Exploring survival datasets using dimensionality reduction

Lovedeep Gondara
lgondara@sfu.ca

March 6, 2017

1 Introduction

Before fully committing to a study, researchers want to explore their data. This is especially the case with physician researchers, where greater monetary and human resources are required and study outcomes have a profound effect. Most researchers at present use rudimentary exploratory analysis while exploring the data. Most often using a point and click statistical software suite such as SPSS. Problem arises when researchers don't exactly know where to look and most options in such packages are *too advanced* for a statistically naive user. Researchers are most of the times interested in finding heterogeneous groups of patients and to investigate if this heterogeneity has any effect on the survival, i.e. is the survival of patients different in different subgroups? Any indications of different survival patterns in different subgroups motivates researchers for further exploration and analysis. But as this is not straightforward using available tools, researchers end up using cross tabs and some basic scatter plots to get an idea of the information contained in the dataset before they approach an analyst. Which can sometimes result in not so clear hypothesis or unstructured questions.

2 Proposed vis solution

This project intends to aid physician researchers interested in data exploration by building a visualization application that will help researchers explore the dataset in a meaningful way. Using publicly available survival datasets, we will use t-SNE [Maaten and Hinton, 2008] for projecting the data to a lower dimensional manifold, which will then be plotted for visual inspection. Clustering methods such as K-means can then be used to cluster the results from t-SNE to assign cluster membership to individual observations. Plots of survival probability can then be constructed by different groups specified by the cluster membership. If any interesting patterns emerge, further exploration of clusters using vis tools such as scatterplot matrix, scagnostics, heat maps and a summary table is made available. If time permits, I would also like to add a heat map of the dataset ordered by empirical survival probability of the complete dataset.

This is in line with my work as a statistician at British Columbia Cancer Agency, where we routinely support physicians in their research. With my prior experience with dimensionality reduction and survival analysis, I believe this can be a helpful tool.

3 Usage Scenario

Figure 1 shows the rough layout of our app. A Researcher wants to explore a survival dataset. Opening the app, researcher is presented with the landing page (Figure 1-a), which has an option to upload a dataset in *csv* format (Figure 1-b). After uploading the dataset, researcher specifies which variables are *time* and *survival status* in the dataset using dropdown lists ((a2,a3), Figure 1-b) as time and status information is vital for modelling survival data. Researcher then proceeds to identify the ”id” variable (a4, Figure 1-b), so it can be excluded from any further modelling.

