Preprocessing Experimentation for Breast Cancer Classification

Timothy Stockton, Brandon Peddle, Angelica Gaulin, Emma Wiechert

Keene State College, timothy.stockton, brandon.peddle, angelica.gaulin, emma.wiechert@keene.edu

***Abstract* - Breast cancer is the 2nd most diagnosed cancer. It will affect about 10% of women at some point in their lives. The incidence rate is on the rise, and according to the data, the survival rate 10 years after diagnosis is 80%. When a cancerous mass forms, it is important to know whether it is malignant or benign as the former can spread throughout the body and cause further harm. Early detection saves lives, as it allows for more treatment options. However, 50% of breast cancers were not detected when the women screened had very dense breast tissue. It is also difficult to observe lesions less than 2 mm with mammograms. Computer aided diagnosis can improve the prediction accuracy and has been adopted for the classification of breast cancer. There are several different datasets, machine learning models, pre-processing techniques, and more available today, with each combination yielding different results. Our research is focused on the use of different feature selection and extraction techniques. The methods we have chosen to use are correlation based feature selection, recursive feature elimination, linear discriminant analysis, principal component analysis, and combination approaches of these algorithms. We hypothesize that at least one of these combination algorithms will outperform other feature selection/extraction methods and/or combinations.**

*Index Terms* - Correlation based Feature Selection (CFS), linear Discriminant Analysis (LDA), Principal Component Analysis (PCA), Recursive Feature Elimination (RFE)

Introduction

Here we will give a detailed overview of what breast cancer is and how it affects women worldwide.

We will also discuss in detail what machine learning is and generally how it is being applied to breast cancer diagnosis and classification.

Related Research

Here we will succinctly summarize the multiple papers that we have read in preperation for our own research. This research should be discussed with a focus on preprocessing techniques and methodology because that is our focus.

One of the case studies that was read in preperation for our research was [] this focused on breast cancer detection using the following machine learning models;

Method

I. Data Set

The data set we used is the Wisconsin Diagnostic Breast Cancer data set. It contains xxxxxxxxx features including: xxxxxx xxxxxxx xxxxxx xxxxx xxxx xxxxxx xxxxxx xxxxxxxxx xxxxxxxx xxxxxxxxxxx xxxxxxx xxxxxxxx xxxxxxxx xxxxxxxx xxxxxxxxxx xxxxxx xxxxxxxxx xxxxxxxxxxx xxxxxxxxxx. These features are derived from mammograms such that xxxxxxx xxxxxxxxxxxx xxxxxxxxx xxxxxxx xxxxxxxxx xxxxxx xxxxxxxxx xxxxxxxxx xxxxxxx xxxxxxxx xxxxxxxxxxxx xxxxxxxxxxx xxxxxxx xxxxxxx xxxxxx xxxxxx xxxxxxxxxx xxxxxxx xxxxxxx xxxxxxxxx xxxxx xxxxxxxxx xxxxxxxxx xxxxx xxxxxx xxxxxxxxxx.

II. Cleaning

The data set must be “cleaned” in order to account for things such as: missing data, incomplete data, erroneous values, and duplicates. The way we did this was xxxx xxx xx x x xxxx xxx xxxxxx xxxxxx xxx xxxx xxxxx xx xx xxxxxx xxx xxxxxxxxxxxxx xxxxxx xxxxxxxxx xxxx xxxxxxxx xxx xxxxxx xxxxxxx xxxxx xxxx xx xxxxxx.

This cleaning is performed on the original data set, and is performed prior to making dublicates so that each data set recieves the same cleaning.

II. Univariate Analysis

We created plots and graphs looking at xxxxxx which showed us xxxxxxxxxxxxx.

III. Bivariate Analysis

We created plots and graphs looking at xxxxxx which showed us xxxxxxxxxxxxx.

IV. Feature Selection

In feature selection, only features with correlations useful for classification are kept while the others are dropped. This is determined by the technique used. Our research utilizes two different feature selection techniques which are correlation based feature selection (CFS) and recursive feature elimination (CFE). We created nine duplicates of the cleaned data set. Three duplicates have CFS applied, while another three duplicates have CFE applied.

