دوره دهم، شماره اول، بهار و تابستان ۱۳۹۶

مجله فناوری اطلاعات در طراحی مهندسی
Information Technology in Engineering Design
http://ited.sinaweb.net

کاهش بعد داده‌ها ضمن حفظ خوشه‌های داده

محبوبه حقایقی‌پور[۱] یحیی فرقانی*[۲] سیدمحمدحسین معطر[۳]

(۱) گروه مهندسی کامپیوتر، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران

(۲) گروه مهندسی کامپیوتر، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران*

(۳) گروه مهندسی کامپیوتر، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران

چکیده

در این مقاله، با توجه به موفقیت روش‌های خوشه‌بندی مبتنی بر k-means، یک روش کاهش ویژگی بر پایه k-means وزن دار ارائه می‌شود. در روش پیشنهادی، نخست با استفاده از روش k-means وزن دار، به ویژگی‌های داده‌ها وزن داده می‌شود. ویژگی‌های وزین تر ضرورتا ویژگی‌های مهمتر نیستند و وزن هر ویژگی، تنها بازه هر ویژگی را به نحوی تغییر می‌دهد که خوشه‌بندی بهتری صورت بگیرد. لذا، سپس با استفاده از یک مدل ریاضی جدید، کسری از ویژگی‌های وزندار شده داده‌های هر خوشه انتخاب می‌شود به نحوی که کمترین تغییر در خوشه ها حاصل شود. تعداد ویژگی‌های منتخب هر خوشه در روش پیشنهادی، برخلاف روش‌های مشابهی چون k-means تنک و fuzzy c-means تنک بصورت صریح تعیین می‌شود. درضمن، آزمایش‌های تجربی روی چهار مجموعه داده واقعی نشان می‌دهد که روش پیشنهادی، از دقت بیشتری نسبت به روش‌های L1PCA, LLE و روش K-means تُنُک برخوردار است.

واژه‌های کلیدی: کاهش ویژگی، خوشه بندی، K-means وزن دار

* عهده دار مکاتبات

نشانی: گروه مهندسی کامپیوتر، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران

تلفن: ۰۹۳۵۲۱۶۱۸۵۷ **پست الکترونیکی:** yforghani@mshdiau.ac.ir

۱- مقدمه

بطورکلی دو روش کاهش ویژگی وجود دارد: استخراج ویژگی و انتخاب ویژگی. در روش انتخاب ویژگی، کسری از ویژگی‌ها به عنوان ویژگی‌های موثر انتخاب می‌شوند و در روش استخراج ویژگی، تعداد اندکی ویژگی موثر تولید می‌شود که هر یک، تابعی از یک یا چند ویژگی از ویژگی‌های اولیه است.

از جمله روش‌های استخراج ویژگی می‌توان به روش PCA (Principle Component Analysis) [۱] اشاره کرد. در این روش، بهترین ویژگی استخراج شده، آن ویژگی است که بیشترین واریانس را داشته باشد و ویژگی با واریانس نزدیک به صفر، ویژگی بی اهمیت یا نویز محسوب می‌شود. واریانس داده، متناسب است با مجموع مربع اختلاف داده‌ها تا میانگین داده‌ها. در صورت وجود داده نویزی، مربع اختلاف داده نویزی تا میانگین داده‌ها، اثر زیادی بر روی واریانس داده‌ها و عملکرد PCA خواهد گذاشت. برای کاهش اثر داده‌های نویزی، در روش L1PCA [۲] از قدرمطلق اختلاف داده تا میانگین داده‌ها بجای مربع اختلاف داده تا میانگین داده‌ها استفاده می‌شود.

روش استخراج ویژگی Locally Linear Embeding (LLE) [۳] شامل دو مرحله است. در مرحله نخست، هر داده بصورت ترکیب خطی سایر داده ها نوشته می‌شود و ضرایب این ترکیب خطی ذخیره می‌گردد. سپس بردارهایی با بعد کمتر از داده های اولیه، به نحوی تولید می‌شوند که همچون داده‌های اولیه، هر کدام، یک ترکیب خطی با همان ضرایب ذخیره شده از سایر بردارهای با بعد کمتر باشد.

هدف از کاهش ویژگی، کاهش حجم داده‌ها، حذف ویژگی‌های نامرتبط و کشف ویژگی‌های موثر در تحلیل داده‌هاست. ویژگی‌های نامرتبط در داده‌ها حتی می‌تواند نتیجه تحلیل داده‌ها را منحرف سازد. به آن دسته از ویژگی‌ها، ویژگی‌های موثر گفته می‌شود که بتواند به خوبی ساختار داده‌ها مثل خوشه‌های موجود در داده‌ها را حفظ کند. لذا، حفظ ساختار داده‌ها یا خوشه‌های داده باید در زمان کاهش ویژگی مدنظر باشد. در روش‌های انتخاب ویژگی k-means تنک [٤] و fuzzy c-means تنک [٥]، کسری از ویژگی‌ها به نحوی انتخاب می‌شوند که فاصله درون خوشه ای داده‌ها نسبت به فاصله داده‌ها از کل داده‌ها کمتر شود و امکان خوشه‌بندی بهتر میسر گردد. متاسفانه، امکان تعیین صریح تعداد ویژگی‌های منتخب در این دو روش وجود ندارد. در این تحقیق، با توجه به موفقیت روش‌های خوشه‌بندی مبتنی بر k-means [٤, ٥]، یک روش انتخاب ویژگی بر پایه روش خوشه‌بندی k-means وزن دار [٥] ارائه می‌شود. به بیان دقیقتر، در روش پیشنهادی، نخست با استفاده از روش k-means وزن دار، به ویژگی‌های داده‌ها وزن داده می‌شود. ویژگی‌های وزین‌تر ضرورتا ویژگی‌های مهم‌تر نیستند و وزن هر ویژگی، تنها بازه هر ویژگی یا شکل خوشه‌ها را به نحوی تغییر می‌دهد که خوشه‌بندی بهتری صورت بگیرد. لذا، سپس با استفاده از یک مدل ریاضی جدید، کسری از ویژگی‌های وزندار شده هر خوشه انتخاب می‌شود به نحوی که کمترین تغییر در خوشه‌ها حاصل شود. تعداد ویژگی‌های منتخب هر خوشه در روش پیشنهادی، برخلاف روش‌های k-means تنک و fuzzy c-means تنک بصورت صریح تعیین می‌شود. درضمن، آزمایش‌های تجربی روی چهار مجموعه داده واقعی نشان می‌دهد که روش پیشنهادی، از دقت بیشتری نسبت به روش‌های L1PCA, LLE و روش K-means تُنُک برخوردار است.

