# Developer manual

Author: Tao Gong 225787

## Data selection (at least 50 words)

There are 57 features and 1 class label. First, all the features were used, after some necessary preprocessing and transformation, 6 models were used, the result seemed good as you can see in data evaluation part. After that, I observed the dataset used weka but I could not find any feature that can be deleted. So I used the function “select attributes”in WEKA. And I got 15 features use best first algorithm which are attributes 4,5,7,16,21,23,24,25,27,42,44,26,52,53,55. But after evaluation, only the result of IBK model got improved. So I still used all the features.

## Data Preprocessing (at least 50 words)

According to the detailed information of the dataset online, this dataset contains missing values but those value are set to be zero. Because it contains no noise data, it is not necessary to cope with it.

No new features are used because the current features are sufficient to build a model. Maybe words like “iphone” will be considered in the future.

## Data Transformation (at least 50 words)

The dataset contains 54 attributes which range from 0 to 100 and 3 attributes which have a very broad range like from 1 to 15841. So the data was scaled and used function “sd” to all the margins of the column of attributes. By doing this, more accurate data mining results can be achieved.

## Data Mining (at least 100 words)

The six classification techniques used are

1. IBK k-nearest neighbors classifier Aha & Kibler (1991).
2. SMO John C. Platt’s sequential minimal optimization algorithm for training a support vector classifier using polynomial or RBF kernels.
3. J48 C4.5 decision trees (Quinlan, 1993)
4. JRip propositional rule learner
5. PART decision lists using the approach of Frank and Witten (1998)
6. AdaBoost M1 method of Freund and Schapire (1996).

The reason to use IBK is that it acts as a baseline for all other five methods. It is one of the easiest methods to interpret the results.

Also, SMO is based on function and a commonly used method. J48 bases on decision trees. JRip and PART are rule based methods. These four methods stands for the main stream of classification methods.

The last method used is AdaBoost which boost the result of base classifier. It was used to observe the performance improvement.

## Pattern Evaluation (at least 100 words with tables or graphs that represent each error rate or performance)

To evaluate the six model, I used 10-fold cross validation. From table 1, the best model is AdaBoostM1 for base classifier PART whose error rate is 4.39.

**Table 1.** Error Rate of Selected Methods

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | IBK | SMO | J48 | JRip | PART | AdaBoostM1(PART) |
| Error rate for all the features | 9.21 | 9.58 | 7.02 | 7.17 | 6.41 | 4.39 |
| Error rate for best first features | 8.87 | 13.52 | 7.31 | 7.96 | 7.2 | 6.67 |

And the measures of performance for each model can be seen in Table 2. The AdaBoostM1(PART) is best.

**Table 2.** Performance measure of Selected Methods

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | IBK | SMO | J48 | JRip | PART | AdaBoostM1(PART) |
| Precision | 0.908 | 0.905 | 0.930 | 0.928 | 0.936 | 0.956 |
| Recall | 0.908 | 0.904 | 0.930 | 0.928 | 0.936 | 0.956 |
| F-Measure | 0.908 | 0.903 | 0.930 | 0.928 | 0.936 | 0.956 |
| MCC | 0.807 | 0.798 | 0.853 | 0.849 | 0.866 | 0.908 |
| ROC Area | 0.906 | 0.891 | 0.939 | 0.937 | 0.956 | 0.983 |
| PRC Area | 0.879 | 0.862 | 0.917 | 0.927 | 0.938 | 0.983 |