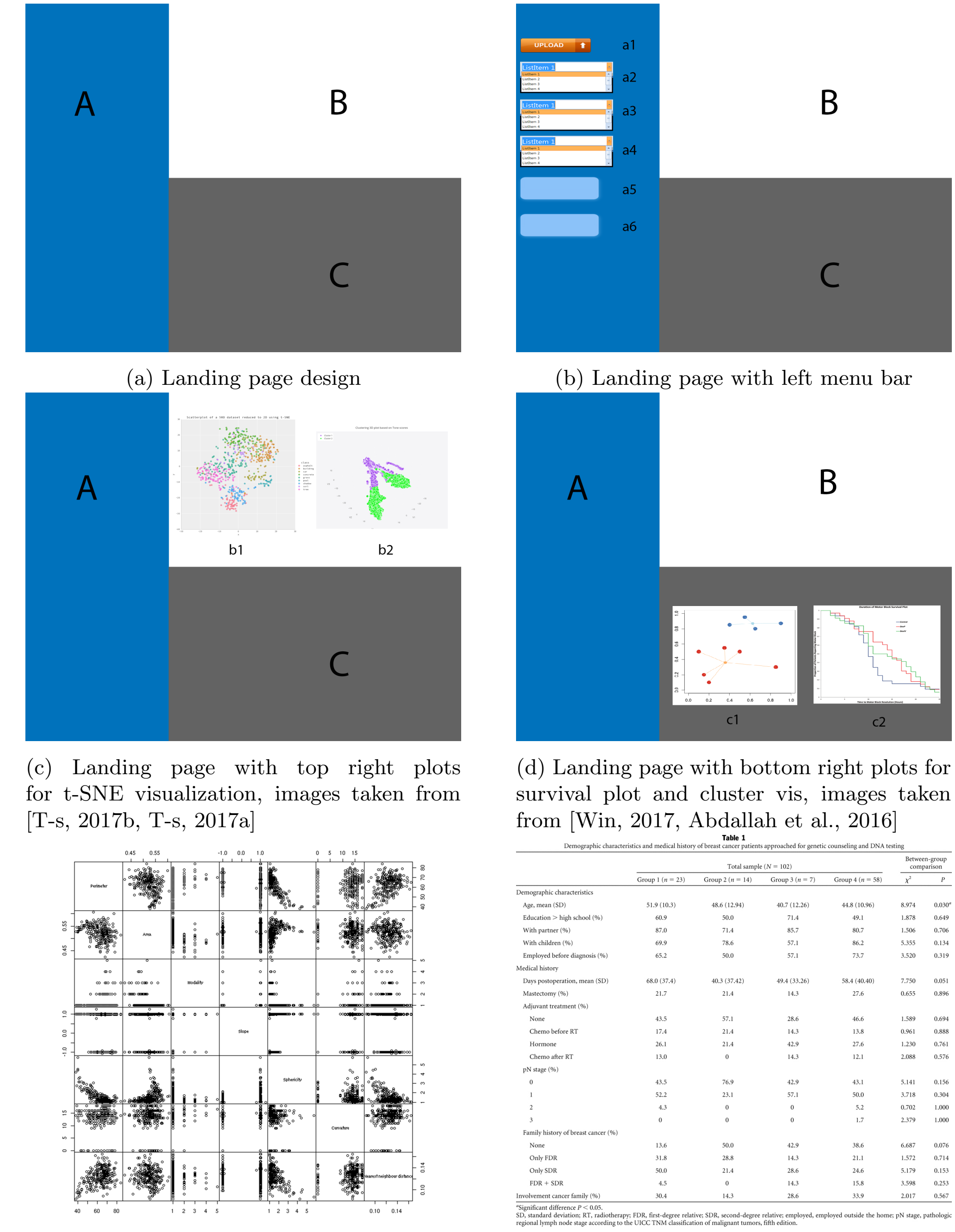
Researcher then proceeds to apply t-SNE by clicking on the button ((a5,a6), Figure 1-b). Model returns dataset dimensionality reduced to two and three dimensions, results of running t-SNE twice; once with two dimensions and once with three dimensions as required output. Returned dimensions are visualized using a two dimensional scatter plot and a three dimensional scatter plot ((b1,b2), Figure 1-c). Researcher then selects the number of groupings (clusters) required from running a clustering algorithm such as K-means on the t-SNE output to assign cluster membership to individual observations, based on t-SNE vis, researcher selects three as number of required clusters. Results of clustering along with Kaplan-Meir plot [Kaplan and Meier, 1958] of survival probability grouped by cluster membership is displayed on bottom left ((c1,c2), Figure 1-d).

After looking at the results from clustering, researcher sees an interesting pattern that needs more exploring. Using a button on the left pane ((a5,a6), Figure 1-b), researcher navigates to an ”explore clusters” page, where researcher is presented with a two sectioned layout, i.e. a small left menu pane and a large right pane. From left pane researcher has the options to select different visualizations such as scatterplot matrix, scagnostics (Figure 1-e), heat map or a summary table (Figure 1-f), all constructed based on cluster membership. Researcher decides to use summary table as it provides tabulated vis of data by cluster membership, making it easier to decipher what variables drove the cluster formation.

This app will be built using R [R Core Team, 2017] with Shiny and/or Plotly.

3.1 Milestones

- Design of landing page with working t-SNE: March 15



(a) Landing page design

(b) Landing page with left menu bar

(c) Landing page with top right plots for t-SNE visualization, images taken from [T-s, 2017b, T-s, 2017a]

(d) Landing page with bottom right plots for survival plot and cluster vis, images taken from [Win, 2017, Abdallah et al., 2016]

(e) Second page for data exploration using scagnostics or scatterplot matrix, image taken from [sca, 2017]

Table 1

Demographic characteristics and medical history of breast cancer patients approached for genetic counseling and DNA testing

