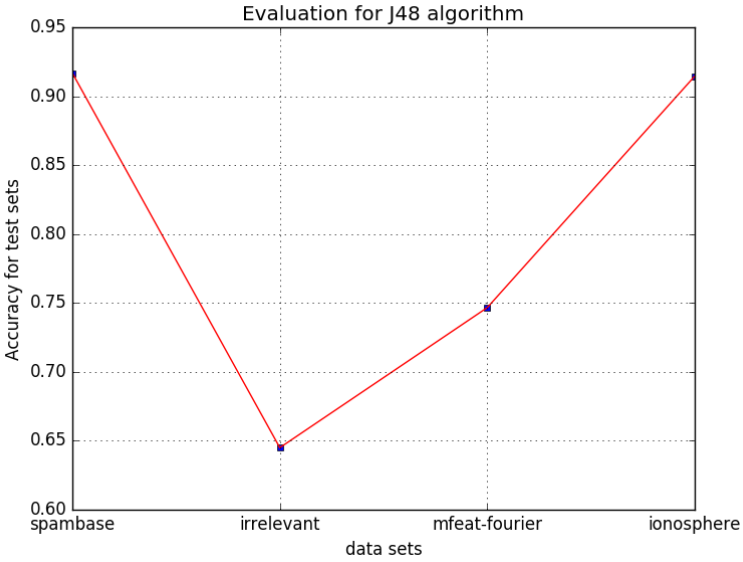
COMP135 Project 1 Report

By beibei du

3.1 Evaluating Decision trees

> login to homework.eecs.tufts.edu
> bash
> go to comp135/project1/data
> export CLASSPATH=/r/aiml/ml-software/weka-3-6-11/weka.jar:$CLASSPATH
> export WEKADATA=/r/aiml/ml-software/weka-3-6-11/data/

```
bash-4.2$ java weka.classifiers.trees.J48 -t spambase_train.arff -T
spambase_test.arff
bash-4.2$ java weka.classifiers.trees.J48 -t irrelevant_train.arff -T
irrelevant_test.arff
bash-4.2$ java weka.classifiers.trees.J48 -t mfeat-fourier_train.arff
-T mfeat-fourier_test.arff
bash-4.2$ java weka.classifiers.trees.J48 -t ionosphere_train.arff -T
ionosphere_test.arff
```

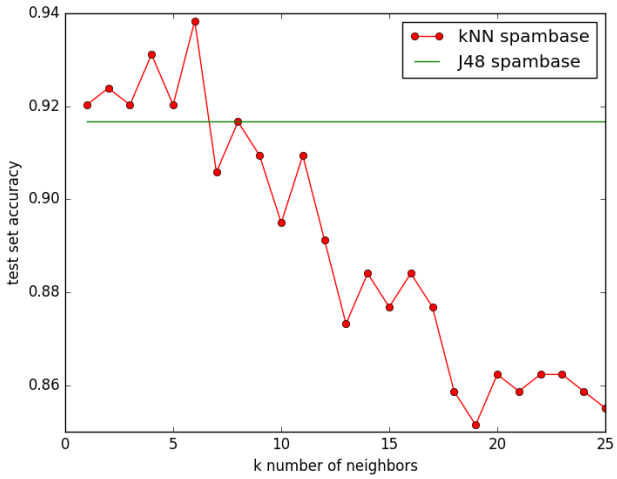


accuracy record:

File name	Train accuracy	Test accuracy
Spambase_*.arff	97.2938%	91.6785%
Irrelevant_*.arff	97.6%	64.5%
Mfeat-fourier_*.arff	97.2993%	74.6627%
Ionosphere_*.arff	99.5726%	91.453%

