A Comparison of Selected Parametric and Non-Parametric Statistical Approaches for Candidate Genes Selection in Transcriptome Data

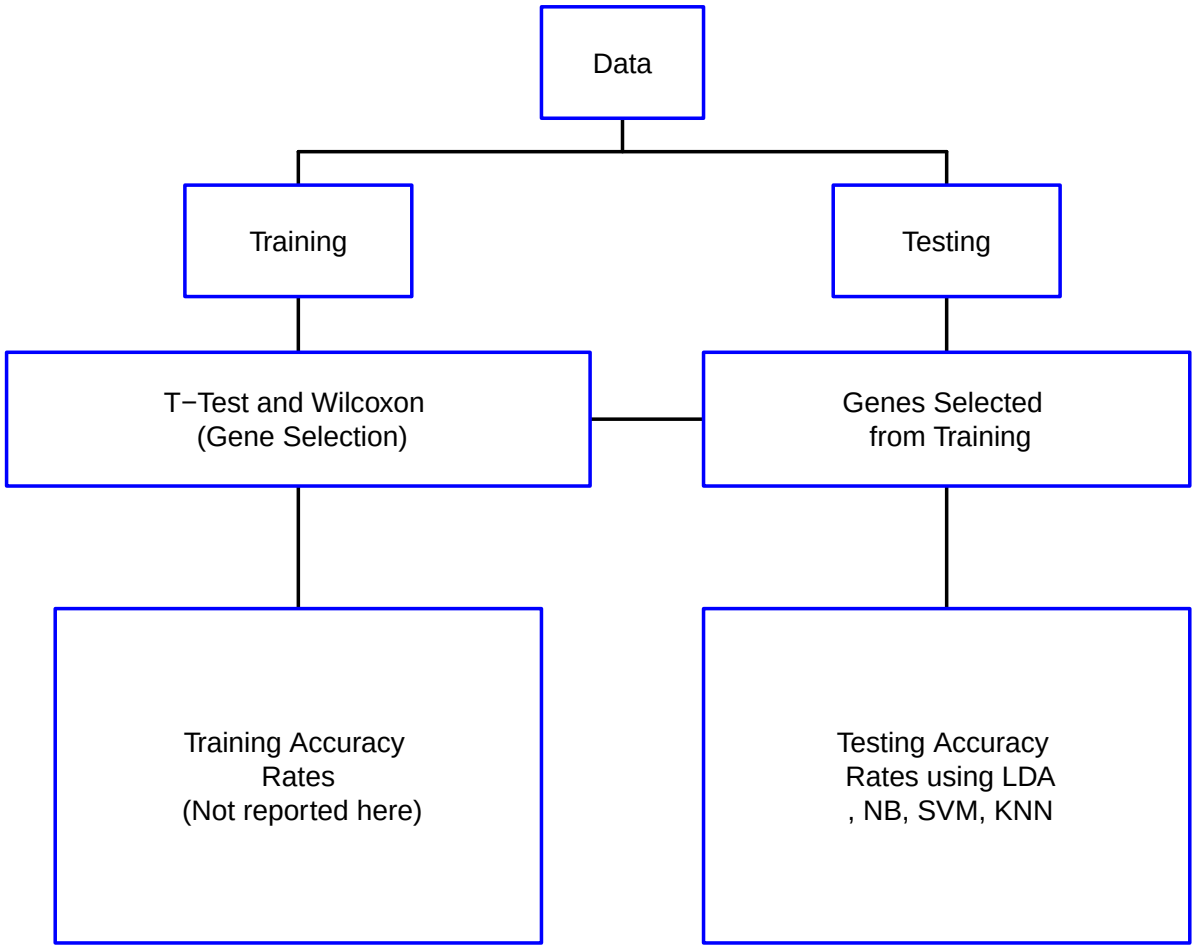
Dawit G. Tadesse, PhD
Cincinnati Children's Hospital Medical Center
dawit.tadesse@cchmc.org

