به کارگیری الگوریتم k-means وزنی هوشمند با متریک مینکوفسکی به منظور خوشه‌بندی مشتریان

(مورد مطالعاتی مشتریان یک شرکت ارائه دهنده خدمات ارتباطی)

نرگس احمدزاده گلی

شرکت مخابرات ایران، منطقه تهران،

سید علی ملک جعفریان

شرکت مخابرات ایران، منطقه تهران،

حسن متقی محمودی

شرکت مخابرات ایران، منطقه تهران،

امیرحسن شیخ انصاری

شرکت مخابرات ایران، منطقه تهران،

چکیده

خوشه‌بندی k-means وزنی هوشمند مینکوفسکی تعمیمی از خوشه‌بندی k-میانگین است که در آن تعداد خوشه‌ها و مراکز مربوطه را می‌توان مشخص کرد درحالی‌که تابع زیان مینکوفسکی به‌عنوان معیار عدم تشابه در نظر گرفته می‌شود. بنابراین برخلاف الگوریتم های k-میانگین ساده، انتخاب مراکز در هر خوشه تصادفی نمی باشد، در نتیجه نیازی به تکرار چندین مرتبه الگوریتم به منظور حذف اثرات تصادفی مراکز اولیه نیست. انتخاب معیار عدم تشابه مینکوفسکی به جای اقلیدسی یا سایر معیارهای عدم تشابه، به پژوهشگر در بهینه سازی بهتر مراکز بویژه زمانی که میان داده ها، داده پرت وجود داشته باشد، کمک می کند و درنتیجه دقت خوشه بندی افزایش می یابد. در این پژوهش، کارکرد الگوریتم یادشده بر برخی پایگاه داده‌های شبیه‌سازی‌شده که خوشه های آنها، از پیش تعیین شده است، بررسی می‌شود. به منظور سنجش دقت الگوریتم در خوشه بندی داده ها از معیارهای اطلاعات نرمال شده (NVI) و همچنین درصد داده هایی که به درستی خوشه بندی شده‌اند، استفاده می شود. در نهایت عملکرد الگوریتم بر پایگاه داده های مشترکین یک شرکت ارائه دهنده خدمات ارتباطی به منظور شناسایی مشتریان خاص و ارائه بسته های تشویقی در راستای توسعه سرویس ها و همچنین نگهداشت مشتریان ارزشمند سنجیده می شود.

واژگان کلیدی: معیار عدم تشابه، k-میانگین هوشمند وزنی، خوشه‌بندی ، تابع زیان مینکوفسکی

۱- مقدمه

۱-۱-الگوریتم k-میانگین

الگوریتم های k-میانگین یکی از محبوب‌ترین و مورد استفاده‌ترین الگوریتم‌های خوشه‌بندی می‌باشند و برای اولین بار توسط مک کویین (Macqueen، ۱۹۶۷) معرفی شد. این الگوریتم ها، از الگوریتم‌های سلسله‌مراتبی بسیار کارآمدتر می باشند و بدین‌منظور طراحی شده اند تا داده‌های عددی را خوشه‌بندی کنند به‌طوری که هر خوشه دارای یک مرکز به نام میانگین است. در این الگوریتم فرض برآن است که تعداد خوشه‌ها یعنی K ثابت باشد، همچنین در آن یک تابع خطا وجود دارد. نتایج الگوریتم به معیار فاصله‌ای که درآن به‌کار برده می‌شود، می تواند بستگی داشته‌باشد. الگوریتم مرسوم k-میانگین به‌صورت زیر تعریف می شود:

فرض کنید $\boldsymbol{X}$ یک مجموعه داده با n مشاهده باشد، به‌عبارتی، $\underline{\boldsymbol{X}} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)'$، همچنین هر یک از مشاهدات دارای m ویژگی باشد، یعنی، $\boldsymbol{X}_i = (X_{i1}, \dots, X_{im})$ .k-میانگین الگوریتمی است که پایگاه داده $\underline{\boldsymbol{X}}$ را به K خوشه مجزای $\boldsymbol{S} = \{S_1, \dots, S_K\}$ از مشاهدات مشابه تقسیم می‌کند به‌طوری‌که در هر خوشه عدم‌تشابه بین هر مشاهده و مرکز آن (که به طور تصادفی از مجموعه داده‌ها انتخاب می‌شود) مینیمم می‌شود. تابع زیان بین مشاهده $\boldsymbol{X}_i$ در خوشه k و مرکز مربوط به آن یعنی $\boldsymbol{C}_k$ به‌صورت زیر است:

$$J_{k-\text{میانگین}} = \sum_{k=1}^{K} \sum_{\boldsymbol{X}_i \in S_k} L(\boldsymbol{X}_i, \boldsymbol{C}_k), \qquad (۱-۱)$$

که در آن، $\boldsymbol{C}_k = (C_{k1}, \dots, C_{km})$ برای $k = 1, \dots, K$ و L فاصله اقلیدسی بین $\boldsymbol{X}_i \in S_k$ و $\boldsymbol{C}_k$ می‌باشد. هر مشاهده به نزدیکترین مرکز $\boldsymbol{C}_k$ یعنی میانگین آن تخصیص داده می‌شود و تابع زیان $J_{k-\text{میانگین}}$ محاسبه می‌شود. مراکز خوشه‌ها به میانگین S_k به‌روزرسانی می‌شود تا زمانی‌که مراکز جدید در مقایسه با مراکز مرحله قبل تغییر نکنند. برای یک مقدار اولیه K، با قرار دادن داده‌های به‌جای مانده به نزدیک‌ترین خوشه‌ها و سپس تکرار تغییر اعضای آن خوشه با توجه به تابع زیان، تا زمانی که مقدار تابع زیان دیگر به‌طور معنادار تغییری نکند و اعضای خوشه‌ها ثابت بمانند، ادامه می‌یابد. رابطه (۱-۱) را می‌توان به‌صورت زیر بازنویسی کرد.

$$G_{k-\text{میانگین}}(\boldsymbol{H}, \boldsymbol{C}) = \sum_{k=1}^{K} \sum_{i=1}^{n} h_{ik} L(\boldsymbol{X}_i, \boldsymbol{C}_k), \qquad (۱-۲)$$

که در آن $\boldsymbol{C} = (\boldsymbol{C}_1, \dots, \boldsymbol{C}_K)'$ و $\boldsymbol{H}$ یک ماتریس $n \times K$ است به‌طوری‌که برای هر $i = 1, \dots, n$، $h_{ik} \in \{0,1\}$ و $\sum_{k=1}^{K} h_{ik} = 1$ می باشد. اجرای الگوریتم به پیدا کردن تعداد خوشه‌ها و مراکز اولیه بستگی دارد. بنابراین الگوریتم می‌بایست به‌دفعات بسیار تکرار شود تا تأثیر مرکز اولیه کمتر شود. فرآیند تا زمانی ادامه می‌یابد که هیچ مشاهده‌ای خارج از خوشه‌ها باقی نماند.

در یک الگوریتم خوشه‌بندی، ویژگی‌هایی که بار اطلاعات کمی دربر دارند یا از اهمیت برخوردار نیستند، ممکن است باعث ایجاد سردرگمی و طبقه‌بندی اشتباه شوند. زیرا، در مقایسه با سایر ویژگی‌ها ممکن است در حمل اطلاعات، به آنها درجه اهمیت یکسانی اختصاص یابد. چنین

مجموعه داده‌هایی که دارای ویژگی‌های زائد می‌باشند، زیاد مشاهده می‌شوند. در الگوریتم k-میانگین، می‌توان ویژگی وزن را به‌منظور کاهش تأثیر ویژگی‌های اضافه یا خطا، اضافه نمود. هانگ و همکاران (Haung et. Al.، ۲۰۰۸)، الگوریتم خوشه‌بندی k-میانگین وزنی را که به هر یک از ویژگی‌ها، وزن‌های متفاوتی اختصاص می‌دهد، معرفی نمودند.

یک تابع زیان با توجه به ساختار پایگاه داده انتخاب می‌شود. احمدزاده و همکاران (Ahmadzadehgoli, ۲۰۱۹) الگوریتم خوشه‌بندی k-میانگین لاینکس بر پایه تابع زیان نامتقارن لاینکس به جای اقلیدسی را ارائه کردند. آنها نشان دادند که زمانیکه بیش برآوردی یا کم برآوردی مراکز خوشه‌ها دارای اهمیت باشد، می توان به یاری آن مشاهداتی را به خوشه‌ای خاص هدایت کرد. آمورین و کامیزرکزاک (Komisarczuk ، Amorim، ۲۰۱۲) نیز در الگوریتم k-میانگین وزنی به جای تابع زیان اقلیدسی، از تابع مینکوفسکی به‌عنوان معیار عدم‌تشابه استفاده کردند که به آن پرداخته می‌شود.

