مهندسی صنایع و مدیریت شریف، (تابستان ۱۳۹۶)
دوره‌ی ۱-۳۳، شماره‌ی ۱/۱، ص. ۱۳-۱۹

Original Article

ارائه‌ی یک الگوریتم ترکیبی برای خوشه‌بندی داده‌ها با استفاده از الگوریتم‌های K-means و الکترومغناطیس

اسماعیل مهدی‌زاده* **(دانشیار)**
محمد تیموری (کارشناس ارشد)
آرش زارع‌طلب (کارشناس ارشد)
دانشکده‌ی مهندسی صنایع و مکانیک، دانشگاه آزاد اسلامی واحد قزوین

خوشه‌بندی یکی از روش‌های پرکاربرد در بسیاری از زمینه‌های علمی است که در آن تلاش می‌شود داده‌ها داخل گروه‌ها براساس درجه‌ی شباهت قرار گیرند. الگوریتم‌های ابتکاری و فرابتکاری زیادی برای حل مسئله‌ی خوشه‌بندی ارائه شده است. یکی از روش‌های ابتکاری پرکاربرد، K-means است. این روش، به‌دلیل وابستگی به حالت اولیه، معمولاً به بهینه‌ی محلی همگرا می‌شود. در این مقاله به‌منظور فرار از بهینه‌ی محلی، الگوریتم K-means با الگوریتم فرابتکاری الکترومغناطیس ترکیب شده و الگوریتم جدیدی با عنوان الگوریتم K-EM برای حل مسئله‌ی خوشه‌بندی ارائه می‌شود. به‌منظور بررسی کارایی الگوریتم پیشنهادی، پنج مجموعه داده انتخاب و حل شده و نهایتاً جواب‌های حاصله با جواب‌های حاصل از الگوریتم‌های مطرح در ادبیات خوشه‌بندی مقایسه می‌شود. نتایج محاسباتی نشان می‌دهد که الگوریتم پیشنهادی در دست‌یابی به جواب‌های مطلوب از کارایی مناسبی برخوردار است.

emehdi@qiau.ac.ir
m66.teimouri@gmail.com
arash_zaretalab@yahoo.com

واژگان کلیدی: خوشه‌بندی، الگوریتم K-means، الگوریتم الکترومغناطیس.

۱. مقدمه

خوشه‌بندی داده‌ها یکی از روش‌های مهم طبقه‌بندی بدون نظارت است. در این فرایند، داده‌ها داخل گروه‌ها یا خوشه‌های خاص براساس درجه شباهت بین‌شان مرتب می‌شوند. در سال‌های اخیر خوشه‌بندی در زمینه‌های مختلف ــ نظیر پردازش تصاویر،[۱] تشخیص ناهنجاری،[۲] پزشکی،[۳] مدیریت ساخت‌وساز،[۴] بازاریابی،[۵] قابلیت اطمینان،[۶] انتخاب تأمین‌کننده[۷] و تحلیل پوششی داده‌ها[۸] ــ مورد توجه محققین قرار گرفته است.

در ادبیات موضوع، دسته‌بندی‌های مختلفی برای روش‌های خوشه‌بندی در نظر گرفته می‌شود اما، به‌طور کلی می‌توان این روش‌ها را چنین تقسیم‌بندی کرد:[۷] روش‌های خوشه‌بندی سلسله‌مراتبی،[۹] مدل ترکیبی خوشه‌بندی،[۱۰] خوشه‌بندی شبکه یادگیری،[۱۱] روش‌های براساس تابع هدف و خوشه‌بندی پارتیشن.[۱۲]

عموماً در بیشتر روش‌های خوشه‌بندی، نظیر الگوریتم K-means، هدف کمینه‌سازی کل عدم تشابه است.[۱۳] این الگوریتم کاربردی‌ترین روش خوشه‌بندی داده‌هاست که در مسائل بزرگ بسیار کارایی دارد. الگوریتم K-means سریع، کارا و آسان با پیچیدگی زمان خطی است. این الگوریتم مجموعه داده را داخل خوشه‌هایی که از قبل تعدادشان مشخص شده، دسته‌بندی می‌کند و هدف آن کمینه‌سازی فاصله‌ی درون خوشه‌یی است. اگرچه، K-means سریع و ساده است ولی دو مشکل اساسی دارد: اول این که به موقعیت اولیه‌اش بسیار وابسته است و به‌همین دلیل اغلب به بهینه‌ی محلی همگرا می‌شود. دوم این که در این روش تعداد خوشه به‌عنوان ورودی الگوریتم مورد نیاز است و برای اجرای الگوریتم باید از قبل مشخص شود. البته روش‌های مختلفی برای حل این مشکل وجود دارد.[۱۴]

در سال‌های اخیر محققین برای غلبه بر مشکل بهینه‌ی محلی تلاش کردند تا با استفاده از الگوریتم‌های ابتکاری و فرابتکاری، که بسیاری از آن‌ها برگرفته از پدیده‌های اجتماعی و طبیعی‌اند، تابع هدف K-means را بهینه کنند. مثلاً در سال ۱۹۹۵، روشی مبتنی بر جست‌وجوی ممنوع[۱] برای مسائل خوشه‌بندی معرفی شد.[۱۵] در سال ۲۰۰۴ نیز الگوریتمی مبتنی بر بهینه‌سازی کلونی مورچگان[۲] برای آنالیز خوشه‌بندی معرفی شد.[۱۶] سپس در سال ۲۰۰۷ برای حل مسائل خوشه‌بندی، کاربرد بهینه‌سازی جفت‌گیری زنبور عسل[۳] مطرح شد.[۱۷] در سال ۲۰۰۸ محققین الگوریتمی ترکیبی براساس الگوریتم ژنتیک[۴]، K-means و حداکثر انتظار رگرسیون لگاریتمی[۵] ارائه کردند.[۱۸] همچنین یک الگوریتم ترکیبی براساس K-means، بهینه‌سازی توده ذرات[۶] و جست‌وجوی سیمپلکس نیلدرمید برای حل مسائل خوشه‌بندی معرفی شد.[۱۹] در سال ۲۰۱۰، از یک روش کارآمد ترکیبی براساس الگوریتم‌های بهینه‌سازی توده ذرات، بهینه‌سازی کلونی مورچگان و K-means برای تحلیل خوشه استفاده

