دسترسی در سایتِ http://jnrm.srbiau.ac.ir
سال چهارم، شماره چهاردهم، تابستان ۱۳۹۷
شماره شاپا: ۲۵۸۸-۵۸۸X

دانشگاه آزاد اسلامی، واحد علوم و تحقیقات

JNRM
پژوهش‌های نوین در ریاضی

خوشه‌بندی با الگوریتم k- میانگین لاینکس هوشمند

نرگس احمدزاده‌گلی[۱]، محمدحسن بهزادی[۲*] ، عادل محمدپور[۳]

[۱و۲] گروه آمار، دانشگاه آزاد اسلامی واحد علوم و تحقیقات، تهران، ایران (نویسنده مسئول).
[۳] گروه آمار، دانشکده ریاضی و علوم کامپیوتر، دانشگاه صنعتی امیرکبیر، تهران، ایران.

تاریخ دریافت مقاله: ۹۶/۰۹/۲۸ تاریخ پذیرش مقاله: ۹۷/۰۴/۱۳

چکیده

خوشه‌بندی k- میانگین لاینکس هوشمند یک تعمیم از خوشه‌بندی k- میانگین است که در آن تعداد خوشه‌ها و مراکز مربوطه را می‌توان مشخص کرد در حالی که تابع زیان لاینکس به‌عنوان معیار عدم‌تشابه در نظر گرفته می‌شود. بنابراین انتخاب مراکز در هر خوشه تصادفی نیست. انتخاب معیار عدم‌تشابه لاینکس به پژوهش‌گر کمک می‌کند تا مراکز را در صورت نیاز بیش برآورد یا کم برآورد نماید که سبب می‌شود برخی مشاهدات به خوشه‌ای خاص هدایت شوند. در این پژوهش، کارکرد الگوریتم یاد شده بر برخی پایگاه داده‌های واقعی و شبیه‌سازی شده بررسی می‌شود و نتایج با توجه به برخی معیارهای درونی و بیرونی ارزیابی می‌شود.

واژه‌های کلیدی: معیار عدم‌تشابه، k- میانگین هوشمند، خوشه‌بندی، تابع زیان لاینکس.

*عهده‌دار مکاتبات: Email: Behzadi@srbiau.ac.ir

۱- مقدمه

در طول ۵۰ سال گذشته، فنون تحلیل خوشه‌ای در دامنه وسیعی گسترش یافته است. هارتیگان خوشه‌بندی داده‌ها را به‌عنوان گروه‌بندی مشاهدات مشابه تعریف کرد[۵]. الگوریتم‌های k- میانگین یکی از محبوب‌ترین و مورد استفاده‌ترین الگوریتم‌های خوشه‌بندی می‌باشند و برای نخستین بار توسط مک کویین معرفی شد[۷]. این الگوریتم‌ها، بدین منظور طراحی شده‌اند تا داده‌های عددی را خوشه‌بندی کنند به طوری 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 ممکن است به معیار فاصله‌ای که به‌کار برده می‌شود بستگی داشته باشد. رایج‌ترین معیار به‌کار رفته در الگوریتم خوشه‌بندی، فاصله اقلیدسی است که پیش‌تر تعریف شد. هارتیگان[۵] الگوریتم مرسوم k- میانگین را به‌صورت زیر تعریف کرد: فرض کنید X یک مجموعه داده با n مشاهده باشد، به عبارتی،

$\underline{X} = (X_1. \ldots . X_n)'$، همچنین هر یک از مشاهدات دارای m ویژگی باشد، یعنی،

$X_i = (X_{i1}. \ldots . X_{im})$ k- میانگین الگوریتمی است که پایگاه داده $\underline{X}$ را به K خوشه مجزای $S = \{S_1. \ldots . S_K\}$ از مشاهدات مشابه تقسیم می‌کند به طوری‌که 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 X_i در خوشه k و مرکز مربوط به آن یعنی $\boldsymbol{C}_k$ به‌صورت زیر است: (۱-۱)

$$J_{\text{میانگین}-k} = \sum_{k=1}^{K} \sum_{X_i \in S_k} L(X_i. C_k)$$

که در آن، $C_k = (C_{k1}. \ldots . C_{km})$ برای $k = 1. \ldots . K$ و L فاصله اقلیدسی بین $X_i \in S_k$ و $\boldsymbol{C}_k$ می‌باشد. هر مشاهده به نزدیکترین مرکز C_k یعنی میانگین آن تخصیص داده می‌شود و تابع زیان $J_{\text{میانگین}-k}$ محاسبه می‌شود. مراکز خوشه‌ها به میانگین S_k به روزرسانی می‌شود تا زمانی‌که مراکز جدید از مرکز مرحله قبل تغییر نکند. برای یک مقدار اولیه K، با قرار دادن داده‌های بجای 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، به عبارتی اعضای خوشه‌ها ثابت بمانند، ادامه می‌یابد. رابطه (۱-۱) را می‌توان به‌صورت زیر بازنویسی کرد. (۱-۲)

$$G_{\text{میانگین}-k}(H. C) = \sum_{k=1}^{K} \sum_{i=1}^{n} h_{ik} L(X_i. C_k)$$

