Unsupervised Learning and Market Basket Analysis in Market Segmentation

Iromi R Paranavithana, *Member, IAENG*, Thashika D Rupasinghe, and Daniel D Prior

***Abstract*—Identifying and classifying market segments often involves a 'best-guess' for marketing and sales managers. This leads to a set of assumptions regarding the type and importance of customer variables and associated data. The age of 'big data' now provides an opportunity to start with the data and to then work backwards – i.e. to 'let the data tell you'. While these techniques have the advantages of reducing user bias, they are still at their infancy. In this study, we develop a two-stage approach using a large set of point-of- sale (POS) data to i) segment a retail market and ii) identify relative segment purchase probabilities. Stage one involves an unsupervised learning approach based on three purchase characteristics (Recency, Frequency and Monetary value – RFM) and product attributes to identify segments in the customer dataset. Stage two involves Market Basket Analysis (MBA) to determine the likely probabilities of purchase behaviors for each segment. Given these outcomes, we argue that marketing and sales managers have a more robust method for identifying market segments and associated purchase behaviors through predictive analytics.**

***Index Terms*—artificial intelligence; big data; customer relationship management; marketing analysis; retail**

I. INTRODUCTION

MANY firms seek alignment between themselves and their operating environments. This involves harmonizing the market offering of the firm with market requirements. However, without a coherent target, this process is challenging. To address this issue, Marketing Managers have traditionally engaged in market segmentation, whereby they identify a set of groups of individuals with consistent preferences and buying profiles [1] as bases for targeting their marketing efforts.

Recent market segmentation approaches seek to eliminate these issues through inductive methods. Prime examples of this include discrete choice-based approaches [2][3] and probabilistic purchase behavior market segmentation frameworks [5]. However, most of these approaches rely on limited consumer samples and a priori specification of segmentation variables.

Manuscript received March 02, 2020; revised December 28, 2020.

M. P. I. R. Paranavithana is a PhD candidate of University of Wollongong, Australia. She is also with the Department of Information and Communication Technology, University of Ruhuna, Sri Lanka. (phone: +94713329428; e-mail: irpmp897@uowmail.edu.au).

T. D. Rupasinghe is a former senior lecturer in the Department of Industrial Management, University of Kelaniya, Sri Lanka. She is now with the Silueta (PVT) Ltd as a General Manager-Industrial Engineering. (e-mail: ThashikaR@masholdings.com).

D. D. Prior is a Senior Lecturer in Management in the School of Business at UNSW Canberra. He is also a member of UNSW Cyber and has held visiting appointments at Cranfield University, UK, and the University of Texas at Austin, USA. (e-mail: d.prior@adfa.edu.au).

In this study, we develop a market segmentation approach based on unsupervised learning and market-based analysis. Unsupervised learning is a type of machine learning concept used to draw inferences from datasets consisting of input data without labeled responses [1]. Market Basket Analysis (MBA), also known as association rule mining, is one of the widely used methods to discover customer purchasing patterns. MBA can be used to extract co-occurrences from the transactional databases in retail stores [3].

The study represents an application of soft computing techniques to address a business need faced by the decision makers, and consequently, identifies a new approach which is an indirect method of market segmentation and analysis.

II. RELATED WORK

A. Customer Segmentation

Existing customer segmentation methods can be classified into methodology-oriented and application-oriented approaches. Methodology oriented approaches use techniques such as statistical methods, Genetic Algorithms, and fuzzy sets to classify data into homogeneous clusters. Most studies use methodology-oriented approaches and modify a data clustering technique or use two or more data mining techniques to obtain more accurate homogeneous segments [2][5][6][7][8]. As study reported by the reference [9] developed a two-phase approach for cross national market segmentation strategy by integrating data mining and statistical methods. The first phase used a multi group confirmatory factor analysis which is a statistical method to test the variance between national clustering factors. The second phase used a Self- Organizing Maps (SOM) which is a data mining method to develop clusters within each nation [9]. Further, the reference [7] introduced a novel clustering algorithm based on a Genetic algorithm to segment the online shopping market effectively.

