Proposal

Qingya Zhao, Runan Peng, Guirong Luo
Stevens Institute of Technology

1) Problem Statement

Cancer begins with uncontrolled division of one cell, which results in a visible mass called a tumor. Since the causes of breast cancer still remain unknown, early detection is the key to reduce the death rate. Compared to the conventional methods which rely on detecting the presence of particular signal features by a human observer, computer-aided diagnosis approaches for automated diagnostic systems have been developed in the past ten years to attempt to solve this problem. Obtained cell samples by fine-needle aspirates (FNA) from patients, using computer based analytical techniques to define nuclear size, shape, and texture features and then using these features to distinguish benign and malignant breast cytology has become a widely accepted method.

The importance of classifying cell samples into benign and malignant has led many research teams, from the biomedical and the bioinformatics field, to study the application of machine learning methods such as Bayes classifiers, artificial neural networks and supervised fuzzy clustering.

2) Description of data set

This breast cancer database originates from the University of Wisconsin Hospitals, Madison, from Dr. William H. Wolberg in 1991. The total number of instances is 699, with 10 attributes plus the class attribute. It includes: 1) sample code number 2) clump thickness 3) uniformity of cell size 4) uniformity of cell shape 5) marginal adhesion 6) single epithelial cell size 7) bare nuclei 8) bland chromatin 9) normal nucleoli 10) mitoses 11) class. Except for the id number, all of the variables are categorical, and for class, 2 denotes benign and 4 symbolizes malignant. Further, there are 16 missing values (2% of the total) marked by a question mark (?).

The research questions may include:

- *How accurate the prediction outcome will be in terms of precision and recall using the logistic model we developed?*
- *What are the distinguishing features that can be observed among the various patient groups?*
- *Which attributes will play a significant role in information gain, or, which factors affect the most in predicting the results when drawing a decision tree?*

3) Implementation Plan

The first step is to collect data and preprocess it, which includes identifying missing values and outliers. If the missing values are less than 5% of the total samples, we may consider dropping these rows. Otherwise, techniques such as interpolation may be adopted. The second phase focuses on the implementation of three algorithms. The logistic regression model will be significant in predicting new instances. And the clustering model will provide insight into the distinct characteristics that each group possesses. Decision tree uses a white box model. If a given situation is observable in a model, the explanation for the condition is easily explained by boolean logic. The third stage concentrates on the model's accuracy, including precision and recall. We may evaluate these algorithms to see what their advantages and disadvantages are in terms of accuracy and computational cost. Finally, there will be a section devoted to discussing the limitations and future research directions.

4) Team members & Task allocation

Our project will be divided into the following parts, with each team member carrying out the tasks that are allocated to them:

Team members	Task allocation
Qingya Zhao (WS)	Decision Tree
Guirong Luo (A)	Clustering
Runan Peng (A)	Logistic Regression
All	Comparison & Interpretation
All	Report Writing