Understanding Health Expenditure through Cluster Analysis and Naive Bayes Classification.

Anthony Lighterness

02-05-2020

Abstract

Healthcare is not only one of the most resource-demanding industries globally, but it also produces some of the most highly dimensional and heterogenous data. With the rise in global health expenditure, the ability to utilise data mining techniques effectively to extract actionable insights and highlight predictable patterns related to cost has never been more important. The current report seeks to emphasise the challenges of unsupervised machine learning for describing real-world, heterogenous medical data. Specifically, the study attempts to identify the relationship between cost and a range of attributes, such as diagnosis, treatment, and length of stay from both unsupervised and supervised perspectives. To do this, a Gower distance matrix followed by principal coordinate analysis (PCoA) and t-stochastic neighbour embedding (tSNE) were performed to visualise partition around medoids (PAM) cluster models in lower dimensional spaces, which were then compared to Wards hierarchical clustering and then lastly to a supervised Naive Bayes (NB) classifier. Visual analysis overall showed that both PAM and Wards hierarchical clustering were able to group some clusters of patients with similar cost outcomes in an unsupervised protocol. The NB model produced 85% and 75% overall accuracy with an AUC of 0.92 and 0.87 when cross validating used 30% of the dataset split and a 10-fold protocol, respectively. The study suffered from the curse of dimensionality during analysis, but is able to showcase the challenges of unsupervised machine learning when attempting to describe nuanced, real world medical data.

1 Introduction

Amidst the escalating data revolution comes the need to understand large, complex, and information rich datasets in all fields of technology, business, and science. Health data is no exception to this as it exemplifies the challenges that come with analysing highly dimensional and heterogenous data. With the growing concerns in global health expenditure being ~10% of the total gross domestic product for most highly-income countries such as Australia and the UK, and the rise in ageing populations and chronic disease, comes the need to extract actionable insights hidden within medical data (Mikulic, 2019). Data mining techniques including unsupervised clustering and supervised classification can be used to describe and predict patterns that may highlight the contributing factors to public health expenditure.

Datasets are heterogeneous in nature when multiple datatypes exist, such as continuous, discrete, and multi-level factor variables. While the literature widely reports the use of homogenous, quantitative datasets, mixed data is an undeniable reality in today's technologically driven society. From an unsupervised learning perspective, analysts face a myriad of challenges when searching for relevant patterns in highly dimensional and heterogenous data. One of the only solutions for this is to compute the Gower distances between objects, thus allowing observations of similar attributes to exist in clusters. Gower's method is unique in that it can measure the (dis)similarities between observations that are described by any data types, including categorical, continuous, or multi-level factor variables. As such, it is consistently supported in the literature when preparing complex datasets for partitional and hierarchical cluster analysis (Akay & Yuksel, 2017; Gower, 1966; Gower, 1971).

While Gower's (dis)similarity metric overcomes the challenge of heterogenous data, it does not solve the curse of dimensionality. Therefore, the literature shows its use in combination with a dimensionality reduction technique (Belkina et al. 2018; Filaire, 2018; Martin, 2016). Principal component analysis (PCA) is the most widely known and applied tool for dimensionality reduction. Its prevalence dominates the literature such that very few describe its counterpart, principal coordinate analysis (PCoA) – most likely due to the bias against mixed type datasets. Gower (1966) suggested the name PCoA, also known as classical multidimensional scaling (CMDS), to distinguish it from PCA. While both techniques lead to the same solution, PCA is based on a *qxq* matrix of associations between variables, whereas PCoA is based on an *nxn* matrix of (dis)similarities between observations (Gower, 1966).

While PCoA is more adept at preserving global structure of the data, t-stochastic neighbour embedding (tSNE) is another dimensionality reduction tool that better represents local interactions (Nguyen & Holmes, 2019). As such, the literature reports tSNE to be effective for visualising k-protocol cluster models in lower dimensional spaces (Belkina et al., 2018; Filaire, 2018; Martin, 2016). However, it is cautioned that tSNE does not preserve long-range interactions between data points and therefore may generate visualisations in which the arrangement of non-neighbouring groups of observations is not informative (Nguyen & Holmes, 2019). It is however, supported by the literature in the use for visualisation of unsupervised clustering models such as partition around medoids (PAM) (Belkina et al., 2018; Filaire, 2018; Martin, 2016).

Unsupervised machine learning functions are evaluated based on their ability to describe patterns within unlabelled data (Jothi et al., 2015). Its counterpart domain, supervised machine learning, can be more quantitatively specific. Indeed, predictive models are trained using a target variable with known labels to then predict and classify new incidences of the same variable (Jothi et al., 2015). Its evaluation therefore produces quantitative metrics, such as ROC-AUC and overall accuracy that describe its ability to correctly predict new labels (Vihinen, 2012). The current report centres its discussions and analyses around two examples of each of supervised and unsupervised learning functions.