	Total sample (N = 102)				Between-group comparison	
	Group 1 (n = 23)	Group 2 (n = 14)	Group 3 (n = 7)	Group 4 (n = 58)	χ^2	P
Demographic characteristics						
Age, mean (SD)	51.9 (10.3)	48.6 (12.94)	40.7 (12.26)	44.8 (10.96)	8.974	0.030[a]
Education > high school (%)	60.9	50.0	71.4	49.1	1.878	0.649
With partner (%)	87.0	71.4	85.7	80.7	1.506	0.706
With children (%)	69.9	78.6	57.1	86.2	5.355	0.134
Employed before diagnosis (%)	65.2	50.0	57.1	73.7	3.520	0.319
Medical history						
Days postoperation, mean (SD)	68.0 (37.4)	40.3 (37.42)	49.4 (33.26)	58.4 (40.40)	7.750	0.051
Mastectomy (%)	21.7	21.4	14.3	27.6	0.655	0.896
Adjuvant treatment (%)						
None	43.5	57.1	28.6	46.6	1.589	0.694
Chemo before RT	17.4	21.4	14.3	13.8	0.961	0.888
Hormone	26.1	21.4	42.9	27.6	1.230	0.761
Chemo after RT	13.0	0	14.3	12.1	2.088	0.576
pN stage (%)						
0	43.5	76.9	42.9	43.1	5.141	0.156
1	52.2	23.1	57.1	50.0	3.718	0.304
2	4.3	0	0	5.2	0.702	1.000
3	0	0	0	1.7	2.379	1.000
Family history of breast cancer (%)						
None	13.6	50.0	42.9	38.6	6.687	0.076
Only FDR	31.8	28.8	14.3	21.1	1.572	0.714
Only SDR	50.0	21.4	28.6	24.6	5.179	0.153
FDR + SDR	4.5	0	14.3	15.8	3.598	0.253
Involvement cancer family (%)	30.4	14.3	28.6	33.9	2.017	0.567

[a] Significant difference $P < 0.05$.

SD, standard deviation; RT, radiotherapy; FDR, first-degree relative; SDR, second-degree relative; employed, employed outside the home; pN stage, pathologic regional lymph node stage according to the UICC TNM classification of malignant tumors, fifth edition.

(f) Second page for data tabulation by cluster, image taken from [Schlich-Bakker et al., 2007]

Figure 1: Crude layout design for visualizing survival data

- Design of landing page with survival plots and clusters based on k-means: March 25
- Design of second page with scagnositics or scatterplot matrix: April 5
- Design of second page with summary table based on cluster membership: April 10
- Further modifications/additions/test runs/bug reports: April 20

4 Related work

T-Distributed Stochastic Neighbor Embedding (t-SNE) [Maaten and Hinton, 2008] is a dimensionality reduction technique suitable for high dimensional data visualization, it works by constructing a probability distribution on pairs of high dimensional objects in a way that similar objects have high probability of getting selected. Several versions of t-SNE have been proposed since, including Barnes-Hut-SNE [Van Der Maaten, 2013], which is a faster version.

T-SNE has been used to identify prognostic tumor sub-populations using mass spectrometry imaging data [Abdelmoula et al., 2016] and to visualize SNPs [Platzer, 2013].

References

[sca, 2017] (2017). Blog grammar of graphics. http://zoonek.free.fr/blosxom/R/2006-08-27$_{T}he_{G}rammar_{o}f_{G}raphics.html.Accessed : 2017 - 03 - 05$.

[Win, 2017] (2017). Origin lab km estimator. http://www.originlab.com/doc/Tutorials/Kaplan-Meier-Estimator. Accessed: 2017-03-05.

[T-s, 2017a] (2017a). Plotly 2d. http://blog.applied.ai/visualising-high-dimensional-data/. Accessed: 2017-03-05.

[T-s, 2017b] (2017b). Plotly 3d. https://plot.ly/ ahajibagheri/27.embed. Accessed: 2017-03-05.

[Abdallah et al., 2016] Abdallah, F. W., Dwyer, T., Chan, V. W., Niazi, A. U., Ogilvie-Harris, D. J., Oldfield, S., Patel, R., Oh, J., and Brull, R. (2016). Iv and perineural dexmedetomidine similarly prolong the duration of analgesia after interscalene brachial plexus blocka randomized, three-arm, triple-masked, placebo-controlled trial. *The Journal of the American Society of Anesthesiologists*, 124(3):683–695.

[Abdelmoula et al., 2016] Abdelmoula, W. M., Balluff, B., Englert, S., Dijkstra, J., Reinders, M. J., Walch, A., McDonnell, L. A., and Lelieveldt, B. P. (2016).

Data-driven identification of prognostic tumor subpopulations using spatially mapped t-sne of mass spectrometry imaging data. *Proceedings of the National Academy of Sciences*, page 201510227.

[Kaplan and Meier, 1958] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

[Maaten and Hinton, 2008] Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

[Platzer, 2013] Platzer, A. (2013). Visualization of snps with t-sne. *PloS one*, 8(2):e56883.

[R Core Team, 2017] R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Schlich-Bakker et al., 2007] Schlich-Bakker, K. J., ten Kroode, H. F., Wárlám-Rodenhuis, C. C., van den Bout, J., and Ausems, M. G. (2007). Barriers to participating in genetic counseling and brca testing during primary treatment for breast cancer. *Genetics in Medicine*, 9(11):766–777.

[Van Der Maaten, 2013] Van Der Maaten, L. (2013). Barnes-hut-sne. *arXiv preprint arXiv:1301.3342*.