School of Computing and Information Systems

UNCOVERING RETAIL CUSTOMER SEGMENTATION FROM LARGE TRANSACTION RECORDS: A NUANCED COMPARISON OF CLUSTERING ALGORITHMS USING ROUGH SET REDUCED DATASET

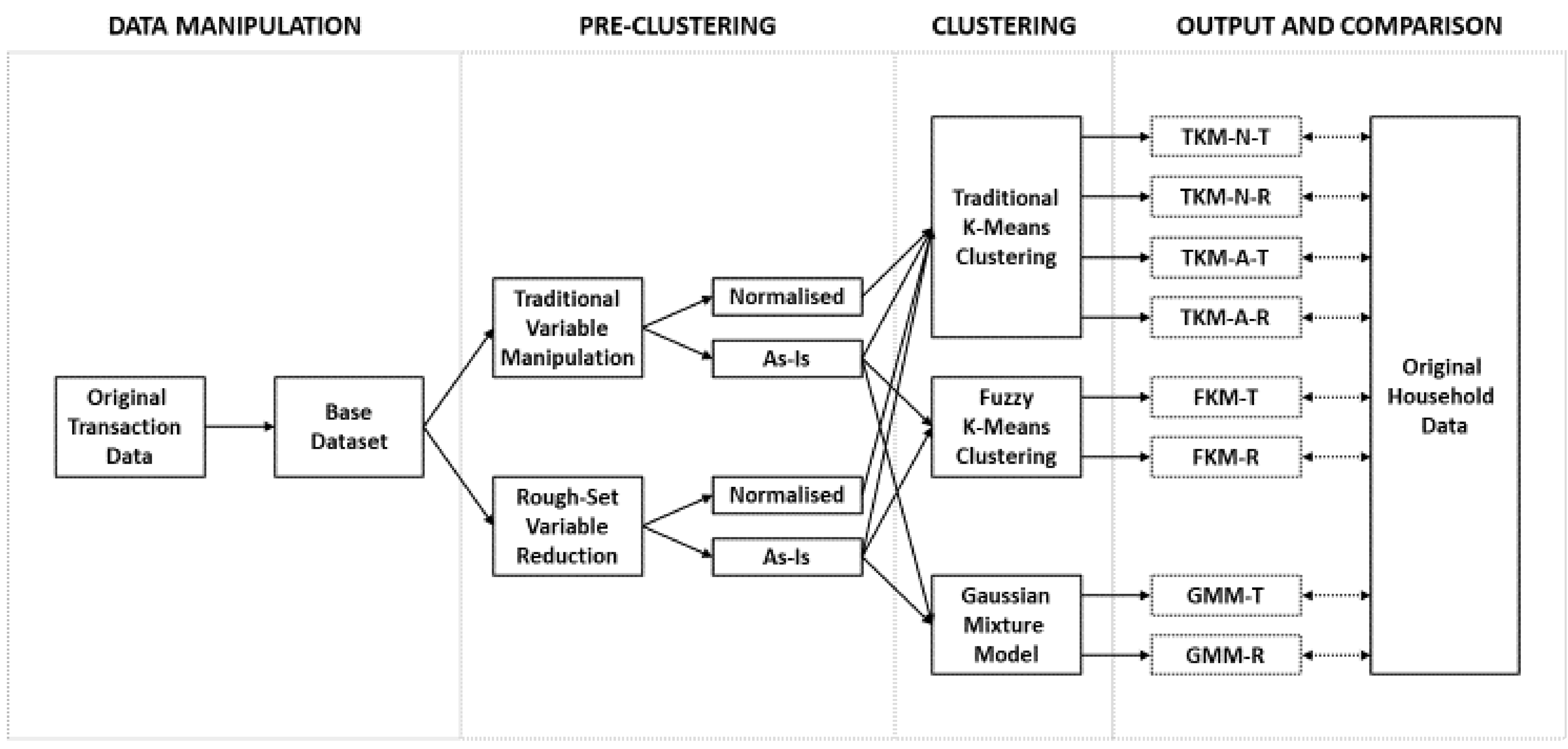
Syed Ahmad Zaki Bin Syed Sakaf Al-Attas
Supervisor: Prof. Kam Tin Seong

Motivation

Motivation of this project lies in that traditional clustering methods are not designed to address inherent inconsistencies in today's real-world data. This project aims to explore and compare the traditional *k*-means clustering (TKM) against alternative clustering methods ie. fuzzy *k*-means clustering (FKM) and Gaussian mixture models (GMM). At the same time, this project will explore the use of Rough Set's reduct as a feature reduction algorithm in understanding its impact on overall clustering accuracy.