مشخص نبودن تعداد خوشه‌ها از پیش و همچنین مراکز اولیه، یکی از ضعف های بسیاری از الگوریتم‌های خوشه‌بندی است. الگوریتم k-میانگین هوشمند که توسط میرکین (Mirkin، ۲۰۰۵) معرفی شد، به‌منظور تعیین تعداد خوشه‌ها و مراکز دقیق آنها مفید است. در این روش مشاهدات خوشه‌بندی نشده، در دورترین نقطه از گرانیگاه اولیه قرار می‌گیرد و دورترین نقطه به عنوان یک مرکز آزمایشی در نظر گرفته می‌شود. آنگاه خوشه توسط همه مشاهداتی که به مرکز آزمایشی نسبت به مرکز اولیه نزدیک‌تر است، پر می‌شود. پس از آنکه همه مشاهدات خوشه‌بندی شدند، خوشه‌های کوچک با استفاده از یک مقدار آستانه‌ای از پیش تعیین‌شده، حذف می‌شوند. این روش به علت سادگی، حتی برای افرادی که پیش‌زمینه‌ای از آمار و علوم کامپیوتر ندارند، نیز خوشایند است و نیازی به تکمیل چندین بار الگوریتم برای یافتن بهترین تعداد خوشه را ندارد و درواقع یک الگوریتم قطعی است که تنها به‌یکبار اجرا نیاز دارد.

۲-۱-الگوریتم خوشه‌بندی k-میانگین وزنی مینکوفسکی

هانگ و همکاران (۲۰۰۸)، الگوریتم خوشه‌بندی k-میانگین وزنی را که به هر یک از ویژگی‌ها، وزن‌های متفاوتی اختصاص می‌دهد، معرفی نمودند. این الگوریتم تابع هدف زیر را مینیمم می‌سازد:

$$G_{\beta,۲}(\boldsymbol{H},\boldsymbol{C},\boldsymbol{W}) = \sum_{k=۱}^{K}\sum_{i=۱}^{n}\sum_{j=۱}^{m} h_{ik} w_j^{\beta} \left(\boldsymbol{X}_{ij} - \boldsymbol{C}_{kj}\right)^{۲} \qquad (۱-۲),$$

که در آن، w_j، وزن ویژگی ها، مقداری نامنفی می‌باشد و $\sum_{j=۱}^{m} w_j = ۱$. همچنین β نشانگر تأثیر وزن و $\boldsymbol{C} = (\boldsymbol{C}_۱, \dots, \boldsymbol{C}_K)'$ و $\boldsymbol{H}$ یک ماتریس $n \times K$، به‌طوریکه برای هر $i = ۱, \dots, n$، $h_{ik} \in \{۰,۱\}$ و $\sum_{k=۱}^{K} h_{ik} = ۱$ می‌باشد. هدف مینیمم نمودن تابع (۲-۱) بین $\boldsymbol{X}_i$ و مرکز خوشه مربوطه در هر خوشه با در نظر گرفتن قید $\sum_{j=۱}^{m} w_j = ۱$ می‌باشد. هر یک از وزن‌ها با استفاده از رابطه زیر به‌روزرسانی می‌شود،

$$w_j^* = \begin{cases} 0 & if \;\; E_j = 0 \\ \frac{1}{\sum_{u \in M} (\frac{E_j}{E_u})^{\frac{1}{\beta - 1}}} & if \;\; E_j \neq 0 \end{cases} \qquad (1\text{-}3)$$

که در آن

$$E_j = \sum_{k=1}^{K} \sum_{i=1}^{n} h_{ik} \left| X_{ij} - C_{kj} \right|^2 .$$

الگوریتم k-میانگین وزنی به‌صورت زیر می‌باشد:

۱- تعداد خوشه‌ها یعنی K، مشخص می‌شود. مراکز اولیه خوشه‌ها و وزن‌های ویژگی‌ها، به‌صورت تصادفی انتخاب می‌شود به گونه‌ای که $\sum_{j=1}^{m} w_j = 1$. وزن‌های تصادفی ابتدایی می‌توانند به‌این‌صورت درنظر گرفته شوند، $w_j = \frac{1}{m}$.

۲- هر شئ به نزدیکترین مرکز با حداقل نمودن (۲-۱) اختصاص می‌یابد.

۳- تمام مراکز به میانگین خوشه مربوطه به‌روزرسانی می‌شوند. فرآیند ادامه می‌یابد و اگر تغییری در مراکز خوشه‌های مرحله ۲ ایجاد نشود، متوقف و سپس $\boldsymbol{S} = \{S_1, \dots, S_K\}$ خوشه‌های نهایی خواهدبود.

٤- با استفاده از رابطه (۳-۱)، وزن‌ها با در نظر گرفتن شرط $\sum_{j=1}^{m} w_j = 1$ به‌روزرسانی می‌شوند.

آمورین و کامیزرکزاک (۲۰۱۲) این الگوریتم را به k-میانگین وزنی مینکوفسکی تعمیم دادند که درآن فاصله مینکوفسکی به‌عنوان معیار عدم‌تشابه به‌جای فاصله اقلیدسی به صورت زیر به‌کار می‌رود.

$$d_{Min}(\boldsymbol{X}, \boldsymbol{Y}) = [\sum_{j=1}^{m} (X_j - Y_j)^p]^{\frac{1}{p}} \quad , \quad p \geq 1,$$

که در آن p، مرتبه فاصله مینکوفسکی می‌باشد. اگروال و همکاران (Agrawal،۱۹۹۳) و یی و فالوتسای (Yi، Faloutsos، ۲۰۰۰) و لی (Li، ۲۰۰۰) برخی از کاربردهای این فاصله را بررسی نمودند. اگر p برابر با ۱، ۲ و ∞ باشد آنگاه به‌ترتیب فاصله منهتن، اقلیدسی و فاصله ماکسیمم نتیجه می‌شود.

حال در این الگوریتم تابع هدف زیر (با توجه به $\sum_{j=1}^{m} w_j = 1$) مینیمم می‌شود.

$$J_{\beta,\beta}(\boldsymbol{H}, \boldsymbol{C}, \boldsymbol{W}) = \sum_{k=1}^{K} \sum_{i=1}^{n} \sum_{j=1}^{m} h_{ik} w_j^{\beta} \left| X_{ij} - C_{kj} \right|^{\beta} \qquad (1\text{-}4)$$

اگر قرار دهیم

$$E_j = \sum_{k=1}^{K}\sum_{i=1}^{n} h_{ik}\left|X_{ij} - C_{kj}\right|^{\beta}.$$

با استفاده از E_j، می‌توان (۴-۱) را بصورت زیر بازنویسی نمود،

$$J_{\beta,\beta}(\boldsymbol{H},\boldsymbol{C},\boldsymbol{W}) = \sum_{j=1}^{m} w_j^{\beta}\, E_j$$

حال به‌منظور حداقل نمودن آن باتوجه به قید $\sum_{j=1}^{m} w_j = 1$، می‌بایست با مشتق گرفتن از تابع لاگرانژ زیر نسبت به w_j و قرار دادن آن با صفر

$$w_j = \left(\frac{\lambda}{\beta E_j}\right)^{\frac{1}{\beta-1}}$$

از آنجا که مشتق مرتبه دوم برای $\beta \geq 1$ نامنفی می باشد لذا نقطه بدست آمده، مینیمم می باشد. حال

با جمع بستن این عبارت بر روی j،

$$1 = \sum_{j=1}^{m}\left(\frac{\lambda}{\beta E_j}\right)^{\frac{1}{\beta-1}}$$

$$w_j^{*} = \frac{1}{\sum_{u\in M}\left[\frac{E_j}{E_u}\right]^{\frac{1}{\beta-1}}} \qquad (۱-۵)$$

که مقدار بهینه وزن ویژگی ها در الگوریتم است.

