Comparative Studies on The Application of PCA and RFM to Clustering Algorithms for Customer Segmentation

Johannes Tebalo Kokozela (714188), Supervisor(s): Dr Martins Arasomwan
School Of Computer Science & Applied Mathematics

1. Abstract

Customer segmentation methods are highly dependent on the feature selection methods, this is because selected features determine how the segments will be interpreted. **In this study we compared the popular feature selection methods in customer segmentation, which are principal component analysis (PCA) and recency, frequency and monetary (RFM) analysis**. The comparison was done using three popular clustering algorithms which are K-Means, Gaussian Mixture Model (GMM) and K-Medoids and three cluster validation methods which are Davies Bouldin Index (DBI) ,Calinski-Harabasz Index (CHI) and Silhouette Coefficient (SC). **We found that RFM is the better feature selection method for customer segmentation**. For a large data set GMM is the recommended clustering algorithm and for smaller data K-Means performed much better.

2. Introduction

The solution to better understanding customers is use data mining techniques such as customer segmentation, which groups customers into distinct segments based on customer value, customer lifetime and customer needs and want. The process of grouping customers into homogeneous groups is called customer segmentation (also known as market segmentation for wide range observations). These groups of customers are called clusters and these clusters have customers that are similar in each cluster (inter-cluster) and different in the between clusters (intra-clusters), allowing the business to work on this groups separately and in a more suitable and somewhat personalised way.

3. Research Problem

In this study we will be investigating if customer segmentation requires that we go through an in-depth feature selection phase (using dimensionality reduction) or the use of RFM analysis as feature selection method is sufficient. The aim of this study is to compare PCA with RFM on customer segmentation, using different clustering algorithms.

5. Results

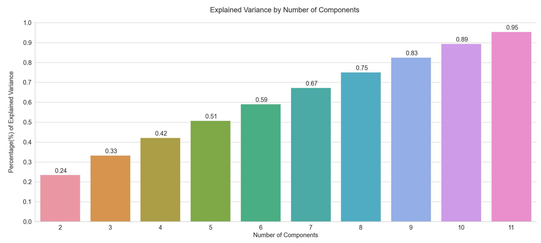


Figure 1: Number of Components and The Corresponding Variation

Figure 1 shows the relationship between the number of selected components and the total variance retained. We see that it takes 9 components to reach variation $\geq 80\%$.

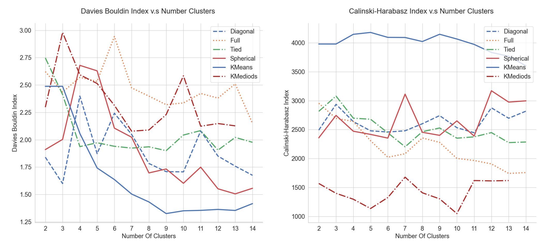


Figure 2: Example of The Evaluation of Different K Values For All Clustering Algorithms

In Figure 2 we observe how the different K-values were evaluated for different clustering algorithms.

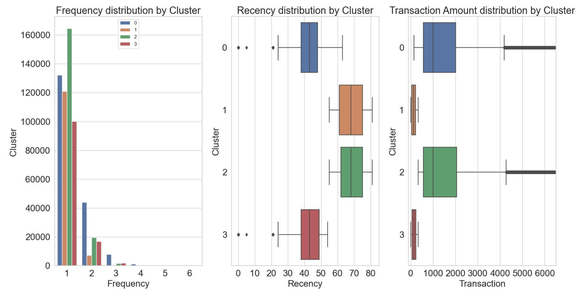


Figure 3: RFM: K-Means with K = 4 Clusters Results

In Figure 3 We see the following outcomes:

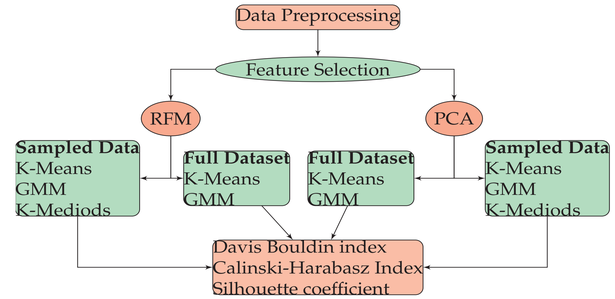
- **Frequency** : We see that in our study frequency is not a good separator if clusters.
- **Recency** : We see that clusters 1 and 2 show clients who haven't recently made a transaction and clusters 0 and 3 show clients who made transactions recently.
- **Monetary** : We see further that cluster 0 would be high transaction amount and transacted more recently (Most valuable clients) and clusters 1 and 3 are low value clients.

Table 1: Summary of Results Table

Design		Model	K	DBI	CHI	SC
RFM	Full Dataset	GMM	4	**0.864**	**242, 176.749**	0.404
	Sampled Dataset	K-Means	4	0.872	27, 846.136	**0.405**
PCA	Full Dataset	K-Means	9	1.327	33, 414.504	0.206
	Sampled Data	K-Means	9	1.325	4, 145.710	0.218

- From Table 1 we see that RFM gives better results than PCA using any cluster validation index.
- GMM and K-Means have very similar results and K-Means however, still performs well even with the sampled dataset.

4. Methodology & Data



The figure above shows the different configurations that were evaluated in this study. The data comes from Kaggle: a data science competitions site. This data comes from an unspecified bank in India and is collected over three months in 2016.

4. Conclusion & Future Work

In this study we determined which feature selection method is appropriate for customer segmentation with a number of experiments. **RFM, K-Means on sampled dataset with 4 clusters** was the optimum configuration. Future work in this subject could focus on determining the most influential variables by applying logistic regression using the resultant clusters.

6. References

[1] Raquel Florez-Lopez and Juan Manuel Ramon-Jeronimo. Marketing segmentation through machine learning models: An approach based on customer relationship management and customer profitability accounting. *Social Science Computer Review*, 27(1):96–117, 2009.

[2] Kishana R Kashwan and CM Velu. Customer segmentation using clustering and data mining techniques. *International Journal of Computer Theory and Engineering*, 5(6):856, 2013.

[3] Sulekha Goyat. The basis of market segmentation: a critical review of literature. *European Journal of Business and Management*, 3(9):45–54, 2011.