که $C = (C_1. \ldots . C_K)'$ و H یک ماتریس $n \times K$ است به طوری‌که برای هر $i = 1. \ldots . n$ ، $h_{ik} \in \{0.1\}$ و $\sum_{k=1}^{K} h_{ik} = 1$ است. اجرای الگوریتم به پیدا کردن تعداد خوشه‌ها و مراکز اولیه 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. انتخاب خودکار تعداد خوشه‌ها، یکی از ویژگی‌های مطلوب الگوریتم‌های خوشه‌بندی است. الگوریتم k- میانگین هوشمند که توسط میرکین[۸] معرفی شد، به منظور 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 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 آن که همه مشاهدات خوشه‌بندی شدند، خوشه‌های کوچک با استفاده از یک مقدار آستانه‌ای از پیش تعیین شده، حذف می‌شوند. این روش به علت سادگی، حتی برای افرادی که پیش زمینه‌ای از آمار و علوم کامپیوتر ندارند، نیز خوشایند است و نیازی به تکمیل چندین بار الگوریتم برای یافتن بهترین تعداد خوشه را ندارد و در واقع یک الگوریتم قطعی است که تنها به یک بار اجرا نیاز دارد.

دودا و همکاران[۳]، یک روش متفاوت ولی رایج از انتخاب مقدار K، ارائه کرده است که در آن، آزمایش‌ها با مقادیر متفاوت برای $k = (1. \ldots . K)$ اجرا می‌شود و نتایج آن تحلیل می‌گردد، اما این روش مناسب نیست زیرا ممکن است لازم باشد تعداد آزمایش‌های بسیار زیادی انجام شود.

الگوریتم خوشه‌بندی k- میانگین محدب توسط مودها و اسپنگلر[۹] پیشنهاد شد که در آن تابع زیانی که به عنوان معیار عدم‌تشابه به کار می‌رود، می‌بایست نامنفی، متقارن و محدب باشد. آنها k- میانگین را به الگوریتم k- میانگین محدب تعمیم دادند. زمانی که خطاهای مثبت و منفی دارای ارزش

یکسان می‌باشند توابع زیان متقارن جهت سنجش عدم‌تشابه به کار می‌رود ولی در بسیاری از حالات به معیاری نامتقارن نیاز است. کامامارا و همکاران[۶] کلاسی از معیارهای عدم‌تشابه نامتقارن را معرفی کرد.

یک تابع زیان با توجه به ساختار پایگاه داده انتخاب می شود. پارسیان و کرمانی[۱۰] بیان داشتند که زمانی که بیش‌برآوردی و کم‌برآوردی دارای ارزش یکسانی نمی باشند، تابع زیان متقارن مناسبی نیست. هریس و واریان[۴ و ۱۱] نمونه‌هایی در این باره ارائه کردند. از جمله آن که، بیش پر نمودن کانتینرها در صنایع فرآوری خوراک، ناخوشایندتر از کمتر پر کردن آنها است یا در زمان احداث سد، کم‌برآوردی سطح آب از بیش‌برآوردی آن اهمیت بیشتری دارد. تابع زیان لاینکس تابعی نامتقارن و محدب است و در یک سمت صفر، تقریباً نمایی و در سمت دیگرخطی می‌باشد. احمدزاده و همکاران الگوریتم خوشه‌بندی k- میانگین لاینکس بر پایه تابع زیان نامتقارن لاینکس را ارائه کردند[۱]. آنها نشان دادند که زمانی‌که بیش‌برآوردی یا کم‌برآوردی مراکز خوشه‌ها دارای اهمیت باشد و بتوان به یاری آن مشاهداتی را به خوشه‌ای خاص هدایت کرد، این الگوریتم در مقایسه با الگوریتم k- میانگین دارای دقت بالاتری است. حال در این پژوهش انگیزه معرفی الگوریتم k- میانگین هوشمند بر پایه تابع زیان لاینکس است. در بخش۲، بر تابع زیان لاینکس و الگوریتم k- میانگین لاینکس مرور می‌شود. در بخش۳ الگوریتم k- میانگین لاینکس هوشمند معرفی می‌گردد و سپس در بخش۴ چند پایگاه داده معتبر که در آن خوشه‌ها از پیش تعیین شده و مشاهدات آنها دارای برچسب می باشند را انتخاب کرده و الگوریتم پیشنهادی بر آنها اجرا می‌شود. همچنین به‌کمک این الگوریتم، تعدادی پایگاه داده که به کمک داده‌های شبیه‌سازی‌شده از چند توزیع آماری تولید می‌شود، خوشه‌بندی می‌گردد. به کارگیری مشاهدات برچسب‌دار سبب می‌شود تا دقت الگوریتم به کمک برخی معیارهای بیرونی سنجش اعتبار خوشه‌بندی، سنجیده شود. در پایان نتایج در بخش۵ جمع‌بندی می‌شود.

