خوشه‌بندی ترکیبی با استفاده از یک فضای ویژگی جدید

منیره عبدوس[۱]، جواد عظیمی[۲]، علیرضا صابری[۳]، مرتضی آنالویی[۴]

چکیده

خوشه‌بندی ترکیبی عبارتست از ترکیب نتایج چندین الگوریتم خوشه‌بندی برای به دست آوردن خوشه‌هایی با دقت بالاتر. خوشه‌بندی ترکیبی با چندین بار اجرای یک الگوریتم در حالت های مختلف می‌تواند نتایج بهتری چه از لحاظ استحکام و چه از لحاظ پایداری و انعطاف پذیری تولید کند. در این مقاله، روشی برای خوشه‌بندی ترکیبی بر مبنای ایجاد فضای ویژگی جدید ارائه شده‌است. در این روش از نتایج الگوریتم های خوشه‌بندی پایه جهت ایجاد ویژگی های جدید استفاده کرده‌ایم. نتایج الگوریتم های پایه با گراف کامل وزن دار مدل سازی شده‌اند. روشی حریصانه برای پیمایش گراف و ایجاد درخت، جهت تعیین مقادیر ویژگی ها معرفی شده‌است. ویژگی های به دست آمده، خصوصیات بهتری نسبت به ویژگی های اصلی دارند، که نمونه‌های هر خوشه را نسبت به یکدیگر به خوبی متمایز می‌سازد. در این مقاله به بررسی روش ارائه شده بر روی چهار مجموعه داده Iris، Wine، Thyroid و Soybean پرداخته شده است. بررسی‌های تجربی نشان می‌دهند روش مذکور به سرعت همگراست و با افزایش تعداد تکرار الگوریتم پایه رفتار مناسبی از خود نشان می‌دهد.

کلمات کلیدی

خوشه بندی ترکیبی، الگوریتم خوشه‌بندی پایه، گراف کامل وزن دار، پیمایش گراف، فضای ویژگی جدید.

Clustering Ensemble Using a Novel Feature Space

Monireh Abdoos, Javad Azimi, Alireza Saberi, Morteza Analoui
Computer Engineering Department- Iran University of Science and Technology

Abstract

Clustering Ensemble is combination of the results of multiple clustering algorithms to achieve more accurate clusters. Clustering Ensembles have emerged as a powerful method for improving both the robustness and the stability of the results. In this paper, we propose a new method based on generating a new feature space. In this method, new features are created by using the output results of the initial clustering algorithms. The results of the initial clustering algorithms are modeled by a complete weighted graph. New feature space is created by applying a greedy method that is introduced for graph spanning and tree generating. The new feature space has a property that differentiates the samples of each cluster. In this paper, we study the results of the proposed method on Iris, Wine, Thyroid and Soybean. Experimental results show the fast convergency and good behavior of the proposed method by increasing the ensemble size.

Keywords

Clustering ensembles, Initial clustering algorithm, complete weighted graph, graph spanning, new feature space.

[1] دانشجوی کارشناسی ارشد، دانشگاه علم و صنعت ایران، abdoos@mail.iust.ac.ir
[2] دانشجوی کارشناسی ارشد، دانشگاه علم و صنعت ایران، ja_azimi@comp.iust.ac.ir
[3] دانشجوی کارشناسی ارشد، دانشگاه علم و صنعت ایران، a_saberi@comp.iust.ac.ir
[4] استادیار دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، analoui@iust.ac.ir

۱- مقدمه

خوشه بندی ترکیبی عبارتست از ترکیب نتایج چندین الگوریتم خوشه‌بندی برای به دست آوردن خوشه‌هایی با دقت بالاتر .

خوشه بندی ترکیبی با چندین بار اجرای یک الگوریتم در حالت‌های مختلف می‌تواند نتایج بهتری چه از لحاظ استحکام و چه از لحاظ پایداری و انعطاف پذیری تولید کند [1,2,3] . به طور خلاصه خوشه‌بندی ترکیبی شامل تولید زیر مجموعه های متفاوت از کل نمونه‌های داده شده، خوشه‌بندی کردن بر اساس اعمال الگوریتم های متفاوت خوشه بندی بر روی زیر مجموعه های ایجاد شده از نمونه‌های اصلی و در نهایت ترکیب نتایج به دست آمده از خوشه‌بندی های متفاوت برای تولید خوشه‌بندی نهایی است. به طور کلی دو مسأله مهم در خوشه‌بندی ترکیبی وجود دارد:

۱- تنوع الگوریتم های خوشه‌بندی مختلف به طوریکه هر کدام از خوشه‌بندی ها بر ویژگی های خاصی از داده‌ها تأکید کنند.