در ادامه، در بخش ۲ به بیان پیش نیازهای تحقیق پرداخته می‌شود. در فصل ۳، روش پیشنهادی ارائه می‌شود. در فصل ٤ با استفاده از چهار مجموعه داده واقعی، روش پیشنهادی با سه روش کاهش ویژگی دیگر بصورت تجربی مقایسه می‌شود. در فصل ۵ به جمع بندی نتایج و کارهای آتی می‌پردازیم.

۲- پیش نیازها

۲-۱- روش خوشه بندی k-means

فرض کنید می‌خواهیم n داده $\{x_1, x_2, \dots, x_n\}$ را به c خوشه تقسیم کنیم که $x_i \in R^m$ و m بعد هر داده است. مدل خوشه بندی k-means به صورت زیر است:

(۱)

$$\min_{u,z} \sum_{k=1}^{c}\sum_{i=1}^{n} u_{ik}\|x_i - z_k\|_2^2$$
$$\text{subject to } \begin{cases} \sum_{k=1}^{c} u_{ik} = 1, \ \ i = 1,2,\dots,n; \\ u_{ik} \in \{0,1\}, \ \ i = 1,2,\dots,n; k = 1,2,\dots,c. \end{cases}$$

که z_k مرکز خوشه k–ام است و u_{ik} تعلق داده i–ام به خوشه k–ام را مشخص می‌کند و

$$\|x_i - z_k\|_2^2 = \sum_{j=1}^{m}(x_{ij} - z_{kj})^2$$

هدف این مدل، پیداکردن مراکز خوشه‌ها یعنی z_kها و u_{ik}ها به نحوی است که فاصله داده‌های هر خوشه تا مرکز آن خوشه، حداقل شود. اولین قید این مدل مشخص می‌کند که داده i–ام به یک خوشه تعلق دارد. مساله خوشه‌بندی، یک مساله NP-Complete است. الگوریتم تکراری خوشه بندی k-means برای به دست آوردن مقدار بهینه محلی مدل k-means بصورت الگوریتم ۱ است.

الگوریتم ۱:

۱. مراکز اولیه خوشه‌ها (θ) را به صورت تصادفی درنظربگیر.
۲. مراکز خوشه‌ها (θ) را ثابت فرض کرده و مقدار بهینه مدل k-means را نسبت به درجات عضویت (u) بدست آور.
۳. درجات عضویت (u) را ثابت فرض کرده و مقدار بهینه مدل k-means را نسبت به مراکز خوشه (θ) بدست آور.
٤. درصورت عدم حصول شرط همگرایی برو به ۲.

۲-۲ روش خوشه بندی k-means وزندار (wk-means)

مدل ریاضی خوشه بندی k-means وزن دار به صورت زیر است:

(۲)

$$\min_{u,z,w} \sum_{i=1}^{n}\sum_{k=1}^{c} u_{ik}\sum_{j=1}^{m} w_{kj}^{\beta}(x_{ij} - z_{kj})^2$$
Subject to
$$\begin{cases} \sum_{k=1}^{c} u_{ik} = 1, \ \ i = 1,2,\dots,n; \\ \sum_{j=1}^{m} w_{kj} = 1, \ \ k = 1,2,\dots,c; \\ u_{ik} \in \{0,1\}, \ \ i = 1,2,\dots,n; k = 1,2,\dots,c; \\ w_{kj} \geq 0, \ \ k = 1,2,\dots,c; j = 1,2,\dots,m. \end{cases}$$

در این مدل، برای ویژگی j–ام خوشه k–ام، وزن متفاوتی درنظر گرفته شده است که با w_{kj} نشان داده شده است و مقدار بهینه آن در طی فرآیند خوشه‌بندی، تعیین می‌شود. در این روش، به ویژگی‌هایی از یک خوشه، وزن بیشتری داده می‌شود که بازه کوچکتری در آن خوشه داشته باشند. لذا، در این مدل، سعی می‌شود با تغییر بازه ویژگی‌های هر خوشه یا تغییر شکل خوشه‌ها از طریق وزندار کردن ویژگی‌های هر خوشه، خوشه‌بندی بهتری انجام شود. پارامتر β نحوه اختصاص وزن به ویژگی‌ها را کنترل می‌کند. اگر $\beta = 1$ باشد، تنها یک ویژگی از هر خوشه، وزنی غیرصفر خواهد داشت. الگوریتم ۲ روش خوشه‌بندی wk-means را نشان می‌دهد.