۲- الگوریتم k- میانگین لاینکس

فرض کنید، $\delta_j(X)$ یک برآوردگر پارامتر θ_j و $\Delta_j=\delta_j(X)-\theta_j$ برای $j=1.\dots.m$ باشد. در این صورت تابع زیان لاینکس چندمتغیره به‌صورت زیر است،

$$L(\Delta)=\sum_{j=1}^{m}\{\exp(a_j\Delta_j)-a_j\Delta_j-1\},$$

که در آن $a_j\neq 0$ و $\Delta=(\Delta_1.\dots.\Delta_m)$. اگر $a>0$، تابع زیان لاینکس، برای مقادیر مثبت Δ نمایی و برای مقادیر منفی Δ و به ازای $a<0$، خطی است. زمانی‌که $a=1$ باشد، $L(\Delta)$ کاملا نامتقارن بوده و بیش‌برآوردی از کم‌برآوردی پرهزینه‌تر خواهد بود. برای مقادیر کوچک و نزدیک به صفر $|a|$ تابع زیان تقریباً متقارن و به تابع زیان مربع اقلیدسی نزدیک می‌باشد. از این‌رو برآورد بهینه‌ای که توسط تابع زیان مربع اقلیدسی و لاینکس با $|a|$ نزدیک به صفر به دست می‌آید، دارای تفاوت زیادی نمی‌باشد. اما برای مقادیر بیشتر $|a|$، نقاط بهینه کاملا متفاوت است.

اکنون، تابع زیان لاینکس، به‌عنوان معیار عدم‌تشابه در خوشه‌بندی k- میانگین به کار می‌رود، زمانی که کم‌برآوردی و بیش برآوردی، دارای ارزش یکسان نباشند. از این‌رو الگوریتم k- میانگین لاینکس هوشمند به صورت زیر است.

نخست مقدار K مشخص می‌شود و مراکز اولیه $C_1.\dots.C_K$ به صورت آزمایشی در نظر گرفته می‌شود. معمولاً مراکز اولیه به‌تصادف از میان مشاهدات انتخاب می‌شود. تعیین مقدار K، یک مسئله سخت است که ممکن است از اطلاعات پژوهش‌گر یا هر اطلاعاتی پیرامون داده‌ها یا تجربیات به دست آید. همه مشاهدات به نزدیک‌ترین مرکز مربوطه با استفاده از تابع زیان لاینکس به‌عنوان معیار عدم‌تشابه، اختصاص می‌یابند. حال مسئله بهینه‌سازی زیر را در نظر بگیرید، $J_{\text{لاینکس}}(H.C)=\sum_{k=1}^{K}\sum_{i=1}^{n}h_{ik}L_{\text{لاینکس}}(X_i-C_k)$، که در آن $C=(C_1.\dots.C_K)'$ و $C_k=(C_{k1}.\dots.C_{km})$ برای $k=1.\dots.K$ و $L_{\text{لاینکس}}$ تابع زیان لاینکس در رابطه(۱-۲) است. H یک ماتریس $n\times K$ است به گونه‌ای که برای هر $i=1.\dots.n$،

- $h_{ik}\in\{0.1\}$
- $\sum_{k=1}^{K}h_{ik}=1$.

اکنون تابع $J_{\text{LINEX}}(H.C)$ با توجه به شرایط ۱ و ۲، به‌صورت زیر مینیمم می‌شود. دو گام زیر را در نظر بگیرید:

گام ۱: $C = \hat{C}$ ثابت در نظر گرفته می‌شود و معادله $J_{\text{LINEX}}(H.\hat{C})$ حل می‌شود، این معادله مینیمم می‌شود اگر و تنها اگر:

$$h_{ik} = \begin{cases} 1. & L_{\text{لاینکس}}(\boldsymbol{X}_i - \boldsymbol{C}_k) < L_{\text{لاینکس}}(\boldsymbol{X}_i - \boldsymbol{C}_t) \text{ for } 1 \le t \le K. \\ 0. & \text{سایر} . \end{cases}$$

گام ۲: $H = \hat{H},$ ثابت در نظر گرفته می‌شود، آنگاه $J_{\text{LINEX}}(\hat{H}.C)$ مینیمم می‌شود، اگر و تنها اگر برای هر عضو (۲)

$$C_{kj} = \frac{1}{a}\log\frac{\sum_{i=1}^{n} h_{ik}e^{aX_{ij}}}{\sum_{i=1}^{n} h_{ik}}, \quad j = 1. \ldots . m.$$

برای اثبات آن کافی است جمع داخلی زیر را برای k ثابت، مینیمم کرد.

$$\sum_{i=1}^{n} h_{ik}\left(e^{a(X_{ij}-C_{kj})} - a(X_{ij} - C_{kj}) - 1\right).$$

بدین منظور، نسبت به C_{kj} مشتق گرفته می‌شود و با قرار دادن آن برابر با تهی، بازده به دست می‌آید. الگوریتم k- میانگین بر پایه پردازه زیان لاینکس (به جای تابع زیان مربع اقلیدسی) که k- میانگین لاینکس نامیده می‌شود، مشابه الگوریتم k- میانگین بوده که تنها تفاوت آن در معیار عدم‌تشابه و مراکز خوشه‌ها است. مرکز هر خوشه به (۲-۱) به روز می‌شود و دوباره هر مشاهده به نزدیک‌ترین مرکز به روز شده، اختصاص می‌یابد. اگر هیچ تغییری در اعضای خوشه‌ها ایجاد نشود، خوشه‌بندی پایان می‌یابد و افرازهای به وجود آمده $S_1. \ldots . S_K$ نهایی می‌شوند.