الگوریتم k-میانگین وزنی مینکوفسکی در برخی مراحل با الگوریتم k-میانگین وزنی تفاوت دارد که بصورت زیر می‌باشند:

۱- در مرحله به‌روزرسانی خوشه‌ها، برای یک مجموعه داده شده از مراکز و یک مجموعه از وزنها به ازاء هر ویژگی، خوشه‌ها با استفاده از قانون مینیمم سازی فاصله‌ها، با فاصله مینکوفسکی با توان β، به‌روزرسانی می‌شوند.

۲- در مرحله به‌روزرسانی مراکز، برای یک مجموعه داده شده از خوشه‌ها و یک مجموعه از وزن‌ها، برای هر ویژگی وخوشه، مرکز هر خوشه به مرکز مینکوفسکی آن باتوجه به مینیمم نمودن معادله (۳-۲) به‌روزرسانی می‌شود.

۳- در مرحله به روزرسانی وزن‌ها، برای یک مجموعه داده شده از خوشه‌ها، با توجه به مراکزشان، وزن‌ها با توجه به رابطه (۵-۱) به شرط آنکه جمع آنها برابر با ۱ باشد، به‌روزرسانی می‌شوند.

برای آنکه الگوریتم بالا کاربردی باشد، به یک فرآیند امکان پذیر محاسباتی به‌منظور بدست آوردن مرکز مینکوفسکی ویژگی‌ها برای هر $\beta \geq 1$، نیاز است. همچنین مقدار مینیمم برای $\beta = 1$ که معادل با فاصله بلوک شهری است، برابر با میانه X_i ها می‌باشد. لذا در اینجا جهت مینیمم سازی کافی است $\beta > 1$ در نظر گرفته شود. مقادیر X_i به‌صورت صعودی $X_1 \leq X_2 \leq \cdots \leq X_{N_k}$ مرتب می‌شوند که N_k تعداد اشیا در خوشه k می‌باشد. در ابتدا ثابت

می‌شود که حداقل مقدار یعنی c، در فاصله $[X_i, X_{N_k}]$ قرار دارد. اگر حداقل مقدار در خارج ازاین محدوده باشد، برای مثال $c > X_{N_k}$ ، آنگاه از آنجایی که برای هر $i = ۱,۲, \dots, N_k$ ، $|X_i - X_{N_k}| < |X_i - c|$ می‌شود، لذا $d(X_{N_k}) < d(c)$، که این یک تناقض بوده، پس c در فاصله $[X_i, X_{N_k}]$ است. بنابراین،

$$d(c) = \sum_{i \in I^+} (c - X_i)^\beta + \sum_{i \in I^-} (X_i - c)^\beta$$

که در آن I^+ نشانگر اندیس‌هایی از $i = ۱,۲, \dots, N_k$ است که به ازای آن $c > x_i$ بوده و I^- اندیس‌هایی است که $c \le x_i$ می‌شود. مشتق اول $d'(c)$ به‌صورت زیر می‌باشد.

$$d'(c) = \beta(\sum_{i \in I^+} (c - X_i)^{\beta - ۱} - \sum_{i \in I^-} (X_i - c)^{\beta - ۱})$$

و مشتق دوم نیز به‌صورت زیر است.

$$.d''(c) = \beta(\beta - ۱)(\sum_{i \in I^+} (c - X_i)^{\beta - ۲} + \sum_{i \in I^-} (X_i - c)^{\beta - ۲})$$

برای $\beta > ۱$ معادله بالا مقدار مثبت دارد که محدب بودن $d(c)$ را نشان می‌دهد. محدب بودن این رابطه برای $\beta > ۱$، به ویژگی‌های مفید دیگری نیز منجر می‌شود. فرض کنید $d(X_i^*)$ مینیمم $d(X_i)$ برای هر $i = ۱,۲, \dots, N_k$ باشد و X_i' مقداری از X_i باشد که در میان تمام مقادیری که از X_i^* کوچکترند به X_i^* نزدیک‌تر باشد. به همین ترتیب، اگر X_i''، مقداری از X_i باشد که درمیان تمام مقادیری‌که از X_i^* بزرگترند، به X_i^* نزدیکتر باشد، آنگاه مینیمم $d(c)$ در فاصله $[X_i', X_i'']$ قرار می‌گیرد. به کمک این ویژگی می‌توان الگوریتم زیررا به‌منظور یافتن مرکز مینکوفسکی مجموعه $\{X_i\}$ از مقادیر $X_۱ \le X_۲ \le \cdots \le X_{N_k}$ برای $\beta > ۱$ به‌کار برد. مرکز مینکوفسکی برای $\beta > ۱$ به شرح زیر است:

۱-مقداردهی اولیه: قرار دهید $c_. = X_{i^*}$، که مینیمم $d(c)$ بر مجموعه $\{X_i\}$ است. همچنین یک مقدار مثبت λ که تقریباً می‌تواند ۱۰٪ از فاصله $X_{N_k} - X_۱$ را اختیار کند، تعریف می‌شود.

۲-به‌روزرسانی $c_۱$: $c_. - \lambda d'(c_.)$ محاسبه می‌شود. اگر حاصل آن در فاصله $[X_i', X_i'']$ ، قرار گیرد، آنگاه $c_۱$ درنظر گرفته می‌شود. در غیراین‌صورت، λ حدوداً ۱۰ درصد کاهش می‌یابد و این مرحله دوباره تکرار می‌شود.

۳- شرط توقف: آزمایش می‌شود که آیا $c_.$ و $c_۱$ همزمان در آستانه تعریف‌شده قرار دارند یا خیر؟ اگر در محدوده تعریف‌شده باشند، فرآیند متوقف و $c_۱$ به‌عنوان مقدار بهینه c در نظر گرفته می‌شود. در غیر این صورت به مرحله بعد ارجاع داده می‌شود.

۴-به‌روزرسانی $c.$: اگر $d(c_1) \leq d(c.)$ ، آنگاه $c._=c_1$ و $d(c.) = d(c_1)$ را قرار داده و به مرحله ۲ باز می‌گردد. در غیراین‌صورت λ تقریباً ۱۰٪ کاهش یافته و سپس به مرحله ۲ بازگشت داده می‌شود.

حال الگوریتم خوشه‌بندی k-میانگین وزنی را می توان به نسخه هوشمند تعمیم داد. در این الگوریتم که توسط میرکین (Mirkin) معرفی شده، مراکز اولیه تصادفی نمی‌باشد و خوشه‌های اولیه توسط الگوریتم k-میانگین هوشمند که پیشتر توضیح داده شد، بوجود می‌آید و درادامه با الگوریتم k-میانگین وزنی ادغام می‌شود. الگوریتم فوق به‌صورت زیر می‌باشد:

۱- هریک از وزن‌های اولیه ویژگی‌ها، بصورت مساوی مقداردهی اولیه می‌شوند، به طوریکه جمع آنها برابر با ۱ شود. در این الگوریتم یک پارامتر خارجی به‌کار می رود تا حداقل اندازه خوشه‌ها تعریف شود و مقدار آن به‌طور معمول برابر با ۲ می‌باشد.

۲- اعضایی که دورترین فاصله را تا مرکز ثقل کل مجموعه داده‌ها دارند، به‌عنوان مراکز آزمایشی خوشه‌ها قرار می‌گیرند. خوشه S که از اعضایی که به مرکز آزمایشی نسبت به مرکز ثقل، نزدیکترند، ساخته می‌شود. همچنین وزن‌ها در محاسبات به کمک معیار فاصله زیر دخالت داده می‌شوند،

$$\sum_{j=1}^{m} w_j^{\beta} (\boldsymbol{X}_{ij} - \boldsymbol{C}_{kj})^2$$

مرکز آزمایشی به مرکز ثقل خوشه S یعنی میانگین آن، به‌روزرسانی می‌شود. با توجه به مرکز مربوطه وزن‌ها، با در نظر گرفتن رابطه (۵-۱) به‌روزرسانی می‌شوند. اگر مرکز جدید با مرکز قبلی متفاوت باشد این فرآیند از ابتدای مرحله ۲ تکرار می‌شود. در غیر این صورت متوقف شده و خوشه S از پایگاه داده‌ها حذف می‌شود.

۳- گام ۲ تا زمانی‌که تمام مشاهدات خوشه‌بندی شوند، ادامه می‌یابد.

٤- خوشه‌هایی که تعداد اعضای آن ها کمتر از مقدار آستانه‌ای تعریف شده باشند، حذف می‌شوند.

٥- حال الگوریتم k-میانگین وزنی با مراکز و وزن‌های ایجاد شده، اجرا می‌شود.

