Category role aided market segmentation approach to convenience store chain category management

Shuihua Han, Yongjie Ye, Xin Fu*, Zhilong Chen

Department of Management Science, School of Management, Xiamen University, Xiamen 361005, China

ARTICLE INFO

Article history:
Received 5 December 2012
Received in revised form 11 September 2013
Accepted 21 September 2013
Available online 16 October 2013

Keywords:
Category management
Market segmentation
Convenience store chain management
Strategic decision support

ABSTRACT

Category management (CM) plays an increasingly important role in retailing management, as it aids retailers to increase their core competitiveness, maximise profits and ensure a good long-term customer relationship. This technique has been successfully applied to diverse large manufacturers and wholesale retailers. However, it remains a challenging task to directly employ the CM technique in convenience store (CVS) chain(s). This is because CVS chains are often distributed in a variety of areas, each store has impulsive consumers, and the traditional market segmentation attributes (e.g. consumer age, salary, and background) are difficult to collect under such circumstances. This makes it impractical to apply one general CM solution to all CVS chains. Hence, it is crucial to segment a market region and then apply customised CM solutions to the corresponding segments. This paper presents an innovative market segmentation model which is driven by category-role (CR), for the first time, to support CM in CVS chains. A new similarity measure (named HCSim()) and an improved weighted fuzzy K-means clustering algorithm (WFKM) are developed in an effort to cluster the CVSs. The usefulness and applicability of this study is illustrated by means of an empirical study to provide marketing strategy decision support. The derived results are also discussed and compared with existing methods.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

Convenience store (CVS) refers to a small store that stocks everyday items such as groceries, beverages and snacks. Convenience is its distinguished strength that attracts consumers. Modern CVSs normally utilise store chain management methods and represent one of the fastest growing segments in the retail market. Such rapid growth brings increasing competition amongst CVS retailers. In recent years, category management (CM) [3,21,38] has gradually gained popularity in these companies to increase their core competitiveness, maximise profits and ensure long-term healthy customer relationship.

CM is a business fundamental that uncovers untapped potential to be explored. It aims to analyse consumer purchasing behaviour and stock the products that consumers are most likely to buy. The basic idea of CM is managing product categories as business unit and customising them to satisfy customers’ needs [18]. More specifically, the defined categories can be used to target the consumer groups and to gain a better understanding of their needs. From early 1990s, CM can be divided into two types: product-centric CM [6,51] and consumer-centric CM [1,14,31]. Product-centric CM uses historical transaction data to gain insights into product movements, so as to identify and employ product attributes to determine a product category. This widely used approach straightforwardly reflects the movement of products, and historical transaction data is relatively easy and cheap to collect. However, this approach reflects the overall movement of products, and it is hard to further identify customers’ future needs, if no detailed consumer data is available. In consumer-centric CM, the demographics, attitudes, interests, shopping occasions and historical purchase records of consumers are taken into account to segmenting and targeting consumers who share similar purchase patterns. Subsequently, a marketing plan for each distinct consumer group is developed. This approach is driven by the understanding that consumer behaviour can change the design of a product category. A distinct advantage of this approach is that product category is designed by considering diverse customer groups, so that the derived category structure is transparent and interpretable. Also, marketing plans based on this approach can be more efficient and accurate. However, this in turn naturally requires the presence of consumer data, and if such data is not available, it is quite costly to collect.

Although CM attracts increasing attentions and some initial results have been obtained [11,14,26,43], applying conventional CM methods to CVS chain management remains a challenging task due to its distinguished features. First, CVSs are often distributed in a variety of areas and each store has impulsive consumers. For example, petrol station CVSs can be located at a city centre, the outskirts of a city or on a motorway. Within the same category, different CVSs may have to provide different products to meet the diverse requirements of their consumers. Second, due to the mobility of CVSs' consumers, especially for the CVSs located at the outskirts of a city or on a motorway, it is difficult to collect...
consumer related data (e.g., the demographics, interests, and shopping occasions) by employing the royalty membership programme. Hence, the conventional consumer-centric CM approaches become less appropriate to analyse and group consumers of CVS chains. Third, CVSs are often small and have very limited storage space, such that it cannot provide a large number of product categories. This leads suppliers to invest most of their time and effort on their large retailers, rather than on CVS chains. This is evident in [1] that ACNielsen mainly provides CM support to wholesalers and supermarket chains, i.e. General Mills, Big Y, SUPERVALU, etc.

To overcome the above difficulties, this paper proposes an innovative market segmentation model which employs category roles (CRs), for the first time, to support CM in CVS chains. Due to the mobility of CSVs' consumers, the consumer related data (e.g., the demographics, interests, and shopping occasions) are difficult to collect. The proposed method is still a product-centric approach, but it is equipped with some new features: a) it is a value-centric approach that recognises the values of a category to retailers, consumers and the marketplace; b) it provides a dynamic view, since the growing potential of a category is considered rather than just focusing on a static view.

This new method takes three CR dimensions (importance to retailers, consumers, and the marketplace) into account to segment the market. Distinct market segments can receive a customised CM and marketing strategy. Additionally, the retailer can perceive the market with finer granularity, and better analyse different behaviour patterns. Initially, the historical transaction data provides a full view of the market. Such data are then used to derive a global category index (CIX) for each store. CVSs that share similar CIXs are clustered into the same group by using a new similarity measure, namely HCSim(), and an improved weighted fuzzy K-means clustering algorithm (WEKM). Based on the obtained clustering results, the retailer can design varied CM and marketing strategies for different CVS clusters.

The remainder of this paper is organised as follows: Section 2 reviews the existing approaches in CM and market segmentation. In Section 3, a novel method, the CR-based market segmentation model, is proposed to cluster CVSs. In addition, an improved clustering algorithm which employs HCSim() is also introduced in this section. The applicability and utility of the proposed method is demonstrated in Section 4 via an empirical study. The derived results are analysed and discussed in Section 5. The final section concludes this paper and points out future work directions.

2. Background

2.1. Category management

CM was introduced in 1990 as an organised process to enable retailers to effectively reach consumers, while maximising profits. It aims to find appropriate products and prices for the target customers, so that the customers are satisfied and remain loyal to retailers. Many retailers today have adopted the fundamental eight-step process [3] to conduct CM, including Define category, Assign a category role, Assess a category, Set performance and scorecard, Create market strategy, Choose tactics, Implement the plan and Review.

The biggest change over the years in CM is moving from product-centric [11,23,51] to consumer-centric [1,10,24]. A few researchers [10,11,31,40] have posited that different category characteristics can reflect different consumer needs. They believe that a key tenet of CM is that the retailer should decide the role each category plays in the retailer's overall portfolio. One popular classification schema to assign CR is cross-category quantitative analysis [14,26] which considers the importance to consumers, retailers and the marketplace. The typical four CRs are:

- Destination category: make the store the primary category provider. When consumers would like to purchase products in a certain category, the store is their main choice. This category plays a key role in retailing and distinguishes the store from other competitors. It involves 5%–10% of categories in the store.
- Routine category: make the stores one of the preferred category providers. The price of daily necessities is not sensitive and reasonably stable. This category generates solid profits for retailers, while meeting customers' various daily needs. It involves 50%–70% of categories in the store.
- Occasional/seasonal category: make the store a major category provider when consumers would like to buy a given occasional/seasonal product. The demands for such a category are not stable and are normally short-term. It involves 10%–15% of categories in the store.
- Convenience category: this category helps the store to provide convenience to consumers, so that consumers can buy all they need from a store in one stop. It involves 10%–15% of categories in the store.