۳-الگوریتم k- میانگین لاینکس هوشمند

الگوریتم k- میانگین لاینکس هوشمند تلفیقی از دو الگوریتم k- میانگین لاینکس و k- میانگین هوشمند است که در حقیقت با به کار گرفتن معیار عدم‌تشابه لاینکس در الگوریتم k- میانگین هوشمند به دست می‌آید. در این الگوریتم تعداد خوشه‌ها و مراکز دقیق آنها مشخص می‌شود. در نتیجه انتخاب مراکز خوشه‌ها به‌صورت تصادفی نیست. انتخاب معیار عدم‌تشابه لاینکس به پژوهش‌گر در بیش‌برآوردی یا کم‌برآوردی مراکز در صورت لزوم، کمک می‌کند. در این الگوریتم ابتدا مشاهدات خوشه‌بندی نشده، در دورترین نقطه از مرکز ثقل اولیه (که با مینیمم کردن تابع زیان لاینکس به دست می‌آید) قرار می‌گیرد و دورترین نقطه به‌عنوان یک مرکز آزمایشی در نظر گرفته می‌شود. سپس خوشه با مشاهداتی که به مرکز آزمایش نزدیک‌تر می‌باشند تکمیل می‌شود. نزدیک‌تر بودن مشاهدات به مراکز آزمایشی با تابع زیان لاینکس تشخیص داده می‌شود. همچنین یک مقدار آستانه‌ای که از پیش تعیین شده است کمک می‌کند تا پس از خوشه‌بندی همه مشاهدات، خوشه‌های کوچک‌تر از آن مقدار آستانه‌ای، حذف شوند. در این روش به تکرار چندین بار الگوریتم برای یافتن بهترین تعداد خوشه نیازی نیست و در واقع یک الگوریتم قطعی است که تنها به یک بار اجرا نیاز دارد، زیرا مراکز اولیه دیگر تصادفی نمی‌باشند. مراحل اجرای الگوریتم k- میانگین لاینکس هوشمند را می‌توان به صورت زیر بیان کرد:

۱- در این مرحله یک مقدار آستانه تعریف می‌شود تا همه خوشه‌هایی که اندازه آن کمتر از مقدار آستانه‌ای است، حذف شوند.

۲- گرانیگاه $\boldsymbol{C'} = (C'_1. \ldots . C'_m)'$ تعریف می‌شود که در آن m، بعد داده‌ها است. از آنجا که معیار عدم‌تشابه لاینکس است از این‌رو گرانیگاه، میانگین مجموعه داده‌ها نیست و در واقع با مینیمم کردن تابع زیر مشابه آنچه در بخش پیشین عنوان شد، به دست می‌آید

$$\sum_{k=1}^{K}\sum_{i=1}^{n} h_{ik}(\exp(a(X_{id} - C_{kj})) - . \\ a(X_{ij} - C_{kj}) - 1)$$

موقعیت این مرکز در هر مرحله از الگوریتم تغییر نمی‌کند. اعضایی که دورترین فاصله را تا گرانیگاه کل مجموعه داده‌ها دارند، به‌عنوان مراکز آزمایشی خوشه‌ها قرار می‌گیرند. خوشه S به وجود می‌آید که از اعضایی که به مرکز آزمایشی نسبت به گرانیگاه، نزدیک‌ترند، ساخته می‌شود.

۳- مرکز آزمایشی به گرانیگاه خوشه S یعنی

$$C_{kj} = \frac{1}{a}\text{Log}\frac{\sum_{i=1}^{n} h_{ik}e^{aX_{ij}}}{\sum_{i=1}^{n} h_{ik}}, \\ j = 1. \ldots . m$$

به‌روزرسانی می‌شود. توجه کنید که در گام دوم، گرانیگاه، از همه داده‌ها با مینیمم کردن معیار عدم‌تشابه لاینکس به دست می‌آید ولی در مرحله سوم، مرکز خوشه S با توجه به رابطه یادشده محاسبه می‌شود.

۴- یک خوشه S، که دربرگیرنده مشاهداتی است که فاصله آن تا C' کمتر است تا C تولید می‌شود. مشاهده X_i به خوشه S تعلق می‌گیرد اگر و تنها اگر $L(X_i. C) \leq L(X_i. C')$ باشد.