But a range of methodological decisions in segmentation studies can undermine the validity of results. For an example underlying data can be of low quality, the sample can be insignificant, and the number of variables can be high [10][11]. These facts cause for entirely random grouping of consumers [11]. Further, the relation between sample size and the number of variables causes methodological problems [10]. The random grouping of consumers leads to inaccurate customer segmentation Application-oriented research studies look for optimum methods for segmenting problems based on applications domains [2]. Based on the past literature, application- oriented approaches define and create new variables for clustering methods or use different combinations of variables in the clustering process [2][5][12][13].

According to the literature, most widely applicable input variables are Lifetime Value (LTV), Recency, Frequency, and Monetary (RFM) [5] [6][12] [14].

RFM model has been employed in many practical scenarios which includes financial and nonprofit organizations such as banking industry [6], on-line industries [15], and telecommunication industries [16]. Furthermore, RFM model can be applied to calculate customer value, customer lifetime value and can be used to segment customers. Numerous studies have used the RFM model to calculate CLV. The reference [17] introduced a new methodology for product recommendation that combined group decision-making and facts mining techniques by utilizing Analytic Hierarchy Process (AHP), clustering and association rule mining techniques. Apart from that, they implemented RFM to evaluate CLV. Four strategies had been compared in their study, particularly the weighted-RFM approach, non-weighted RFM approach, the non-clustering method, and the standard collaborative filtering (CF) approach. Weighted-RFM technique considers the relative importance of the RFM variables through AHP, even as the non-weighted RFM technique does no longer. The non-clustering technique makes an association rule-based totally advice before clustering.

B. Association Rules

Association is a well-known data mining technique; where a pattern is extracted from a dataset by using the relationship of the product on other products which exist in the same transaction [3]. For example, MBA which is an association technique used to identify the products purchased by a customer together. The association rules can be categorized based on the strategy to traverse the search space and its strategy to determine the support values of the item sets [18].

MBA helps retailers to find out commonly purchased products to manage inventory, optimize store layout and identify cross selling opportunities available to maximize profit [19][20]. When considering an application point of view, the MBA applications are uncommon within retail and marketing contexts [21]. The studies carried out previously used unsupervised learning- based artificial neural network (ANN) to map categories onto a set of market basket prototypes. These studies represent consumer-oriented representation of cross category dependencies or purchase interrelationships [22][20][23].

This study uses the association rules to make probabilistic decisions of customers purchasing a product as the second phase of the study.

III. METHODOLOGY

A. Dataset

The study used UCI Machine Learning Repository which contains 541,909 instances with 8 attributes [4]. The dataset consists of both numerical and categorical attributes. It contains a transnational dataset which has transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers. The dataset has multivariate, sequential, and time-series characteristics while attributes are integers and real [26].

B. Overview of the Methodology

The description of the product, quantity of the items purchased, price of the items, and the date of the purchase were extracted from the dataset. Then, the dataset split into product related attributes and non-product related attributes. Text mining was performed to extract the product categories from the product descriptions. The recency, frequency, and monetary values of each product were computed using the attributes. For example, customers with high frequency, high monetary, and high recency of buying a product implies that the customer has a greater purchasing experience with that product.

The non-product related data which is RFM values used to identify customer clusters by performing clustering techniques. The literature suggests, K-means, Ward's Minimum Variance method, and EM clustering algorithms as the most appropriate techniques for clustering [27][28]. These algorithms used to identify the possible clusters in the selected POS dataset. Then, the performance evaluates using the internal validation indices - connectivity, silhouette width, and Dunn index [29]. The customer buying behavior profiles created for each the clustering algorithm based on the results. Subsequently, on the customer- product based clusters, MBA was performed to choose the product(s) that customer may purchase after t time of their previous purchase. The Fig. 1 illustrates the overview of the methodology.

Product based Clusters

The data preprocessed as follows prior to text mining.

Removal of Punctuation: All punctuation marks added to the description removed. The set of forbidden characters includes all punctuation marks and brackets.

Removal of Digits: All the numeric associated with the descriptions removed because the numbers do not contribute to the meaning of the text.

Removal of Stop words: Stop words such as articles (a, an, the), prepositions (in, for, from, with, within etc.) and conjunctions (and, or, but etc.) removed.