۲- الگوریتم ترکیب کننده نتایج برای تولید خوشه‌های نهایی.

برای اولین مسأله، یعنی به دست آوردن نتایج متنوع که هر کدام بر ویژگی خاصی از داده‌ها تاکید کند می‌توان از روشهای زیر استفاده نمود:

۳- استفاده از الگوریتم های متفاوت خوشه بندی [4].

۴- تغییر مقادیر اولیه و یا سایر پارامترهای الگوریتم خوشه‌بندی انتخاب شده [3,5].

۵- انتخاب بعضی از ویژگی داده‌ها [1,6,7].

۶- تقسیم بندی داده‌های اصلی به زیر مجموعه‌هایی متفاوت و مجزا [8,9,10,11,12,13].

مسأله اصلی در خوشه‌بندی ترکیبی، انتخاب تابع ترکیب یا الگوریتم ترکیب خوشه‌بندی های مختلف برای ایجاد خوشه‌بندی نهایی می‌باشد. ترکیب خوشه‌بندی های متفاوت را می توان به عنوان یافتن یک خوشه‌بندی میانگین از خوشه‌بندی های موجود در نظر گرفت که از جمله مسائل NP-Complete محسوب می‌شود [14].

تاکنون توابع ترکیب متفاوتی ارائه شده‌است: تقسیم بندی ابرگراف [1,6] ، روش رأی‌گیری[5,8,15] و روشهای مبتنی بر همبستگی [2,16,17].

در این مقاله، روشی برای خوشه‌بندی ترکیبی ارائه شده‌است. روش ارائه شده یک فضای جدید ویژگی برای نمونه‌ها ایجاد می‌نماید و از آن فضای ویژگی جدید برای خوشه‌بندی نهایی استفاده می‌کند.

اغلب روشهای خوشه‌بندی ترکیبی، الگوریتم k-means را به عنوان الگوریتم خوشه‌بندی پایه به کار می‌برند. خوشه‌هایی که با به کارگیری الگوریتم k-means به دست می‌آیند، وابستگی زیادی به انتخاب مراکز اولیه خوشه‌ها دارند [18,19]. تاکنون روش های زیادی جهت انتخاب هوشمندانه مراکز اولیه خوشه‌ها، پیشنهاد شده‌است [18,20,21]. این روشها، تمام فضای ویژگی را جهت انتخاب هوشمندانه مراکز اولیه خوشه‌ها، مورد بررسی قرار می‌دهند. در این مقاله روشی هوشمند بر مبنای k-means ارائه شده‌است. این روش صرفاً برای خوشه‌بندی ترکیبی معرفی و در این مقاله به عنوان الگوریتم پایه خوشه‌بندی استفاده شده‌است.

روش خوشه‌بندی ارائه شده در این مقاله دارای چهار مرحله زیر می‌باشد:

۱- اجرای روش هوشمندانه k-means به عنوان الگوریتم پایه.

۲- ایجاد گراف کامل طبق خروجی الگوریتم پایه.

۳- ایجاد فضای ویژگی جدید با استفاده از پیمایش گراف.

۴- اجرای k-means نهایی بر روی داده ها در فضای ویژگی جدید جهت به دست آوردن خوشه‌های نهایی.

در این مقاله به بررسی و مقایسه روش ارائه شده و سه روش CSPA[1]، HGPA[2] و CAL[3] پرداخته شده‌است.

تقسیم بندی ابرگراف: در این روش، خوشه‌ها با ابرلبه های یک گراف نمایش داده می‌شوند. رأسهای گراف معادل نمونه‌هایی هستند که باید خوشه‌بندی شوند. دو روش از این زیر مجموعه به نامهای CSPA و HGPA مورد مقایسه قرار گرفته است، این روشها در [1] شرح داده شده‌اند.

توابع ترکیب مبتنی بر ماتریس همبستگی: فرض کنید مجموعه داده شامل N نمونه هر کدام دارای d بعد می باشد فرض کنید $S=\{S_1, \ldots, S_{B1}\}$ مجموعهٔ زیر مجموعه نمونه‌های ماست که از نمونه‌های اولیه استخراج شده‌اند. هر یک از الگوریتم های انتخابی هنگامی که بر روی زیر مجموعه نمونه‌های موجود در S اجرا شوند نتایج $P=\{P_1,\ldots, P_{B1}\}$ را تولید می‌کنند. هر P_i مجموعه ای از خوشه‌هاست یا به عبارت دیگر $P_i=\{C_1^i, C_2^i, \ldots C_{k(i)}^i\}$ و $X_i = C_1^i \cup C_2^i \ldots \cup C_{k(i)}^i$ به طوریکه $k(i)$ تعداد خوشه‌ها در i-امین خوشه بندی می‌باشد. در اولین گام الگوریتم k-means را بر روی $X=\{X_1 \ldots X_{B1}\}$ اجرا می‌کنیم تا بتوانیم با استفاده از P_i های تولید شده ماتریس همبستگی را به صورت زیر به دست آوریم:

