Segmenting and Targeting Customers Through Clusters Selection & Analysis

Ilung Pranata
School of Design, Communication & IT
The University of Newcastle, Australia
University Drive, Callaghan
Ilung.Pranata@newcastle.edu.au

Geoff Skinner
School of Design, Communication & IT
The University of Newcastle, Australia
University Drive, Callaghan
Geoff.Skinner@newcastle.edu.au

***Abstract*—This paper investigates the use of machine learning clustering technique to segment and target customers of a wholesale distributor. It describes the selection, analysis, and interpretation of clusters for evaluating customers annual spending on the products. We show how circular statistics can categorize customers by looking at the annual spending on six essential product categories. Several clusters were created using k-means clustering algorithm and an in-depth analysis on these clusters were performed using several techniques to carefully select the best cluster. Automated clustering was able to suggest groups that these customers fall into. The evaluation and interpretation of clusters were able to provide insights into various purchase behaviors and to nominate the best customer group to target.**

Keywords*— *cluster selection, cluster evaluation, k-means, customer categorization, data mining.

I. Introduction

The terms Big Data, Data Analytics, and Business Informatics are contemporary terms and technologies of increasing significance. The technologies are becoming widely accessible and common place in many domains and enterprise portfolios. However, along with cloud computing infrastructure to support these technologies, we find the organizational data management requirements and techniques becoming increasingly complex but necessary to stay competitive in a data centric business evolution. The building block of many of these ingenious data manipulation techniques have derived from a more traditional, but now quickly dating, field of data mining. Data mining is in turn built upon the foundation of database exploration and the general collection of a diverse array of data for potential commercial gain and understanding.

It is commonly stated "information is the new power". However the real power base is recognizing the fact that information is simply data analyzed and used in the right context. Hence, a key to many successful contemporary businesses is in large part attributed to how effective their data mining processes and strategies are. But data mining is more than this, and is best explained through a well-used formal definition. That is, data mining (DM) "…is the process of discovering interesting patterns and knowledge from large amounts of data" [1]. Further, it is seen as "…an essential process where intelligent methods are applied to extract data patterns" [1]. In a wider systems process perspective, data mining is frequently seen as an important phase or step in a broader defined process, that being knowledge discovery (KD) [2]. As such it is often difficult to separate the two concepts due to their close inter-dependence. Early applications of data mining were primarily focused on customer relationship management (CRM), such as detecting patterns in customer behavior for better supporting marketing, sales and support [3]. As this paper demonstrates, the focus has not changed to any great degree, however the methods and processes of DM and KD have becoming increasingly more complex.

In contemporary data mining we now use terms such as segmenting, cluster analysis, and customer domains. Again these technologies and techniques are founded upon data mining and primarily used for customer profiling for increased revenue opportunities. Specifically, application of data mining allows better cross selling methods along with improved customer retention [4]. This is due to data mining activities covering three key areas: discovery; predictive modelling; and forensic analysis [5]. From this six basic models can be formulated: classification, regression, time series, clustering, association analysis, and sequence discovery [6]. These models used in various combinations can assist in applying data mining to retail and more specifically customer relationship management. Four key retail applications include: Performing basket analysis; Sales forecasting; and Merchandise planning and allocation [7]. KD process coupled with the above mentioned applications allows an entity to undertake effective customer segmentation, which is an ability to discover discrete segments in their customer bases by considering additional variables beyond traditional analysis. Applied targeted customer selection through segmentation and cluster selection is the focus of this paper. However, before detailing our innovative application of these technologies, the following section provides a small review of the theoretical background and related works.

The core elements of the paper are focused on an innovative application of cluster segmentation and data analytics technologies to a wholesale distributor's large data set. Essentially, this paper applies advanced data mining analysis in an e-commerce environment. As such, this paper is presented as follow: this section provides an introduction to data mining and its usage for customer analysis and segmentation. This is followed by section II that provides review on the literature for several data mining and clustering approaches in this area. Section III describes the customer annual expenditure dataset that we use for analyzing and categorizing customers. Section

978-1-5090-0363-1/15/$31.00

IV present the clustering techniques used to provide segmentation of customers. Section V discuss the selection criteria and the process to select the best number of segments/groups based on a given dataset. Three pronged measures were employed to determine the best number of customer groups. This is then followed by section VI that interprets the results and provide insights to each customer segment. Section VII concludes this paper.

