EECS E6690: Statistical Learning for Biological and Information Systems

Final Project

lz2684 sl4401 jw3816

Paper

Scaling up the accuracy of Naive-Bayes classifiers: a decision-tree hybrid

Content

- Compare Naive-Bayes and Decision-Tree
- A new algorithm: NBTree
- Test on 29 datasets

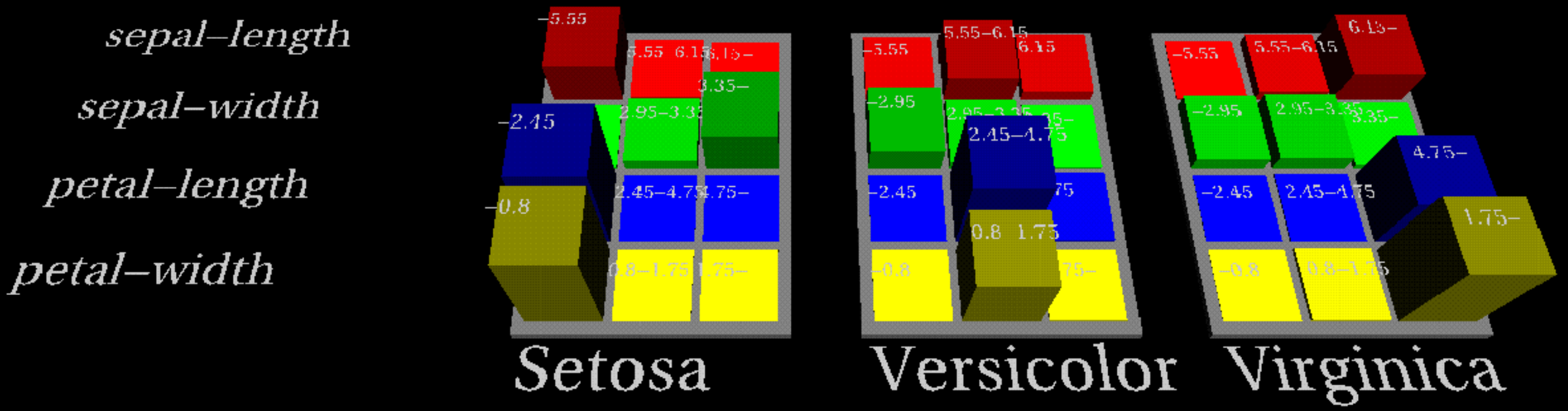
Our work

- Reproduce the result
- Compare the accuracy of NBTree and C5.0 Tree

Naive Bayes Classifier

- Based on independence assumption
- Good interpretability

- Good interpretability



Reference:Kohavi, Ron. (1997). Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid. KDD.