$$Co\text{-}association(x,y)=\sum_{i=1}^{B1} l(P_i(x),P_i(y)) \quad (۱)$$

$$l(a,b)=\begin{cases}1 \text{ if } a=b\\ 0 \text{ if } a\neq b\end{cases} \quad (۲)$$

تابع $l(P_i(a),P_i(b))$ در صورتی که دو عنصر a و b در خوشه‌بندی P_i در یک خوشه قرار گرفته باشند مقدار ۱ و در غیر این صورت مقدار ۰ بر می‌گرداند . مقدار پارامتر B_1 نمایانگر تعداد زیر مجموعه ها می‌باشد و یا به بیان دیگر تعداد دفعات تکرار الگوریتم پایه k-means است . بعد از به دست آوردن ماتریس همبستگی با استفاده از الگوریتم ساده سلسله مراتبی نظیر AL(Average-Link)

اقدام به استخراج خوشه‌های نهائی از ماتریس همبستگی می‌گردد. این روش تحت عنوان CAL ، در این مقاله مورد مقایسه قرار گرفته است.

روش پیشنهادی دربخش ۲ معرفی شده‌است. در این بخش روش هوشمندانه k-means و مکانیزم ایجاد فضای ویژگی جدید معرفی می‌شوند. در بخش ۳، نتایج تجربی بر روی چهار مجموعه داده استاندارد Iris، Wine، Soybean و Thyroid آورده شده‌است. این مقاله در بخش ۴ نتیجه‌گیری شده‌است.

۲- روش پیشنهادی

در این مقاله، روشی برای خوشه‌بندی ترکیبی ارائه شده‌است. در این روش، الگوریتم هوشمندانه k-means به عنوان الگوریتم خوشه‌بندی پایه مورد استفاده قرار گرفته است. این روش در بخش بعد معرفی می‌شود. روش پیشنهادی با استفاده از نتایج الگوریتم های خوشه‌بندی پایه، فضای ویژگی جدیدی ایجاد می‌نماید که به عنوان ویژگی های اصلی جهت خوشه‌بندی نهایی مورد استفاده قرار می‌گیرند.

فرض کنید X، مجموعه داده شامل N نمونه باشد:

$$X = \{ x_1, x_2, \ldots, x_N \} \quad (۳)$$

اگر $p_j(x_i)$، خروجی متناظر با j-امین الگوریتم خوشه‌بندی پایه بر روی نمونه x_i باشد، داریم:

$$x_i \rightarrow \{p_1(x_i), p_2(x_i), \ldots, p_H(x_i)\} \quad (۴)$$

$p_j(x_i)$، $j=1, 2, \ldots, H$ و $i=1, 2, \ldots, N$ خروجی های حاصل از اجرای k-means هوشمند بر روی x_i می‌باشد.

هر بار اجرای الگوریتم خوشه‌بندی پایه متناظر با یک بعد در فضای ویژگی جدید خواهد بود، بنابراین فضای ویژگی جدید دارای H بعد می‌باشد. اگر مجموعه X شامل N نمونه با m ویژگی باشد، مجموعه جدید ایجادشده، X'، مجموعه‌ای متشکل از N نمونه با H ویژگی خواهد بود:

$$\begin{aligned} X &= \{ x_1, x_2, \ldots, x_N \} \\ x_i &= (x_{i1}, x_{i2}, \ldots, x_{im}) \\ i &= 1,2,\ldots,N \end{aligned} \Rightarrow \begin{aligned} X' &= \{ x'_1, x'_2, \ldots, x'_N \} \\ x'_i &= (x'_{i1}, x'_{i2}, \ldots, x'_{iH}) \end{aligned} \quad (۵)$$

مقادیر ویژگیها در فضای جدید با به کارگیری مکانیزمی تعیین می‌شود که در بخش ۲-۲ معرفی می‌گردد.

بعد از اینکه نمونه‌ها در فضای ویژگی جدید ایجاد شدند، خوشه‌بندی نهایی با به کارگیری ساده یکی از روش های خوشه‌بندی انجام می‌گیرد. در روش ارائه شده در این مقاله الگوریتم k-means جهت خوشه‌بندی نهایی استفاده شده‌است.