الگوریتم۲

۱. مراکز اولیه خوشه‌ها (z) و وزن‌ها (w)را بصورت تصادفی در نظر بگیر.
۲. مراکز خوشه‌ها (z) و وزن‌ها (w) را ثابت فرض کرده و مقدار بهینه مدل wk-means را نسبت به درجات عضویت (u) با استفاده از رابطه زیر بدست آور:

$$u_{ik} = \begin{cases} 1 & if\ \forall l: \sum_{j=1}^{m} w_{kj}^{\beta}(x_{ij} - z_{kj})^2 \leq \sum_{j=1}^{m} w_{lj}^{\beta}(x_{ij} - z_{lj})^2 \\ 0 & otherwise. \end{cases}$$

۳. مراکز خوشه‌ها (z) و درجات عضویت (u) را ثابت فرض کرده و مقدار بهینه مدل wk-means را نسبت به وزن‌ها (w) با استفاده از رابطه زیر بدست آور:

$$w_{kj} = \begin{cases} \frac{1}{m_i} & if \sum_{i=1}^{n} u_{ik}(x_{ij} - z_{kj})^2 = 0 \; and \; m_i = \left|\left\{t: \sum_{i=1}^{n} u_{ik}(x_{it} - z_{kt})^2 = 0\right\}\right|; \\ 0 & if \sum_{i=1}^{n} u_{ik}(x_{ij} - z_{kj})^2 \neq 0 \; and \; \sum_{i=1}^{n} u_{ik}(x_{it} - z_{kt})^2 = 0 \; for \; some \; t; \\ \frac{1}{\sum_{t=1}^{m} \left[\frac{\sum_{i=1}^{n} u_{ik}(x_{ij} - z_{kj})^2}{\sum_{i=1}^{n} u_{ik}(x_{it} - z_{kt})^2}\right]^{\frac{1}{\beta-1}}} & if \sum_{i=1}^{n} u_{ik}(x_{it} - z_{kt})^2 = 0 \; for \; each \; t. \end{cases}$$

٤. وزن‌ها (w) و درجات عضویت (u) را ثابت فرض کرده و مقدار بهینه مدل wk-means را نسبت به مراکز خوشه ها (z) با استفاده از رابطه زیر بدست آور:

$$z_{kj} = \frac{\sum_{i=1}^{n} u_{ik} x_{ij}}{\sum_{i=1}^{n} u_{ik}}$$

۵. درصورت عدم حصول شرط همگرایی برو به ۲.

۳- روش پیشنهادی (استخراج ویژگی مبتنی بر روش خوشه بندی k-means وزندار)

روش کاهش ویژگی پیشنهادی، شامل دو مرحله است. در مرحله اول، با استفاده از روش wk-means، خوشه‌بندی صورت می‌گیرد. همانطور که قبلا گفته شد در روش wk-means، سعی می‌شود با وزن‌دار کردن ویژگی‌های داده‌های هر خوشه یا تغییر شکل خوشه‌ها، خوشه‌بندی بهتری صورت گیرد. در مرحله دوم، یعنی پس از تعیین مقدار بهینه وزن ویژگی‌های هر خوشه (w_{kj}ها)، با استفاده از مدل زیر، S ویژگی هر خوشه به نحوی انتخاب می‌شود که کمترین تغییر در خوشه بندی یا کمترین تغیر در تابع هدف مدل wk-means حاصل شود:

$$\min_{u,z,\tilde{w}} F = \sum_{i=1}^{n}\sum_{k=1}^{c} u_{ik} \sum_{j=1}^{m} \tilde{w}_{kj} w_{kj}^{\beta} (x_{ij} - z_{kj})^2 \quad (۳)$$

$$\text{subject to} \begin{cases} \sum_{j=1}^{m} \tilde{w}_{kj} = S, \;\; k = 1,2,\dots,c; \\ \sum_{k=1}^{c} u_{ik} = 1, \quad i = 1,2,\dots,n; \\ \tilde{w}_{kj} \in \{0,1\}, \;\; k = 1,2,\dots,c; j = 1,2,\dots,m; \\ u_{ik} \in \{0,1\}, \;\; i = 1,2,\dots,n; \; k = 1,2,\dots,c. \end{cases}$$

که w_{kj} وزن ویژگی j–ام از خوشه k–ام است که قبلاً با استفاده از مدل wk-means به دست آمد. $\tilde{w}_{kj}$ متغیری برای انتخاب ویژگی است. اگر $\tilde{w}_{kj} = 1$ باشد یعنی j–امین ویژگی برای k–امین خوشه حفظ می‌شود وگرنه حذف می‌شود. برای حل مدل (۳)، الگوریتم زیر پیشنهاد می‌شود:

الگوریتم۳

۱. مراکز اولیه خوشه‌ها (z) و مقدار متغیرهای انتخاب ویژگی $(\tilde{w})$ را بصورت تصادفی درنظربگیر.

۲. مراکز خوشه‌ها (z) و مقدار متغیرهای انتخاب ویژگی $(\tilde{w})$ را ثابت فرض کرده و مقدار بهینه مدل (۳) را نسبت

به درجات عضویت (u) با استفاده از رابطه زیر بدست آور:

$$u_{ik} = \begin{cases} 1 & if\ \forall l: g_{ik} \le g_{il} \\ 0 & otherwise. \end{cases}$$

که $g_{il} = \sum_{j=1}^{m} \tilde{w}_{lj} w_{lj}^{\beta} (x_{ij} - z_{lj})^2$.