Log probabilities are evidences to determine the class

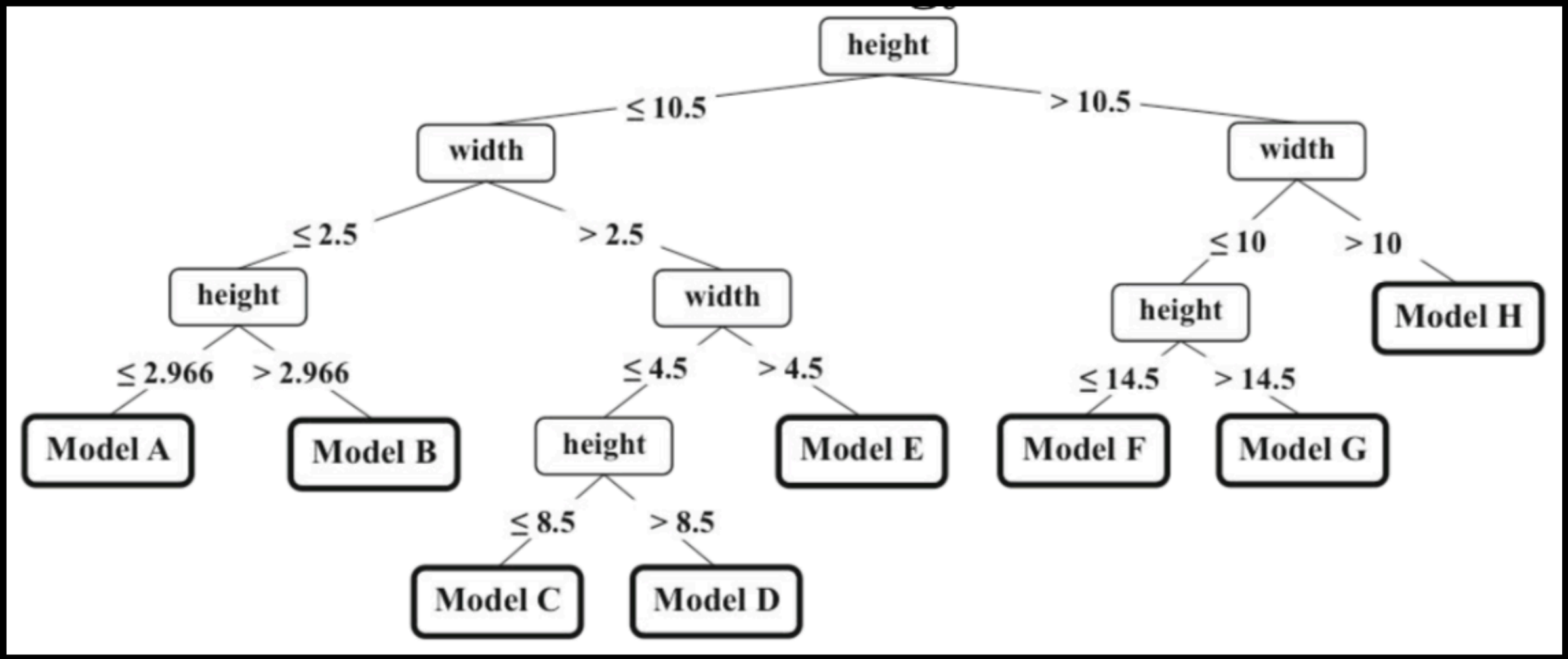
Naive Bayes Classifier

- Based on independence assumption
- Good interpretability
- Downside: Accuracy asymptotes

Decision Tree

- Good interpretability
- Accuracy doesn’t asymptote
- Segments the data

NB-Tree



Reference: Gajderowicz, Bart & Sadeghian, Alireza & Soutchanski, Mikhail. (2013). Ontology Enhancement Through Inductive Decision Trees. 10.1007/978-3-642-35975-0_14.