۵- اگر مرکز جدید با مرکز قبلی متفاوت باشد این فرآیند از ابتدای مرحله ۳ تکرار می‌شود، در غیر این‌صورت متوقف شده و خوشه S از پایگاه داده‌ها حذف می‌شود.

۶- تا زمانی که همه مشاهدات خوشه‌بندی شوند، فرایند ادامه می‌یابد.

۷- همه خوشه‌هایی که از مقدار آستانه حذف خوشه کوچک‌ترند حذف می‌شوند.

۸- در پایان با توجه به مراکز به دست آمده، الگوریتم k- میانگین لاینکس اجرا شده و خوشه‌هایی که در مرحله پیشین حذف شده‌اند دوباره خوشه‌بندی می‌شوند.

٤- ارزیابی الگوریتم‌ها بر پایگاه داده‌ها

٤-١- پایگاه داده‌های واقعی

مجموعه داده‌هایی طبیعی که مورد بررسی قرار می‌گیرند در پایگاه UCI در دسترس می‌باشند[۲]. همه این داده‌ها دارای برچسب می‌باشند، به این معنی که خوشه‌های آنها از پیش مشخص شده است. بنابراین می‌توان به کمک آنها الگوریتم‌ها را به‌راحتی ارزیابی کرد. استاندارد کردن داده‌ها قبل از اجرای الگوریتم دقت را افزایش می‌دهد. در ادامه الگوریتم‌های k- میانگین، k- میانگین لاینکس (هر یک ۵۰۰ بار) و k- میانگین لاینکس هوشمند بر این پایگاه داده‌ها اجرا می‌شود و متوسط دقت (AM) و سایر معیارهای ارزیابی سنجیده می‌شود.

۱- پایگاه داده آیریس که مجموعه‌ای از ۱۵۰ نمونه گل است به گونه‌ای که هر یک دارای ۴ ویژگی است و در سه خوشه، طبقه‌بندی می‌شود و در هر خوشه ۵۰ مشاهده قرار می‌گیرد.

۲- پایگاه داده‌ی برجستگی لرزه‌ای دربرگیرنده ۲۵۸۴ داده با ۱۸ ویژگی عددی و کیفی است که مسئله‌ی پیش‌بینی انرژی بالا (بالاتر از^{۴}۱۰ ژول) را در یک معدن زغال توصیف می‌نماید. این داده‌ها در دو دسته خوشه‌بندی می‌شوند "برجستگی لرزه‌ای با انرژی بالا" و "برجستگی لرزه‌ای با انرژی کم". اگر یک برجستگی لرزه‌ای با انرژی بالا در خوشه‌ی انرژی پایین خوشه‌بندی شود، می‌تواند بسیار خطرناک باشد.

۳- بقاء هابرمن، پایگاه داده‌ای با ۳۰۶ داده و سه ویژگی عددی از یک مطالعه بر بیماران زنده مانده از سرطان سینه است که عمل جراحی را تحمل کرده‌اند. این داده‌ها به دو دسته بخش می‌شوند، "بیمارانی که پنج سال یا بیشتر زنده مانده‌اند" و "بیمارانی که کمتر از پنج سال عمر کرده‌اند".

۴- تلسکوپ گاما در برگیرنده ۱۹۰۲۰ داده و ۱۱ ویژگی است که ذرات گاما با انرژی بالا را با فن پردازش تصویر در یک زمین جوی با تلسکوپ گاما چرنکوف توصیف می‌کند. داده‌ها به دو خوشه تقسیم می‌شوند، "تکی" و "پس زمینه". خوشه‌بندی یک مشاهده "پس زمینه به‌عنوان "تکی" بدتر از خوشه‌بندی یک مشاهده "تکی" به‌عنوان "پس زمینه" است. از این‌رو الگوریتم k- میانگین مرسوم برای این پایگاه داده مناسب نیست.

٤-٢- پایگاه داده‌های شبیه‌سازی‌شده

تابع چگالی‌ها و روابط بین متغیرهای تصادفی انتخاب شده از آنها را در جدول ۱ در نظر بگیرید. اکنون پنج پایگاه داده از توزیع‌های یادشده، به طوری‌که هرکدام دارای ۱۰۰ متغیر وابسته با m ویژگی بوده، تولید کرده به گونه‌ای که در دو گروه، خوشه‌بندی شوند. فرآیند تولید داده‌ها در جدول ۲، نشان داده می‌شود.

در پایگاه داده‌های تولید شده، به ۵۰ متغیر نخست، برچسب ۱ و به ۵۰ متغیر دوم، برچسب ۲، تعلق می‌گیرد. حال الگوریتم‌های k- میانگین و k- میانگین لاینکس، k- میانگین هوشمند و k- میانگین لاینکس هوشمند را بر پایگاه داده‌ها ۵۰۰ بار اجرا کرده و متوسط دقت در این تکرارها سنجیده می‌شود.

٤-٣-ارزیابی الگوریتم

در این بخش، الگوریتم k- میانگین هوشمند لاینکس، با خوشه‌بندی پایگاه داده‌های واقعی و شبیه‌سازی شده، مورد ارزیابی قرار می‌گیرد و نتایج آن با الگوریتم‌های k- میانگین، k- میانگین لاینکس و k- میانگین هوشمند مقایسه می‌شود.

