

**A comparison of naïve Bayes and logistic regression classifiers for mass
spectrometry data**

Jie Zheng

25/05/2007

Introduction

Proteomic pattern analysis using mass spectrometry (MS) has been employed for early diagnosis of many diseases such as Ovarian cancer (Petricoin *et al.*, 2002; Veenstra and Yates 2006). Various well-known classification algorithms have been developed for complex data analysis such as MS data, for example, linear discriminant analysis, k-nearest neighbor, random forest, and support vector analysis. Comparisons between these algorithms have been investigated both theoretically and empirically (Wu *et al.*, 2003; Boos *et al.*, 2005; Ng and Jordan 2002; Oh *et al.*, 2006). However, there is no single best classifier for all given problem. Classifier performance depends greatly on the characteristics of the data.

In this report, we compare the performance of two linear classifiers of Naïve Bayesian (NB) and Logistic Regression (LR) based on empirical MS spectra, and search for the characteristics of the data that determine the performance. The comparison between LR and NB has been studied theoretically by Ng and Jordan (2002).

Material and preprocessing

The dataset was downloaded from clinical proteomics program databank (Ovarian Dataset 8-7-02, online data source). This low resolution data were produced by using the WCX2 protein chip and an upgraded PBSII surface enhance laser desorption ionization (SELDI) time of flight (TOF) mass spectrometer. The sample sets consist of 91 controls and 162 ovarian cancers.

Preprocessing is an important step in the analysis of MS data. The data were preprocessed by the computer program PrepMs 1.0 (Morris, et al. 2005). The data preprocessing includes calibration, spectral de-noising, baseline correction and normalization, peak detection, and peak quantification. Default parameter setting was used. Matrix noise was eliminated with parameters the matrix saturation point of 2000 Da and signal-to-noise of 5, which means less conservative and will eliminate less peaks. The spectra were aligned by using the top 5 peaks detected in the mean spectrum, and smoothed with smoothing threshold of 10 for peak detection and 4 for peak quantification.

After preprocessing, we obtained 326 discrete peaks (features) for each spectrum. Fig. 1 shows a comparison of two example spectra (one from control, and the other from cancer) before and after preprocessing. As shown in Fig. 1E&F, the differences of MS intensities of control and cancer are signals when m/z is close to zero, and the signals after preprocessing become stronger around the m/z value of 9000. The total 253 samples with 326 features per sample at same locations were used as training and testing data.

Method

NB and LR algorithms are mostly common used and extensively studied, and they have many variants (Huang *et al.*, 2005). We use the most basic ones focusing on binary classification with discrete feature values and describe them briefly as follows (see e.g. Mitchell 2006 for detail). The two classifiers were implemented by Mathematica 5.1.

The NB algorithm assumes the features x_i ($i=1, \dots, n$) are all conditionally independent of one another when given the class y ($y=0$ for control, and 1 for cancer).

Applying Bayes rule, we have

$$p(y | x_1, \dots, x_n) \sim \prod_{i=1}^n p(x_i | y) p(y),$$

where $p(y|x)$ denotes conditional probability of y given x , and $p(y)$ the prior probability of y . We assume equal prior probability for each class ($p(y) = 0.5$). We assume that $p(x_i | y)$ follows normal distribution with mean μ_y and standard deviation σ_y , which was substituted by estimates from training data. The most probable value of y is the value that maximizes $p(y | x_1, \dots, x_n)$ with respect to y .

LR assumes

$$p(y = 1 | x_1, \dots, x_n) = \frac{1}{1 + \exp(\beta_0 + \sum_{i=1}^n \beta_i x_i)},$$

and $p(y = 0 | x_1, \dots, x_n) = 1 - p(y = 1 | x_1, \dots, x_n)$. We assign the value y that maximizes

$p(y | x_1, \dots, x_n)$. Thus, we set $y=0$ if $\frac{p(y = 0 | x_1, \dots, x_n)}{p(y = 1 | x_1, \dots, x_n)} > 1$ i.e. $\beta_0 + \sum_{i=1}^n \beta_i x_i > 0$,

otherwise $y = 1$.

