Linking Gene-Expression Experiments with Survival-Time Data

Liat Ben-Tovim Jones[1], Shu-Kay Ng[1], Katrina Monico[1] and Geoff McLachlan[1]

[1] Department of Mathematics and Institute for Molecular Bioscience, University of Queensland, Brisbane 4072, Australia

Abstract: We apply a model-based clustering approach to classify tumour tissues on the basis of microarray gene expression. The association between the clusters so formed and patient survival (recurrence) times is examined. The approach is illustrated using the lung cancer data set of Wigle et al. (2002). We show that the prognosis clustering is a powerful predictor of the outcome of disease, in addition to the stage of disease at presentation.

Keywords: Mixture models; EMMIX-GENE algorithm; Microarrays; Survival analysis; Cox proportional hazards.

1 Introduction

In clinical medicine, accurately determining the stage of disease is crucial in the management of cancer patients. Stage is defined using a combination of clinical parameters (tumour size, lymph node involvement and the presence of metastases). However, patients with the same stage of a particular cancer can have very different treatment responses and also clinical outcome. There is much interest in determining whether microarrays can be used as better indicators for outcome. Here we demonstrate how model-based clustering in conjunction with survival analysis can be used to assess the prognostic information in microarray data. We report in detail our results for the lung cancer data set of Wigle et al. (2002). This data set formed part of the CAMDA'03 challenge, and a fuller description of the methods is given in Ben-Tovim Jones et al. (2004), and also their application to the three other CAMDA'03 lung cancer data sets.

2 Cluster Analysis

Wigle et al. (2002) used cDNA microarrays to measure the gene expressions for 39 tumour samples from patients diagnosed with various types of lung cancer. We downloaded the data at `http://www.camda.duke.edu/camda03`, and used the set of 2880 genes as in Wigle et al. (2002). For each patient, the

clinical outcome was given as the time between surgery and the recurrence. We label 1 to 24 the patients for which there has been a recurrence of the cancer, while those labelled 25-39 had no recurrence before the end of the study (their times to recurrence are censored). We input the data into the EMMIX-GENE algorithm of McLachlan et al. (2002). In the first screening step, 766 genes remained and these were then clustered into 20 groups. The means of these 20 groups (the metagenes) were used to cluster the tissues in the final step of EMMIX-GENE. Given the very small number of tumours (39) available here relative to the number of genes or indeed metagenes, some constraints had to be imposed on the component-covariance matrices in fitting a normal mixture model to cluster these tumours. We considered fitting to all 20 metagenes (a) mixtures of normals with equal component-covariance matrices; (b) mixtures of normals with (unrestricted) diagonal component-covariance matrices; and (c) mixtures of factor analyzers with equal component-covariance matrices for $q = 6$ factors. All three models led to two clusters, represented as

$$C_1 = \{15, 30-32, 34, 35, 37, 39\} \text{ and } C_2 = \{1-14, 16-29, 33, 36, 38\}.$$

Cluster C_1 corresponds to the good-prognosis group with 7 patients who are recurrence-free plus 1 patient who had experienced relapse of the tumour. This patient, however, was still alive at the end of the follow-up period. Cluster C_2 corresponds to the poor-prognosis group as it contains 23 of the 24 patients with recurrence, plus 7 patients with censored recurrence times. To further show that the first cluster C_1 corresponds to a recurrence-free group, we considered the long-term survival model

$$S(t) = \pi_1 + \pi_2 S_2(t), \tag{1}$$

where t is the time to recurrence, $S_2(t)$ is the conditional survival function for time to recurrence given recurrence will occur, and $\pi_2 = 1 - \pi_1$ is the probability of a recurrence. Under (1), a proportion π_1 of the patients will not have a recurrence; that is, their recurrence time is at infinity. The survival function $S_2(t)$ is taken to have the Weibull form,

