Classification of Breast Cancer Subtypes via Gene Expression

Kevin O'Connor
Department of Statistics and Operations Research
University of North Carolina - Chapel Hill
Chapel Hill, NC
koconn@live.unc.edu

Abstract

Breast cancer subtype identification is an important problem with high clinical relevance. In this paper we use various machine learning methods to predict patients' breast cancer subtype from the measured expression levels of a subset of their genes. It is shown that multinomial logistic regression can achieve a moderate level of accuracy whereas modern classification methods like svm and a fully-connected neural network can achieve far greater results. The performance of each method is then evaluated and compared.

1 Introduction

For many years now, successful treatments for various types of cancers have eluded re- searchers. Breast cancer in particular continues to affect millions of women and men. An important characteristic in the study of breast cancer is the subtype of disease, identified as Luminal A, Luminal B, Basal, HER2, or Normal. For a single tumor, the subtype determines both how the disease may develop in the patient as well as the proper course of treatment. Thus, there is significant incentive for researchers to find distinctive characteristics of each subtype to more clearly identify them in future patients.

In the past decade, the cost of obtaining genetic information from an individual has de- creased substantially. This has led to a boom in the amount of genetic data available in many areas of medicine. In 2005, a database called The Cancer Genome Atlas (TCGA) was established in order to organize this data and make it widely available to researchers around the world for analysis. At this point, the database contains genetic expression levels for over 20,000 genes from patients with tumors in each of the 5 breast cancer subtypes. It offers an excellent opportunity to look for patterns in the genetic data that distinguish the subtypes.

1.1 Code

The code used to perform the analyses was written in R and Python/Keras and can be found at

`https://github.com/oconnor-kevin/comp755`

1.2 Data

The data can be downloaded from the following website,

`https://portal.gdc.cancer.gov`

After filtering to remove genes with low variances and log-transforming, we are left with 1201 samples and 4101 genes. The code for filtering and transforming is included in the repository above.

2 Experiments

2.1 Multinomial Logistic Regression

In our first experiment, we will fit a multinomial logistic regression model to the data.

2.1.1 The Model

Given a sample $X_i \in \mathbb{R}^p$ with label $y_i \in \{1, ..., K\}$ and K coefficient vectors, $\{\beta_1, ..., \beta_K\}$, we can write our model as

$$\mathbb{P}(y_i = k) = \frac{\exp\{X_i\beta_k\}}{\sum_{k'} \exp\{X_i\beta_{k'}\}}$$

For n independent observations, $X_1, ..., X_n$, this gives us a joint likelihood of the correct classes,

$$\mathcal{L}(\beta_1, ..., \beta_K; \{X_i, y_i\}_{i=1}^n) = \prod_{i=1}^n \frac{\exp\{X_i\beta_{y_i}\}}{\sum_{k'} \exp\{X_i\beta_{k'}\}}$$

and log-likelihood,

$$\mathcal{NLL}(\beta_1, ..., \beta_K; \{X_i, y_i\}_{i=1}^n) = -\sum_{i=1}^n \left[X_i\beta_{y_i} + \log\left(\sum_{k'} \exp\{X_i\beta_{k'}\} \right) \right]$$

2.1.2 Training

In order to fit the model to the data, we minimize the negative log-likelihood via gradient descent to find the maximum likelihood estimator of the parameters $\{\beta_1, ..., \beta_K\}$. Note that care has to be taken when computing the second term in the log-likelihood to avoid numerical overflow.

2.2 K-means

Next we consider an unsupervised learning approach to identify subgroups. We will apply the K-means algorithm to our data to produce 5 clusters which we hope will correspond to breast cancer subtypes.

2.2.1 The Algorithm

In this algorithm, we start by randomly initializing the means of each cluster, $\{\hat{m}_1^0, ..., \hat{m}_5^0\}$. Then at iteration t, assign point i to the cluster with the closest mean,

$$\hat{y}_i^t = \text{argmin}_k \|X_i - \hat{m}_k^{t-1}\|_2$$

recomputing the means of the newly clustered data after each iteration,

$$\hat{m}_k^t = \frac{1}{n_k^t} \sum_{i:\hat{y}_i^t = k} X_i$$

where n_k^t is the number of points in cluster k at iteration t. This is repeated until convergence or some maximum number of iterations is reached.

2.2.2 Purity

As this is an unsupervised clustering algorithm, the output won't specify which cluster corresponds to which subtype. If we did know this, it would be straightforward to compute the classification accuracy. But as this is not known, we need to think more carefully about how to evaluate the performance of the algorithm.

One option is to compute the purity. From a high level, the purity measures how homogeneous the clusters that have been found are. Specifically, let n_{kl} be the number of elements in cluster k which belong to subtype l. Then associate a label with cluster k,

$$c_k = \text{argmax}_l n_{kl}$$

Table 1: Characteristics of clusters resulting from K-means clustering algorithm. (Purity=0.7535)

K-means Clustering Results		
Cluster	Majority Label	Size
1	Luminal B	222
2	Normal	112
3	Luminal A	340
4	Basal	196
5	Luminal A	331

Table 2: Binary classification accuracy for SVM on train and test data sets.