Cluster models aim to group data into sets such that intra-cluster similarity is maximised while inter-cluster similarity minimised (Akay & Yuksel, 2017). These can be divided into two categories: partitional and hierarchical, where the latter can be further classified into agglomerative and divisive algorithms (Akay & Yuksel, 2017). The current investigation concentrates on PAM and Ward's hierarchical clustering – an agglomerative type. While Ward's method outperforms its hierarchical counterparts, PAM is overshadowed in the literature by its quantitative equivalent, k-means (Akay & Yuksel, 2017; Park, 2009). Despite this, PAM is reportedly more rigorous against outliers than k-means and is able to handle distance matrices computed from complex mixed type datasets (Budiaji & Leisch, 2019). This is because in PAM models, the centre of a cluster is identified as the object in that specific cluster that lies closest to all other objects (Budiaji & Leisch, 2019). Therefore, it is less impacted by outliers compared to k-means. Agglomerative hierarchical algorithms, in contrast, start by merging smaller clusters until one large cluster is left (Akay & Yuksel, 2017). Although PAM is known to be the most powerful algorithm for k-medoids, its main shortcoming is that it works inefficiently for large datasets due to its time complexity being $O(k(n-k)^2)$ (Park & Jun, 2009). Due to the limitations of unsupervised techniques, researchers favour reporting of supervised machine learning, particularly in the medical domain (Johti et al., 2015).

Supervised machine learning is much more popular for many reasons, including seamless application, reliable assessment, and utility in predictions (Jothi et al., 2015). One of the most common tasks in supervised learning functions is that of classification (Jothi et al., 2015). A widely used framework for classification is provided by a simple theorem of probability known as Bayes' rule (Lewis, 1998). Bayesian classifiers take into account all available evidence from mixed data-type explanatory variables to make final predictions and provide transparent explanations of these outcomes (Jothi et al., 2015; Al-Aidaroos et al., 2012). This process is natural and familiar to the decision-making protocols in medicine, ranging from patient diagnosis to billing (Al-Aidaroos et al., 2012). Moreover, it is simple, computationally efficient, and naturally robust to missing and noise-filled data. As such, it is clear why Naïve Bayes is one of the most studied supervised algorithms in the literature.

The current investigation sought to explore the utility of unsupervised clustering on complex, real hospital data. Specifically, we asked if it is possible to identify similar patterns of health cost when patients are grouped by their (dis)similarities across a range of attributes other attributes. We then wanted to compare these findings with the predictability of cost using the same variables applied in a supervised Naïve Bayes (NB) classifier. The study aims to highlight the challenges of pattern analysis from an unsupervised machine learning perspective using real world, heterogeneous data.

2 Data

The dataset of interest was sourced privately through a contact from St Peter's hospital in Surrey, United Kingdom (NHS Ashford and St Peter's Hospital, 2014). The data was collected by the National Health Service (NHS) during an observational study from 2010 to 2011. Initially, 28 data items and 14879 observations of patients is presented. To the best of our knowledge, no known interventions or pre-processing were applied to this data. However the current study pursued extensive pre-processing so that only 10 variables were selected for analysis. Briefly, NA values were omitted and two variables were calculated, including *Length.of.Stay* and *RTT.Period*. The relevant variables with their descriptions are summarised below in Table 2.1 (NHS, 2010).

Table 2.1: Data items and definitions.

Variable	Data.Type	Description
Adm.Type	Factor, 3 levels	Admission of non-elective, elective day case or in-patient.
HRG4	Factor, 881 levels	Grouping of procedure codes.
ICD10	Factor, 1899 levels	International classification of diseases code.
OPCS4	Factor, 1223 levels	Classification of interventions and procedures.
PbR	Binary	Payment by result received.
Patient.Age	Numeric, Discrete	Age of patient in years.
Specialty.Code	Factor, 45 levels	Code of main specialty involved.
Length.of.Stay	Numeric, Discrete	Difference between admission and discharge dates.
RTT.Period	Numeric Discrete	Perior from GP referral to commencement of treatment.
Cost.Class	Factor, 6 levels	Discretized cost levels.

3 Methods

R Studio was used to perform all pre-processing and analyses in the current report (R Studio, 2016). Overall, the current study sought to identify the challenges of knowledge discovery of health data from unsupervised and supervised perspectives. To commence unsupervised clustering, a real-world hospital dataset containing multi-level factor, categorical, and discrete numeric variables, was prepared by computing its Gower distances without the cost level variable (Muhaimin, 2018). This was achieved using the `daisy()` function including the Gower metric. Before clustering the distances with PAM, a silhouette width plot at various k values, ranging from 2 to 8, was plotted to select an appropriate number of clusters (Muhaimin, 2018; Kaoungky et al., 2018; Murugananthi & Ramyachitra, 2014). This led to the selecting k-values of 3 and 6.

To visualise the PAM models' ability to cluster a sizeable Gower distance matrix, two dimensionality reduction techniques were used, including tSNE, and PCoA (Belkina et al., 2018; Filaire, 2018; Martin, 2016). For tSNE, the `Rtsne()` function was used and adjusted for *is_distance=TRUE*. This output was visualised using `ggplot()` in multiple plots for comparisons. The `geom_point()` colour aesthetic was used in one plot to highlight points that were clustered by the PAM model, and in another to highlight the true cost label to see

the effectiveness of this protocol in clustering similar patients' cost outcomes. The same procedure was repeated, using the PCoA protocol. However, to perform PCoA, `cmdscale()` was used with *eig=TRUE* to enable calculation of the variance explained by the two principal co-ordinate dimensions.