CRs are essential to maintain a consistent strategic and tactic plan that provides cohesive market approaches regarding price, promotion and assortment. It is pointed out that the same category may act as a different CR in different stores [43], so that customised category strategies should be implemented in different market segments.

2.2. Market segmentation

Market segmentation was initially defined by Smith in 1956 [42]. It has been well recognised that a mass market strategy has little chance of success in a highly competitive market. The objective of market segmentation is to divide a heterogeneous market into a number of smaller homogeneous markets in which consumers share sufficiently similar characteristics. This helps the retailer to identify the consumer groups, and use resources more effectively to match consumers' needs.

In general, existing market segmentation methods can be categorised into two main approaches [17,47,48]: the first is the priori approach [16,17]. It segments the market according to prior knowledge or speculated factors that are associated with consumers, services or products (such as demographic characteristics, purchase amounts and geographic areas). The second is a widely used approach called post-hoc segmentation. Different from the priori approach, it segments the market by analysing market data. More and more segmentation techniques are becoming available for the post-hoc approach [2,46], including clustering [12], classification and regression trees (CART) [15], self-organising map (SOM) [27], and multi-objective evolutionary algorithms (MOEA) [34].

The selection of segmentation variables is an essential step in building segmentation models. Many scholars generalise segmentation variables from many perspectives, ranging from demographic, geographic, purchasing behaviours to consumer values. Segmentation variables can be classified into two groups: general variables [33,48] and product-specific variables [44,47]. General variables focus on neighbours' attributes (such as geographic, lifestyles and demographics) rather than on individual consumers. Product-specific variables concern consumers' preferences and their responses to products. These variables involve customer demands, intentions and purchasing behaviours.

Since 1956, market segmentation has dramatically expanded its applications to various domains, such as the online shopping market [28], the printed circuit board industry [8], and the tourist market [30]. However, the application of market segmentation is still at a relatively early stage for the retail industry [5,35], particularly when taking consumer demands into consideration. Currently, retailers normally employ unified or traditional segmentation attributes to segment the market, while ignoring underlying and potential consumer characteristics and demands. In particular, little research has been conducted in CVS segmentation. For CVSs, the concept of market segmentation can be employed and expanded to cluster stores into...
different market segments. This is because the store features (e.g., store location, and historical transaction data) usually carry useful but hidden information regarding consumer behaviours and preferences that cannot be easily exploited by traditional market segmentation approaches. By employing clustered market segmentation, different marketing strategies can be applied to market segments that consist of stores that share similar features, so that consumer requirements can be better satisfied and more profits can be achieved. This paper mainly concerns the applicability of using market segmentation techniques to segment CVSs to achieve better CM.

2.2.1. RFM segmentation model

RFM (recency, frequency, and monetary) is a purchase-behaviour based segmentation model [39,44]. It utilises consumers' historical transaction data to analyse and segment the market. This model is easy to use and has been widely employed in a variety of marketing areas [19,25,43]. The foundation of the RFM model includes three purchase attributes:

- The most recent purchase timestamp, “R” (recency): this is the time difference between the last purchase timestamp and the current timestamp of a given product. A smaller time interval indicates that this product is more likely to be purchased by this consumer.
- The purchase frequency, “F” (frequency): given a period of time, the number of purchase times of a given product. A higher frequency implies that this product is more popular.
- The total purchase volume, “M” (monetary): given a period of time, the total amount of money this consumer spent. Obviously, a higher value indicates that a consumer is more valuable to a retailer, as he/she contributes more towards the retailer’s profits.

In the RFM model, the involved attributes are based on historical transaction data; hence they are relatively handy to collect. However, there exist several drawbacks when applying RFM to CVS market segmentations. First, the analysis of the RFM model is not based on an individual category but instead reveals consumer behaviour for the whole store. Second, for a CVS, the profits of different categories can be quite different. Some products show low profit which only exists for providing convenience to consumers. The RFM model only concerns the total purchase volume, without considering the profits in each category. Therefore, it is hard to reflect the importance of a given category to retailers. Third, given that the analysis is based on historical transaction data, the model may not be suitable to predict future consumer behaviour. To sum up these drawbacks, the RFM is more appropriate to analyse consumer behaviour for individual stores rather than CVS chains.

3. CR-based market segmentation approach

Considering the drawbacks of the RFM model and the problem at hand, this section proposes an innovative CR-based market segmentation model, which is particularly suitable for CVS chains.

3.1. Category index (CIX)

Cross-category quantitative analysis [3,14,26] has been widely used in the assignment of CRs. In this method, three dimensions (as shown in Table 1) should be taken into consideration and they are described below:

- The importance to consumers: refers to consumers’ demands for this category. If a category is more desirable to consumers, then it is more essential and important to consumers. Some indicators exist to measure this dimension, such as the purchase frequency, annual expenditure and purchase volume.
- The importance to retailers: refers to the sales performance of a given category. Since CVSs are often small and have very limited storage space, the categories that are included in CVSs should be carefully selected by retailers. The selection of categories can result in diverse CVS sales performances. Therefore, sales volume, sales gross profits, and efficiency plateau are often used to measure this dimension.
- The importance to the marketplace: while the first two dimensions rely on the current status of categories, this dimension reflects the importance of a given category in the future. It refers to consumers’ future demand and purchase propensity. In order to seize market opportunities, CVSs should not only focus on different categories current market status, but also need to consider the categories future development. Some indicators can be used to measure this dimension, such as the average growth rate of a given category, purchase trends, and market share trends.

Due to the availability of dimension measures, five segmentation attributes are selected to aggregate a new category index (CIX):

$$CIX_i = \omega_F F_i + \omega_N N_i + \omega_S S_i + \omega_R R_i + \omega_G G_i$$

where CIX is the index value of category i, and $F_i, N_i, S_i, R_i$ and $G_i$ represent the average sales frequency, average sales volume, average sales revenue, average gross profit and average growth rate of category i, respectively. $\omega_F, \omega_N, \omega_S, \omega_R$ and $\omega_G$ represent the corresponding attribute weights and $\omega_F + \omega_N + \omega_S + \omega_R + \omega_G = 1$. Different from the conventional RFM model, this CR-based model not only reveals consumer desirability and purchase behaviour for a given category, but also takes the retailer’s mission and market development direction into consideration to provide a global index value (CIX), and to measure category contributions for retailers.

The new CIX that considers consumers, retailers and the marketplace is employed to build an innovative market segmentation model (as shown in Fig. 1). In the new model, the market is divided into several segments where similar CVSs are grouped together.

3.2. CR-based CVS clustering

Clustering is a well-established unsupervised learning method used to explore patterns in a given problem. It aims to group a set of objects that share similar attributes into the same cluster. This technique has been widely used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. Clustering is also a commonly used technique in market segmentation. Thanks to the rapid development of IT in retailing, transaction data becomes easier to collect and this makes it possible to employ clustering techniques to build the newly proposed CR-based market segmentation model.

