Islamic Azad University , Shiraz Branch
نشریه تحلیل مدارها، داده ها و سامانه ها
Journal of Circuits, Data and Systems Analysis
sanad.iau.ir/journal/jcdsa

Data Analysis of Marketing Companies using Improved K-Means Clustering and LRFMM$_2$ Model

Atieh Mirzaei[1] , Zahra Rezaei [2*]

[1] Department of Computer Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran
tyhmirzaii@yahoo.com
[2] Department of Computer Engineering, Marvdasht Branch, Islamic Azad University, Marvdasht, Iran
zrezaei@iau.ac.ir

Abstract: Clustering, especially *k-means*, is one of the most important data mining techniques for identifying and monitoring customer behavior. In classical k-means, the optimality depends on the initial selection of the centers; therefore, it is not optimal. Another problem is determining the number of clusters and making the clusters zero. Customers' behavioral characteristics are considered in customer clustering, and a method is used to find the optimal number of clusters and the initial values of cluster centers to achieve more accurate results and predict customer lifetime. The results of this research show the customer behavior of each cluster. The proposed improved k-means method has been compared with the classical K-means once including the M_2 parameter (customer satisfaction level) and once without M_2. The Normalized Mutual Information (NMI) criterion was calculated on the proposed method and the compared method, and in both cases, considering or missing the M_2 index, the NMI of the proposed method was higher. Also, the adjusted rand index with the M_2 parameter has recorded the highest number. In terms of time, it is faster than classical K-means, which shows that the proposed method has the best performance in terms of speed and performance accuracy compared to classical K-means.

Keywords: Clustering, K-means, Optimization, Marketing, Analysis

JCDSA, Vol. 2, No. 6, Summer 2024	Online ISSN: 2981-1295	Journal Homepage :https://sanad.iau.ir/en/Journal/jcdsa
Received: 2024-06-01	Accepted: 2024-08-23	Published: 2024-09-15

CITATION

Mirzaei, A., and Rezaei, Z., *"Data Analysis of Marketing Companies using Improved K-Means Clustering and LRFMM2 Model"*, Journal of Circuits, Data and Systems Analysis (JCDSA), Vol. 2, No. 6, pp. 19-29, 2024.
DOI: 00.00000/0000

* Corresponding author

Extended Abstract

1- Introduction

Today's new economy primarily focuses on providing better services, and the present era is called the customer-oriented economy [1]. Sales transaction data generated by retail and e-commerce businesses is rapidly increasing today [2]. Network marketing companies sell their products online by recruiting network marketers. It is believed that keeping a customer is more important than finding a new customer, and this can be considered a job and profession that helps generate employment. The existence of great competitors in the field of business has led to escalating competition between organizations. Regardless of the size of the company, excellent customer service is critical to gaining new customers and retaining existing customers [3]. Effective customer knowledge management leads to effective customer relationship management (CRM). The interaction between organizations and customers has changed significantly, so there is no long-term guarantee of business continuity with customers. However, organizations must adequately identify their customers and anticipate their needs and expectations [4]. With a deep understanding of customer clustering, companies can create specific and targeted marketing strategies for each customer group [5]. Customer segmentation is the process of dividing the entire market into smaller customer groups, which makes it possible to understand the different needs and behaviors of customers and adapt the appropriate marketing approach or product recommendations to them [6]. This research is a step toward improving the marketing business in Iran, which will be used through the detection of violations from the data of marketing companies and for investigation by the institutions that supervise these businesses.

2- Methodology

The data collected from the database of network marketing companies was randomly extracted and included customer IDs and purchase amounts for each month from April 1996 to December 1996. The customer ID, the number of purchases made, the price paid, the date of purchase, and customer satisfaction have been calculated and extracted from the existing dataset. The data of each marketing company has a different structure, and to provide aggregated reports, it was necessary to aggregate all the data in a data warehouse with BI tools and to unify the data in terms of structure before doing anything. In the data mining phase, the transaction data is fragmented, including transaction date, purchase frequency (number of months the purchase was made), monetary value of the transaction, and customer satisfaction. To extract the L, R, F, M, and M_2 parameters of each customer in this research, the distance between the first and last purchase of the customer was calculated as the L feature. The algorithm's first step in the proposed method is determining each cluster's initial element. This method first starts with a cluster containing an element, and in other steps, it calculates the distance of other elements from the center of the specified clusters. However, the data closest to the data mean is determined as the only member of the first cluster. After assigning each element to a cluster, the average data of that cluster is updated based on the Customer life cycle value (CLV).

3- Results and discussion

The *Custno* variable should be a number between 2000 and 5000. In this research, 2450 to 4000 classes were considered to increase the number of courses due to the small number of samples. Data that are placed in the same class have the same CLV value. Then, the total average of these CLVs is calculated according to *Custno*. The center point in K-means clustering is the point that represents a particular cluster and is the average of all data points in the cluster at each step (up to convergence) change. In determining the initial data of each cluster, suitable initial centers can speed up the convergence of K-means. The final results of the k-means algorithm depend on the initialization of the center. Two different clusters have been done on the data. The first clustering was done by considering the M_2 parameter (marketers' satisfaction level), and the second was done without considering M_2. In the first clustering, customers in all clusters except the first to fifth clusters in this pyramid are in a favorable situation regarding L, R, F, M, and M_2 characteristics. Therefore, to retain these customers, the company is suggested to convert their loyalty behavior into a loyalty display through communicating and interacting with them. Based on this, customers in clusters 7 to 10 have the best conditions in terms of characteristics L, R, F, M, M_2. In the second clustering, the data is divided into 3 clusters.

4- Conclusion

The main goal of the research is to use the k-means algorithm and find the length of customer life using parameters such as the distance between the first and last purchase of the marketer as L feature, the distance between the customer's last purchase and the end of the period as R feature, the number of months customer purchases in a specific time period as feature F, the total monetary value between the first and last purchase of a customer in a specific time period as feature M and the level of satisfaction of marketers as feature M_2 and predicting the probability of buying in the coming months and To find points with a big jump; the data is first sorted by CLV in ascending order. Then, the Euclidean distance value of each data is obtained from its previous data in ascending order of CLV. The distance between some data and the center of the first cluster and the obtained optimal center values are calculated. After that, these steps are repeated to obtain all the cluster centers. The optimal number of clusters has been calculated using the equations as 10. The lower the level of marketers' satisfaction, the CLV variable, which represents the customer's lifetime, also decreases. The results show that the proposed method correctly identified the number of clusters, but it was done in a shorter time.

Islamic Azad University , Shiraz Branch
نشریه تحلیل مدارها، داده ها و سامانه ها
Journal of Circuits, Data and Systems Analysis
sanad.iau.ir/journal/jcdsa

تحلیل داده‌های شرکت‌های بازاریابی با استفاده از خوشه‌بندی K-Means بهبودیافته و مدل $LRFMM_2$

عطیه میرزایی[1]، زهرا رضایی[2*]

۱- گروه مهندسی کامپیوتر، واحد تهران جنوب، دانشگاه آزاد اسلامی، تهران، ایران (tyhmirzaii@yahoo.com)
۲- گروه مهندسی کامپیوتر، واحد مرودشت، دانشگاه آزاد اسلامی، مرودشت، ایران (zrezaei@iau.ac.ir)

چکیده: خوشه‌بندی به ویژه k میانگین، یکی از مهمترین روش‌های داده‌کاوی است که در شناسایی و رصد رفتار مشتریان مورد استفاده قرار می‌گیرد. در k میانگین کلاسیک، بهینگی وابسته به انتخاب اولیه مراکز بوده و در نتیجه، این روش، بهینه نیست. مشکلات دیگر آن تعیین تعداد خوشه‌ها و صفر شدن خوشه‌ها است. در خوشه‌بندی مشتری، ویژگی‌های رفتاری مشتری به عنوان روشی برای یافتن تعداد بهینه خوشه‌ها و مقادیر اولیه مراکز خوشه برای دستیابی به نتایج دقیق‌تر و پیش‌بینی طول عمر مشتریان در نظر گرفته شده است. نتایج این پژوهش رفتار مشتریان هر یک از خوشه‌ها را نشان می‌دهد و روش پیشنهادی بهبود یافته k میانگین یک بار با احتساب پارامتر M_2 (میزان رضایت مشتریان) و یک بار بدون M_2 با k میانگین کلاسیک مقایسه شده است. معیار اطلاعات متقابل نرمالیزه بر روی روش پیشنهادی و روش مورد مقایسه محاسبه شده و در هر دو حالت، با در نظر گرفتن یا فقدان شاخص M_2، معیار اطلاعات متقابل نرمالیزه روش پیشنهادی بالاتر بوده است. همچنین شاخص رند تعدیل یافته برای حالت روش پیشنهادی با احساب پارامتر M_2 بالاترین عدد را ثبت کرده است. روش پیشنهادی از نظر زمان محاسباتی نیز سریعتر از k میانگین کلاسیک بوده که نشان می‌دهد این روش عملکرد بهتری را از نظر سرعت و دقت عملکرد نسبت به k میانگین کلاسیک دارد.