۳. مراکز خوشه‌ها (z) و درجات عضویت (u) را ثابت فرض کرده و مقدار بهینه مدل (۳) را نسبت به متغیرهای انتخاب ویژگی$(\tilde{w})$ با استفاده از رابطه زیر بدست آور:

$$\tilde{w}_{kj} = \begin{cases} 1 & f_{kj} \le S\min_{l}\{f_{kl}\}; \\ 0 & otherwise. \end{cases}$$

که $f_{kl} = \sum_{i=1}^{n} u_i w_{ik}^{\beta} (x_{il} - z_{kl})^2$ و منظور از Smin، Sامین کوچک‌ترین مقدار می باشد.

٤. مقدار متغیرهای انتخاب ویژگی $(\tilde{w})$ و درجات عضویت (u) را ثابت فرض کرده و مقدار بهینه مدل (۳) را نسبت به مراکز خوشه‌ها (z) با استفاده از رابطه زیر به دست آور:

$$z_{kj} = \frac{\sum_{i=1}^{n} u_{ik} x_{ij}}{\sum_{i=1}^{n} u_{ik}}$$

۵. درصورت عدم حصول شرط همگرایی برو به ۲.

شرط همگرایی هر یک از الگوریتم‌های ۱ تا ۳، عدم تغییر مراکز خوشه ها در دو تکرار متوالی است. در ادامه، درستی هر یک از مراحل الگوریتم۳ اثبات می‌شود.

قضیه ۱- اگر مراکز خوشه‌ها (z) و مقدار متغیرهای انتخاب ویژگی $(\tilde{w})$ را ثابت فرض کنیم مقدار بهینه مدل (۲٦) نسبت به درجات عضویت (u) با استفاده از رابطه زیر بدست می‌آید:

$$u_{ik} = \begin{cases} 1 & if\ \forall l: g_{ik} \le g_{il} \\ 0 & otherwise. \end{cases}$$

که $g_{il} = \sum_{j=1}^{m} \tilde{w}_{lj} w_{lj}^{\beta} (x_{ij} - z_{lj})^2$.

اثبات.

اگر مراکز خوشه‌ها (z) و مقدار متغیرهای انتخاب ویژگی $(\tilde{w})$ را ثابت فرض کنیم مدل (۳) بصورت زیر در می‌آید:

$$\min_{u} \sum_{i=1}^{n} \sum_{k=1}^{c} u_{ik} \sum_{j=1}^{m} \tilde{w}_{kj} w_{kj}^{\beta} (x_{ij} - z_{kj})^2$$

$$\text{subject to} \begin{cases} \sum_{k=1}^{c} u_{ik} = 1, & i = 1,2,\dots,n; \\ u_{ik} \in \{0,1\}, & i = 1,2,\dots,n;\ k = 1,2,\dots,c. \end{cases}$$

از آنجا که متغیرهای مدل فوق فقط u_{ik} ها هستند و در هر عبارت از مدل فوق، فقط یک متغیر وجود دارد مدل فوق را نیز می‌توان بصورت جمع چند مدل کوچکتر و بصورت زیر نوشت:

$$\sum_{i=1}^{n} \left(\begin{array}{l} \min_{u} \sum_{k=1}^{c} u_{ik} \sum_{j=1}^{m} \tilde{w}_{kj} w_{kj}^{\beta} (x_{ij} - z_{kj})^2 \\ \text{subject to} \begin{cases} \sum_{k=1}^{c} u_{ik} = 1; \\ u_{ik} \in \{0,1\}, \quad k = 1,2,\dots,c. \end{cases} \end{array} \right)$$

i–امین مدل از این مدل‌های کوچکتر را درنظربگیرید:

$$\min_{u} \sum_{k=1}^{c} u_{ik} \sum_{j=1}^{m} \tilde{w}_{kj} w_{kj}^{\beta} (x_{ij} - z_{kj})^2$$

$$\text{subject to} \begin{cases} \sum_{k=1}^{c} u_{ik} = 1; \\ u_{ik} \in \{0,1\}, \quad k = 1,2,\dots,c. \end{cases}$$

با توجه به قیود این مدل، فقط یکی از u_{ik} برابر با یک و سایر u_{ik} برابر با صفر می‌شوند. بنابراین برای حداقل شدن تابع هدف مدل، $u_{ik} = 1$ می‌شود اگر ضریب آن در تابع هدف، یعنی $g_{ik} = \sum_{j=1}^{m} \tilde{w}_{kj} w_{kj}^{\beta} (x_{1j} - z_{kj})^2$، حداقل باشد.

پایان اثبات.

قضیه ۲- اگر مراکز خوشه‌ها (z) و درجات عضویت (u) را ثابت فرض کنیم مقدار بهینه مدل (۳) نسبت به متغیرهای انتخاب ویژگی$(\tilde{w})$ با استفاده از رابطه زیر به دست می‌آید:

$$\tilde{w}_{kj} = \begin{cases} 1 & f_{kj} \le S\min_{l}\{f_{kl}\}; \\ 0 & otherwise. \end{cases}$$

که $f_{kl} = \sum_{i=1}^{n} u_i w_{ik}^{\beta} (x_{il} - z_{kl})^2$ و منظور از Smin، Sامین کوچکترین مقدار می‌باشد.

اثبات.