Suppose that a retailer has N CVSs distributed in different areas and each store has M product categories. Clustering analysis aims to group similar CVSs into the same cluster with regard to their CIXs that are represented by the aforementioned five segmentation attributes, including $F, N, S, R$ and $G$. Although a CVS sells a relatively small number of categories, it is still often required to provide 20–30 categories to meet consumer needs. Therefore, the problem at hand is an $N \times M$ high-dimensional clustering problem. Solving such a high-dimensional problem, conventional clustering methods may easily suffer from the curse of dimensionality [50]. The traditional algebraic distance
space, a revised category similarity measure, similarity between category structures and the similarity in data

\[ HCsim_{sj} \]

total number of sub-categories or products of

\[ De \]

subset of category set \( C \).

\[ c_{ij} \]

where

\[ \text{definition 1.} \]

given a category set \( \mathcal{C} = \{ c_1, c_2, \ldots, c_M \} \), the number of included categories is \( |\mathcal{C}| \). for each category \( c_i \in \mathcal{C} \), its category structure is represented as \( c_i = \{ p_{i1}, p_{i2}, \ldots, p_{in}, \ldots, p_{ik}\} \), where \( p_{ij} \) indicates the \( j \)th sub-category or product of the category \( c_i \) and \( k = |c_i| \) is the total number of sub-categories or products of \( c_i \).

**Definition 2.** Suppose that \( S = \{ s_1, s_2, \ldots, s_N \} \) is a set of \( \text{CIX} \) values of all \( \text{CVSs} \) that belong to a retailer. given a \( \text{CVS} j \), its \( \text{CIX} \) values of its included categories in a given period of time can be represented as:

\[ s_j = \{ c_j, \text{CIX}_j \} = \{ (c_{j1}, \text{CIX}_{j1}), (c_{j2}, \text{CIX}_{j2}), \ldots, (c_{jm}, \text{CIX}_{jm}) \} \]

(2)

where \( m \) is the number of categories of \( \text{CVS} j \) and \( \mathcal{C} \) is a non-empty subset of category set \( \mathcal{C} \).

**Definition 3.** Given two \( \text{CVSs} j \) and \( k \), their \( \text{CIX} \) values are written as \( s_j = \{ c_j, \text{CIX}_j \} \) and \( s_k = \{ c_k, \text{CIX}_k \} \). \( p_{jk}(c_i) \) represents the ratio of common sub-categories or products in category \( c_i \) that are shared by store \( j \) and \( k \).

\[ p_{jk}(c_i) = \frac{|c_j \cap c_k|}{|c_j|} \]

(3)

where \( c_j \) and \( c_k \) are the sub-category or product collection of category \( c_i \) in stores \( j \) and \( k \), respectively, and \( c_j \subset C_j \) and \( c_k \subset C_k \).

Based on the features of \( \text{CVS} \) category data, coupled with both the similarity between category structures and the similarity in data space, a revised category similarity measure, \( HCsim() \), is defined as:

\[ HCsim(s_j, s_k) = \frac{\sum_{i=1}^{L} p_{jk}(c_i) \cdot |\text{CIX}_j - \text{CIX}_k|}{\sum_{i=1}^{L} (1 - X(c_i))} \]

(4)

where \( L = |C_j| = |C_k| \) is the total number of categories that are sold by stores \( j \) and \( k \). This similarity measure has the following properties:

- The greater similarity value indicates that the two \( \text{CVSs} \) are more similar to each other.
- \( 0 \leq HCsim(s_j, s_k) \leq 1 \), if and only if two \( \text{CVSs} \) are identical, then \( HCsim(s_j, s_k) = 1 \). In contrast, when \( HCsim(s_j, s_k) \rightarrow 0 \), \( \text{CVSs} \) \( j \) and \( k \) are totally different.
- This measure is symmetric, i.e., \( HCsim(s_j, s_k) = HCsim(s_k, s_j) \).

This distance function differs from traditional distance functions. Its value reflects the similar dimensions (categories, in this case) between any two stores and is less affected by dissimilar categories, thus, the more categories shared by two stores, the more similar the two stores are, and vice versa. It matches the natural human-being perceptions well.

By applying the properties of \( HCsim() \), for simplicity, the \( HCsim() \) measure can be rewritten as:

\[ HCsim(s_j, s_k) = \frac{\sum_{i=1}^{L} p_{jk}(c_i) \cdot (1 - X(c_i))}{\sum_{i=1}^{L} (1 - X(c_i))} \]

(5)

where \( X(c_i) = \begin{cases} 0 & c_i \in C_j \text{ and } c_i \in C_k \\ 1 & \text{otherwise} \end{cases} \)

Guided by the CR theory, different categories play different roles in a \( \text{CVS} \). Additionally, even the same category can play different roles in different \( \text{CVSs} \). Therefore, when clustering the \( \text{CVSs} \) it is not reasonable to assume a given category is equally important to all \( \text{CVSs} \). It is required to dynamically assign different weights to different categories, according to their own category attributes.

An improved weighted fuzzy \( K \)-means clustering algorithm (\( \text{WFKM} \)) that employs the previously proposed \( HCsim() \) similarity measure is...
proposed herein to cluster CVSs. The objective function, that WFKM aims to minimise, can be defined as follows:

\[
J_{\text{WFKM}} = \sum_{i=1}^{K} \sum_{j=1}^{N} \mu_{ij}^{m} HCsim_{\nu_{ij}}(s_{j}, q_{i})
\]

\[
= \sum_{i=1}^{K} \sum_{j=1}^{N} \sum_{l=1}^{L} a_{ijl}^{\mu_{ij}} \left( 1 - \frac{p_{ij}(\mu_{ij})}{1 + |C_{X_{ij}} - C_{X_{kl}}|} \right) X(c_{i}) \sum_{i=1}^{K} \sum_{j=1}^{N} (1 - X(c_{j}))
\]

This objective function satisfies:

\[
\mu_{ij} \in [0, 1], \sum_{j=1}^{N} \mu_{ij} = 1, \text{ and } 1 \leq j \leq N.
\]

\[
\omega_{ik} \in [0, 1], \sum_{k=1}^{K} \omega_{ik} = 1, \text{ and } 1 \leq i \leq K.
\]

where \( K \) is the number of clusters, \( L \) is the number of categories; \( q_{i} \) is the centroid of the \( i \)-th category, \( \mu_{ij} \) indicates the membership degree of \( s_{j} \) belonging to cluster \( i \). \( \omega_{ik} \) is the newly introduced important weight of category \( k \) to cluster \( i \), and \( p_{ij}(\mu_{ij}) \) is the category similarity between stores \( s_{j} \) and \( q_{i} \), also \( m > 1 \) and \( \beta > 1 \).

By applying Eq. (7) and the Lagrangian function, the objective function can be rewritten as:

\[
J_{\text{WFKM}} = \sum_{i=1}^{K} \sum_{j=1}^{N} \mu_{ij}^{m} HCsim_{\nu_{ij}}(s_{j}, q_{i}) = \sum_{i=1}^{K} \lambda_{i} \left( \sum_{j=1}^{N} \mu_{ij} - 1 \right).
\]

where \( \lambda_{i} \) is the Lagrangian parameter.

By applying Eq. (9), \( \mu_{ij} \) can be written as:

\[
\mu_{ij} = \left( \sum_{r=1}^{K} HCsim_{\nu_{ij}}(s_{j}, q_{r}) \right)^{- \frac{1}{m}}.
\]

In this equation, when \( s_{j} = q_{i} \), it indicates that sample point \( s_{j} \) and cluster centroid \( q_{i} \) are identical, i.e., \( HCsim_{\nu_{ij}}(s_{j}, q_{i}) = 0 \). In this case, the membership degree of \( s_{j} \) belonging to cluster \( i \) is equal to 1, whereas the membership degrees of \( s_{j} \) belonging to other clusters are equal to 0.

