Graph Analysis on Breast Cancer Data using Bayesian Networks Project Proposal

Kalea Sebesta*

University of Texas at San Antonio
k_sebesta@yahoo.com

May 17, 2018

Abstract

The purpose of this project is to create a Bayesian Network and perform graphical analysis on breast cancer data. The goal is to visualize and better understand the predictors and their relationship to a breast cancer diagnosis being benign or malignant. The graphical analysis approach will enable the identification of probabilities between predictors and their affects on the cancer outcome.

I. Introduction

Cancer is a terrifying diagnosis for any individual and along with fear, the diagnosis brings a number of uncertainties. The research to find answers for cancer patients is on the rise and special dedication to understanding the dynamics of the disease from a micro level are being conducted. For the purpose of this project, the scope will be narrowed to breast cancer and ten features that have been computed for each cell nucleus.

II. Data

The data for this project can be found at Kaggle or on UCI Machine Learning Repository: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data/data. The features in this dataset are originally computed from a digitized image of a fine needle aspirate (FNA) breast mass. This dataset contains the patient or image identification number, the diagnosis (malignant or benign), and ten specific features that have been calculated in three different ways (mean, standard error, and worst) thus resulting in 30 features. There are no missing values and the distribution of outcomes is 212 malignant and 357 benign. The ten cell nucleus features are:

- Radius
- Texture
- Perimeter
- Area
- Smoothness
- Compactness

*A special thank you to Dr. Han for the guidance and mentoring on this project

- Concavity
- Concave Points
- Symmetry
- Fractal Dimension

III. Methods

For this project R programming language will be leveraged to preform and display analytical and graphical results. Specifically the package bnlearn will be utilized to create the Bayesian Network.

IV. Challenges

1. Understanding and determining the correct distribution algorithm for the Bayesian Network (Gaussian, Min-Max Hill Climbing, etc.)

V. Anticipated Results & Timeline

Table 1: *Milestone Timeline*

Expectations		
Date	Milestone	Status Report Week
May 25	Preliminary Research on Past Work	1
June 1	Creating Custom Fitted Bayesian Network	2
June 8	Creating Custom Fitted Bayesian Network	3
June 15	Structured Learning Computing and Comparing	4
June 22	Structured Learning Computing and Comparing	5
June 29	Parameter Learning	6
July 6	Parameter Learning	7
July 13	Model Validation	8
July 20	Model Validation	9
July 27	Final Report	10

References

[Nagarajan, Scutari, and Lebre, 2013] Nagarajan, R., Scutari, M., and Lebre, S., (2013). Bayesian Networks in R with Applications in Systems Biology *Springer (US)*, Vol. 48.

[Scutari, M. and Denis, J. B., 2014] Scutari, M. and Denis, J. B., (2014). Bayesian Networks in R with Examples in R *Texts in Statistical Science, Chapman & Hall/CRC (US)*.