NB-Tree

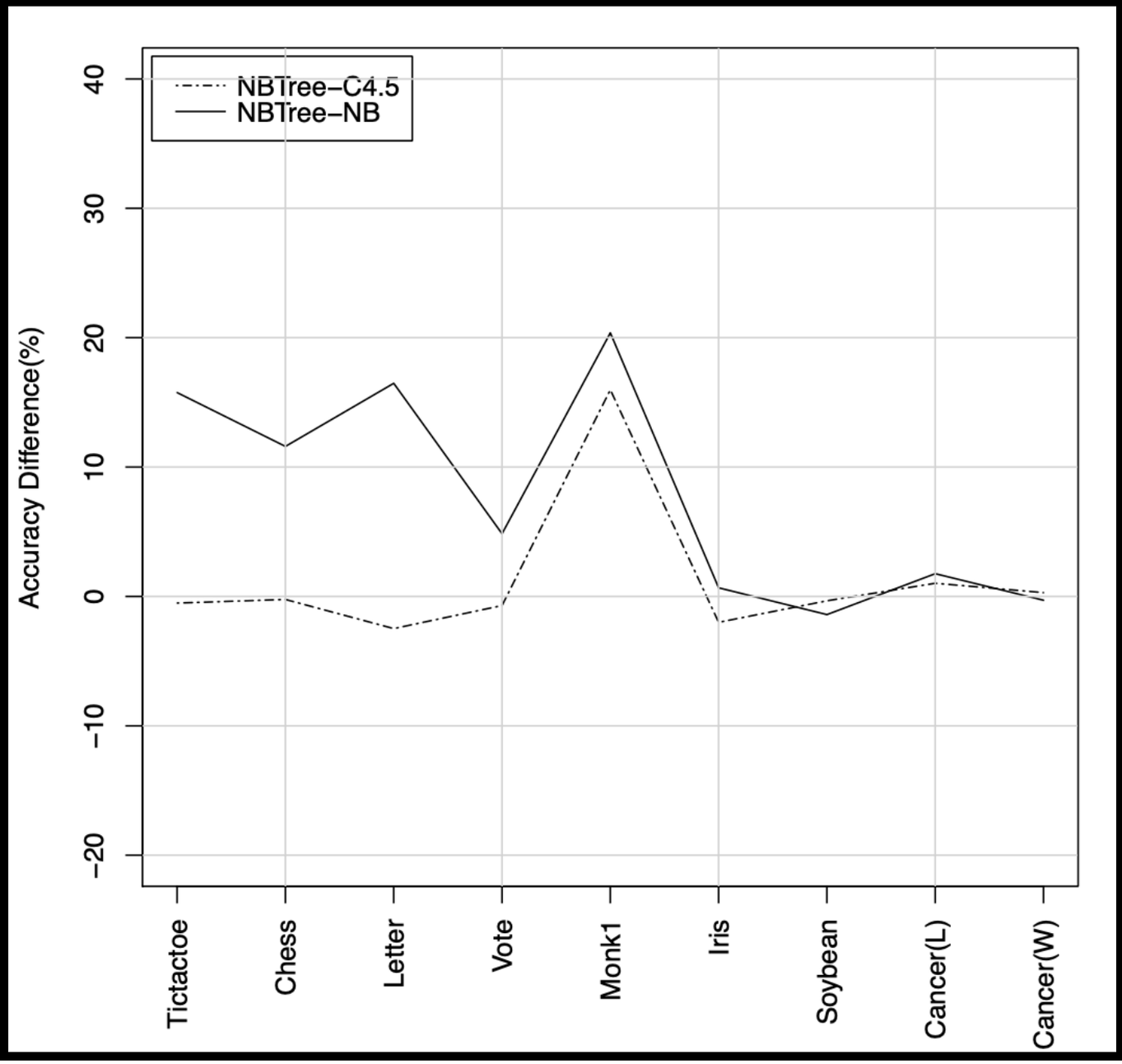
Why NB-Tree

- Compared to Naive Bayes
- Compared to Decision Tree

Dataset

Dataset	No. Attributes	Train Size	Test Size
tic-tac-toe	9	958	CV-10
chess	36	2130	1066
letter	16	15000	5000
vote	16	435	CV-10
monk1	6	124	432
iris	4	150	CV-10
soybean-large	35	562	CV-10
breast-cancer(L)	9	277	CV-10
breast-cancer(W)	10	683	CV-10