Similarly, \( \omega_{ik} \) can be calculated as:

\[
\omega_{ik} = \frac{1}{\sum_{r=1}^{N} \sum_{j=1}^{N} \mu_{ij}^{m} HCsim_{\nu_{ij}}(s_{j}, q_{r})}\left( 1 - \frac{\sum_{r=1}^{N} \mu_{ij}^{m} HCsim_{\nu_{ij}}(s_{j}, q_{r})}{1 - X(c_{i})} \right)^{- \frac{1}{m}},
\]

when \( X(c_{i}) = 1 \), this indicates category \( c_{i} \) is not available in both CVSs, so that this category contributes less in regard to clustering. Set

\[
1 - \frac{p_{ij}(\mu_{ij})}{1 + |C_{X_{ij}} - C_{X_{kl}}|} X(c_{i}) = 1,
\]

for the objective function \( J_{\text{WFKM}} \), then set:

\[
R_{i} = \sum_{j=1}^{N} \mu_{ij}^{m} \omega_{ik} \left( 1 - \frac{p_{ij}(\mu_{ij})}{1 + |C_{X_{ij}} - C_{X_{kl}}|} \right) X(c_{i}) \sum_{i=1}^{K} \sum_{j=1}^{N} (1 - X(c_{j}))
\]

\[
T_{ij} = \sum_{j=1}^{N} \mu_{ij}^{m} \omega_{ik} \left( 1 - \frac{p_{ij}(\mu_{ij})}{1 + |C_{X_{ij}} - C_{X_{kl}}|} \right) X(c_{i}) \sum_{i=1}^{K} \sum_{j=1}^{N} (1 - X(c_{j}))
\]

where \( R_{i} \) reflects the total fuzzy similarity degree between data objects and the centroid of category \( i \), and \( T_{ij} \) represents the fuzzy similarity degree between \( s_{j} \) and the centroid of category \( i \).

To obtain the minimal value of \( J_{\text{WFKM}} \), \( R_{i} \) is required to obtain the minimal value in each category. Given the category vector \( \mathbf{q}_{i} = \{q_{1}, q_{2}, \ldots, q_{m} \} \), the purpose is to find a \( q_{i} \) that satisfies the following equation to be the new cluster centroid.

\[
R_{i} = \min_{q_{i}} \sum_{j=1}^{N} \mu_{ij}^{m} \omega_{ik} \left( 1 - \frac{p_{ij}(\mu_{ij})}{1 + |C_{X_{ij}} - C_{X_{kl}}|} \right) X(c_{i}) \sum_{i=1}^{K} \sum_{j=1}^{N} (1 - X(c_{j}))
\]

Given a dataset \( \mathbf{q}_{i} \), the algorithm to find the cluster centroid, \( q_{i} \), can be summarised below:

**Algorithm 1. Find_centric(\( \mathbf{q}_{i} \))**

1: Randomly choose \( n_{t} \) sample points from \( \mathbf{q}_{i} \) as the training dataset.
2: Calculate the sum of \( T_{ij} \), between \( n_{t} \) and the remaining data points in \( \mathbf{q}_{i} \), which can be represented as:

\[
ST_{t} = \sum_{j=1}^{N} T_{ij} = \sum_{j=1}^{N} \sum_{i=1}^{K} \mu_{ij}^{m} \omega_{ik} \left( 1 - \frac{p_{ij}(\mu_{ij})}{1 + |C_{X_{ij}} - C_{X_{kl}}|} \right) X(c_{i}) \sum_{i=1}^{K} \sum_{j=1}^{N} (1 - X(c_{j}))
\]

3: Choose \( n_{t} \) with minimal \( ST_{t} \) as the testing dataset.
4: Calculate the sum of \( T_{ij} \), between testing dataset and the remaining data points in \( \mathbf{q}_{i} \), which can be represented as:

\[
ST_{t} = \sum_{j=1}^{N} T_{ij} = \sum_{j=1}^{N} \sum_{i=1}^{K} \mu_{ij}^{m} \omega_{ik} \left( 1 - \frac{p_{ij}(\mu_{ij})}{1 + |C_{X_{ij}} - C_{X_{kl}}|} \right) X(c_{i}) \sum_{i=1}^{K} \sum_{j=1}^{N} (1 - X(c_{j}))
\]

5: Make the testing dataset with the minimal \( ST_{t} \) to be the centroid of category \( i \).

Similar to the conventional fuzzy C-means clustering algorithm, WFKM can be quickly converged. However, the cluster results are quite sensitive to the initial chosen cluster centroids. Thus, the selection of the initial centroid plays a crucial role for clustering. Randomly choosing the initial centroids may result in diverse clustering results that may converge to local optima. To solve this, this work employs a data density based algorithm [32] to select the initial cluster centroids. This algorithm is based on the assumption that the centroid of each cluster has more intensive data density than other data points. It aims to find the top \( K \) intensive data density points as the initial cluster centroids.

**Algorithm 2. WFKM**

1: \( m = 0, J_{\text{WFKM}} = \xi (\xi \text{ is a large positive constant}) \)
2: choose \( K \) initial cluster centroids \( q_{i}^{m}, \omega_{ik}^{m} = \frac{1}{m} (1 \leq i \leq K, 1 \leq k \leq L) \)
3: set \( m = m + 1 \)
4: cluster data points to each cluster, based on \( HCsim(\) \)
5: calculate \( \mu_{ij}^{m} \), by applying Eq. (10)
6: update \( \omega_{ik}^{m} (1 \leq i \leq K, 1 \leq k \leq L) \) by applying Eq. (11)
7: while \( J_{\text{WFKM}} - J_{\text{WFKM}}^{m} > \alpha (\alpha \text{ is a pre-defined small constant}) \) and \( m < M \) do
8: \( m = m + 1 \)
9: cluster data points to each cluster, based on \( HCsim(\) \)
10: calculate \( \mu_{ij}^{m} \), by applying Eq. (10)
11: update \( \omega_{ik}^{m} (1 \leq i \leq K, 1 \leq k \leq L) \) by applying Eq. (11)
12: end while
In summary, the CR-based CVS clustering approach can be illustrated in Fig. 2. The data processing step aims to filter out missing and noisy data points, and derives a reliable dataset for further use. After that, by the joint use of \( HCsim() \) and WFKM, the CVSs are grouped into different clusters, each cluster reflecting a different market segment. The derived clustering results can then be used to support the development of effective marketing strategies. Ultimately, analysis of the clustering results helps to gain a better understanding of market segments and consumers. It also provides solid evidence for the implementation of CM and appropriate marketing strategies for CVSs.

4. Data and empirical results

An empirical study is conducted to verify the proposed model in this section. A dataset from PetroChina is used. It contains transaction records that are collected from 82 petrol CVSs during January–June, 2009 in Guangdong province, China. The dataset contains 2 category levels: 21 major categories and 95 sub-categories, and 3456 products. In addition, QlikView and MATLAB are employed in this work to support the clustering analysis.

