Iranian Association of Information and Communication Technology
IKT 2013
Shiraz University
Regular Paper

ارایه روشی جدید برای انتخاب خوشه بندی ترکیبی با استفاده از معیارهای پراکندگی و استقلال

محمد یوسف نژاد[1]، حسین علیزاده[2]، بهروز مینایی بیدگلی[3]

[1]دانشکده فناوری اطلاعات، دانشگاه علوم و فنون مازندران، myousefnezhad@ustmb.ac.ir
[2] دانشکده کامپیوتر، دانشگاه علم و صنعت ایران، halizadeh@iust.ac.ir
[3]دانشکده کامپیوتر، دانشگاه علم و صنعت ایران، b_minaei@iust.ac.ir

چکیده

خوشه بندی ترکیبی مبتنی بر انتخاب روشی جهت انتخاب و ادغام نتایج خوشه بندی پایه روی یک داده خاص بر اساس یک معیار توافقی است. در سال های اخیر از معیار پراکندگی جهت انتخاب نتایج اولیه در این روش استفاده شده است. هدف این مقاله معرفی شاخص استقلال به عنوان روشی مکمل برای انتخاب نتایج اولیه مطلوب می باشد. در این مقاله جهت اندازه گیری پراکندگی دو خوشه بندی معیاری جدید بر اساس معیار APMM ارائه می شود و همچنین برای تشخیص درجه استقلال دو الگوریتم خوشه بندی اولیه مشابه و تاثیرات آن بر روی نتایج نهایی، یک معیار مکاشفه ای معرفی شده است. نتایج تجربی روی چندین مجموعه داده استاندارد نشان می‌دهد که روش پیشنهادی این مقاله به طور موثری نتایج نهایی را بهبود می بخشد. همچنین، مقایسه نتایج به دست آمده با سایر روش‌های خوشه‌بندی ترکیبی نشان از کارایی بالای روش پیشنهادی دارد.

کلمات کلیدی

خوشه بندی ترکیبی، خوشه بندی مبتنی بر انتخاب، درجه استقلال الگوریتم، پراکندگی خوشه بندی

۱- مقدمه

خوشه بندی[1] وظیفه کاوش الگوهای پنهان در داده های بدون برچسب[2] را بر عهده دارد [1]. به خاطر پیچیدگی مسئله و ضعف روش‌های خوشه بندی پایه، امروزه روش‌های خوشه بندی ترکیبی مورد استفاده قرار می گیرند. در مطالعات اخیری که در مورد خوشه بندی ترکیبی انجام شده، کیفیت نتایج اولیه خوشه‌بندی و پراکندگی[3] در نتایج اولیه مورد توجه مقالات زیادی قرار گرفته است اما پاسخ به بعضی سوالات در این زمینه همچنان با ابهامات زیادی روبروست. خوشه بندی ترکیبی مبتنی بر انتخاب روشی است که در آن از زیر مجموعه ی منتخب از نتایج اولیه برای ترکیب و ساخت نتایج نهایی استفاده می شود [1, 2, 3]. پراکندگی درنتایج اولیه یکی از مهمترین عواملی است‌که می‌تواند در کیفیت نتایج نهایی اثرگذار باشد. همچنین، کیفیت نتایج اولیه نیز عامل دیگری است که در کیفیت نتایج حاصل از ترکیب موثر است. هر دو عامل در تحقیقات اخیر خوشه‌بندی ترکیبی مورد توجه قرار گرفته‌اند اما پاسخ به بعضی سوالات دراین زمینه همچنان با ابهامات زیادی روبروست [1, 2, 3, 4, 5].