Result



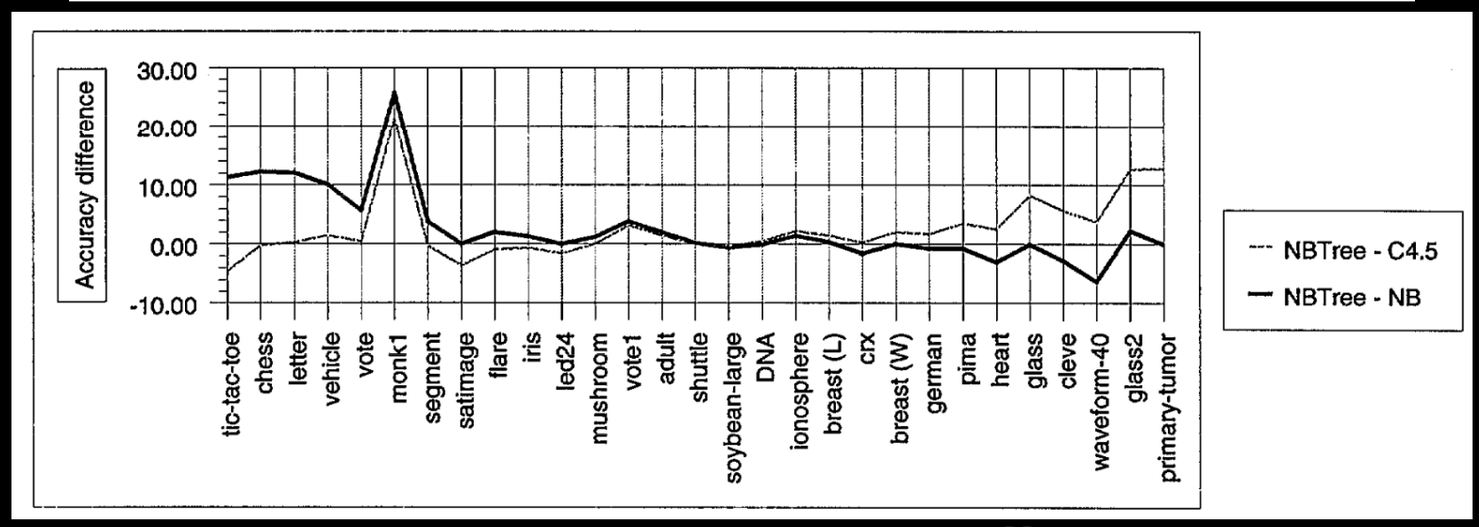
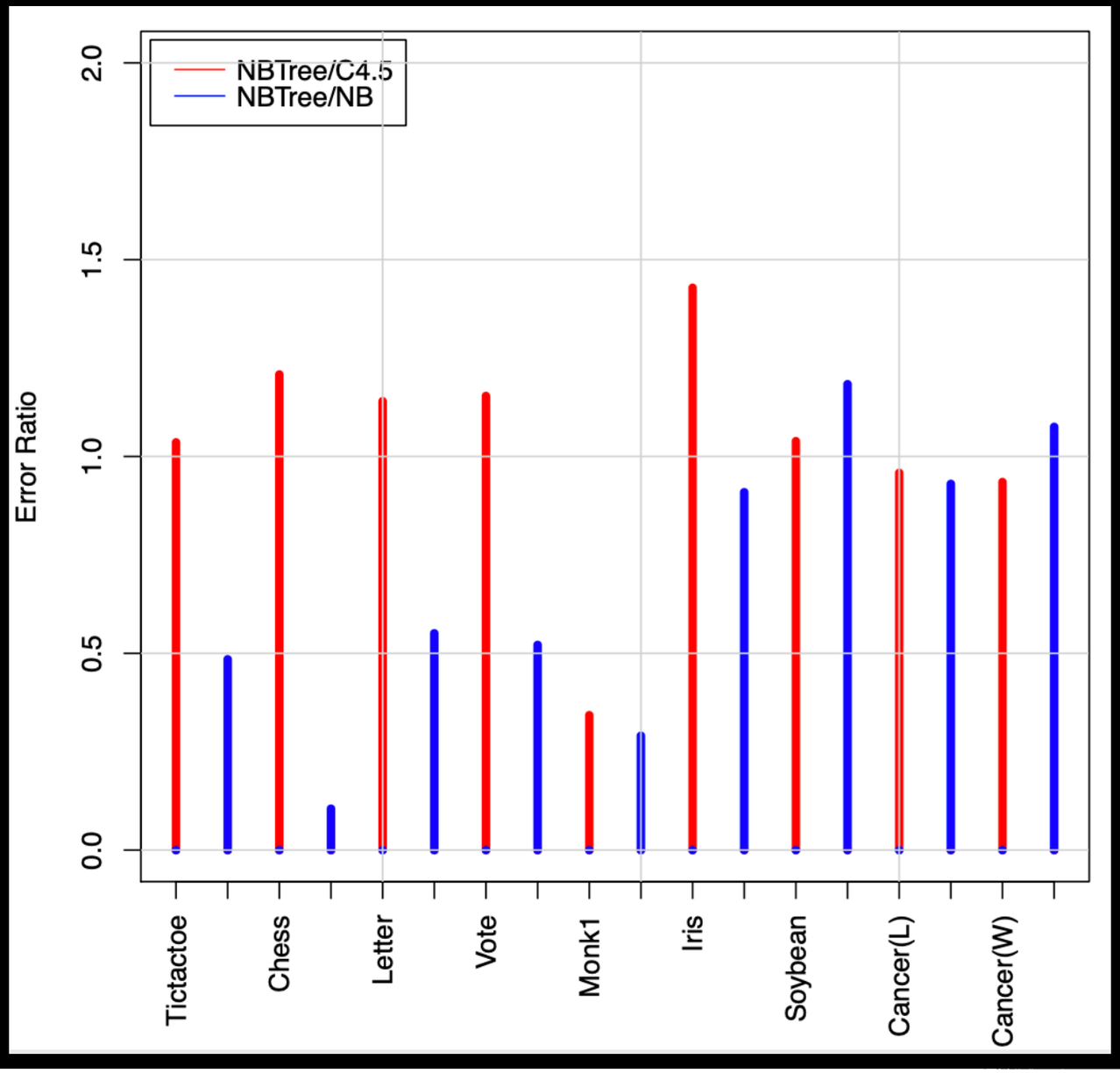


