

A PHASE TYPE SURVIVAL TREE MODEL FOR CLUSTERING PATIENTS’ HOSPITAL LENGTH OF STAY

Lalit Garg¹, Sally McClean², Brian Meenan³, Peter Millard⁴

^{1,2,3}University of Ulster, Coleraine, Co. Londonderry, BT52 1SA, UK

⁴St. George’s Hospital Medical School,

12 Cornwall Road, Cheam, Sutton, Surrey, SM2 6DR, UK

E-mail: ¹Garg-l, ²si.mcclean, ³bj.meenan}@ulster.ac.uk; ⁴phmillard@tiscali.co.uk

Abstract. Clinical investigators, health professionals and managers are often interested in developing criteria for clustering patients into clinically meaningful groups according to their expected length of stay. In this paper, we propose phase-type survival trees which extend previous work on exponential survival trees. The trees are used to cluster the patients with respect to length of stay where partitioning is based on covariates such as gender, age at the time of admission and primary diagnosis code. Likelihood ratio tests are used to determine optimal partitions. The approach is illustrated using nationwide data available from the English Hospital Episode Statistics (HES) database on stroke-related patients, aged 65 years and over, who were discharged from English hospitals over a 1-year period.

Keywords: Survival trees, survival analysis, CART, recursive partitioning, patient clustering, statistical learning, stroke care, phase type distributions.

1. Introduction

Decision trees in survival analysis are popularly known as survival trees and are type of classification and regression tree (Breiman et al. 1984, Davis and Anderson, 1989). Survival tree based analysis is a powerful non-parametric method of clustering survival data for prognostication i.e. to determine importance and effect of various covariates (such as patient’s characteristics) and their interrelation on patient’s survival, treatment outcome, disease risk, disease progress or hospital length of stay (Davis and Anderson, 1989, Gao et al., 2004). In this paper, we illustrate how phase type survival trees can be constructed and used for clustering length of stay data.

2. Background

Phase type distributions are among popular choices to fit spell length of stay data (Fackrell, 2009). Fackrell (2009) compares five subclasses of phase type distributions and based on log-likelihood values he identified that the general phase type distributions provide the best fit followed by Coxian phase type distributions. However, general phase type distributions are over-parameterized and parameter estimation is difficult (Fackrell, 2009, Marshall and McClean, 2004). On the other hand, Coxian phase type distributions do not present such problems and also provide a simple interpretation of fit for the length of stay data (Fackrell, 2009). We model patient flow in the care system as an n state Markov process (Fig. 1) with Coxian phase type distributions (Cox, 1955, Marshall and McClean, 2004, McClean et al., 2007). A patient can be admitted to the care system only in the first state (state 1). Sequential transitions are possible from any state k (where $k = 1, 2, \dots, n$) to the next state $k+1$ with a transition rate λ_k . Also transition is possible from any state k to the absorbing state $n+1$ with a transition rate μ_k . The absorbing state represents the event discharge or death of the patient. The time spent in the hospital before discharge or death has the probability density function:

$$f(t) = \mathbf{p} \exp(\mathbf{Q}t) \mathbf{q} \quad (1)$$

where the row vector \mathbf{p} , the initial state probability distribution is defined as:

$$\mathbf{p} = (1 \ 0 \ 0 \ \dots \ 0 \ 0) \quad (2)$$

the transition matrix \mathbf{Q} is defined as

$$\mathbf{Q} = \begin{pmatrix} -(\lambda_1 + \mu_1) & \lambda_1 & 0 & \dots & 0 & 0 \\ 0 & -(\lambda_2 + \mu_2) & \lambda_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & -(\lambda_{n-1} + \mu_{n-1}) & \lambda_{n-1} \\ 0 & 0 & 0 & \dots & 0 & -\mu_n \end{pmatrix} \quad (3)$$

and the column vector \mathbf{q} represents absorption probabilities and is defined as

$$\mathbf{q} = (\mu_1 \ \mu_2 \ \dots \ \mu_{n-2} \ \mu_n)^\top . \quad (4)$$

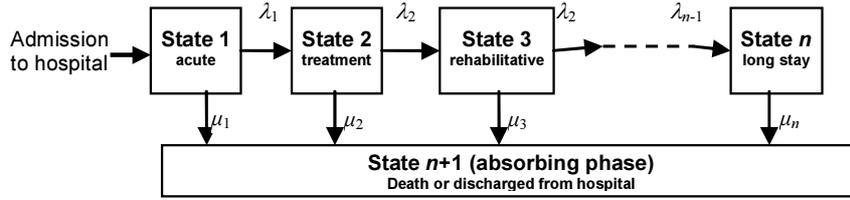


Fig. 1. Stroke care system modelled as an n state Markov process with Coxian phase type distribution