II. Theoretical Background

Clustering and Segmentation are two primary techniques heavily utilized in analytical CRM procedures [8]. The division of the customer data set into distinct and homogenous groups is the purpose of Segmentation [8]. Likewise, clustering is the division of data into similar objects [9]. While each respective technique sounds similar to the other, there are some fundamental differences when we introduce the term borders. Borders are what divides the data sets into groups. As such finding the borders between the groups is the role of clustering while segmentation uses the borders to actually form the groups [10]. So in a simple example where one might have an extremely homogenous data set, then segmentation would be possible however clustering would be very difficult due to the lack of density differences [10]. Generally however, we use clustering to form initial segments that are not readily identifiable [11].

Due to the ongoing success of data mining for CRM, there has been a rapidly increasing volume of literature in this domain. Researchers and professionals alike are investing considerable time investigating and experimenting in this space with some very interesting results. As such, this section is only able to touch upon the vast amount of literature relevant to the papers specific focus, that being segmentation and clustering for customer targeting. The authors in [12] provide a solid starting point with their review of some popular analytical methods for direct marketing segmentation. They detail the following three approaches: RPM (Recency, Frequency, and Monetary value), CHAID (Chi-square Automatic Interaction Detection), and logistic regression. A similar work to our own but in a slightly different context is provided in [13]. Here the authors provide segmentation analysis on U.S. grocery shoppers using three clustering approaches: k-means; Beale's F-type statistic; and Pseudo-Hotteling's. Like this paper the main research contributions were from the clustering analysis. As such the remainder of this section reviews clustering techniques.

As mentioned and recognized by the authors, application of data mining clustering techniques to large consumer data sets is not new. There is an extensive amount of literature in this area, however the differentiators between the works, is the application and combination of techniques applied, and of course the data sets being examined. Some similar works to our own are detailed in [14-17]. Common among these works is the application of various forms of the K-means clustering algorithm. K-means seems to be the most widely used and accepted means of clustering [18]. The reasons for this is best described by the authors in [18] where they state the following about K-Means: “Ease of implementation, simplicity, efficiency, and empirical success are the main reasons for its popularity. K-means is so widely adopted in the data mining field. Numerous textbooks and thesis have been written about it [19-21]. The authors like many researchers before them have chosen to use the k-means clustering algorithm due to its extensive and well accepted history of use and acceptance.

To complement the basic k-means approach, a number of quantitative measures are required to enhance the analyst’s selection of the best data partitions. Some common measures that have also been applied in this work are: Elbow [22]; Davies-Bouldin [23]; and Silhouette Width [24]. Further, we apply the use of circular statistics, similar to the work conducted in [25]. The final clarification of clustering terms as it relates to this papers work is the difference between supervised and unsupervised clustering [26]. Our work is best exemplified through the application of unsupervised clustering using k-means analysis with the above mentioned measures and statistical approach.

III. Dataset Description

The data used in this study refers to the customers of a wholesale distributor in Portugal and can be obtained from [27]. These customers are surveyed for their annual expenditure in monetary units (m.u.) on six essential product categories:

- fresh products,
- milk products,
- grocery products,
- frozen products,
- delicatessen, and
- detergents and paper products.

The data contains 440 instances: 298 customers from the Horeca (Hotel/Restaurant/Café) channel and 142 customers from the retail channel. These customers were distributed in two large Portuguese cities, i.e. Lisbon and Oporto and another region. Table 1. Shows the distribution of customers in the region.

TABLE I. Customer Distributions based on Regions

Region	No of Instances	Percentage
Lisbon	77	17.5%
Oporto	47	10.5%
Other region	316	31.8%
Total	440	100%

IV. Clustering Technique

Clustering is normally seen as the process which begins with raw data and ends with new knowledge of the particular area being studied [1]. Authors in [28] describe clustering as an unsupervised classification of patterns (i.e. observations, items, feature vectors, etc.) into groups (normally named clusters). In this paper, we focus at finding the natural groups of customers based on their annual spending on six product categories. It is expected that these natural groups could provide insights into customer spending behavior. These insights further provide means to target customers with specific products.

