Distinguish Malignant From Benign Breast Cancer

Xin Liu, Jingzhu Zhou and Mengshu Shao Department of Biostatistics & Bioinformatics

Introduction and Objectives

Objective:The goal of this work is to use machine learning methodology to develop digitalized method to minimize subjectivity of the visual diagnosis and increase accuracy of classification for each breast cancer patient.
Introduction: Breast cancer stage can be diagnosed by visually examining fine needle aspirate cancer cell images. The accuracy of the diagnosis is over 90%. However, there exists large standard error for the mean sensitivity and the mean specificity of the diagnosis. This indicates the accuracy varies greatly within individual series and visual diagnosis involves a great deal of subjectivity.

Materials & Methods

The following classification methods were considered in this research:

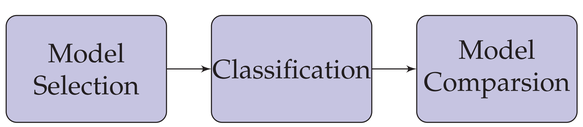
- Logistic regression
- Penalized logistic regression (LASSO plus Ridge)
- Linear / Qudratic Discriminant Analysis
- Support Vector Machine

The dataset contains records of 569 breast cancer instances, in which:

- 212 malignant
- 357 benign

Thirty variables are computed for each cell nucleus for breast cancer diagnosis prediction including mean(_m), sd(_sd), mean of largest three(_w) for the following features:

1) radius
2) texture
3) perimeter
4) area
5) smoothness
6) compactness
7) concavity
8) concave points
9) symmetry
10) fractal dimension



Results 1

Model Selection for LDA,QDA,SVM:

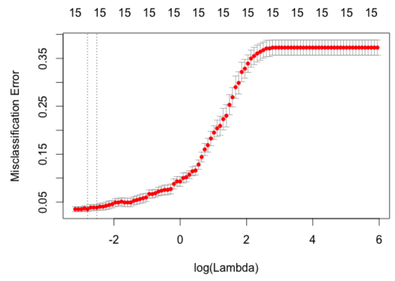
- 1.Pairwise correlations of the 30 variables were used to delete the variables that are highly correlated. Variable deleted: radius_m,perimeter_m,area_w,perimeter_w, radius_w,area_sd,perimeter_sd
- 2.Stepwise selection(logistic regression) were used to select the best subset of variables based on lowest AIC. Variable selected: concave_points_w, area_m, texture_w,radius_sd,compactness_sd, smoothness_w,concavity_w ,texture_sd

Model Selection for Penalized Regression:

- Lasso uses the constraint of the overall magnitude of the coefficients, thus important predictors are included in the model, and less important predictors shrink, potentially to zero. 10-folds cross validation selects the tuning parameter which makes the smallest cross validation error.
- The variables selected are texture_m, concavity_m, concave_points_m, fractal_m, radius_sd, texture_sd, smoothness_sd, compactness_sd, fractal_sd, radius_w, texture_w, smoothness_w, concavity_w, concave_points_w and symmetry_w.

Results 2

1. Ridge parameter selection



The best tuning parameter:0.00304. Then Ridge regression was used to fit the model based on the tuning parameter.

$$logit(P(Y_i = 1|X = x)) = -9.35 + 0.073texture_m + 4.74concavity_m + 14.72concave_points_m - 39.34fractal_m + 1.81radius_sd - 0.13texture_sd - 8.03smoothness_sd - 2.52compactness_sd - 56.07fractal_sd + 0.15radius_w + 0.06texture_w + 12.44smoothness_w + 1.58concavity_w + 8.59concave_points_w + 4.41symmetry_w$$

2.Varying SVM input parameters

In another set of experiments we studied the dependence of the training error found by the SVM algorithm on the parameter C. Table 2 shows the variation of the training error corresponds with the parameter C. It can be seen that the smallest error is obtained for C=0.9.

Cost	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
Accuracy	93.85	95.08	96.31	96.49	96.84	97.19	97.19	97.54	97.72	97.36

Table 1: Parameter selection for SVM

Conclusion

Model	**Logistic**	**Ridge**	**LDA**	**QDA**	**SVM**
Accuracy	96.14	96.49	95.96	94.91	97.72
Predictor numbers	8	15	8	8	24

Table 2: Model Comparison

- All models give an accuracy rate above 95 %. Support vector machines gave the best accuracy, however it used the highest number of parameters.
- We would recommend logistic regression since it gave comparably high accuracy with the lowest number of parameters.
- Future Research:The various fitted models can be tested for accuracy, sensitivity and specificity on testing data (future patients prognosis) based on those cell measurements.Other classification methods such as extended bases LDA, QDA, plus interaction terms could be considered.

References

[1] William H. Wolberg, W. Nick Street, and Olvi L. Mangasarian. Computerized diagnosis of breast fine-needle aspirates. *The Breast Journal*, 3(2):77–80, 1997.

[2] William H Wolberg, W.Nick Street, Dennis M Heisey, and Olvi L Mangasarian. Computer-derived nuclear features distinguish malignant from benign breast cytology. *Human Pathology*, 26(7):792 – 796, 1995.

[3] William H Wolberg, W Nick Street, and Olvi L Mangasarian. Image analysis and machine learning applied to breast cancer diagnosis and prognosis. *Analytical and Quantitative cytology and histology*, 17(2):77–87, 1995.

Contact Information

Email xin.liu1@duke.edu
mengshu.shao@duke.edu
jingzhu.zhou@duke.edu