این مقاله دو هدف را دنبال می کند، هدف اول معرفی شاخص استقلال به عنوان روشی مکمل جهت انتخاب نتایج اولیه مطلوب و هدف دوم ارائه معیاری ترکیبی بر اساس پراکندگی و استقلال می باشد. در این راستا این مقاله جهت اندازه گیری پراکندگی دو خوشه بندی به ارائه معیاری جدید بر اساس معیار APMM[4] که توسط علیزاده و همکاران جهت اندازه گیری پراکندگی دو خوشه ارائه شده می پردازد [6] و جهت تشخیص درجه استقلال دو الگوریتم خوشه بندی اولیه مشابه (همانند دو K-Means) به معرفی یک معیار مکاشفه ای می پردازد. در ادامه مقاله ابتدا در بخش دوم به بررسی پیش زمینه های مورد نیاز پرداخته شده و در بخش سوم کارهای انجام شده در این زمینه بیان می شود. سپس در بخش چهارم، مدل پیشنهادی این مقاله ارائه می شود و در بخش پنجم به ارزیابی و بررسی فواید و مشکلات مدل پیشنهادی پرداخته می شود و در نهایت در بخش ششم نتایج ارائه این مقاله و خط و مشی کارهای آتی بیان می شود.

۲- پیش زمینه

ایده اصلی خوشه‌بندی اطلاعات، جداکردن نمونه‌ها از یکدیگر و قرار دادن آنها در گروه‌های شبیه به هم می‌باشد. به این معنی که نمونه‌های شبیه به هم باید در یک گروه قرار بگیرند و با نمونه‌های گروه‌های دیگر حداکثر تفاوت را دارا باشند [1, 7]. در واقع خوشه‌بندی داده‌ها یک ابزار ضروری برای یافتن گروه‌ها در داده‌های بدون برچسب است [3].

از آنجایی که اکثر روش‌های خوشه‌بندی پایه روی جنبه‌های خاصی از داده‌ها تاکید می‌کنند، درنتیجه روی مجموعه داده‌های خاصی کارآمد

می‌باشند. به همین دلیل، نیازمند روش‌هایی هستیم که بتوانند با استفاده از ترکیب این الگوریتم‌ها و گرفتن نقاط قوت هر یک، نتایج بهینه‌تری را تولید کنند. هدف اصلی خوشه‌بندی ترکیبی، جستجوی نتایج بهتر و مستحکم‌تر با استفاده از ترکیب اطلاعات و نتایج حاصل از چندین خوشه‌بندی اولیه است [2, 3]. خوشه‌بندی ترکیبی می‌تواند جواب‌های بهتری از نظر استحکام[۵]، نو بودن[۶]، پایداری[۷] و انعطاف پذیری[۸] نسبت به روش‌های پایه ارائه دهد [3, 4, 8, 9]. به طور خلاصه خوشه‌بندی ترکیبی شامل دو مرحله اصلی زیر می‌باشد [1, 3]:

۱- تولید نتایج متفاوت از خوشه‌بندی‌ها، به عنوان نتایج خوشه‌بندی اولیه بر اساس اعمال روش‌های مختلف که این مرحله را، مرحله ایجاد تنوع یا پراکندگی می‌نامند.

۲- *ترکیب نتایج* به دست آمده از خوشه‌بندی‌های متفاوت اولیه برای تولید خوشه نهایی؛ که این کارتوسط تابع توافقی[۹] (الگوریتم ترکیب کننده) انجام می‌شود.

۳- کارهای انجام شده

روش‌های خوشه‌بندی ترکیبی سعی می‌کنند تا با ترکیب افرازهای[۱۰] مختلف تولید شده از روش‌های خوشه‌بندی پایه، یک افراز مستحکم[۱۱] از داده‌ها تولید کنند [3, 10, 11, 12]. در اکثر تحقیقات اخیر، همه افرازها با وزن برابر در ترکیب نهایی حاضر می‌شوند و همه خوشه‌های موجود در افرازها نیز با وزن برابر درترکیب نهایی شرکت می‌کنند[3, 13] و یک معیار برای انتخاب از میان ترکیبات ممکن ارائه شده که مبتنی بر کیفیت کلی یک خوشه‌بندی است. برای این کار، آنها میزان ثبات بین افراز ترکیبی و افرازهای پایه را در نظر گرفته‌اند و با استفاده از یک قاعده ترکیبی ثابت، یک معیار شباهت دو به دو[۱۲] را روی فضای ویژگی‌های d-بعدی به کار برده‌اند.