There are several clustering algorithms proposed in the literature. We chose to use k-means clustering algorithm [29] for this particular study. K-means was chosen due to its exceptional performance with a large set of data and also due to the simple features used in our study. K-means aims to minimize the distance between the data points and the cluster. Suppose a set of data points $x_1, x_2, \ldots, x_n$ where each data point is a d-dimensional real vector, k-means aims to partition the *n* data points to $k \leq n$ sets $c = \{c_1, c_2, \ldots, c_k\}$ so as to minimize their distance with the cluster center. The objective of k-means is to find

$$\arg\min_{c} \sum_{i=1}^{k} \sum_{x \in c_i} \| x - \mu_i \|^2$$

where c_i is the set of data points that belong to cluster *i* and μ_i is the mean of points in c_i. K-means employs the square of Euclidean distance $d(x, \mu_i) = \| x - \mu_i \|^2$ to measure the distance of each data point to the centroid in Euclidean space. The algorithm works as follow:

1. First, decide the number of clusters *k*.
2. Initialize the center of the clusters ($\mu_i, i = 1, \ldots, k$)
3. Each data point x_i is assigned to its closest cluster center
4. Set the position of each cluster center c_i to the mean of its constituent data points.
5. Repeat step 3-4 until convergence.

When the algorithm converges, it does not necessarily mean the minimum sum of squares but rather it is just a heuristic due the problem is non-convex. Convergence means the assignments of data points do not change from one iteration to another.

In this work, RapidMiner [30] predictive analytics platform is used to perform data clustering with k-means implementation. The clustering was performed for customer's annual spending on all six product categories. A total of 9 clustering experiments from *k*=2 to *k*=10 were performed. Statistics for each *k* clustering were collected to evaluate the effectiveness of the *k* number clustering and to select the best *k* for grouping the customers.

V. Clusters Selection

One crucial issue in clustering is to pick the best number of clusters (*k*) for further analysis and evaluation. A number of quantitative measures are available for assessing clusters quality, such as in [31-33]. These measures evaluates the validity of data partitions produced by a clustering algorithm and suggests partitions that worth researcher's attention. However, this crucial evaluation is often neglected in clustering analysis [34]. The results of these quantitative measures would be expected to reflect the selected partitions, features, clustering algorithms, and other parameters. In this paper, we employ multi-pronged quantitative measures, namely Elbow method [22], Davies-Bouldin [23], and Silhouette Width [24] to select the best partitions for further analysis and evaluation.

A. The Elbow Method

Elbow method looks at the intra and inter cluster similarity using a Euclidean distance. It normalizes the intra-cluster sums of squares to measure the compactness of clusters. It is based on the principle that the first cluster will add much information and create greater variance but as more and more clusters added, at some point of time the marginal gain will be dropped.

The intra-cluster sum of squared distance between each member of a given cluster c_i containing n_i data points and its centroid μ_i is computed mathematically using:

$$D_i = 2n_i \sum_{x \in c_i} \| x - \mu_i \|^2$$

The measure of compactness of the clusters can be measured by normalizing the intra-cluster sums of squares:

$$\overline{W}_k = \sum_{i=1}^{k} \frac{1}{2n_i} D_i$$

$\overline{W}_k$ (i.e. the average within centroid distance) will then be plotted in the graph to determine the optimal number of clusters by looking at the angle in the graph. This angle is normally called as "elbow" which shows the optimal number of clusters (*k*) to be picked. Fig. 1 shows $\overline{W}_k$ from clusters *k*=2 to *k*=10 in our dataset.

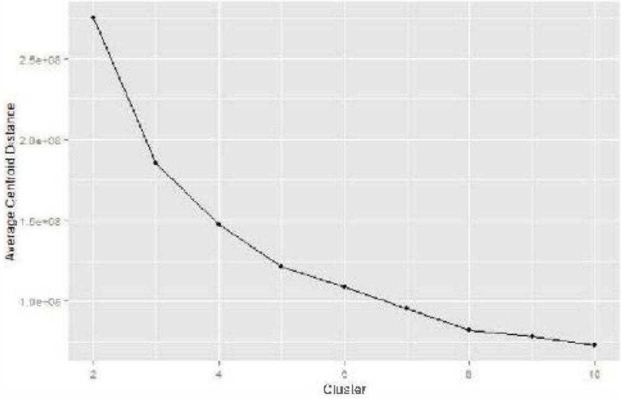

Fig. 1. The average within centroid distance from k=2 to k=10

At some situations, the "elbow" cannot always be unambiguously identified or there may be multiple elbows exist [35]. As seen in fig. 1, there are at least three elbows can be identified: *k* = 3, 5 and 8. Therefore, further clusters analysis need to be performed to obtain the optimal *k*.