* نویسنده مسئول
تاریخ: دریافت ۱۳۹۳/۲/۲۷، اصلاحیه ۱۳۹۴/۴/۲۹، پذیرش ۱۳۹۴/۵/۲۵.

شد.[۲۰] در سال ۲۰۱۰ نیز یک الگوریتم کلنی زنبورعسل مصنوعی[۷] برای اجرای خوشه‌بندی ارائه شد.[۲۱] در سال ۲۰۱۱ نیز یک الگوریتم ترکیبی هرج و مرج توده ذرات[۸] برای خوشه‌بندی داده‌ها معرفی شد.[۲۲] محققین در سال ۲۰۱۱، یک الگوریتم کارآمد ترکیبی براساس الگوریتم‌های رقابت استعماری و برایش شده[۹] و K-means برای خوشه‌بندی داده‌ها ارائه کردند.[۲۳] آنان در سال ۲۰۱۲ یک الگوریتم جست‌وجوی باینری[۱۰] معرفی، و کاربردش را روی مجموعه داده‌های واقعی آزمودند.[۲۴] هاتملو در سال ۲۰۱۳، یک الگوریتم سیاهچاله[۱۱] را برای خوشه‌بندی داده‌ها معرفی کرد.[۲۵]

الگوریتم‌های ذکر شده برای بهبود کاستی‌های الگوریتم K-means، چه با ترکیب با آن و چه به تنهایی، می‌کوشند ولی این نکته را باید پذیرفت که برخی از آن‌ها نیز از کاستی‌هایی رنج می‌برند؛ کاستی‌هایی نظیر کیفیت نتایج، سرعت همگرایی پایین (مانند آنچه در الگوریتم‌های جست‌وجوی ممنوع و ژنتیک شاهدیم) و برخی دارای ساختار پیچیده، و نرخ همگرایی پایین مانند بهینه‌سازی توده ذرات‌اند.

یکی از روش‌های فراابتکاری که در سال‌های اخیر مورد توجه قرار گرفته، الگوریتم الکترومغناطیس است. الگوریتم الکترومغناطیس یک روش فراابتکاری مبتنی بر جمعیت است.[۲۶و۲۷] این الگوریتم، از سازوکار جاذبه ـ دافعه‌ی نظریه‌ی الکترومغناطیس برای تعیین جواب بهینه استفاده می‌کند. یکی از مهم‌ترین مزایای این الگوریتم تعداد کم پارامتری است که برای تنظیم شدن دارد. علاوه بر این، جواب‌های به دست آمده توسط این الگوریتم به‌راحتی به بهینه‌ی محلی گرفتار نمی‌شوند. از آنجا که الگوریتم الکترومغناطیس، یک الگوریتم مبتنی بر جمعیت است، کیفیت جواب به دست آمده توسط آن به جمعیت اولیه وابسته است. بنابراین در این مطالعه از جواب‌های اولیه‌ی تولید شده توسط K-means به‌عنوان جواب کاندید برای الگوریتم الکترومغناطیس استفاده می‌کنیم. ما نشان خواهیم داد که الگوریتم ترکیبی K-EM قوی و مناسب برای خوشه‌بندی داده‌هاست و همچنین کیفیت جواب‌های به دست آمده توسط این الگوریتم دارای کیفیت بالایی است؛ برای این کار از چندین مجموعه داده واقعی و مشهور برای ارزیابی استفاده می‌شود.

درادامه این نوشتار، به تحلیل خوشه‌یی داده (بخش ۲)، و سپس معرفی الگوریتم الکترومغناطیس (بخش ۳) خواهیم پرداخت. در بخش ۴ الگوریتم ترکیبی K-EM معرفی می‌شود و سپس درخصوص نتایج آزمایشات در بخش ۵ بحث می‌شود. نهایتا در بخش ۶ نتیجه‌گیری کلی مقاله ارائه می‌شود.

۲. تحلیل خوشه‌بندی داده

خوشه‌بندی یکی از شاخه‌های یادگیری بدون نظارت است و فرایند خودکاری است که در طی آن، اشیا به دسته‌هایی تقسیم می‌شوند که اعضای آن از نظر شاخص‌های مورد نظر مشابه یکدیگرند. بنابراین، برای سنجش شباهت بین اشیا از اندازه‌گیری فاصله استفاده می‌شود. روش‌های مختلفی برای اندازه‌گیری فاصله بین دو شی وجود دارد که فاصله اقلیدسی معروف‌ترین و پرکاربردترین گونه‌ی فاصله است. این فاصله از فاصله‌ی مینکوفسکی مشتق شده است و به‌صورت رابطه ۱ تعریف می‌شود:

$$d(x,y)=\left(\sum_{i=1}^{m}|x_i-y_i|^r\right)^{1/r}\Rightarrow d(x,y)=\sqrt{\sum_{i=1}^{m}(x_i-y_i)^2} \quad (1)$$

