Comparison of Behavioral Customer Segmentations for Private Labels using Clustering Algorithms

Carlos Hernández
Departamento de Procesos Industriales
Universidad Católica de Temuco
Temuco, Chile
carlos.hernandez.zavala@uct.cl

Magaly Sandoval
Departamento de Ingeniería Industrial y Sistemas
Universidad de La Frontera
Temuco, Chile
magaly.sandoval@uforntera.cl

Abstract

An increasingly common practice among retailers is the creation of their own brands. The so-called private labels allow them to compete on price to attract new customers. The goal is to offer a quality similar to that of the traditional brands but at a lower price range. In this research, clustering algorithms based on machine learning techniques are applied to carry out a behavioral customer segmentation for private labels.
The research has been completed in four stages: analysis, design, development, and discussion. During the analysis, 1.073 customer loyalty surveys are preprocessed and analyzed. During the design, 23 questions are selected to design experiments. The clustering algorithms used in the investigation are Simple K-Means Algorithm (SKMA) and Expectation-Maximization Clustering (EMC).
Experimental results reveal clear differences in the performance of the selected algorithms. For example, when 4 clusters are predefined, SKMA distributes the instances according to the following proportions: 43%, 26%, 15%, and 15%. EMC, instead, distributes the instances in 18%, 23%, 31%, and 29%.
In conclusion, the results show that both algorithms, SKMA and EMC, are effective and useful to segment customers based on their preferences. However, the peculiarities of their clustering strategies gives rise to significantly different clusters.

Keywords
Behavioral Segmentation, Machine Learning, Clustering Algorithms, Private Labels

1. Introduction
As a result of the recent coronavirus crisis has accelerated the digital transformation in many companies making business environment more complex than ever before. The incorporation of new technologies and process digitization has increased the competitiveness in every aspect and across different industries. Nowadays, knowing and understanding customers has become crucial to survive in this new post-pandemic era. The first step to build a successful relationship with customers is characterizing them according to specific criteria. Thus, by means of gathering basic information about them it would be possible to develop marketing strategies that are more focused and personalized, and consequently, more effective. Currently, there are informatics tools and algorithms that can be used to accomplished the customer characterization.
According to Kotler (2001), companies must identify the most attractive market segments to design and carry out effective marketing campaigns. They also have to consider that customers have specific needs, desires, attitudes, purchasing power and practices. Kotler (2001) states that market segments can be discovered by means of analyzing the attribute hierarchy that consumers employ when choosing a given brand.
An appropriate segmentation might lead to the right customers for the company. In other words, those ones that can help produce more profit. Focusing on profitable customers is an effective manner to optimize the resources allocated for attracting, retaining, and regaining customers (Khajvand et al., 2011). In particular, behavioral

customer segmentation classifies them into different homogeneous groups to represent specific behaviors. This works analyses the relationship between customers and private labels.
According to Gómez et al. (2016) studies on private labels have been mostly focused on characterizing customers to identify the factors influencing labels' perception and purchase intention. The authors identified two groups. First are the personal characteristics and then the characteristics of products, private labels, and store".
Based on both the investigation by Fernández and Martínez (2004) and the research by Soberman and Parker (2004), Buil et al. (2007) argue that at the beginning private labels were associated with products of low quality and low sale prices. However, the situation has changed over the years and today private labels are well positioned, having a positive perception and being associated to products of a quality comparable to that offered by traditional brands but at lower prices.
Private labels are no longer targeted exclusively to low-income consumers that are highly sensitive to sale prices but also to consumers whose main motivation is not necessarily economical (Hansen et al. 2006).
According to Buil et al. (2007), the reason consumers purchase private labels is because the benefit exceed the cost and that deal prone customers seeking for special offers are likely to purchase private labels.

1.1 Objective
To compare behavioral customer segmentations for private label products by means of applying machine learning techniques and clustering algorithms.