B. Davies-Bouldin

Davies-Bouldin (DB) index [23] evaluates clusters performance using dataset quantities and features. It is a function of the ratio of the sum of within-cluster scatter to between-cluster separation. The measure of scatter within *i*th cluster, S_i is computed mathematically using:

$$S_i = \frac{1}{|c_i|} \sum_{x \in c_i} \| x \ - \mu_i \|$$

while the measure of separation ($M_{i,j}$) between clusters c_i and c_j is computed using $M_{i,j} = \| \mu_i - \mu_j \|$. The two clusters performance is then measured in a similarity measure ($R_{i,j}$) where

$$R_{ij} = \frac{(S_i + S_j)}{M_{i,j}}$$

Once similarity measures are computed for all clusters, the DB index is defined as

$$DB = \frac{1}{k} \sum_{i=1}^{k} D_i$$

$$D_i = \max_{j \neq i} R_{ij}$$

DB index is the average similarity between each cluster and its most similar one. Thus, it is desirable to choose the minimal *DB* as it shows better inter-clusters separation and better tightness inside the clusters. We run DB index on our dataset for cluster k=2 to k=10. The results are plotted in fig. 2. It is clear from the plot that clusters with the lowest DB index are clusters $k = 3$, 4, 5, and 8 respectively.

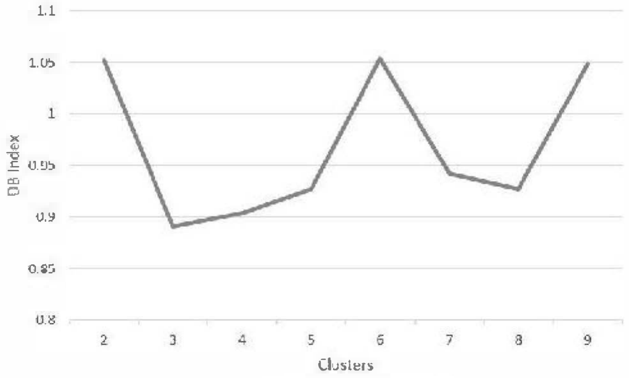

Fig. 2. The DB index of clusters k=2 to k=10

C. Silhouette Width

Authors in [24] introduced Silhouette Width to provide validation of centroid clusters. It since has been used in many studies, such as in genome expression [34] and in serious games analytics [36]. The Silhouette Width (SW_x) can be calculated for each data point x using the following mathematical equation:

$$SW_x = \frac{separation(x) - cohesion(x)}{\max\{cohesion(x), separation(x)\}}$$

where $cohesion(x)$ is a measure of average distance between a data point x and other data points in the same cluster c_i, and it can be computed mathematically using:

$$cohesion(x) = \frac{1}{|c_i|} \sum_{x,y \in c_i} dist_{x,y}$$

$separation(x)$ is a measure of the average distance between data point x and all data points in the nearest cluster c_i, and it can be computed using mathematical equation shown below:

$$separation(x) = \min(\frac{1}{|c_i|} \sum_{x,y \in c_i} dist_{x,y})$$

The distance $dist_{x,y}$ is calculated as the Euclidean distance between data point x and y. If multiple m features are used in the clustering, the difference must be calculated for each feature, then squared and summed, as depicted in the following equation:

$$dist_{x,y} = \sqrt{\frac{1}{N} \sum_{j=1}^{m} (x_j - y_j)^2}$$

The validation of the overall clusters is then obtained from the average Silhouette Width ($\overline{SW}$) over all N records in the dataset as shown below:

$$\overline{SW} = \frac{1}{N} \sum_{x=1}^{N} SW_x$$

The value of $\overline{SW}$ ranges from -1 (completely meshed clusters) to +1 (well separated clusters). We applied Silhouette Width in our dataset and obtain the results as plotted in fig. 3. The best three performers of k clusters are when $k = 2, 3, 4$.