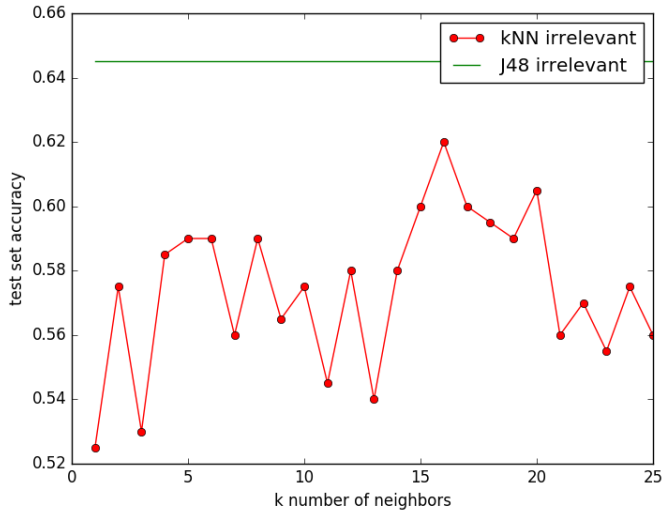
3.4 Evaluating kNN with respect to k

1) Figure 1 --- Compare with kNN and J48 on spambase dataset



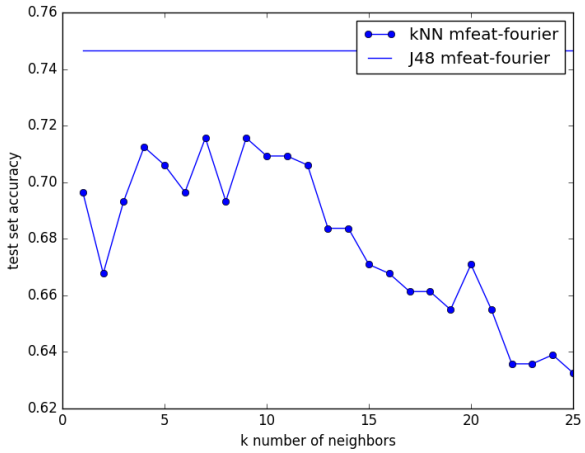
From this figure, we can find for the spambase dataset, when k <= 6, kNN performance better than J48 and also when k = 6 kNN has the highest test accuracy, but after that, it's clearly kNN test accuracy decreases and J48 performance better with higher accuracy, when k = 25, kNN algorithm has lowest test accuracy. The accuracy decreases may because more noisy data was included in the neighbor.

2) Figure 2 -- Compare with kNN and J48 on irrelevant dataset



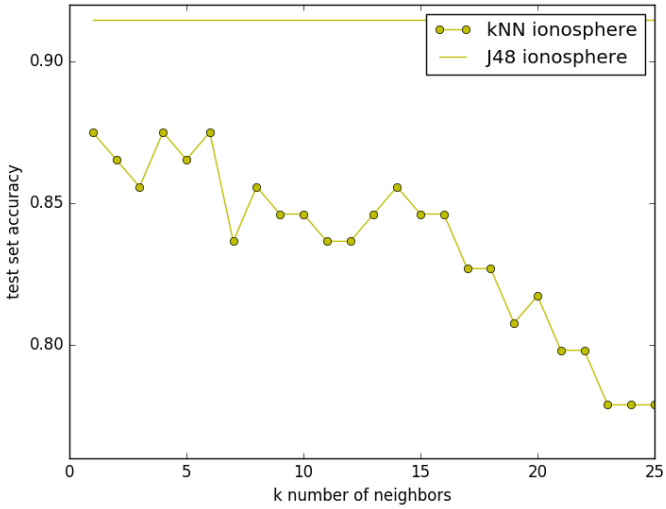
From this figure, we can find for irrelevant dataset, J48 has higher test accuracy than kNN algorithm for all k neighbors from 1 to 25. And when k = 16, it has highest accuracy but still lower than J48, when k = 1, it has lowest accuracy.

3) Figure 3 -- Compare with kNN and J48 on mfeat-fourier dataset



According to figure above, for mfeat-fourier dataset, when k = 7 or k = 9, kNN algorithm has highest test accuracy but still lower than J48 algorithm, and when k = 25, it has lowest test accuracy. After k = 9, even though the test accuracy increases in some point as like k = 20, but in big direction, it still decreases.

4) Figure 4 -- Compare with kNN and J48 on ionosphere dataset



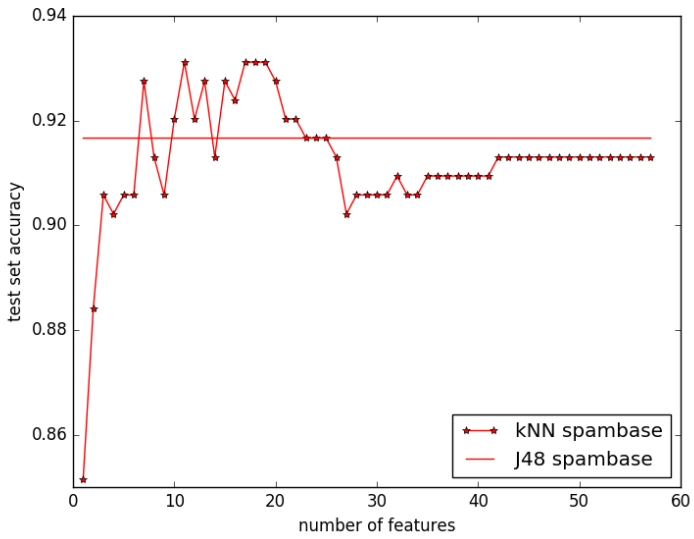
According to figure above, J48 has higher test accuracy than kNN algorithm for all k from 1 to 25 for ionosphere data set. When k = 1 or k = 4 or k = 6, it has higher test accuracy, after that, the big direction of test accuracy decreases and when k = 25, it has lowest test accuracy.

Conclude from all four figures above, J48 performance better than kNN except for spambase dataset. And for every dataset, it has higher test accuracy in some point(k neighbors).