عظیمی و همکاران [14] از مفهوم پراکندگی برای هوشمند نمودن خوشه‌بندی ترکیبی استفاده کرده اند. این روش به صورت پویا اقدام به انتخاب زیرمجموعه بهینه‌ای از نتایج اولیه در ترکیب نهایی می‌کند. نتایج تجربی صورت گرفته نیز نشان داده است که ترکیب خوشه بندی‌های اولیه با بیشترین، کمترین و میزان متوسطی از تطبیق با خوشه‌بندی ترکیبی اولیه، نتیجه بهتری را به ترتیب، در مجموعه داده‌های راحت، سخت و متوسط می‌دهد. روش فوق در هر مجموعه داده سعی می‌کند تا نتایج خوشه‌بندی اولیه‌ای که موجب منحرف شدن نتایج نهایی می‌شود را از ترکیب نهایی خارج کند و به این ترتیب خوشه‌بندی‌های ترکیبی اولیه‌ای را که دارای دقت نسبتاً مناسبی هستند، وارد ترکیب نهایی کند [14, 15, 16]. چندین روش اعتبارسنجی خوشه، مبتنی بر ایده استفاده از پایداری پیشنهاد شده است [17]. بن هور و همکاران [18] نیز روشی برای محاسبه پایداری ارائه کرده‌اند که بر مبنای شباهت بین نمونه‌ها درخوشه‌بندی‌های مختلف عمل می‌کند. در این روش، ابتدا ماتریس همبستگی با استفاده از روش بازنمونه برداری به دست می‌آید [18].

فرد و جین یک روش خوشه‌بندی ترکیبی ارائه کرده اند که در آن با استفاده از معیار پایداری خوشه، شباهت دو به دو آموزش داده می‌شود. در این روش، به جای استفاده از معیارهای ارزیابی مبتنی بر افراز نهایی، افرازهای حاصل از الگوریتم‌های پایه، در نواحی مختلف از فضای ویژگی d-بعدی مورد ارزیابی قرار گرفته‌اند [13].

فرن و لین روشی برای خوشه‌بندی ترکیبی پیشنهاد کرده‌اند که از زیرمجموعه‌ی موثرتری از افرازهای اولیه در ترکیب نهایی استفاده می‌کند. در این روش اگرچه تعداد اعضای شرکت کننده در ترکیب نهایی کمتر از یک خوشه‌بندی ترکیبی کامل[۱۳] است، به دلیل انتخاب افرازهای با کارایی بالاتر، نتایج نهایی بهبود می‌یابند. پارامترهایی که در این روش مورد توجه قرار گرفته‌اند، عبارتند از: کیفیت و پراکندگی [19]. در این روش از معیار مجموع اطلاعات متقابل نرمال شده (SNMI[۱۴]) (برای یک افراز در مقایسه با افرازهای دیگر ترکیب) برای اندازه‌گیری کیفیت یک افراز استفاده شده است. همچنین، معیار اطلاعات متقابل نرمال شده (NMI[۱۵]) (بین تمام افرازهای موجود درترکیب) برای اندازه گیری پراکندگی لازم برای ترکیب به کار رفته است [19]. فرن و لین نشان می دهند که روش پیشنهادیشان نسبت به خوشه‌بندی ترکیبی کامل و یا روش انتخاب تصادفی از کارایی بهتری برخوردار است [19]. علیزاده و همکاران [20] روشی جهت انتخاب خوشه بر اساس معیار پایداری ارائه داده اند. در این روش به معرفی معیار APMM و روش ماکزیموم جهت رفع مشکل تقارن در معیار NMI پرداخته شده است. این روش، روشی نوین جهت تشکیل ماتریس همبستگی بدون نیاز به تمامی نتایج خوشه های خوشه‌بندی اولیه می باشد [6, 20].

