Phase -1 Project Report

Ripu Singla (Y6387)

Patha Rama Krishna(Y6318)

Most of the methods like J48, logistic Regression, SMO and Decision Stump give the default classifier which classify class B+E with approx 92 % accuracy because there are 8072 instance out of 8695 are labeled as B+E . This classification is not good because it gives 100% false alarms which is not good for the relations between different countries.

First, We have separated the data for each station respectively in different files (vdata, xdata , ydata, wdata, zdata) because different stations are located at different places so there is no reason to consider the whole data together as one. Then we have applied different algorithms on each of these data files.

Results are shown in the following table

- Decision Stumps, combined with Adaboost

Location	Hits (%)	Correct Rejection(%)	Miss (%)	False Alarm (%)	AUC	Accuracy (%)
V	100	0	0	100	0.801	93.25
W	100	0	0	100	0.881	98.19
X	100	0	0	100	0.675	90.86
Y	100	0	0	100	0.778	96.69
Z	100	0	0	100	0.575	90.98

- Logistic Regression, combined with Adaboost

Location	Hits (%)	Correct Rejection(%)	Miss (%)	False Alarm (%)	AUC	Accuracy (%)
V	100	0	0	100	.803	93.19
W	100	0	0	100	0.875	98.19
X	100	0	0	100	0.646	90.86
Y	100	0	0	100	0.804	96.69
Z	100	0	0	100	0.551	90.97

- Random Tree, combined with Adaboost

Location	Hits (%)	Correct Rejection(%)	Miss (%)	False Alarm (%)	AUC	Accuracy (%)
V	99	20	0.44	80	o.938	94.19
W	100	6	0	94	0.962	98.32
X	99.6	39.5	0.3	60.47	0.969	94.17
Y	99.57	25	0.43	75	0.959	97.1
Z	99	97.5	1	2.5	0.999	98.89

- Nearest Neighbour (IB1), uncombined

Location	Hits (%)	Correct Rejection(%)	Miss (%)	False Alarm (%)	AUC	Accuracy (%)
V	92.57	49	7.42	50.43	.711	89.67
W	98.8	21.4	1	78.57	0.602	97.55
X	95.4	62.3	4.6	37.7	0.789	92.39
Y	97.7	30	2.3	70	0.639	95.45
Z	99.47	92.6	0.53	7.4	0.96	98.85

- As we can see from above tables, for z station the IB1 and Random Tree, combined with Adaboost give very good results.

As the given data set is highly imbalanced (90% of B+E category) so we tried under-sampling technique. For example Station V's data contains 1704 instances including 115 B category and 1589 B+E category. So we divided B+E labeled instances into 5 equal parts and sampled the each part with 115 B category and used randomized algorithm in weka to shuffle them. Then we made 5 classifiers from these above 5 samples, applied on the whole station data and took the vote. We used matlab for voting process.

After trying all the above algorithms for these samples we got the best results for the Adaboost Algorithm with Random Tree as base classifier.

Results are given in following table

- **Random Tree Combined with Adaboost**

Samples	Hits (%)	Correct Rejection(%)	Miss (%)	False Alarm (%)	AUC	Accuracy (%)
vSample1	83.63	74.78	16.37	25.22	0.88	83.04
vSample2	79.86	86.95	20.14	13.05	0.886	80.34
vSample3	77.22	88.69	22.78	11.31	0.882	77.99
vSample4	84.77	65.21	15.23	34.79	0.869	83.45
vSample5	71.24	71.37	28.76	28.69	0.88	82.98
Average	83.64	74.78	**16.36**	**25.22**	**0.88**	83.04

- As we can see from the above table this final classifier reduces the amount of false alarm from 100% in the default case to 25% with compromising 16% of miss cases.
- The result shown above is for V station but similar procedure can be used for the other stations in the same way.

Reference:

ICDM Data Mining Contest '08 booklet
(http://www.cs.uu.nl/groups/ADA/icdm08cup/booklet.pdf)