واژه های کلیدی: خوشه‌بندی، k میانگین ، بهینه‌سازی، بازاریابی، تحلیل

نوع مقاله: پژوهشی
DOI: 00.00000/0000
تاریخ ارسال مقاله: ۱۴۰۳/۰۳/۱۲ تاریخ پذیرش مقاله: ۱۴۰۳/۰۶/۰۲ تاریخ چاپ مقاله: ۱۴۰۳/۰۶/۲۵

۱- مقدمه

اقتصاد جدید امروز به طور گسترده بر ارائه خدمات بهتر متمرکز شده و عصر حاضر را اقتصاد مشتری مداری می‌نامند [۱]. داده‌های تراکنش‌های فروش که توسط کسب‌وکارهای خرده‌فروشی و تجارت الکترونیک ایجاد می‌شوند؛ در دنیای امروزی به سرعت در حال افزایش است [۲]. شرکت‌های بازاریابی شبکه‌ای از طریق جذب بازاریاب به صورت شبکه‌ای اقدام به فروش محصولات خود به صورت آنلاین می‌نمایند. اعتقاد بر این است که حفظ مشتری مهمتر از یافتن مشتری جدید است و بیانگر آن است که این می‌تواند به عنوان یک شغل و حرفه محسوب شده و به اشتغال‌زایی کمک نماید. وجود رقبای فوق العاده در حوزه کسب و کار منجر به رقابت‌های تصاعدی بین سازمان‌ها شده است. صرف نظر از اندازه شرکت، خدمات عالی به مشتریان برای به‌دست آوردن مشتریان جدید و حفظ مشتریان موجود، از اهمیت زیادی برخوردار است [۳]. مدیریت موثر دانش مشتری، منجر به مدیریت کارآمد ارتباط با مشتری[1] می شود. تعامل بین سازمان‌ها و مشتریان به طور قابل توجهی تغییر کرده است؛ به طوری که هیچ تضمین بلندمدتی برای تداوم تجارت با مشتریان وجود ندارد. با این حال، سازمان‌ها باید مشتریان خود را به درستی شناسایی کنند و نیازها و انتظارات آن‌ها را پیش‌بینی کنند [۴]. با درک عمیق از خوشه‌بندی مشتریان، شرکت‌ها می‌توانند برای هر گروه مشتری، استراتژی های بازاریابی خاص و هدفمند را ایجاد کنند [۵]. تقسیم‌بندی مشتری، فرآیند تقسیم کل بازار به گروه های مشتریان کوچکتر است که باعث می‌شود تا نیازها و رفتارهای مختلف مشتریان، درک شده و رویکرد بازاریابی مناسب یا توصیه‌های محصول را با آنها تطبیق داد [۶].

این تحقیق گامی است در جهت سالم‌سازی کسب و کار بازاریابی در ایران که از طریق شناسایی موارد تخلف از داده‌های شرکت‌های بازاریابی، کشف و جهت بررسی توسط نهادهای ناظر بر این کسب و کارها، مورد استفاده قرار خواهد گرفت. از جمله تخلفات موجود در این حوزه که برای کسب بالاسری بیشتر از فروش زیرمجموعه‌ها رخ می‌دهد، مجبور کردن بازاریابان زیرمجموعه به خرید محصولاتی است که قادر به فروش آن‌ها به مشتریان نیستند. طبق قوانین کشور, اجبار در خرید بایستی منجر به غیرفعال شدن کد بازاریابی فرد متخلف گردد. به همین جهت با استفاده از مدل $LRFMM_2$ و محاسبه طول عمر مشتری[2] به کشف موارد تخلف

[1] Customer Relationship Management (CRM)

[2] Customer Life Cycle (CLV)

پرداخته خواهد شد. بدین ترتیب که معمولاً در اجبار در خرید، بازاریابان در ماه‌های ابتدای عضویت‌شان خریدهایی با مبالغ نه چندان کم انجام می‌دهند و در ماه‌های بعد از آن به علت عدم توانایی در فروش محصولات به مشتری، خریدی انجام نمی‌دهند. یافتن این افراد از طریق خوشه‌بندی صحیح انجام می‌شود. روش‌های خوشه‌بندی می‌تواند داده‌های ارزیابی‌های مشتریان را در خوشه‌های مختلف تقسیم‌بندی کند. سپس، مدیران، مدیریت ارتباط با مشتری خود را توسعه خواهند داد و با توجه به ترجیحات[1] مشتریان، آن‌ها را خوشه‌بندی کنند. در این مقاله، خوشه‌بندی بهینه بازاریابان بر اساس تعداد ماه‌هایی که خرید کرده‌اند، فاصله زمانی از آخرین خریدشان و میزان مبلغ خریدی که انجام داده‌اند، صورت می‌پذیرد. به کمک روش دیوس بولدین تعداد خوشه‌های بهینه تعیین می‌شوند. افزایش سرعت و دقت اجرای الگوریتم k میانگین[2] و مقایسه‌ی آن با k میانگین کلاسیک از اهداف این تحقیق می‌باشد. در مرحله بعد اعضا، بر اساس پارامترهای $LRFMM_2$ خوشه‌بندی می‌شوند. در ادامه ارزش طول عمر مشتری برای هر یک از خوشه‌های موجود محاسبه و خوشه‌ها بر اساس ارزش طول عمر مشتری آنها رتبه‌بندی می‌شوند. درنهایت با استفاده از الگوهای پیشین و مدل ارزش طول عمر مشتری در هر خوشه و انجام سعی و خطا، الگوهای سفارشی به دست می‌آید. نوآوری دیگر این پژوهش استفاده از پارامتر M_2 که نارضایتی مشتریان است، می‌باشد که برای شناسایی بازاریابانی که توسط بالاسری های خود اجبار در خرید داشته‌اند، استفاده شده است. در این تحقیق روش پیشنهادی از طریق معیار اطلاعات متقابل نرمالیزه، معیار ارزیابی اگوریتم جهت تشخیص دقت و صحت و همچنین شاخص رند تعدیل یافته با احتساب M_2 و بدون M_2 با k میانگین کلاسیک مقایسه شده است.

در ادامه این مقاله، بررسی جامعی از روش‌های بازاریابی انجام می‌شود. در بخش سوم، مدل پیشنهادی تحلیل داده‌های شرکت‌های بازاریابی بیان می‌شود. بخش چهارم نتایج شبیه‌سازی و تجزیه و تحلیل نتایج را نشان می‌دهد. در نهایت در بخش پنجم، نتیجه‌گیری مطرح می‌شود.