۴- مدل پیشنهادی

نو بودن و پایداری در نتایج اولیه از مهمترین خواصی هستند که این مقاله به دنبال رسیدن به آن است. در اینجا نو بودن یعنی رسیدن به افراز جدیدی در خوشه بندی های اولیه که تا به حال در سایر نتایج به این حالت نرسیده ایم که این امر کمک بسزایی در کشف الگوهای[۱۶] جدید (دانش ضمنی[۱۷]) از داده می کند. پایداری نیز تضمین می کند تا با تکرار مکرر یک روش روی یک داده نتایج مشابه داشته باشد. در صورتی که فقط رسیدن به پایداری نتایج نهایی خوشه بندی ترکیبی مهم باشد ممکن است این دو خصوصیت (نو بودن و پایداری) در خلاف راستای همدیگر عمل کنند به این معنی که هر چقدر نو بودن جواب‌ها در تکرار مکرر یک خوشه بندی بیشتر باشد جواب های غیر پایداری مشاهده شود و هر چقدر پایداری بیشتری مد نظر باشد خیلی از جواب های نو را از دست داده و نهایتاً درصد پیش بینی الگوی درست کمتری مشاهده خواهد شد [3, 6, 14, 15, 16]. راه حلی که این مقاله به دنبال آن است روشی است که تضمین کند با تکرار هر بار اجرای فرآیند خوشه بندی ترکیبی، متنوع ترین مجموعه از نتایج خوشه بندی اولیه تولید شده و در ترکیب از آنها استفاده می شود. این مقاله با اندازه گیری و کنترل درجه استقلال الگوریتم های خوشه بندی اولیه به دنبال رسیدن به این هدف می باشد.

۴-۱- محاسبه درجه استقلال دو الگوریتم

در بیشتر روش های خوشه بندی ترکیبی جهت ایجاد پراکندگی و رسیدن به نتایج نوتر و انعطاف پذیرتر، از تکرار مکرر یک الگوریتم پایه خوشه بندی روی داده (برای مثال K-Means) بهره گرفته می شود. در این الگوریتم ها عموماً جهت ایجاد نتایج متفاوت در بخشی از روش حل مسئله از مقادیر قابل برنامه ریزی یا تصادفی استفاده می شود. برای مثال در K-Means، مقادیر اولیه مراکز خوشه ها یا مقدار K یا تعداد دفعات تکرار[۱۸] الگوریتم جزء این پارامترها می باشد. لازم به ذکر است که برخی از الگوریتم ها همانند Linkage که با تکرار مکرر بر روی یک داده همیشه یک جواب معین را تکرار می کنند (معمولاً از مولد اعداد تصادفی استفاده نمی کنند) شامل این قانون نمی شوند و معمولاً در ساخت نتایج اولیه خوشه بندی ترکیبی، از هر یک از انواع آن فقط یک بار استفاده می شود [3, 6, 10, 21].

این مقاله، ماتریسی از مقادیر اولیه را به عنوان عامل محرک الگوریتم بر اساس روش کار هر الگوریتم در نظر می گیرد (به عنوان مثال مقدار تصادفی اولیه مراکز خوشه ها در الگوریتم هایی مثل K-means و FCM و ... یا پارامترهای داخلی الگوریتم ها همانند ماتریس فاصله در روش های Spectral و هر نوع مقادیر اولیه ای که می تواند روش کار الگوریتم‌ها را تغییر دهند) که به آن پارامترهای اساسی الگوریتم گفته می شود. بدیهی است که چون روش کار هر الگوریتم ثابت است، اگر مقادیر ثابت بمانند جواب های نهایی الگوریتم پایه یکی خواهد بود یا به عبارتی نتایج هر الگوریتم پایه به مقادیر این ماتریس *وابسته* می باشد. از این رو بر اساس تعریف ذیل، درجه استقلال دو الگوریتم به صورت شبه کد شکل (۱) محاسبه می شود:

```
Function BIndependency (C1, C2, P1, P2)
  If type of cluster C1 and C2 is equal then
     Distance-Matrix is distance between P1 and P2
     Do until Distance-Matrix is not null
        Find minimum cell of Distance-Matrix
        Store cell in Temp-Array
        Remove Row and Column of founded cell
        Create new Distance-Matrix
     End loop
     Return average of Temp-Array
  Else
     Result = 1
  End If
End Function
```