Objectives of this project are: (1) Detail different clustering outcomes (2) Uncover merits and shortcomings of each clustering methods (3) Suggest situations where each approach would excel (4) Incorporate use of R code in SAS Enterprise Miner (EM) environment

Workflow



Data & Data Preparation

Data Source

Obtained from Dunnhumby's The Complete Journey (Retail Shopping). Used only the demographic and transaction data

Method of Execution

Data manipulation is done on SAS JMP Pro, whereas the step starting from pre-clustering onwards is done on SAS EM 14.1. Given that the above clustering algorithms and Rough Set are not included in EM, this capstone takes advantage of the Open-Source Integration Node within EM and utilise R. The above clustering algorithms already exist as R packages.

Base Dataset

Using the RFM model (Recency, Frequency and Monetary value), Dunnhumby's transaction table was distilled into 28 continuous and 1 nominal household ID variable.

Pre-Clustering

Base Dataset are split into the following two datasets:

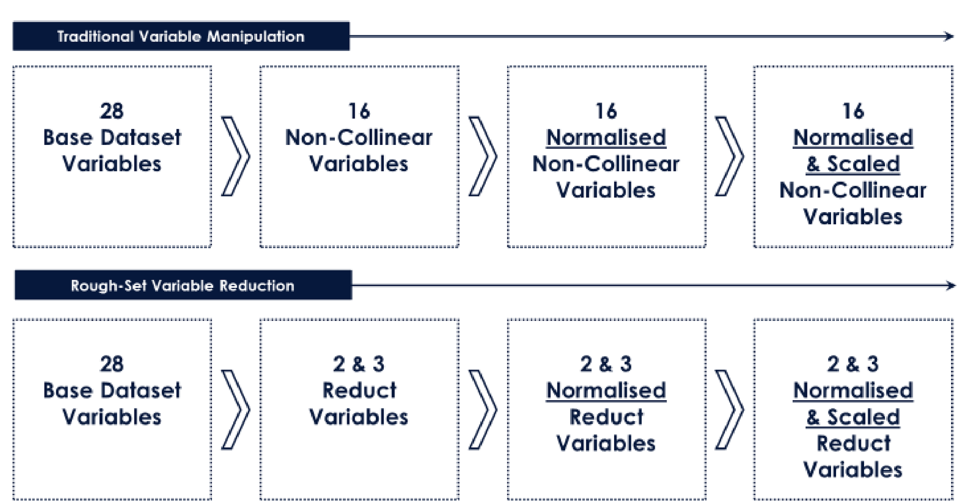
1) Traditional Variable Manipulation

Employed Principal Components Analysis (PCA) techniques to distill 16 non-collinear variables. This is specifically fed into traditional *k*-means clustering, given its need for non-collinear variables.

2) Rough-Set Variable Reduction

Employed Rough Set's reduct algorithm to reduce features to two and/or three significant variables.

Both datasets are then further split into two separate datasets ie. normalised and as-is, creating a total of four datasets.



These four datasets are then scaled, before separately being fed into the three clustering algorithms.

Pre-Clustering				*Clustering Method & Output*		
Data Manipulation Method	**Normalised?**	**Scaled?**	**Variable Count**	**TKM**	**FKM**	**GMM**
Traditional Variable	Yes	Yes	16	TKM-N-T	-	-
	No	Yes	16	TKM-A-T	FKM-T	GMM-T
Rough Set Reduct	Yes	Yes	2 & 3	TKM-N-R	-	-
	No	Yes	2 & 3	TKM-A-R	FKM-R	GMM-R

Analysis & Results

Scoring Methods To Identify Optimal Cluster Counts

Traditional *k*-means clustering uses the silhouette index to identify the optimal cluster counts. It measures how similar a data-point is within-cluster (cohesion), compared to other clusters (separation). Fuzzy *k*-means uses a variant of this same silhouette index, where it incorporates fuzzy logic. Conversely, GMM uses Bayesian Information Criterion (BIC) to identify optimal cluster *k*.

