Report

Introduction

This project uses Feature Selection with the Nearest Neighbor Algorithm to find the best features in a given data set. I am tasked with finding the best features for Small Data Set 43 and Large Data Set 52. The search algorithms implemented are Forward Selection and Backward Elimination. Forward Selection starts with no features and adds a feature at every level. Backward Elimination starts with all the features and removes a feature until none are left.

Data

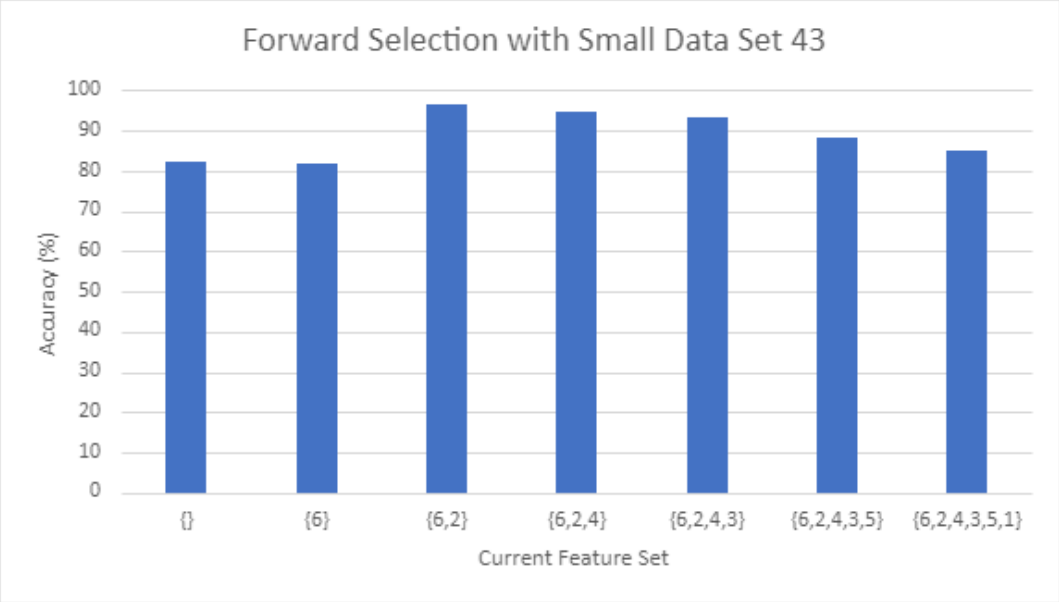

Figure 1. Accuracy of subsets of data using Forward Selection on Small Data Set 43

Figure 1 shows the results of running Forward Selection on Small Data Set 43. At the beginning of the search, there are no features, which report an accuracy value of 82.4%. At the next level, adding any feature decreased the accuracy value, but adding feature 6 decreased the accuracy value the least, leaving the accuracy at 81.8%. Afterward, feature 2 is added which dramatically increased the accuracy value to 96.6%. Feature 4 is added after but lowers the accuracy slightly to 94.8%. However, since this is a minor decrease in accuracy and floating point operations are known to be inaccurate, feature 4 should be further looked into. The process of adding features continues until all features are added, with an accuracy value of 84.8%. The set with the greatest accuracy value is {2,6}, with set {2,4,6} coming in close.