اگر در این الگوریتم، به‌جای فاصله اقلیدسی، فاصله مینکوفسکی به‌کار گرفته شود، به الگوریتم حاصل k-میانگین وزنی مینکوفسکی هوشمند گفته می‌شود. با روشی مشابه آنچه قبلا گفته‌شد، یک ثابت مثبت بسیار کوچک به E_j و E_u اضافه می‌شود تا دقت افزایش یابد. سایر قسمت‌های الگوریتم مشابه الگوریتم k-میانگین وزنی مینکوفسکی می‌باشد.

الگوریتم k-میانگین مینکوفسکی وزنی هوشمند نسبت به سایر الگوریتم‌های k-میانگین، ویژگی‌های مثبتی دارد ازجمله آنکه از آنجا که مراکز اولیه تصادفی نمی باشند، لذا یک بار اجرای الگوریتم کفایت می‌کند و نیازی به تکرار چندین مرتبه آن به‌منظور حذف اثرات مراکز تصادفی نمی‌باشد.

پیش از اجرای الگوریتم بر پایگاه داده ها، نرمال سازی اطلاعات به نتایج بهتر منجر می شود.جین و دوبز (Jin، ۱۹۸۸) اظهار داشتند که در بسیاری از مطالعات تحلیل خوشه‌ای، داده‌های خام به‌طور مستقیم مورد استفاده قرار نمی‌گیرند و آماده‌سازی داده‌ها برای تحلیل خوشه‌ای مستلزم برخی تبدیلات مانند استانداردسازی یا نرمال سازی می‌باشد. به کمک استانداردسازی داده‌ها می‌توان اندازه داده‌ها را کم نمود. پس از استانداردسازی می‌توان از داده‌ها یک اندازه مکانی را کسر نمود و حاصل را به یک اندازه مقیاسی برای هر متغیر تقسیم کرد. اگر فرض کنید $\underline{\boldsymbol{X}}^{*\prime} = (\boldsymbol{X}^{*}_{\boldsymbol{1}}, \dots, \boldsymbol{X}^{*}_{\boldsymbol{n}})$ دلالت بر داده‌های استانداردشده داشته‌باشد. آن گاه ماتریس داده‌های استانداردشده یک ماتریس $n \times m$ به‌صورت زیر می‌باشد:

$$(\boldsymbol{X}^{*}_{\boldsymbol{1}}, \dots, \boldsymbol{X}^{*}_{\boldsymbol{n}}) = \begin{bmatrix} X^{*}_{11} & \cdots & X^{*}_{1m} \\ \vdots & \ddots & \vdots \\ X^{*}_{n1} & \cdots & X^{*}_{nm} \end{bmatrix}$$

که در آن

$$X^{*}_{ij} = \frac{X_{ij} - \mu_j}{\sigma_j} \qquad (۱-۶)$$

و X^{*}_{ij} مقدار استانداردشده، μ_j اندازه مکانی و σ_j اندازه مقیاسی می‌باشد. می‌توان روش‌های استانداردسازی مختلفی را با انتخاب متفاوت μ_j و σ_j در معادله (۶-۱) به‌کار گرفت. در برخی از روش‌های استانداردسازی معروف μ_j را میانگین یا میانه و σ_j را انحراف‌معیار یا دامنه‌تغییرات درنظر می‌گیرند.

۳-روش تحقیق

در این پژوهش انگیزه اصلی، استفاده از الگوریتم k-میانگین وزنی هوشمند بر پایه تابع زیان مینکوفسکی به منظور دسته بندی مشتریان یک شرکت ارائه دهنده خدمات ارتباطی است. در این بخش ۲، چند پایگاه داده‌های شبیه‌سازی‌شده از چند توزیع آماری تولید می‌شود و به کمک الگوریتم k-میانگبن وزنی هوشمند با معیار عدم تشابه مینکوفسکی در نرم افزار MATLAB نسخه ۲۰۲۳ خوشه‌بندی می‌گردد. به کارگیری مشاهدات برچسب‌دار سبب می‌شود تا دقت الگوریتم به کمک برخی معیارهای بیرونی سنجش اعتبار خوشه‌بندی (که در بخش بعدی عنوان می‌شود)، ارزیابی گردد. در ادامه نتایج الگوریتم بر پایگاه داده واقعی با بیش از ۳۲۴۰۰۰ مشتری (هر یک با ۷ ویژگی) مورد بررسی قرار می‌گیرد و نتایج جمع بندی می‌شود.

۳-۱- پایگاه داده های شبیه سازی شده

۵ پایگاه داده از توزیع‌های لاگ نرمال، نرمال، گاما و پواسن هرکدام دارای ۱۰۰ متغیر وابسته و m ویژگی، تولید می شود به‌گونه‌ای که در دو گروه، خوشه‌بندی شوند. فرآیند تولید داده‌ها در ادامه نشان داده می‌شود.

در پایگاه داده‌های تولید شده، به ۵۰ متغیر نخست، برچسب ۱ و به ۵۰ متغیر دوم، برچسب ۲، تعلق می‌گیرد. حال الگوریتم k-میانگین هوشمند وزنی مینکوفسکی را بر پایگاه داده‌ها اجرا کرده و با معیار بیرونی تغییر اطلاعات نرمال شده (NVI[1]) (Duda، ۲۰۰۳) دقت آن سنجیده می‌شود. اگر NVI در فاصله [۰،۱] قرار گیرد، بدان معناست که کارکرد الگوریتم مناسب است. هر چقدر مقدار آن کوچک‌تر باشد خوشه‌ها همگن‌ترند.

جدول ۱: معرفی توزیع‌ها و روابط بین متغیرهای تصادفی و مستقل تولید شده از این توزیع‌ها

	$f(x)$	روابط میان متغیرهای تصادفی مستقل
لاگ نرمال (LN)	$, x > ۰, \sigma > ۰, \mu \epsilon R \frac{۱}{x\sigma\sqrt{۲\pi}} \exp\left(-\frac{(\log(x)-\mu)}{۲\sigma^۲}\right)$	ضرب n متغیر تصادفی مستقل و هم توزیع LN دارای توزیع LN است.
نرمال	$, x, \mu \in R, \sigma > ۰ f(x) = \frac{۱}{\sigma\sqrt{۲\pi}} \exp\left(-\frac{(x-\mu)^۲}{۲\sigma^۲}\right)$	جمع n متغیر تصادفی مستقل و هم توزیع نرمال، گاما و پواسن به ترتیب دارای همان توزیع‌ها هستند.
گاما	$, x \geq ۰, \alpha, \beta > ۰ \frac{\beta^\alpha x^{\alpha-۱} e^{-x\beta}}{\Gamma(\alpha)}$	
پواسن	$, x = ۰,۱, \ldots \frac{e^{-\lambda}\lambda^x}{x!}$	

[1] Normalized variation information

جدول ۲: فرآیند تولید چند پایگاه داده از ۱۰۰ متغیر وابسته هرکدام با m ویژگی که در دو گروه خوشه‌بندی می‌شوند

$\boldsymbol{Y} = (\boldsymbol{Y_1}, \dots, \boldsymbol{Y_{100}})'$	**پایگاه داده**	
$\boldsymbol{Y}_i = (X_{i1} + X_{\cdot}, X_{i2} + X_{\cdot}, X_{i3} + X_{\cdot}) = (Y_{ij}) \quad \text{for } j = ۱,۲,۳$		**چند متغیره با ۳ ویژگی**
$, Y_{ij} \overset{id}{\sim} \text{Normal}(۰,۲) \text{ for } i = ۱, \dots, ۵۰ X_{\cdot}, X_{ij} \overset{iid}{\sim} \text{Normal}(۰,۱)$ $X_{ij} \overset{iid}{\sim} \text{Normal}(۴,۱), Y_{ij} \overset{id}{\sim} \text{Normal}(۴,۲) \text{ for } i = ۵۱, \dots, ۱۰۰$	**نرمال**	
$, Y_{ij} \overset{id}{\sim} \text{Gamma}(۲,۳) \text{ for } i = ۱, \dots, ۵۰ X_{\cdot}, X_{ij} \overset{iid}{\sim} \text{Gamma}(۱,۳)$ $X_{ij} \overset{iid}{\sim} \text{Gamma}(۶,۳), Y_{ij} \overset{id}{\sim} \text{Gamma}(۷,۳) \text{ for } i = ۵۱, \dots, ۱۰۰$	**گاما**	
$\boldsymbol{Y}_i = (X_{i1} X_{\cdot}, X_{i2} X_{\cdot}, X_{i3} X_{\cdot}) = (Y_{ij}) \quad \text{for } j = ۱,۲,۳$ $, Y_{ij} \overset{id}{\sim} \text{LN}(۲,۲) \text{ for } i = ۱, \dots, ۵۰ X_{\cdot}, X_{ij} \overset{iid}{\sim} \text{LN}(۱,۱)$ $, Y_{ij} \overset{id}{\sim} \text{LN}(۲۱,۲) \text{ for } i = ۵۱, \dots, ۱۰۰ X_{ij} \overset{iid}{\sim} \text{LN}(۲۰,۱)$	**لاگ نرمال**	
$, \ Y_i = X_i + X_{\cdot} \qquad \boldsymbol{Y} = (Y_{۱}, \cdots, Y_{۱۰۰})$ $, Y_i \overset{id}{\sim} \text{Poisson}(۲) \text{ for } i = ۱, \dots, ۵۰ X_{\cdot}, X_i \overset{iid}{\sim} \text{Poisson}(۱)$ $X_i \overset{iid}{\sim} \text{Poisson}(۱۰), Y_i \overset{id}{\sim} \text{Poisson}(۱۱) \text{ for } i = ۵۱, \dots, ۱۰۰$	**پواسن**	**تک متغیره**