فرض کنید مجموعه‌یی از N شی $O=\{O_1,O_2,...,O_N\}$ به‌وسیله‌ی مجموعه‌ی O نشان داده شود که در آن شی O_i به‌وسیله‌ی ویژگی‌های $\{o_{i1},o_{i2},...,o_{id}\}$ نشان داده شود. همچنین، مجموعه‌یی از K خوشه‌ی $C=\{C_1,C_2,...,C_K\}$ در فضای اقلیدسی N بعدی R^N را در نظر بگیرید. هدف در یک مسئله‌ی خوشه‌بندی، تقسیم مجموعه‌ی O درون مجموعه‌ی K است. از این‌رو، شرایط خوشه‌ها باید عبارت باشد از:

- هر خوشه حداقل باید شامل یک شیء باشد، یعنی:

$$C_i\neq\phi \quad \forall i\in\{1,2,...,K\}$$

- دو خوشه‌ی مختلف نباید اشیاء مشترک داشته باشند، یعنی:

$$i,j\in\{1,2,...,K, C_i\cap C_j=\phi \quad \forall i\neq j\}$$

- هر شیء باید به یک خوشه تخصیص پیدا کند، یعنی:

$$\bigcup_{i=1}^{K}C_i=O$$

الگوریتم K-means کاربردی‌ترین روش خوشه‌بندی داده‌هاست. برای پیدا کردن K خوشه، مسئله به‌عنوان بهینه‌سازی (حداقل‌سازی) یک تابع عملکرد، روی داده‌ها و مکان خوشه‌ها تعریف می‌شود. تابع عملکرد مورد استفاده برای این هدف، به‌صورت رابطه‌ی ۲ تعریف می‌شود:[۲۰]

$$f(O,C)=\sum_{i=1}^{N}\min\left\{\|O_i-C_l\|^2\,\middle|\,l=1,2,...,K\right\} \quad (2)$$

۳. الگوریتم الکترومغناطیس

الگوریتم الکترومغناطیس، یک الگوریتم فراابتکاری مبتنی بر جمعیت است که اولین بار توسط بیر بیل و فنگ در سال ۲۰۰۳ برای جست‌وجوی جواب بهینه در مسائل بهینه‌سازی پیوسته ارائه شد.[۲۶] در سال‌های اخیر، این روش برای حل مسائل در زمینه‌های مختلف از جمله: زمان‌بندی تولید کارگاهی دوره‌یی،[۲۸] مسیریابی وسایل نقلیه‌ی تجهیز شده،[۲۹] بهینه‌سازی کلی خطی محدود،[۳۰] طراحی چیدمان سیستم‌های تولیدی قابل تنظیم مجدد،[۳۱] مورد استفاده قرار گرفته است.

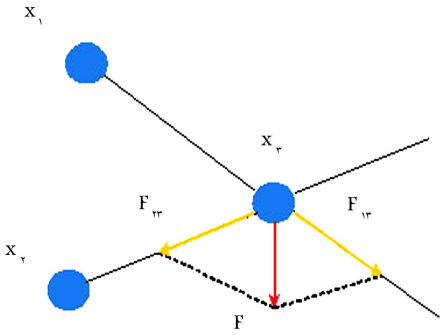
الگوریتم الکترومغناطیس از سازوکار جاذبه ـ دافعه نظریه‌ی الکترومغناطیس برای تعیین جواب بهینه استفاده می‌کند. این روش عمدتاً دارای چهار فرایند اصلی است: مقداردهی اولیه‌ی الگوریتم، انجام جست‌وجوی محلی برای یافتن بهینه‌ی محلی، محاسبه‌ی نیروی کل وارده بر هر ذره و حرکت در طول جهت نیرو. طرح کلی نسخه‌ی ابتدایی الگوریتم را می‌توان در شکل ۱ مشاهده کرد.

در مرحله‌ی اول فرایند، اندازه ذرات[۱۲] به‌طور کاملاً تصادفی از ناحیه‌ی موجه انتخاب می‌شود. مقادیر اولیه بین یک حد بالا و پایین انتخاب می‌شود که این حدود بین بازه [۰,۱] تنظیم می‌شوند. سپس مقدار تابع هدف نقاط محاسبه شده و ذره‌ی دارای بهترین مقدار ذخیره می‌شود.

در مرحله‌ی دوم فرایند، جست‌وجوی محلی برای یافتن بهینه‌ی محلی برای هر ذره، اجرا می‌شود. در این مرحله می‌توان از هر روش جست‌وجوی محلی برای افزایش کارایی الگوریتم استفاده کرد. مرحله‌ی سوم فرایند، به محاسبه‌ی نیروی کل وارده بر هر ذره اختصاص دارد. نیروی الکتروستاتیک بین دو نقطه‌ی بارها به‌طور مستقیم متناسب با مقدار هر بار و به‌طور معکوس متناسب با مربع فاصله‌ی بین

ALGORITHM 1. EM (*popsize*, *MAXITER*, *LSITER*, δ)
popsize: number of sample points
MAXITER: maximum number of iterations
LSITER: maximum number of local search iterations
δ: local search parameter, $\delta \in [0, 1]$
1. Initialize ()
2. iteration=1
3. **while** iteration < *MAXITER* **do**
4. Local (*LSITER*, δ)
5. **F**= CalcF ()
6. Move (**F**)
7. iteration= iteration +1
8.**end while**

شکل ۱. طرح کلی الگوریتم الکترومغناطیس.