۲- مرور ادبیات

به طور سنتی مشتری بخش‌بندی با استفاده از روش‌های مبتنی بر آمار که مجموعه‌ای از آمار را از داده‌های مشتری محاسبه می‌کند و مشتریان را گروه می‌کند، به‌دست می‌آید. اعمال الگوریتم‌های خوشه‌بندی مبتنی بر فاصله در فضای این آمار، به بخش‌ها تقسیم می‌شوند. تیانی جیانگ و الکساندر توژیلین [۷] رویکرد مبتنی بر گروه‌بندی را پیشنهاد دادند که برای محاسبه بخش‌های مشتری، بر اساس ترکیب بهینه داده‌های تراکنشی چندین مشتری برای ایجاد یک مدل داده‌کاوی به‌کار گرفته شد. آنها دریافتند که یافتن یک بخش مشتری، بهینه NP-hard[3] است؛ بنابراین چندین گروه‌بندی مستقیم زیر بهینه را پیشنهاد کردند. تجزیه و تحلیل خوشه‌بندی یک کار توصیفی است که به دنبال شناسایی گروه‌های همگن از اشیا، بر اساس مقادیر ویژگی‌های آنهاست. سوراب شاه و مانموهان سینگ [۸] الگوریتم خوشه‌بندی جدیدی ارائه‌دادند که شبیه الگوریتم k میانگین و K-medoids است و چندین روش را برای انتخاب خوشه اولیه آزمایش کردند. الگوریتم k میانگین اصلاح شده از نظر تعداد خوشه‌ها و مقایسه زمان اجرا در مقایسه با الگوریتم‌های k میانگین و K-Mediod نتایج بهتری به‌دست آورد. سانگ چو و سونگ چول [۹] یک الگوی وزنی جدید استخراج فرکانس مبتنی بر امتیاز RFM[4] مشتری برای سیستم توصیه‌های تجاری u-commerce شخصی شده را پیشنهاد کردند. سیستم‌های سنتی دارای مشکلاتی مانند تاخیر در سرعت پردازش داده، در نظر گرفتن وزن برابر برای هر آیتم هستند. در این مقاله، برای حل این مشکلات، داده‌های متداول خریداری شده از کل داده‌ها استخراج شده، اهمیت ویژگی اقلام در نظر گرفته شده تا با تأکید بر موارد مهم با قابلیت خرید بالا، روندهای متوالی در حال تغییر، پیش‌بینی شود. برای تأیید عملکرد روش بهبود یافته، آزمایش‌هایی با مجموعه داده‌های جمع‌آوری شده در یک مرکز خرید اینترنتی لوازم آرایشی انجام شده است.

ایکس او و سی لی [۱۰] یک مدل تقسیم‌بندی سه بعدی مشتری بر اساس ارزش طول عمر مشتری، رضایت مشتری و فعالیت مشتری پیشنهاد دادند که مشتریان را با دقت بیشتری به گروه‌های مختلف تقسیم می‌کند. متغیرهای مربوطه توسط مدل RFM، مدل کانو و مدل BG/NBD به دست آمده است. مدل تقسیم‌بندی مشتری، ده گروه از مشتریان را با استراتژی‌های بازاریابی مربوطه ارائه می‌کند؛ به طوریکه می‌تواند به شرکت‌ها در به حداکثر رساندن سود کمک کند. فروشگاه‌های اینترنتی با به‌کارگیری مفهوم مدیریت ارتباط با مشتری می‌تواند مشتریان را شناسایی کند، بنابراین مشتریان می‌توانند با توجه به نیاز خود، از تسهیلات ویژه در استراتژی‌های بازاریابی مناسب برخوردار شوند. لیسنا زهروتون [۱۱] با استفاده از مفهوم مدیریت ارتباط با مشتری برای خرید آنلاین، مشتریان بالقوه را با تقسیم‌بندی مشتریان شناسایی کردند. بنابراین برای تقسیم‌بندی مشتری و بازاریابی دقیق، از روش خوشه‌بندی فازی استفاده کردند که به افزایش سود شرکت کمک می‌کند. علیزاده و همکاران [۱] برای تحلیل ویژگی‌های رفتار خرید مشتریان، به منظور بهبود عملکرد سیستم مدیریت ارتباط با مشتری، یک روش سیستماتیک را پیشنهاد دادند. برای این منظور، از مدل بهبود یافته LRFM (شامل شاخص‌های طول، تازگی، فرکانس و پولی) استفاده شد که در حال حاضر مدل رایج‌تری نسبت به مدل پایه RFM برای تحلیل ارزش طول عمر مشتری است. از آنجایی که مدل RFM وفاداری مشتریان را در نظر نمی‌گیرد؛ از مدل LRFM برای ایجاد اصلاحات استفاده شده است. برخلاف بسیاری از مطالعات گذشته که در آن از روش‌های خوشه‌بندی آماری در کنار مدل RFM یا LRFM استفاده شده است؛ مطالعه حاضر امکان تحلیل خوشه‌بندی را با وارد کردن شاخص‌های LRFM در چارچوب یک سیستم استنتاج فازی فراهم کرده است. نتایج به‌دست‌آمده

[1] preferences
[2] k-Means algorithm
[3] nondeterministic polynomial time problem
[4] Recency, Frequency, Monetary Value

برای یک شرکت عمده‌فروشی بر اساس رویکرد پیشنهادی نشان داد که بین خوشه‌ها از نظر چهار شاخص LRFM تفاوت معناداری وجود دارد. بنابراین می‌توان از این رویکرد برای خوشه‌بندی مشتریان و بررسی ویژگی‌های آنها استفاده کرد. نقطه قوت این رویکرد در مقایسه با رویکردهای قدیمی، انعطاف پذیری بالای آن است، زیرا در آن نیازی به خوشه‌بندی مجدد مشتریان و تدوین مجدد استراتژی ها در صورت افزایش یا کاهش تعداد مشتریان نیست.

خوشه‌بندی چندهسته‌ای[1] به دلیل عملکرد رقابتی در یادگیری بدون نظارت، نظرات قابل توجهی را به خود جلب می‌کند. با این حال، مشاهده می‌شود که بسیاری از رویکردهای موجود خوشه‌بندی چندهسته‌ای ارتباط بین بخش‌های خوشه‌بندی مختلف را به اندازه کافی در نظر نمی‌گیرند. در نتیجه، روش‌های موجود باعث ایجاد تنوع زائد و کم بخش‌های خوشه‌بندی انتخاب شده می‌شوند که عملکرد خوشه‌بندی را بدتر می‌کنند. برای پرداختن به این موضوعات، یک روش خوشه‌بندی چند هسته‌ای k میانگین موثر و کارآمد توسط جی هو و همکاران [۱۲] ارائه شد. در این روش، تراز حداکثری بخش توافقی با بخش‌های پایه وزنی پیشنهاد می شود. الگوریتم مورد نظر، به طور مشترک بخش‌های خوشه‌بندی پایه و نتیجه خوشه‌بندی توافقی بهینه را بهینه می‌کند. در [۴]، ویژگی‌های رفتاری مشتریان (ویژگی مخرب) را در نظر گرفتند که برای خوشه‌بندی مشتریان، تعیین بهینه تعداد خوشه‌ها و مقادیر اولیه مراکز خوشه برای به‌دست آوردن نتایج دقیق‌تر بود. در نهایت، با توجه به نیاز سازمان‌ها به استخراج دانش از دیدگاه مشتریان از طریق رتبه-بندی مشتریان بر اساس عوامل موثر ارزش مشتری، روشی برای مدل-سازی رفتار آن‌ها و استخراج دانش برای ارتباط با مشتری پیشنهاد شده است. نتایج ارزیابی مشتریان شرکت همکاران سیستم نشان می‌دهد که روش بهبود یافته k میانگین ارائه شده در این مقاله از نظر سرعت و دقت نسبت به الگوریتم k میانگین برتری دارد. در مورد بیمه عمر، ضریب نفوذ یکی از اهداف اصلی هر صنعت بیمه توسعه یافته است. از این نظر بازاریابی سیستماتیک جزء مهمی در برنامه استراتژیک شرکت‌های بیمه است. برای دستیابی به هدف، بیمه‌گران باید مشتریان خود را در گروه‌های مختلف گروه‌بندی کنند که در آن برخی ویژگی‌های مشترک است و افراد الگوی مشابهی را نشان می‌دهند. در [13]، از خوشه‌بندی k میانگین به عنوان یک الگوریتم یادگیری بدون نظارت به منظور تقسیم مشتریان به تعدادی خوشه استفاده می‌کند. خوشه‌ها بر اساس دو متغیر مستقل ساخته شده‌اند: حق بیمه ماشین و عمر. سپس آمار توصیفی سایر ویژگی‌های تعیین‌کننده ارائه می‌شود که با آن بیشترین تمایل در خرید بیمه عمر ارائه می‌شود.