Stemming: Stemming is used to reduce the words to its 'root' or 'stem'. For an example, include, includes, including - all these words can be reduced to the root form "include". This helps to determine the similarities and differences of items.

Feature Frequency: Feature Frequency is the number of items which have the same features. This is used as weights for both clustering methods.

Then, Ward's minimum variance method, K-means method and EM algorithms applied to the sub-datasets to identify the optimal clustering algorithms and the number of clusters. The most appropriate clustering algorithm selected and was used to identify the existing product categories.

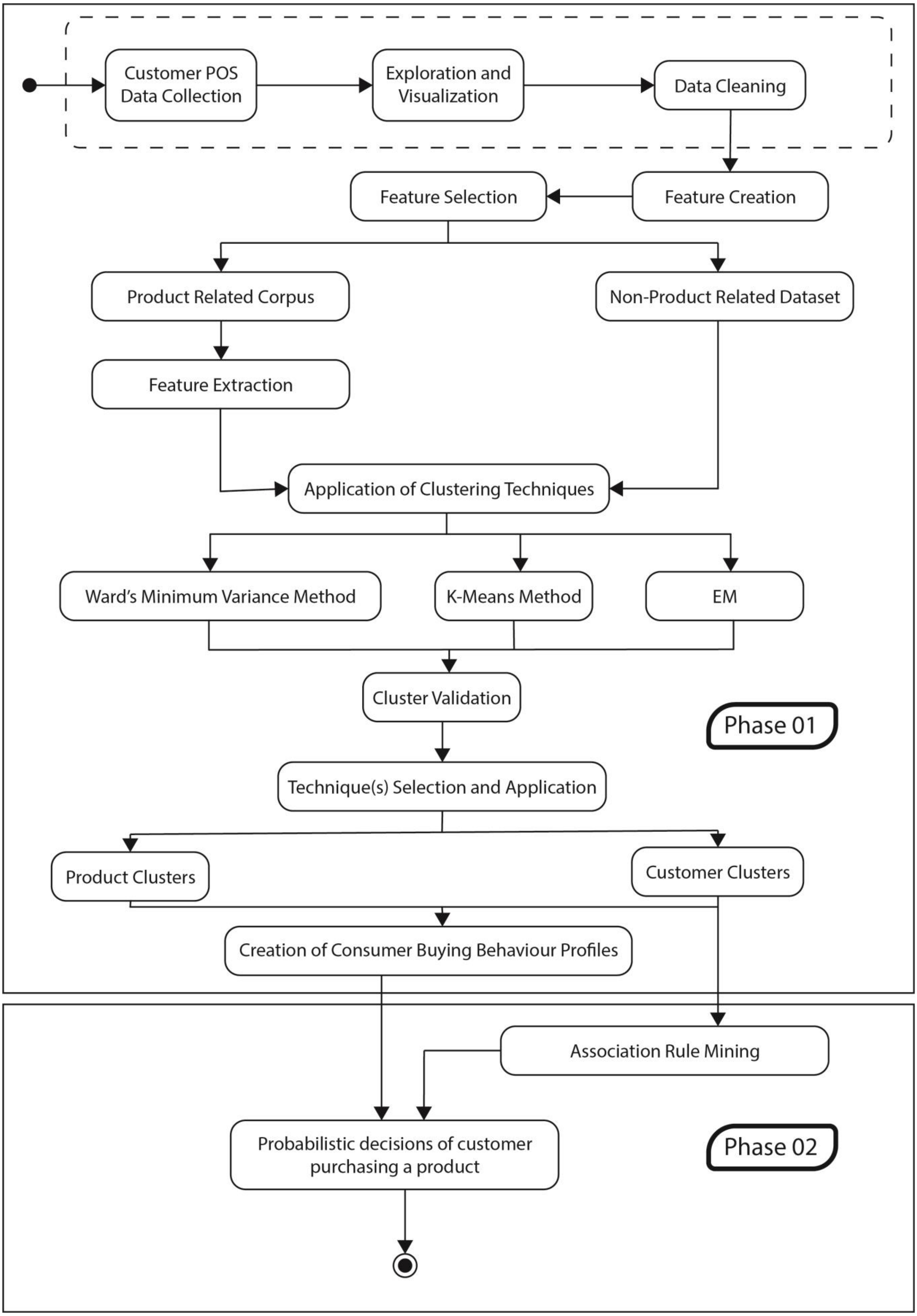

Fig. 1 The overview of the methodology

ISBN: 978-988-14049-2-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Customer based clusters.