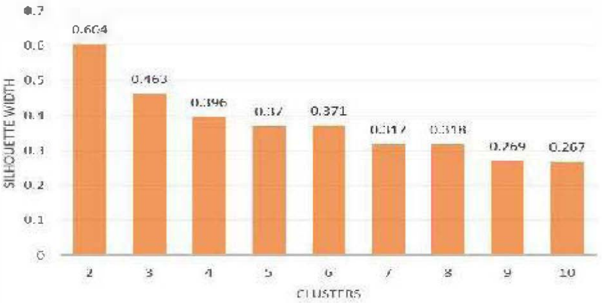

Fig. 3. The Silhouette Width of clusters k=2 to k=10

Now that the dataset has been clustered and various cluster validation analyses have been performed, the next step is to select the best partition for further evaluation and interpretation. Table II shows the top three k partitions for the three performed methods. It is clear that $k = 3$ partition should be selected for cluster interpretation and evaluation as this partition was the best performer in the three validation methods. Hence, this partition was selected.

TABLE II. BEST K PARTITIONS BASED ON ELBOW METHOD, DAVIES-BOULDIN INDEX, AND SILHOUETTE WIDTH

No	Elbow Method	Davies-Bouldin	Silhouette Width
1	$k = 3$	$k = 3$	$k = 2$
2	$k = 5$	$k = 4$	$k = 3$
3	$k = 8$	$k = 5, k = 8$	$k = 4$

VI. CLUSTERS INTERPRETATION

Clusters formed by the clustering algorithms are not normally accompanied by the reasons of why the data points are grouped in that way. Hence, practitioners are required to identify the trends and draw conclusions from the clustering results. Trends of some clusters are obvious and do not require in-depth domain expertise to identify and to draw the conclusion from. However, other clusters can be hard to interpret without sufficient information and knowledge of the domain.

Clusters of our customer annual expenditure dataset are shown in table III. This table also shows the annual expenditure on various product categories and the total number of customers in each cluster. A conclusion can be drawn that the clustering algorithm has clustered these customer data based on their total annual expenditure. Cluster_0 from table III is a group that contains high value customers with an average total annual expenditure of 93806.83 m.u. This is then followed by Cluster_2 (i.e. middle value customers) with an average total annual expenditure of 51995.52 m.u. and Cluster_1 (i.e. low value customers) with 24015.56 m.u average total expenditure. As Cluster_0 has the highest annual expenditure amongst all groups, the wholesale distributor may want to pay more attention into the 28 customers of the group.

TABLE III. CLUSTERING RESULTS OF ANNUAL CUSTOMER EXPENDITURE DATASET WHEN K=3

Product Categories	Cluster_0	Cluster_1	Cluster_2
Fresh	11849.18	7390.96	32768.01
Milk	24717.11	4439.77	4827.68
Grocery	33887.71	6292.20	5723.15
Frozen	3409.32	2495.53	5535.92
Detergents_Paper	15459.71	2238.65	1074.12
Delicassen	4483.86	1158.44	2066.64
Average Total Aunual Expediture	93806.83	24015.56	51995.52
Total Customers	28	337	75

A further analysis into customer annual spending on each product category, as shown in fig. 4, shows that high value customers normally spend most amount of their shopping budget on grocery products and then followed by milk and detergents paper. The middle value customers spend most shopping budget on fresh products, grocery products and frozen products respectively. On the other hand, the low value customers spend most in the fresh products, grocery products, and milk.

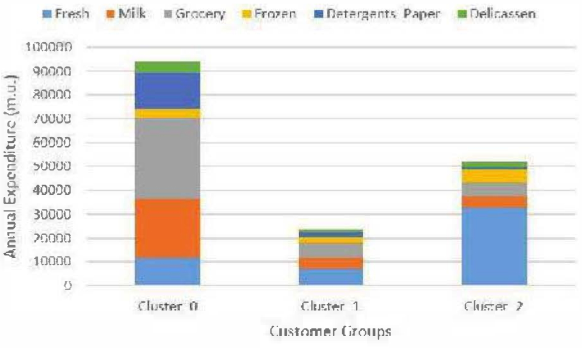

Fig. 4. The Average Annual Expenditure of the Three Clusters

Fig. 5. presents the annual spending ratio of each product category in each cluster against the total average annual expenditure of the cluster. This figure is paramount for wholesale distributor to determine which customer groups should be targeted for each product. For example, frozen products are best targeted to the middle and low value customers groups rather than to the high value group while fresh products are best to be marketed to middle value customers (Cluster_2) as they tend to buy a large amount of these products.