با ارائه داده‌های مرتبط و به موقع به واحدهای تجاری در صنعت خرده‌فروشی، استفاده از هوش تجاری در شناسایی مشتریان بالقوه را در [۱۴] بررسی کردند. به منظور اجرا و اعمال رویکرد علمی با استفاده از الگوریتم k میانگین، داده‌های معاملاتی و خرده‌فروشی به‌صورت آنی، تجزیه و تحلیل شدند. در این مطالعه، بر اساس RFM مجموعه داده توسط k میانگین خوشه‌بندی شد. آلیچا راچوال و همکاران [۶] روش جدیدی را برای مقابله با مجموعه داده‌های مختلط برای تقسیم‌بندی مشتری پیشنهاد کردند. متغیرهای طبقه‌بندی شده با استفاده از یک مدل ابتکاری بدون نظارت بر اساس رمزگذار خودکار تعبیه شدند. سپس، مشتریان با استفاده از الگوریتم‌های مختلف خوشه‌بندی بر اساس ماتریس‌های شباهت به گروه‌ها تقسیم شدند. علاوه بر روش کلاسیک k میانگین و DBSCAN جدیدتر، سه الگوریتم دیگر شامل: الگوریتم لووین، الگوریتم حریص و الگوریتم انتشار برچسب استفاده شد. این تحقیق بر روی دو مجموعه داده انجام شد که یکی شامل مشتریان خرده‌فروشی و دیگری شامل مشتریان عمده فروشی بود.

کارایی الگوریتم خوشه‌بندی k میانگین به عنوان روشی برای تقسیم‌بندی کارآمد مصرف‌کننده توسط مالایی و همکاران [۳] بررسی شد . الگوریتم k میانگین ادغام شده با تجزیه و تحلیل RFM در سطح جهانی به عنوان یک روش خوشه‌بندی بخش‌بندی عمیق معتبر است که ثابت کرده در تنظیمات مختلف تجاری بسیار کارآمد است. نتایج تجربی شواهد متقاعدکننده‌ای از عملکرد الگوریتم از نظر تقسیم‌بندی مصرف‌کننده را ارائه کردند. مقدار خلوص بالای به‌دست آمده (۰/۹۵) کارایی الگوریتم خوشه بندی k میانگین را از نظر تقسیم بندی و طبقه‌بندی دقیق مشتریان نشان می‌دهد و استراتژی‌های بازاریابی هدفمند و رویکردهای شخصی‌سازی شده را تسهیل می‌کند. محمد فرهان و جری هیکل [۵] استفاده از خوشه‌بندی k میانگین را مورد بحث قرار دادند. این فرآیند شامل گروه بندی داده‌های مشتری بر اساس عوامل مختلفی مانند نوع انتقال، رضایت مشتری، روش پرداخت و شعبه بانک است. خوشه‌های به دست آمده مورد تجزیه و تحلیل قرار می‌گیرند تا مشخصات و نیازهای مختلف مشتریان را شناسایی کنند. هدف از تحقیق بانکیت ایندارماوان و همکاران [۲]، توسعه مدل تقسیم‌بندی مشتری با استفاده از k میانگین بهینه بود. الگوریتم خوشه‌بندی برای گروه‌بندی دقیق‌تر مشتریان بر اساس داده‌های معاملات فروش، ارائه شد. الگوریتم گروه‌بندی، مشتریان را به سه خوشه بر اساس خرید تقسیم کرد. نتایج، اثربخشی الگوریتم k میانگین بهینه را نشان داد.

۳- روش پیشنهادی

در این تحقیق، روش پیشنهادی در مراحل زیر اجرا می‌شود:

۳-۱- آماده‌سازی و پیش‌پردازش داده‌ها

استخراج داده‌ها، مقیاس گذاری پارامترهای L، R، F، M، M_2، داده‌های مشتری و تشخیص فاصله و پاکسازی از مراحل آماده‌سازی و پردازش داده‌ها است. اطلاعات داده‌های جمع‌آوری شده از پایگاه داده شرکت‌های بازاریابی شبکه‌ای به‌صورت

[1] Multiple kernel clustering (MKC)

تصادفی استخراج شده و شامل شناسه مشتری و مبلغ خرید در هر ماه از فروردین ۱۳۹۶ تا دی ماه ۱۳۹۶ می‌باشد. از پایگاه داده موجود شناسه مشتری، دفعات خریدهای انجام شده، قیمت پرداخت شده، تاریخ خرید و رضایت مشتری محاسبه و استخراج شده است. داده‌های شرکت‌های بازاریابی دارای ساختاری متفاوت می‌باشند که جهت ارائه گزارشات تجمیعی لازم بود که کلیه‌ی داده‌ها در یک انباره داده با ابزارهای BI تجمیع شده و پیش از انجام هر کاری، یکسان‌سازی داده‌ها از لحاظ ساختار صورت پذیرد.

این اختلاف ساختار شامل موارد ذیل بوده است:

- برخی تاریخ‌ها میلادی و برخی شمسی ثبت شده‌اند.
- برخی مقادیر به ریال و برخی به تومان درج شده‌اند.
- با توجه به اینکه در شرکت‌های بازاریابی بالاسری‌ها به صورت ماهانه واریز می‌گردد؛ لذا تعداد ماه‌های خرید ملاک است. برای به‌دست آوردن تعداد ماه‌هایی که بازاریاب خرید انجام داده، تاریخ‌های خریدی که به صورت میلادی ذخیره شده، بایستی ابتدا به شمسی تبدیل و سپس فقط ماه خرید از تاریخ فوق‌الذکر استخراج شود.

۳-۲- استخراج داده

در مرحله استخراج داده‌ها، داده‌های معامله به‌طور جزئی وجود دارد که شامل تاریخ معامله، تعداد دفعات خرید (تعداد ماه‌هایی که خرید انجام شده)، ارزش پولی معامله و رضایت مشتری است. برای استخراج پارامترهای L، R، F، M و M_2 هر مشتری در این تحقیق، فاصله بین اولین و آخرین خرید مشتری به عنوان ویژگی L محاسبه شد. فاصله بین آخرین خرید مشتری و پایان دوره زمانی به عنوان ویژگی R محاسبه می‌شود. تعداد دفعات خریدها بر حسب ماه، بین اولین و آخرین خرید مشتری به عنوان ویژگی F محاسبه می‌شود. ارزش کل پولی بین اولین و آخرین خرید مشتری در یک بازه زمانی خاص به عنوان ویژگی M محاسبه می‌شود. در این تحقیق، سوءاستفاده از مشتری (ویژگی M_2) نیز به عنوان یکی از پارامترهای تأثیرگذار در مدیریت ارتباط با مشتری در نظر گرفته شده که از رضایت مشتری در هر رکورد از مجموعه داده‌ها محاسبه می‌شود و میزان حداقل و حداکثر سواستفاده از مشتری است.

شکل (۱): مراحل مدل‌سازی

۳-۳- مدل‌سازی

روشهای مختلفی برای اجرای پروژه‌های داده‌کاوی ارائه شده‌اند؛ اما یکی از قدرتمندترین روشها، روش CRISP[1] است این روش به عنوان یک روش فرآیندی، چرخه حیات یک پروژه داده‌کاوی را دربرمی‌گیرد. روش CRISP همچنین یک روش چرخشی است که شش مرحله عملیات داده‌کاوی را به طور کامل هدایت می‌کند. این شش مرحله شامل شناخت کسب وکار، شناخت داده، پیش‌پردازش، مدلسازی، ارزیابی مدل و توسعه مدل است. مراحل مدل‌سازی در شکل (۱) نشان داده شده است.

