Learning from Data: Cleft Lip and Palate Patients in the West Coast of Sabah

Z A Omar¹, S N Chin², Norhafiza Hamzah¹ and Fouziah Yassin³

¹ Mathematics with Computer Graphics Programme
² Mathematics with Economy Programme
³ Physic with Electronics Programme

Faculty of Science and Natural Resources, Universiti Malaysia Sabah, 88400 Kota Kinabalu Sabah, Malaysia

Email: zatur@ums.edu.my

Abstract. Analysing data can be quite a challenge sometimes due to the nature of the data and the vast options of methods and techniques that can be used on the data. In this study, for example, a six years Cleft Lip and Palate dataset were gathered on these patients’ conditions in the quest to identify the contributing factors for a successful pre-graft orthodontic treatment. The challenges faced was in the small number of datasets and imbalance sample class. Therefore, this study had taken a step back and tried to approach the dataset with a combination of unsupervised and supervised learning methods to tackle the challenges by incorporating clustering – for testing records creation and; resampling – for balancing sample class. We also observed if the auto-created testing records are replaceable with the manually selected testing records by looking at the performances of the classification models. Based on the feature that was selected, k-Means and PAM were implemented as the clustering algorithm using the Euclidean formula as the distance measure. Resampling was done using SMOTE and Random Forest as the classification model. When the comparison was done on the models, the ones that were fed by resampled training records showed an increase in the AUC values and decrease in the OOB error. Comparable results were also achieved between the training records produced by PAM and by manual selection as both models, based on the AUC values, was classified as excellent classification models.

1. Introduction

Cleft Lip and Palate (CLP) is an orofacial abnormality that occurs roughly 1 in 700 live births worldwide and the rate varies upon regions [1]. The CLP is a condition where a cleft (gap) occurred at the upper lip and the upper mouth of these babies. Normally, the affected babies will be committed to different treatments depending on the severity of their CLP conditions and these conditions can be read further here [1]. If the cleft had affected the child’s alveolar bone, they might have to undergo a secondary alveolar bone graft (SABG) surgery to fill up the gap in the gum so that proper teeth eruption could occur. In most cases, pre-graft orthodontic (PGO) treatments were required in preparation for the SABG surgery.

Conducting the SABG surgery at the right time is critical. The normal adopted practice is, during the mixed dentition stage before the eruption of the permanent canine [2]. Therefore, some studies find
that the right age for SABG surgery is between the age of 9 to 12 years old [2]. This implies that, ideally, the PGO treatment should end right before the child is 9 to 12 years old. In real life, however, this is usually not the case.

Take for example the dataset taken from the Specialised Orthodontics Clinic of Queen Elizabeth Hospital at Kota Kinabalu, Sabah. This is a six-year CLP patient record taken from Jan 2010 to December 2015. It consisted of 33 records with various CLP conditions based on the cleft lateral (unilateral or bilateral), the completeness of the cleft lateral (complete or not complete) and the affected palate (soft or hard or both). Information on the age of the patient first referred to the clinic (ageR) and the age when they did the surgery (ageP) were also recorded along with their gender and ethnicity [2, 3].

Based on these two age categories shown in the boxplot (figure 1), the youngest patient referred (ageR) was at the age of 5 and the oldest was 25 and half of them were referred after the age of 11 years old. Depending on the success of the PGO treatments, the youngest patients went through the SABG surgery (ageP) was at the age of 8 and the oldest at 27. The median for ageP was also 11 years old indicating the same situation where 50% of the patients, went through the SABG surgery after the age of 11. There were also a couple of cases where the age of the surgeries was out of the normal range.

![Figure 1. Boxplot of CLP Patients age first reviewed (ageR) and age at SABG surgery (ageP).](image)

Realizing that the duration of the PGO treatments may help to expedite the timing for the surgery, we had investigated the dataset and tried to identify the patients’ significant conditions that might contribute to what we had defined as a success (the duration that took less than a year) of the PGO treatments. The challenges in both studies were, 1) the number of training records was small and 2) the training records were imbalanced. One of the inclusion criteria by expert’s judgement was to only include patients that were first referred below the age of 14 years old (based on the current patients’ data and mixed dentition age is around 6-13 years old [4]). That criterion had further reduced the training records to only 23 with 15 success cases and 8 failures.