اگر مراکز خوشه‌ها (z) و درجات عضویت (u) را ثابت فرض کنیم مدل (۳) بصورت زیر در می‌آید:

$$\min_{\tilde{w}} \sum_{k=1}^{c}\sum_{j=1}^{m} \tilde{w}_{kj} \sum_{i=1}^{n} u_{ik} w_{kj}^{\beta} \left(x_{ij} - z_{kj}\right)^2$$

$$\text{subject to} \begin{cases} \sum_{j=1}^{m} \tilde{w}_{kj} = S, \;\; k = 1,2,\dots,c; \\ \tilde{w}_{kj} \in \{0,1\}, \;\; k = 1,2,\dots,c; j = 1,2,\dots,m; \end{cases}$$

از آنجا که متغیرهای مدل فوق فقط $\tilde{w}_{kj}$ ها هستند و در هر عبارت از مدل فوق، فقط یک متغیر وجود دارد، مدل فوق را نیز می‌توان بصورت جمع چند مدل کوچکتر و بصورت زیر نوشت:

$$\sum_{k=1}^{c} \left(\begin{array}{l} \min_{\tilde{w}} \sum_{j=1}^{m} \tilde{w}_{kj} \sum_{i=1}^{n} u_{ik}\, w_{kj}^{\beta} \left(x_{ij} - z_{kj}\right)^2 \\ \text{subject to} \begin{cases} \sum_{j=1}^{m} \tilde{w}_{kj} = S; \\ \tilde{w}_{kj} \in \{0,1\}, \;\; j = 1,2,\dots,m; \end{cases} \end{array} \right)$$

k–امین مدل از این مدل‌های کوچکتر را درنظربگیرید:

$$\min_{\tilde{w}} \sum_{j=1}^{m} \tilde{w}_{kj} \sum_{i=1}^{n} u_{ik}\, w_{kj}^{\beta} \left(x_{ij} - z_{kj}\right)^2$$

$$\text{subject to} \begin{cases} \sum_{j=1}^{m} \tilde{w}_{kj} = S; \\ \tilde{w}_{kj} \in \{0,1\}, \;\; j = 1,2,\dots,m; \end{cases}$$

با توجه به قیود این مدل، فقط S تا از $\tilde{w}_{kj}$ برابر با یک و سایر $\tilde{w}_{kj}$ برابر با صفر می‌شوند. بنابراین برای حداقل شدن تابع هدف مدل، $\tilde{w}_{kj} = 1$ می‌شود. اگر ضریب آن در تابع هدف، یعنی $f_{kj} = \sum_{i=1}^{n} u_i w_{ik}^{\beta} \left(x_{ij} - z_{kj}\right)^2$، کمتر از S–امین کوچکترین ضریب $\tilde{w}_{lj}$ باشد.

پایان اثبات.

قضیه ۳– اگر مقدار متغیرهای انتخاب ویژگی $(\tilde{w})$ و درجات عضویت (u) را ثابت فرض کنیم مقدار بهینه مدل (۲۶) نسبت به مراکز خوشه‌ها (z) با استفاده از رابطه زیر به دست می‌آید:

$$z_{kj} = \frac{\sum_{i=1}^{n} u_{ik} x_{ij}}{\sum_{i=1}^{n} u_{ik}}$$

اثبات.

هنگامی که متغیرهای u و $\tilde{w}$ را ثابت در نظر می‌گیریم برای بدست آوردن مقدار بهینه متغیر z کافیست مشتق تابع هدف مدل (۳) نسبت به z را برابر مقدار صفر قرار دهیم:

$$\frac{\partial F}{\partial z_{kj}} = 0 \qquad \rightarrow \sum_{i=1}^{n} -2u_{ik} w_{kj}^{\beta} \tilde{w}_{kj} \left(x_{ij} - zkj=0;\right.$$

$$\rightarrow 2w_{kj}^{\beta} \tilde{w}_{kj} \sum_{i=1}^{n} u_{ik} x_{ij} =$$

$$2w_{kj}^{\beta} \tilde{w}_{kj} z_{kj} \sum_{i=1}^{n} u_{ik}; \rightarrow z_{kj} = \frac{\sum_{i=1}^{n} u_{ik} x_{ij}}{\sum_{i=1}^{n} u_{ik}}.$$

پایان اثبات.

٤– آزمایش‌ها و ارزیابی

در این تحقیق از مجموعه داده‌های واقعی ذیل از مخزن UCI برای مقایسه تجربی روش کاهش ویژگی پیشنهادی با روش‌های کاهش ویژگی LLE، L1PCA و روش K-means تُنُک استفاده می‌گردد:

۱)Wine: این مجموعه داده دارای ۳ خوشه با مجموعا ۱۷۸ داده و ۱۳ ویژگی است.

۲)Vertebral2: این مجموعه داده دارای ۲ خوشه با مجموعا ۳۱۰ داده و ٦ ویژگی است.

۳)Vertebral3: این مجموعه داده دارای ۳ خوشه با مجموعا ۳۱۰ داده و ٦ ویژگی است.

٤)Parkinsons: این مجموعه داده دارای ۲ خوشه با مجموعا ۱۹۵ داده و ۲۲ ویژگی است.

در هر آزمایش، یکی از روش‌های کاهش ویژگی بر روی یک مجموعه داده اعمال شده و سپس مجموعه داده کاهش بعد یافته، با استفاده از روش wk-means خوشه‌بندی می‌شود. روش‌های خوشه‌بندی و کاهش ابعاد مبتنی بر wK-means به شدت به مقادیر اولیه مرکز خوشه‌ها، که به صورت تصادفی انتخاب می شوند، وابسته است. لذا هریک از آزمایش‌ها ۲۰ بار تکرار شده و میانگین و انحراف معیار نتایج گزارش می‌شود. درضمن، مقدار پارامتر ß در مدل wk-means و فاز اول روش پیشنهادی برابر با ٦ در نظر گرفته شده است.