Specifically, what CFS does is xxxxx xxxxxx xx xxxxxx xxxx xxxxx xxx xx xxxxxx xxxx xxx xxxx xxxxxx xxxxx xxxxxxxxxxxx xx xxxxxx xx x xxxxxx xxx xxx xxxxx xxxxxxx xx x xx xxxxxx.

What CFE does is xxx xx xxxxxx xxxx xxxxx xxxxxx xx xxxxxx xxxxxxx xxxxxxxx xxxxxxxx xxxxxxxx xxxxxx xxxxxxx xxxxxx xxx x xxxxx xxxx.

V. Feature Extraction

Feature extraction involves the combination of features and reduction of dimenson. The techniques we use are principle component analysis (PCA) and linear discriminant analysis (LDA). First we apply PCA to one of the data set duplicates that did not have a feature selection technique applied, and then we do the same for LDA. Next, we apply PCA to a duplicate that had CFS applied and then do the same for LDA on another CFS duplicate. Finally, we use PCA on a duplicate that has RFE applied and then use LDA on another RFE duplicate.

Specifically, what PCA does is xxxxx xxxxxx xx xxxxxx xxxx xxxxx xxx xx xxxxxx xxxx xxx xxxx xxxxxx xxxxx xxxxxxxxxxxx xx xxxxxx xx x xxxxxx xxx xxx xxxxx xxxxxxx xx x xx xxxxxx.

What LDA does is xxx xx xxxxxx xxxx xxxxx xxxxxx xx xxxxxx xxxxxxx xxxxxxxx xxxxxxxx xxxxxxxx xxxxxx xxxxxxx xxxxxx xxx x xxxxx xxxx.

VI. Further Cleaning

At this stage, we have nine different data sets derived from the cleaned original. Specifically we have one with only CFS applied, one with only RFE applied, one with only LDA applied, and one with only PCA applied. We also have combination data sets including: one with CFS and LDA applied, one with CFS and PCA applied, one with RFE and LDA applied, and one with RFE and PCA applied. Finally we have a dataset that had no feature selection or feature extraction techniques used which acts as a control data set.

All nine of these datasets must undergo further cleaning and processing before they can be used to train our models. This further cleaning includes normalizing or scaling via xxxxxxxxxxxxxxx, encoding via LabelEncoder. Outliers were not removed because xxx xxxxxx xxx xxxxxx xxxxx xxx xxxxx xxxxx xxxxxx xxxx xxxx xxx xxx.

Normalization is used because otherwise one feature may hold more weight in the classification process due to large or small numbers. It puts all features on the same scale between 0 and 1 so that although a number may be large or small, it is processed relative to it’s position within that feature only.

Encoding is used because the labels in the data set stating whether the patients cancer was malignant or benign are strings. The machine learning models can not utilize strings as input, so the LabelEncoder changes those strings into integers. In this case M becomes xxxxxxx and B becomes xxxxxxx.

VII. Train-Test Split

Each of our nine data sets are now split into an X variable which contains all features other than the labels and a Y variable which contains only the labels. These variables are used to build the training and testing sets. We used a 70-30 split which means the training set contains 70% of the records and the testing set contains 30% of the records.

VIII. Logistic Regression

Our first ML model used is linear regression which xxxxx xxxx xxx xx xxxxx xxx xx xx xxxx xxxxxx xxxx xxx xx xxxxx xx xxx xx xx x xxx.

We created nine seperate linear regression models, each using a different one of our data sets to train it and test it.

IX. Decision Tree

Our second ML model used is decision tree which xxxxx xxxx xxx xx xxxxx xxx xx xx xxxx xxxxxx xxxx xxx xx xxxxx xx xxx xx xx x xxx.

We created nine seperate decision tree models, each using a different one of our data sets to train it and test it.

X. Random Forest

Our third ML model used is random forest which is comprised of multiple decision trees and xxxxx xxxx xxx xx xxxxx xxx xx xx xxxx xxxxxx xxxx xxx xx xxxxx xx xxx xx xx x xxx.

We created nine seperate random forest models, each using a different one of our data sets to train it and test it.

