Preprocessing

Siyu Zhou

Data Preprocessing has five steps. Describing data summarization is the first step. The number of numeric type attribute was 16 while nominal type was 2. The value of minimum was 10, maximum was 26, mean was 13.468 and StdDex was 3.316. The second step is data cleaning. The Strikeouts attribute had missing data, so I applied replace-data filter to dataset. Then I got outlier and removed the data which were out of boundary. The third step is data transform. I chose Z-score normalization. The fourth step is reduction.I removed attribute of player because the unique value of this attribute was 0. I also removed the attribute of outlier and extremevalue. Then I used PCA to reduce dimension to 12. The fifth step, I used filter to do discretization. Until now, I got a new dataset. But when I applied J48 to this new dataset, there was no tree. So I discarded discretization.

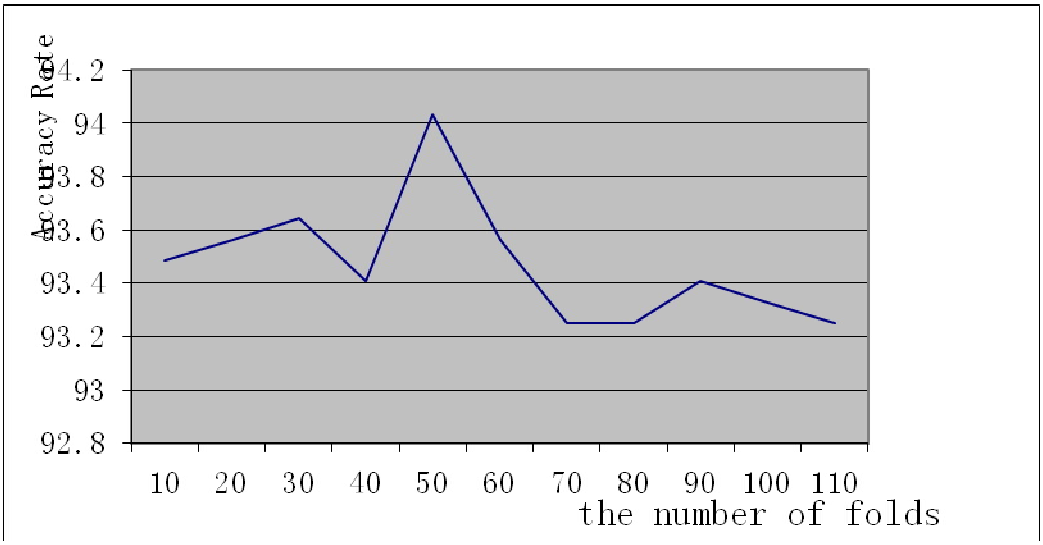
The new dataset had 12 attributes which I thought having close relations. Now, the number of sample was 1274. I thought there was no need to do sampling because 1274 wasn’t too larger and I wanted to keep them to a more accurate analysis.

J48&Decisin Tree

Siyu Zhou

In J48 method, the default value of confidencefactor is 0.25. With the increasing of this value, the tree will be more inaccurate. When confidencefactor is 0.04, I got the balanced point for correctly classified instances, relative absolute error and root relative squared error. Similarly, seed equaling to two also was the balanced point. I tried 11 times for the various number of folds and I found the peak value was around 50 folds, as Graph-1 showing. Then I tried twenty-one consecutive integers from 40 to 60, in order to find the highest number of correctly classified instances. Eventually, I saw 50 folds giving maximum accuracy rate.

In conclusion, the best decision tree should have 50 cross-validation folds, the value of confidencefactor in J48 algorithm equals to 0.04 and seed equals to two.


Accuracy Rate
94.2
94
93.8
93.6
93.4
93.2
93
92.8
10
20
30
40
50
60
70
80
90
100
110
the number of folds

Graph-1