Comp 135 hw1

Fanying Ye

Due: September 29 2015

1 Introduction

In this assignment I experimented with the k nearest neighbors algorithm and the decision tree learning algorithm, and evaluate the Relief algorithm for feature weighting and selection. I used the weka system for decision tree algorithm evaluating and write Java code for others.

2 Plots and Conclusion

2.1 Evaluating Decision Trees

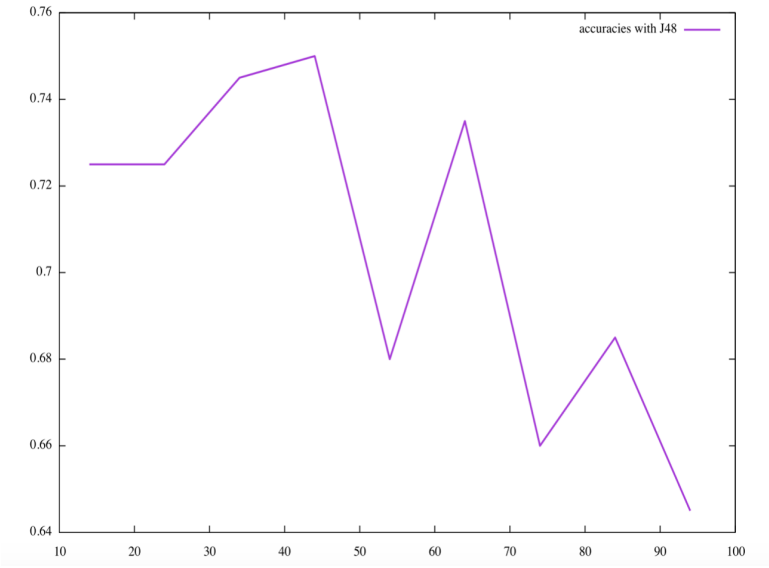

Figure 1: Accuracies of J48 algorithm

Conclusion:
The accuracy of J48 decrease after taking into account of more irrelevant fea-

tures. The reason behind this is that added features are not relevant to the classification. As a result, taking these features into account adds noise to the data, which leads to the decrease of accuracy of the classification.

2.2 Evaluating kNN

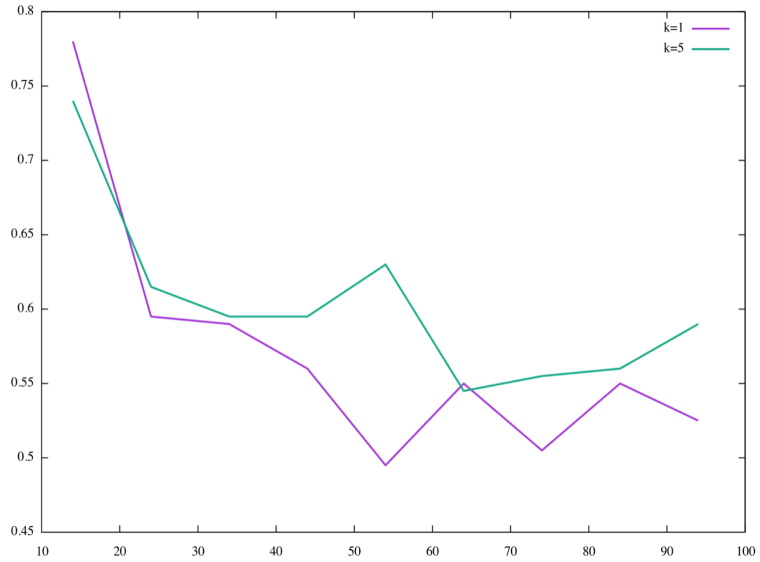

Figure 2: Accuracies of kNN algorithm

Conclusion:
The accuracy of kNN with k=5 is higher than with k=1 nearly all the time. Because larger k address the problem of "noisy" labels in training data as well as the problem of not smooth label map. The accuracy drops as adding irrelevant features for both k=1 and k=5. The reason is that kNN treats all dimensions equally and sensitive to irrelevant features.
Accuracy of kNN is higher than J48 when the number of irrelevant attribute is small, while lower comparing with J48 when adding more irrelevant features. Thus, we can conclude that kNN algorithm is more sensitive to irrelevant features than decision tree algorithm.