The three variables collectively known as RFM, are often used in customer segmentation for marketing purposes. This study utilizes the customer segmentation based on the RFM along with the three algorithms namely K-means, Ward's Minimum Variance method and EM. Subsequently, the Pareto principle applied to determine the top customers. Therefore, the segments have been created by looking at the top customers who have contributed to 80% of annual sales for the year. The customer data has been preprocessed and transformed into the three input variables to reduce positive skewness and then was standardized.

A. *Customer-Product based clusters*

After identifying the product categories, the customers have been clustered according to the products they purchased by considering their RFM values. The three algorithms have been used and the best technique was selected based on the performances. The MBA was used to determine the products purchased by each customer.

After identifying the clusters of the customer with respect to RFM, it is worthwhile to identify the products that customer may purchase during their next visit to the retail outlet. For that purpose, association rule mining-based Apriori algorithm was implemented to identify the associated rules specific to the products in concern. Then associations between products have been identified and the probabilistic decisions of customers have been made.

IV. RESULTS

A. *Product based clusters*

These three different measures calculated for each clustering technique for different numbers of clusters (basically 2-6 clusters). The best clustering technique and the number of clusters selected based on the performance.

At first, product descriptions in the dataset converted into a DTM to identify the possible product categories. After that, K-means and Ward's Minimum Variance Method applied to the DTM. Then, connectivity of the clusters identified based on the distance of the terms in the DTM that placed in the same cluster as their nearest neighbors in the data space. The cluster with minimum connectivity was selected because the minimum connectivity emphasizes the closeness of the data point to the cluster which it belongs to.

The ratio of minimum distance of the observations to the maximum distance between the observations in particular clusters used to calculate the Dunn Index. The cluster with maximum ratio was selected using the Dunn Index. Because the maximum value of the ratio emphasizes that the number of times the cluster consists of the observations which has a minimum distance contained within the maximum distance between the observations in particular cluster is high. This shows that the cluster which has maximum Dunn index is more homogeneous than others.

The average Silhouette width calculated using the average distance between an observation in the cluster and the observations in the nearest neighboring cluster. The silhouette value of the observation is calculated using the average distance between an observation and all other observations in the same cluster, and the average distance between an observation and the observations in the "nearest neighboring cluster". The average silhouette width which lies between the -1 shows poorly clustered observations while the silhouette width which lies between the 1 shows well clustered observations. Therefore, the cluster which has maximum average silhouette width selected. After that, the optimal number of clusters has been selected which has minimum connectivity, maximum Dunn index and the maximum average silhouette width.

The measures of those indices are illustrated in Table I.

TABLE I
UNITS FOR INTERNAL VALIDATION OF THE PRODUCTS

Clustering Method	Measurement	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Hierarchical	Connectivity	2.92	5.85	8.95	13.14
	Dunn	0.78	0.81	0.70	0.64
	Silhouette	0.18	0.14	0.06	0.05
K-means	Connectivity	7.26	9.54	13.89	23.89
	Dunn	0.62	0.55	0.57	0.57
	Silhouette	0.06	0.07	0.07	0.07

According to the internal validation measures, the optimal method is the Hierarchical method (Connectivity = 2.92, Dunn index = 0.81 and, Silhouette width = 0.18).

B. *Customer clusters*

The customers have been clustered according to the RFM value to identify the customer profiles. First, data has been visualized to understand the data. The Figure 5 depicts the outcome that is probably most interested in, and customer monetary value is plotted on the y-axis. Frequency of purchases is on the x-axis, and chose to represent the third variable, recency of purchase, by color- coding the data points. Finally, included the 80/20 rule segments to map those designations on to customer monetary value, frequency, and recency. The clusters obtained for the customers after the visualization have been evaluated by the internal validation indices which are connectivity, Dunn index and silhouette width. The measures of those indices are illustrated in Table II.