The group memberships set by PAM were compared with Ward's hierarchical method. To perform Ward's hierarchical clustering, the Gower distance matrix was applied to the `hclust()` function with the *ward.D2* metric. Next, the resulting dendrogram was visualised, using `as.dendrogram()`, alongside four colour-coded bars below its leaflets. These represent the colours of observations clustered by Ward's method, PAM at K=3 and K=6, and the true cost label. Together, this allowed visual evaluation of the clustering models' ability to group patients of similar attributes, and to see if these grouped observations are similar to the cost label. Ward's method and PAM (K=6) group membership of observations were also compared in frequency crosstabs, which were plotted in heatmap visuals using `geom_tiles()`.

Finally, the current study sought to compare the unsupervised learning pathway with a supervised model on the same dataset. To achieve this, a NB classifier was trained in duplicate; the first using a typical 70:30 split for training and testing validation, and the second using a 10-fold cross validation protocol. Laplace smoothing was applied as some variables' factors may not have been represented in the training subset. A confusion matrix for both models was calculated using the `confusionMatrix()` function. Furthermore, the `multiclass.roc()` and `plot.roc()` functions were combined in a custom function to visualise the ROC-AUC plots. Lastly, the `varImp()` function was used on the output of the 10-fold cross-validated NB model to identify the contributions of the predictor variable to predict the cost. Overall, this allowed adequate assessment and comparisons of unsupervised clustering and supervised classification on real hospital data, thereby highlighting the challenges that come with these methods.

4 Results and Discussion

Overall, the results show some congruency and feasibility between the different unsupervised cluster and NB classifier models. Firstly, to validate the resulting Gower distance matrix, the output distances were cross-referenced with the raw data to find two pairs of observations deemed most (dis)similar. The result of this query, as seen in Table 4.1 below, indicate that the Gower method worked well in calculating the distanced between mixed type attributes. With a validated Gower distance matrix, PAM and Ward's hierarchical clustering were performed.

Table 4.1: Pairs of most similar and dissimilar observations.

	Adm.Type	HRG4	ICD10	OPCS4	PbR	Patient.Age	Specialty.Code	Length.of.Stay	RTT.Period
1931	2	HB23B	M232	W822	1	43	110	0	63
237	2	HB23B	M232	W822	1	43	110	0	64
13517	3	WD22Z	F250		0	53	710	324	0
1457	2	SA13Z	D561	X339	1	10	420	0	289

Note:
Pairs of observations that are most similar (gold) and most dissimilar (blue).

To visualise the PAM clustering models in lower dimensional spaces, the Gower distance matrix was applied to both tSNE and PCoA separately. To select an appropriate number of clusters for PAM, a Silhouette width plot over a range of 2 to 8 clusters was visualised. As seen in Figure 4.1 below, the optimal K-values for PAM are 3. However, both 3 and 6 were included in this study as it was hypothesised that 6 clusters of PAM may relate more closely to the 6 levels of the cost variable. The 6-level cost variable was then applied as a colour aesthetic for the PAM cluster scatterplots, which can be seen below in Figure 4.2.

The results of the PAM clustering scatterplots show that the data is somewhat clusterable. In Figure @(fig:clustervisuals) below, PAM (k=6) is visualised using t-SNE and PCoA with the true cost label. Clusters of similar colours indicate similar cost ranges, meaning that PAM was able to group patients based on a select range of attributes that then are in fact somewhat associated with similar costs. Although the two dimensions of PCoA only account for approximately 30% of the total variance, clusters of similar cost-related colours are reliably distinct from one another. It appears that PCoA produces more distinguishable clusters than t-SNE in this particular outcome. The outcomes of PAM (k=3) and scatterplots without the class label can be seen in Appendix 7.1 below, which show that no observable differences in PAM (k=3) and PAM (k=6) as seen when the true cost label is applied.

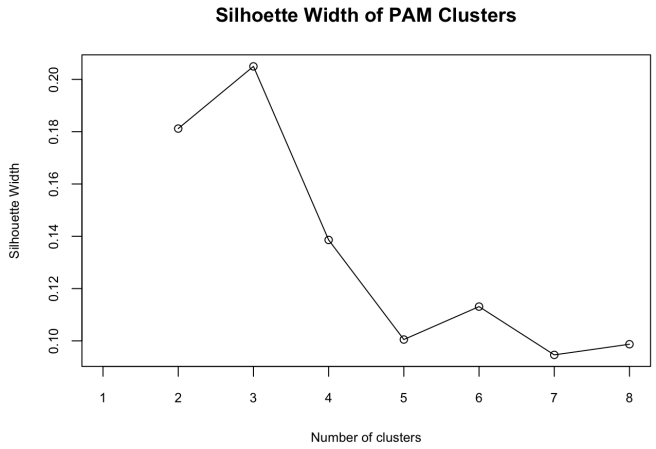

Figure 4.1: Silhouette width of PAM at different k values.

