AI Based Classification of Cancer Data

Emmett Forrestel[1], Alan Zaoxing Liu[2]

Clarence High School, 9625 Main Street[1], BU Photonics Center, 8 St Mary's St[2]

Introduction

- The purpose of this research is to predict cancer risk.
- The features are more than thirteen hundred proteins, the subject's age, a urinary PSA test, and a serum PSA test.
- The given data is very sparse with many of the protein tests only having been done on a small amount of the total one hundred ninety three subjects.
- Many different machine learning models were used to try to accurately classify the subjects.

Testing

- Models include: Gradient boosted decision tree, Adaboost decision tree, neural net, random forest, etc.
- Preparation techniques: standard scaler, principal component analysis, truncated singular value decomposition, dropping null columns
- Parameter tuning: grid Search CSV, bagging models.
- These algorithms were trained on a set of data that gave the cancer risk of each patient, and then tested and evaluated on data without being given the cancer risk for each patient.

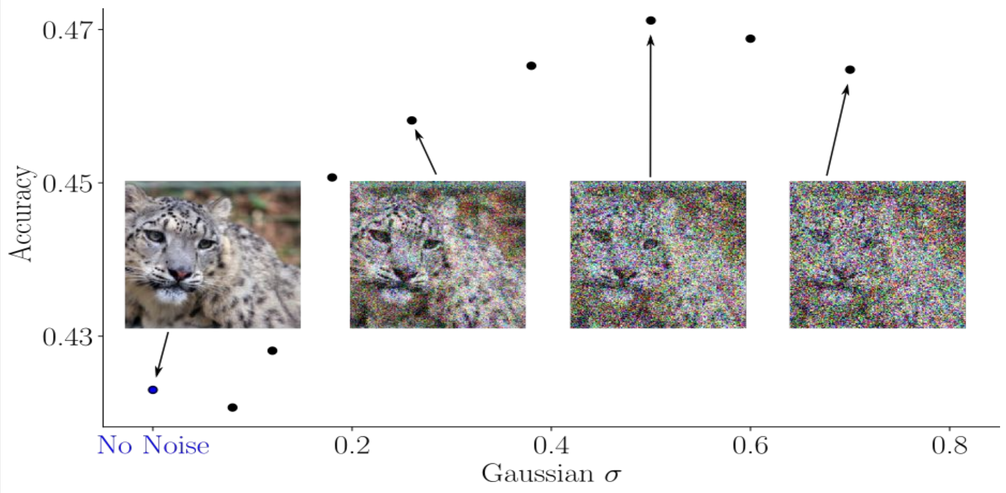
Problems and Solutions

- To work around the sparsity any test that is more than half empty values was not considered in training or testing.
- Reducing the amount of tests to 444 from 1347.

	F33	F34	F35	F36	F37	F38	F39
1	0	0	9544.169699	0	0	0	0
2	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0
15	0	0	1831.773015	1644.919305	0	0	0
16	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0
23	0	0	15405.49193	0	0	0	0
24	0	0	9000.061294	0	0	0	0
25	0	0	0	0	0	0	0
26	0	0	0	0	0	3359.190812	0
27	0	0	0	0	0	0	0

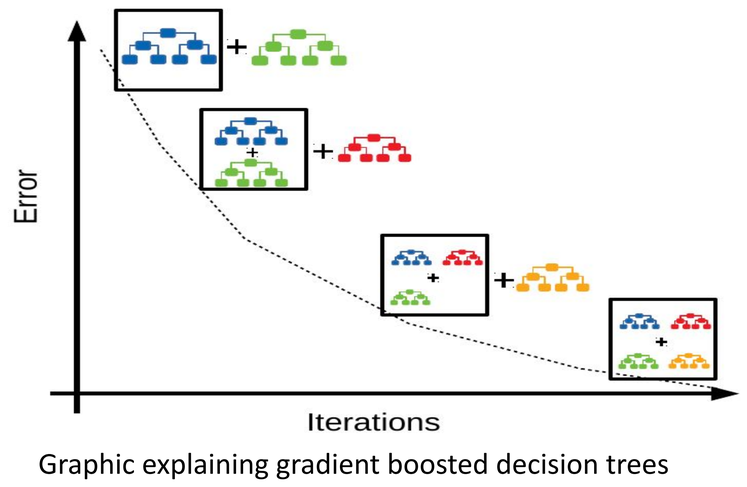
An example of the sparsity of the data

- Standard scaler was applied to fix all values on a scale from 0-1 making them easily comparable.
- To make the algorithm better at generalizing gaussian noise was added to the data, slightly augmenting the data.
- Sparse principal component analysis was used to discover the most important features in predicting cancer risk.



Graphic explaining gaussian noise

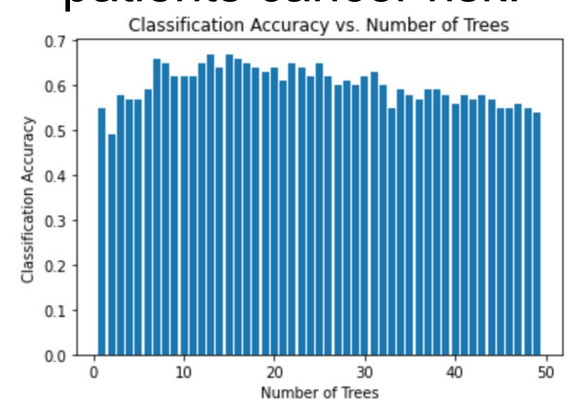
- A gradient boosted decision tree was the most successful model in classifying the data.
- A maximum depth of 3, minimum samples per leaf of 5, and 13 trees provided the best results.



Graphic explaining gradient boosted decision trees

Conclusions

- Gradient boosted decision tree is the final algorithm selected.
- Data preparation techniques: standard scaler, Sparse PCA, dropping mostly void columns, adding gaussian noise, and grid search CSV.
- This combination of techniques provides a tool to help identify a patients cancer risk.



Variation of accuracy based of number of trees used

- The algorithm had a peak accuracy of 0.71
- Dropping null columns to reduce sparsity, and using PCA were the most helpful tools to improve model performance.
- These preliminary results stand as a step for future AI based cancer risk assessment.

Acknowledgements

Assistant Professor, Alan Zaoxing Liu, Department of Electrical and Computer Engineering