۳-۴-خوشه‌بندی

با استفاده از مدل خوشه‌بندی k میانگین، شرکت‌ها می‌توانند بخش‌های مختلف بازار را شناسایی کنند [۵]. برای پیش‌بینی دقیق رفتار مشتری، الگوریتم‌های خوشه‌بندی، به‌ویژه k میانگین یکی از مهم‌ترین روش‌های داده‌کاوی مورد استفاده در بازاریابی مدیریت ارتباط با مشتری است که با آن می‌توان رفتار مشتریان را شناسایی کرد. با این حال، در مطالعات مختلف در مورد خوشه‌بندی k میانگین مشاهده شده که مشتریان با شاخص‌های رفتاری متفاوت در خوشه‌بندی ممکن است یکسان به نظر برسند؛ به این معنی‌که شاخص‌های رفتاری مشتری نقش مهمی در خوشه‌بندی مشتریان ندارند. بنابراین اگر میزان مشارکت مشتری به پارامترهای رفتاری مانند میزان رضایت آنها بستگی داشته باشد، می‌تواند بر روی خوشه های k میانگین اثر منفی داشته باشد و نتیجه قابل قبولی نداشته باشد [۴] و استراتژی‌های بازاریابی متناسب با هر یک از آنها را توسعه دهد. در تعیین داده‌های اولیه هر خوشه باید توجه داشت که مراکز اولیه مناسب می‌توانند همگرایی k میانگین را تسریع کنند. در روش پیشنهادی، اولین قدم در الگوریتم، تعیین عنصر اولیه هر خوشه است. این روش در ابتدا با یک خوشه حاوی یک عنصر آغاز می‌شود و در مراحل دیگر، فاصله سایر عناصر از مرکز خوشه‌های مشخص شده را محاسبه می‌کند. با این حال، نزدیکترین داده به میانگین داده به عنوان تنها عضو خوشه اول تعیین می‌شود. پس از اختصاص هر عنصر به یک خوشه، میانگین داده آن خوشه براساس طول عمر مشتری مطابق با (۱) به روز می‌شود. این معادله برای هر ویژگی مرکز خوشه و داده‌های جدید اعمال می‌شود:

$$Ave_{new} = \frac{data_{new} + (c.Ave_{old})}{size_{old} + 1} \qquad (۱)$$

که در آن $data_{new}$ عنصر ورودی جدید به خوشه است، ave_{old} میانگین عناصر قبلی در خوشه است و $size_{old}$ تعداد عناصر خوشه قبل از ورود عنصر جدید به خوشه است. در ادامه، فاصله بین تمام عناصر مجموعه داده و مرکز خوشه محاسبه می‌شود. سپس، دو عنصر استخراج می‌شود. اولین عنصری که کمترین فاصله را با مرکز خوشه دارد عضو جدید خوشه اول است و عنصر دوم که بیشترین فاصله را با مرکز خوشه دارد اولین عنصر خوشه دوم است. روند افزودن دو عنصر به خوشه‌ها تکرار می‌شود به طوری که هر خوشه k، حداقل یک عضو دارد. در نهایت، فاصله بین

[1] CRoss Industry Standard Process

داده‌ها و مرکز خوشه‌ها محاسبه می‌شود و داده‌هایی که بیشترین میانگین فاصله را از کل مرکز خوشه دارند به عنوان اولین عنصر خوشه جدید مطابق (۲) تعیین می‌شوند، و داده‌هایی که کمترین فاصله را از مرکز هر خوشه دارند به عنوان عضوی از آن خوشه در نظر گرفته می‌شوند.

$$b = \max\left\{\frac{\sum_{k\geq ave\geq 1} dist(x,ave)}{k}\right\}_{p\geq x\geq 1} \quad (۲)$$

که در آن x داده جدید است، ave میانگین داده‌های موجود در خوشه‌ها، k تعداد خوشه‌های بهینه و p تعداد داده‌های باقیمانده در مجموعه داده‌ها است. مرکز هر خوشه، پس از افزودن عناصر به آنها، مطابق با (۳) به روز می‌شود. برای دریافت خوشه k با حداقل یک عضو. سپس، با استفاده از (۴)، نزدیکترین داده به مرکز هر خوشه پیدا شده و به آن اضافه می‌شود:

$$a = \min\left\{\frac{\sum_{k\geq ave\geq 1} dist(x,ave)}{k}\right\}_{p\geq x\geq 1} \quad (۳)$$

$$c = \min\{dist(x,a)\}_{x\in m(a)} \quad (۴)$$

که در آن m(a) عدد خوشه مورد نظر (a) در (۳) است و x عنصری از مجموعه داده‌ها است. بعد از اینکه همه عناصر مجموعه داده اصلی خوشه‌بندی شدند، میانگین هر خوشه نیز موجود است و می‌تواند به عنوان مرکز خوشه‌ها در k میانگین استفاده شود.

برای به‌دست آوردن تعداد خوشه در روش پیشنهادی، شناسایی نقاطی که در آنها یک جهش بزرگ در طول عمر مشتری مشتری ایجاد می‌شود، ضروری است. برای یافتن نقاط با یک پرش بزرگ، داده‌ها ابتدا از نظر طول عمر مشتری به ترتیب صعودی مرتب می‌شوند. سپس، مقدار فاصله اقلیدسی هر داده از داده‌های قبلی آن به ترتیب صعودی طول عمر مشتری به‌دست می‌آید.

۴- ارزیابی

۴-۱- آمار توصیفی متغیرهای تحقیق

برای ارزیابی الگوریتم k میانگین در این تحقیق، از روی اطلاعات فروش ۱۳۲۲ بازاریاب استفاده شده است؛ که از ۷۵ نفر این افراد نظرسنجی در مورد میزان رضایت آنان از همکاری با شرکت‌های بازاریابی، صورت پذیرفته است. نتیجه این نظرسنجی تحت عنوان پارامتر M_2 استفاده شده است. جدول (۱) مجموعه داده‌های مورد استفاده در آزمایش‌ها را نشان می‌دهد. با طبقه‌بندی مشتریان در ۱۰ خوشه، دانش لازم از این خوشه‌ها باید استخراج شود تا خوشه‌ها، ارزیابی و تحلیل شوند. از آنجا که بازاریابان در یک خوشه از ویژگی‌های رفتاری مشابهی برخوردارند، دانش استخراج شده می‌تواند راهنمای خوبی برای اتخاذ استراتژی‌های بهینه متناسب با هر خوشه برای مدیریت بهتر روابط مشتری باشد.

۴-۲- بررسی طول عمر مشتری

مقدار انتخاب شده برای *Custno* به تعداد سوابق موجود در مجموعه داده بستگی دارد. این متغیر باید عددی بین ۲۰۰۰ و ۵۰۰۰ باشد [۴].

در این تحقیق به دلیل کم بودن تعداد نمونه ها، مقدار این متغیر ۲۴۵۰ تا ۴۰۰۰ در نظر گرفته شده تا تعداد کلاس‌ها زیاد شود. داده‌هایی که در یک کلاس قرار می‌گیرند مقدار طول عمر مشتری یکسانی دارند. سپس میانگین کل این طول عمر مشتری‌ها برحسب *Custno* محاسبه شده است. نقطه مرکز در روش k میانگین، نقطه‌ای است که به عنوان نماینده خوشه عمل می‌کند و میانگین همه نقاط داده در خوشه است که در هر مرحله (تا زمان همگرایی) تغییر می‌کند. در تعیین داده‌های اولیه هر خوشه، مراکز اولیه مناسب می‌توانند همگرایی k میانگین را تسریع کنند. نتایج نهایی به مقداردهی اولیه مرکز بستگی دارد.