بدین‌منظور، از معیار بیرونی معیار تغییر اطلاعات نرمال شده (NVI) و معیار درونی دیویس – بالدین (DB) استفاده می‌شود. اگر NVI در فاصله $[0.1]$ قرار گیرد، بدان معناست که کارکرد الگوریتم مناسب است. هر چقدر مقدار آن کوچک‌تر باشد خوشه‌ها همگن‌ترند. همچنین مقادیر کمتر DB نشان می‌دهد که خوشه‌ها به خوبی جداسازی شده‌اند.

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

روابط میان متغیرهای تصادفی مستقل	$f(x)$	
ضرب n متغیر تصادفی مستقل و هم توزیع LN دارای توزیع LN است.	$\frac{1}{x\sigma\sqrt{2\pi}}\exp\left(-\frac{(\log(x)-\mu)}{2\sigma^2}\right),\ x>0, \sigma>0, \mu\epsilon R$	لاگ نرمال (LN)
جمع n متغیر تصادفی مستقل و هم توزیع نرمال، کوشی، گاما و پواسن به ترتیب دارای همان توزیع‌ها هستند.	$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),\ x,\mu\in R, \sigma>0$	**نرمال**
	$\frac{\beta^{\alpha}x^{\alpha-1}e^{-x\beta}}{\Gamma(\alpha)}, x\geq 0,\ \alpha,\beta>0$	گاما
	$\frac{e^{-\lambda}\lambda^{x}}{x!}, x=0,1,\ldots$	پواسن

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

	پایگاه داده	$Y=(Y_1.\ldots.Y_{100})'$
چند متغیره با ۳ ویژگی		$Y_i=(X_{i1}+X_0.X_{i2}+X_0.X_{i3}+X_0)=(Y_{ij})\quad \text{for } j=1.2.3$
	نرمال	$X_0.X_{ij}\overset{iid}{\sim}\text{Normal}(0.1),\ Y_{ij}\overset{id}{\sim}\text{Normal}(0.2) \text{ for } i=1.\ldots.50$ $X_{ij}\overset{iid}{\sim}\text{Normal}(4.1),\ Y_{ij}\overset{id}{\sim}\text{Normal}(4.2) \text{ for } i=51.\ldots.100$
	گاما	$X_0.X_{ij}\overset{iid}{\sim}\text{Gamma}(1.3),\ Y_{ij}\overset{id}{\sim}\text{Gamma}(2.3) \text{ for } i=1.\ldots.50$ $X_{ij}\overset{iid}{\sim}\text{Gamma}(6.3),\ Y_{ij}\overset{id}{\sim}\text{Gamma}(7.3) \text{ for } i=51.\ldots.100$
	لاگ نرمال	$Y_i=(X_{i1}X_0.X_{i2}X_0.X_{i3}X_0)=(Y_{ij})\quad \text{for } j=1.2.3$ $X_0.X_{ij}\overset{iid}{\sim}\text{LN}(1.1),\ Y_{ij}\overset{id}{\sim}\text{LN}(2.2) \text{ for } i=1.\ldots.50$ $X_{ij}\overset{iid}{\sim}\text{LN}(20.1),\ Y_{ij}\overset{id}{\sim}\text{LN}(21.2) \text{ for } i=51.\ldots.100$
تک متغیره		$Y=(Y_1.\cdots.Y_{100}),\ Y_i=X_i+X_0$
	پواسن	$X_0.X_i\overset{iid}{\sim}\text{Poisson}(1),\ Y_i\overset{id}{\sim}\text{Poisson}(2) \text{ for } i=1.\ldots.50$ $X_i\overset{iid}{\sim}\text{Poisson}(10),\ Y_i\overset{id}{\sim}\text{Poisson}(11) \text{ for } i=51.\ldots.100$

جدول۳ - نتایج الگوریتم‌های k- میانگین، k- میانگین لاینکس، k- میانگین هوشمند و k- میانگین لاینکس هوشمند، در پایگاه داده‌های شبیه‌سازی شده