Although PCoA reduced the distance matrix to two dimensions that explain collectively approximately only 31% of the variance, it was sufficiently effective in preserving the local distances between observations. Indeed, the most interesting outcome overall is seen when comparing subplots D and F, where one can identify clusters of similar cost labels. This means that without the cost label, PAM, visualised using either tSNE or PCoA, is able to cluster patients of similar attributes together, which then in turn pertains to specific costs. This observation concurs with the hypothesis that the select attributes collectively contribute to a specific cost outcome as they describe a patient's pathway from medical diagnosis to treatment. As such, these preliminary findings show that this complex dataset is describable using PAM coupled with tSNE or PCoA.

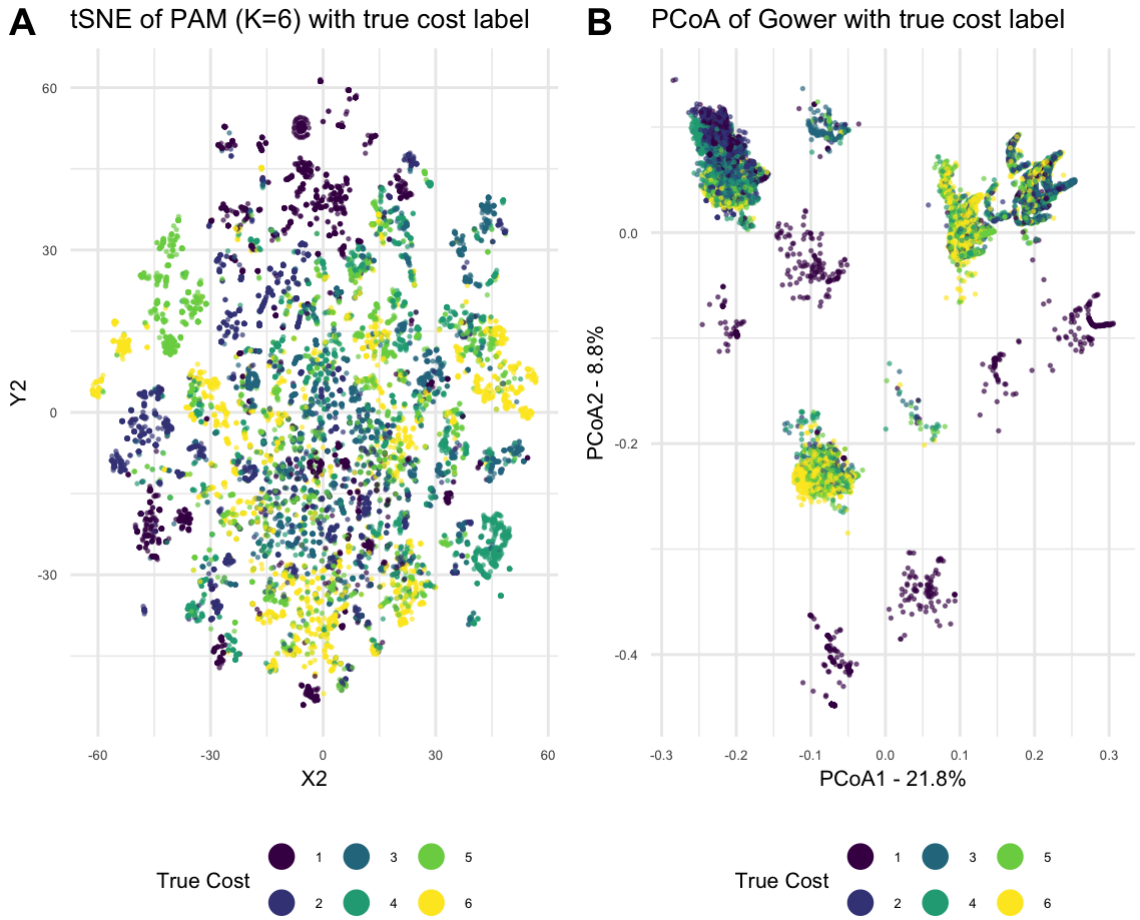

Figure 4.2: Dimensionally reduced solutions from PCoA and tSNE to visualise PAM clusters with true cost label.

To further assess the clusterability of this dataset, Ward's method of hierarchical clustering was applied to compare with the PAM model. The group memberships of Ward's method and PAM were compared with one another alongside the true cost class. These results are visualised as heatmaps, as seen in Figure 4.3 below. The frequencies of group memberships differ somewhat when the cluster models are compared with the true cost class. Both PAM and Ward's method showed some exclusive observations grouped nearly exclusively into cluster #6 when compared to the true cost label. Ward's cluster number 4 has only members of cost class 1. However, its cluster numbers 2 and 3 contained costs of all levels. Similarly, PAM's cluster numbers 1, 2, and 4 also contained members of all cost levels to varying extents. This shows the difficulty in unsupervised clustering.