The likelihood function is defined as follows (Marshall and McClean, 2004):

$$l = \prod_{i=1}^N (\mathbf{p} \exp \{ \mathbf{Q} t_i \} \mathbf{q}) \quad (5)$$

where N is the total number of patients in the care system and t_i is the spell length of stay of a patient i ($i = 1, 2, 3, \dots, N$). It is more convenient to work with log likelihood function which can be defined as:

$$L = \sum_{i=1}^N (\log (\mathbf{p} \exp \{ \mathbf{Q} t_i \} \mathbf{q})) . \quad (7)$$

This can also be written as:

$$L = \sum_{i=1}^N f(t_i) \quad (8)$$

where

$$f(t_i) = \log (\mathbf{p} \exp \{ \mathbf{Q} t_i \} \mathbf{q}) \quad (9)$$

This order n Coxian phase type fit of spell length of stay data has $2n-1$ free parameters (degrees of freedom) to be estimated. We used freely available downloadable package EMpht (Asmussen et al., 1996) developed by Asmussen et al. (1996) and Olsson (1996), which implements maximum likelihood parameter estimation using the expectation-maximization (EM) algorithm.

3. Tree construction

In this section, we will describe the criteria used for tree construction.

Splitting criteria: Tree construction can be achieved by recursively partitioning into sub groups by one of the covariates based on some splitting criteria maximizing either within node homogeneity or between node separation (Gao et al., 2004). We used splitting criteria to maximize within node homogeneity based on improvement of log-likelihood functions (David and Anderson, 1989). A covariate a can have any of the l values such that

$$N = N_{a1} + N_{a2} + \dots + N_{al} = \sum_{i=1}^l N_{ai} . \quad (10)$$

Therefore equation 8 can also be written as follows:

$$L = \sum_{j=1}^l \sum_{i=1}^{N_{aj}} f(t_{iaj}) = \sum_{i=1}^{N_{a1}} f(t_{ia1}) + \sum_{i=1}^{N_{a2}} f(t_{ia2}) + \dots + \sum_{i=1}^{N_{al}} f(t_{ial}) \quad (11)$$

or

$$L = L_{a1} + L_{a2} + \dots + L_{al} = \sum_{i=1}^l L_{ai} . \quad (12)$$

Covariate a splits the dataset into l subgroups and each subgroup is separately fitted to the Coxian phase type distribution where the total log-likelihood

$$L = \sum_{j=1}^l \sum_{i=1}^{N_{aj}} f^{(iaj)}(t_{iaj}) = \sum_{i=1}^{N_{a1}} f^{(ia1)}(t_{ia1}) + \sum_{i=1}^{N_{a2}} f^{(ia2)}(t_{ia2}) + \dots + \sum_{i=1}^{N_{al}} f^{(ial)}(t_{ial}) \quad (13)$$

and
$$f^{(iaj)}(t_{iaj}) = \log(\mathbf{p}^{(iaj)} \exp\{\mathbf{Q}^{(iaj)} t_{iaj}\} \mathbf{q}^{(iaj)}) . \quad (14)$$

In other words the total log-likelihood is the sum of individual log-likelihoods of each sub-group partitioned by covariate a .

Selection criteria: Cross-validation, bootstrap re-sampling and other popular pruning techniques are extremely expensive for large datasets (Gao et al., 2004). Therefore, we are using a simpler approach of determining if a node is a terminal node. If it is not then we select the best possible partition by exploring all possible splits. A terminal node is the node at which within node homogeneity cannot significantly be improved by any possible split. At each node we will apply one covariate at a time and record the total log-likelihood for partitioning by that covariate. Then we will repeat this with other covariates. The covariate which maximizes the total log-likelihood of sub-groups is determined and L_{\max} is calculated as follows.

$$L_{\max} = \max(L_a, L_b, \dots, L_c) . \quad (15)$$

Now we compare this log-likelihood with the log-likelihood of the node before partition and calculate the value of chi-square statistic $\chi^2_{(df)}$:

$$\chi^2_{(df)} = 2(L_{\max} - L_p) \quad (16)$$

where degrees of freedom

$$df = df_{\max} - df_p \quad (17)$$

where df_{\max} is the sum of the degrees of freedom of each of the subgroups partitioned by the log-likelihood maximizing covariate. We used 0.05 significance level ($\chi^2_{(df)}(p < 0.05)$) to determine if the node is a terminal node.