Result

We divided the data into training data and testing data. Training data were sampled by randomly choosing N_T different MS control spectra and N_T different MS cancer spectra. The rest data were used as testing. Prediction accuracies for control, cancer, and total testing MS spectra were calculated separately for different values of N_T . The results are shown in Fig. 2A obtained by normal NB classifier, and Fig. 2B by LR.

Fig. 2 shows that the prediction accuracies are different in several aspects. (1) The accuracy obtained by LR increases with the training sample size, while there is no clear trend for NB. (2) The asymptotic prediction accuracy (when N_T is large) obtained by LR is larger than that by NB for control, cancer, or total testing MS spectra. (3) The accuracy obtained by LR is smaller than NB when the training sample size is small ($N_T \sim 20$). (4) As to NB the accuracy for control spectra is larger than the accuracy for cancer spectra, whereas there is no such obvious difference for LR.

The fluctuation in prediction accuracy was caused by both the classifier and the sparse of the testing data which getting less with N_T .

Discussion

When there is limited amount of data, NB performs better than LR. This is because of independent feature assumption of NB which does not require the entire covariance matrix. Whereas LR overfits the training data especially when there are many features and training data is sparse. Regularization in LR to reduce overfitting has been suggested (Mitchell 2006). Theoretically, the linear classifiers NB and LR are

identical in the limit of the number of training spectra approaching infinity, provided the NB assumptions hold (Mitchell 2006; Boos *et al.*, 2005). Therefore, that asymptotic accuracy predicted by NB being smaller than by LR indicates that features are either non-normal distributed or correlated. The accuracy for control spectra predicted by NB is larger than for cancer spectra, probably because the variance among cancer spectra is larger than that among control spectra, as shown in Fig. 3. This makes sense since the cancer spectra might be in different periods. On the other hand, LR is a discriminative classifier which estimates directly the class of any given spectrum, and it thus does not depend on variance of any feature.

Our results suggest that the performance of NB and LR depends on the amount of training data, correlation between features, and probability distribution of each feature, which may vary with empirical data.

Reference

- Huang, K., Zhou, Z., King, I., and Lyu, M.R.. 2005. Improving Naïve Bayesian classifier by discriminative Training. *International Conference on Neural Information Processing*, Taipei, October 30 - November 2, pp. 49-54.
- Mitchell, T.M. 2006. *Machine learning*. Chapter 1.
<http://www.cs.cmu.edu/~tom/mlbook.html>.
- Morris, J.S., Coombes, K.R., Koomen, J., Baggerly, K.A. and Kobayashi, R. 2005. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum, *Bioinformatics*, **21**, 1764-1775.

-
- Ng, A.Y., and Jordan, M.. 2002. On Discriminative vs. Generative Classifiers: A comparison of logistic regression and Naive Bayes. *Neural Information Processing Systems*, **14**.
- Oh, J.H., Nandi, A., Gurnani, P., Knowles, L., Schorge, J., Rosenblatt, K.P., and Gao, J.X.. 2006. Proteomic biomarker identification for diagnosis of early relapse in ovarian cancer. *Journal of Bioinformatics and Computational Biology*, **4**, 1159-1179.
- Ovarian Dataset 8-7-02. Dowload from clinical proteomics program databank with website address: <http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp>.
- Petricoin III, E.F, Ardekani, A.M., et al.. 2002. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359**, 572–577.
- Roos, T., Wettig, H., Grünwald, P., Mullymäki, P., and Tirri, H.. 2005. On discriminative bayesian network classifiers and logistic regression. *Machine Learning*, **59**, 267–296.
- Veenstra, T.D., Yates, J.R. 2006. *Proteomics for biological discovery*. Chapter 13.
- Wu, B., Abbott, T., Fishman, D., et al.. 2003. Comparison of statistical methods for lassification of ovarian cancer using mass spectrometry data. *Bioinformatics* **19**, 1636–1643.

