Breast Cancer Recurrence Prediction

Chang Liu
College of Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
cliu5@andrew.cmu.edu

Kangning Chen
Tepper Business School
Carnegie Mellon University
Pittsburgh, PA 15213
kangninc@tepper.cmu.edu

Abstract

Breast cancer is considered to be the second leading cause of cancer deaths in women today. Nowadays, as the development on machine technique, the data mining skills are widely used in medical industry to effectively predict and diagnosis diseases, including cancer-curing. In this essay, we will use semi-supervised machine learning to study the probability and time gaps for the potential recurrence.

1 Introduction

Breast cancer is the most common cancer among women excluding non-melanoma skin cancers. Occasionally, breast cancer can return after primary treatment. Consequently, the main problem under these circumstances is to predict such a recurrent event, because only expertise from experience is not enough.

2 Dataset and Preprocessing

We use two datasets for this project. The first Wisconsin Prognostic Breast Cancer data set contains 198 instances labeled for recurrence of breast cancer. Along with the label for recurrence, this data set records the time before recurrence or the time when a patient remains disease free. Along with the recurrence and time, this data set contains 10 relevant attributes describing the characteristics of the cell nuclei present in the digitized image of a fine needle aspirate (FNA) of the breast mass. The second Wisconsin Prognostic Breast Cancer dataset contains 569 instances with the ID number and the diagnosis outcome (malign or benign) and the same 10 relevant attributes as the prognostic dataset. Each attribute has three values: The mean, standard error, and "worst" or largest. In addition, the prognostic dataset has more attributes: Time, Tumour size and Lymph node status. The detailed information is displayed in Table 1.

The data set relatively clean and complete. However, we need to do two steps before implement our data. The first step is to fill the missing data. There are a few unobservable values in the attribute Lymph node status of Prognostic DataSet. We simply use the formula $\mu + \sigma * Z$ to generate these field, where Z is a standard normal distribution. The reason of adding this random penalty is that we don't want to significantly reduce the variance. The second step is to merge Diagnostic and Prognostic datasets according to patient ID. We use Excel Function VLOOKUP here.

Table 1: The main characteristics of the Wisconsin breast cancer dataset

Attribute	Range
*Time (recurrence time if field 2 = R, disease-free time if field 2 = N) 4-33)	(1,125)
Radius (mean of distances from the centre to all points on the perimeter)	(10.95, 27.22)
Texture (standard deviation of gray-scale values)	(10.38, 39.28)
Perimeter	(71.90,182.10)
Area	(361.60, 2250)
Smoothness (local variation in radius lengths)	(0.075, 0.145)
Compactness (perimeter^2/area $- 1.0$)	(0.046, 0.311)
Concavity (severity of concave portions of the con- tour)	(0.024, 0.427)
Concave points (number of concave portions of the contour)	(0.020, 0.201)
Symmetry	(0.131, 0.304)
Fractal dimension ("coastline approximation" - 1)	(0.050, 0.097)
*Tumour size - diameter of the excised tumour in centimetres	(0.400, 10.00)
*Lymph node status - number of positive auxiliary lymph nodes	(0, 27)

The attributes with * are the attributes that only Prognostic Dataset has.

3 Methods

3.1 Baseline

In *Diana's* paper, it uses Wisconsin Prognostic Breast Cancer dataset as the training data and apply Naive Bayesian classification on this fully labeled dataset. This methodology can serve as the baseline of our work with a testing accuracy of 74.24%, but we may experiment on prospect of improving the accuracy by including unlabeled instances. More importantly, sensitivity and specificity should be concerned as other baseline accuracy besides the testing accuracy (27.78% and 91.67% in *Diana's* paper). Also, we will try to implement the semi-supervised classifier by merging a supervised classifier (i.e. SVM) and an unsupervised classifier(i.e. K-means clustering) introduced in *D.Rajakumari's* paper.

