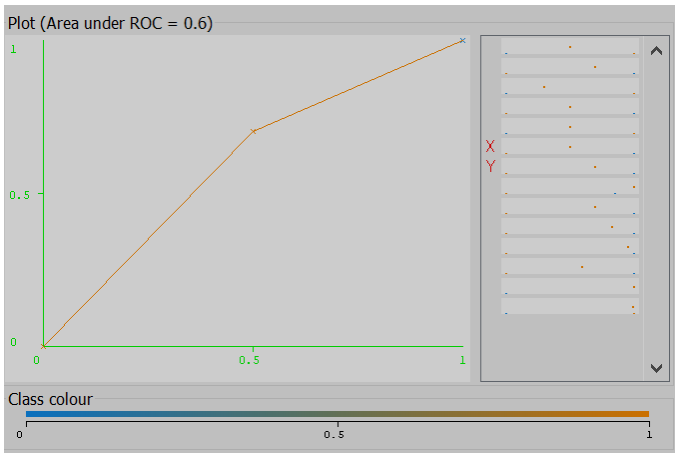
10 fold cross validation(default) was used.

These are the results and ROC curves of the 2 classifier methods that were requested in a comparative tabular format. The ROC is plotted for label value = 1 i.e the CPU can play games.

J48



1NN



Accuracy	75%	60%
Time	0.01s	0s
ROC area	0.75	0.6

Clearly, the 1NN is faster than the J48 classifier which is to be expected as input sample size is very small and 1NN can rapidly (almost instantaneously classify).

This comes of with a tradeoff in accuracy though as the J48 classifier is more accurate (75%) as opposed to 1NN(60%). The Receiver Operating Characteristics curve is similarly proportioned.