Using the reduced training records, in the first study [2], we tackled the problem as a classification problem hence producing a decision tree based on the C4.5 algorithm [5]. The produced decision tree chose the affected cleft palate (acp) as the root node and produced kappa value = 0.5148 after 10-fold cross-validation. The tree suggested a modest classification model but not up to the requirement of a clinical study (of kappa > 0.8). We then produced an assemble of decision trees (random forests) [3] to explore the contributing features with even less training records (18) to balance out the number of cases in each class. The out-of-bag (OOB) build in error checking mechanism was used to assess the goodness of fit of the produced random forests model [6]. The results showed an improved model (based on the decrease in OOB error) when only the important variables were used. When we try to eliminate the chances of variable selection bias, the conditional forest (cforest) [7] indicated that only the acp variable is important.

In this paper, we have taken yet another approach in tackling the challenges that were faced before by incorporating clustering, data resampling and random forests. Clustering will be based on ageR and
age $P$ and cluster that has a higher average value of age $P$ will be excluded from the training records. What we try to observe from this study is, if the chosen produced records from the clustering algorithm would be comparable with the manual records selection. If that was the case, the processes can easily be automated and might be suitable to be implemented in a future dataset. The approach of this study is discussed further in the following section followed by comparative findings of random forests model when the fed training data was produced by the clustering algorithms and the manually selected records. Since the approach had incorporated many well-established methods, each method will be informed briefly and further information about the methods can be followed from the given references.

2. Methodology

2.1. Clustering

Clustering falls under the unsupervised learning problems that can be roughly defined as an algorithm that inspect the intrinsic characteristics of each data points and try to group homogenous data together while making sure each group are unrelated from one another [8, 9, 10, 11]. Such grouping helps to organise and uncover hidden structures of the data as well as its natural grouping [9]. There are two approaches to the cluster which are hierarchical and partitional. For the purpose of our study, we had selected k-Means and Partitioning around Medoid (PAM) as the clustering algorithms that fall under the partitional approach. The decision was mainly due to the nature of the features we had selected to be clustered and the number of records available. Studies done by Jain et.al [8, 9] are some of the most cited when it comes to this topic and we would like to direct readers to their studies for an in-depth discussion about clustering.

2.1.1. k-Means. This is the most widely used algorithms so far and has been the regular algorithm used in most of the online tutorial about clustering. The basic idea in k-Means is to randomly place $k$ number of centroids in the input space, as far away possible from each other, the nearest data points are associated with the centroid. This process will be repeated by placing centroid at different locations and stops when it stabilizes (no data points moved group or not much changes in the sum of square error) [8, 9, 10]. What’s important to note here is the number of $k$ and the distance calculation needs to be determined by the researcher. For simplicity purposes, Euclidean distance was used and the value of $k$ to be tested was set to 2, 3 and 4.

2.1.2. Partitioning around Medoid (PAM). k-Means has been reported to perform well in some studies such as [12] and [13], however, the algorithms are known to be sensitive to outliers [11]. Using the same principle as k-Means, the sensitivity is diminished by using medoid instead of mean. That is why PAM is also known as k-Medoids. The medoid is the most central data point in a cluster. This is done by arbitrarily selecting $k$ data points as the initial representative of the cluster. The rest of the data points are associated with cluster based on the nearest $k$. Then a different data point in the cluster was selected as the medoid and if the newly appointed medoids produced the minimum sum of distances, then the new medoid is promoted to be the representative. This process is repeated until it stabilized [14]. Implementing PAM in this study used the same Euclidean distance calculation and $k$ was set at 2, 3 and 4.

2.1.3. The optimal number of clusters. There are several methods that are available to help in deciding the optimal number of clusters and these methods can be categorized into internal and external measures. The measure used in this study was the Silhouette Index (internal) where the measure is based on pairwise distances on all data points in the cluster (compactness), and pairwise distances between all data points in the cluster and all data points in the closest other clusters (separation) [15]. The value closes to 1 considered to be desirable.
2.1.4. Cluster Tendency. To ensure the validity of the clusters, the dataset tendency to cluster can be determined using Hopkins Statistics [16]. The statistics, or H value, can be considered as a hypothesis test of spatial randomness with $H_0$ being the dataset is uniformly distributed. Therefore, highly clusterable datasets will have an $H$ value that is close to 1 and 0.5 for completely random data [17].