۲-۱- الگوریتم خوشه‌بندی پایه

بسیاری از روشهای خوشه‌بندی ترکیبی ارائه شده الگوریتم k-means را به عنوان روش خوشه‌بندی پایه مورد استفاده قرار می‌دهند. در الگوریتم k-means ابتدا k نقطه به طور تصادفی انتخاب می‌گردند و به عنوان مراکز اولیه خوشه‌ها به کار می‌روند. بررسی های انجام شده بر روی k-means نشان می‌دهند که نتایج حاصل از اجرای k-means وابستگی زیادی به انتخاب مراکز اولیه خوشه‌ها دارد[19,20]. تاکنون روش های زیادی مبتنی بر انتخاب هوشمندانه مراکز اولیه k-means ارائه شده‌است. این روش ها کل فضای ویژگی را جهت انتخاب هوشمندانه مراکز مورد بررسی قرار می‌دهند. در روش پیشنهادی ما، به طور ساده، از نتایج الگوریتم خوشه‌بندی قبلی، جهت انتخاب مراکز اولیه خوشه‌ها بهره می‌برد. این روش صرفاً در خوشه‌بندی ترکیبی قابل استفاده می‌باشد و به این صورت است:

در اولین اجرای k-means، مراکز اولیه به طور تصادفی انتخاب می‌گردند. در اجراهای بعدی k-means از نمونه‌های به دست آمده از هر خوشه، یک نمونه به طور تصادفی انتخاب می‌گردد و به عنوان مراکز اولیه خوشه مربوطه در نظر گرفته می‌شود. به عبارت دیگر، خروجی حاصل از اجرای j-امین k-means جهت انتخاب مراکز اولیه خوشه‌ها در $(j+1)$-امین k-means مورد استفاده قرار می‌گیرد.

این روش به سادگی قابل پیاده سازی است و تأثیری در پیچیدگی زمانی خوشه‌بندی ترکیبی نخواهد داشت.

۲-۲- تعیین گراف و ایجاد فضای ویژگی جدید

الگوریتم های خوشه‌بندی پایه به دو منظور انجام می‌گیرند، یکی تعیین مراکز خوشه‌ها و دیگری تعیین نمونه‌های هر خوشه.

مراکز خوشه‌ها جهت تعیین مقادیر ویژگی های جدید مورد استفاده قرار می‌گیرند[22]. به هر مرکز خوشه برچسبی[۴] تخصیص داده می‌شود و نمونه‌های هر خوشه مقداری معادل با برچسب مرکز همان خوشه می‌گیرند. به این ترتیب یک بعد در فضای ویژگی جدید ایجاد خواهد شد. تعیین مقدار برچسب های مراکز خوشه‌ها، با استفاده از تئوری گراف انجام می‌گیرد.

فرض کنید $G=<V,E>$، گراف کامل وزن دار باشد که V و E به ترتیب نشان دهنده رأسها و یالهای گراف می‌باشند. مراکز خوشه‌ها به عنوان رأسهای گراف، V، در نظر گرفته می‌شوند. خط واصل بین مراکز خوشه‌ها به عنوان یالهای گراف و فاصله اقلیدسی مراکز در فضای ویژگی مربوطه به عنوان وزن یال در نظر گرفته می‌شود.

جهت تعیین مقادیر برچسب ها، با شروع از یک رأس گراف به پیمایش گراف پرداخته و مقادیر برچسب ها حین پیمایش گراف تعیین می‌شوند. پیمایش گراف به صورتی انجام می‌شود که در آن هر رأس گراف فقط یک بار ملاقات شود. الگوریتم حریصانه[۵] زیر جهت پیمایش گراف و تعیین مقادیر برچسب ها به کار گرفته شده‌است.

۱- یکی از رئوس به عنوان رأس آغازین، v_i ، انتخاب می‌گردد.

۲- $label(v_i)=1$ (۶)

۳- از میان رأسهای مجاور v_i ، رأس v_j دارای یال با کمترین وزن انتخاب می‌گردد.

۴- $label(v_j)=label(v_i)+f(w(i,j))$ (۷)

که f تابعی از وزن یالها می‌باشد.

۵- رأس v_j به عنوان رأس مشاهده شده علامت می‌خورد.

۶- $v_j = v_i$.

۷- الگوریتم از مرحله ۳ تکرار می‌شود تا جائیکه همه رأسها مشاهده شوند.