جدول ۱: مقایسه دقت خوشه بندی به روش wk-means با استفاده از داده‌های کاهش ابعاد یافته به روش‌های LLE, L1PCA و K-means تُنُک و روش پیشنهادی

تعداد ویژگی پس از کاهش ویژگی	الگوریتم کاهش ابعاد	Wine	Parkinsons	Vertebral3	Vertebral2
D = 2	روش پیشنهادی	۰,۶۵۱۰±۰,۰۵۵	۰,۷۱۵۳±۰,۰۵۰	۰,۴۹۳۵±۰,۰۰۸	۰,۶۲۶۰±۰,۰۳۰
	Sparse K-means[6]	۰,۵۶۱۸±۰,۰۱۰	۰,۶۱۵۳±۰,۰۱۰	۰,۶۲۹۴±۰,۰۱۰	۰,۵۳۵۵±۰,۰۶۰
	LLE[3]	۰,۵۴۳۸±۰,۰۱۰	۰,۶۲۱۳±۰,۰۱۰	۰,۵۴۱۶±۰,۰۱۳	۰,۶۲۰۲±۰,۰۲۲
	L1PCA[۲]	۰,۶۳۲۱±۰,۰۱۴	۰,۴۹۱۲±۰,۰۱۲	۰,۴۹۴۵±۰,۰۲۴	۰,۵۷۸۱±۰,۰۱۲
D = 3	روش پیشنهادی	۰,۶۰۸۹±۰,۰۸۶	۰,۷۴۷۶±۰,۰۲۳	۵۳۳۵±۰,۰۳۰	۰,۶۱۰۶±۰,۰۱۶
	Sparse K-means[6]	۰,۴۹۶۵±۰,۰۱۰	۰,۷۱۳۳±۰,۰۲۰	۰,۳۹۵۷±۰,۲۰	۰,۶۷۱۰±۰,۰۲۰
	LLE[3]	۰,۴۰۴۵±۰,۰۱۱	۰,۷۲۲۸±۰,۰۱۰	۰,۳۹۶۰±۰,۰۱۰	۰,۶۸۸۳±۰,۰۴۲
	L1PCA[۲]	۰,۸۱۱۲±۰,۰۲۲	۰,۷۴۶۸±۰,۰۱۳	۰,۳۵۲۱±۰,۰۱۹	۰,۶۹۰۱±۰,۰۲۹
D = 4	روش پیشنهادی	۰,۶۹۱۵±۰,۰۵۳	۰,۷۶۷۶±۰,۰۲۰	۰,۵۵۸۰±۰,۰۱۵	۰,۶۳۳۴±۰,۰۱۱
	Sparse K-means[6]	۰,۶۷۴۰±۰,۰۴۲	۰,۷۱۳۳±۰,۰۲۰	۰,۵۴۳۰±۰,۰۳۰	۰,۶۸۷۰±۰,۰۳۰
	LLE[3]	۰,۶۸۱۷±۰,۰۴۳	۰,۷۵۳۵±۰,۰۱۴	۰,۵۰۶۵±۰,۰۱۰	۰,۶۶۲۵±۰,۰۲۰
	L1PCA[۲]	۰,۶۷۸۴±۰,۰۱۹	۰,۷۳۹۸±۰,۰۱۱	۰,۵۳۰۹±۰,۰۱۲	۰,۶۳۹۶±۰,۰۲۲
D = 5	روش پیشنهادی	۰,۶۳۱۴±۰,۰۹۳	۰,۷۹۱۷±۰,۰۳۴	۰,۵۱۸۳±۰,۰۱۸	۰,۶۴۶۴±۰,۰۱۱
	Sparse K-means[6]	۰,۷۵۴۲±۰,۰۲۰	۰,۷۲۳۳±۰,۰۱۰	۰,۵۵۷۷±۰,۰۱۰	۰,۶۹۹۰±۰,۰۳۰
	LLE[3]	۰,۶۶۰۴±۰,۰۱۰	۰,۷۶۳۵±۰,۰۲۰	۰,۵۰۶۵±۰,۰۱۰	۰,۶۶۴۵±۰,۰۲۰
	L1PCA[۲]	۰,۶۵۹۷±۰,۰۲۱	۰,۷۵۶۲±۰,۰۱۱	۰,۵۴۹۷±۰,۰۱۲	۰,۶۳۹۸±۰,۰۱۹
D = 6	روش پیشنهادی	۰,۶۶۷۹±۰,۰۵۴	۰,۷۱۸۴±۰,۰۳۳	–	–
	Sparse K-means[6]	۰,۶۵۱۶±۰,۰۲۶	۰,۶۳۹۰±0.018	–	–
	LLE[3]	۰,۴۹۲۱±۰,۰۳۵	۰,۶۹۰۲±۰,۰۳۷	–	–
	L1PCA[۲]	۰,۵۶۲۰±۰,۰۵۱	۰,۵۶۲۷±۰,۰۳۲	–	–

جدول ۲: مقایسه نتایج خوشه بندی به روش wk-means بدون کاهش بعد داده‌ها.