شکل 1: محاسبه درجه استقلال دو خوشه بندی

در شکل (۱)، C1 و C2 دو الگوریتم خوشه بندی می باشند که قرار است درجه استقلال آنها با هم مقایسه شود و همچنین P1 و P2 به ترتیب ماتریس های پارامترهای اساسی این دو الگوریتم می باشند. در صورت یکی نبودن نوع دو الگوریتم استقلال آنها یک محاسبه می شود که به معنی کاملاً مستقل می باشد در غیر این صورت ماتریس $n \times n$ فاصله Max-Distance بر اساس P1 و P2 تشکیل می شود در اینجا n برابر با حداکثر اندازه ماتریس های P1 و P2 است. لازم به ذکر است که در این حالت هر چه فاصله (در این مقاله ما از فاصله اقلیدسی استفاده کردیم ولی با این حال می توان از هر معیار فاصله دیگری استفاده کرد) بیشتر باشد، درجه استقلال بهتری بدست خواهد آمد، از این رو در حلقه شکل (۱) می‌بایست هر بار مقدار حداقل پیدا شده، و در Temp-Array نگهداری شود و سطر و ستونی که در آن این مقدار وجود دارد حذف شود و برای ماتریس جدید به وجود آمده مجدداً همین کار تکرارشود. نهایتاً مقدار درجه استقلال، میانگین مقادیر حداقل ها در ماتریس فاصله خواهد بود. خروجی شبه کد شکل (۱) همواره یک مقدار بین صفر و یک خواهد بود که در آن یک به معنی کاملاً مستقل و صفر به معنی کاملاً وابسته می باشد. جهت محاسبه مقدار استقلال یک الگوریتم نسبت به کل الگوریتم های دیگر در نتایج اولیه خوشه بندی کافی است که میانگین مقادیر BIndependency برای آن الگوریتم نسبت به بقیه الگوریتم‌ها، مطابق با رابطه (۱) محاسبه شود:

$$Independence(C) = \frac{1}{M}\sum_{i=1}^{M} \text{BIndependence}(C, C_i) \quad (۱)$$

در رابطه (۱) مقدار M برابر با کل الگوریتم هایی است که قرار است با الگوریتم C مقایسه شوند. محاسبه درجه استقلال باعث حذف جواب های مشابه خواهد شد، از این رو فقط آن الگوریتم هایی که مقدار استقلال آنها نسبت به سایر الگوریتم ها بالاتر از یک مقدار آستانه قابل برنامه ریزی است در جواب نهایی شرکت می کنند که این باعث حفظ خاصیت نو بودن و پایداری همزمان خواهد شد. رابطه (۲) را شرط استقلال می نامیم که در آن iT مقدار آستانه جهت پذیرش یا رد یک الگوریتم در جواب نهایی خواهد بود:

$$Independence(C) \geq iT \quad (۲)$$

۴-۲- محاسبه پراکندگی دو خوشه بندی

ما در این مقاله از معیار APMM برای محاسبه مقدار پراکندگی استفاده می کنیم چون این معیار هم از لحاظ پیچیدگی زمانی سریعتر از NMI می‌باشد و هم مشکل تقارن ندارد [6]. در این روش برای محاسبه تراکم خوشه C_i از رابطه (۳) استفاده می کنیم:

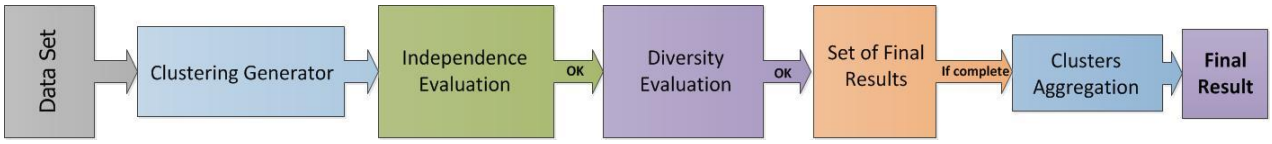
$$AAPMM(C_i) = \frac{1}{M}\sum_{j=1}^{M} APMM\left(C_i, P_j^{b*}\right) \quad (۳)$$

در رابطه (۳) پارامتر P_j^{b*} نشان دهنده j-امین افراز از مجموعه مرجع است و همچنین تابع APMM در این رابطه را از رابطه (۴) محاسبه می کنیم:

$$APMM\ (C,P) = \frac{-2n_c \log\left(\frac{n}{n_c}\right)}{n_c \log\left(\frac{n_c}{n}\right) + \sum_{i=1}^{kp} n_i^p \log\left(\frac{n_i^p}{n}\right)} \quad (۴)$$