الگوریتم فوق، مقادیر برچسبهای متناظر با هر یک از خوشه‌ها را تعیین می‌نماید. هر نمونه با توجه به اینکه متعلق به چه خوشه‌ای باشد، مقدار بر چسب متناظر با مرکز همان خوشه را به عنوان ویژگی جدید خود می‌پذیرد. این فرآیند پس از هر بار اجرای الگوریتم خوشه‌بندی پایه انجام می‌گیرد. بدین ترتیب با اجرای H مرتبه الگوریتم خوشه‌بندی پایه به H ویژگی جدید برای هر نمونه دست می‌یابیم.

شبه کد مربوط به پیمایش گراف و تعیین مقادیر برچسب ها در شکل (۱) آورده شده‌است.

```
Procedure Spanning_Labeling(V,E)
 {V: The graph vertices, a vector with n
     elements
  E: The graph edges, an n*n matrix whose
     elements show the weight of the
     corresponding edge
   }

Begin
 Current=min(distance(V(j),O));{j=1,2,…,n}
 Nodes=Current;
 Rest=V-Current;
 Label(Current)=1;
 While Rest<>[]
   Begin
    Next=min(E(Current, j));{j=1,2,…,n}
    Edge=Edge+E(Current, Next);
    Label(Next)=Label(Current)+
                f(E(Current, Next));
    Current=Next;
    Nodes=Nodes+Current;
    Rest=Rest-Current;
   End
End
```

شکل (۱): شبه کد مربوط به پیمایش و برچسب گذاری گراف

۳- مطالعات تجربی

در این بخش نتایج عملی اعمال الگوریتم پیشنهادی بررسی گردیده است. در این آزمایشات چهار مجموعه داده معروفUCI به نامهای Iris، Wine، Thyroid و Soybean بررسی گردیده است. خصوصیات این مجموعه داده ها در جدول(۱) آمده است.

جدول(۱): خلاصه ای از خصوصیات مجموعه داده ها

مجموعه داده	تعداد کلاسها	تعداد ویژگیها	تعداد نمونه‌ها	تعدادنمونه‌ها به ازای هر خوشه
Iris	۳	۴	۱۵۰	۵۰ -۵۰ -۵۰
Wine	۳	۱۳	۱۷۸	۴۸-۷۱-۵۹
Thyroid	۳	۵	۲۱۵	۳۰ -۳۵ -۱۵۰
Soybean	۴	۳۵	۴۷	۱۷ -۱۰ -۱۰ -۱۰

از آنجا که برای تمامی مجموعه داده‌ها، تعداد خوشه‌ها و نوع خوشه اختصاص یافته به هر نمونه مشخص شده‌است، می توان نرخ خطای خوشه بندی ترکیبی نهایی را به راحتی به دست آورد و کارایی روش استفاده شده را آزمود. از آنجا که الگوریتم k-means به عنوان الگوریتم خوشه‌بندی پایه مورد استفاده قرار گرفته است، لذا از داده‌های نرمال استفاده شده‌است.

میانگین خطای مجموعه داده های فوق با استفاده از روش پیشنهادی و سه روش HGPA، CSPA و CAL در جدول های (۲)- (۵) آورده شده‌است. نتایج ارائه شده میانگین حاصل از ۵۰ بار اجرای مستقل روش های فوق می‌باشد.

جدول (۲): مقایسه میانگین خطای روش های مختلف (٪) بر روی Iris

تعداد تکرار الگوریتم پایه(H)	روش CAL	روش CSPA	روش HGPA	روش پیشنهادی
۵	۱۲،۷	۶،۳۸	۱۹،۸۱	۸،۳۶
۱۰	۹،۹۷	۵،۲۳	۷،۹۷	۴
۱۵	۷،۷۳	۴،۳۲	۵،۰۵	۴
۲۰	۶،۱۷	۴،۲۳	۴	۴،۸۹
۲۵	۵،۰۳	۴،۳	۴	۴
۳۰	۵،۵۷	۴،۳	۴	۴
۳۵	۵،۰۷	۴،۳۳	۴	۴
۴۰	۵،۵۳	۴،۲۷	۴	۴
۴۵	۵،۶	۴،۴	۴	۴،۸۷
۵۰	۵،۵	۴،۵	۴	۴

جدول (۳): مقایسه میانگین خطای روش های مختلف (٪) بر روی Wine

تعداد تکرار الگوریتم پایه(H)	روش CAL	روش CSPA	روش HGPA	روش پیشنهادی
۵	۱۱،۷	۱۰،۶۵	۱۵،۹۸	۱۱،۸۷
۱۰	۱۳،۳	۹،۹۷	۱۰،۰۶	۹،۹۱
۱۵	۹،۱۹	۱۰،۰۳	۸،۵۷	۱۲،۴۹
۲۰	۱۱،۳	۱۰،۳۹	۷،۴۲	۸،۱۱
۲۵	۱۰،۶	۱۰،۳۷	۹،۰۴	۱۰،۹۷
۳۰	۴۰،۵	۱۰،۲۵	۸،۱۵	۱۰،۸۲
۳۵	۹،۹۷	۱۰،۵۳	۷،۳۹	۸،۶۹
۴۰	۱۰،۱	۱۰،۶۷	۸،۰۹	۹،۱۸
۴۵	۹،۶۵	۱۰،۵۱	۷،۸۴	۹،۴
۵۰	۹،۸۸	۱۰،۱۱	۷،۵۳	۹،۵۹