2. Literature Review

2.1 Market segmentation
Market segmentation is usually referred as the process of classifying or grouping customers with different characteristics and behavior to implement more efficient marketing strategies and tactics (Kotler and Armstrong, 1999). Some of the most studied segmentation types in the literature are:

- Behavioral : brand loyalty, buyer journey stage, price sensitivity, purchasing style, etc.
- Benefit : customer service, quality, etc.
- Demographic : age, education level, gender, income, family members, status, religion, etc.
- Geographic : country, city, district, etc.
- Psychographic : hobbies, interests, lifestyle, etc.

2.2 Behavioral segmentation
Behavioral segmentation groups customers based on pattern of their habits and practices. There are four mayor types of behavioral segmentation: purchase behavior, occasion-based purchases, benefits sought, and customer loyalty

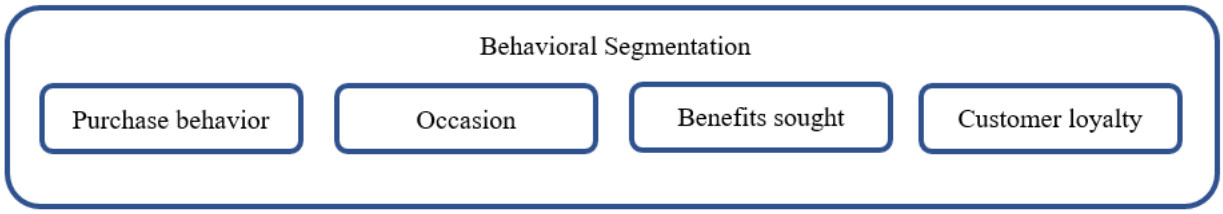

Figure 1. Behavioral segmentation

Table 1. Types of behavioral segmentation

Segmentation type	Description
Purchase and usage behavior	This type of segmentation is useful to understand the stage of the buyer's journey in which customers are, and therefore to determine appropriate purchase triggers (Figure 2).
Occasion or timing	This segmentation classifies customers based on the specific dates or moments of the day they purchase.
Sought benefits	This segmentation groups customers based on value proposition they look for. Understanding the benefits consumers expect can help redefined strategies and tactics to satisfice each segment.

Customer loyalty	This segmentation classifies customers based on the level of loyalty, which can be measured in terms of the purchase frequency.

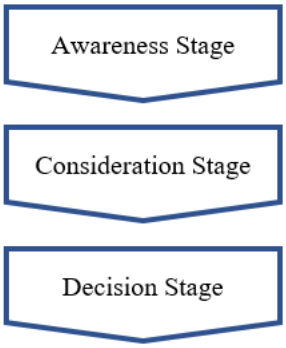

Figure 2. Consumer's journey-stage model

2.3 Machine learning

Machine learning is usually referred as the branch of artificial intelligence (AI) that uses algorithms to find patterns and to learn from datasets through experience. There several types of machine learning algorithms: supervised, unsupervised, and reinforcement algorithms. In supervised learning, the training is carried out using labelled datasets. This means that the class or the value to be predicted is included in the dataset so it can be used for training. In the case of unsupervised learning, instead, the desired class is not known. The machine learning algorithms used in this work have been implemented with WEKA 3.8.5. (Witten et al., 2017)

2.4 Clustering algorithms

Clustering algorithms are used to discover patterns and to group data points. They are a particular case of machine learning algorithms employed to analyzed unlabeled datasets. Some of the most popular clustering algorithms are:

- Agglomerative hierarchical clustering
- Density-based spatial clustering
- Expectation-maximization clustering (EMC)
- Simple K-Means algorithm (SKMA)
- Mean-shift clustering