SDSS2018

May 18, 2018

Introduction

Flow Chart for Data Analysis



Methods

Gene Selection Methods

- Two-sample t-test
- Wilcoxon Mann-Whitney
- Common Genes from the above two

Discriminant Functions

- Linear Discriminant Analysis (LDA)
- Naive Bayes Discriminant Function (NB)
- K-Nearest Neighbor (KNN)
- Support Vector Machine (SVM)

Data Set

Atopic dermatitis is a skin disease characterized by areas of severe itching, redness, scaling, and loss of the surface of the skin.

Training Data (GEO ID: GSE36842)

- There are 24 AD cases and 15 normal

Testing Data (GEO ID: GSE16161)

- There are 9 AD cases and 9 normal

Results

- The genes selected by t-test reach 100% testing accuracy rates for all methods.
- Only NB reach 100% accuracy for the genes selected by WMW.

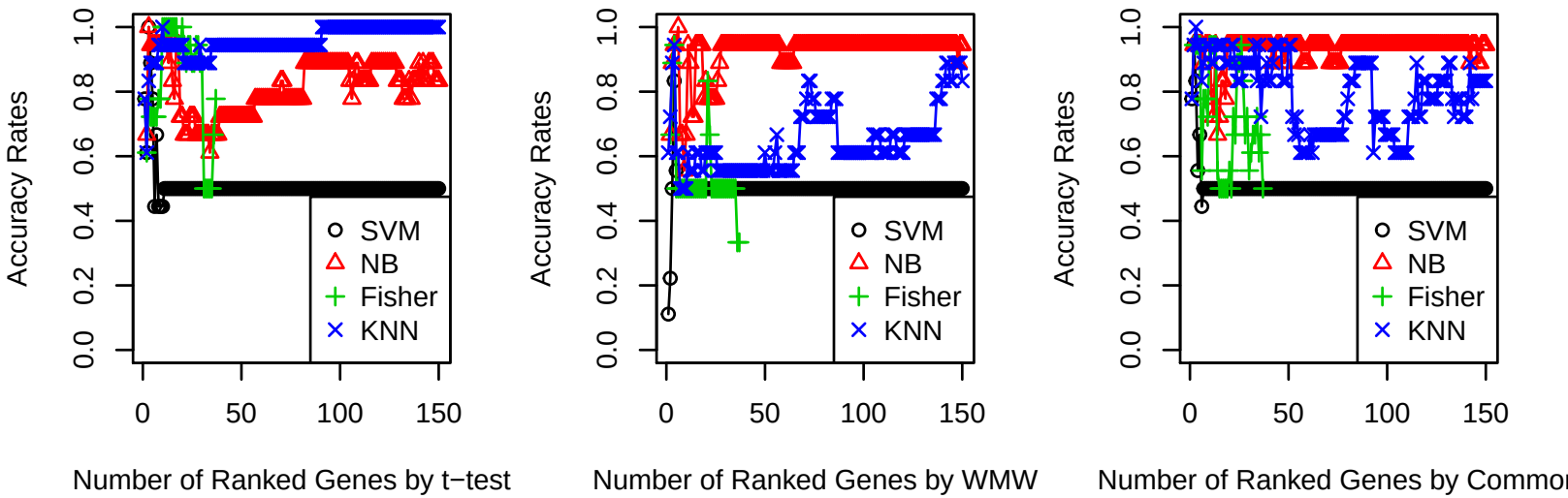


Figure 1: Testing accuracy rates

Results

- KNN for the genes selected by t-test has a robust 100% accuracy rate.