4. Application

To illustrate the phase type survival tree method for clustering patients according to their hospital length of stay, we used the dataset available from the English Hospital Episode Statistics (HES) database representing the first episode of care of 105765 patients with a stroke related code anywhere in their diagnosis chain and discharged between April 1st 1994 and March 31st 1995 from all English hospitals (Vasilakis and Marshall, 2005). All patients were aged 65 or over. No information that identified individual patients was supplied. For this application we identified one continuous covariate i.e., patient's age at the time of admission to hospital and two categorical covariates i.e., patient gender and type of stroke diagnosed. For the continuous covariate we used cut-points that divide patients into three almost equal subgroups. For the categorical covariate patient gender, HES dataset has four different values 1 for male, 2 for female, 3 and 4 for other or unspecified. Values 3 and 4 do not have prognostic significance. Therefore, we discarded daughter nodes created by patient gender covariates having value 3 or 4. The value of the covariate type of stroke diagnosed is determined by the presence of a particular ICD-9 code (World Health Organisation, 1977) anywhere in the diagnostic chain. It can have any of the 4 values. Hemorrhagic Stroke (ICD-430-ICD-432), Ischemic Stroke (ICD-433, ICD-434, ICD-436, ICD-437), Transient Ischemic Attack (TIA) (ICD-435) and other strokes (ICD-438).

5. Results

Figure 2 is the schematic diagram of the final tree we constructed. Table 1 lists the nodes and the possible splits of the tree we constructed. Bold faced covariates represent the splits selected for creating daughter nodes. Node 9, 10, 11 and 12 are nodes created through splitting node 8 by diagnosis covariate. The total gain in the homogeneity by clustering into leaf nodes in terms of log-likelihood value is the difference between root node log-likelihood before clustering and the total log-likelihood of the leaf nodes:

$$G_{Total} = -(L_{root} - (L_4 + L_5 + L_6 + L_7 + L_9 + L_{10} + L_{11} + L_{12}) - L_{discard})$$

where $L_{discard}$ is the log-likelihood of the sub-groups which were discarded (with covariate patient's gender value '3' or '4'). The total gain in log-likelihood is 3793.631635 with 35 extra free parameters ($p=1$).

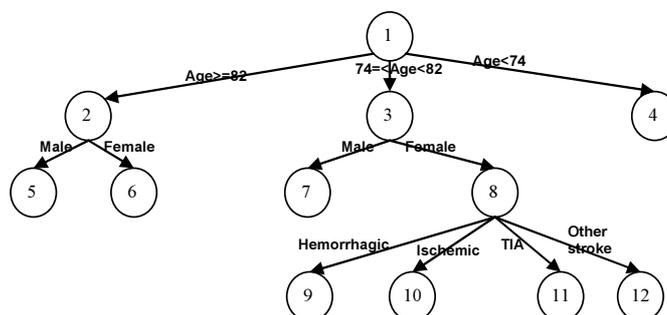


Fig. 2. Phase-type survival tree for HES database on stroke-related patients

Table 1. Tree construction for HES database on stroke-related patients

Node	Covariate	Covariate value	Number of patients	Loglikelihood	Number of phases	Degrees of freedom (df_{max})	Significance (p)	
All	Complete dataset		105765	-372202.5986	14	27		
1 (Root node)	Gender	Male	47136	-159483.7033	7	47	<0.000001	
		Female	58109	-210420.5992	9			
		Unspecified	520	-1827.076608	9			
	Age	Age<=73	34995	-108847.6535	7	51	<0.000001	
		73 <Age< 82	35393	-124625.1882	11			
		Age>=82	35377	-135081.0997	9			
	Diagnosis	Hemorrhagic	5593	-19510.25135	9	66	<0.000001	
		Ischemic	67190	-236434.3777	8			
		TIA	11196	-39030.51958	9			
		Other	21786	-77126.05211	9			
	2 (Age >=82)	Gender	Male	11720	-43505.86672	4	21	<0.000001
			Female	23485	-90841.97838	4		
Unspecified			172	-673.126948	4			
Diagnosis		Hemorrhagic	1860	-7111.186087	4	28	1.000000	
		Ischemic	22605	-86486.26874	4			
		TIA	3777	-14364.33872	4			
		Other	7135	-27159.12134	4			
3 (73 <Age< 82)	Gender	Male	16419	-56342.3129	4	21	<0.000001	
		Female	18808	-67631.89135	4			
		Unspecified	166	-582.499592	4			
	Diagnosis	Hemorrhagic	1871	-6611.551198	4	30	1.000000	
		Ischemic	22351	-78619.50522	4			
		TIA	3697	-12928.67402	5			
		Other	7474	-26500.37126	4			
4 (Age<=73)	Gender	Male	18997	-58518.12493	4	21	1.000000	
		Female	15816	-49962.72347	4			
		Unspecified	182	-549.700237	4			
	Diagnosis	Hemorrhagic	1862	-5634.446937	5	34	1.000000	
		Ischemic	22234	-69026.23065	5			
		TIA	3722	-11338.31477	4			
		Other	7177	-22994.7595	5			