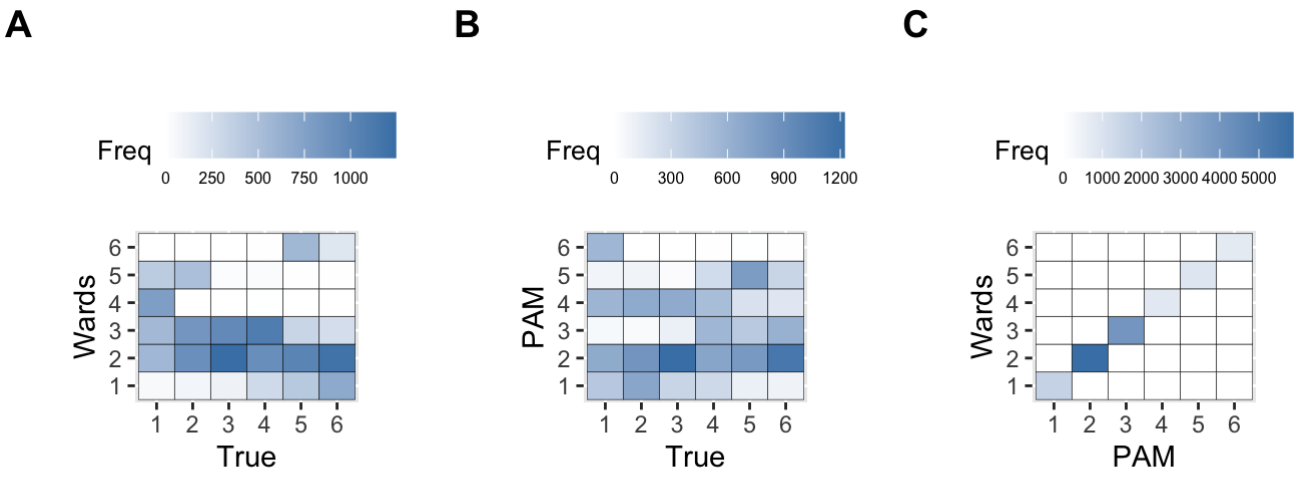

Figure 4.3: Frequencies of group memberships across clusters defined by Ward's Hierarchical and PAM compared with the ground truth.

The aforementioned results concur with Ward's dendrogram. In Figure 4.4 below, the colour-coded group memberships of PAM at k=3 and k=6, Ward's method, and the true cost label are compared. Overall, it can be seen that both PAM and Ward's method share similar clusters of groups. However, Figure 4.4 clearly shows the contrast in colours of the true cost label, indicating that observations of different class levels were clustered together. This was seen also in Figure 4.3 above. Upon close inspection, it can be noted that some groups of similar cost levels can be seen, which means that Ward's method was able to cluster some observations close to their real cost levels in unsupervised. PAM (k=6) also follows some patterns with the grouping seen in the true label, suggesting that the clusters made by Ward are similar to those in both PAM models. As such, the cluster assessment overall shows somewhat satisfactory unsupervised pattern recognition of the data. Although this assessment and analysis is somewhat subjective, the results are impressive considering patterns are observable looking at a highly dimensional and heterogeneous dataset using an unsupervised approach.

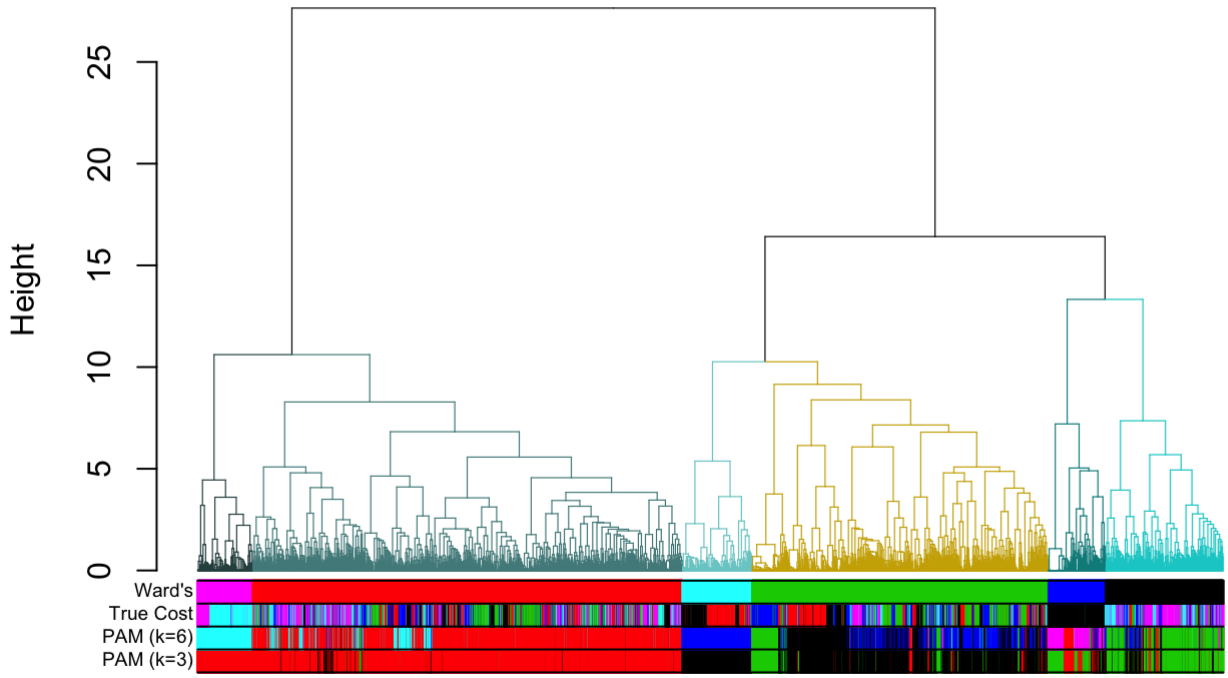

Figure 4.4: Ward's hierarchical clustering dendrogram with three colour-coded bars for label and cluster comparisons.