۲-۳- پایگاه داده های مشتریان یک شرکت ارائه دهنده خدمات ارتباطی

شرکت مورد مطالعه دارای ۳۲۴۶۱۹ مشترک است که از سرویس های ارتباطی بهره مند هستند و در ۲۳ محدوده جغرافیایی در کشور ایران واقع شده اند. به دلیل محدودیت های امنیتی اسامی این ۲۳ محدوده از A۱ تا A۲۳ مشخص شده است. هر مشترک دارای ۷ ویژگی کلیدی استفاده از سرویس ها است که با B۱ تا B۷ مشخص شده است. تمام ۷ ویژگی مورد اشاره (شامل میزان استفاده از سرویس ها، سرعت سرویس مورد استفاده، میزان شارژ سرویس، درآمد حاصل از مصرف سرویس ها و صورتحساب و بدهکاری و بستانکاری) برای تمام مشترکین مشخص است و جامعه فاقد هر گونه داده گم شده می باشد. مراحل تجزیه و تحلیل اطلاعات به شرح زیر است:

۱- نرمال سازی داده ها و کاهش بُعد داده ها به کمک تکنیک PCA

۲- خوشه بندی داده ها بر اساس ویژگی های انتخاب شده به کمک الگوریتمk -میانگین وزنی هوشمند مینکوفسکی با برنامه‌نویسی در محیط MATLAB نسخه ۲۰۲۳

۳- انتخاب نهایی برچسب ها بر اساس های معیارهای ارزیابی

۴-توصیف خوشه ها به کمک مراکز بهینه هر خوشه و معیارهای تشابه و عدم تشابه و نمودارها

۵-برنامه ریزی به منظور توسعه سرویس ها و نگهداشت مشتریان بر اساس نتایج خوشه بندی

پس از نرمال سازی داده ها کاهش بُعد داده ها با روش PCA انجام گرفت. با توجه به حجم بالای داده ها، کاهش بُعد سبب افزایش دقت الگوریتم و سرعت پردازش اطلاعات می شود. به منظور انتخاب تعداد خوشه ها به ترتیب داده ها را در ۲، ۳، ۴ و ۵ دسته خوشه‌بندی نموده و با استفاده از ۱۰ معیار ارزیابی خوشه ها که بیشترین همبستگی درون خوشه ای و بیشترین فاصله بین خوشه ها را در نظر می گیرد، تعداد ۳ خوشه تصمیم گیری می شود.

۴-یافته ها

با اجرای الگوریتم بر پایگاه داده های شبیه سازی شده بخش ۳-۱، و تفکیک داده ها در ۲ خوشه، و مقایسه نتایج با برچسب های اولیه، متوسط درصد مشاهداتی که به درستی خوشه بندی شده اند (MA) و معیار تغییر اطلاعات نرمال شده (NVI) در جدول ۳ اندازه گیری شده است.

جدول۳: نتایج الگوریتم k-میانگین وزنی مینکوفسکی هوشمند، در پایگاه داده‌های شبیه‌سازی شده

NVI	AM	پایگاه داده
۰.۰۰	۱۰۰.۰	نرمال
۰.۰۰	۱۰۰.۰	لاگ نرمال
۰.۲۱	۹۸.۰	گاما
۰.۴۲	۹۴.۰	پواسن

نتایج اجرای الگوریتم بر ۴ پایگاه داده شبیه سازی شده گویای دقت بالای الگوریتم در تفکیک سازی خوشه ها است. مقادیر کوچک NVI نشان دهنده خوشه بندی دقیق داده هاست.

در ادامه مشترکین شرکت ارائه دهنده خدمات ارتباطی مورد بررسی قرار می گیرد. به منظور تصمیم گیری جهت انتخاب تعداد خوشه بهینه برای خوشه بندی ۳۲۴۶۱۹ مشترک شرکت هر کدام با ۷ ویژگی، به ترتیب در ۲، ۳ ، ۴ و ۵ خوشه تقسیم می شوند و بر اساس نتایج خوشه ها ۱۰ معیار ارزیابی در جدول ۴ سنجیده می شود.

جدول ۴: نتایج معیارهای ارزیابی به منظور انتخاب بهینه تعداد خوشه ها

validity index	>	<	۲ خوشه	۳ خوشه	۴ خوشه	۵ خوشه
Dunn index	*		۰.۰۶۴۹۰	**۰.۱۱۱۹۱**	۰.۰۰۰۰۶	۰.۰۰۰۰۷
Generalized Dunn-۳۳ criterion	*		۰.۵۶۶۰۳	**۰.۹۵۳۴۸**	۰.۶۵۴۲۲	۰.۰۰۳۲۵
Score Function criterion	*		۰.۰۰۰۰۰	**۰.۲۹۰۷۵**	۰.۰۰۰۰۰	۰.۰۰۰۰۰
Silhouette index.	*		۰.۹۷۸۳۱	**۰.۹۹۶۰۳**	۰.۲۰۳۸۹	۰.۱۲۵۹۹
C-criterion		*	۰.۴	**۰.۱**	۰.۸	۰.۸
COP INDEX		*	**۰.۴**	۱.۳	۹.۶	۷.۴
CS index		*	۳.۴	**۲.۶**	۱۰.۵	۹.۳
Davies-Bouldin criterion		*	۱.۱	**۰.۷**	۲.۷	۳.۵
Davies-Bouldin criterion		*	۱.۱	**۰.۷**	۲.۷	۳.۵
Xie-Beni criterion.		*	**۰.۱**	۱.۰	۶.۴	۴.۷

در جدول ۴، مقادیر بالاتر ۴ شاخص اول و مقادیر پایین تر ۶ شاخص دوم نشان دهنده نتایج بهتر است. در هر ردیف بهترین نتایج پررنگ شده است. مطابق نتایج انتخاب ۳ خوشه بهترین معیارهای ارزیابی را نتیجه می دهد. لذا مشتریان شرکت مورد مطالعه در ۳ خوشه اصلی دسته بندی خواهند شد.