$$S_2(t) = \lambda t^{\alpha-1} \exp(-\lambda t^{\alpha}). \tag{2}$$

The exact recurrence and survival times of two patients in C_2 were unknown and so they were excluded from all the survival analyses, leaving 37 patients with 15 of these censored. In Figure 1, we plot the fitted Weibull-based long-term survival model $\hat{S}(t)$ along with the Kaplan-Meier estimate. This shows excellent agreement between the nonparametric estimate as given by the Kaplan-Meier estimate and the parametric estimate $\hat{S}(t)$. In particular, from the asymptote of the curves, the probability π_1 of a patient being recurrence-free is approximately 0.2. Thus on average, one would expect to have approximately 8 recurrence-free patients in a set of 39. Here the cluster C_1, which is conjectured as corresponding to the

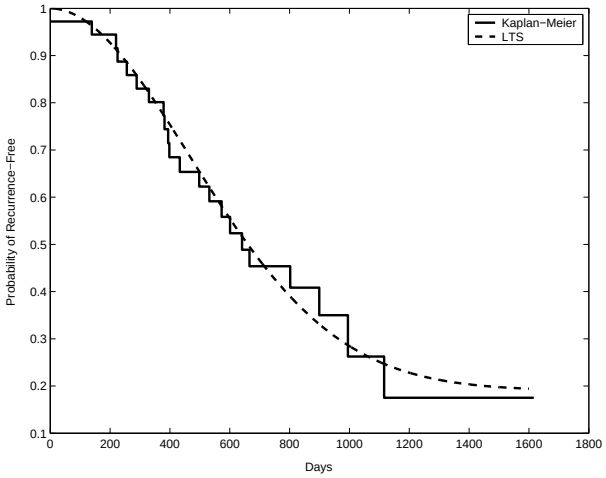

FIGURE 1. Fitted LTS model versus Kaplan-Meier.

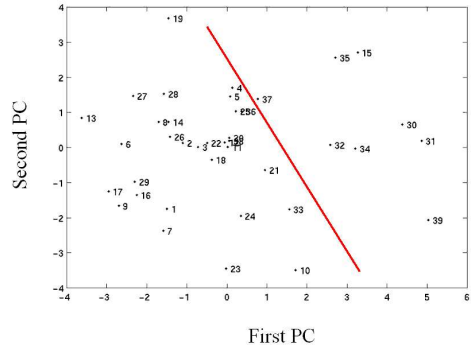

FIGURE 2. PCA of tissues based on 20 metagenes.

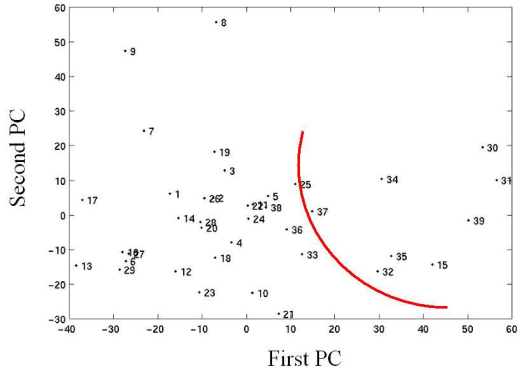

FIGURE 3. PCA of tissues based on all genes (via SVD).

recurrence-free group, has indeed 8 members in it. Interestingly, 5 of the censored patients clustered into C_2 were also put together in a cluster corresponding to early recurrence in the hierarchical clustering of Wigle et al. (2002). This long-term survival model (1) can be used also to estimate the posterior probability that a patient with a censored recurrence time will be recurrence-free. Unfortunately, unless the censored time is very long, these estimated posterior probabilities are equal, being around 0.5. Patient (P81 AC) who has a censored time of 1,161 days has a high posterior probability of being recurrent-free so her membership of cluster C_1 would appear to be atypical. To further investigate the validity of our clustering of the 39 tumours, we considered a plot of the first two principal components (PCs) of the tumours obtained by a singular-value decomposition based on (a) the 20 metagenes and (b) all the genes, as given in Figures 2 and 3, respectively. In each of these two figures, we have imposed the allocation boundary that will give the clustering that we have obtained above. In each case, it can be seen that this boundary represents a resaonable partition of the data into two clusters in the space of the first two PCs.