شکل ۲. یک مثال از اثر جاذبه ـ دافعه بر روی ذره x_3.

بارهاست. بار هر ذره‌ی i، قابلیت تعیین قدرت جذب یا دفع ذره‌ی iام را (q^i) دارد. این بار به‌صورت رابطه‌ی ۳ محاسبه می‌شود:

$$q^i = \exp\left(-n\frac{f(x^i)-f(x^{\text{best}})}{\sum_{j=1}^{\text{popsize}}(f(x^j)-f(x^{\text{best}}))}\right), \quad i = 1, 2, \ldots, \text{popsize} \qquad (3)$$

که در آن $f(x^i)$، $f(x^j)$ و $f(x^{\text{best}})$ به‌ترتیب مقدار هدف ذره‌ی iام، مقدار هدف ذره‌ی j و بهترین جواب جمعیت است. با استفاده از رابطه‌ی ۳، نقاط با تابع هدف کم‌تر مقدار بار بیشتری دارند. با توجه به این که در ابعاد بالا تعداد ذرات در جمعیت تمایل به بزرگ شدن دارند، این رابطه در متغیر n ضرب می‌شود. بعد از مقایسه‌ی تابع هدف ذرات، جهت نیروی خاص بین دو ذره تعیین می‌شود. سرانجام، نیروی کل F^i وارده بر ذره‌ی i طبق رابطه‌ی ۴ محاسبه می‌شود.

$$F^i = \sum_{j\neq i}^{\text{popsize}} \begin{cases} (x^j - x^i)\frac{q^i q^j}{\|x^j - x^i\|^2} & \text{if } f(x^j) < f(x^i) \\ (x^i - x^j)\frac{q^i q^j}{\|x^j - x^i\|^2} & \text{if } f(x^j) \geq f(x^i) \end{cases}, \quad \forall i \qquad (4)$$

یک مثال دوبعدی برای سه ذره از مرحله‌ی سوم فرایند در شکل ۲ به تصویر کشیده شده است. نیروی وارد بر x_3 توسط x_1، F_{13} است. اگر مقدار تابع هدف x_3 مطلوب تر از x_1 باشد، x_3، x_1 را دفع می‌کند. نیروی وارد بر x_3 توسط x_2، F_{23} است. اگر مقدار تابع هدف x_3 بدتر از x_2 باشد، x_2، x_3 را جذب می‌کند. در نتیجه کل نیروی وارده بر x_3 برابر F است.

بعد از محاسبه‌ی کل نیروی F یک ذره، مرحله‌ی نهایی (مرحله‌ی چهارم) به حرکت در طول جهت نیرو اختصاص دارد. ذره‌ی i در جهت نیرو با یک طول گام تصادفی با استفاده از معادله‌ی ۵ حرکت می‌کند. طول گام تصادفی λ فرض می‌شود که از توزیع یکنواخت بین بازه صفر و ۱ تولید شده است. به‌منظور حفظ موجه بودن، نیروی اعمال شده روی هر ذره، نرمالایز می‌شود.

$$x^i = x^i + \lambda\frac{F^i}{\|F^i\|}(\text{RNG}) \qquad i = 1, 2, \ldots, \text{popsize} \qquad (5)$$

مرحله‌ی دوم تا چهارم آنقدر تکرار می‌شود تا الگوریتم به یک معیار توقف برسد. در مطالعه‌ی ما، الگوریتم الکترومغناطیس با استفاده از یک حداکثر تکرار متوقف می‌شود.

۴. الگوریتم ترکیبی K-EM

مطالعات انجام شده در سال‌های اخیر توسط محققین و همچنین کاربردهای مختلف الگوریتم الکترومغناطیس در زمینه‌های مختلف، بر این نکته تأکید می‌کند که می‌توان این الگوریتم را به‌عنوان یک الگوریتم قدرت‌مند در نظر گرفت. از طرف دیگر، الگوریتم K-means به‌دلیل برآورد کم‌تر تابع نسبت به الگوریتم الکترومغناطیس بسیار زودتر همگرا می‌شود و بیشتر مواقع دارای نتایج خوشه‌بندی با دقت پایین است. بنابراین می‌توان با ترکیب این دو الگوریتم از مزایای هر دو آن‌ها بهره‌مند شد.

چنان که در قسمت قبل ذکر شد، الگوریتم الکترومغناطیس یک الگوریتم مبتنی بر جمعیت است و کیفیت جواب به دست آمده توسط آن به جمعیت اولیه وابسته است. بنابراین، می‌توان کیفیت جواب به دست آمده توسط این الگوریتم را با جواب‌های اولیه خوب بالا برد. از این رو، در این مطالعه ما از جواب‌های اولیه تولید شده توسط K-means به‌عنوان جواب کاندید برای الگوریتم الکترومغناطیس استفاده می‌کنیم. هدف اصلی الگوریتم پیشنهادی این است که از مزایای هر دو روش برای بهبود جواب نهایی استفاده کند.