۴-۳- متوسط مقادیرمشخصه‌ها

جدول‌های (۲، ۳) متوسط مقادیرمشخصه‌ها در هر خوشه را با احتساب M_2 نشان می‌دهد. با مقایسه این دو جدول، می‌توان دریافت که روش پیشنهادی و روش k میانگین صرفاً به ازای خوشه اول، مقادیر مشخصه‌های همسانی حاصل نموده‌اند. همچنین در روش پیشنهادی، مقادیر طول عمر مشتری نسبت به روش k میانگین کلاسیک در هریک از خوشه‌ها بیشتر است. رابطه میان میزان چرخه عمر مشتری به ازای داده‌های هریک از خوشه‌ها در روش پیشنهادی و روش مورد مقایسه (k میانگین کلاسیک) با احتساب و بدون احتساب M_2 در شکل-های (۲-۵) نشان داده شده است. از شکل‌ها مشخص است که ارزش طول عمر مشتری در هر دو روش پیشنهادی و مورد مقایسه بادر نظر گرفتن معیار M_2 در بازه [0-0.6]، و بدون M_2 که در بازه [0-0.4]، است. به عبارتی دیگر می‌توان گفت در هر دو روش، وقتی معیار M_2 مورد ملاحظه قرار گیرد، با روند افزایشی در بازه [0-0.6] مواجه خواهیم بود؛ اما درصورتیکه در هر دو روش، معیار M_2 مورد ملاحظه قرار نگیرد، آنگاه شاهد روند افزایشی در بازه [0-0.4] خواهیم بود. از طرفی در خصوص مقایسه میان دو این روش در حالت در نظر گرفتن M_2، می‌توان ادعا کرد که به ترتیب با ۱۰ و ۸ خوشه مواجه هستیم به طوریکه روش پیشنهادی از خوشه ۸ ام به بعد با افزایش چرخه عمر مشتری بیش از ۰/۴ مواجه است؛ اما در روش k میانگین ساده، صرفا داده‌ها در ۸ خوشه قرار گرفته‌اند و از خوشه ۷ام به بعد تا خوشه ۸ ام، روند افزایشی بیشتر از ۰/۴ مواجه است.

در خصوص مقایسه میان دو روش پیشنهادی و روش k میانگین در حالت فقدان معیار M_2، می‌توان ادعا کرد که به ترتیب برای هریک از روش‌ها با ۱۰ و ۸ خوشه مواجه هستیم به طوریکه روش پیشنهادی از خوشه ۸ ام به بعد با افزایش چرخه عمر مشتری بیش از ۰/۳ مواجه است اما در روش k میانگین ساده، صرفا داده‌ها در ۸ خوشه قرار گرفته‌اند و از خوشه ۷ام به بعد تا خوشه ۸ام، روند افزایشی بیشتر از ۰/۳ مواجه است.

جدول (۱) : داده های مورد استفاده در آزمایش (با احتساب M2)

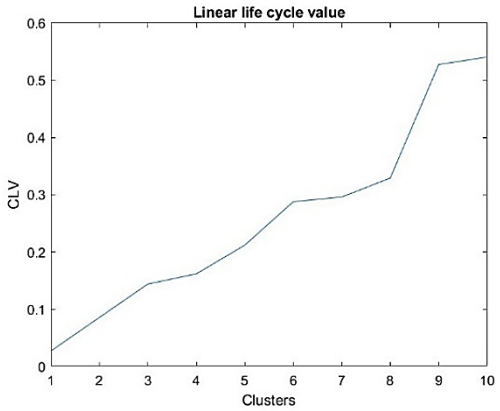
id	M_2	L	R	F	M	CLV	Distance	میانگین
91493295	0/000	0/224181818	0/01	0/139636364	0/014450618	0/388	0	
91532794	0/000	0/224181818	0/01	0/122181818	0/034399852	0/391	0/002495	0/002495
91357635	0/000	0/249090909	0	0/104727273	0/037178989	0/391	0/002728	0/001364
91607842	0/000	0/274	0	0/104727273	0/012628859	0/391	0/003087	0/001029
91732934	0/000	0/199272727	0	0/139636364	0/053092157	0/392	0/003732	0/000933
91638719	0/000	0/274	0	0/052363636	0/066387202	0/393	0/004482	0/000896
91582741	0/000	0/199272727	0/03	0/139636364	0/02518098	0/394	0/005821	0/00097
91722254	0/000	0/174363636	0/02	0/122181818	0/078402874	0/395	0/00668	0/000954
91746476	0/000	0/199272727	0	0/122181818	0/073726335	0/395	0/006912	0/000864
91417752	0/000	0/224181818	0/02	0/122181818	0/028919224	0/395	0/007014	0/000779
91701381	0/000	0/249090909	0	0/122181818	0/02493784	0/396	0/007942	0/000794
91520394	0/000	0/249090909	0/01	0/122181818	0/018594896	0/400	0/011599	0/001054
91767664	0/000	0/149454545	0	0/104727273	0/147480363	0/402	0/013393	0/001116
91689041	0/000	0/199272727	0/03	0/139636364	0/03324273	0/402	0/013883	0/001068
91424376	0/000	0/249090909	0/01	0/069818182	0/082112004	0/411	0/022752	0/001625
91644570	0/000	0/249090909	0	0/139636364	0/025876877	0/415	0/026335	0/001756
91720654	0/000	0/174363636	0/02	0/122181818	0/102055029	0/419	0/030332	0/001896
91007783	0/000	0/224181818	0/02	0/157090909	0/027525764	0/429	0/04053	0/002384
91664278	0/000	0/274	0	0/139636364	0/015235158	0/429	0/040603	0/002256

جدول (۲): متوسط مقادیر L، R، F، M، M2 و طول عمر مشتری در هر خوشه (روش پیشنهادی با احتساب M2)

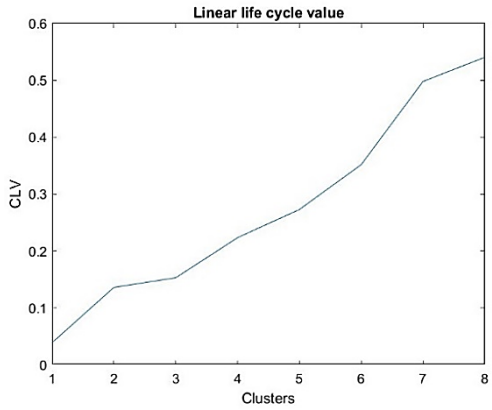
شماره خوشه	L	R	F	M	M_2	CLV
۱	0.180303	0.04097	0.104212	0.048414	0.166667	0.540565
۲	0.054225	0.08043	0.028609	0.016011	0.032864	0.21214
۳	0.160709	0.047914	0.084666	0.030971	0.20339	0.527649
۴	0.007419	0.004654	0.004511	0.010772	0	0.027355
۵	0.107386	0.065844	0.058481	0.020553	0.035714	0.287978
۶	0.025352	0.02174	0.014298	0.013749	0.010695	0.085836
۷	0.136631	0.05821	0.069497	0.02565	0.039216	0.329205
۸	0.043022	0.06414	0.021408	0.01467	0.018779	0.162019
۹	0.044263	0.042651	0.02303	0.016529	0.017778	0.14425
۱۰	0.105822	0.066288	0.055661	0.02355	0.044944	0.296265

جدول (۳): متوسط مقادیر L، R، F، M، M2 و طول عمر مشتری در هر خوشه (روش k میانگین ساده با احتساب M2)

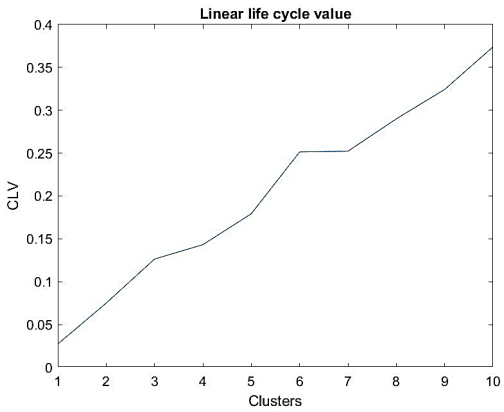
شماره خوشه	L	R	F	M	M_2	CLV
۱	0.180303	0.04097	0.104212	0.048414	0.166667	0.540565
۲	0.059724	0.079007	0.03177	0.016167	0.036885	0.223554
۳	0.15427	0.050138	0.081967	0.029628	0.181818	0.497821
۴	0.012255	0.007985	0.007397	0.011569	0	0.039206
۵	0.106294	0.066286	0.055349	0.023131	0.021978	0.273037
۶	0.041183	0.035538	0.021756	0.016565	0.021277	0.136319
۷	0.13595	0.057873	0.070347	0.02697	0.060606	0.351747
۸	0.040336	0.063055	0.019993	0.014	0.015748	0.153133



