Data Project 1

Breast Cancer Analysis

Fall 2018 CIS 331

Grand Valley State University

Allison Bolen

2. **Abstract**:

To explore the recurrence of breast cancer given certain traits.

3. Project: Breast Cancer Recurrence Analysis

3.1 Problem description.

I would like to find what affects the recurrence of breast cancer. What traits are the best predictors and how does one prevent recurrence. A short description for the attributes that might not be self explanatory are as follows. Inv-nodes: the number (range 0 - 39) of axillary lymph nodes that contain metastatic breast cancer visible on histological examination. When cancer spreads to the lymph nodes it becomes more dangerous because lymph nodes filter and out harmful substances throughout the body, if cancer spreads to a lymph node that cancer can now spread to anywhere in the body. I would expect that this will be a huge factor in the recurrence of cancer. The attribute 'degMalig' is the degree of malignancy. A tumor can be benign or malignant, a malignant tumor is cancer. The more malignant a tumor is the more dangerous it is. Malignancy is typically measure on a scale of 1 to 4, ie: stage 1 cancer. The attribute irradiate refers to the presence of cells exposed to radiation.

3.2 Review of publications. Provide brief discussion on approaches used in these research papers. CITE YOUR SOURCES!

3.3 Software tools used for predictive data mining: Description of a tool or tools that will be used (if comparative analysis was made). Give a short explanation of applied algorithms. Make emphasis on the parameters you will tune during the mining process.

3.4 Preprocessing of data

When I used R to analyze the initial data set there were some reading issues on the column names so I renamed them to more readable values. I also decided to do some minor alteration for the data. I found that of the almost 300 instances only 9 instances were missing some attributes. Since only 9 of almost 300 were affected I decided to just remove them from the set instead of replacing those values that were missing. Removing those 9 instances would affect the overall quality of the data set.

I also decided to regroup the ‘invNodes’ attribute by less than 2 and more than 2 because most of the data came in less than or equal to 2. So we would lose no information by reducing the levels of that attribute to 2 as long as we remember that we reduced it down. Though instead of overwriting the ‘invNodes’ column I just saved the resulting data to a new attribute column called ‘invNodesModif’.

After cleaning and reorganizing the data I used Weka to decide what features to reduce my data set down to. Using the CfsSubset Algorithm and the Best First search I found that the most important features were ageGroup, invNodesModified, and degMalig.

3.5 Describe all steps of a data mining process, and all activities with selected data mining tools. If possible, present and interpret some intermediate results. Describe and discuss a selection of system’s parameters for data mining including selection of training and testing data set for all applied techniques. Describe the criteria for ending iterations in a data mining process.

3.6 Results (types of output reports and contents, interpretation of each output value): Presentation and interpretation. Give the detailed discussion of obtained results. Perform a comparative analysis of different methodologies. Conclusions and future work: problems, constraints

Sources:

https://www.nejm.org/doi/full/10.1056/NEJMoa041588

https://ww5.komen.org/BreastCancer/TumorSizeandStaging.html