جدول (۴): مقایسه میانگین خطای روش های مختلف (٪) بر روی **Thyroid**

تعداد تکرار الگوریتم پایه(H)	روش CAL	روش CSPA	روش HGPA	روش پیشنهادی
۵	۲۴.۳۵	۴۹.۳۵	۴۳.۷۷	۱۲.۲۵
۱۰	۲۰.۸۸	۴۹.۲۶	۳۹.۷۲	۱۲.۵۱
۱۵	۱۹.۸۶	۴۸.۶	۴۰.۴۷	۱۳.۳۲
۲۰	۱۷.۱۹	۴۸.۴۷	۳۸.۲۶	۱۳.۰۴
۲۵	۱۶.۴۷	۴۸.۸۸	۳۷.۴۷	۱۲.۸۵
۳۰	۱۵.۸۸	۴۸.۶	۳۸.۱۲	۱۳.۵۸
۳۵	۱۶.۰۵	۴۸.۸۴	۳۸.۴	۱۳.۲۲
۴۰	۱۶.۴۹	۴۹.۰۹	۳۷.۵۶	۱۳.۱۸
۴۵	۱۶.۵۸	۴۹.۱۴	۳۹.۸۷	۱۳.۰۴
۵۰	۱۵.۸۶	۴۸.۶۵	۳۶.۶۵	۱۳.۱۳

جدول (۵): مقایسه میانگین خطای روش ها (٪) بر روی **Soybean**

تعداد تکرار الگوریتم پایه(H)	روش CAL	روش CSPA	روش HGPA	روش پیشنهادی
۵	۷.۰۲	۱۵.۴۷	۱۸.۵۱	۱۴.۰۴
۱۰	۷.۰۱	۱۳.۴	۱۵.۹۶	۶.۳۸
۱۵	۷.۵۲	۱۲.۵۵	۱۴.۵۷	۸.۰۹
۲۰	۶.۵۵	۱۳.۰۹	۱۴.۵۷	۹.۰۸
۲۵	۶.۸۸	۱۳.۱۹	۱۵.۲۱	۷.۹۴
۳۰	۶.۲۱	۱۴.۲۶	۱۵	۹.۸۶
۳۵	۴.۵۵	۱۴.۱۵	۱۴.۴۷	۵.۱۱
۴۰	۵.۲۱	۱۳.۹۴	۱۵.۱۱	۷.۸
۴۵	۴.۲۲	۱۳.۹۴	۱۵.۸۵	۵.۰۴
۵۰	۴.۵۱	۱۳.۵۱	۱۵.۸۵	۳.۴۸

نتایج فوق نشان می‌دهند که روش پیشنهادی به سرعت همگرا می‌باشد.

تابع f در رابطه (۷)، تابعی از وزن یالها در گراف خوشه‌ها می‌باشد. نتایج تجربی نشان داده‌اند که استفاده از توابع متفاوت، تأثیر چندانی در نتایج و سرعت همگرایی نخواهد داشت. برای نشان دادن این امر سه تابع کاملاً متفاوت زیر را در نظر بگیرید:

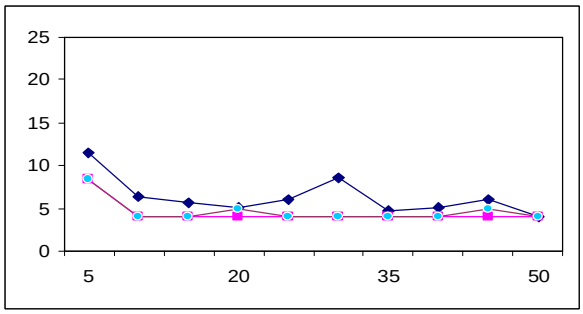
$$f_1(w(i,j)) = \frac{1}{w(i,j)} \qquad (۸)$$

$$f_2(w(i,j)) = w(i,j) \qquad (۹)$$

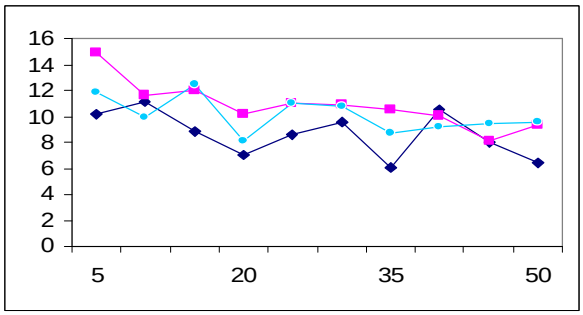
$$f_3(w(i,j)) = \frac{w(i,j)}{\arg\min_{i,j} w(i,j)} \qquad (۱۰)$$