Figure legends

Figure 1: Comparisons of MS spectra before (left side) and after (right side) preprocessing. Subscript “ctr” denotes an example of control spectra, and “can” an example of cancer spectra.

Figure 2: Plot of the prediction accuracy obtained by Naïve Bayesian method (A) and Logistic regression (B) verse training sampling size (N_T). Green rectangles represent the prediction accuracy for test spectra of control, blue triangles for cancer, and red stars for all spectra. Equal number (N_T) of spectra from control and cancer was used in the test.

Figure 3: Standard deviations of cancer spectra minus that of control spectra at each feature location. All the spectra were included.

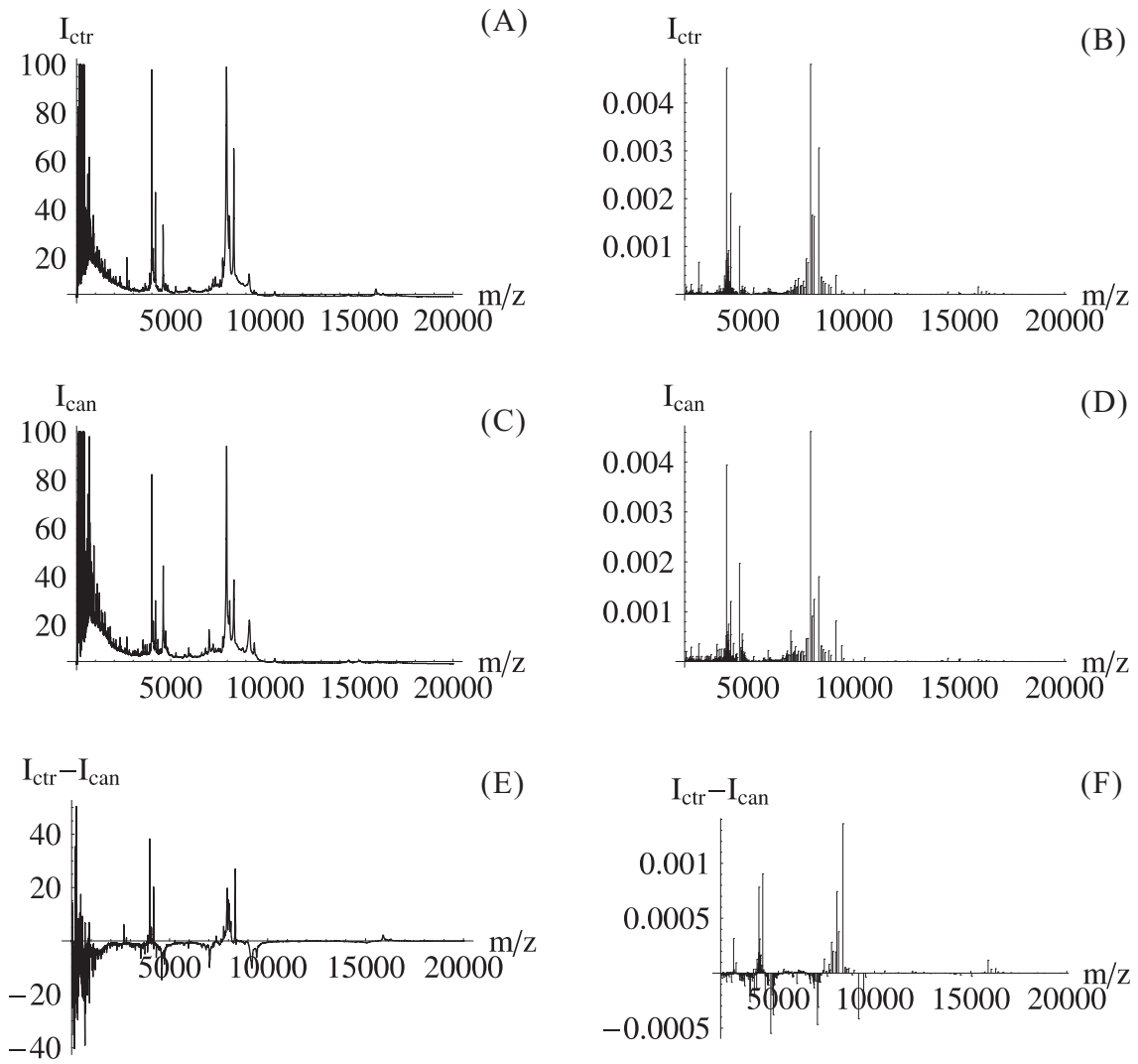


FIGURE 1

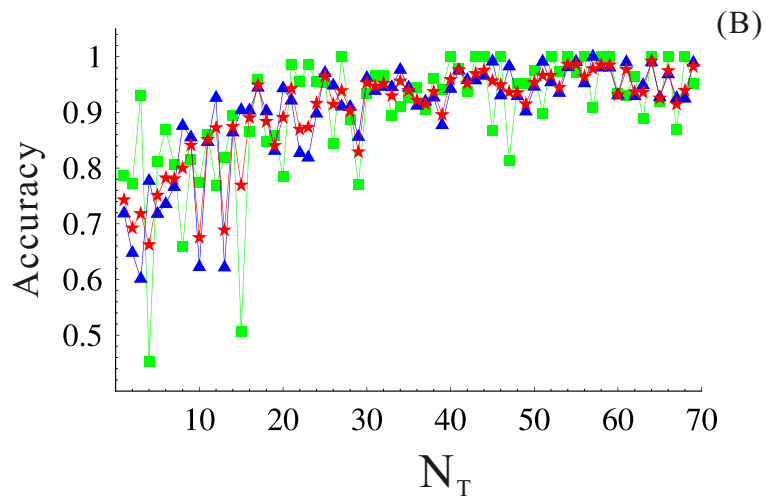
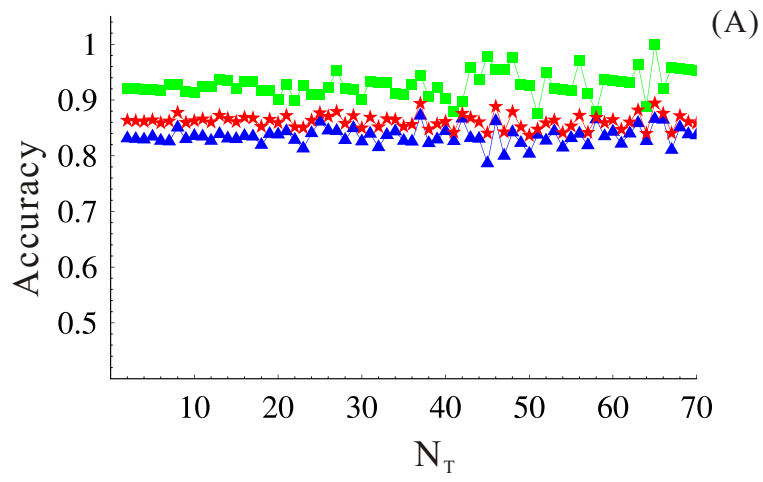


FIGURE 2

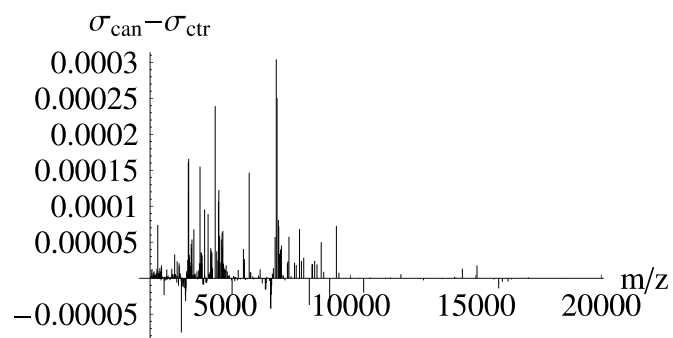


FIGURE 3