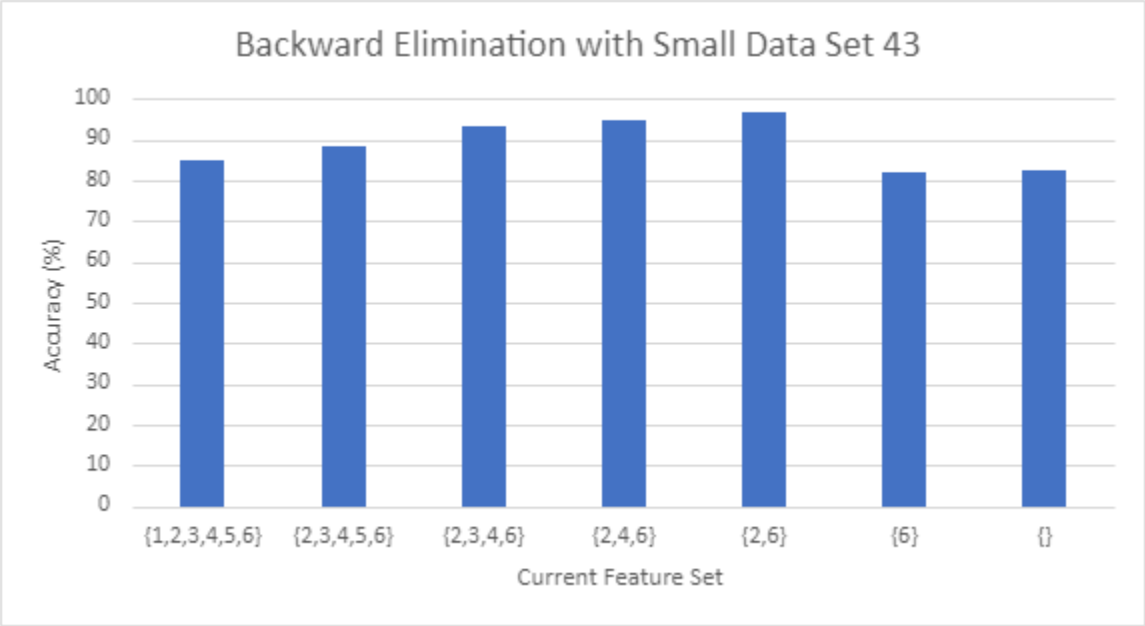

Figure 2. Accuracy of subsets of data using Backward Elimination on Small Data Set 43

Figure 2 shows the results of running Backward Elimination on Small Data Set 43. At the beginning of the search, all features are included, which reports an accuracy value of 84.8%. At the next level, removing feature 1 yields the best accuracy at 88.4%. Afterward, feature 5 is removed which increases the accuracy value to 93.4%. This process is continued until there are no features left in the set, with a default rate of 82.4%. The set with the greatest accuracy value is {2,6}, with set {2,4,6} coming in close. As stated previously, due to possible inaccuracies in accuracy calculations, feature 4 should be looked into further. This result supports the results found in Figure 1.

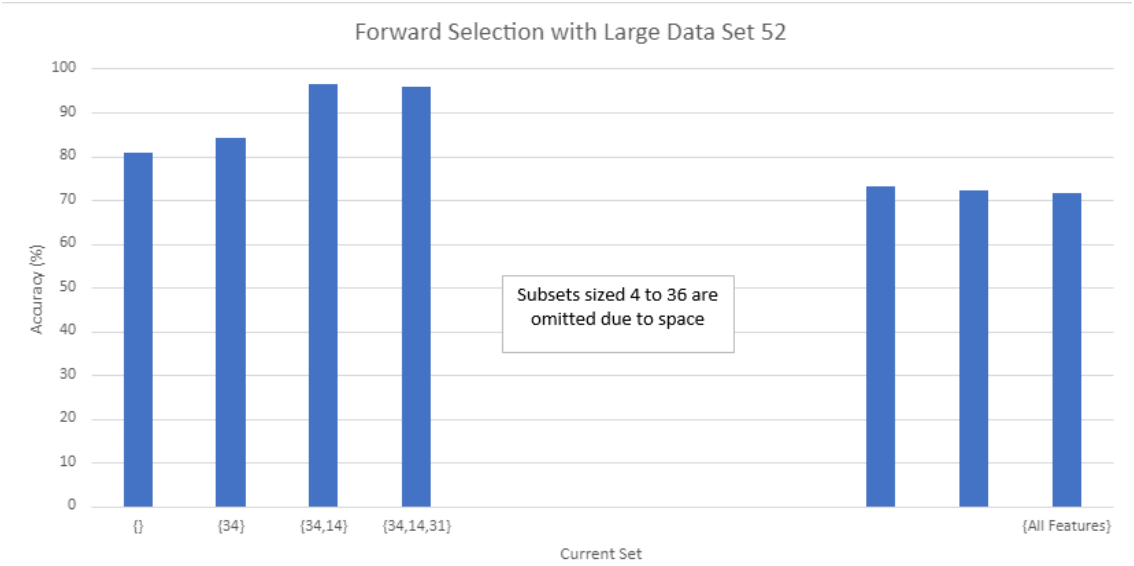

Figure 3. Accuracy of subsets of data using Forward Selection on Large Data Set 52

Figure 3 shows the results of running Forward Selection on Large Data Set 52. Subsets sized 4 to 36 are omitted from the graph due to space. At the beginning of the search with no features, there is an accuracy value of 80.7%. Adding feature 34 increases the accuracy value to 84.2%. Adding feature 14 to that set yields a 96.4% accuracy. And adding feature 31 to the set results in a 95.7% accuracy. Since these values are so close, feature 31 should be looked into further. The set with the greatest accuracy contains features {34,14} with {34,14,31} close behind. At the end of the search, it is found that the accuracy with all features is 71.7%.