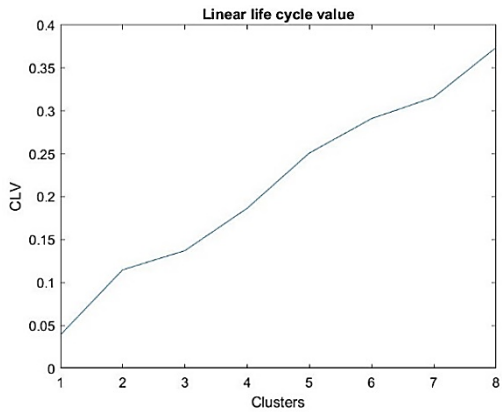
شکل (۲): روش پیشنهادی با احتساب M_2



شکل (۳): k میانگین کلاسیک با احتساب M_2



شکل (۴): روش پیشنهادی بدون M_2



شکل (۵): k میانگین کلاسیک بدون M_2

۴-۴- مقایسه کیفیت خوشه

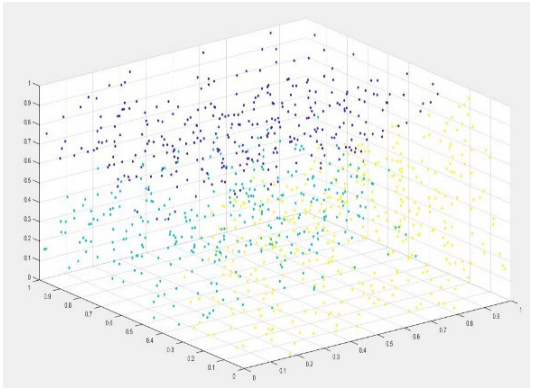
معیار خوشه‌بندی که در آزمون‌ها برای اندازه‌گیری کیفیت خوشه‌ها استفاده می‌شود به‌صورت زیر تعریف شده است [۱۵]:

$$J(x,y)=\sum_{i=1}^{n}\|x_n-\mu(x_i)\|^2 \qquad (۵)$$

که در آن x_i نقطه داده iام است است و $\mu(x_i)$ نزدیک‌ترین مرکز برای نقطه داده i را نشان می‌دهد. برای هر نقطه، مجذور فاصله بین نقطه و نزدیکترین مرکز آن محاسبه می‌شود و این مقدار برای تمام نقاط داده جمع می‌شود. این معیار که خطای خوشه‌بندی نیز نامیده می‌شود، روشی بسیار مؤثر برای اندازه‌گیری کیفیت خوشه‌بندی خانواده الگوریتم‌های خوشه‌بندی k میانگین است. هرچه مقدار معیار بیشتر باشد، کیفیت خوشه‌های تولید شده پایین تر است و بالعکس.

۵-۴- تجزیه و تحلیل نتایج خوشه‌بندی

هدف از این تحقیق طبقه‌بندی بازاریابان بر اساس طول عمر مشتری آن‌ها با استفاده از ویژگی‌های $LRFMM_2$ و روش بهبود یافته k میانگین است. همچنین، تجزیه و تحلیل خوشه‌ها به منظور ارزیابی دقیق‌تر ویژگی‌های مشتریان و تقسیم‌بندی نهایی مشتریان در قالب هرم طول عمر مشتری از موارد خاص مورد استفاده در این مطالعه برای جلب رضایت بازاریابان است. دو خوشه‌بندی متفاوت بر روی داده‌ها صورت گرفته است. خوشه‌بندی اول با در نظر گرفتن پارامتر M_2 (میزان رضایت بازاریابان) و خوشه‌بندی دوم بدون در نظر گرفتن M_2، صورت گرفته است. در خوشه‌بندی اول مشتریان در همه خوشه‌ها به غیر از خوشه‌های اول تا پنجم در این هرم از نظر ویژگی‌های L، R، F، M، M_2 در وضعیت مطلوبی قرار دارند. بنابراین، برای حفظ این مشتریان، به شرکت پیشنهاد می‌شود که از طریق برقراری ارتباط و تعامل با آن‌ها، رفتار وفاداری این مشتریان را به نمای وفاداری تبدیل کند. بر این اساس، مشتریان در خوشه‌های ۷ تا ۱۰ از نظر ویژگی‌های L، R، F، M، M_2 بهترین شرایط را دارند. در خوشه‌بندی دوم، داده‌ها به سه خوشه تقسیم شده‌اند. ماکزیمم خوشه‌ها با توجه به تعداد دسته‌بندی تغییر می‌کند؛ ولی به طور کلی عدد بهینه در یک بازه محدود است. ماکزیمم و نقطه اکسترمم یا ۳ یا ۴ یا ۵ است. شکل (۶)، خوشه‌بندی بازاریابان را نشان می‌دهد.



شکل (۶): خوشه بندی بازاریابان

۴-۶- سرعت k میانگین بهبود یافته

در k میانگین بهبود یافته از شاخص دیویس بولدین[1] استفاده شده است که این امکان را می‌دهد از ابتدا تعداد خوشه‌ها مشخص باشد و با توجه به داده‌ای که موجود است اندازه‌گیری بهتری بدهد و سرعت را افزایش دهد. چنانچه از روش دیویس بولدین استفاده نشود یک بار باید کل مجموعه را در نظر گرفت و دفعه بعد برای k=2 الگوریتم k میانگین اجرا می‌شود و به همین ترتیب تا k=10 بایستی محاسبه شود و درنهایت ۱۰ نتیجه با هم مقایسه شوند که کدامیک برای خوشه‌بندی است. ولی در روش دیویس بولدین تنها دو بار الگوریتم اجرا می‌شود؛ بنابراین محاسبات کاهش یافته و به تبع آن سرعت الگوریتم افزایش می‌یابد.

۴-۷- پیش‌بینی احتمال خریدکردن در ماه‌های آتی

در این تحقیق، خرید ۱۰ ماه بازاریابان در سال ۱۳۹۶ مورد بررسی قرار گرفت. حال برای پیش‌بینی وفاداری بازاریابان, آخرین ماه خرید در نظر گرفته نمی‌شود و مجدد پارامترهای F, R,M l, محاسبه می‌شوند. سپس روش دیویس بولدین بر روی آن انجام می‌پذیرد تا تعداد بهینه‌ی خوشه‌ها مشخص گردد. حال خوشه‌بندی مشخص می‌کند که مشتریان وفادار با کدام ID ها هستند. سپس بررسی شده که آیا در ماه آخر، فرد خرید داشته است یا خیر. از تعداد ۱۰۶۹ بازاریابی که در ۹ ماه اول سال ۱۳۹۶ خرید انجام داده‌اند؛ تعداد ۱۵۵ نفر از آنها در دی ماه نیز خرید کرده‌اند؛ که نشان می‌دهد ۱۴.۵ درصد از بازاریابان قبلی تمایل به خرید مجدد در ماه آتی داشته‌اند و این بیانگر وضعیت نه چندان مطلوب وفاداری بازاریابان قدیمی نسبت به شرکت می‌باشد.

۴-۸- معیار اطلاعات متقابل نرمالیزه

یکی از راه‌های سنجش دقت تشخیص روش‌های مبتنی بر خوشه‌بندی با آزمایش بر روی داده‌های مختلف به‌کارگیری معیار اطلاعات متقابل نرمالیزه[2] می‌باشد [16]. فرم کلی این معیار در (۶) آورده شده و در ادامه به شرح جزئیات آن پرداخته می‌شود.

$$NMI = \frac{2\sum_i \sum_j n_{ij} \log \frac{n_{ij} n}{b_i d_j}}{-\sum_i b_i \log \frac{b_i}{n} - \sum_j d_j \log \frac{d_j}{n}} \qquad (۶)$$

که در آن، عبارت صورت نشان‌دهنده اطلاعات متقابل و عبارت مخرج نشان‌دهنده آنتروپی می‌باشد. جدول سازگاری از سنجش شباهت میان نتیجه تشخیص و بخش درستی از هر خوشه موجود به دست می‌آید؛ بدین صورت که یک مجموعه خوشه V با n گره و دو بخش $C = \{c_1, c_2, \dots, c_n\}$ (نتیجه تشخیص) و $P = \{p_1, p_2, \dots, p_i\}$ (بخش درستی یا صحت) در نظر گرفته شده و همپوشانی‌های بین دو بخش C و P در جدول امکان است. جدول (۴) یافته‌های حاصل از معیار NMI بر روی روش پیشنهادی را نشان می‌دهد. جدول (۵)، یافته‌های معیار مجموع مربعات خطا[3] را نشان می‌دهد که حاکی از بهبود روش پیشنهادی در مقایسه با روش k میانگین کلاسیک از نظر ملاحظه و عدم ملاحظه M_2 است. همچنین روش پیشنهادی با روش مطرح شده در [۴] مقایسه شده است.