با لحاظ‌کردن توأمان همه‌ی موارد یادشده، الگوریتم ترکیبی K-EM را می‌توان به دو فاز اصلی تقسیم کرد: در فاز اول الگوریتم K-means به تعداد اندازه ذرات اجرا شده، زیرا ممکن است در مجموعه داده‌های بزرگ به جواب بهینه‌ی محلی همگرا شود؛ و همچنین با هر بار اجرای الگوریتم K-means برای یک مجموعه داده ممکن است به جواب‌های متفاوت برسیم. بنابراین برای اجتناب از جواب‌های بهینه‌ی محلی تولید شده K-means، آن را در فاز اول به تعداد اندازه ذرات اجرا می‌کنیم تا بهترین جمعیت ابتدایی را تولید کند. در فاز دوم بهترین جواب تولید شده در فاز اول برای تعیین یک جواب بهینه برای مسائل خوشه‌بندی توسط الگوریتم الکترومغناطیس به کارگرفته می‌شود. گام‌های اصلی الگوریتم پیشنهادی عبارت است از:

ـ فاز ۱. اجرای الگوریتم و تولید جمعیت اولیه

۱. انتخاب تصادفی K نقطه به‌عنوان مراکز ثقل اولیه؛
۲. تخصیص هر نقطه به نزدیک‌ترین مرکز ثقل؛
۳. به‌روزرسانی مکان هر مرکز ثقل به‌وسیله‌ی محاسبه‌ی میانگین مقدار نقاط تخصیص داده شده به مرکز؛
۴. تکرار مراحل ۲ و ۳ تا زمانی که شرایط توقف ارضا شود، یعنی بیشترین تعداد تکرارها حاصل شود یا تغییری در محل مراکز ثقل حاصل نشود؛
۵. استفاده از خروجی فاز اول به‌عنوان جواب کاندید برای فاز دوم.

ـ فاز ۲. اجرای الگوریتم الکترومغناطیس

```
ALGORITHM 2. Initialize()
    1:Best solution ()
    2: for i = 1 to popsize do
    3:        Use K-means algorithm for x^i
    4:        f(x^i)=evaluate (x^i)
    5: end for
    6: Best solution= argmin {f(x^i), ∀ i}
ALGORITHM 3. Local(LSITER, δ)
    1: counter=1
    2: Length= δ(max_k) {u_k - l_k}
    3: for i = 1 to popsize do
    4:        for k = 1 to n do
    5:             λ=U(0, 1)
    6:             while counter < LSITER do
    7:                y= x^i
    8:                λ_2=U(0, 1)
    9:                if λ_1<0.5 then
    10:                    y_k=y_k+ λ_2
    11:               else
    12:                    y_k=y_k - λ_2
    13:               end if
    14:               if f (y) <f(x^i) then
    15:                   x^i = y
    16:                   counter = LSITER−1
    17:               end if
    18:               counter = counter +1
    19:            end while
    20:       end for
    21: end for
    22: x^best = argmin {f(x^i), ∀ i}
ALGORITHM 4. CalcF():F
    1: for i = 1 to popsize do
    2:        calcute (q^i)
    3:        F^i = 0
    4: end for
    5: for i = 1 to popsize do
    6:    for j = 1 to popsize do
    7:          if f(x^j) <f(x^i) then
    8:        F^i=F^i+(x^j−x^i) q^i q^j / ||x^j−x^i||   {Atraction}
    9:          else
    10:       F^i=F^i−(x^j−x^i) q^i q^j / ||x^j−x^i||   {Repulsion}
    11:         end if
    12:   end for
    13: end for
ALGORITHM 5. Move(F)
    1: for i = 1 to popsize do
    2:    if i≠1best then
    3:        λ=U(0, 1)
    4:        F^i = F^i / ||F^i||
    5:          for k=1 to n do
    6:             if F_k^i > 0 then
    7:                z_k=x_k^i +λ F_k^i (u_k− x_k^i)
    8:             else
    9:                z_k=x_k^i +λ F_k^i (x_k^i − l_k)
    10:            end if
    11:         end for
    12:     end if
    13: end for
```

شکل ۳. شبه کد الگوریتم پیشنهادی.

۱. محاسبه‌ی مقدار برازندگی برای همه‌ی جواب‌های کاندید؛

۲. محاسبه‌ی نیروی کل وارد بر هر ذره براساس روابط ۳ و ۴؛

۳. حرکت ذرات با توجه به نیروی برآیند وارد بر آنها براساس رابطه‌ی ۵؛

۴. تکرار مراحل ۲ و ۳ تا زمانی که شرایط توقف ارضا شود.

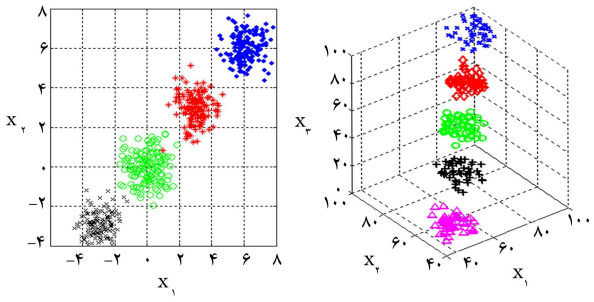
شبه کد الگوریتم ترکیبی K-EM در شکل ۳ نشان داده شده است.

۵. نتایج محاسباتی

در این قسمت، عملکرد الگوریتم ترکیبی K-EM با چندین الگوریتم مشهور که در ادبیات اخیر گزارش شده است، شامل K-means، ژنتیک، شبیه‌سازی تبرید، بهینه‌سازی کلونی مورچگان، بهینه‌سازی توده ذرات و الکترومغناطیس، مقایسه می‌شود.[۲۰] مقایسه‌ی نتایج برای هر مجموعه داده براساس بهترین جواب پیدا شده در بیش از ۳۰ شبیه‌سازی متفاوت برای هر الگوریتم است. عملکرد الگوریتم‌ها براساس معیار مجموعه فواصل درون خوشه‌یی (مطابق رابطه‌ی ۲) مقایسه می‌شود. این معیار برابر است با فاصله‌ی بین هر داده و مرکز خوشه‌ی مربوطه‌اش. واضح است اگر مجموع فاصله‌ی درون خوشه‌یی کم باشد، کیفیت بالاتر خوشه‌بندی به دست می‌آید. نتایج از لحاظ بهترین جواب، میانگین جواب و انحراف استاندارد داده شده است.