2.3 Evaluating kNN with Relief

Conclusion:
Accuracies of kNN with Relief are higher than without Relief. As we can see from the plot, both improvement approaches: feature selection and weighted distance, improve the accuracy of kNN algorithm. Since we select top 14 features and the first dataset contains 14 features, so the accuracy of feature selection

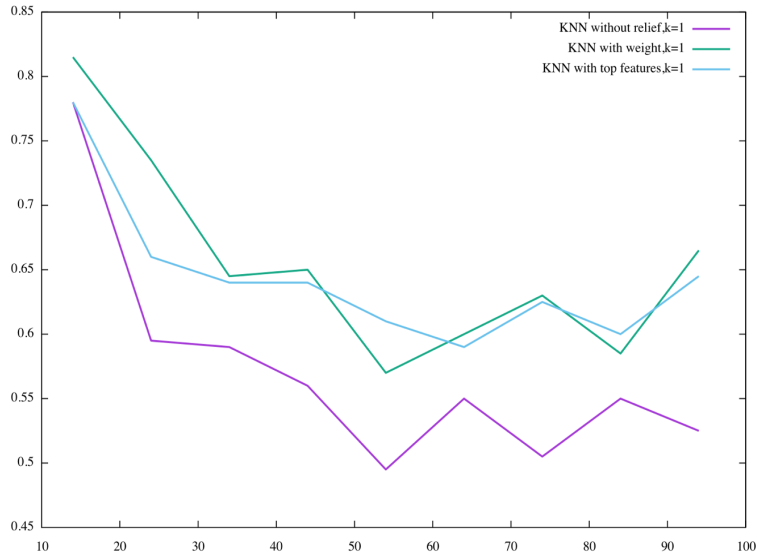

Figure 3: Accuracies with and without Relief, k=1

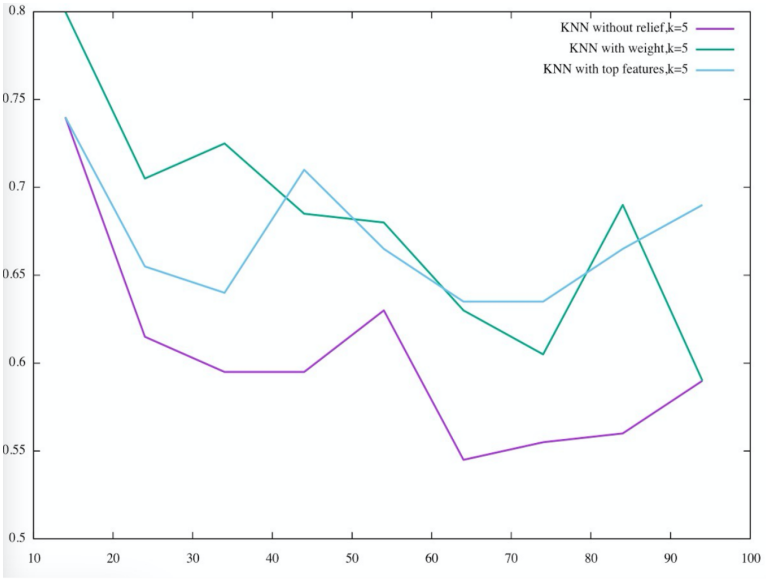

Figure 4: Accuracies with and without Relief, k=5

improvement method is the same with kNN whiout Relief at the first point. Although the Relief algorithm improve the performance of kNN, accuracy of these two methods still tends to drop when adding more irrelevant features.

2.4 Evaluating the effect of m

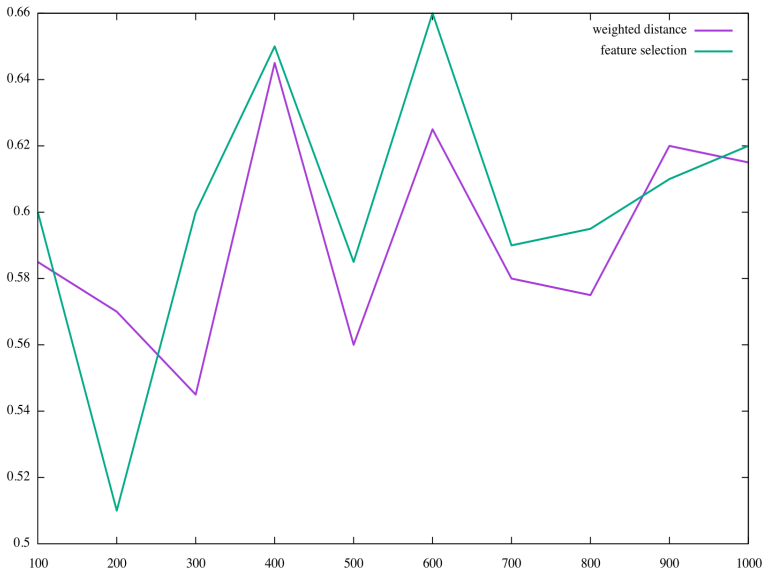

Figure 5: Accuracies of kNN(k=1) with Relief under different m

Conclusion:
In this experiment, we consider the dataset with 94 features only and repeat the evaluation for m = 100, 200, ... , 1000. From the plot, we can see that the accuracy tends to increase with the increase of m generally. However, the accuracy drops sometimes when increasing m. The reason behind this is that we only have 94 features in the dataset, choosing a larger m means that we select some instances repeatedly, which might lead to taking account of a noisy point multiple times.