Figure 1. The accuracy differences

Result



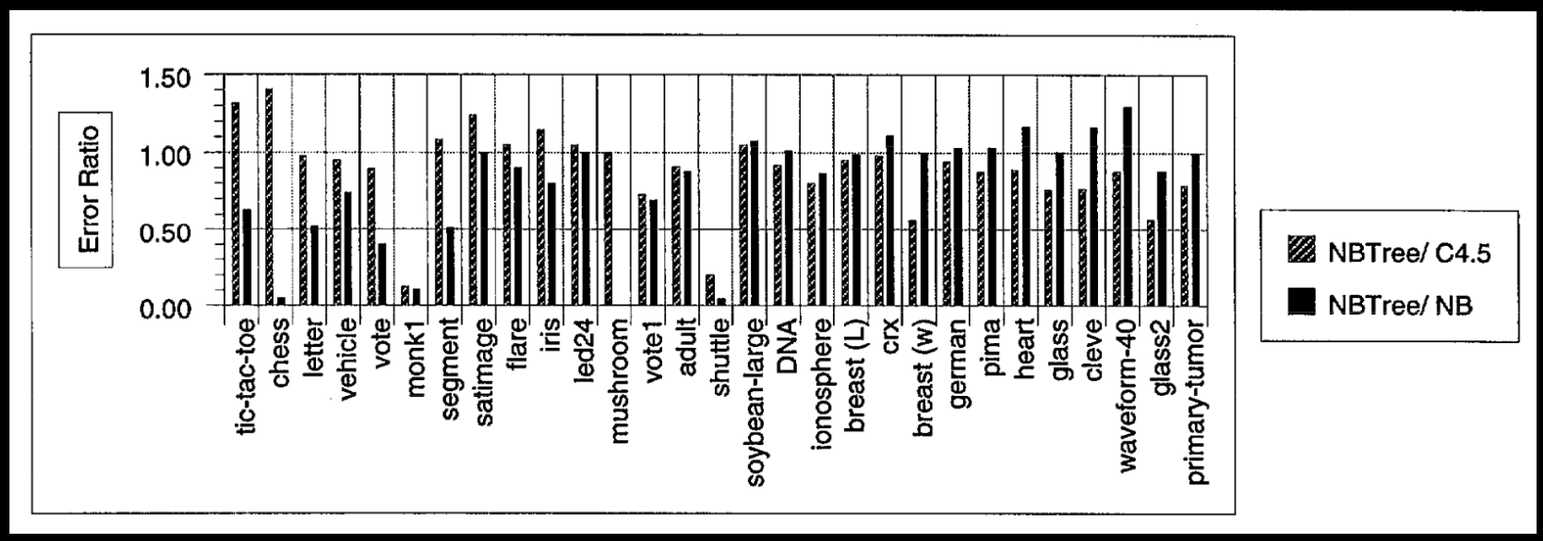


Figure 2. The error ratios of NBTree to C4.5 and Naive-Bayes.