3.5 Feature Selection for kNN

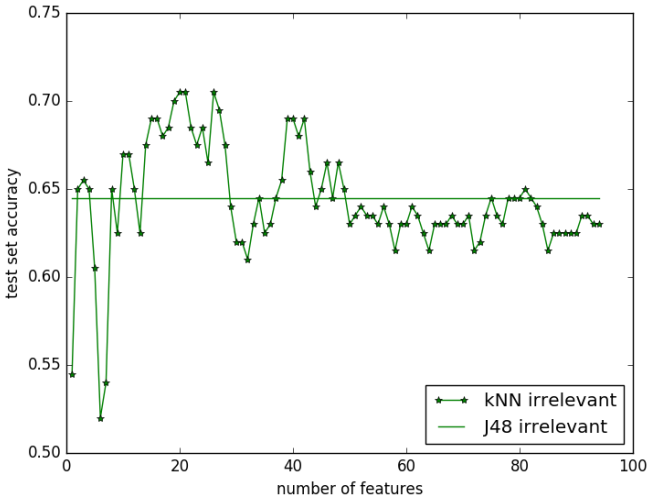
> All figures are function of n (number of features)

1) Figure 1 --- Compare with kNN and J48 on spambase dataset



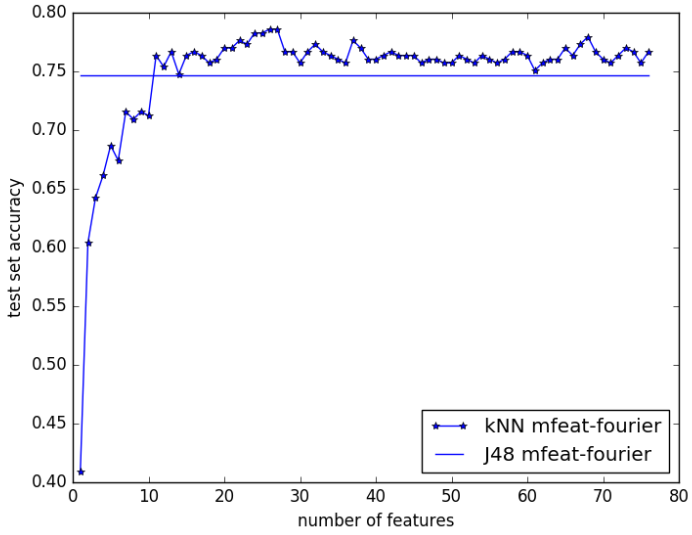
Use feature selection kNN algorithm, we can find that when n = 1, it has lowest test accuracy, and n = 7, its test accuracy is higher than J48 algorithm, after n = 26, it performance poor than J48 algorithm and also after that in big direction, the test accuracy don't have big change even the number of selected features increase.

2) Figure 2 -- Compare with kNN and J48 on irrelevant dataset



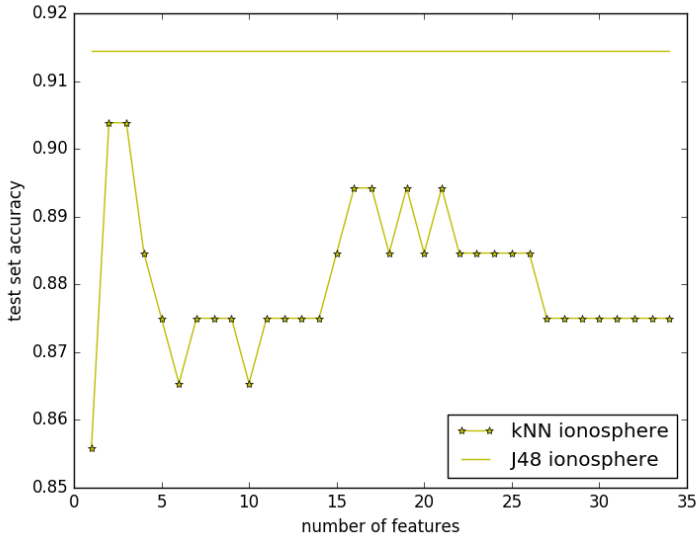
For irrelevant dataset, the kNN selection algorithm will increase first and will have high test accuracy between some number of features. Then it will decrease and below the J48 but very close J48 and don't have big change until all the feature were selected. When k = 6, it has lowest test accuracy.

3) Figure 3 -- Compare with kNN and J48 on mfeat-fourier dataset



For mfeat-fourier dataset, we can find that when n = 1, it has lowest test accuracy, the test accuracy will increase first with the number of selected features increase and will have higher test accuracy than J48 starting at some point(n), after that, it will change slowly and keep above J48.

4) Figure 4 -- Compare with kNN and J48 on ionosphere dataset



For ionosphere dataset, we can find that when n = 1, kNN has lowest test accuracy, when n = 2 or 3, it has highest test accuracy, after that, it will first decrease thane increase, then decrease and keep at some value. But compare with J48, J48 has much higher test accuracy than kNN for all number of n feature selection.

Conclude, for all of these four datasets, compare kNN selection algorithm with J48 algorithm when k = 5, we can find that except for ionosphere dataset, the other three data set will perform better than J48 in some point. For every dataset, in big direction, the test accuracy will increase then decrease with some small fluctuation and then they will keep the value with small change. Also we can know that kNN selection algorithm performs better than kNN algorithm with no feature weight. Since for kNN selection algorithm, every feature's weight is different, and the most relevant feature weights more. So in some point, it performs better than J48 also. It's not feasible to automatically select k and n, since for all these four datasets, k has same value 5, but the highest test accuracy for every dataset, n is different.