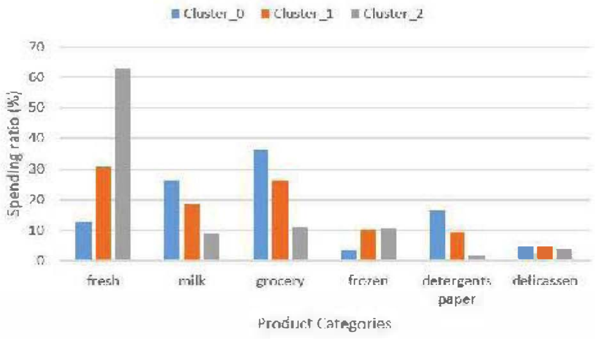

Fig. 5. The Average Annual Expenditure on Each Product Category

While Fig. 4 shows the overall customer group that the distributor should target (i.e. high expenditure customers), fig. 5 shows the best customer group to target for each product category. These customer insights could potentially help the distributor to allocate their marketing campaign more effectively. As such, returns from the marketing investments will be greater as the distributor targets the right customers with the right products. This, in turn, increases customer retention [4].

VII. Conclusion

This paper has presented techniques for evaluating customer annual expenditure across various product categories. These techniques were proven to group similar customers from a myriad of non-related data. More importantly, they allow practitioners and domain experts to develop insights into these groups. Six product categories were used as features for categorizing customers and a machine learning clustering algorithm, k-means, is employed to find correlation between these features and to group similar customers. In order to determine the best number of clusters, three validation measures, i.e. Elbow Method, Davies-Bouldin Index and Silhouette Width, were used. Results show that it is possible to segment customers by their annual product expenditure. The cluster $k = 3$ is chosen after series of evaluation measures. This cluster was the top performer in both Elbow Method & Davies Bouldin (index = 0.89) while it is a second performer in Silhouette Width (index = 0.463). Three groups were created as results of the clustering process showing high, middle, and low value customer groups. The clustering identified high value group customers who have, on average, more than $90,000 annual expenditure. Further drill-down analysis shows that this customers group spend most on groceries. Further analysis can also be performed on other clusters. With such insights, distributor can better target customers with the right products. The success of this work can further facilitates better data mining and customer segmentation techniques for businesses in retail, logistic, e-commerce and many other areas.

References

[1] Han, J., Kamber, M., and Pei, J.: 'Data mining: Concepts and techniques. Morgan Kaufmann series in data management systems' (Morgan Kaufmann, 2011. 2011)

[2] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.

[3] Berry, M. J., & Linoff, G. (1997). Data mining techniques: for marketing, sales, and customer support. John Wiley & Sons, Inc..

[4] W. Zhao, X.Y. Li & L.P. Fu, "Research on clustering analysis and its application in customer data mining of enterprise,"Proceedings of the 2014 Pacific-Asia
Application (CSIA 2014), Bangkok, Thailand, November 17-18, 2014.

[5] Rygielski, C., Wang, J. C., & Yen, D. C. (2002). Data mining techniques for customer relationship management. Technology in society, 24(4), 483-502.

[6] Edelstein H. Data mining: exploiting the hidden trends in your data. DB2 Online Magazine. http://www.db2mag.com/9701edel.htm

[7] Decision Support Solutions: Compaq. Object relational data mining technology for a competitive advantage. http://www.tandem.com/brfs_wps?odmadvwp/odmadvwp.htm

[8] Tsiptsis, K. K., & Chorianopoulos, A. A. (2011). Data mining techniques in CRM: inside customer segmentation. John Wiley & Sons.

[9] Berkhin, P. (2006). A survey of clustering data mining techniques. In Grouping multidimensional data (pp. 25-71). Springer Berlin Heidelberg.

[10] Mixotricha, "The Difference between Segmentation and Clustering," July 17, 2010, https://zyxo.wordpress.com/2010/07/17/the-difference-between-segmentation-and-clustering/

[11] Dominique Levin, "The Difference Between Segmentation and Clustering," Aug, 29 2014, http://www.agilone.com/blog/the-difference-between-segmentation-and-clustering.