2.2. Scaling and Resampling
Scaling was done to normalize the dataset as highly suggested by [10] and [8] to name a few. Scaling in our case is essential since we decided not to remove the outliers and the distance measure used was the Euclidean distance that easily can be affected by large scaled feature. Resampling was conducted using Synthetic Minority Over-sampling Technique (SMOTE). Roughly, the distribution of all produced training records had the proportion of 60:40% based on the class label and these were made balanced by creating synthetic samples from the minor class [18], hence the name.

2.3. Random Forests and Model Validation
The random forests is an ensemble of decision trees that take the majority vote for classification. What makes it different from the classic decision tree creation is, the number of variables for node splitting was chosen randomly [3, 6]. Random forests also come with a build in misclassification error checking mechanism known as the Out of Bag (OOB) error rate [6] and the OOB will be used for models comparison along with the Receiver Operating Characteristic (ROC) curve [18]. The ROC curve is considered as a standard technique used to summarize the model performance over a range of tradeoffs between true positive and false positive error rates. In this study, the area under the curve was used as the performance metric for the random forests models.

2.4. The Process
The analysis was done using R version 3.5.2 and each of the package and function used in the program is mentioned as the process is being described. To get the Hopkins statistics $H$ value, we used `get_clust_tendency()` function from `factoextra` package. After getting a satisfactory $H$ value ($H < 0.5$), k-Means (`kmeans()`) and PAM (`pam()`) from `cluster` package were executed on a scaled dataset (`scale()`) to identify clusters in the given dataset, and the produced clusters were examined for decision on which cluster (hence records) to exclude. Since we were dealing with a small dataset, such inspection was still feasible. The exclusion will be based on the cluster that grouped around older `ageP` to maintained only records that closely fall in the ideal age range. To decide this, cluster with the highest `ageP` average will be removed. At this point, there will be three sets of training records, each from records grouped by k-Means, PAM and manual selection respectively.

A copy of all three sets was then produced. The new copied sets will be resampled using `SMOTE()` function from `DMwR` package. This is to balance out the class proportion to 50:50%. Altogether there will be six sets of training records that will be used to produce the random forests models (`randomForest()`). Each of the random forests models `mtry` value will be tuned based on minimizing the OOB error rate and the number of trees grown – `ntree`, were all set to 1000. Each model will be evaluated based on the OOB and AUC values (`ROCR` package).

3. Results and Discussion

3.1. Cluster Tendency
The H value is 0.3067 ($H < 0.5$), indicating there was enough evidence to reject the null hypothesis and the dataset has a clustering tendency. We rather expected the tendency since outliers were not removed and we expected that the outliers might be clustered into its own group.
3.2. The Clusters and the Training Records

3.2.1. K-Means. K of 2, 3, 4 was tested and the results of the alienated data are shown in figure 2. (a). $k = 2$ was chosen based on the silhouette width results (see figure 2. (b)). Given $k=2$, k-Means had grouped those with older age$P$ and the outliers in class one with the average age$P = 15.75$. The average age$P$ for class two was 12. Therefore, all records from class one were removed leaving 25 records for the training.

![Figure 2.](image)

(a). Comparison of k-Means Clusters for $k=2$, 3 and 4.

(b). Comparison of k-Means Average Silhouette Width Plot for varying $k$.

Figure 2. k-Means Clusters Plot for $k=2$, 3 and 4 in (a) and the Average Silhouette With Plot in (b).

3.2.2. PAM. $k$ of 2, 3, and 4 were also tested and the results of the alienated data are shown in figure 3. The average silhouette width was 0.61, 0.36 and 0.41 respectively (see figure 4). $k = 2$ was chosen since it produced the widest width. Based on PAM, the older age$P$ and all the outliers was grouped in class two with the average age$P = 18.56$. While the average age$P$ for class one = 10.79. All the records from class two were removed leaving 24 records for the training.

![Figure 3.](image)

Figure 3. Comparison of PAM Clusters for $k=2$, 3 and 4.
Figure 4. Comparison of PAM Average Silhouette Width Plot for k = 2, 3 and 4.