4.1. Data preprocessing

Initially, noisy, erroneous and missing data is removed from the dataset. In this case, 7 out of 82 CVSs (which only contain transaction data for January and February) are filtered out. Meanwhile, the Fast food category is also removed due to missing transaction data. Thus, a total of 20 major categories and 75 CVSs remain. This dataset is then imported into a database which includes three types of data: 1) category transaction data (reflects consumer demands and purchase behaviour on a certain category), 2) category distribution data (represents the category definition and structure in different CVSs), and 3) store basic data (represents geographical information of CVSs, etc.). The five segmentation attributes used in the proposed model (i.e. \( F, N, S, R \) and \( G \)) are derived on a monthly basis. To start with, normalisation is performed on these attributes by using the Min–Max method, in which the value of each attribute is mapped onto a range of \([0,1]\).

4.2. Clustering analysis of CVS market segmentation

This paper employs \( v = \frac{D_{max}}{D_{mind}} - 1 \) to validate the applicability of the proposed \( HCsim() \) measure in the given problem. Given a \( d \)-dimensional problem domain, \( D_{max} = \max_{1 \leq i \leq N, 1 \leq j \leq N} HCsim(s_i, s_j) \) denotes the largest distance between any two stores, i.e., the ith store \( s_i \) and the jth store \( s_j \). While \( D_{min} = \min_{1 \leq i \leq N, 1 \leq j \leq N} HCsim(s_i, s_j) \) denotes the shortest distance. Mathematically speaking, with the increase of dimensionality(\( d \)), if \( v \) approaches 0, this indicates that \( D_{max} \) and \( D_{min} \) are close to each other, so that the distance measure is not suitable for the problem at hand [13]. For comparison purposes, the relationships between dimensionality and \( v \) of \( HCsim() \) and \( v \) of traditional city-block distance [9] are illustrated in Fig. 3, respectively. With the increase in dimensionality, the \( v \) of city-block distance approaches 0, while the \( v \) of \( HCsim() \) gradually increases. Therefore, \( HCsim() \) is more appropriate to handle high-dimensional problems.
4.2.1. The clustering analysis process

After the data preprocessing and normalisation, the next step is to determine the weights of the five segmentation attributes. There are several methods to determine the weights of attributes, such as AHP, Delphi, and the entropy method. Since AHP and Delphi methods arrive at the attribute weights by the use of manual assignments, the results tend to be more subjective and more likely to cause manual biases. Also, in this work, it is difficult to find a sufficient number of qualified professional CM managers to manually evaluate these attributes. In contrast, the entropy method is a purely data-driven approach, and the obtained results can be more objective. In information theory, entropy is a measure of the uncertainty in a random variable [41]. In [41], entropy quantifies the expected value of the information contained in a message, and Shannon entropy is the average unpredictability in a random variable, that is equivalent to its information content. Therefore, it is the method used in this paper to determine the weights of the five attributes. The detailed calculation step is as follows:

1. Based on the normalised data in Section 4.1, an \( m \times n \) matrix \( R_{ij} = (r_{ij})_{m \times n}; (i = 1, 2, \ldots, m; j = 1, 2, \ldots, n) \) of the five attributes is built.

2. Convert \( R_{ij} \) into a normalised matrix \( R_{ij}^{'} = (r_{ij}^{''})_{m \times n} \) by using the following equation:

\[
\begin{align*}
    r_{ij}^{''} &= \frac{r_{ij}}{\sum_{n} r_{ij}^{''}}, \quad i = m, \quad j = n. \\
\end{align*}
\]

3. Calculate the output information entropy.

\[
E_{j} = -\frac{1}{\ln m} \sum_{i=1}^{m} r_{ij} \ln r_{ij}, \quad j = n, \\
\text{when } r_{ij}^{''} = 0, \text{ set } r_{ij} \ln r_{ij} = 0.
\]

4. Calculate the five attribute weights \( \omega = (\omega_1, \omega_2, \ldots, \omega_n) \), where

\[
\omega_j = 1 - E_{ij},
\]

The obtained results are shown in Table 2.

Based on Table 2 and collected transaction data, the CIDs (which indicate the global contribution of a given CVS) of different CVSs can be calculated, by the use of Eq. (1). The results are listed in Table 3.

To perform clustering based on the proposed WFKM algorithm, some parameters need to be identified. As shown in Table 4, \( m \) is the number of membership functions in the WFKM, and it controls the degree of overlapping between two clusters (in other words, \( m \) is the smoothing parameter which controls the fuzziness of the clusters). Guided by [22], this paper sets \( m = 2 \) to achieve the optimal results; \( \beta \) is a parameter for attribute weight \( \omega \), and it is required to be greater than 1 [20]. To simplify the calculation in Eq. (11), \( \beta \) is set to be 2; \( M \) and \( \alpha \) are the termination parameters of Algorithm 1. Considering the computation efficiency, \( M \) is set to be 1000 and \( \alpha \) is set to be 0.00005; for parameter \( r \), according to Algorithm 1, in order to find the cluster centroid, \( n_c \) and \( n_t \) sample points are chosen from \( Q \) to be the training dataset and testing dataset, respectively, and usually \( n_c = r \times Q \). As indicated in Fig. 4, when \( r \) is set to be 5%, 10%, 20%, or 30%, with the iterations increases, the objective functions taking different \( r \) values all converge to a stable value (when the number of iterations is \( >5 \)). Thus, this work randomly selected 20% to be the value of \( r \).

For clustering analysis, it is crucial to select the number of clusters \( K \). In the literature, three measures, namely partition coefficient [4], fuzzy entropy [29] and XB coefficient [49] are often employed to assess the clustering performance. For the partition coefficient, a greater index value indicates better performance, while for the XB coefficient and fuzzy entropy, a smaller index value indicates better performance. In this work, based on the collected dataset, these three indexes are calculated by using different \( K \) values, respectively. The obtained results are depicted in Fig. 5. With the increase of \( K \), the index value of the partition coefficient gradually decreases; the index value of fuzzy entropy increases at the beginning, reaches the peak index value when \( K = 3 \), and then slowly decreases; the index value of the XB coefficient has a downward trend when \( K = \{2, 4\} \) and then tends to become stable when \( K = \{4, 10\} \). In addition, the CM manager of PetroChina pointed out that it is quite expensive and challenging to handle and especially implement a large number of different marketing strategies (which reflect different market clusters) concurrently. Guided by Fig. 5 and CM expertise, although the overall effectiveness of \( K = 4 \) and \( K = 9 \) are comparable, \( K = 4 \) is chosen in an effort to achieve optimal clustering results.