To compare the efficacy of using unsupervised clustering in deriving patterns from the current dataset, an NB classifier was applied. Firstly, the supervised learning protocol was significantly faster, simpler, and more reliable in its assessment. The main outcome, shown in Figure 4.5 below, was that when trained and validated using a typical 70:30 split of the dataset, the NB classifier produced an AUC of 0.921 and an impressive average predictive accuracy of ~85%. When cross-validated by a 10-fold scheme, an AUC of 0.87 and accuracy of 76% was achieved. Overall, this is considered a success of predictive power, especially considering that the cost variable was discretized into a 6-level factor, which would have led to some loss of information and precision (Quinn, 2004).

The NB classifier was also able to show the variable contributions in predicting each level of the cost variable, which is visualised in Figure 4.6. Interestingly, Length.of.Stay is most relevant in the mid range cost of £660-1,490, but least important in the lowest cost range. Furthermore, HRG4 is most important for predicting the lowest and highest cost levels of £0-529 and £2,840-167,000. This variable is only second to PbR for the

remaining levels. Collectively, these results show the predictability of cost changes at different increments and is affected by different variables at these varying levels. Therefore, it is clear why these complex relationships would be difficult to reveal using unsupervised clustering techniques. Overall, not only did the NB classifier show satisfactory predictability of cost, but it also offered a glimpse into the intricate relationships between the explanatory variables at different levels. This further clarifies why it is difficult to rely solely on PAM and Ward's hierarchical clustering to describe patients' cost from an unsupervised perspective.

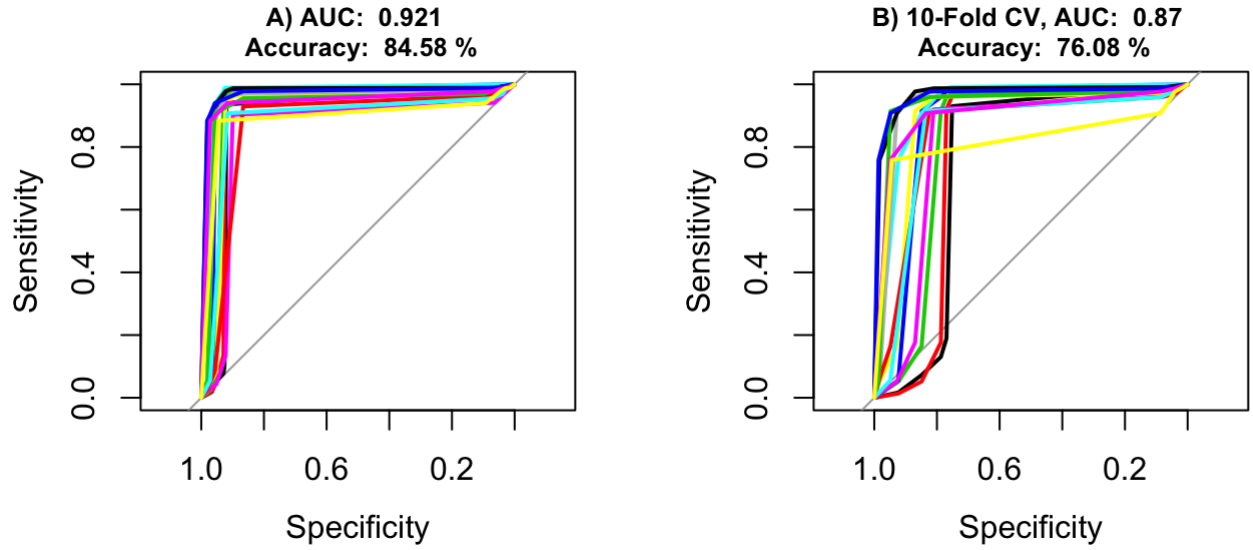

Figure 4.5: ROC-AUC and accuracy metrics for predicting each cost level using two Naive Bayes models.

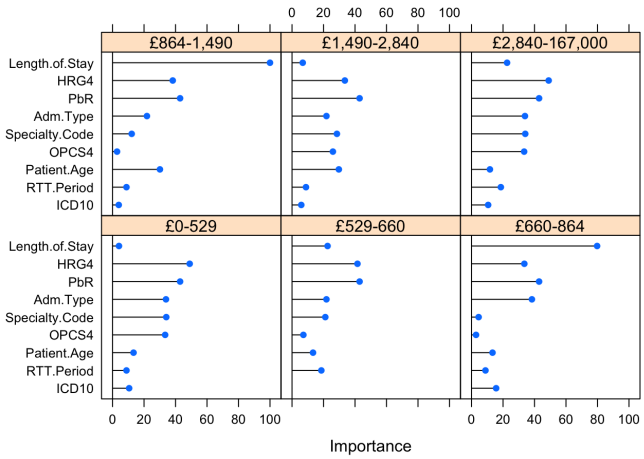

Figure 4.6: Variables' contributions in predicting each cost level using Naive Bayes with 10-fold cross-validation.