TABLE II
INTERNAL VALIDATION OF THE CUSTOMERS

Clustering Method	Measurement	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Hierarchical	Connectivity	7.44	9.76	9.760	11.26
	Dunn	0.02	0.02	0.03	0.03
	Silhouette	0.59	0.56	0.60	0.59
K-means	Connectivity	1.77	7.48	13.67	1.77
	Dunn	0.02	0.02	0.02	0.02
	Silhouette	0.63	0.63	0.63	0.62
EM	Connectivity	8.78	1.93	16.22	34.78
	Dunn	0.01	0.02	0.01	0.01
	Silhouette	0.29	0.16	0.47	0.52

According to the internal validation measures, the optimal method is the K-means (Connectivity = 1.77, Dunn index = 0.02 and, Silhouette width = 0.63).

C. *Market Basket Analysis of the Products*

The optimal three customer profiles are high value customers, medium value customers and low value customers. These customer clusters are employed in the MBA process to retrieve the products that customers may purchase, after purchasing a relevant product. For that, Apriori algorithm is utilized to find the associations that exist among the products purchased by customers in the clusters. Products listed by the invoice numbers for association rule mining. After that, Apriori algorithm applied to identify the associate rules under the minimum support of 0.001 and the minimum confidence of 0.8.

TABLE III
TOP 5 RULES DISCOVERED USING MBA

Customer ID	LHS	RHS	Support	Confidence	Lift
567935	{picture dominoes}	{please one-person metal sign}	0.0017	1.00	12.5
543300	{red hanging heart light holder}	{hanging heart zinc t-light holder}	0.0017	1.00	53.4
555243	{red hanging heart t-light holder}	{set of 12 forks set of 3 butterfly cookie cutters}	0.0017	1.00	21.7
552453	{herb marker chives}	{watering can blue elephant}	0.0026	1.00	10.9
560448	{heart of wicker small}	{heart filigree dove large}	0.0068	0.89	3.5

LHS = Left hand Side, RHS = Right Hand Side

The plot shows that rules with a high lift typically have low support. The top 5 rules among the identified 75709 rules are as depicted in the Table III.

For example, the first top association rule can be described as “If a customer with customer ID 567935 buys PICTURE DOMINOES, the customer has 100% of likely to buy PLEASE ONE-PERSON METAL SIGN also”. These association rules determine the product that customer may purchase, after the customer purchases one or more products. Further, association rules used to find the target items that customers may purchase after purchasing a certain product. The Table IV shows the top 5 results of the products that customers may purchase after purchasing a product x.

TABLE IV
TOP 5 PRODUCTS THAT CUSTOMER MAY PURCHASE AFTER PURCHASING A PRODUCT X

Customer ID	LHS	RHS	Support	Confidence	Lift
543300	{set of 12 forks set of 3 butterfly cookie cutters}	{heart filigree dove large}	0.0213	0.46	1.8
567846	{set of 12 forks set of 3 butterfly cookie cutters}	{no singing metal sign}	0.0153	0.33	2.4
552479	{set of 12 forks set of 3 butterfly cookie cutters}	{heart decoration painted zinc}	0.0153	0.33	1.7
552453	{set of 12 forks set of 3 butterfly cookie cutters}	{area patrolled metal sign}	0.0145	0.31	2.3
551269	{set of 12 forks set of 3 butterfly cookie cutters}	{filigree heart with butterfly}	0.0136	0.30	1.8

LHS = Left hand Side, RHS = Right Hand Side

V. DISCUSSION

The analysis of the impact of clusters and its quality by using partitioning method centroid algorithm K-Means, Ward’s Minimum Variance method, and Expectation Maximization is important to consider. The study considers the online retail dataset for the clustering of two to five cluster groups which is done by taking the results of K-Means, Ward’s Minimum Variance method and Expectation Minimization for the need of the comparison. The experimental results show that the K-means algorithm performs better than Ward’s Minimum Variance method and Expectation Maximization algorithms in the customer clustering and customer and product-based clustering. Further, the results show that Ward’s Minimum Variance method performs better than K-means algorithm and Expectation Maximization algorithms in the product clustering which is taken text mining into consideration. This work convinces that the optimal clustering method is depending on the dataset and the characteristics of that dataset which is going to apply the clustering technique. The study evaluates the performance of Support Vector Machine, Random Forest Classifier and Naive Bayes algorithms in classification of the customers into different customer segments. It reveals that Naive Bayes algorithms perform better in terms of accuracy, precision, and recall. The study conducted to identify the potential of different machine learning techniques to perform market segmentations. Here, both supervised and unsupervised learning techniques have been compared and based on the nature of the dataset available retailers can use the optimal algorithm. For an example, where the dataset has predefined labels to identify the categories, supervised learning approach can be used.