جدول ۱. مقادیر پارامترهای الگوریتم پیشنهادی.

پارامتر	Popsize	MAXITER	LSITER	δ
مقدار	۵	۱۰۰۰	۵	۰٫۰۵



شکل ۴. شمای مجموعه داده‌های مصنوعی.

پارامترهای اصلی در الگوریتم K-EM عبارت است از: بیشترین تعداد تکرار (MAXITER)، اندازه ذرات (popsize)، تعداد جست‌وجوی محلی (LSITER) و پارامتر جست‌وجوی محلی (δ). مقدار هریک از این پارامترها در جدول ۱ ارائه شده است. در این مطالعه، از نرم افزار MATLAB برای کدکردن الگوریتم پیشنهادی استفاده شده که برای این کار از یک لپ‌تاپ با پردازنده‌ی ۲ GHz به همراه ۶ GB رم استفاده شده است.

برای ارزیابی روش پیشنهادی، ما از ۵ مجموعه داده استفاده کرده‌ایم. از مجموعه‌داده‌های پیشنهادی، ۲ مجموعه‌داده مصنوعی است (شکل ۴).[۱۹] سایر مجموعه‌داده‌های در نظرگرفته شده واقعی‌اند که از ماشین آزمایشی آزمایشگاه گرفته شده است.[۳۲] ویژگی این مجموعه داده‌ها در جدول ۲ خلاصه شده است. این مجموعه داده‌ها از نظر تعداد اشیا، تعداد خوشه‌ها و تعداد شاخص‌ها متفاوت‌اند.

نتایج شبیه‌سازی به دست آمده برای مقایسه‌ی روش پیشنهادی K-EM با الگوریتم‌های مختلف موجود در ادبیات، روی مجموعه داده‌های مختلف در جدول ۳ نشان داده شده است. اعداد پررنگ شده نشان‌دهنده‌ی مقادیر با تابع هدف بهتر هستند. با توجه به مقادیر به دست آمده در این جدول، مشاهده می‌شود که مقدار تابع هدف الگوریتم پیشنهادی K-EM روی مجموعه داده Artset۱ برابر ۵۱۵٫۸۷ است. الگوریتم پیشنهادی روی مجموعه داده‌ی Artset ۲، بهترین نتیجه را مانند الگوریتم‌های GA، SA، ACO، PSO و EM به دست آورده است که مقداری معادل ۱۷۴۳٫۲۰ است. خروجی الگوریتم پیشنهادی روی مجموعه داده Iris تقریبا برابر و نزدیک به عدد ۹۶٫۵۴ است، که به‌طور قابل توجهی از سایر الگوریتم‌ها بهتر است. نتایج به دست آمده روی مجموعه داده‌ی Wine نشان می‌دهد که الگوریتم پیشنهادی به بهینه‌ی کلی ۱۶۲۹۲٫۲۹ همگرا شده است. در حالی که بهترین جواب برای سایر الگوریتم‌های K-means، GA، SA، ACO، PSO و

جدول ۲. ویژگی مجموعه داده‌ها.

مجموعه داده‌ها	تعداد خوشه‌ها	تعداد شاخص‌ها	اندازه‌ی مجموعه داده
Artset۱	۴	۲	۲۵۰
Artset۲	۵	۳	۶۰۰
Iris	۳	۳	۱۵۰
Wine	۳	۱۳	۱۷۸
CMC	۳	۱۰	۱۴۷۳

EM برابر ۱۶۵۵۵/۶۸، ۱۶۵۳۰/۵۳، ۱۶۴۷۳/۴۸، ۱۶۵۳۰/۵۳، ۱۶۳۴۵/۹۶ و ۱۶۲۹۵/۳۱ است. در نهایت، الگوریتم پیشنهادی روی مجموعه داده CMC به بهینه‌ی کلی ۵۶۹۳/۸۲ رسیده است.

این در حالی است که سایر الگوریتم‌ها حتی با بیش از ۳۰ اجرا نتوانستند به این مقدار دست یابند.

به‌طور خلاصه، با توجه به نتایج به دست آمده، می‌توان نتیجه گرفت که الگوریتم پیشنهادی K-EM دقیق و قابل اعتماد است و می‌تواند جواب‌های با کیفیت را با انحراف استاندارد پایین به دست بیاورد، در حالی که سایر الگوریتم‌ها ممکن است به بهینه‌ی محلی گرفتار شوند. از طرف دیگر باعث می‌شود که الگوریتم K-means به بهینه‌ی محلی گرفتار نشود و همچنین باعث افزایش کیفیت جواب به دست‌آمده نسبت به الگوریتم K-means می‌شود.