4.2.2. Results

In this work, 100 trials are carried out to clustering the CVSs and each of which uses the same initial cluster centroids (38, 1, 4, and 10) that are selected by using a data density based algorithm [32]. To report the best results, only the clustering results of the trial which achieves the minimal XB coefficient are listed in Table 5. In this trial, the

---

**Table 2**

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Category sales frequency (F)</th>
<th>Category sales volume (N)</th>
<th>Category sales revenue (S)</th>
<th>Category sales gross profit (R)</th>
<th>Category sales growth rate (G)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weights</td>
<td>0.2022</td>
<td>0.1834</td>
<td>0.1870</td>
<td>0.2137</td>
<td>0.2137</td>
</tr>
</tbody>
</table>

**Table 3**

<table>
<thead>
<tr>
<th>CVS no.</th>
<th>Category 1</th>
<th>Category 2</th>
<th>Category 3</th>
<th>...</th>
<th>Category 20</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.3895</td>
<td>0.5226</td>
<td>0.4591</td>
<td>...</td>
<td>0.6574</td>
</tr>
<tr>
<td>2</td>
<td>0.3129</td>
<td>0.5377</td>
<td>0.4729</td>
<td>...</td>
<td>0.7216</td>
</tr>
<tr>
<td>3</td>
<td>0.3447</td>
<td>0.5383</td>
<td>0.4844</td>
<td>...</td>
<td>0.7256</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>0.3059</td>
<td>0.3658</td>
<td>0.3356</td>
<td>...</td>
<td>0.4094</td>
</tr>
</tbody>
</table>

**Table 4**

<table>
<thead>
<tr>
<th>Maximum # of iterations (M)</th>
<th># of memberships (m)</th>
<th>Attribute weight (β)</th>
<th>Threshold of initial cluster centre optimisation</th>
<th>Threshold of selected parameters (r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>2</td>
<td>2</td>
<td>0.00005</td>
<td>0.1</td>
</tr>
</tbody>
</table>

---

Fig. 3. The relationship between \( n \) and dimensionality.
clustering parameters that are reported in Table 4, the selected K value, and four optimal selected initial cluster centroids are used to derive such results. The CVSs are grouped into four market segments (Cluster1, Cluster2, Cluster3, and Cluster4). For each cluster, the distribution of included CVSs in different regions is summarised in Table 6.

In this work, each category is assigned a CR according to its CIX, and the occasional/seasonal category takes 10% of categories in a store, the routine category holds 50%, and the convenience category holds 5% of categories, the occasional/seasonal category takes 10%–15% of categories, and the convenience category takes 10%–15% of categories. As a result, 2 out of 20 (10%) categories are identified as the destination category, 12 categories (60%) are considered as the routine category, 3 categories (15%) are the occasional/seasonal category, and 3 categories (15%) are defined as the convenience category. The detailed results, including CIX, category ranking, and CRs are listed in Table 8.

For Cluster1, most CVSs are distributed in the middle and west areas of Guangdong province. Also, most of them are located on a motorway. In terms of CIX, this cluster has the second highest average CIX (0.3780), such that it plays a relatively important role in the CVS chain market.

For Cluster2, although it only consists of 8 CVSs, it achieves the highest average CIX (0.4579). This indicates that this type of CVS outperforms other CVSs in the market. Geographically speaking, 7 out of 8 CVSs are located in the middle of Guangdong province, an economically developed area. This is one possible reason for their good performance. Unlike the other three clusters, the Tobacco category appears in the destination category. Hence, the consumers may include more smokers than in other clusters. Further investigations on smoking consumers may help this segment to develop valuable marketing strategies.

For Cluster3, only one store is located in a city centre, while over half of the CVSs are located in the east area of Guangdong. The average CIX (0.3514) ranks in the third place, with the performance not being as good as Cluster1 and Cluster2. It is interesting to observe that the Household goods category that normally appears in the routine category jumps to the seasonal category in this cluster.

For Cluster4, most of the CVSs are located on motorways, and the majority of CVSs is located in the middle and east of Guangdong province. This cluster performs the worst in the market, as it has the lowest average CIX (0.3364).

In summary, the destination category achieves a high level of consistency in this study. With the exception of Cluster2, the destination category consists of Drinks and Sweets categories. The routine category mainly contains ready-to-eat categories (e.g., Sweets, Crisps/snacks, Biscuits, Milk, and Hot meals). This indicates the clustering results are meaningful, since the derived CR accurately reflects the reality. Such categories consist of routine products in daily life. Moreover, the routine category highlights the distinguished convenience strength of CVS chains. In the seasonal category, the included categories are quite diverse and none of the market segments consist of identical categories. The convenience category mainly contains Newspaper/magazine and Ice cream, and once again reveals the reliability of the proposed clustering algorithm. As it is well-known that categories such as Newspaper/magazine and Ice cream are low-profit product, but are essential to consumers.

### 5. Discussion

This section discusses and analyses the obtained clustering results. First, the newly proposed WFKM clustering algorithm is compared against the original fuzzy K-means algorithm. Second, the performance of the CR model is evaluated via a comparison with the traditional RFM model by employing the new WFKM algorithm in both models. Third, the CR model segments the market from a new CIX perspective, revealing derived market segments that are different from the ones...
obtained by using traditional segmentation attributes (i.e., geographical areas and price regions). These differences are summarised in Section 5.2. Based on the clustering results, some market strategies and tactics, including pricing, promotions, shelf space allocation and category-specific strategies, are suggested in the remaining part of this section.

5.1. Model evaluation

5.1.1. Comparison between WFKM and unweighted clustering algorithm

Three measures, partition coefficient, fuzzy entropy, and XB coefficient are used throughout this paper to assess the clustering performance. In this section, the clustering results obtained from WFKM and unweighted FKM algorithms are reported in Table 7. In WFKM, each category is dynamically assigned a weight according to its category attributes, whereas the unweighted FKM treats all attributes as equally important. As aforementioned in Section 4.2.1, for the partition coefficient, the greater index value indicates better performance, while for the XB coefficient and fuzzy entropy, the smaller index value indicates better performance. Therefore, Table 7 shows that the newly proposed WFKM algorithm outperforms the unweighted FKM algorithm in all three measures.

5.1.2. The comparison between the CR model and the RFM model

The obtained results are shown in Table 9, two models derive the same cluster centroid in Cluster2 and Cluster3. More specifically, the partition coefficient, fuzzy entropy, and XB coefficients in Table 9 are calculated as follows:

The partition coefficient evaluates the degree of overlapping between two clusters, given a cluster number c, and a fuzzy partition matrix U, the partition coefficient is defined as:

\[ V_{PC}(U, c) = \frac{1}{n} \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^2 \]

where \( u_{ij} \) indicates the membership degree of CVS j belonging to Cluster i, and it can be calculated by the use of Eq. (10), and n is the number of CVSs. In this work, c = 4 and n = 75.

For fuzzy entropy, the partition coefficient is defined as:

\[ V_{FX}(U, c, V) = -\frac{1}{n} \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_j^2 \ln \mu_j \]

where \( \mu_j \) indicates the membership degree of CVS j belonging to Cluster i, and \( \mu_j \) can be calculated by the use of Eq. (17), and n is the number of CVSs.

The XB coefficient can be defined as follows:

\[ V_{XB}(U, V, c) = \frac{\sum_{i=1}^{c} \sum_{j=1}^{n} m_i [v_i - v_j]^2}{n \times \min_{i=1, k=1, j=1} \| v_i - v_k \|^2} \]

Apart from the same parameters as the previous two measures, V is the collection of cluster centroids, \( m \) is the smoothing parameter that controls the fuzziness of the clusters, \( \| v_i - v_j \|^2 \) is the HCSim() distance between cluster centroid (CVS i) and a common CVS j, i.e., \( \| v_i - v_j \|^2 = HCSim(q(i), s(j)) \); \( \| v_i - v_j \|^2 \) is the HCSim() distance between two cluster centroids - CVS i and CVS k, i.e., \( \| v_i - v_k \|^2 = HCSim(q(i), q(k)) \). The XB coefficient finds a balance point between the internal compaction and the external separation.