The associated products for different customer profiles can be identified using MBA. Therefore, MBA with the help of association rules could be employed to identify the consumer buying behavior. This would be helpful for retailers to set up the retail shop and expand the business in future. The study reveals that the association rule learning can be used in MBA to probabilistic decisions regarding the customer purchases. The association rules are used to identify relations when there are no predictor variables exist in the data available for analysis. Further, customer satisfaction is at the center point in Customer Relationship Management (CRM) and to achieve this, it is necessary to find the main interest of the customer in a particular product. The K-means clustering used in the study creates the clusters by considering the customers' RFM values with which retailers easily can identify the high, medium, and low value customers from a large dataset. The clustering of customers with K-means helps to identify the products purchased by customers in different levels. Further, the placement of products according to the customer levels in retail with the help of clustering will not only be effective and impressive but also be helpful to achieve the goal of market-basket analysis.

The proposed approach can be differentiated from the previous literature because this study is useful in failure classification in retail stores or in supermarkets. The current failure classification methods do not have a definitive way to determine the number of clusters into which a set of program executions should be divided. This study further opens paths to identify the consumer purchasing behavior by combining the clustering and market basket analysis and can be reported as the first instance where several forms of supervised and unsupervised learning techniques has been applied together for a large retail dataset to address a business need rather than using one learning category. Therefore, this approach reduces the effect of bias and provides marketing and sales managers a robust method to segment their customers in terms of their business needs.