The total number of records in each set of training records only varies by 1 or 2 records as compared to the original training set. However, the ageP average was different in all three with the highest in k-Means training and less difference in PAM and the original set. The average differences should be due to records classification with k-Means having much older ageP value in class two. While PAM had classified older ageP and younger ageP in the cluster that will be removed. This should reduce the range of the PAM dataset which translated in much lower ageP average.

3.3. Random Forests

Based on the six training records sets, the summary of the results is shown in table 1. It is rather obvious that the original small datasets produced much higher OOB error rate which will correlate with less accuracy in the model. Even the AUC value falls below 0.5 for k-Means and PAM except for the manually selected dataset that shows AUC = 0.6 that suggest the model was just better than random guessing.

There was a huge improvement, however when the training records were resampled. Especially with the manually selected records where OOB rate reduced to 8.33% (from 34.78%) and AUC = 0.95, which suggest an excellent model. This is followed by the PAM training records - OOB = 15.62%, AUC = 0.92, and finally the k-Means training records - OOB = 25%, AUC = 0.78.

Table 1. Comparison of Random Forests Model Performance between Original and Resampled Training Records.

|                  | Original |          |         | Resample |          |
|------------------|----------|----------|---------|----------|----------|
|                  |          |          |         |          |          |
|                  | k-Means  | PAM      | Manual  | k-Means  | PAM      | Manual  |
| OOB              | 44%      | 37.5%    | 34.78%  | 25%      | 15.62%   | 8.33%   |
| AUC              | 0.2698   | 0.4453   | 0.6349  | 0.7832   | 0.9219   | 0.9506  |

Of all three training records, k-Means was considered the worst before and even after resampled and comparable performances between PAM and the manual selected records after resampled. Looking at the variable important ranks shown in table 2, the mean decreased accuracy values increased on all resampled and the range varies in all six training records. Some of the variable ranks also changed before and after resampled as shown in k-Means and Manual. Much more stable ranking can be seen in PAM as the top two variables (acp and ethnicity) were the same before and after resampling.

When the models’ performances were considered, the interest would be on models using the resampled training records between PAM and Manual selection. If only the top three variables were considered (disregarding the rank), both models had identified acp, ethnicity and ageR. Interestingly the top three variables in PAM (resampled) were the same as the results in [3] which were not resampled but kept balanced.
Table 2. Comparison of Variables Importance Ranks between Original and Resampled Training Records.

| Variables | Original | Resampled |
|-----------|----------|-----------|
| k-Means | PAM | Manual | k-Means | PAM | Manual |

4. Conclusion

This is a study where we returned to our previous studies done [2, 3] on the same dataset of Cleft Lip and Palate patients from the west coast of Sabah. In this study, we had tried to incorporate the unsupervised learning in producing the training records and supervised learning in model prediction. The challenges that we faced before were in the small number of the dataset and it was imbalanced. These problems were overcome by introducing synthetic samples created by SMOTE technique based on the study done by [19]. We also observed if clustering would be feasible to produce an acceptable training record and, in this study, PAM was able to produce a much better training record as compared to k-Means. We speculated that our decision to not remove the outliers might affect k-Means performances as k-Means combined with Euclidean distance were known to be sensitive [10]. The outliers, however, were grouped in the cluster that was removed from the training records in both algorithms and it was further scaled to equalize the dimension, magnitude and variability [9] of the selected features.

The resampled training records showed to have decreased the OOB and increased the AUC on all sets with substantial results in the manually selected records and PAM. This was aligned with [19] findings that SMOTE resampling gives better AUC and balances the sensitivity and specificity measures. PAM also showed stable variables ranks for the first top two before and after resampled and its top three variables rank in the resampled dataset were the same as in [3]. This means that given the case study, random forests were not influenced by the small number of the dataset [3] but rather with the imbalance of the dataset class [20].

In this study, clusters produced by PAM managed to group all outliers and the much younger patients in a group that was being excluded, leaving a cluster with ageP average = 10.79. This was really close with the original dataset where the ageP average is 10.57 [2]. Even though the misclassification error rate was higher in PAM, but the difference in the AUC results between PAM and the manual selection was only off by 0.03, which was insignificant as both models would be classified as an excellent classification model [21]. Using PAM clustering algorithm, therefore, in our case was feasible to produce a training record to be used in the classification model. Though informed judgement should be considered on choosing the right data representation to cluster, as different features can produce different clusters even by the same clustering algorithm [9].