Wine	Parkinsons	Vertebral3	Vertebral2
۰,۸۹۹۸±۰,۰۳۰	۰,۶۶۱۵±۰,۰۱۰	۰,۴۸۵۹±۰,۰۳۰	۰,۶۵۳۷±۰,۰۱۰

با توجه به جدول ۱ می‌توان به نکات زیر اشاره کرد:

همانطور که در جدول ۱ مشخص است، دقت خوشه‌بندی هر مجموعه داده پس از کاهش ویژگی، بستگی به تعداد ابعاد پس از کاهش ویژگی یا میزان کاهش ویژگی، روش کاهش ویژگی ونوع مجموعه داده دارد. به عنوان مثال، میانگین دقت خوشه‌بندی مجموعه داده ۱۳ بعدی Wine پس از کاهش ویژگی به ۳ ویژگی با استفاده از روش L1PCA، ۸۱,۱۲٪ شده است، حال آنکه میانگین دقت

خوشه‌بندی مجموعه داده ٦ بعدی Vertebral3 پس از کاهش ویژگی به ۳ ویژگی با استفاده از روش L1PCA، ۳۵,۲۱٪ شده است. در جدول ۱ بهترین روش کاهش ویژگی به ازای هر مجموعه داده که بیشترین دقت خوشه بندی را پس از کاهش ویژگی به دست آورده است به صورت پررنگ تر مشخص شده است.

در حالت کلی کاهش بیش از حد ابعاد در مجموعه داده‌های مختلف می‌تواند دقت خوشه‌بندی را به شدت کاهش دهد. برخی نتایج در ابعاد ۲ و ۳ تقریبا غیر قابل اتکا می‌باشد، لیکن با کاهش ابعاد داده به میزان مناسب در هر مجموعه داده، می‌توان امیدوار بود که روش پیشنهادی علاوه بر تامین اهداف روش‌های کاهش ابعاد (از جمله کاهش فضای ذخیره سازی، فشرده سازی و افزایش سرعت پردازش بر روی داده‌ها) دقت مناسبی را در حل مسائل داده کاوی ارائه خواهد داد.

مقایسه جدول ۱ و ۲ نشان می‌دهد که در بسیاری موارد، انتخاب ویژگی مناسب در مجموعه‌های داده می‌تواند نتایج بهتری را در مسائل داده کاوی ارائه دهد. به این معنا که وجود بعضی ویژگی‌ها در برخی از مجموعه داده‌ها می‌تواند نتایج تحلیل داده‌ها را منحرف کند و دقت خوشه بندی را کاهش دهد.

برای مقایسه کارایی چند خوشه بند بر روی بیش‌از دو مجموعه داده، استفاده از میانگین آنها نادرست است [۷]. لذا از روش رتبه‌بندی خوشه‌بندها برای هر مجموعه داده به‌طور مستقل از سایر مجموعه داده‌ها استفاده می‌گردد [۸]. روش پیشنهادی در مقایسه با روش L1PCA از ۱۸ حالت آزمایش شده ۱۳ بار رتبه بهتری به دست آورده است. لازم به ذکر است خوشه‌بندی داده‌های کاهش ابعاد یافته به روش پیشنهادی در ابعاد بالاتر عملکرد بهتری از خود نشان می‌دهد. روش پیشنهادی در ۲۲ حالت از ۱۳ حالت آزمایش نسبت به روش LLE عملکرد بهتری داشته است. همانطور که در جدول ۱ مشاهده می‌شود، روش پیشنهادی در اکثر آزمایش‌ها در ابعاد ۵ و ٦ از روش LLE پیشی گرفته است. اگر بخواهیم روش پیشنهادی را با روش K-means تُنک، که یک روش مبتنی بر K-means می‌باشد، مقایسه کنیم، در جدول ۱ برتری روش پیشنهادی مشاهده می‌شود. چرا که روش پیشنهادی در ۱۲ مورد از ۱۸ مورد آزمایش این مقاله، عملکرد بهتری از خود نشان داده است. بنابراین، می‌توان گفت که دقت خوشه‌بندی پس از کاهش ویژگی با روش پیشنهادی بهتر از سایر روش‌های کاهش ویژگی مورد مقایسه در این پایان نامه است.

آزمایش‌های مذکور با استفاده از پردازنده ٦٤ بیتی Core i5 2.5Ghz به همراه حافظه اصلی 6GB و با استفاده از سیستم عامل Windows 7 Ultimate آزمایش شده است. جدول ۳ سرعت اجرای الگوریتم‌های کاهش ابعاد مورد آزمایش را نشان می‌دهد. براین اساس می‌توان گفت که سرعت اجرای الگوریتم پیشنهادی بهتر از روش LLE و بدتر از روش L1PCA و k-means تنک است.

۵- نتیجه گیری

در این مقاله، با توجه به موفقیت روش‌های خوشه‌بندی مبتنی بر k-means، یک روش کاهش ویژگی بر پایه k-means وزن دار ارائه شد. به بیان دقیقتر، در روش پیشنهادی، نخست با استفاده از روش k-means وزن دار، به ویژگی‌های داده‌ها وزن داده شد. ویژگی‌های وزین‌تر ضرورتا ویژگی‌های مهمتر نیستند و وزن هر ویژگی، تنها بازه هر ویژگی را به نحوی تغییر می‌دهد که خوشه‌بندی بهتری صورت بگیرد. لذا، سپس با استفاده از یک مدل ریاضی جدید، کسری از ویژگی‌های وزن‌دار شده داده‌های هر خوشه انتخاب شد به نحوی که کمترین تغییر در خوشه‌ها حاصل شود. تعداد ویژگی‌های منتخب هر خوشه در روش پیشنهادی، برخلاف روش‌های مشابهی چون k-means تنک و fuzzy c-means تنک بصورت صریح تعیین می‌شود. درضمن، آزمایش‌های تجربی روی چهار مجموعه‌داده واقعی نشان می‌دهد که روش پیشنهادی، از دقت بیشتری نسبت به روش‌های L1PCA, LLE و روش

K-means تُنُک برخوردار است.

