CS378 - Data Mining

Final Project Proposal

Donald Li, Peter Zheng and Yifeng Xu

1 Motivation and objective

Cancer of unknown primary (CUP) is defined as malignancy without known origin at the time of the initial diagnosis, thus representing a heterogeneous group of tumors with varying clinical features. This creates many challenges when physicians are treating patients with CUP. Recent advances in high throughput sequencing technologies have given rise to huge amount of mutational data in many cancer patients. This creates many opportunities for data mining that will give clinicians treatment directions according to each patients mutational profiles.

Using random forest and other classification algorithms, we hope to create a model that can effectively predict cancer types according to sequencing data. We intend on obtaining sequencing data from cBioPortal (http://www.cbioportal.org/), an online database containing Next Generation Sequencing (NGS) from many cancer centers and large scale multi-center sequencing projects such as TCGA. For training data, we decided to train our classification model with the 10,000 patient MSK-IMPACT cohort (labeled with cancer types), which contains sequencing data from the targeted NGS sequencing panel on 341 oncogenes. For testing data, we plan on using TCGA datasets (labeled with cancer types) to validate our models.

Our hypothesis is that we can predict cancer types using random forest on sequencing information. Using this classification model to predict the cancer type of CUP may offer valuable and life-saving insights into decision and treatment plans for clinicians in clinics.

2 Related Work and Methods

- **Unsupervised Learning (clustering)**
  - **Frequent Itemset Mining**
    - Method: Our primary method of unsupervised learning so as to establish the existence of classes of clusters in the data.
    - How it is related: Identify co-currency of likely subsets of the attributes before we get into the details. If time permits, we want to add some of the findings in the mutation signature to test our result.
- **Supervised Learning**
  - **Naive Bayesian Network**
    - Method: Convert Naive Bayes method to a directed graph.
    - How it is related: readily handle incomplete features.
  - **Decision Tree**
    - Method: C4.5, partitioning on continuous variables.
    - How it is related: Using decision tree on biological significant information can allow us to create a clinically relevant decision tree for cancer type prediction.
- **Ensemble Methods**

- **Adaboost**
  - Method: weighted vote with a collection of classifiers. If a tuple is mis-classified, its weight is increased.
  - How it is related: More robust model for decision tree training.
- **Random Forest (Bagging + Decision Tree)**
  - Method: Used to correct overfitting habit of decision tree.
  - How it is related: OUR MAIN METHOD! More robust model for decision tree training.

3 Proposed Work

Data Cleaning ⇒ Decision tree implementation using training and test set ⇒ adaboost and random forest implementation using training and test set

4 Evaluation

- Dataset: all datasets are labeled for cancer types (except for CUP). Model is trained on predicting the right cancer type using random forest. The model is then tested by using sequencing cohort from other projects.
  - Training set: we will be using MSK-IMPACT Clinical Sequencing Cohort, this dataset contains profiles of 10,946 patients. For each patient, cancer type, related body details (e.g. mutated genes name, gene location, mutation type, etc.) and other personal information (e.g. smoker or not).
  - Test set: we will be using TCGA dataset available on the database. Every TCGA dataset containing Whole Genome Sequencing (WGS) information, which includes more information than MSK-IMPACT. However, filters has been applied for MSK-IMPACT included oncogenes. This allows for readily available training sets.
- Method of Evaluation: Evaluation metric can be set as prediction accuracy, precision, recall, F-measure, .632 bootstrapping (since we have limited samples on the patients), holdout method, random subsampling, cross-validation.

5 Plan of Action

- Week 1 (3/25 - 3/31) Modify and read online dataset to our demand
  - *Donald*: Implement method to integrate desired dataset (e.g. Connect patient in data to case lists which contains cancer type information);
  - *Peter*: Investigate potential relationship between attributes and outcomes using biological knowledge, in preparation for Bayesian network;

  - *Yifeng*: Implement method to read dataset using metadata, prepare categorized data for decision tree;

- Week 2 (4/1 - 4/7) Implement decision tree / Bayesian network
  - *Donald and Peter*: Implement decision tree;
  - *Yifeng*: Implement Bayesian Network;
  - *All*: Evaluate outcome of decision tree and Bayesian network to choose the one with better outcome, using small-scale training and testing data. Then implement the chosen method, workload will be divided once method is determined;
- Week 3 (4/8 - 4/14) Implement methods to enhance robustness
  - *Donald*: Implement adaboost method;
  - *Peter and Yifeng*: Implement random forest method;
- Week 4 (4/15 - 4/21) (wk3 continued) **In class project presentation ready**
  - *Donald*: Prepare slides for presentation;
  - *Peter and Yifeng*: Prepare draft and bulletins;
  - *All*: Continue work on adaboost and random forest;
- Week 5 (4/22 - 4/28)
  - *All*: Continue improving Random Forest, Adaboost. Implement clustering techniques if time permits. Evaluate results;
- Week 6 (4/29 - 5/5) **Project report and deliverable ready**
  - *All*: Project report and deliverable preparation, final tweak on deliverable;
- Deadline (5/6 - 5/7) **Get everything done**
  - *All*: Package project and publish as executable, finalize report.