Mining Oncology Data: Knowledge Discovery in Clinical Performance of Cancer Patients

By John Hayward (Computer Science), Prof. Carolina Ruiz (CS), and Dr. Giles Whales (UMass Medical School)

Cancer research: “the intense scientific effort to understand the development of cancer and identify potential therapies” [Cancer Genome Project]

Database and data mining techniques have become vital tools in studying clinical oncology data

Our Major ML & Data Mining Objectives

- Patient Longevity
- Quality of Life/Performance Status
- Surgical Prospects
- Imaging and Diagnostic Test Accuracy
- Patient Demographics (breast cancer)

Our joint WPI-UMass Medical School Research on Gastrointestinal Cancer and Breast Cancer:

(1) Design and program cancer database modules and (2) Discover patterns in clinical data using data mining and machine learning techniques

Study health patterns in along population divisions: Tumor Resection/No Resection, Chemo/Radiotherapy, Tumor Progression, Redeveloped Symptoms

Initial Data Analysis Results: Contributions of Endoscopic Ultrasound and CT Scan to Tumor Staging and Preoperative Assessment of Pancreatic Tumor Resectability

Further Data Analysis Using Machine Learning (ML) & Data Mining Techniques

ML algorithms have demonstrated higher accuracy rates over major oncology diagnostic tools like Endoscopic Ultrasound (EUS)

Data Mining & ML Analysis of Quality of Life Performance using Decision Trees

Data Mining & ML Analysis of Age vs. Longevity using Regression Trees

Data Set Algorithm Percent Correct

Conclu J48 84.35%
All NaiveBayes 86.25%
Conclu.CFS Bagging.NaiveBayes 92.00%
All.CFS Bagging.NaiveBayes 92.00%
ZeroR (Benchmark) 79.00%
OneR (Benchmark) 70.25%

Naive Bayes and \(t\) tests for patient survival

Regression Trees, Model Trees, Decision Trees

Comparative Accuracy of Assessing Tumor Malignancy

Comparative Accuracy of T-Staging Pancreatic Tumors

More concern on accuracy after Whipple

Database Captures Heterogeneous Structure of Patient Narrative

Study health patterns in along population divisions: Tumor Resection/No Resection, Chemo/Radiotherapy, Tumor Progression, Redeveloped Symptoms

Initial Data Analysis Results: Contributions of Endoscopic Ultrasound and CT Scan to Tumor Staging and Preoperative Assessment of Pancreatic Tumor Resectability

Further Data Analysis Using Machine Learning (ML) & Data Mining Techniques

ML algorithms have demonstrated higher accuracy rates over major oncology diagnostic tools like Endoscopic Ultrasound (EUS)

Data Mining & ML Analysis of Quality of Life Performance using Decision Trees

Data Mining & ML Analysis of Age vs. Longevity using Regression Trees

Data Set Algorithm Percent Correct

Conclu J48 84.35%
All NaiveBayes 86.25%
Conclu.CFS Bagging.NaiveBayes 92.00%
All.CFS Bagging.NaiveBayes 92.00%
ZeroR (Benchmark) 79.00%
OneR (Benchmark) 70.25%

Naive Bayes and \(t\) tests for patient survival

Regression Trees, Model Trees, Decision Trees

Comparative Accuracy of Assessing Tumor Malignancy

Comparative Accuracy of T-Staging Pancreatic Tumors

More concern on accuracy after Whipple

Database Captures Heterogeneous Structure of Patient Narrative