برای مقایسه روش کاهش ویژگی پیشنهادی با سایر روش‌های کاهش ویژگی، در هر آزمایش، یکی از روش‌های کاهش ویژگی بر روی یک مجموعه داده واقعی اعمال شده و سپس مجموعه داده کاهش بعد یافته، با استفاده از روش wk-means خوشه بندی شدند. آزمایش‌های تجربی روی چهار مجموعه داده واقعی نشان داد که روش پیشنهادی در مقایسه با روش L1PCA از ۱۸ حالت آزمایش شده ۱۳ بار رتبه بهتری بدست آورده است. روش پیشنهادی در ۲۲ حالت از ۱۳ حالت آزمایش نسبت به روش LLE عملکرد بهتری داشته است. روش پیشنهادی در ۱۲ مورد از ۱۸ مورد آزمایش این مقاله، عملکرد بهتری نسبت به روش k-means تنک داشته است. بنابراین، می‌توان گفت که دقت خوشه‌بندی پس از کاهش ویژگی با روش پیشنهادی بهتر از سایر روش‌های کاهش ویژگی مورد مقایسه در این مقاله است. درضمن، سرعت اجرای الگوریتم پیشنهادی بهتر از روش LLE و بدتر از روش L1PCA و k-means تنک است.

جدول ۳: مقایسه سرعت اجرای الگوریتم‌های کاهش ابعاد LLE, L1PCA و K-meansتُنُک و روش پیشنهادی (به ثانیه).

اندازه ابعاد آزمایش	الگوریتم کاهش ابعاد	Wine	Parkinsons	Vertebral3	Vertebral2
D = 2	روش پیشنهادی	۰,۰۵۸	۰,۰۷۷	۰,۰۴۸	۰,۰۴۰
	Sparse K-means[6]	۰,۰۲۵	۰,۰۳۶	۰,۰۳۸	۰,۰۲۷
	LLE[3]	۰,۰۸۰	۰,۱۶۸	۰,۱۱۱	۰,۰۹۴
	L1PCA[۲]	۰,۰۱۳	۰,۰۲۵	۰,۰۲۸	۰,۰۱۷
D = 3	روش پیشنهادی	۰,۰۵۵	۰,۰۷۷	۰,۰۴۵	۰,۰۴۰
	Sparse K-means[6]	۰,۰۲۱	۰,۰۳۶	۰,۰۳۱	۰,۰۲۸
	LLE[3]	۰,۰۸۰	۰,۱۵۳	۰,۱۱۱	۰,۱۱۱
	L1PCA[۲]	۰,۰۱۷	۰,۰۲۵	۰,۰۲۶	۰,۰۱۷
D = 4	روش پیشنهادی	۰,۰۶۱	۰,۰۹۴	۰,۰۴۸	۰,۰۴۰
	Sparse K-means[6]	۰,۰۲۵	۰,۰۳۶	۰,۰۳۶	۰,۰۳۳
	LLE[3]	۰,۰۷۵	۰,۱۷۰	۰,۱۰۸	۰,۱۱۱
	L1PCA[۲]	۰,۰۲۳	۰,۰۲۱	۰,۰۱۸	۰,۰۱۹
D = 5	روش پیشنهادی	۰,۰۵۷	۰,۰۹۷	۰,۰۴۹	۰,۰۴۰
	Sparse K-means[6]	۰,۰۲۶	۰,۰۴۵	۰,۰۳۹	۰,۰۳۵
	LLE[3]	۰,۰۷۰	۰,۱۷۲	۰,۰۹۶	۰,۱۱۶
	L1PCA[۲]	۰,۰۱۹	۰,۰۳۲	۰,۰۳۱	۰,۰۲۹
D = 6	روش پیشنهادی	۰,۰۶۴	۰,۰۹۴	–	–
	Sparse K-means[6]	۰,۰۲۷	۰,۰۴۵	–	–
	LLE[3]	۰,۰۷۳	۰,۱۷۲	–	–
	L1PCA[۲]	۰,۰۲۱	۰,۰۳۹	–	–

٦- مراجع

1. Ding, C. and T. Li. *Adaptive dimension reduction using discriminant analysis and k-means clustering*. in *Proceedings of the 24th International Conference on Machine Learning*. 2007. Corvallis.
2. Markopoulos, P.P., et al., *Efficient L1-Norm Principal-Component Analysis via Bit Flipping.* IEEE Transactions on Signal Processing, 2017. **65**(16): p. 4252-4264.
3. Rovis, S.T. and L.K. Saul, *Nonlinear Dimentionality Reduction by Locally Linear Embedding.* Sience, 2000. **290**: p. 2323-2326.
4. Qiu, X., et al., *A sparse fuzzy c-means algorithm based on sparse clustering frame work.* Jornal of Elsevier on Neurocomputing, 2015. **157**: p. 290-295.
5. Huang, X., Y. Ye, and H. Zhang, *Extensions of Kmeans-Type Algorithms: A New Clustering Framework by Integrating Intracluster Compactness and Intercluster Separation.* IEEE Transactions on Neural Networks and Learning Systems, 2014. **25**(8): p. 1433 - 1446.
6. Tibshirani, R. and D.M. Witten, *A framework for feature selection in clustering.* NIH-PA Author Manuscript, 2010. **105**(490): p. 713–726.
7. Demsar, J., *Statistical comparison of classifier over multiple data sets.* Journal of Machine Learning, 2006. **7**: p. 1-30.
8. García, S., A. Fernández, and J. Luengo, *Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power*. Information Sciences, 2010. **180**: p. 2044-2064.