EECS 445: PROJECT PROPOSAL

NAND DALAL
JEFFREY LIN
CHARLES LEWIS
ERIC TASEKI

Contents

1. Problem statement

DNA sequences are encoded into mRNA sequences which are translated into amino acid sequences composing proteins which perform cellular functions. Array based methods enable measurements of expression levels of thousands of genes and has accelerated the collection of gene expression pattern information. Mutations in these genes cause normal cells to evolve into cancer cells. Current methods of analyzing which permutations of gene expressions produces cancer cells involves tedious testing of thousands of genes, which calls for an efficient method of extracting the relevant information from these large data sets.

2. Significance

Cancer affects millions, and often times survival depends on time of diagnosis. However, a large portion of cancers are detected at late stages. Gene expression levels can tell us how various genes interact and about their contribution to cellular functions that may produce cancerous tissue. Machine learning algorithms can be applied in this situation to understand what combinations and expression levels of genes contribute to gene mutation and production of cancer cells. Thus, given a better understanding of the effects of gene expression levels diagnosis/detection of cancer can be accomplished earlier.

3. Related work/Novelty

Supervised and unsupervised learning has been applied in the form of classification algorithms in an attempt to group genes according to cellular functionality. This project will aim to apply similar classification algorithms with two key differences. First, multiple sets of data on multiple types of cancer currently exist, so our project will aim to analyze a data set that has not been studied in depth. Second, because humans have thousands of genes, and a mutation in any one of these has the capability to produce cancer cells. This means for each example data point we have, there exists thousands of feature gene expression levels. Our project will aim to perform dimensionality reduction techniques such as principal component analysis to group related feature gene expression levels.

4. Proposed method/approach

The following proposed method is a rough idea about one method in which to analyze cancer data. First, k means or k nearest to classify genes Second, PCA Third, real analysis

5. Evaluation

6. Data and Resources

- `http://www.inf.ed.ac.uk/teaching/courses/dme/html/datasets0405.html`
- `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html#colon-cancer`
- `http://www.cs.huji.ac.il/~nirf/Papers/BBFNSYFull.pdf`
- `http://users.soe.ucsc.edu/~karplus/abe/Science_Fair_2012_report.pdf`