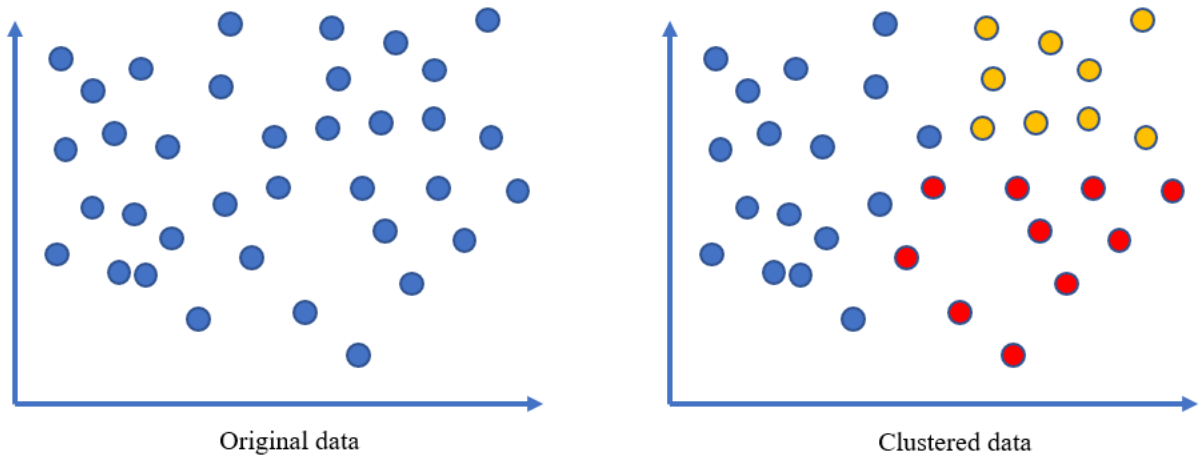

Figure 3. Clustering data

2.5 Simple k-means algorithm (SKMA)

K-means clustering is an unsupervised machine learning algorithm that is used to categorize unlabeled data. The algorithm works iteratively and assign every new instance to one of the existing K clusters. The classification criteria is based on the feature similarity of the instances.

2.5.1 Number of clusters

Finding the best clustering scheme might be useful when optimization is the goal. It can be found by means of varying k, distance measures, and clustering method. There are several methods to determine the optimal number of clusters. Some of the most common are: average silhouette method, elbow method, and gap statistic method.

2.6 Expectation-maximization clustering (EMC)

The expectation-maximization (EM) algorithm is an iterative procedure for the maximum likelihood estimate of a parametric distribution. A particular case of this algorithm is the parameter estimation of a Gaussian Mixture Model (GMM) when the generating Gaussian of each observation is unknown, commonly known as Expectation-Maximization Clustering (EMC) (Garriga et al., 2016; Jung et al., 2014).

3. Methods

This investigation is carried out following a 4-stage model: analysis, design, construction, and discussion (Figure 4).

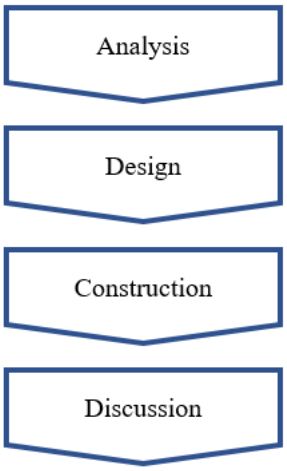

Figure 4. Four-stage model

3.1 Analysis

The work commenced with a complete review of a survey about purchasing habits and preferences supermarket customers. The survey was conducted in Temuco (Chile), approximately 221.000 inhabitants, and it considered five supermarket chains each of them targeting different demographic market segments.

The original questionnaire was based on previous well-documented surveys created by renowned authors. It has 69 questions from which 24 were selected for the present investigation. Selected questions are exclusively related to private labels and are organized in 5 domains according to their nature (Table 2). All answers to the questions are given in a scale from 1 to 7.

Table 2. Domains for private labels

	Domain	Questions
D.1	Quality perception	7
D.2	Price	7
D.3	Value proposition	3
D.4	Purchase Satisfaction	2

D.5	Familiarity with private labels	5

Although having gathered data from 5 supermarket chains, only one of them is being considered in this investigation. The selected chain is the middle of the price range, it counts with stores in different districts of the city and it has a well-established private label with a variety of products.