پایگاه داده	الگوریتم	a	AM	DB	NVI
نرمال	k- میانگین	–	۸۲,۹	۰,۷۸	۰,۲۴
	k- میانگین لاینکس	$۱۰^{-۳}$	۸۷,۳	۰,۶۳	۰,۲۲
	k- میانگین هوشمند	-	۱۰۰,۰	۰,۵۷	۰
	k- میانگین لاینکس هوشمند	$۱۰^{-۳}$	۱۰۰,۰	۰,۵۷	۰
لاگ نرمال	k- میانگین	–	۰,۷۱	۰,۷۳	۰,۸۳
	k- میانگین لاینکس	$۱۰^{-۶}$	۷۳,۰	۰,۸۶	۰,۷۳
	k- میانگین هوشمند	-	۷۲,۰	۰,۸۱	۰,۸۳
	k- میانگین لاینکس هوشمند	$۱۰^{-۶}$	۷۹,۰	۰,۷۱	۰,۷۸
گاما	k- میانگین	–	۹۸,۰	۰,۶۲	۰,۲۱
	k- میانگین لاینکس	$۱۰^{-۶}$	۹۸,۰	۰,۶۲	۰,۲۱
	k- میانگین هوشمند	-	۹۸,۰	۰,۶۲	۰,۲۱
	k- میانگین لاینکس هوشمند	$۱۰^{-۶}$	۹۸,۰	۰,۶۲	۰,۲۱
پواسن	k- میانگین	–	۹۴,۰	۰,۴۴	۰,۵۲
	k- میانگین لاینکس	$۱۰^{-۶}$	۹۵,۸	۰,۴۵	۰,۴۵
	k- میانگین هوشمند	-	۹۴,۰	۰,۴۴	۰,۴۲
	k- میانگین لاینکس هوشمند	$۱۰^{-۶}$	۹۶,۰	۰,۴۵	۰,۴۴

در هر پایگاه داده بهترین نتایج در هر ستون پررنگ نشان داده‌شده است. از آنجا که در پایگاه داده‌های شبیه‌سازی شده در جدول ۳، بیش‌برآوردی و کم‌برآوردی دارای اهمیت نیست، لذا مقدار a در الگوریتم خوشه‌بندی k- میانگین لاینکس و لاینکس هوشمند، به صفر بسیار نزدیک باشد و این سبب می‌شود تا مراکز خوشه‌ها (که با مینیمم کردن تابع زیان لاینکس به‌دست آمده است) به مراکز حاصل از مینیمم کردن تابع زیان مربع اقلیدسی بسیار نزدیک شود. نتایج جدول، نشان می‌دهد که الگوریتم خوشه‌بندی k- میانگین لاینکس هوشمند در مواردی در مقایسه با سایر الگوریتم‌ها دارای دقت بیشتری است. در ادامه عملکرد الگوریتم‌ها را بر پایگاه داده‌های واقعی معرفی شده مقایسه می‌شود. در جدول۴، در پایگاه داده‌ی آیریس، بیش برآوردی و کم برآوردی دارای اهمیت نیست. ولی در دیگر پایگاه داده‌ها، خوشه‌بندی اشتباه یک مشاهده ممکن است دارای هزینه و آسیب زیاد باشد لذا بیش برآوردی یا کم برآوردی دارای اهمیت هستند. نتایج جدول۴، بیان‌گر آن است که الگوریتم k- میانگین لاینکس هوشمند در این مجموعه داده‌ها، در مقایسه با سایر الگوریتم‌ها ازجمله الگوریتم k- میانگین، عملکرد بهتری دارد، به‌ویژه زمانی که بیش‌برآوردی و کم‌برآوردی دارای اهمیت یکسان نباشند. الگوریتم‌های k- میانگین و k- میانگین هوشمند که بر پایه معیارهای عدم‌تشابه متقارن بنا شده، برای داده‌هایی مناسب می‌باشند که در آنها خطاهای مثبت و منفی دارای اهمیت یکسان باشد.

جدول ٤ - نتایج الگوریتم‌های k- میانگین، k- میانگین لاینکس، k- میانگین هوشمند و k- میانگین لاینکس هوشمند، در پایگاه داده‌های واقعی

پایگاه داده	الگوریتم	a	AM	D	NVI
بقاءهابرمن	k- میانگین	-	۵۱,۲	۰,۹۶	۰,۹۹
	k- میانگین لاینکس	۰,۹	۷۲,۸	۰,۹۹	۰,۸۷
	k- میانگین هوشمند	-	۶۶,۲	۰,۹۶	۰,۹۹
	k- میانگین لاینکس هوشمند	۰,۹	۷۳,۳	۰,۷۲	۱,۰۰
تلسکوپ گاما	k- میانگین	-	۵۸,۹	۱,۴۱	۰,۹۹
	k- میانگین لاینکس	۰,۱	۷۰,۷	۱,۷۴	۰,۹۵
	k- میانگین هوشمند	-	۶۴,۹	۱,۲۶	۰,۹۸
	k- میانگین لاینکس هوشمند	۰,۱	۷۱,۰	۱,۶۱	۰,۹۲
برجستگی لرزه‌ای	k- میانگین	-	۹۰,۸	۰,۴۲	۰,۹۸
	k- میانگین لاینکس	۰,۱	۹۳,۴	۰,۴۰	۱,۰
	k- میانگین هوشمند	-	۹۰,۹	۰,۴۲	۱,۰
	k- میانگین لاینکس هوشمند	۰,۱	۹۳,۴	۰,۴۰	۰,۹۸
آیریس	k- میانگین	-	۷۸,۳	۰,۶۳	۰,۵۲
	k- میانگین لاینکس	$۱۰^{-۳}$	۸۴,۸	۰,۵۳	۰,۴۹
	k- میانگین هوشمند	-	۸۰,۲	۰,۶۸	۰,۵۴
	k- میانگین لاینکس هوشمند	$۱۰^{-۳}$	۸۴,۰	۰,۵۴	۰,۵۱
هپاتیت	k- میانگین	-	۶۴,۹	۱,۶۲	۰,۸۵
	k- میانگین لاینکس	**۰,۱**	۷۵,۷	۰,۶۶	۰,۵۴
	k- میانگین هوشمند	-	۷۵,۳	۰,۷۱	۰,۵۹
	k- میانگین لاینکس هوشمند	۰,۱	۷۶,۰	۰,۶۶	۰,۵۴