5 Conclusion

In conclusion, the current investigation aimed to shed light upon the challenges and feasibility of using unsupervised clustering on real world, sizeable and heterogeneous hospital data. The computation of a Gower distance matrix allowed the application of PAM and Ward's hierarchical clustering of mixed type data, which was visualised in lower dimensions using t-SNE and PCoA. While PCoA showed clear distinctions between clusters indicated by the cost label, it was still difficult to identify clear relationships even in conjuction with Ward's hierarchical clustering. One significant shortcomming in the current study is that processing the Gower distance matrix and PCoA is nearly impractical due to excessive computation time required. Although the study was also limited in quantitative assessment of the clustering techniques, by introducing the cost label following PAM and Ward's clustering, we were able to show that these methods could visualise some groups of similar cost levels. Furthermore, an NB classifier was able to show complex relationships between the predictor variables and cost at an overall accuracy ranging from 75-85%, which is considered a success given the complex nature of the dataset in question.

6 References

Akay, O. and Yuksel, G. (2017). Clustering the Mixed Panel Dataset using Gower's Distance and K-Prototypes Algorithms. Communications in Statistics – Simulation and Computation. Doi: 10.1080/03610918.2016.1367806

Al-Aidaroos, K. M., Bakar, A. A., & Othman, Z. (2012). Medical Data Classification with Naive Bayes Approach. Information Technology Journal. doi:10.3923/itj.2012 (doi:10.3923/itj.2012)

Belkina, A.C., Ciccolella, C.O., Anno, R., Halpert, R., Spidlen, J., Cappione, J.E.S. (2019). Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. Nat Commun 10, 5415. https://doi.org/10.1038/s41467-019-13055-y (https://doi.org/10.1038/s41467-019-13055-y)

Budiaji, W., & Leisch, F. (2019). Simple K-Medoids Partitioning Algorithm for Mixed Variable Data. Algorithms, 12(177). doi:10.3390/a12090177 (doi:10.3390/a12090177)

Filaire, T. (2018). Clustering on mixed type data. Retrieved from https://towardsdatascience.com/clustering-on-mixed-type-data-8bbd0a2569c3 (https://towardsdatascience.com/clustering-on-mixed-type-data-8bbd0a2569c3)

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika, 53(3-4), 325-338. doi:10.1093 (doi:10.1093)

Gower, J.C. (1971). A general coefficient of similarity and some of its properties. Biometrics, 27:4, pp. 857-874.

Jothi, N., Rashid, N. A. A., & Husain, W. (2015). Data Mining in Healthcare – A Review. Procedia Computer Science, 72, 306-313. doi:10.1016/j.procs.2015.12.145 (doi:10.1016/j.procs.2015.12.145)

Kaoungku, N., Suksut, K., Chanklan, R., Kerdprasop, K., & Kerdprasop, N. (2018). The Silhoette Width Criterion for Clustering and Association Mining to Select Image Features. International Journal of Machine Learning and Computing, 8(1), 69-73. doi:10.18178/ijmlc.2018.8.1.665 (doi:10.18178/ijmlc.2018.8.1.665)

Lewis, D.D. (1998). Naïve (Bayes) at Forty: The Independence Assumption in Information Retrieval. ECM '98: Proceedings of the 10th European Conference on Machine Learning, pp. 4-15.

Martin, D.P. (2016). Clustering Mixed Data Types in R. Wicked Good Data, from https://dpmartin42.github.io/posts/r/cluster-mixed-types (https://dpmartin42.github.io/posts/r/cluster-mixed-types)

Mikulic, M. (2019). Health expenditure as a percentage of GDP in select countries 2018. Retrieved from https://www.statista.com/statistics/268826/health-expenditure-as-gdp-percentage-in-oecd-countries/ (https://www.statista.com/statistics/268826/health-expenditure-as-gdp-percentage-in-oecd-countries/)

Muhaimin, M. (2018). Clustering categorical and numerical datatype using Gower distance. Medium from https://medium.com/@rumman1988/clustering-categorical-and-numerical-datatype-using-gower-distance-ab89b3aa90d9 (https://medium.com/@rumman1988/clustering-categorical-and-numerical-datatype-using-gower-distance-ab89b3aa90d9)

Nguyen, L.H., and Holmes, S. (2019). Ten quick tips for effective dimensionality reduction. PLOS Computational Biology, 15(6). Doi: 10.1371/journal.pcbi.1006907

NHS Ashford and St. Peter's Hospital. (2014, May 20). St Peter's Hospital. Retrieved from Ashford St. Peter's NHS: http://www.ashfordstpeters.nhs.uk/st-peter-s-hospital/82-st-peter-s-hospital (http://www.ashfordstpeters.nhs.uk/st-peter-s-hospital/82-st-peter-s-hospital)