Table 9 indicates that the XB coefficient and the fuzzy entropy of the CR model are smaller than that of the RFM model, respectively. In contrast, the partition coefficient of CR is greater than the partition coefficient of RFM model. The above results reveal that, compared with the RFM model, the newly proposed CR model outperforms the RFM model as it achieves more accurate and more effective clustering results. In addition, the CR model considers three dimensions (i.e., importance to consumers, retailers and the marketplace). However, this revised RFM only considers the consumer dimension (i.e., average monthly sales frequency and the average monthly sales volume) and the retailer dimension (i.e., average monthly gross profit).
5.2. Segmentation comparison

Given the problem at hand, the market segmentation method which is currently used by PetroChina segments the market either by the CVSs’ geographic area (i.e., east, middle and west areas of Guangdong province) or their price region (i.e., city centre or motorway). For example, when considering the geographic area, if one market segment is grouped into east area of Guangdong, then naturally all CVSs in this cluster should be located in east Guangdong. Similarly, when considering price region, CVSs are only grouped into either the city centre cluster or the motorway cluster.

However, the clustering results obtained from the newly proposed CR model are quite different from the existing ones. For each cluster, the distributions of CVSs included in different regions are summarised in Table 6. Since the new model is based on CIXs, the CVSs in one cluster have similar CIX performance, but they can be distributed in different geographic areas and price regions. For example, not all CVSs in Cluster1 are located on a motorway, 13% of them are located in a city centre. Therefore, category strategy should be ideally designed according to individual CVS cluster performance, rather than using simple geographic/price attributes.

5.3. CR-based category strategy

As aforementioned in Section 4.2.2, each category is assigned a CR with respect to their CIXs in the corresponding cluster. Due to their different CRs, customised category strategies should be applied accordingly. For example, in general, the destination category is used to build store image and attract consumers. In this work, the Drinks category appears in the destination category in all clusters and the Sweets category appears in the destination category in Cluster1, Cluster3, and Cluster4. These two categories contribute most towards the store performance. It is natural to pay more attentions to these sub-categories and provide a wider variety of products for consumers to choose. Also, frequent and diverse promotional offers on these categories are necessary to gain more profits for CVSs. In terms of shelf space allocation, normally a fixed and sufficient place needs to be provided to each destination category. For Cluster2, the Tobacco category appears in the destination category, which is quite different from other market segments. To better satisfy consumers’ needs in Cluster2, other data mining techniques (e.g., rule association mining) can be employed to identify other popular products that are associated with Tobacco. This may require adjustment of shelf space allocation.

For the same category, CRs may play different roles in different market segments. To provide a more effective market strategy, it is important to analyse their CRs under different circumstances. For example, the Household goods category is defined as a routine category in Cluster1 and Cluster3, while it appears in the seasonal category in Cluster2 and the convenience category in Cluster4. Due to their different CRs, different category strategies for the Household goods category are provided in Fig. 6.

5.4. Category tactics

For the obtained four market segments, to grab more market opportunities, it is necessary to establish an effective tactic. The first step is to analyse the features of each cluster. Regarding the average CIX, Cluster1 > Cluster2 > Cluster3 > Cluster4. According to Table 6, CVSs in Cluster2 are mainly distributed in the central region of Guangdong province where the economic development is higher than the rest part of this province; and the majority of CVSs in Cluster3 is located on the motorway. By analysing products’ average price trends (as

Table 8
The derived CR of categories in different clusters.

<table>
<thead>
<tr>
<th>Ranking</th>
<th>CR</th>
<th>Cluster1 (average CIX: 0.3780)</th>
<th>Cluster2 (average CIX: 0.4579)</th>
<th>Cluster3 (average CIX: 0.3514)</th>
<th>Cluster4 (average CIX: 0.3364)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Destination category (10%)</td>
<td>Drinks 0.5124</td>
<td>Drinks 0.6687</td>
<td>Drinks 0.4852</td>
<td>Drinks 0.4419</td>
</tr>
<tr>
<td>2</td>
<td>Routine category (10%–70%)</td>
<td>Sweets 0.4553</td>
<td>Tobacco 0.5930</td>
<td>Sweets 0.4111</td>
<td>Sweets 0.3834</td>
</tr>
<tr>
<td>3</td>
<td>Household good (70%)</td>
<td>Crisps/snacks 0.4499</td>
<td>Sweets 0.5377</td>
<td>Crisps/snacks 0.4098</td>
<td>Crisps/snacks 0.3823</td>
</tr>
<tr>
<td>4</td>
<td>Stationery (70%)</td>
<td>Biscuit 0.4445</td>
<td>Tobacco 0.5306</td>
<td>Biscuit 0.3983</td>
<td>Biscuit 0.3718</td>
</tr>
<tr>
<td>5</td>
<td>Food (70%)</td>
<td>Hot meal 0.4194</td>
<td>Hot meal 0.5229</td>
<td>Hot meal 0.3981</td>
<td>Hot meal 0.3680</td>
</tr>
<tr>
<td>6</td>
<td>Milk (70%)</td>
<td>Milk 0.4152</td>
<td>Milk 0.5072</td>
<td>Milk 0.3839</td>
<td>Milk 0.3670</td>
</tr>
<tr>
<td>7</td>
<td>Tobacco (70%)</td>
<td>Tobacco 0.4068</td>
<td>Milk 0.5022</td>
<td>Milk 0.3816</td>
<td>Milk 0.3533</td>
</tr>
<tr>
<td>8</td>
<td>Engine oil (70%)</td>
<td>Engine oil 0.3920</td>
<td>Engine oil 0.4883</td>
<td>Stationery 0.3472</td>
<td>Engine oil 0.3483</td>
</tr>
<tr>
<td>9</td>
<td>Household good (70%)</td>
<td>Household good 0.3808</td>
<td>Health &amp; beauty 0.4777</td>
<td>Bread 0.3387</td>
<td>Health &amp; beauty 0.3203</td>
</tr>
<tr>
<td>10</td>
<td>Stationery (70%)</td>
<td>Stationery 0.3694</td>
<td>Bread 0.4725</td>
<td>Tobacco 0.3354</td>
<td>Bread 0.3196</td>
</tr>
<tr>
<td>11</td>
<td>Health &amp; beauty (70%)</td>
<td>Health &amp; beauty 0.3681</td>
<td>Engine oil 0.4699</td>
<td>Auto supplies 0.3315</td>
<td>Auto supplies 0.3158</td>
</tr>
<tr>
<td>12</td>
<td>Fruits (70%)</td>
<td>Fruits 0.3634</td>
<td>Fruits 0.4400</td>
<td>Tobacco 0.3276</td>
<td>Tobacco 0.3132</td>
</tr>
<tr>
<td>13</td>
<td>Fruits (70%)</td>
<td>Fruits 0.3460</td>
<td>Stationery 0.4300</td>
<td>Wine 0.3207</td>
<td>Fruits 0.3128</td>
</tr>
<tr>
<td>14</td>
<td>Auto supplies (70%)</td>
<td>Auto supplies 0.3317</td>
<td>Auto supplies 0.4277</td>
<td>Fruits 0.3170</td>
<td>Pharmacy 0.3100</td>
</tr>
<tr>
<td>15</td>
<td>Seasonal category (70%–85%)</td>
<td>Pharmacy 0.3297</td>
<td>Wine 0.3876</td>
<td>Household good 0.3163</td>
<td>Wine 0.3085</td>
</tr>
<tr>
<td>16</td>
<td>Convenience category (85%–100%)</td>
<td>Wine 0.3286</td>
<td>Family planning 0.4767</td>
<td>Pharmacy 0.3096</td>
<td>Family planning 0.3075</td>
</tr>
<tr>
<td>17</td>
<td>Sox gift (85%–100%)</td>
<td>Toy/gift 0.3161</td>
<td>Pharmacy 0.3464</td>
<td>Ice cream 0.3063</td>
<td>Toy/gift 0.3059</td>
</tr>
<tr>
<td>18</td>
<td>Convenience category (85%–100%)</td>
<td>Ice cream 0.3140</td>
<td>Ice cream 0.3438</td>
<td>Family planning 0.3059</td>
<td>Ice cream 0.3059</td>
</tr>
<tr>
<td>19</td>
<td>Newspaper/magazine (85%–100%)</td>
<td>Newspaper/magazine 0.3096</td>
<td>Newspaper/magazine 0.3361</td>
<td>Toy/gift 0.3059</td>
<td>Household good 0.3006</td>
</tr>
<tr>
<td>20</td>
<td>Ice cream (85%–100%)</td>
<td>Ice cream 0.3068</td>
<td>Toy/gift 0.3277</td>
<td>Newspaper/magazine 0.2984</td>
<td>Newspaper/magazine 0.2987</td>
</tr>
</tbody>
</table>