REFERENCES

[1] M. Celebi and K. Aydin, Unsupervised learning algorithms. Springer, 2016.
[2] CHAN, "Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer", Expert Systems with Applications, vol. 34, no. 4, pp. 2754-2762, 2008. Available:
[3] 10.1016/j.eswa.2007.05.043.
[4] Y. Chen, K. Tang, R. Shen and Y. Hu, "Market basket analysis in a multiple store environment", Decision Support Systems, vol. 40, no. 2, pp. 339-354, 2005. Available: 10.1016/j.dss.2004.04.009.
[5] "UCI Machine Learning Repository: Online Retail Data Set", Archive.ics.uci.edu, 2020. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Online Retail.
[6] Cheng and Y. Chen, "Classifying the segmentation of customer value via RFM model and RS theory", Expert Systems with Applications, vol. 36, no. 3, pp. 4176-4184, 2009. Available: 10.1016/j.eswa.2008.04.003.
[7] N. Hsieh, "An integrated data mining and behavioral scoring model for analyzing bank customers", Expert Systems with Applications, vol. 27, no. 4, pp. 623-633, 2004. Available: 10.1016/j.eswa.2004.06.007.
[8] K. KIM and H. AHN, "A recommender system using GA K- means clustering in an online shopping market", Expert Systems with Applications, vol. 34, no. 2, pp. 1200-1209, 2008. Available: 10.1016/j.eswa.2006.12.025.
[9] Liu, C. Lai and W. Lee, "A hybrid of sequential rules and collaborative filtering for product recommendation", Information Sciences, vol. 179, no. 20, pp. 3505-3519, 2009. Available: 10.1016/j.ins.2009.06.004.
[10] S. Lee, Y. Suh, J. Kim and K. Lee, "A cross-national market segmentation of online game industry using SOM", Expert Systems *with Applications*, vol. 27, no. 4, pp. 559-570, 2004. Available: 10.1016/j.eswa.2004.06.001.
[11] S. Dolnicar, "Using cluster analysis for market segmentation -typical misconceptions, established methodological weaknesses and some recommendations for improvement", Australian Journal of Market Research, vol. 11, no. 2, pp. 5-12, 2003. [Accessed 20 February 2020].
[12] Ernst and S. Dolnicar, "How to Avoid Random Market Segmentation Solutions", Journal of Travel Research, vol. 57, no. 1, pp. 69-82, 2017. Available: 10.1177/0047287516684978.
[13] H. Hwang, T. Jung and E. Suh, "An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry", Expert Systems with Applications, vol. 26, no. 2, pp. 181-188, 2004. Available: 10.1016/s0957-4174(03)00133-7.
[14] S. Kim, T. Jung, E. Suh and H. Hwang, "Customer segmentation and strategy development based on customer lifetime value: A case study", Expert Systems with Applications, vol. 31, no. 1, pp. 101-107, 2006. Available: 10.1016/j.eswa.2005.09.004.
[15] J. McCarty and M. Hastak, "Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression", Journal of Business Research, vol. 60, no. 6, pp. 656-662, 2007. Available: 10.1016/j.jbusres.2006.06.015.
[16] Y. Li, C. Lin and C. Lai, "Identifying influential reviewers for word-of-mouth marketing", Electronic Commerce Research and Applications, vol. 9, no. 4, pp. 294-304, 2010. Available: 10.1016/j.elerap.2010.02.004.
[17] S. Li, L. Shue and S. Lee, "Business intelligence approach to supporting strategy-making of ISP service management", Expert Systems with Applications, vol. 35, no. 3, pp. 739-754, 2008.
[18] Available: 10.1016/j.eswa.2007.07.049.
[19] Liu and Y. Shih, "Integrating AHP and data mining for product recommendation based on customer lifetime value", Information & Management, vol. 42, no. 3, pp. 387- 400, 2005. Available: 10.1016/j.im.2004.01.008.
[20] J. Hipp, U. Güntzer and G. Nakhaeizadeh, "Algorithms for association rule mining --- a general survey and comparison", ACM SIGKDD Explorations Newsletter, vol. 2, no. 1, pp. 58-64, 2000. Available: 10.1145/360402.360421.
[21] R. Blattberg, B. Kim and S. Neslin, Database marketing. 2008.
[22] N. Gutierrez, "Demystifying Market Basket Analysis DM Review Special Report", 2006.
[23] Vindevogel, D. Van den Poel and G. Wets, "Why promotion strategies based on market basket analysis do not work", Expert Systems with Applications, vol. 28, no. 3, pp. 583-590, 2005. Available: 10.1016/j.eswa.2004.12.019.
[24] R. Decker, "Market Basket Analysis by Means of a Growing Neural Network", The International Review of Retail, Distribution and Consumer Research, vol. 15, no. 2, pp. 151- 169, 2005. Available: 10.1080/09593960500049332.
[25] R. Decker and K. Monien, "Market basket analysis with neural gas networks and self-organising maps", Journal of Targeting, Measurement and Analysis for Marketing, vol. 11, no. 4, pp. 373-386, 2003. Available: 10.1057/palgrave.jt.5740092.
[26] I. Cadez and P. Smyth, Bayesian Predictive Profiles with Applications to Retail Transaction Data, 1st ed. 2016.
[27] J. Qiu, "A Predictive Model for Customer Purchase Behavior in E-Commerce Context", 2014. [Online]. Available: http://aisel.aisnet.org/cgi/viewcontent.cgi? article =1018&context=pacis2014.
[28] Z. Wang and X. Lei, "Study On Customer Retention Under Dynamic Markets", in Second International Conference on Networks Security, Wireless Communications and Trusted Computing, 2010.
[29] Nathiya and Punitha, "An Analytical Study on Behavior of Clusters Using K Means, EM and K* Means Algorithm", International Journal of Computer Science and Information Security, vol. 7, no. 3, , pp. 185-190, 2010.
[30] K. Tsiptsis and A. Chorianopoulos, Data Mining Techniques in CRM. Chichester, West Sussex, U.K: Wiley, 2009.
[31] X. Wu et al., "Top 10 algorithms in data mining", Knowledge and Information Systems, vol. 14, no. 1, pp. 1-37, 2007. Available: 10.1007/s10115-007-0114-2.