نتایج مربوط به به کارگیری توابع f_1، f_2 و f_3 در شکل های (۲) تا (۵) آورده شده‌است.

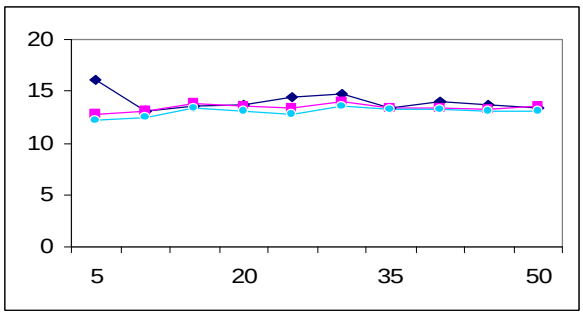
در نتایج گزارش شده در جدول های (۲)-(۵) از تابع f_3 در رابطه (۷) استفاده شده‌است.



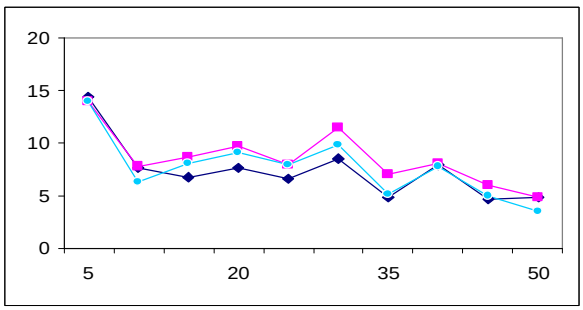
شکل (۲): درصد خطای روش ارائه شده بر روی مجموعه داده **Iris** بر حسب تعداد تکرار الگوریتم پایه



شکل (۳): درصد خطای روش ارائه شده بر روی مجموعه داده **Wine** بر حسب تعداد تکرار الگوریتم پایه



شکل (۴): درصد خطای روش ارائه شده بر روی مجموعه داده **Thyroid** بر حسب تعداد تکرار الگوریتم پایه



شکل (۵): درصد خطای روش ارائه شده بر روی مجموعه داده **Soybean** بر حسب تعداد تکرار الگوریتم پایه

$f=f_1$
$f=f_2$
$f=f_3$

۴- نتیجه گیری

در این مقاله، روشی برای خوشه‌بندی ترکیبی بر مبنای ایجاد یک فضای ویژگی جدید ارائه دادیم. در این روش نتایج الگوریتم های خوشه‌بندی پایه با به کارگیری تئوری گراف مدل سازی شدند. مقادیر ویژگی های جدید طی پیمایش گراف تعیین شدند. طی این فرآیند، هر تکرار الگوریتم خوشه‌بندی پایه منجر به ایجاد یک بعد در فضای ویژگی جدید می‌شود. به این ترتیب، فضای ویژگی جدید دارای ابعادی به تعداد تکرار الگوریتم خوشه‌بندی پایه می‌شود. الگوریتم ساده و کارایی تحت عنوان k-means هوشمند به عنوان الگوریتم خوشه‌بندی پایه معرفی شد. استفاده از این روش تضمین می‌کند که با افزایش تعداد تکرار الگوریتم پایه، نتایج بهبود می‌یابند. این الگوریتم تأثیری در افزایش هزینه محاسباتی الگوریتم ندارد.

بررسی نتایج حاصل از اجرای این روش بر روی چهار مجموعه داده استاندارد Iris، Wine، Thyroid و Soybean رفتار مناسب و همگرایی سریع این روش را نشان می‌دهند. همچنین این روش بهترین نتایج را برای دو مجموعه داده Soybean و Thyroid داشته است.

در تعیین مقادیر ویژگی ها از رابطه (۷) استفاده کردیم. در این رابطه تابعی از وزن یالهای گراف به کار گرفته شده. بررسی‌های انجام شده نشان داده‌اند که استفاده از توابع مختلف تأثیر چندانی در نتایج خوشه‌بندی نهایی نداشته‌است.