3.2 First Approach: Supervised Learning

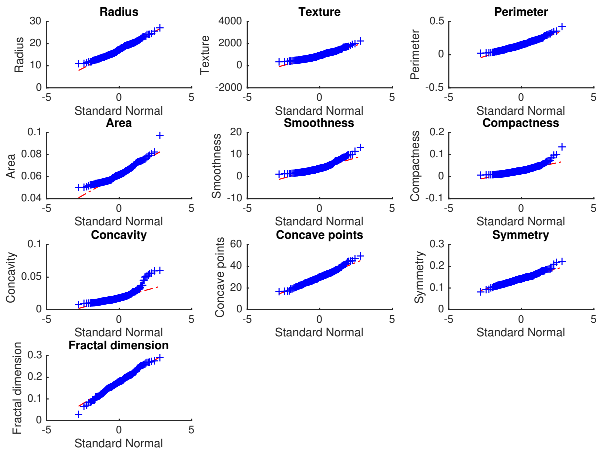
3.2.1 Gaussian Naive Bayes

Gaussian Naive Bayes is the methodology introduced in the *Diana's* paper. The essence of this methodology is to model the generative distribution of (X, Y) based on the training set, and predict the outcome by choosing the highest possible label. Despite its name and its naive intuition, it is still one of the most effective and efficient algorithms for data mining.

Naive Bayesian Classifier learns the conditional probability of each attributes X_k given the class label Y. After that, by applying Bayes Rule, the conditional probability of Y given the attributes can be computed and then we can predict $\hat{Y}$ by choosing the maximum.

Notice that the data type of each attribute is the numerical number. The following qqplots suggest that their distribution can be well fit by Gaussian Distribution. Hence we can assume the probability density function is :

$$\mathbb{P}(X_i|Y = y_k) = \frac{1}{\sqrt{2\pi}\sigma_{ik}} e^{-\frac{(X_i-\mu_{ik})^2}{2\sigma_{ik}^2}}$$



Gaussian Naive Bayes algorithm
1. Estimate μ_{ik} = mean of X_i given $Y = Y_k$, and σ_{ik}^2 = variance of X_i given $Y = Y_k$. Also, from training set, we can estimate $\mathbb{P}(Y = y_k) = \frac{\text{number of records with } Y = y_k}{\text{sample size}}$
2. By applying bayesian rule, we want to compute:

$$\mathbb{P}(Y = y_k | X = x) = \frac{\mathbb{P}(x|y_k)\mathbb{P}(y_k)}{\mathbb{P}(x)}$$

Since $\mathbb{P}(x)$ is the same for all y_k, we have:

$$\begin{aligned}\hat{Y} &= argmax_y \frac{\mathbb{P}(x|y)\mathbb{P}(y)}{\mathbb{P}(x)} \\ &= argmax_y \mathbb{P}(x|y)\mathbb{P}(y)\end{aligned}$$

3. Assume that the attributes are independent Gaussian distributed. Then:

$$\mathbb{P}(x|y) = \prod_i \mathbb{P}(X_i = x_i|y) = \prod_i \frac{1}{\sqrt{2\pi}\sigma_{ik}} e^{-\frac{(X_i - \mu_{ik})^2}{2\sigma_{ik}^2}}$$

4. Plug the results in step 3 back to the formula in step 2. We can compute our prediction $\hat{Y}$

3.2.2 Improvements

We can improve the Naive Bayesian by a couple of ways. Firstly, in the methodology introduced in the *Diana's* paper. The Naive Bayesian assumed independence of the attributes. However, this assumption is obviously wrong. For example, Compactness = perimeter^2/area $- 1.0$, which means the two attributes are highly correlated by other two fields. Also, in the training set, $corr(\text{Radius}, \text{Texture}) = 0.9929$ which is highly correlated. We can resolve this issue by applying Multivariate Gaussian Distribution. Specifically, in step 3 of *Naive Bayes algorithm*, the formula becomes:

$$\mathbb{P}(x|y) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{(X-\mu)^T\Sigma^{-1}(X-\mu)}{2}}$$

where Σ is the covariance matrix trained by training set.