3.2 Design

The original unlabeled dataset, a matrix of 1073 rows (instances) by 24 columns, is prepared to be clustered by means of applying the algorithms SKMA and EMC. During the experiments, different clustering schemes will be tested and compared. Survey domains and their corresponding questions are presented below (Table 3 and Table 4).

Table 3. Questions per domain

Domain	Q.8	Q.9	Q.10	Q.11	Q.12	Q.13	Q.14	Q.15	Q.24	Q.25	Q.26	Q.27
D.1	✓	✓	✓						✓			✓
D.2				✓						✓	✓	
D.3												
D.4												
D.5					✓	✓	✓	✓				

Table 4. Questions per domain (continuation)

Domain	Q.28	Q.29	Q.30	Q.31	Q.32	Q.36	Q.37	Q.38	Q.39	Q.41	Q.62	Q.63
D.1	✓	✓										
D.2			✓	✓	✓		✓					
D.3								✓	✓	✓		
D.4											✓	✓
D.5						✓						

3.3 Construction

The goal is to carry out a customer behavioral segmentation for private labels offer by a specific supermarket chain by means of applying clustering algorithms. Since behavioral segmentation differs significantly from other types of segmentation, the experiments considers a range of parameter combinations.

The answers to the original survey are discretized in values from 0 to 7. The summary of the answers to que questions are presented below (Table 5)

Table 5. Questions' answer summary

Question ID	Value 0	Value 1	Value 2	Value 3	Value 4	Value 5	Value 6	Value 7
Q.8	40	149	64	87	215	181	166	171
Q.9	43	106	51	78	257	192	176	170
Q.10	41	117	55	105	246	180	184	145
Q.11	41	63	31	95	220	160	193	270
Q.12	45	134	58	131	272	210	172	51
Q.13	46	161	61	103	240	194	194	74
Q.14	49	175	69	96	226	172	191	95
Q.15	46	208	79	128	250	149	143	70
Q.24	50	76	53	94	263	203	181	153
Q.25	22	53	37	123	218	252	221	147
Q.26	23	40	47	170	220	244	189	140
Q.27	46	83	37	104	289	186	174	154
Q.28	45	113	45	142	284	168	146	130
Q.29	25	116	70	192	235	189	137	109
Q.30	50	35	40	102	238	171	210	227
Q.31	53	106	76	165	323	171	91	88

Q.32	54	90	57	152	316	175	127	102
Q.36	46	185	52	84	213	174	175	144
Q.37	47	55	40	97	276	189	195	174
Q.38	49	95	46	139	323	164	140	117
Q.39	47	112	42	77	296	177	169	153
Q.41	25	92	54	91	144	278	230	159
Q.62	48	96	44	65	266	223	173	158
Q.63	25	13	13	43	68	176	412	323

3.4 Discussion

For the purposes of this investigation tow of the most common clustering algorithms are employed. Namely, SKMA and EMC, which belong to a broader family usually called Gaussian mixture models.

The first algorithm studies, SMKA, requires the definition of K centroids and the iterations until certain degree of convergence to a local minimum is achieved. The latter, EMC, is meant to solve some of the weaknesses of SKMA.

Rather than focusing on the accuracy of the classification, due to the nature of the behavioral segmentation the interest is set on the number of clusters and on the number of data points in each of them.

4. Data Collection

Finding the optimal number of clusters for a given dataset requires the application of optimization algorithms. However, it might be difficult to handle when many clusters are defined. The following tables present the data (instances) classification distribution when clustering algorithms are force to generate 1, 2, 3, and 4 clusters (Table 6 and Table 7).