مراجع

[1] Strehl, A., Ghosh, J.,"cluster ensembles—a knowledge reuse framework for combining partitioning", Proceeding of 11-th National Conference on Artificial Intelligence, Edmonton, Alberta, Canada, pp. 93–98, 2002.

[2] Fred, A.L.N., Jain, A.K., "Data Clustering Using Evidence Accumulation", Proceeding of the 16th International Conference on Pattern Recognition, ICPR 200, Quebec City, pp.276 – 280, 2002.

[3] Topchy, A., Jain, A.K., Punch, W., "Combining Multiple Weak Clustering", Proceeding of 3d IEEE International Conference on Data Mining, pp.331-338, 2003.

[4] Hu, X., Yoo, I., "Cluster ensemble and its applications in gene expression analysis", Y.-P.P. Chen (Ed.), Proc. 2-nd Asia-Pacific Bioinformatics Conference, Dunedin, New Zealand, pp. 297–302, 2004.

[5] Fern, X.Z, Brodley, C.E., "Random projection for high dimensional data clustering: a cluster ensemble approach", Proceeding of 20th International Conference on Machine Learning, ICML, Washington DC, pp.186–193, 2003.

[6] Strehl, A., Ghosh, J., "Cluster ensembles a knowledge reuse framework for combining multiple partitions", Journal on Machine Learning Research, pp. 583-617, 2002.

[7] Greene, D., Tsymbal, A., Bolshakova, N., Cunningham, P., "Ensemble clustering in medical diagnostics", R. Long et al. (Eds.), Proceeding of 17th IEEE Symp. on Computer-Based Medical Systems, pp. 576– 581, 2004.

[8] Dudoit, S., Fridlyand, J., "Bagging to improve the accuracy of a clustering procedure", Bioinformatics 19, pp.1090–1099, 2003.

[9] Fischer, B., Buhmann, J.M., "Bagging for path-based clustering", IEEE Transactions on Pattern Analysis and Machine Intelligence, pp.1411–1415, 2003.

[10] Fred, A.L.N., Jain, A.K., "Robust data clustering", Proceeding of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, USA, vol. II, pp. 128–136, 2003.

[11] Minaei, B., Topchy, A., Punch, W. F., „Ensembles of Partitions via Data Resampling", Proceeding of International Conference on Information Technology, ITCC 04, Las Vegas, 2004.

[12] Monti, S., Tamayo, P., Mesirov, J., Golub, T., "Consensus clustering: a resampling based method for class discovery and visualization of gene expression microarray data", Machine Learning 52, pp.91–118, 2003.

[13] Topchy, A., Minaei-Bidgoli, B., Jain, A.K., Punch, W., "Adaptive Clustering ensembles", Proceeding of International Conference on Pattern Recognition, ICPR'04, Cambridge, UK, pp.272-275, 2004.

[14] Barthelemy, J.P., Leclerc, B., "The median procedure for partition:", Partitioning Data Sets, AMS DIMACS Series in Discrete Mathematics, pp.3-34, 1995.

[15] Weingessel, A., Dimitriadou, E., Hornik, K., "An ensemble method for clustering. Working paper", http://www.ci.tuwien.ac.at/Conferences/DSC-2003/ , 2003.

[16] Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. 2nd Edition, John Wiley & Sons Inc. New York NY, 2001.

[17] Aarts, E.H.L., Eiben, A.E., VanHee, K.M., A general theory of genetic algorithms. Tech.Rep.89/08, Einndhoven University of Technology, 1989.

[18] Bradley, P., Fayyad, U., "Refining initial points for k-means clustering", Proceedings of 15th International Conference on Machine Learning, San Francisco, CA, pp. 91-99, 1998.

[19] Pena, J., Lozano, J., Larranaga, P., "An Empirical comparison of four initialization methods for the k-means algorithm", Pattern Recognition Letters, Vol. 20, pp. 1027-1040, 1999.

[20] Babu, G., Murty, M., "A near optimal initial seed value selection in k-means algorithm using a genetic algorithm", Pattern Recognition Letters, Vol. 14, pp. 763-769, 1993.

[21] Linde, Y., Buzo, A., Gray, R., "An algorithm for vector quantizer design", IEEE trans. Comm. Vol. 28, pp. 84-95, 1980.

[22] Azimi, J., Davoodi, S.R., Analoui, M., "Fast convergence clustering ensemble", Conference on Data Mining and Data Warehouses, 2006, Ljubljana, Slovenia.

---

1 Cluster-based Similarity Partitioning Algorithm
2 Hypergraph Partitioning Algorithm
3 Co-association and Average Linkage
4 label
5 greedy