SVM Classification Accuracy		
Subtypes	**Train Accuracy**	**Test Accuracy**
Luminal A vs. Luminal B	98.89%	91.77%
Her2 vs. Basal	99.54%	98.18%
Normal vs. Rest	97.91%	96.25%

Then c_k represents a majority vote of the labels in cluster k. Then define purity, P,

$$P = \sum_k \frac{n_{kc_k}}{n}$$

Notice that $P \in [0, 1]$ and P close to 1 corresponds to more homogeneous clusters. We will use this to evaluate the performance of the K-means algorithm.

2.3 SVM

Moving on to more modern classification methods, we apply an SVM. This attempts to find a boundary in the data space which maximizes the margin between the support vectors from different classes. This is achieved by minimizing the *hinge loss*,

$$L_{\text{hinge}}(y, f(\mathbf{x})) = (1 - yf(\mathbf{x}))_+$$

where $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$, along with a penalty for the size of the weights, giving us a complete loss function of

$$L(\mathbf{w}; \{X_i, y_i\}_{i=1}^n) = \sum_{i=1}^n (1 - y_i(\mathbf{w}^T X_i + w_0))_+ + \frac{1}{2}\|\mathbf{w}\|_2^2$$

which is minimized over $\mathbf{w}, w_0 \in \mathbb{R}^p, \mathbb{R}$. Unfortunately as SVM's partition the data space into only two regions, they do not generalize well to multi-class classification problems. In order to apply them to our problem, we try two approaches:

1. Consider a subset of the data with only two classes.
2. Consider a *one vs rest* classification problem.

The two pairs of subtypes considered are Luminal A vs. Luminal B and Her2 vs. Basal. In the one vs rest scheme, we consider predicting the Normal subtype vs. the other four.

2.4 Fully-connected Neural Network

While SVM's were for several decades considered to be state of the art in prediction problems, neural networks have surpassed them in the past decade. Many different kinds of neural networks have been developed for a variety of different prediction problems such as image classification, natural language processing, and time series learning. However, a simple fully-connected network will suffice for our data as there is no sequential or spatial dependencies between the data that might necessitate a more sophisticated network.

Table 3: Classification accuracy for held out test set with 20% of data.

Classification Accuracy	
Model	Accuracy
Mult. Logistic Regression	48.75%
FC Neural Network	53.75%

The network we will use has 4 densely connected hidden layers with sizes (1024, 2048, 1024, 512), batch normalization and dropout (with probability 0.5) at each layer, and ReLU activation functions. On the output layer, we have 5 nodes with a softmax activation. Furthermore, we train with 100 epochs and batch size of 20. The weights for the network can be found in the repository at the link in section 1.1.

3 Discussion

We have seen that the problem of classifying breast cancer subtype using gene expression can be posed and solved in various ways. The simple approach of trying to learn a map between input data points to subtype labels is taken by both the multinomial logistic regression (MLR) model and fully-connected neural network (FC NN). A comparison of their results can be found in Table 3. We see that while the neural network does improve performance, it does not do so significantly. This is especially true when considering that the FC NN model requires many thousands of parameters more than the MLR model. This was a somewhat surprising result as FC NN's often attain much greater classification accuracy with other data. However, the FC NN model used here could almost certainly be improved by altering the architecture and tuning hyperparameters in a systematic way, so these results should not be thought of as an upper bound on the performance of such a model.

Stated slightly differently, one may alternatively think of the problem as trying to distinguish subtypes as clusters in the data space. This is accomplished by the K-means algorithm. Table 1 contains the results from applying the algorithm to the breast cancer data. We see that, based on a majority vote of its members, the clusters roughly correspond to subtype. Further evidence of this is given by the purity which was computed to be 0.7535 indicating that the clusters were substantially homogeneous and thus subtypes were fairly distinguishable.

Finally, we may pose our problem as a binary classification problem, either trying to distinguish between just two of the subtypes, or distinguish one subtype from the rest. This was accomplished by the SVM approach. We found that an SVM was able to achieve classification accuracy greater than 90% in all three test sets. This suggests that an SVM approach is a powerful method for solving the problem stated as a binary classification problem.

4 Conclusion

In this paper, we investigated a number of different classification methods for identifying breast cancer subtype using gene expression. We saw that even older methods like multinomial logistic regression and K-means were able to distinguish subtypes fairly well. Furthermore, it was observed that the more modern methods like an SVM and a fully-connected neural network yielded an improvement in performance as expected.

Acknowledgments

Thank you to my girlfriend, Cambria, for giving up our Friday night together so I can finish this project!

References

[1] Chollet, Francois. "Deep Learning with Python". Manning Publications Co., 2017.

[2] Murphy, Kevin P. “Machine Learning: A Probabilistic Perspective. Adaptive Computation and Machine Learning." (2012).