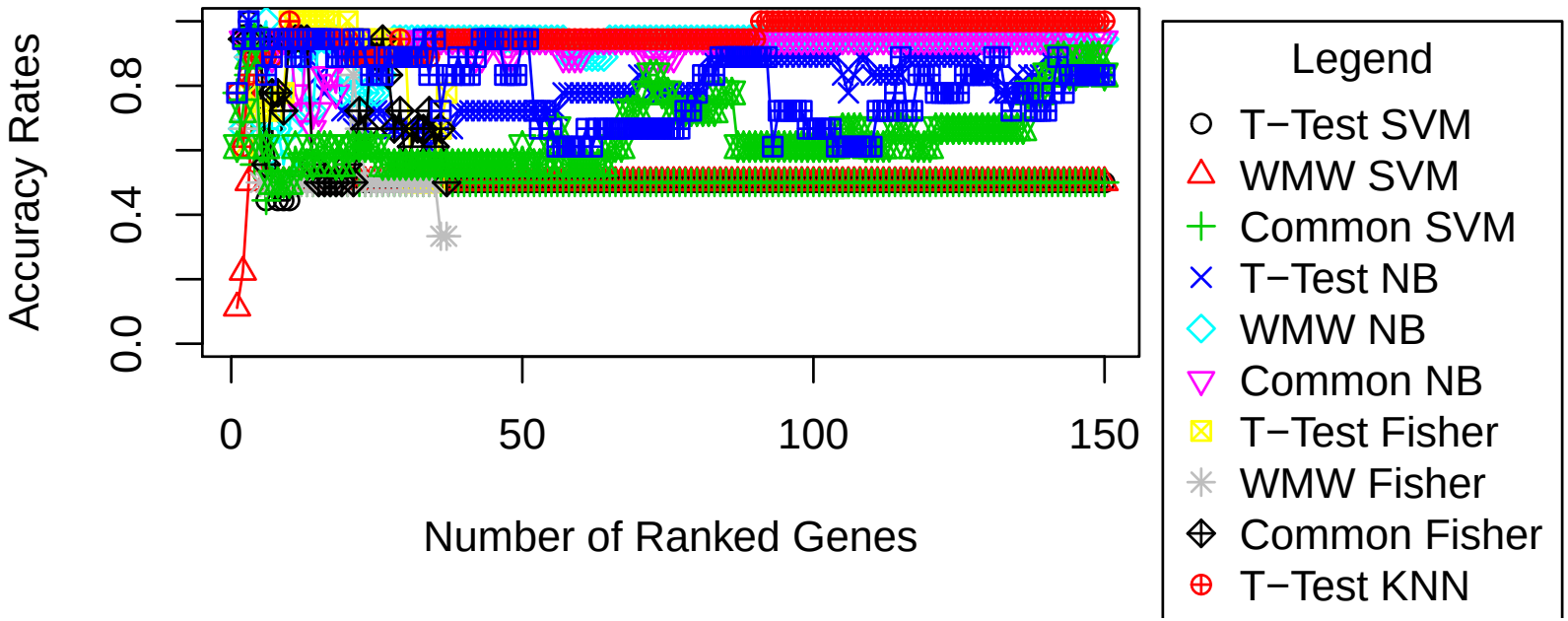


Figure 2: Testing accuracy rates for all methods

Conclusion

- We compared the parametric two sample t-test and the non-parametric Wilcoxon Mann-Whitney gene selection methods.
- We have found that the two gene selection methods choose very different sets of genes.
- We have also found that regradless of the discriminant functions used, two sample t-test choose the most important genes interms of classification.
- Our methods are more robust for other gene expression data sets as we use testing data rather than cross validation.

Selected Bibliography

- Dawit G. Tadesse and Mark Carpenter (2016), On High-Dimensional Classification for Sparse Signals, Applied Probability and Statistics, **11:1**, 01-24.
- Fan, J. and Fan, Y. (2008). High dimensional classification using features annealed independence rules. Ann. Statist., **36**, 2605-2637.
- Lin, D., Jinwen, MA., and Jian (2004), PEI., Rank sum method for related gene selection and its application to tumor diagnosis, Chinese Science Bulletin, volume **49** No. 15, 1652-1657.

Thank You!