Output & Observations

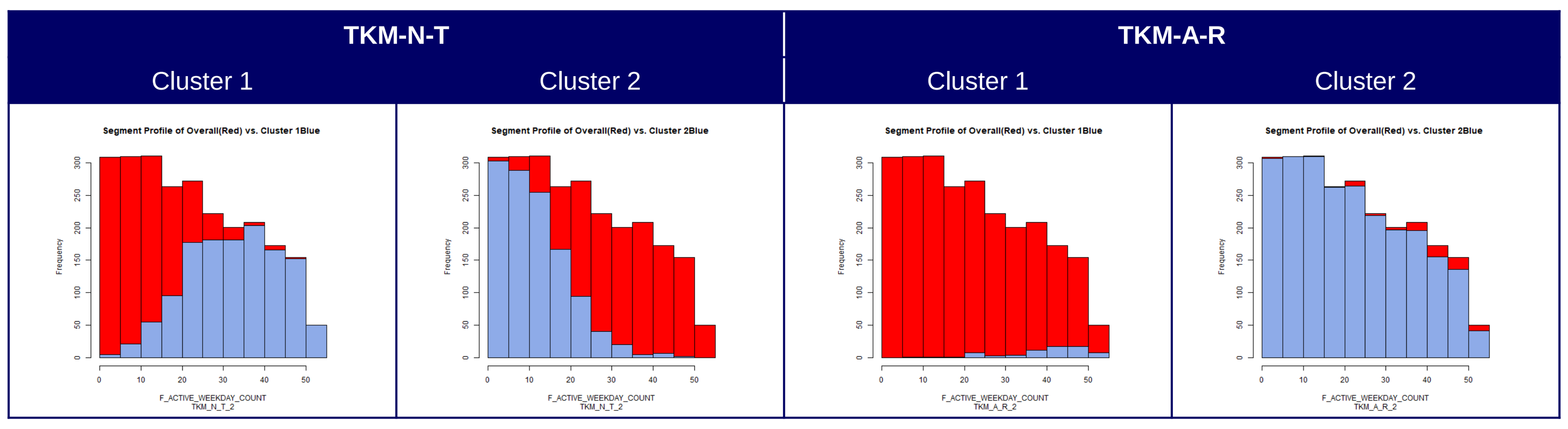
- Outputs, using reduct dataset, tend to score highest, but tend to exhibit uneven distribution amongst its clusters
- GMM outputs exhibited higher optimal cluster counts, compared to TKM and FKM
- FKM-R recorded highest score. Exploring alternative cluster counts with similar high scores yielded cluster counts of 7 and 9 respectively, but distribution still remained uneven
- Since GMM-R's optimal cluster count of 19 is the highest of the lot, superimposing GMM-T's cluster count of 8 onto GMM-R yielded sub-optimal clusters

	TKM-N-T	TKM-N-R	TKM-A-T	TKM-A-R	FKM-T	FKM-R	GMM-T	GMM-R
Variable Count	16	2	16	2	28	2	28	2
Scoring Method	Avg Sil.	Avg Sil.	Avg Sil.	Avg Sil.	Fuzzy Sil.	Fuzzy Sil.	BIC	BIC
Highest Score	~0.19	~0.83	~0.21	~0.90	~0.45	~0.97	-58,608 (VEV)	-2,270 (EEV)
Optimal Cluster Count	2	2	3	2	2	2	8	19
Distribution Between Clusters	Even	Uneven	Somewhat Even	Uneven	Even	Uneven	Even	Uneven

Transaction (Txn) Cluster Observations

- Clusters, using reduct dataset of 2 variables, had the most distinctive split, despite uneven distribution (see TKM-A-R below)
- Clusters, using reduct dataset of 3 variables, had a less significant split, than using reduct dataset of 2 variables
- Some of GMM-T's 8 clusters share similar attributes