جدول۵- توصیف خوشه ها بر اساس ۷ ویژگی مشترکین

خوشه ها	متوسط سرعت سرویس	میزان استفاده از سرویس	میزان خرید بسته	متوسط درآمد	متوسط صورتحساب	بدهکاری	بستانکاری
خوشه ۱	۱۲,۹۸۲	۱۱۹,۴۴۰	۵۸۸,۴۳۵	۳,۹۶۶,۰۲۲	۵,۸۰۸,۲۶۳	۲۰۶,۹۲۵	۴۴,۶۹۸
خوشه۲	۱۸,۳۴۹	۱۶۳,۹۷۷	۳,۱۵۶,۵۱۷	۹,۷۴۷,۲۵۸	۹,۳۴۰,۱۵۹	۳۰۴,۹۹۰	۵۹,۲۲۸
خوشه۳	۲۹,۶۴۹	۳۹۵,۴۲۱	۱۹,۳۰۵,۵۸۹	۳۱,۴۱۷,۶۹۸	۱۵,۲۴۶,۰۲۰	۵۰۶,۷۸۷	۹۰,۲۳۲

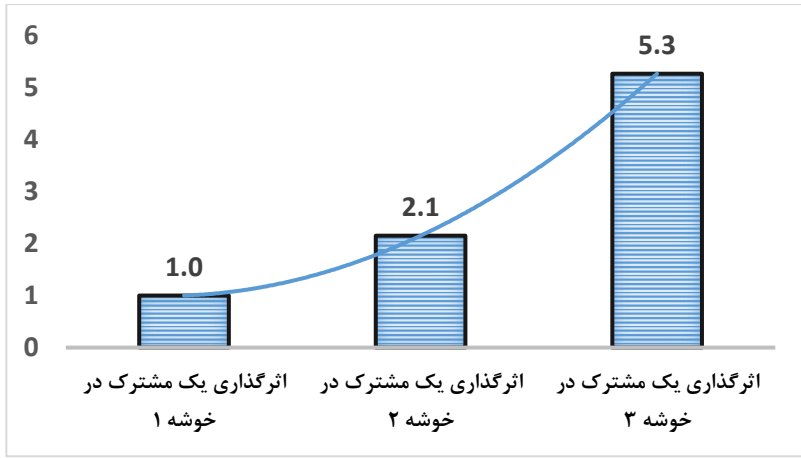
در ادامه در جدول ۶، ۲۳ ناحیه زیر مجموعه شرکت مورد مطالعه در سه خوشه دسته بندی می شوند و به طور کلی خوشه اول با ۵۴ درصد جمعیت، خوشه دوم با ۳۳ درصد و خوشه سوم با ۱۳ درصد از مشترکین شکل می گیرد.

جدول ۶- فراوانی و درصدهای فراوانی خوشه ها به تفکیک ۲۳ ناحیه

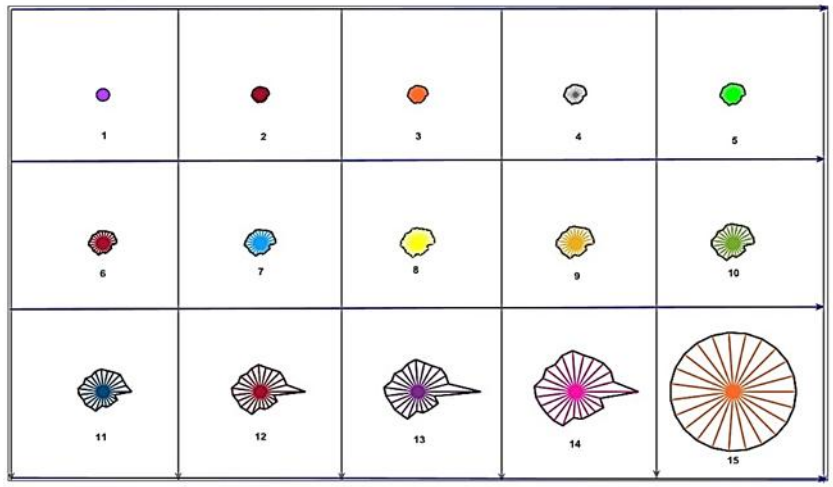
ناحیه	تعداد مشترک خوشه ۱	تعداد مشترک خوشه ۲	تعداد مشترک خوشه ۳	تعداد کل مشترکین	درصد مشترک خوشه ۱	درصد مشترک خوشه ۲	درصد مشترک خوشه ۳
A۱	۳,۶۰۱	۱,۹۹۷	۷۵۵	۶,۳۵۳	۵۷	۳۱	۱۲
A۲	۴,۴۸۴	۱,۶۴۷	۶۸۶	۶,۸۱۷	۶۶	۲۴	۱۰
A۳	۷,۲۹۳	۴,۲۳۷	۲,۰۷۶	۱۳,۶۰۶	۵۴	۳۱	۱۵
A٤	۱,۹۵۹	۱,۲۰۷	۴۷۹	۳,۶۴۵	۵۴	۳۳	۱۳
A٥	۹,۸۷۸	۵,۹۷۵	۳,۲۱۲	۱۹,۰۶۵	۵۲	۳۱	۱۷
A٦	۱۳,۶۳۲	۱۰,۲۶۹	۲,۹۰۰	۲۶,۸۰۱	۵۱	۳۸	۱۱
A۷	۶,۰۳۷	۴,۰۸۶	۲,۲۲۰	۱۲,۳۴۳	۴۹	۳۳	۱۸
A۸	۱۱,۱۱۳	۶,۸۳۷	۲,۶۹۴	۲۰,۶۴۴	۵۴	۳۳	۱۳
A۹	۴,۶۱۳	۲,۶۳۰	۹۹۷	۸,۲۴۰	۵۶	۳۲	۱۲
A۱۰	۱۱,۷۸۱	۷,۹۰۴	۳,۰۱۶	۲۲,۷۰۱	۵۲	۳۵	۱۳
A۱۱	۱۱,۲۴۱	۹,۰۶۰	۴,۸۳۶	۲۵,۱۳۷	۴۵	۳۶	۱۹
A۱۲	۸,۹۵۷	۶,۳۸۷	۲,۹۳۰	۱۸,۲۷۴	۴۹	۳۵	۱۶
A۱۳	۶,۹۰۸	۵,۲۳۰	۱,۷۶۹	۱۳,۹۰۷	۵۰	۳۸	۱۳
A۱٤	۲,۷۲۱	۱,۸۱۱	۶۶۳	۵,۱۹۵	۵۲	۳۵	۱۳
A۱٥	۲۰,۸۷۷	۱۱,۹۱۵	۳,۳۱۸	۳۶,۱۱۰	۵۸	۳۳	۹
A۱٦	۷,۴۳۹	۳,۹۲۰	۱,۳۶۱	۱۲,۷۲۰	۵۸	۳۱	۱۱
A۱۷	۵,۸۱۰	۳,۰۰۶	۹۲۴	۹,۷۴۰	۶۰	۳۱	۹
A۱۸	۱۰,۶۹۰	۵,۹۵۱	۲,۱۰۴	۱۸,۷۴۵	۵۷	۳۲	۱۱
A۱۹	۲,۹۰۷	۱,۲۳۷	۶۰۴	۴,۷۴۸	۶۱	۲۶	۱۳
A۲۰	۸,۰۸۲	۴,۴۷۵	۱,۷۵۳	۱۴,۳۱۰	۵۶	۳۱	۱۲
A۲۱	۸,۸۶۸	۴,۵۳۳	۱,۶۵۲	۱۵,۰۵۳	۵۹	۳۰	۱۱
A۲۲	۳,۹۶۵	۲,۷۳۰	۷۷۴	۷,۴۶۹	۵۳	۳۷	۱۰
A۲۳	۱,۶۸۵	۱,۰۲۷	۲۸۴	۲,۹۹۶	۵۶	۳۴	۹