XI. SVM

Our fourth ML model used is SVM which xxxxx xxxx xxx xx xxxxx xxx xx xx xxxx xxxxxx xxxx xxx xx xxxxx xx xxx xx xx x xxx.

We created nine seperate SVM models, each using a different one of our data sets to train it and test it.

XII. Neural Network

Our fifth ML model used is a neural network which xxxxx xxxx xxx xx xxxxx xxx xx xx xxxx xxxxxx xxxx xxx xx xxxxx xx xxx xx xx x xxx.

We created nine seperate neural network models, each using a different one of our data sets to train it and test it.

XIII. XGBoost Tree

Our sixth and final ML model used is XGBoost tree which xxxxx xxxx xxx xx xxxxx xxx xx xx xxxx xxxxxx xxxx xxx xx xxxxx xx xxx xx xx x xxx.

We created nine seperate XGBoost tree models, each using a different one of our data sets to train it and test it.

XIV. Method Summary

We created a loop in which each data set is used to train and test each ML model. This results in 54 different models in total. Refer to our flowchart.

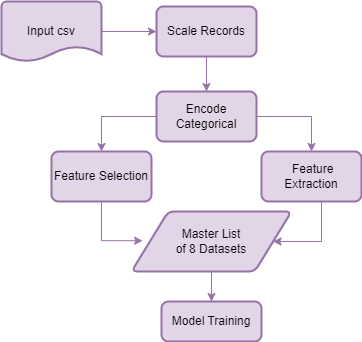


Figure I

Methodology Flowchart

Discussion

Here we will discuss and analyze our findings in detail including the accuracy, precision, etc. Of each ML model using the different preprocessing techniques.

Conclusion

This is where we will discuss which method or combination of methods we believe is best based on the reslts. xxxxxxxxx xxxx xxxxxx xxxxxx xxxx xx xx xxxxxx xxxx xxxxx xxxxxx xxxxxxx xxxxxxxxxx xxxxx xxx xxxxxx xxxx xxxxxxx xxxxxxxx xxxxxx xxxxxx xxx xxxx.

We will also state steps we believe can and/or should be taken in futre research on this topic. xxxxxx xxxxx xxxx xxxxx xx xx x xxxxxx xxxx xxxx xxxxx xxxxxxxxx xxxx xxxx xxx xx x xxxxxx xxxxxx xxxxxx xxxxx xxxx xxxx xxx xx xxxxxxxxx xxxxx.

Acknowledgment

The preferred spelling of the word “acknowledgment” in American English is without an “e” after the “g.” Use the singular heading even if you have many acknowledgments. Please put the sponsor acknowledgments in this section. Do not use a footnote on the first page.

References

Place references in a separate References section at the end of the paper. Number the references sequentially by order of appearance, not alphabetically. List up to three authors’ names in a reference; replace the others by “*et al*.”

* **Reference text**: 8 point, Times New Roman, full justified, hanging .25”, no space between the references, tab between right bracket and the start of the reference

[1] “Today in Science History: Engineering Quotes.” 2012.

todayinsci.com/QuotationsCategories/E\_Cat/Engineering-Quotations.htm. Web. Accessed: April 9, 2012.

[2] Donohue, Susan K. and Richards, Larry G. October 2011. “P-12 Engineering Education: Using Engineering Teaching Kits to Address Student Misconceptions in Science.” *Proceedings* of the 41st Frontiers in Education Conference, Rapid City, SD, pp. F2A-1 – F2A-3.

[3] Dweck, Carol S. 2006. *Mindset: The New Psychology of Success*, New York: Random House, Inc.

[4] Kaplan, Avi and Maehr, Martin L. June 2007. “The Contributions and Prospects of Goal Orientation Theory.” *Educational Psychology Review* 19(2), pp. 141 – 184.

[5] Dweck, Carol S. “Messages That Motivate: How Praise Molds Students’ Beliefs, Motivation, and Performance (In Surprising Ways).” In Aronson, Joshua (ed.), 2006, *Improving Academic Achievement: Impact of Psychological Factors on Education*. New York: Elsevier Science, pp. 37 – 60.

Author Information

An example: **Jane Smith,** Professor, Department of Electrical and Computer Engineering, Big State University.