NHS. (2010). Main Specialty Code. Retrieved from Data Dictionary NHS: https://www.datadictionary.nhs.uk/data_dictionary/attributes/m/main_specialty_code_de.asp?shownav=1 (https://www.datadictionary.nhs.uk/data_dictionary/attributes/m/main_specialty_code_de.asp?shownav=1)

Park, H.S. and Jun, C.H. (2009). A simple and fast algorithm for K-medoids clustering. Expert Systems with Applications, 36, pp. 3336-3341. Doi: 10.1016/j.eswa.2008.01.039

Quinn, K.M. (2004). Bayesian factor analysis for mixed ordinal and continuous responses. Political Analysis, 12: 4. Pp. 338-353. Doi: 1093/pan/mph022

RStudio Team (2016). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL

http://www.rstudio.com/ (http://www.rstudio.com/).
Vihinen, M. (2012). How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. BMC Genomics, 13. doi:10.1186 (doi:10.1186)

7 Appendix

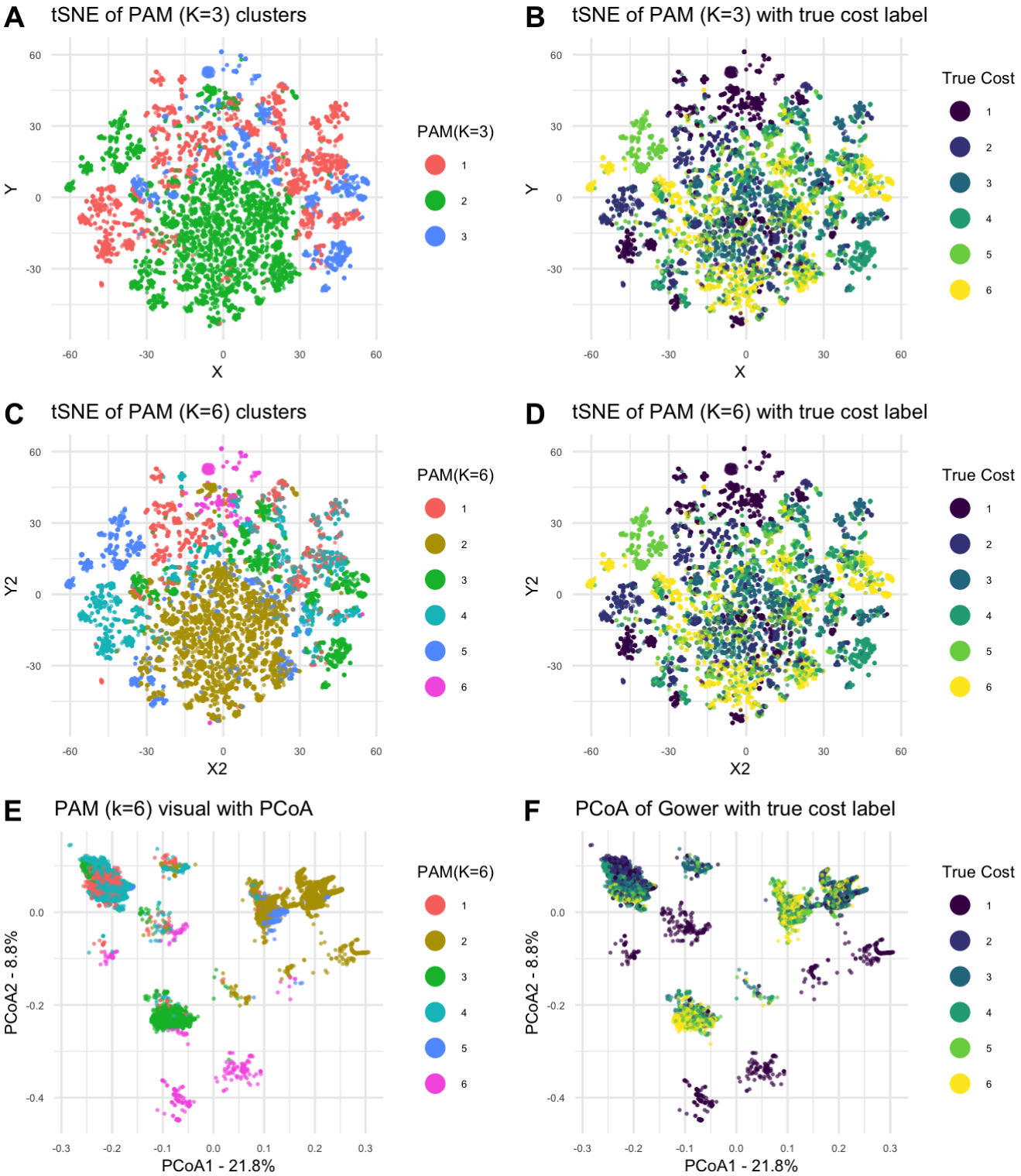

Figure 7.1: Dimensionally reduced solutions from PCoA and tSNE to visualise PAM (k=3 and 6) clusters compared to true cost label.