Table 6. SKMA with k=2, 3, 4, and 5

K	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5	
	%	Instances	%	Instances	%	Instances	%	Instances	%	Instances
2	67	714	33	359						
3	47	499	32	345	21	229				
4	43	461	26	280	15	166	15	166		
5	38	411	27	292	7	76	9	94	19	200

Table 7. EMC with 2, 3, 4, and 5 clusters

K	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5	
	%	Instances	%	Instances	%	Instances	%	Instances	%	Instances
2	48	514	52	559						
3	16	172	43	460	41	441				
4	18	188	23	246	31	328	29	311		
5	30	323	21	227	4	46	26	282	18	195

5. Results and Discussion

SKMA can help determine the number of clusters that minimized distance between each data point and its closest centroid. It is an iterative method. EMC can do it too. However, this approach can give rise to a larger number of clusters. Some of them having just a bunch of data points. For practical purposes, it might not be efficient to design a marketing strategy for too many segments. In the case of SKMA only 4 clusters concentrate between 10% and 20%. Same situation happens with EMC (Table 8).

Table 8. EMC and SKMA with 12 clusters

# Cluster	**EMC**		**SKMA**	
	%	Instances	%	Instances
1	13	138	18	191
2	3	37	2	24
3	18	189	4	46
4	4	45	6	62
5	7	76	14	151

6	11	119	3	36
7	5	49	15	164
8	6	69	10	109
9	8	86	6	66
10	12	126	4	39
11	7	75	6	67
12	4	45	4	45
13	2	19	7	73

5.1 Numerical Results

Since behavioral segmentation is not as evident as other types of segmentation, the characterization of the resulting clusters or market segments demands additional labor. The success of a marketing strategy depends on the understanding of customers' behavior when selecting either private labels or traditional brands.

The original dataset was taken from an extensive supermarket customer survey whose questions are organized in several domains (Table 3 and Table 4). The following tables present the distribution of the clusters generated by SKMA and EMC separated by domain (Table 9, Table 10, Table 11, Table 12, and Table 13).

Table 9. Clustering for domain D.1 (Q.8, Q.9, Q.10, Q.24, Q.27, Q.28, and Q.29)

	SKMA		EMC	
Cluster	**%**	**Instances**	**%**	**Instances**
C.1	54	582	4	46
C.2	30	324	37	398
C.3	6	64	40	427
C.4	10	103	19	202

Table 10. Clustering for domain D.2 (Q.11, Q.25, Q.26, Q.30, Q.31, Q.32, and Q.37)

	SKMA		EMC	
Cluster	**%**	**Instances**	**%**	**Instances**
C.1	38	407	19	200
C.2	21	226	26	279
C.3	25	263	39	415
C.4	16	177	17	179

Table 11. Clustering for domain D.3 (Q.38, Q.39, and Q.41)

	SKMA		EMC	
Cluster	**%**	**Instances**	**%**	**Instances**
C.1	63	673	32	342
C.2	18	195	33	352
C.3	9	93	18	188
C.4	10	112	18	191

Table 12. Clustering for domain D.4 (Q.62, and Q.63)

	SKMA		EMC	
Cluster	**%**	**Instances**	**%**	**Instances**
C.1	69	741	53	564
C.2	15	166	19	208
C.3	3	31	24	253
C.4	13	135	4	48

Table 13. Clustering for domain D.5 (Q.12, Q.13, Q.14, Q.15, and Q.36)

	SKMA		EMC	
Cluster	**%**	**Instances**	**%**	**Instances**
C.1	44	467	4	46

C.2	21	227	25	263
C.3	26	277	41	439
C.4	10	102	30	325

6. Conclusion

Different from demographic segmentation where differences between groups or clusters are evident, behavioral segmentation demands a more extensive analysis to characterize the difference between market segments. This investigation is based on an extensive supermarket customer survey carried out in Temuco (Chile) to establish a market segmentation using the well-known SKMA and EMC clustering algorithms.

Clustering algorithms are widely used to identify pattern and classify unlabeled data by means of grouping similar data points in clusters that shares some degree of similarity.

Both SKMA and EMC are iterative optimization methods to cluster data points. Depending on the needs and the number of iteration is possible to determine the optimal number of clusters. Although it might not be the most practical approach, especially when the number of clusters is too big. In this research the optimal number was found to be 12 clusters, from which only 4 concentrates mora than 10% of the data points. Instead of that, a fixed number of clusters, from K=2 to K=4, was analyzed. In the case of four clusters. While SKMA cluster sizes are 43%, 26%, 15%, and 15%., EMC cluster sizes are 218%, 23%, 31%, and 29%.