New method: C5.0 Decision Tree

- Add Boosting on C4.5
- Improve accuracy

Experiment

- Test on the same datasets
- Boosting parameter trials=5
- Select features before constructing the tree

Result

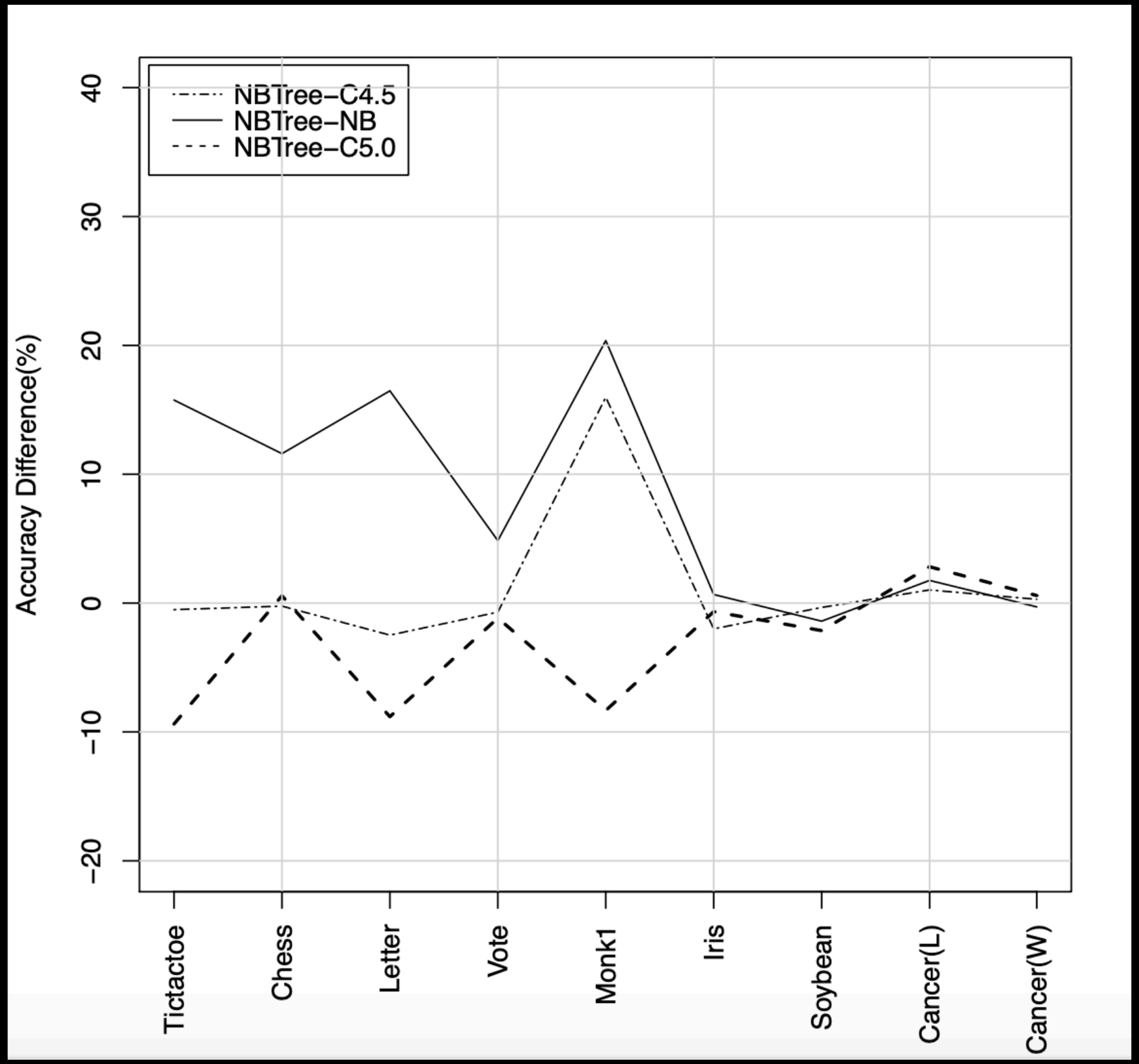


Figure 3. C5.0 Decision Tree

Thank you for your listening!