5 (Age>=82 Male)	Diagnosis	Hemorrhagic	593	-2180.100904	4	26	1.000000
		Ischemic	7387	-27518.89464	4		
		TIA	1246	-4636.007606	4		
		Other	2494	-9187.55383	3		
6 (Age>=82 Fe- male)	Diagnosis	Hemorrhagic	1259	-4888.606315	4	28	1.000000
		Ischemic	15100	-58543.49525	4		
		TIA	2511	-9641.868259	4		
		Other	4615	-17827.22779	4		
7 (73 <Age< 82 Male)	Diagnosis	Hemorrhagic	860	-2933.985556	4	28	0.996027
		Ischemic	10191	-34900.37992	4		
		TIA	1740	-5945.469171	4		
		Other	3628	-12558.59084	4		
8 (73 <Age< 82 Female)	Diagnosis	hemorrhagic	998	-3622.623753	4	28	0.049369
		Ischemic	12061	-43332.41141	4		
		TIA	1934	-6883.738102	4		
		Other	3815	-13776.75566	4		

6. Conclusion

In this paper we illustrate how phase type survival trees can be used to cluster, identify and quantify the significance and effects of various covariates (patient characteristics such as age, gender, disease etc.) and their interaction in prediction of patient's length of stay in hospital. We have used splitting criteria based on improvement of log-likelihood functions. As future work we will determine the effect of using other splitting criteria to develop more efficient clustering. Also for continuous covariates we will develop an automated algorithm which can be used to decide optimum cut points.

7. Acknowledgement

The authors acknowledge support for this work from the EPSRC (Grant References EP/E019900/1 and GR/S29874/01). Any views or opinions presented herein are those of the authors and do not necessarily represent those of RIGHT or MATCH, their associates or their sponsors.

References

- Asmussen, S. Nerman O. and Olsson M. 1996. Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics* 23: 419–441.
- Breiman, L. Friedman, J H. Olshen, R. A. and Stone, C. J. 1984. *Classification and Regression Trees*. Wadworth and Brooks/Cole: Monterey.
- Cox, D.R. 1955. *A use of complex probabilities in the theory of stochastic processes*. *Proceedings of the Cambridge Philosophical Society* 51: 313–319.
- Davis, R. and Anderson, J. 1989. *Exponential Survival Trees*. *Statistics in Medicine* 8: 947–962.
- Fackrell, M. 2009. *Modelling healthcare systems with phase-type distributions*. *Health Care Management Science* 12: 11–26.
- Gao, F. Manatunga, A. K. and Chen S. 2004. *Identification of prognostic factors with multivariate survival data*. *Computational Statistics & Data Analysis* 45: 813–824.
- Marshall, A. H. and McClean S. I. 2004. *Using Coxian Phase-Type Distributions to Identify Patient Characteristics for Duration of Stay in Hospital*. *Health Care Management Science* 7: 285–289.
- McClean, S. I. Garg, L. Meehan, B. and Millard, P. H. 2007. Non-Homogeneous Markov Models for Performance Monitoring in Healthcare. In C.H. Skiadas, Eds. *Recent Advances In. Stochastic Modelling and Data Analysis*, 146–153.
- Olsson, M. 1996. *Estimation of phase-type distributions from censored data*. *Scandinavian Journal of Statistics* 23: 443–460.
- Vasilakis, C. and Marshall, A. H. 2005. *Modelling nationwide hospital length of stay: opening the black box*. *Journal of the Operational Research Society* 56: 862–869.
- World Health Organisation. 1977. *International Classification of Diseases, ninth revision (ICD-9)*. WHO: Geneva.