اما معیار عدم‌تشابه لاینکس در هر دو حالت متقارن و نامتقارن کارآمد است. پارامتر a در پایگاه داده آیریس به دلیل عدم اهمیت بیش‌برآورد یا کم‌برآوردی، به صفر بسیار نزدیک انتخاب می‌شود و در سایر پایگاه داده‌ها، انتخاب a بدین صورت است که در بازه $[-1.1]$ در فواصل ۰,۱ مقداری از a که منجر به مقادیر کمتر معیارهای NVI و DB می‌شود و در عین حال باعث دقت بیشتری می‌شود، انتخاب می‌شود.

۵-نتیجه‌گیری

گاهی به یک معیار عدم‌تشابه غیرمتقارن نیاز است تا فاصله بین داده‌ها محاسبه شود. در این پژوهش تابع زیان لاینکس به‌عنوان معیار عدم‌تشابه به‌جای سایر معیارهای رایج از قبیل مربع اقلیدسی، منهتن و... در خوشه‌بندی k- میانگین هوشمند به‌کار برده شده است. بنابراین مراکز در هر خوشه میانگین یا میانه مشاهدات نیست. در اینجا با معرفی الگوریتم k- میانگین هوشمند لاینکس، کارکرد آن بر شماری از پایگاه داده شبیه‌سازی شده و واقعی که از پیش برچسب‌گذاری شده‌اند، به یاری برخی معیارهای درونی و بیرونی همچون AM، NVI و DB سنجیده شده است. تفاوت اصلی این الگوریتم با دیگر الگوریتم‌های خوشه‌بندی مبتنی بر مرکز در آن است که، زمانی که در خوشه‌بندی بیش‌برآوردی یا کم‌برآوردی مراکز دارای ارزش باشد، به عبارتی خوشه‌بندی نادرست مشاهدات دارای زیان و هزینه زیادی باشد، تا اندازه‌ای می‌تواند در دسته‌بندی درست مشاهدات بهتر عمل کند همچنین آن که 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. مقادیر معیارهای ارزیابی نیز می‌تواند به انتخاب صحیح پارامتر a کمک کند. از ویژگی مهم این الگوریتم آن است که زمانی که مقدار a به صفر نزدیک باشد، عملکرد آن همانند الگوریتم k- میانگین هوشمند رایج است. به‌طور کلی نتایج نشان دهنده آن است که عملکرد الگوریتم k- میانگین هوشمند لاینکس از k- میانگین و k- میانگین هوشمند در بیشتر حالات بهتر است. نتایج الگوریتم‌های k- میانگین لاینکس هوشمند به دلیل تصادفی نبودن مراکز اولیه، معادل چندین مرتبه اجرای الگوریتم‌های k- میانگین رایج است، اما نتایج بهتری ارائه می‌کند.

فهرست منابع

[1] Ahmadzadehgili. N, Mohammadpour. A, Behzadi,. M.H (2017). LINEX k-means: Clustering by an Asymmetric Dissimilarity Measure, To appear: Journal of Statistical Theory and Applications .

[2] Asuncion. A and Newman. D.J (2007). UCI Machine Learning Irvine, CA*:* University of California, School of Information and Computer Science.

[3] Duda. R.O, Hart. P.E, and Stork. D.G (2001). Pattern Classification. John Wiley and Sons, Inc. New York, 2nd Edition.

[4] Harris. T.J (1992). Optimal Controllers for Nonsymmetric and Nonquadratic Loss Functions, Technometrics.

[5] Hartigan. J (1975). Clustering Algorithms. Toronto: JohnWiley & Sons.

[6] Kummamuru. K, Krishnapuram. R and Agrawal. R (2005). On Learning Asymmetric Dissimilarity Measures, ICDM '05 Proceedings of the Fifth IEEE International Conference on Data Mining.

[7] Macqueen. J (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the 5th Berkeley symposium on mathematical statistics and probability.

[8] Mirkin. B (2005). Clustering for Data Mining: A Data Recovery Approach. Computer Science and Data Analysis Series. Boca Raton, FL: Chapman & Hall/CRC.

[9] Modha. D.S and Spangler. W.S (2003). Feature Weighting in k-means Clustering, Machine Learning.

[10] Parsian. A and Kirmani. S.N.U.A (2002). Estimation under LINEX Loss Function, Handbook of Applied Econometrics and Statistical Inference.

[11] Varian. H.R (1975). 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.