۱۳	۳۳	۵۴	۳۲۴,۶۱۹	۴۲,۰۰۷	۱۰۸,۰۷۱	۱۷۴,۵۴۱	**مجموع**

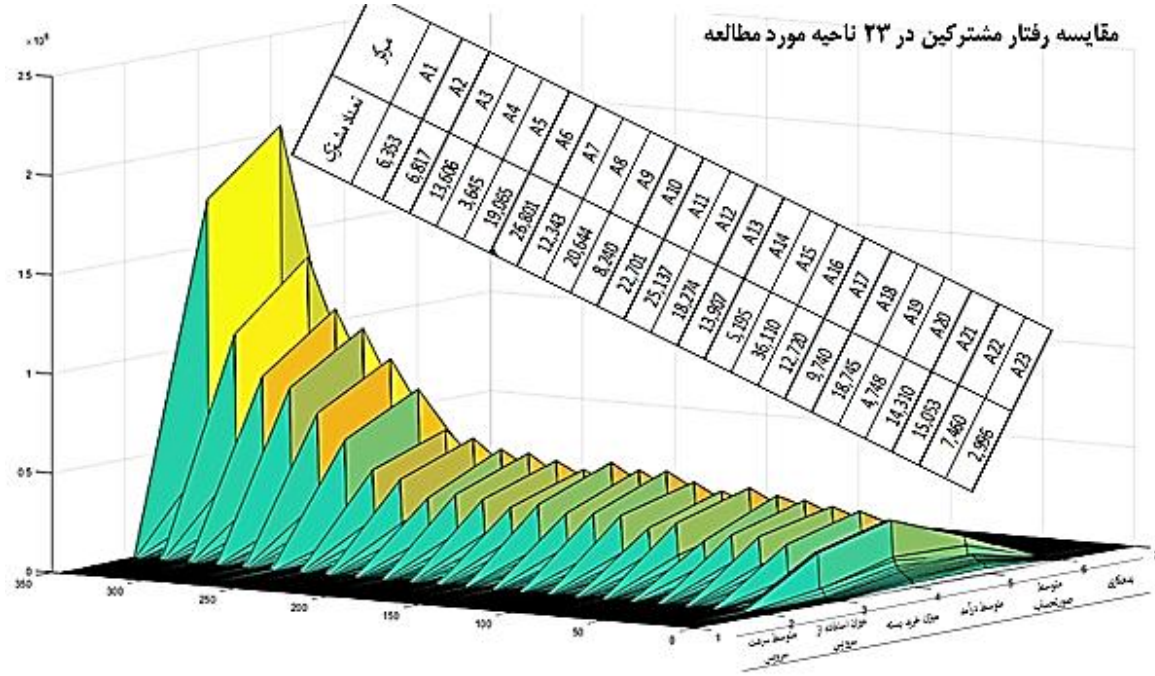
با محاسبه سرانه درآمدزایی هر یک از مشترکین در هریک از خوشه ها و تخصیص ضریب ۱ به سرانه درآمد خوشه ۱ و محاسبه نسبت سرانه سایر خوشه ها به خوشه ۱، مشخص می شود که رشد درآمد در خوشه های دوم و سوم نسبت به خوشه اول دارای رشد نمایی است به‌طوریکه یک مشترک در خوشه سوم حدود ۵.۳ برابر یک مشترک در خوشه اول و یک مشترک در خوشه دوم حدود ۲.۱ برابر یک مشترک در خوشه اول برای شرکت درآمد زایی دارد.



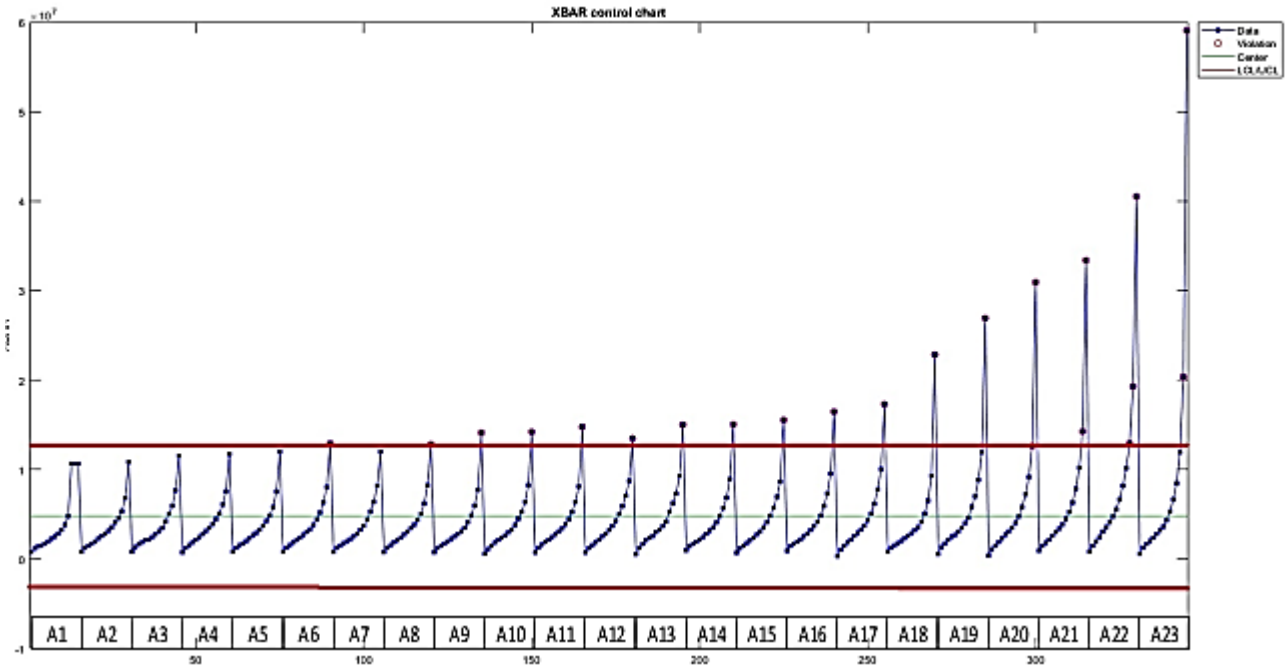
شکل ۱- نسبت تأثیر درآمدی یک مشترک در هر خوشه در مقایسه با خوشه ۱



شکل ۲- سیر تکاملی هر مشترک در خوشه ۱ تا ۳ (هر خوشه به ۵ کران تقسیم شده است)



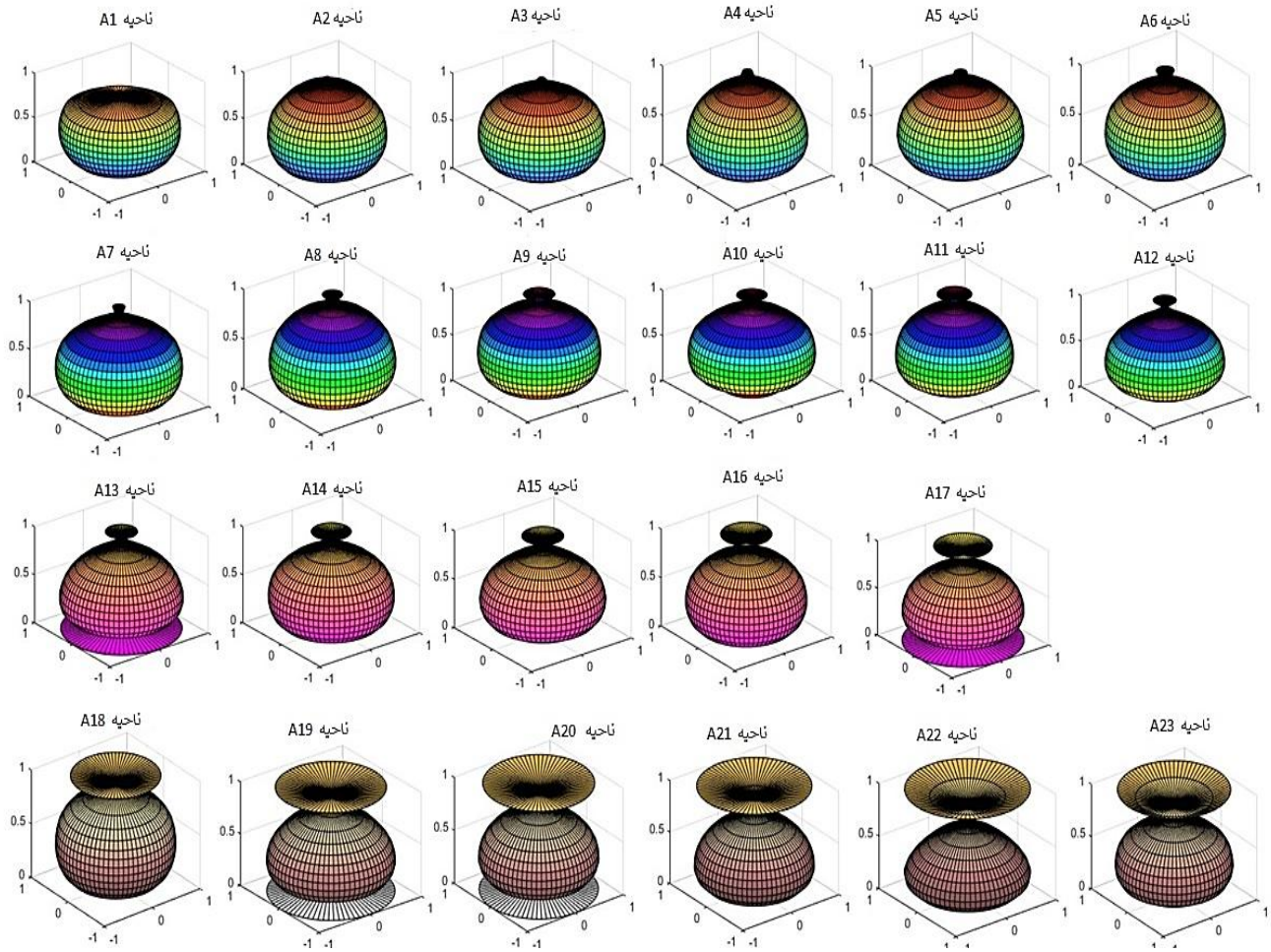
شکل۳- نمودار سه بعدی مقایسه رفتار مشترکین در ۲۳ ناحیه مورد مطالعه



شکل۴- شناسایی ناحیه هایی که مشترکین خاص در آن قرار گرفته اند

در شکل ۳ رشد ویژگی های درآمدی و مصرف خمات تمام مشترکین ۲۳ ناحیه مورد مطالعه به تفکیک ۳ خوشه اصلی و هر خوشه ۵ کران بندی، نشان داده شده است. شکل ۴ با در نظر گرفتن ۱۵ کران بندی برای مشترکین هر یک از ناحیه های ۲۳ گانه، ناحیه هایی که بالاتر از خط بالایی قرار گرفته است، دارای مشترکین بسیار خاص و با درآمد زایی بسیار زیاد است که می بایست به منظور خدمات بهتر و توسعه سرویس ها در اولویت قرار گیرند.