۶. نتیجه‌گیری

خوشه‌بندی یکی از روش‌های پرکاربرد است که در بسیاری از زمینه‌های علمی مورد توجه محققین قرارگرفته است. در سال‌های اخیر، پژوهش‌های متعددی برای توسعه‌ی الگوریتم‌ها و یافتن جواب بهینه در حل مسائل خوشه‌بندی صورت گرفته است. این توسعه گاهی با ترکیب الگوریتم‌ها با یکدیگر و گاهی با ابداع روشی جدید میسر شده است. الگوریتم K-means، یکی از کاربردی‌ترین روش خوشه‌بندی داده‌هاست که از مزایایی همچون سرعت و ساده بودن در اجرا برخوردار است. با این حال، به‌دلیل وابسته بودن به موقعیت اولیه‌اش اغلب به بهینه‌ی محلی همگرا می‌شود. در این مطالعه یک الگوریتم ترکیبی براساس الگوریتم‌های K-means و الکترومغناطیس برای حل مسائل خوشه‌بندی توسعه داده شده است. عملکرد الگوریتم پیشنهادی روی چندین مجموعه داده اجرا و با دیگر روش‌های موجود درادبیات مقایسه می‌شود. نتایج محاسبات نشان می‌دهد که الگوریتم پیشنهادی از نظر کیفیت جواب پیدا شده نسبت به دیگر روش‌ها بهتر است. در تحقیقات آینده، الگوریتم پیشنهادی می‌تواند در زمینه‌های دیگری مورد استفاده قرارگیرد. همچنین می‌توان با ترکیب این الگوریتم با سایر الگوریتم‌ها کارایی آن را افزایش داد.

جدول ۳. مقایسه مقادیر الگوریتم‌های مختلف.

مجموعه داده	معیار	الگوریتم						
		K-means	GA	SA	ACO	PSO	EM	K-EM
Artset۱	بهترین	۵۱۶/۰۴	۵۱۸/۰۹	۵۱۸/۹۵	۵۱۷/۸۷	۵۱۵/۹۳	۵۱۵/۹۵	**۵۱۵/۸۷**
	انحراف استاندارد	۲۹۵/۸۴	۱۸۹/۸۶	۱۹۵/۱۵	۲/۰۱	۱۸۰/۲۴	۰/۰۴	۰/۰۰
	میانگین	۷۲۱/۵۷	۶۳۸/۰۹	۶۸۴/۶۸	۵۱۹/۸۸	۶۲۷/۷۴	۵۱۶/۰۱	۵۱۵/۸۷
Artset۲	بهترین	۱۷۴۶/۹	**۱۷۴۳/۲۰**	**۱۷۴۳/۲۰**	**۱۷۴۳/۲۰**	**۱۷۴۳/۲۰**	**۱۷۴۳/۲۰**	**۱۷۴۳/۲۰**
	انحراف استاندارد	۷۲۰/۶۶	۴۳۷/۰۵	۴۲۹/۰۲	۱۳۴/۰۶	۴۱۵/۰۲	۳۹/۶۹	۲/۰۱
	میانگین	۲۷۶۲/۰۳	۲۶۶۷/۳۰	۲۶۸۶/۸۴	۱۹۴۸/۹۷	۲۵۱۷/۲۰	۱۷۴۹/۱۵	۱۷۴۵/۴۰
Iris	بهترین	۹۷/۳۳۳	۱۱۳/۹۸	۹۷/۴۵	۹۷/۱۰	۹۶/۸۹	۹۶/۵۸	**۹۶/۵۴**
	انحراف استاندارد	۱۴/۶۳۱	۱۴/۵۶	۲/۰۱	۰/۳۶	۰/۳۴	۰/۰۱۴	۰/۰۱
	میانگین	۱۰۶/۰۵	۱۲۵/۱۹	۹۹/۹۵	۹۷/۱۷	۹۷/۲۳	۹۶/۶۱	۹۶/۵۸
Wine	بهترین	۱۶۵۵۵/۶۸	۱۶۵۳۰/۵۳	۱۶۴۷۳/۴۸	۱۶۵۳۰/۵۳	۱۶۳۴۵/۹۶	۱۶۲۹۵/۳۱	**۱۶۲۹۲/۲۹**
	انحراف استاندارد	۷۹۳/۲۱	۰/۰۰	۷۵۳/۰۸	۰/۰۰	۸۵/۴۹	۰/۰۲۱	۰/۰۱
	میانگین	۱۸۶۰۱/۰۰	۱۶۵۳۰/۵۳	۱۷۵۲۱/۰۹	۱۶۵۳۰/۵۳	۱۶۴۱۷/۴۷	۱۶۲۹۸/۵۰	۱۶۲۹۲/۳۱
CMC	بهترین	۵۸۴۲/۲۰	۵۷۰۵/۶۳	۵۸۴۹/۰۳	۵۷۰۱/۹۲	۵۷۰۰/۹۸	۵۶۹۵/۳۸	**۵۶۹۳/۸۲**
	انحراف استاندارد	۴۷/۱۶	۵۰/۳۶	۵۰/۸۶	۴۵/۶۳	۴۶/۹۵	۰/۶۶۷	۲/۱۲
	میانگین	۵۸۹۳/۶۰	۵۷۵۶/۵۹	۵۸۹۳/۴۸	۵۸۱۹/۱۳	۵۸۲۰/۹۶	۵۶۹۶/۰۴	۵۶۸۳/۸۲

پانوشت‌ها

1. tabu search (TS)
2. ant colony optimization (ACO)
3. honey bee mating optimization (HBMO)
4. genetic algorithm (GA)
5. regression expectation maximization (REM)
6. particle swarm optimization (PSO)
7. artificial bee colony (ABC)
8. chaotic particle swarm optimization (CPSO)
9. modified imperialist competitive algorithm (MIKA)
10. binary search
11. black hole
12. popsize

منابع (References)