3 Survival Analysis

For the 37 patients with survival data available, we clustered 29 as poor prognosis (C_2) and 8 as good prognosis (C_1). We use the Kaplan-Meier estimate to provide an estimate of the overall probability of being recurrence-free following surgery. Given that there is only one recurrence in C_1, it should have a significantly better Kaplan-Meier estimate than C_2, and this is confirmed in Table 1. These two Kaplan-Meier estimates are plotted in Figure 4. The Kaplan-Meier curves were compared with the use of the log-rank test.

TABLE 1. Non-parametric Survival Analysis

Cluster	No. of Patients (Censored)	Mean Time to Recurrence ($\pm$ SE)
C_1	8 (7)	1388 $\pm$ 155.7
C_2	29 (8)	665 $\pm$ 85.9

We also fitted the proportional hazards model of Cox (1972), using covariates to represent the clinical data and a zero-one indicator variable to membership of cluster C_1 or not. The fit for the final form of this model is given in Table 2. The significance of estimated hazard ratios were tested using the Wald test. All calculations in the survival analysis were performed with the S Plus statistical package. It can be seen that membership of cluster C_1 (the poor-prognosis cluster) was the only significant factor affecting the event of being recurrence-free ($P = 0.06$).

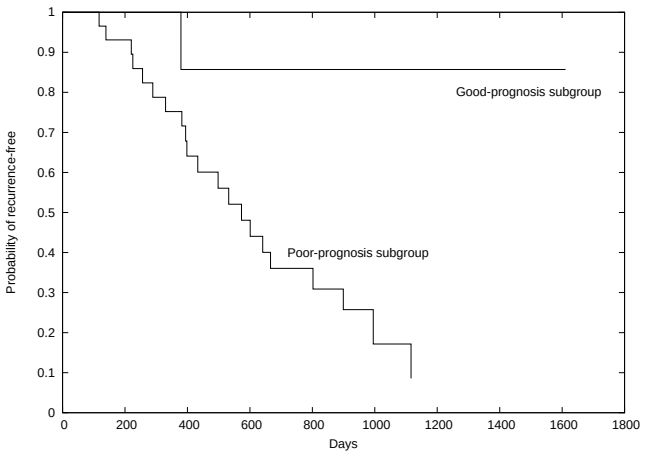

FIGURE 4. Kaplan-Meier curves of recurrence-free for the two clusters.

TABLE 2. Multivariate Cox Hazards Analysis of the Risk of Recurrence

Variable	Hazard Ratio (95%CI)	*P*-Value
Poor (vs. good prognosis cluster)	6.8 (0.9-51.8)	0.06
Stages 2 or 3 (vs. Stage 1)	1.1 (0.4-2.7)	0.88

4 Conclusions

We were able to use a model-based clustering approach to to identify patient clusters with clinical outcomes of recurrence versus non-recurrence of tumour. The gene-expression data provided prognostic information, beyond the clinical indicator of stage. A limiting factor in the analyses was the small numbers of tumours available. Further, the high proportion of censored observations limited the comparison of survival rates.

References

Ben-Tovim Jones, L., Ng, S.K., Ambroise, C., Monico, K., Khan, N., et al. (2004). Use of microarray data via model-based classification in the study and prediction of survival from lung cancer. In: *Methods of Microarray Data Analysis IV*, K.F Johnson and S.M. Lin (Eds.). Dordrecht. Kluwer. To appear.

McLachlan, G.J., Bean, R.W., and Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**, 413–422.

Wigle, D.A., Jurisica, I., Radulovich, N., Pintilie, M., et al. (2002). Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. *Cancer Research*, **62**, 3005–3008.