در ادامه در شکل ۵، مشترکین هر یک از ناحیه ها را به کمک تلفیقی از توابع مثلثاتی، در مقیاس واحد [۱و۱-] بصورت اشکال هندسی با طیف های رنگی ترسیم شده اند. به کمک تصاویر بدست آمده می توان نواحی که مشترکین آن دارای رفتار مشابه (از منظر درآمد زایی برای شرکت و میزان مصرف خدمات) هستند را بر اساس تشابه اشکال هندسی بدست آمده، تفکیک کرد و برای هر ناحیه A_1 تا A_{23} متناسب با شکل بدست آمده (که متأثر از بضاعت مشترکین، فرهنگ و... است)، برنامه ریزی مربوطه را انجام داد. صفحه های زیر اشکال هندسی در سه ناحیه A_{17}، A_{19} و A_{20} نشان دهنده مشترکینی است که به تازگی سرویس را خریده اند یا استفاده آنها از سرویس بسیار کم است.



شکل ۵- تبدیل مشترکین هر ناحیه به اشکال هندسی

شکل های ۱ تا ۵ به رده بندی مشتریان هر ناحیه و شناسایی ناحیه هایی که رفتار مشترکین در آن مشابه است کمک می کند. با مقایسه اشکال بدست آمده، ناحیه های (A۱-A۸) در رده پایین تر، ناحیه های (A۹-A۱۷) در رده متوسط و (A۱۸-A۲۳) در قوی ترین رده قرار دارند لذا می توان ناحیه های جغرافیایی رده های بالاتر را در اولویت برنامه های توسعه قرار داد. همچنین در هر ناحیه مشتریان خوشه سوم با بیشترین درآمدزایی برای شرکت می بایست با ارائه بسته‌های تشویقی و سرویس های خاص در نگهداشت و وفاداری آنها تأثیر گذاشت. مشتریان خوشه دوم با تدوین برنامه های خاص به خوشه سوم سوق داده شوند تا از تنزل آن به خوشه دوم جلوگیری شوند. بسیاری از مشترکین خوشه اول به تازگی سرویس را خریداری کرده اند و مجدد در شش ماه آینده می بایست

مورد بررسی قرار گیرند تا به خوشه های دوم و سوم هدایت شوند. و در مواردی که سرویس در خوشه اول بلااستفاده مانده پس از ۶ ماه می بایست مورد بررسی قرار گیرد.

۵-نتایج

در این پژوهش تابع زیان مینکوفسکی با $p=۱$ به‌عنوان معیار عدم‌تشابه به‌جای سایر معیارهای رایج از قبیل مربع اقلیدسی در خوشه‌بندی k-میانگین هوشمند وزنی، به‌کار برده شده است. با در نظر گرفتن $p=۱$، مراکز در هر خوشه به جای میانگین میانه مشاهدات است لذا داده های پرت تأثیر کمتری بر مرکز داده ها دارند. در اینجا با به کارگیری الگوریتم k-میانگین هوشمند وزنی مینکوفسکی، کارکرد آن بر شماری از پایگاه داده های شبیه‌سازی شده که از پیش برچسب‌گذاری شده‌اند، به یاری برخی معیارهای درونی و بیرونی همچون AM، NVI سنجیده شده است. در ادامه به منظور استفاده از کاربرد الگوریتم در تحلیل مشتریان یک شرکت ارائه دهنده خدمات مخابراتی، مشتریان در ۳ خوشه اصلی دسته بندی می شوند و بهینه بودنن نتایج به کمک ۱۰ معیار ارزیابی سنجیده می شود. از آنجاکه انتخاب مراکز اولیه به تصادف نیست در نتیجه سرعت پردازش اطلاعات (به دلیل نیاز نداشتن به تکرار الگوریتم به منظور حذف اثرات تصادفی مراکز تصادفی) افزایش می‌یابد و تنها یک باز اجرای الگوریتم کافی است. پس از تفسیر هر یک از خوشه ها به کمک شاخص های آماری، نمودارها و اشکال هندسی مدیران سازمان برای توسعه خدمات و نگهداشت مشتریان ارزشمند، بسته های تشویقی و خاص برای هر یک از مشتریان در هر یک از خوشه ها در نظر گرفته اند همچنین با برنامه ریزی به منظور تقویت مشتریان خوشه ۱ جهت ترغیب آنها به استفاده از خدمات شرکت، از ریزش آنها جلوگیری بعمل می آید.

۶-منابع

.۱ Ahmadzadehgoli, N., Mohammadpour, A., Behzadi M.H. (۲۰۱۷). LINEX k-means: Clustering by an Asymmetric Dissimilarity Measure, Journal of Statistical Theory and Applications, volume ۱۷, issue ۱, pages ۲۹-۳۸.

.۲ Davies D., and Bouldin D. (۱۹۷۹). A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, ۱(۲): ۲۲`۴–.۲۲۷

.۳ Duda, R. O., Hart, P. E., and Stork, D. G. (۲۰۰۱). Pattern Classification. John Wiley and Sons, Inc. New York, ۲nd Edition.

.۴ Harris T. J. (۱۹۹۲). Optimal Controllers for Nonsymmetric and Nonquadratic Loss Functions, Technometrics ۳۴, ۲۹۸– .۳۰۶

.۵ Hartigan, J. (۱۹۷۵). Clustering Algorithms. Toronto: JohnWiley & Sons.

.۶ Jiang, D., Tang, C., and Zhang, A. (۲۰۰۴). Cluster analysis for Gene Expression Data: A survey. IEEE Transactions on Knowledge and Data Engineering, ۱۶(۱۱):۱۳۷۰–.۱۳۸۶

.۷ Kummamuru K., Krishnapuram R. and Agrawal R. (۲۰۰۵). On Learning Asymmetric Dissimilarity Measures, ICDM ' ۰۵Proceedings of the Fifth IEEE International Conference on Data Mining, .۶۹۷-۷۰۰

.۸ Macqueen, J. (۱۹۶۷). Some methods for classification and analysis of multivariate observations. In Proceedings of the ۵th Berkeley symposium on mathematical statistics and probability, volume ۱, pages ۲۸۱–۲۹۷. Berkeley, CA: University of California Press .

.۹

.۱۰ Mirkin, B. (۲۰۰۵). Clustering for Data Mining: A Data Recovery Approach. Computer Science and Data Analysis Series. Boca Raton, FL: Chapman & Hall/CRC.

.۱۱ Modha D. S. and Spangler W. S. (۲۰۰۳). Feature Weighting in k-means Clustering, Machine Learning ۵۲, .۲۱۷-۲۳۷

.۱۲ Parsian A. and Kirmani S. N. U. A. (۲۰۰۲). Estimation under LINEX Loss Function, Handbook of Applied Econometrics and Statistical Inference ۱۶۵, .۵۳-۷۶

.۱۳ Reichart R. and Rappoport A. (۲۰۰۹). The NVI Clustering Evaluation Measure. In Proceedings of the Thirteenth Conference on Computational Natural Language Learning, .۱۶۵-۱۷۳

.۱۴ Seidpishe, M., Mohammadpour, A. (۲۰۱۷). Hierarchical Clustering of Heavy-Tailed Data Using a New Similarity Measure .

.۱۵ Varian H. R., (۱۹۷۵). A Bayesian Approach to Real Estate Assessment. Studies in Bayesian Econometrics and Statistics in Honor of Leonard J. Savage (Eds S. E. Fienberg and A. Zellner), ۱۹۵-۲۰۸