Table 9
CR and RFM model comparison.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>CR model</th>
<th>RFM model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cluster1</td>
<td>Cluster2</td>
</tr>
<tr>
<td>Cluster centroid</td>
<td>38</td>
<td>1</td>
</tr>
<tr>
<td>Average membership</td>
<td>0.2855</td>
<td>0.6212</td>
</tr>
<tr>
<td>The partition coefficient</td>
<td>0.3628</td>
<td>0.3217</td>
</tr>
<tr>
<td>The XB-coefficient</td>
<td>0.4029</td>
<td>0.3413</td>
</tr>
</tbody>
</table>
shown in Fig. 7), Cluster3 charges a relatively higher average price on products than other clusters, while Cluster4 charges the lowest average price. Meanwhile, compared with Cluster2 and Cluster3, the price trends of Cluster1 and Cluster4 tend to be relatively stable. Based on the above features of each cluster, an effective tactic is identified below:

• Though Cluster2 performs best amongst all market segments, when it comes to the number of CVSs, this cluster only contains 8 stores, which is much less than other clusters. Hence, one suggestion is to set up more CVSs in Cluster2 to attract more consumers and gain more profits. Furthermore, as illustrated in Fig. 8, compared with other clusters, Cluster2 achieves better sales volume, sales revenue, and sales gross compared with other clusters in the mid-price products. Therefore, it is suggested that Cluster2 should pay more attention to mid-price products, and provide more mid-price products to meet its consumers' needs.

• One possible reason for the higher average price in Cluster3 is that the majority of its CVSs (95.7%) is located on motorways. The current marketing strategy suggests that consumers of motorway CVSs are less sensitive to product price. The motorway CVSs often attract consumers who are travelling and they visit the store by chance. For such consumers, they are very mobile and have relatively less loyalty to stores. Therefore, the charged price is normally higher than those stores in a city centre. However, the results obtained in this work mildly contradict this assumption, since the average CIX of Cluster3 only ranks in the third place. This is because today's travellers are aware that motorway CVSs may charge higher prices than supermarkets, so they often prepare their drinks, sweets, and other essentials in advance. Meanwhile, Fig. 8 shows that Cluster3 gains the highest percentage of sales revenue (around 65%), but the lowest sales volume (around 11%) in high-price products. Since revenue = price × volume, this implies that the price of high-price products in Cluster3 is higher than that in other clusters. The current strategy for Cluster3 may need to be revised, possibly by reducing the price of high-price products.

• The price trends of Cluster1 and Cluster4 tend to be relatively stable. This indicates that the proposed prices are less sensitive to the time series. This is especially true for Cluster4, as its current performance is the worst in the market. Furthermore, it is important to take the price sensitivity into account when adjusting prices. For example, since Cluster2 performs best in terms of CIX, for those CVSs distributed in the same geographic area and price region, the current pricing strategy used by the CVSs in Cluster2 can be jointly considered to revise the price of such CVSs in Cluster4. Moreover, with an attempt to obtain a better price strategy, a more systematical approach can be developed to analyse the price sensitivity in different time periods. Additionally, when it comes to low-price products, as shown in Fig. 8, the overall performance of Cluster3 and Cluster4 are better than Cluster1 and Cluster2. Also, the overall performance of Cluster1 in the high-price products is better than that of the other clusters (it achieves the highest sales volume and gross profits, and the second highest sales revenue). This reveals that low-price products gain more popularity from consumers in Cluster3 and Cluster4, while high-price products are more popular in consumers in Cluster1. One suggestion is that high-price products in Cluster1 and Cluster4 can be replaced...
by low-price products. In contrast, Cluster 1 is suggested to provide more high-price products rather than low-price products.

6. Conclusions

CM techniques form the basis of mature retail management systems, and they help to integrate supply chain and strategic marketing to develop fast consumer reaction systems. Successful CM cases often appear in the cooperation between large manufacturers and supermarkets (e.g., P&G). However, a CVS has the characteristics of being small and scattered, with a limited number of categories and products. Moreover, different types of markets have various kinds of consumers. CVS consumers have diverse needs and they are also mobile. This makes it very difficult to employ the mainstream consumer-centric CM methods to analyse and group consumers of CVS chain. To overcome these difficulties, it is essential to cluster the CVSs, and apply customised CM strategies to each CVS cluster.

Starting from the three dimensions of CR, this work combines the characteristics of the retail market with traditional behaviour-based market segmentation techniques to propose a new CR-based market segmentation model. In particular, an innovative similarity measure (HCsim) and an improved WFKM clustering algorithm are proposed to group CVSs with similar CIX into the same market segment. The applicability and utility of the proposed clustering model are demonstrated via an empirical study on a CVS chain dataset provided by PetroChina. By using the CR model, the current retail market is divided into four clusters, and they are ranked as $\text{Cluster}_2 > \text{Cluster}_1 > \text{Cluster}_4 > \text{Cluster}_3$ with respect to their CIXs. For comparison purposes, the new model is compared with the traditional RFM model, with the comparative results revealing that the CR model achieves more stable and more effective clustering results. Another distinguished feature of the new model is that the CR helps to define appropriate marketing strategies for different categories in different market segments.

Although the proposed approach is promising, much progress can be made through further research. First, when identifying the CR, the importance of certain categories to competitors also needs to be considered, so that the derived model would better reflect the reality. Second, to better simulate the changes in consumers demand and purchasing behaviour, time series analysis will be employed to analyse historical sales data. Moreover, to further validate the utility and efficiency of the CR model, customers’ responses to the newly derived marketing strategies will be collected. This action will take some time, but once such data becomes available, it can be used to build predictive models [7,39].

Acknowledgement

This work is supported by the National Nature Science Foundation of China (Grant Nos. 70971112, 71301133, 71371159) and Humanity and Social Science Youth Foundation of Ministry of Education, China (Grant No. 13YJC630033). Thanks also go to PetroChina for the donation of the dataset in use.

References