An additional analysis was carried out to determine whether the question of each domain had an influence on the cluster scheme. The difference in the sizes of the resulting clusters confirms that the segmentation depends significantly on the nature of the questions taken into consideration.

In conclusion, both clustering algorithms SKMA and EMC can help segmenting customers based on their behavior and preferences towards private labels. However, behavioral segmentation requires an additional work to characterize properly the found segments.

References

Kotler, P. and Keller, K., Marketing Management, 13th edition, Pearson Prentice-Hall, 2009.

Buil, I., Martínez, E., and Montaner, T., El comportamiento del consumidor ante la promoción de ventas y la marca de distribuidor, Universia Business Review, vol. 16, pp. 24-25, 2007.

Fernández, A., and Martínez, E., Las marcas del distribuidor y el consumidor español, Distribución y Consumo, p. 12-25, 2004.

Garriga J., Palmer J., Oltra A., and Bartumeus F., Expectation-Maximization Binary Clustering for Behavioural Annotation, *PLoS ONE*, vol. 11, no. 3 2016.

Gómez, M., Paiva, G., and Schnettler, B., Private Labels in Chile: Influential Factors in the Purchase Intention, Handbook of Research on Strategic Retailing of Private Label Products in a Recovering Economy, Hershey: IGI Global, 2016.

Hansen, K., Singh, V., and Chintagunta, P., Understanding store brand purchase behavior across categories, Marketing Science, vol. 25, no. 1, pp. 75-90, 2006.

Jung Y., Kang m., and Heo M., Clustering performance comparison using K-means and expectation maximization algorithms, *Biotechnology & Biotechnological Equipment*, vol. 28, pp. 44-48, 2014.

Khajvand, M., Zolfaghar, K., Ashoori, S., and Alizadeh, S., Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. Procedia Computer Science, vol. 3, pp. 57-63, 2011.

Kotler, P., Dirección de Mercadotecnia. Análisis, Planeación, Implementación y Control, 8th edition, Pearson Education, 2001.

Soberman, D., and Parker, P., Private label: psychological versioning of typical consumer products, International Journal of Industrial Organization, vol. 22, pp. 849-861, 2004.

Witten, I., Frank, E., Hall, M., and Pal, C., *Data Mining: Practical Machine Learning Tools and Techniques*, 4th Edition, Morgan Kaufmann, Cambridge, 2017.

Biographies

Carlos Hernández is an industrial engineer, consultant, and university professor. He earned Licentiate Degree in Engineering from Universidad de La Frontera, Temuco, Chile, Master of Sciences in Computational Engineering and Doctor of Engineering from Technische Universität Braunschweig, Brunswick, Germany. He is the author of

several scientific and engineering articles. He has taught lectures in Discrete Event Simulation, Supply Chain Management, Engineering Economics, Corporate Finances, Financial Engineering, Business Analytics, Data Mining and Machine Learning for engineering students. He has developed a professional career working for large multinational companies (PricewaterhouseCoopers, BHP Billiton, and Merck Sharp & Dohme). He also worked as a scientific researcher in the Institut für Produktionsmesstechnick at TU Braunschweig, Germany. His research interests include manufacturing process simulation, transportation systems simulation, supply chain design and simulation, and machine learning for finances. He is a member of IEOM.

Magaly Sandoval is an industrial engineer, consultant, and university professor. She earned Licentiate Degree in Engineering from Universidad de La Frontera, Temuco, Chile, and MBA from Universidad Austral de Chile, Valdivia, Chile. She is a Babson SEE certified mentor. She has taught lectures in Project Planning & Management, Project Evaluation, Engineering Economy, Business Management, International Business Management, and Entrepreneurship for engineering students. During her academic tenure she has been appointed in different management positions and has mentored over a hundred students. Her research interests include business models, project management, and entrepreneurship & intrapreneurship.