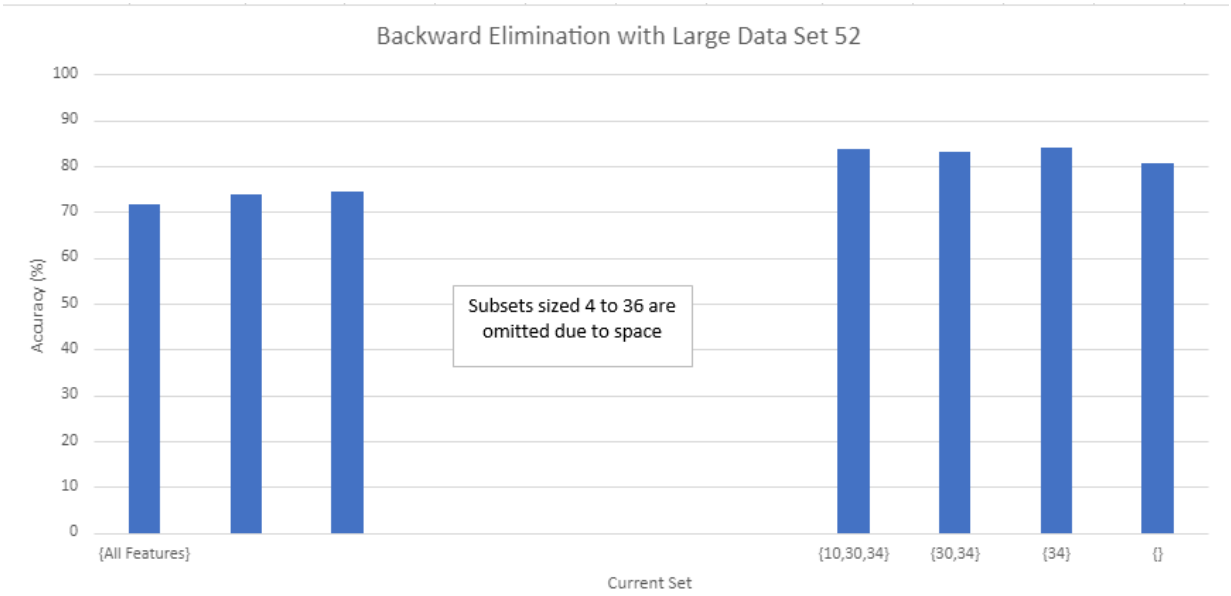

Figure 4. Accuracy of subsets of data using Backward Elimination on Large Data Set 52

Figure 4 shows the results of running Backward Elimination on Large Data Set 52. Subsets sized 4 to 36 are omitted from the graph due to space. At the start of the search, all features are included with an accuracy of 71.7%. Feature 14 is first removed, which yields an accuracy of 73.7%. Towards the end of the search, feature 34 is the last feature remaining, which also yields the highest overall accuracy found in this search with an accuracy of 84.2%. This search only found one feature, which shows that Forward Selection is more accurate, but gives confidence in feature 34.

Conclusion for Small Data Set 43

I believe that features “2” and “6” are the best features. Using these 2 features, we have an accuracy value of 0.966 or 96.6%. Feature “4” should be looked further into.

Conclusion for Large Data Set 52

I believe that features “14” and “34” are the best features. Using these 2 features, we have an accuracy value of 0.964 or 96.4%. Feature “31” should be looked further into.

Computational Effort for Search

Hardware tested on: Laptop with Intel Core i7-11800H Processor, 16 GB of RAM (15.7 GB usable)

Small Data Set 43 (6 features, 500 instances):
- Forward Selection: Approximately 1.9 seconds
- Backward Elimination: Approximately 2.3 seconds

Large Data Set 52 (40 features, 1000 instances):
- Forward Selection: Approximately 24.86 minutes
- Backward Elimination: Approximately 31.42 minutes