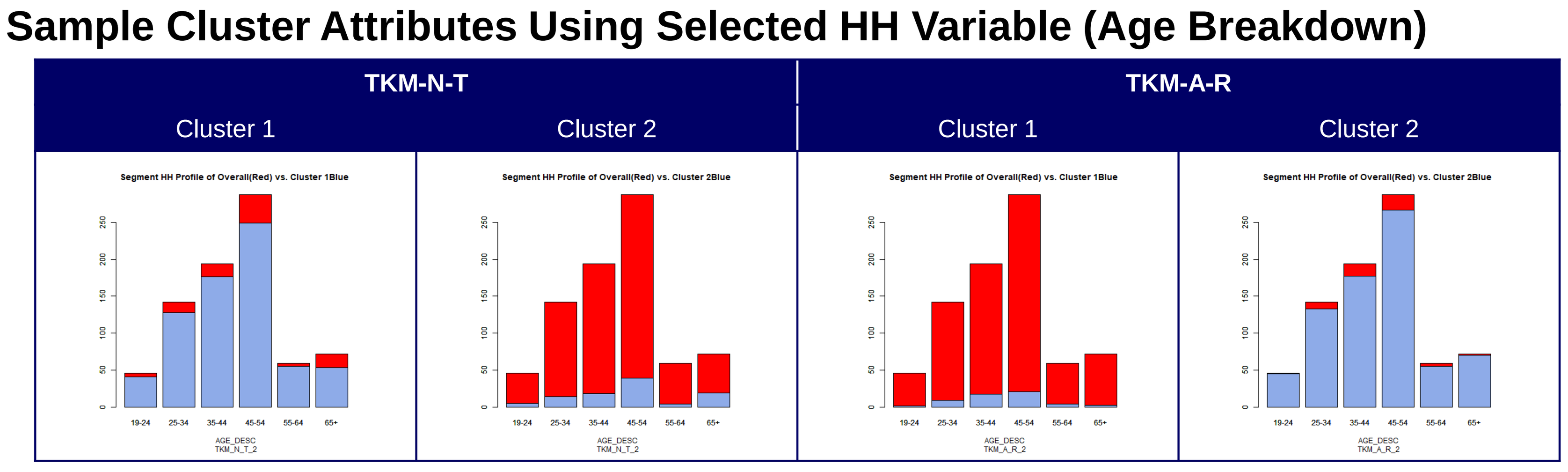
Sample Cluster Attributes Using Selected Txn Variable (Active Weekday Count)



Household (HH) Cluster Observations

- Clusters, using reduct dataset, did not show a similar clear split on household variables
- Household attributes in individual clusters largely match its transactional attributes, except in outputs with large cluster counts

Sample Cluster Attributes Using Selected HH Variable (Age Breakdown)



Conclusion

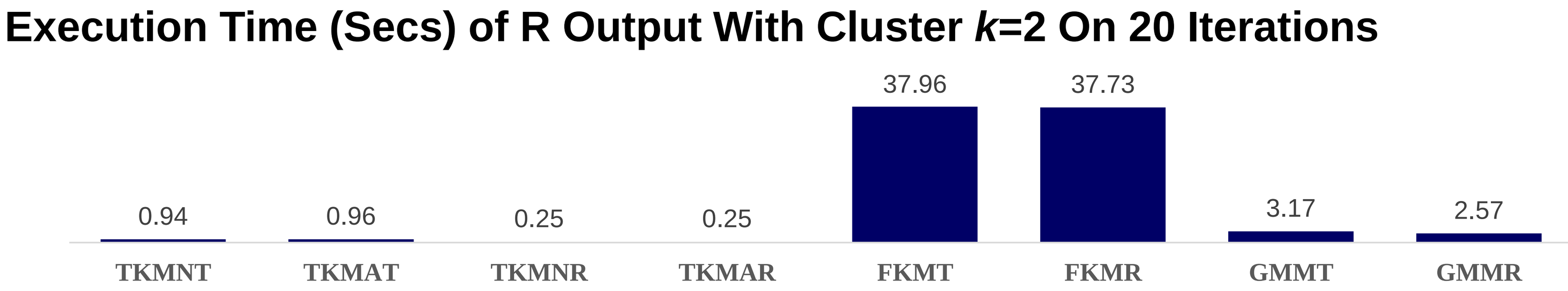
Rough Set's Reduct Impact

- Cluster outputs, using reduct dataset of 2 variables, displayed the most distinctive cluster attributes, especially for FKM and GMM
- Outcome of a rough set's reduct algorithm can neither be known beforehand, nor controlled to select a set number of variables a priori
- Conscious decision in this capstone to apply reduct pre-clustering, as compared to other related journals who have applied it post-clustering to identify significant variables of each cluster

Processing Speed

- FKM took the longest to run, followed by GMM, due to 'soft' clustering calculations for each object within dataset. Speed may also be linked to the use of a lower R version due to SAS EM compatibility requirements

Execution Time (Secs) of R Output With Cluster *k*=2 On 20 Iterations



Use of R Code within SAS EM Environment

- Allows use of statistical algorithms that are not standard in SAS EM
- Key to maintain updated R and SAS EM compatible versions for sustained use