[12] McCarty, J. A., & Hastak, M. (2007). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. Journal of business research, 60(6), 656-662.

[13] Wolfson, P., Kinsey, J., & Senauer, B. (2001). Data mining: A segmentation analysis of US grocery shoppers (pp. 01-01). St. Paul: Retail Food Industry Center, University of Minnesota.

[14] Dennis, C., Marsland, D., & Cockett, T. (2001). Data mining for shopping centres-customer knowledge-management framework. Journal of Knowledge Management, 5(4), 368-374.

[15] Lawrence, R. D., Almasi, G. S., Kotlyar, V., Viveros, M., & Duri, S. S. (2001). Personalization of supermarket product recommendations (pp. 11-32). Springer US.

[16] Huang, S. C., Chang, E. C., & Wu, H. H. (2009). A case study of applying data mining techniques in an outfitter's
Systems with Applications, 36(3), 5909-5915.

[17] Voges, K. E., Pope, N., & Brown, M. R. (2002). Cluster analysis of marketing data examining on-line shopping orientation: A comparison of k-means and rough clustering approaches. Heuristics and Optimization for Knowledge Discovery, 207-224.

[18] Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. Data mining and knowledge discovery, 2(3), 283-304.

[19] Wu, J. (2012). Advances in K-means clustering: a data mining thinking. Springer Science & Business Media.

[20] Tan, P. N., Steinbach, M., & Kumar, V. (2013). Data Mining Cluster Analysis: Basic Concepts and Algorithms.

[21] Aggarwal, C. C., & Reddy, C. K. (Eds.). (2013). Data clustering: algorithms and applications. CRC Press.

[22] The Geomblog, "Choosing the number of clusters I: The Elbow Method," 7th of March 2007, http://blog.geomblog.org/2010/03/this-is-part-of-occasional-series-of.html

[23] Davies, D.L., and Bouldin, D.W.: 'A Cluster Separation Measure', IEEE Transactions on Pattern Analysis and Machine Intelligence, 1979, 2, pp. 224-227

[24] Peter J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," Journal of Computational and Applied Mathematics, Volume 20, November 1987, Pages 53–65.

[25] Thomas, J. C. R. (2013, November). New Version of Davies-Bouldin Index for Clustering Validation Based on Cylindrical Distance. In V Chilean Workshop on Pattern Recognition.

[26] Grira, N., Crucianu, M., & Boujemaa, N. (2004). Unsupervised and semi-supervised clustering: a brief survey. A Review of Machine Learning Techniques for Processing Multimedia Content, 1, 9-16.

[27] https://archive.ics.uci.edu/ml/datasets.html?format=&task=clu&att=&area=&numAtt=&numIns=&type=&sort=nameUp&view=table, accessed 29 June 2015

[28] JAIN, A.K., MURTY, M.N., and FLYNN, P.J.: 'Data Clustering: A Review', ACM Journal Computing Surveys (CSUR), 1999, 31, (3), pp. 264-323

[29] MacQueen, J.B.: 'Some methods for classification
multivariate observations'. Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability1967 pp. Pages

[30] http://www.rapidminer.com/, accessed 6 July 2015

[31] Ackerman, M., and Ben-David, S.: 'Measures of clustering quality: A working set of axioms for clustering.', in D. Koller, D.S., Y. Bengio, & L. Bottou (Ed.): 'Advances in Neural Information Processing Systems'

[32] Jain, A.K.: 'Data clustering: 50 years beyond K-means.', Pattern Recognition Letters, 2010, (31)

[33] Strehl, A., Ghosh, J., and Mooney, R.: 'Impact of similarity measures on web-page clustering'. Proc. the Workshop of Artificial
Web Search, AAAI 2000, Austin, Texas pp. Pages

[34] Bolshakova, N., and Azuaje, N.: 'Cluster validation techniques for genome expression data', Signal Processing, 2003, 83, pp. 825-833

[35] Ketchen, D.J., and Shook, C.L.: 'The application of cluster analysis in Strategic Management Research: An analysis and critique', Strategic Management Journal, 1996, 17, (6), pp. 441-458

[36] Asteriadis, S., Karpouzis, K., Shaker, N., and Yannakakis, G.N.: 'Towards detecting clusters of players using visual and gameplay behavioral cues.', Procedia Computer Science, 2012, 15, pp. 140-147