Forward Selection Example with Small Data Set 96

Note: CS170_SuperSmall_Data__43.txt contains Small Data Set 96's data

```
Welcome to Nathan's Feature Selection Algorithm.
Type in the name of the file to test:
CS170_SuperSmall_Data__43.txt
Type the number of the algorithm you want to run.
1. Forward Selection
2. Backward Elimination
1
This dataset has 6 features (not including the class attribute), with 500 instances.
Beginning Search
On level 1 of the search tree
--Testing feature 1 with current accuracy of: 0.874
--Testing feature 2 with current accuracy of: 0.682
--Testing feature 3 with current accuracy of: 0.734
--Testing feature 4 with current accuracy of: 0.718
--Testing feature 5 with current accuracy of: 0.672
--Testing feature 6 with current accuracy of: 0.746
On level 1, I added feature 1 with accuracy of 0.874
On level 2 of the search tree
--Testing feature 2 with current accuracy of: 0.826
--Testing feature 3 with current accuracy of: 0.866
--Testing feature 4 with current accuracy of: 0.82
--Testing feature 5 with current accuracy of: 0.836
--Testing feature 6 with current accuracy of: 0.948
On level 2, I added feature 6 with accuracy of 0.948
On level 3 of the search tree
--Testing feature 2 with current accuracy of: 0.904
--Testing feature 3 with current accuracy of: 0.94
--Testing feature 4 with current accuracy of: 0.904
--Testing feature 5 with current accuracy of: 0.938
On level 3, I added feature 3 with accuracy of 0.94
On level 4 of the search tree
--Testing feature 2 with current accuracy of: 0.88
--Testing feature 4 with current accuracy of: 0.884
--Testing feature 5 with current accuracy of: 0.892
On level 4, I added feature 5 with accuracy of 0.892
On level 5 of the search tree
--Testing feature 2 with current accuracy of: 0.828
--Testing feature 4 with current accuracy of: 0.864
On level 5, I added feature 4 with accuracy of 0.864
On level 6 of the search tree
--Testing feature 2 with current accuracy of: 0.836
On level 6, I added feature 2 with accuracy of 0.836
------------------------------------------------------------
Best features to use:
1 6
The best accuracy from this data set is: 0.948
Duration: 1891 milliseconds
```

Backward Elimination Example with Small Data Set 96

Note: CS170_SuperSmall_Data__43.txt contains Small Data Set 96's data

```
Welcome to Nathan's Feature Selection Algorithm.
Type in the name of the file to test:
CS170_SuperSmall_Data__43.txt
Type the number of the algorithm you want to run.
1. Forward Selection
2. Backward Elimination
2
This dataset has 6 features (not including the class attribute), with 500 instances.
Beginning Search
On level 1 of the search tree
--Testing removing feature 1 with current accuracy of: 0.736 after removing the feature from the set
--Testing removing feature 2 with current accuracy of: 0.864 after removing the feature from the set
--Testing removing feature 3 with current accuracy of: 0.816 after removing the feature from the set
--Testing removing feature 4 with current accuracy of: 0.828 after removing the feature from the set
--Testing removing feature 5 with current accuracy of: 0.872 after removing the feature from the set
--Testing removing feature 6 with current accuracy of: 0.818 after removing the feature from the set
On level 1, I removed feature 5
On level 2 of the search tree
--Testing removing feature 1 with current accuracy of: 0.73 after removing the feature from the set
--Testing removing feature 2 with current accuracy of: 0.884 after removing the feature from the set
--Testing removing feature 3 with current accuracy of: 0.862 after removing the feature from the set
--Testing removing feature 4 with current accuracy of: 0.88 after removing the feature from the set
--Testing removing feature 6 with current accuracy of: 0.792 after removing the feature from the set
On level 2, I removed feature 2
On level 3 of the search tree
--Testing removing feature 1 with current accuracy of: 0.74 after removing the feature from the set
--Testing removing feature 3 with current accuracy of: 0.904 after removing the feature from the set
--Testing removing feature 4 with current accuracy of: 0.94 after removing the feature from the set
--Testing removing feature 6 with current accuracy of: 0.83 after removing the feature from the set
On level 3, I removed feature 4
On level 4 of the search tree
--Testing removing feature 1 with current accuracy of: 0.708 after removing the feature from the set
--Testing removing feature 3 with current accuracy of: 0.948 after removing the feature from the set
--Testing removing feature 6 with current accuracy of: 0.866 after removing the feature from the set
On level 4, I removed feature 3
On level 5 of the search tree
--Testing removing feature 1 with current accuracy of: 0.746 after removing the feature from the set
--Testing removing feature 6 with current accuracy of: 0.874 after removing the feature from the set
On level 5, I removed feature 6
On level 6 of the search tree
--Testing removing feature 1 with current accuracy of: 0.816 after removing the feature from the set
On level 6, I removed feature 1
------------------------------------------------------------
Best features to use:
1 6
The best accuracy from this data set is: 0.948
Duration: 2036 milliseconds
```

Program Inputs and Outputs

Inputs: The program first asks the user to input the file name containing the data that the user wants to perform feature selection on. Subsequently, the program then asks if the user wants to perform Forward Selection or Backward Elimination and to pick. If the file does not exist in the same directory as the code, the user can input a path to it when prompted for the file name. If the file does not exist, the program will exit.

Outputs: The program first states how many features the file contains and the number of instances. It will then run the search the user chose, displaying the accuracies along the way. After the search is completed, the best features to pick are shown. The corresponding accuracy is also displayed with the run time of the program in milliseconds.