1. Xia, Y., Feng, D., Wang, T., Zhao, R. and Zhang, Y. "Image segmentation by clustering of spatial patterns", *Pattern Recognition Lett,* **28**, pp. 1548-1555 (2007).
2. Friedman, M., Last, M., Makover, Y. and Kandel, A. "Anomaly detection in web documents using crisp and fuzzy-based cosine clustering methodology", *Information Sciences*, **177**, pp. 467-475 (2007).
3. Liao, L., Lin, T. and Li, B. "MRI brain image segmentation and bias field correction based on fast spatially constrained kernel clustering approach" , *Pattern Recognition Letters*, **29**, pp. 1580-1588 (2008).
4. Cheng, Y.M. and Leu, S.S. "Constraint-based clustering and its applications in construction management", *Expert Systems with Applications*, **36**, pp. 5761-5767 (2009).
5. Kim, K.J. and Ahn, H. "A recommender system using GA K-means clustering in an online shopping market", *Expert Systems with Applications*, **34**, pp. 1200-1209 (2008).
6. Taboada, H.A. and Coit, D.W. "Data clustering of solutions for multiple objective system reliability optimization problems", *Quality Technology and Quantitative Management*, **4**, pp. 191-210 (2007).
7. Che, Z.H. "Clustering and selecting suppliers based on simulated annealing algorithms", *Computers and Mathematics with Applications,* **63**, pp. 228-238 (2012).
8. Po, R.W., Guh, Y.Y. and Yang, M.Sh. "A new clustering approach using data envelopment analysis", *European Journal of Operational Research*, **199**, pp. 276-284 (2009).
9. Hartigan, J.A., *Clustering Algorithms*, Wiley, New York (1975).
10. McLachlan, G.J. and Basford, K.E., *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker, New York (1988).
11. Kohonen, T., *Self-Organizing Maps*, third ed. Springer-Verlag, Berlin (2001).
12. Bezdek, J.C., *Pattern Recognition with Fuzzy Objective Function Algorithm*, Plenum Press, New York (1981).
13. Duda, R.O. and Hart, P.E., *Pattern Classification and Scene Analysis*, Wiley, New York (1973).
14. Ray, S. and Turi, R.H. "Determination of number of clusters in K-means clustering and application in colour image segmentation", *Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques (ICAPRDT'99)*, Calcutta, India, pp. 137-143 (1999).
15. Al-Sultan, K.S. "A tabu search approach to the clustering problem", *Pattern Recognition*, **28**, pp. 1443-1451 (1995).
16. Shelokar, P.S., Jayaraman, V.K. and Kulkarni, B.D. "An ant colony approach for clustering", *Analytica Chimica Acta*, **509**, pp. 187-195 (2004).
17. Fathian, M., Amiri, B. and Maroosi, A. "Application of honey-bee mating optimization algorithm on clustering", *Applied Mathematics and Computation*, **190**, pp. 1502-1513 (2007).
18. Cao, D.N. and Krzysztof, J.C. "GAKREM: a novel hybrid clustering algorithm", *Information Sciences,* **178**, pp. 4205-4227 (2008).
19. Kao, Y.T., Zahara, E. and Kao, I.W. "A hybridized approach to data clustering", *Expert Systems with Applications*, **34**, pp. 1754-1762 (2008).
20. Niknam, T. and Amiri, B. "An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis", *Appl. Soft Comput,* **10**, pp. 183-197 (2010).
21. Zhang, Ch., Ouyang, D. and Ning, J. "An artificial bee colony approach for clustering", *Expert Systems with Applications,* **37**, pp. 4761-4767 (2010).
22. Chuang, L.Y., Hsiao, C.J. and Yang, C.H. "Chaotic particle swarm optimization for data clustering", *Expert Systems with Applications*, **38**, pp. 14555-14563 (2011).
23. Niknam, T., TaherianFard, E., Pourjafarian, N. and Rousta, A. "An efficient hybrid algorithm based on modified imperialist competitive algorithm and K-means for data clustering", *Engineering Applications of Artificial Intelligence*, **24**, pp. 306-317 (2011).
24. Hatamlou, A. "In search of optimal centroids on data clustering using a binary search algorithm", *Pattern Recognition Letters,* **33**, pp. 1756-1760 (2012).
25. Hatamlou, A. "Black hole: A new heuristic optimization approach for data clustering", *Information Sciences*, **222**, pp. 175-184 (2013).
26. Birbil, S.I. and Fang, S.C. "An electromagnetism-like mechanism for global optimization", *Journal of Global Optimization,* **25**, pp. 263-282 (2003).
27. Birbil, S.I., Fang, S.-C. and Sheu, R.-L. "On the convergence of a population-based global optimization algorithm", *Journal of Global Optimization*, **30**, pp. 301-318 (2004).

28. Jamili, A., Shafia, M.A. and Tavakkoli-Moghaddam, R. "A hybridization of simulated annealing and electromagnetism-like mechanism for a periodic job shop scheduling problem", *Expert Systems with Applications*, **38**, pp. 5895-5901 (2011).

29. Yurtkuran, A. and Emel, E. "A new hybrid electromagnetism-like algorithm for capacitated vehicle routing problems", *Expert Systems with Applications*, **37**, pp. 3427-3433 (2010).

30. Zhang, C., Li, X., Gao, L. and Wu, Q. "An improved electromagnetism-like mechanism algorithm for constrained optimization", *Expert Systems with Applications*, **40**, pp. 5621-5634 (2013).

31. Guan, X., Dai, X., Qiu, B. and Li, J. "A revised electromagnetism-like mechanism for layout design of reconfigurable manufacturing system", *Computers & Industrial Engineering*, **63**, pp. 98-108 (2012).

32. Newman, D.J., Hettich, S., Blake, C.L Merz, C.J. UCI Repository of Machine Learning Databases. http://www.ics.uci.edu/-mlearn/MLRepository.html. Irvine, CA: University of California, Department of information and computer science (1998).