برای نشان دادن میزان شباهت بین دو روش خوشه‌بندی با روش‌های دیگر داده‌کاوی می‌توان از "شاخص رند متعادل شده" [18] بهره گرفت. برای محاسبه آن باید دو پارامتر را اندازه‌گیری کنیم:

- A: تعداد جفت عبارت‌هایی که در خوشه‌ها و در واقعیت کنار هم هستند. به بیان دیگر هم در خوشه‌ها دارای برچسب یکسان هستند و هم برچسب در واقعیت (داده‌های واقعی) تطبیق دارد.
- B: تعداد جفت‌هایی است که هم در خوشه‌ها و هم در داده واقعی از یکدیگر مجزا می‌باشند. یعنی برچسب نمونه داده‌های در خوشه‌های آن‌ها متفاوت است و همچنین برچسب داده‌های واقعی آن‌ها نیز با یکدیگر فرق دارد.

جدول (۶) یافته‌های حاصل از شاخص رند تعدیل یافته را نشان می‌دهد در این حالت، روش پیشنهادی در حالت درنظر گرفتن معیار M2، بهترین عملکرد را نسبت به تمامی روش‌های دیگر حاصل نموده است و اما از نظر مقایسه میان روش پیشنهادی از دید احتساب و عدم احتساب M2، اولویت با روش پیشنهادی با احتساب M2 است و از نظر مقایسه میان روش k میانگین کلاسیک از دید احتساب و عدم احتساب M2، برتری با روش k میانگین با احتساب M2 است.

جدول (۴): دقت NMI – مقایسه روش پیشنهادی

روش	معیار NMI
روش پیشنهادی با احتساب M2	۰/۹۷۸۵
روش پیشنهادی بدون احتساب M2	۰/۸۹۹۹

جدول (۵): یافته‌های SSE

روش	معیار SSE
روش پیشنهادی با احتساب M2	**۰/۰۷**
روش پیشنهادی بدون احتساب M2	**۰/۰۲**
روش پیشنهادی [۴] بر روی دیتاست ۱	**۰/۳۹**
روش پیشنهادی [۴] بر روی دیتاست ۲	**۰/۷۲**

جدول (۶): شاخص رند تعدیل یافته

روش	شاخص رند تعدیل یافته
روش پیشنهادی با احتساب M2	**۰/۹۵۲۱**
روش پیشنهادی بدون احتساب M2	**۰/۸۶۵۲**

[1] Davies–Bouldin index
[2] Normalized Mutual Information (NMI)
[3] Sum of squared errors (SSE)

۵- نتیجه‌گیری

هدف اصلی تحقیق، استفاده از الگوریتم k میانگین و یافتن میزان طول عمر مشتری با استفاده از پارامترهایی از جمله فاصله بین اولین و آخرین خرید بازاریاب به عنوان ویژگی L، فاصله بین آخرین خرید مشتری و پایان دوره زمانی به عنوان ویژگی R، تعداد ماه‌های خرید مشتری در بازه زمانی مشخص به عنوان ویژگی F، ارزش کل پولی بین اولین و آخرین خرید مشتری در یک بازه زمانی خاص به عنوان ویژگی M و میزان رضایت بازاریابان به عنوان ویژگی M2 است. برای یافتن نقاط با یک پرش بزرگ، داده‌ها ابتدا از نظر طول عمر مشتری به ترتیب صعودی مرتب می‌شوند. سپس، مقدار فاصله اقلیدسی هر داده از داده‌های قبلی به ترتیب صعودی طول عمر مشتری به‌دست می‌آید. فاصله بین برخی داده‌ها و مرکز خوشه اول و مقادیر مراکز بهینه به‌دست آمده محاسبه می‌شود. پس از آن، این مراحل برای به‌دست آوردن تمام مراکز خوشه‌ای تکرار می‌شوند. با استفاده از معادلات، تعداد بهینه خوشه‌ها ۱۰ محاسبه شده است. هرچه که میزان رضایت بازاریابان کمتر باشد، متغیر طول عمر مشتری نیز کاهش می‌یابد. نتایج نشان می‌دهد که روش پیشنهادی تعداد خوشه‌ها را به درستی شناسایی کرده است؛ اما این کار در زمان کوتاه‌تری انجام شده است.

مراجع

[1] Alizadeh Zoeram, A. and A.R. Karimi Mazidi, *New approach for customer clustering by integrating the LRFM model and fuzzy inference system.* Interdisciplinary Journal of Management Studies (Formerly known as Iranian Journal of Management Studies), 2018 :(۲)۱۱ .p. 351-378.

[2] Nugroho, B.I., et al., *Customer segmentation in sales transaction data using k-means clustering algorithm.* Journal of Intelligent Decision Support System (IDSS), 2024. **7**(2): p. 130-136.

[3] Sarkar, M., A.R. Puja, and F.R. Chowdhury, *Optimizing Marketing Strategies with RFM Method and K-Means Clustering-Based AI Customer Segmentation Analysis.* Journal of Business and Management Studies, 2024. **6**(2): p. 54-60.

[4] Zare, H. and S. Emadi, *Determination of Customer Satisfaction using Improved K-means algorithm.* Soft Computing, 2020. 24(22): p. 16947-16965.

[5] Farhan, M. and J. Heikal, *Used Car Customer Segmentation Using K-Means Clustering Model With SPSS Program: Case Study Caroline. Id.* Jurnal Indonesia Sosial Sains, 2024.

[6] Rachwał, A., et al., *Determining the quality of a dataset in clustering terms.* Applied Sciences, 2023. **13**(5): p. 2942.

[7] Jiang, T. and A. Tuzhilin, *Improving personalization solutions through optimal segmentation of customer bases.* IEEE transactions on knowledge and data engineering, 2008. **21**(3): p. 305-320.

[8] Shah, S. and M. Singh. *Comparison of a time efficient modified K-mean algorithm with K-mean and K-medoid algorithm*. in *2012 international conference on communication systems and network technologies*. 20 .۱۲ IEEE.

[9] Cho, Y.S. and S.C. Moon, *Weighted mining frequent pattern based customer's RFM score for personalized u-commerce recommendation system.* JoC, 2013. **4**(4): p. 36-40.

[10] He, X. and C. Li. *The research and application of customer segmentation on e-commerce websites*. in *2016 6th International Conference on Digital Home (ICDH)*. 2016. IEEE.

[11] Zahrotun, L. *Implementation of data mining technique for customer relationship management (CRM) on online shop tokodiapers. com with fuzzy c-means clustering* .in *2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*. 2017. IEEE.

[12] Hu, J., et al., *Consensus multiple kernel K-means clustering with late fusion alignment and matrix-induced regularization.* IEEE Access, 2019. **7**: p. 136322-136331.

[13] Khanizadeh, F., F. Khamesian, and A. Bahiraie, *Customer segmentation for life insurance in Iran using k-means clustering.* International Journal of Nonlinear Analysis and Applications, 2021. **12**(Special Issue): p. 633-642.

[14] Anitha, P. and M.M. Patil, *RFM model for customer purchase behavior using K-Means algorithm.* Journal of King Saud University-Computer and Information Sciences, 2022. **34**(5): p. 1785-1792.

[15] Macqueen, J., *Some methods for classification and analysis of multivariate observations*. 1967: University of California Press.

[16] Danon, L., et al., *Comparing community structure identification.* Journal of statistical mechanics: Theory and experiment, 2005. **2005**(09): p. P09008.

[17] Amelio, A .and C. Pizzuti. *Is normalized mutual information a fair measure for comparing community detection methods?* in *Proceedings of the 2015 IEEE/ACM international conference on advances in social networks analysis and mining 2015*. 2015.

[18] Rand, W.M., *Objective criteria for the evaluation of clustering methods.* Journal of the American Statistical association, 1971. **66**(336): p. 846-850.