References

[1] Hodgkinson P D, Brown S, Duncan D, Grant C, McNaughton A, Thomas P and Mattick C R 2005 Management of Children with Cleft Lip and Palate: A Review Describing the Application of Multidisciplinary Team Working in this Condition Based upon the Experiences of a Regional Cleft Lip and Palate Centre in the United Kingdom Fetal and Maternal Medicine Review 16 pp 1-27
[2] Ali Omar Z, Chin S N, Sentian A and Kong L Y 2018 Decision Tree for Pre-Graft Orthodontic Treatment: A Case Study of Cleft Lip and Palate Patients in the West Coast of Sabah Advanced Science Letters 24 pp 1398-401

[3] Ali Omar Z, Chin S N, Sentian A, Hamzah N and Yassin F 2018 Exploring Contributing Features of Pre-Graft Orthodontic Treatment of Cleft Lip and Palate Patients Using Random Forests Transactions on Science and Technology 5 pp 5-11

[4] Abhilashaa W 2011 Secondary Alveolar Bone Grafting in Cleft on the Lip and Palate Patients Contemporary Clinical Dentistry 2 pp 146-54

[5] Quinlan J R 1996 Bagging, boosting, and C4.5 Proceedings of the thirteenth national conference on Artificial intelligence 1 pp 725-30

[6] Breiman L 2001 Random Forests Machine Learning 45 pp 5-32

[7] Hothorn T, Hornik K and Zeileis A 2006 Unbiased Recursive Partitioning: A Conditional Inference Framework Journal of Computational and Graphical Statistics 15 pp 651-74

[8] Jain A K, Murty M and Flynn P J 1999 Data Clustering: A Review ACM Computing Surveys 31 pp 264-323

[9] Jain A K 2010 Data clustering: 50 years beyond K-means Pattern Recognition Letters 31 pp 651-66

[10] Mohamad I and Usman D 2013 Standardization and Its Effects on K-Means Clustering Algorithms Research Journal of Applied Sciences Engineering and Technology 6 pp 3299-303

[11] Velmurugan T and Santhanam T 2010 Computational Complexity between K-Means and K-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points Journal of Computer Science 6 pp 363-8

[12] Kaltungo A Y and Ibrahim Y 2017 Experimental investigation of the effectiveness of K-means for classifying misalignment and unbalance faults in industrial rotating machines Proceedings of the International Conference on Maintenance Engineering pp 499-510

[13] Bouras C and Tsogkas V 2010 W-kmeans: Clustering News Articles using WordNet 2010 Lecture Notes in Computer Science Series 6278 pp 379-88

[14] Kaufman L and Rousseeuw P J 1990 Finding Groups in Data: An Introduction to Cluster (New York: John Wiley & Sons)

[15] Hassani M and Seidl T 2017 Using internal evaluation measures to validate the quality of diverse stream clustering algorithms Vietnam Journal of Computer Science 4 pp 171–83

[16] Adolfssson A, Ackerman M and Brownstein N C 2016 To Cluster, or Not to Cluster: How to Answer the Question Proceedings of Knowledge Discovery from Data

[17] Banerjee A and Dave R N 2004 Validating clusters using the Hopkins statistic IEEE International Conference on Fuzzy Systems 1 pp 149-53

[18] Chawla N V, Bowyer K W, Hall L O and Kegelmeyer W P 2002 SMOTE: Synthetic Minority Over-sampling Technique Journal of Artificial Intelligence Research 16 pp 321–57

[19] Dubey R, Zhou J, Wang Y, Thompson P M and Ye J 2014 Analysis of sampling techniques for imbalanced data: An n=648 ADNI study NeuroImage 87 pp 220-41

[20] Janitza S and Hornung R 2018 On the overestimation of random forest’s out-of-bag error PLoS ONE 18 p 8

[21] Swets J A, Dawes R M and Monahan J 2000 Psychological Science Can Improve Diagnostic Decisions Psychological Science in the Public Interest 1 pp 1 1-26