Project Proposal

Qinjing Zou 1701213176

1. Introduction

The diagnosis of breast tumors has traditionally been performed by a full biopsy, an invasive surgical procedure. Fine needle aspirations (FNAs) provide a way to examine a small amount of tissue from the tumor. By carefully examining both the characteristics of individual cells and important contextual features such as the size of cell clumps, physicians at some specialized institutions have been able to diagnose successfully using FNAs. However, many different features are thought to be correlated with malignancy, and the process remains highly subjective, depending upon the skill and experience of the physician. In order to increase the speed, correctness, and objectivity of the diagnosis process, we have used image processing and machine learning techniques.

2. Goal

So the goal of this essay is to first train a model using image data from multiple patients. And then try to classify new patient's data to figure out whether her breast cancer is benign or malignant.

3. Process

a) Image Preparation

The imaged area on the aspirate slides has been visually selected for minimal nuclear overlap. Once the nuclei to be analyzed have been identified by the operator, a mouse was used to trace a rough outline of cell nuclei on the computer monitor. From this rough outline, the actual boundary of the cell nucleus was located by an active contour model known as a "snake". The computer calculates ten nuclear features for each nucleus.

b) Data preprocessing

Split the data into in-sample and out-of-sample data.
Feature scaling: Normalization and standardization.
Dimensionally reduction: feature selection or feature extraction.

c) Diagnostic classification procedure using ML

DecisionTree/KNN/logistic/SVM/RandomForest

d) Feature adjustment

K-fold cross valuation.
Learning curve and validation curve.

4. Data

a) Origins:

Breast Cancer Wisconsin (Diagnostic) Data Set
http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

b) Description

This breast cancer databases was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg.

Data Set Characteristics:	Multivariate	**Number of Instances:**	569	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	32	**Date Donated**	1995-11-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	677561

c) Attribute information

1) ID number

2) Diagnosis (M = malignant, B = benign)

3-32) Ten real-valued features are computed for each cell nucleus:

- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness (perimeter^2 / area - 1.0)
- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- symmetry
- fractal dimension ("coastline approximation" - 1)

The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

d) Class distribution

Benign: 357 (62.7%)

Malignant: 212 (37.3%)

e) Sample (part)

	id	diagnosis	radius mean	texture mean	perimeter mean	area mean	smoothness mean	compactness mean	concavity mean	concave points mean	...	radius worst	texture worst	perimeter worst	area worst	smoothness worst
42	855625	M	19.07	24.81	128.30	1104.0	0.09081	0.21900	0.21070	0.09961	...	24.09	33.17	177.40	1651.0	0.1247
540	921385	B	11.54	14.44	74.65	402.9	0.09984	0.11200	0.06737	0.02594	...	12.26	19.68	78.78	457.8	0.1345
509	915460	M	15.46	23.95	103.80	731.3	0.11830	0.18700	0.20300	0.08520	...	17.11	36.33	117.70	909.4	0.1732
138	868826	M	14.95	17.57	96.85	678.1	0.11670	0.13050	0.15390	0.08624	...	18.55	21.43	121.40	971.4	0.1411
311	89382601	B	14.61	15.69	92.68	664.9	0.07618	0.03515	0.01447	0.01877	...	16.46	21.75	103.70	840.8	0.1011

5 rows × 32 columns

5. Desirable Output

diagnosis(benign or malignant)

6. References

a) Mangasarian O L, Street W N, Wolberg W H. Breast Cancer Diagnosis and Prognosis Via Linear Programming[J]. Computational Science & Engineering IEEE, 1995, 2(3):70.

b) Wolberg W H, Street W N, Heisey D M, et al. Computerized breast cancer diagnosis and prognosis from fine-needle aspirates.[J]. Archives of Surgery, 1995, 130(5):511.

c) Wolberg W H, Street W N, Mangasarian O L. Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates.[J]. Cancer Letters, 1994, 77(2-3):163.

d) Goldgof D B. Nuclear feature extraction for breast tumor diagnosis[J]. Proc Spie, 1993, 1993:861-870.