Another improvement for Naive Bayes is choosing proper feature rather than simply include all features in the dataset. In the *Diana's* paper, all features related to the average are brutally included and all other features are brutally expelled. Greedy backward elimination can be applied here to eliminate redundant features.

Greedy backward elimination
1. Add all features to train the classifier.
2. Iterate each feature and train the classifier without this feature. Remove the feature with the highest test error.
3. Repeat step 2 until the optimal feature subset is founded.

3.2.3 Other trials: SVM, Perception and KNN

We have tried soft margin SVM and Perception with kernel tricks. However, the results are extremely bad. We need to investigate more after this midterm report. We will also implement KNN method after the midway report.

3.3 Enhancement: Semi-supervised Learning

In real world applications, labeled data are relatively hard to get while unlabeled data are cheap. The target label y requires human annotation which takes a long time, and these experiments require a lot of resources including experienced experts and special devices. Therefore, a trend to utilize the surplus unlabeled data together with scarcely labeled data is desirable. In this project, we will incorporate unlabeled data set (Diagnostic Dataset) to improve the classifier.

Semi-supervised learningis a class ofsupervised learningtasks and techniques that also make use of unlabeled data for training - typically a small amount oflabeled datawith a large amount of unlabeled data. The learner has both labeled training data $(x_i, y_i)_{i=1}^{l}$ and unlabeled training data $(x_i)_{i=l+1}^{l+u}$, and learns a predictor $f : X \to Y, f \in \mathbb{F}$ Where $\mathbb{F}$ is the hypothesis space. The predictor learned by semi-supervised methods usually predicts future test data better than that learned by supervised learning which only considers the labeled training data.

4 Evaluation and Preliminary Results

We checked the training errors and testing errors respectively, and calculated the expected accuracy. In addition, we looked at confusion matrix and account for sensitivity and specificity as well. Compare the result between Nave Bayes classifier and Semi-supervised learning one. The sensitivity and specificity are defined by:

$$sensitivity = \frac{\text{number of True Positives}}{\text{number of True Positives} + \text{number of F alse N egatives}}$$
$$specificity = \frac{\text{number of True Negatives}}{\text{number of True Negatives} + \text{number of False Negatives}}$$

To limit the overfitting issue, k-fold cross-validation will be applied in the evaluation process. Here we choose k = 10, and each fold has 20 test data. The supervised classifier is trained by 178 labeled records, and semi-supervised classifier incorporates the remaining unlabeled records in Diagnostic.

So far, we have finished Data-preprocessing and Naive Bayesian and its improvements. Here is the preliminary results:
We observed that the sensitivity value is 34.04%, which is higher than the baseline.

Table 2: Gaussian Naive Bayesian with improvement

	$\hat{y} = R$	$\hat{y} = N$	success rate
$y = R$	16	31	0.3404
$y = N$	14	137	0.9073

5 Timeline and Work Division

We roughly finished the midway goal set in the proposal. We plan to implement the Semi-supervised classifier by Nov 20th and use the rest time to try different learning algorithm. Chang will work on algorithm design for the Semi-supervised part. Kangning will continue to improve the existing supervised classifier. The team will work together for final report and presentation.

References

[1] Diana Dumitru. "Prediction of recurrent events in breast cancer using the Naive Bayesian classification" *Annals of University of Craiova* (2009), Vol 36(2).

[2] Chapelle, Olivier; Schlkopf, Bernhard; Zien, Alexander (2006). *Semi-supervised learning.* Cambridge, Mass.: MIT Press.ISBN978-0-262-03358-9.

[3] Zhu, Xiaojin.*Semi-Supervised Learning*University of Wisconsin-Madison.

[4] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell. *Text classification from labeled and unlabeled documents using EM. Machine Learning*, 39(2/3):103?134, 2000.