در رابطه (۴) C_i^a خوشه i-ام در افراز P^a می‌باشد و P^{b*} افراز متناظر با خوشه C_i در خوشه‌بندی P^b است. پارامتر n تعداد کل نمونه‌های مجموعه داده و n_{ij}^{ab} تعداد نمونه‌های مشترک بین خوشه‌های $C_i^a \in P^a$

و $C_i^b \in P^b$ می باشد. همچنین، K_{b*} تعداد خوشه‌های موجود در افراز P^{b*} می‌باشد. از آنجایی که AAPMM فقط تراکم یک خوشه را



شکل 2: خوشه بندی ترکیبی با معیار ترکیبی استقلال و پراکندگی

محاسبه می کند، برای محاسبه یک خوشه بندی، از رابطه (۵) استفاده می شود:

$$A3(P) = \frac{1}{n}\sum_{i=1}^{k} n_i \times AAPMM(C_i) \qquad (۵)$$

در رابطه (۵) معیار A3 تراکم یک افراز که در اینجا نتیجه یک خوشه بندی اولیه است را محاسبه می کند که در آن n تعداد کل داده های افراز P و k تعداد کل خوشه ها و n_i تعداد کل داده های افراز i-ام می‌باشد. این معیار همواره یک عدد بین صفر و یک بر می گرداند که مقدار تراکم (عکس پراکندگی) را نشان می دهد. جهت محاسبه پراکندگی از رابطه (۶) استفاده می کنیم و رابطه (۷) شرط پراکندگی می باشد که در آن dT مقدار آستانه جهت رد یا پذیرش یک نتیجه الگوریتم اولیه به منظور تشکیل نتیجه نهایی می باشد:

$$Diversity(p) = 1 - A3(p) \qquad (۶)$$

$$Diversity(p) \geq dT \qquad (۷)$$

۴-۳- ترکیب نتایج اولیه

در این مرحله، خوشه‌های انتخاب شده، ماتریس همبستگی را تشکیل می‌دهند. در روش انباشت مدارک (EAC[۱۹]) نتایج m خوشه‌بندی روی داده‌های نمونه‌برداری شده، در ماتریس همبستگی $n \times n$ ذخیره می‌شوند. هر داده ورودی از این ماتریس در روش انباشت مدارک، به صورت رابطه (۸) محاسبه می‌شود [10]:

$$C(i,j) = \frac{n_{i,j}}{m_{i,j}} \qquad (۸)$$

در این رابطه $n_{i,j}$ تعداد دفعاتی است که جفت نمونه‌های i و j با هم در یک خوشه گروه‌بندی شده‌اند و $m_{i,j}$ تعداد نمونه‌برداری هایی است که هر دوی این جفت نمونه‌ها به طور همزمان در آن ظاهر شده‌اند [10].

۴-۴- جمع بندی

در شکل (۲) فرآیند شکل گیری نتایج نهایی در روش پیشنهادی این مقاله به تصویر کشیده شده است همانطور که در این شکل می بینید، ابتدا نتایج در بخش مولد تولید شده سپس بر اساس رابطه (۲) شرط استقلال آن چک می شود و شرط پراکندگی افرازهایی که دارای شرایط استقلال مناسب هستند بر اساس رابطه (۷) چک شده و این افرازها مجموعه نتایج نهایی را می سازند. در بخش ترکیب، با استفاده از رابطه (۸) عناصر ماتریس همبستگی تشکیل شده و با استفاده از الگوریتم سلسله مراتبی، اتصال میانگین نتایج خوشه‌بندی نهایی تشکیل می شود [2].

۵- ارزیابی

در این بخش نتایج به کارگیری روش پیشنهادی روی مجموعه داده‌های مختلف و پارامترهای مورد استفاده گزارش می شود. در کلیه آزمایشات، مقادیر iT و dT بر اساس پیچیدگی زمانی تعیین شده تا بر روی یک سیستم با مشخصات معین[۲۰] به مدت ۳۰ دقیقه محاسبات انجام شود. بدیهی است که افزایش مقادیر آستانه، شرایط بهتر و زمان طولانی تری را به وجود می آورند.

۵-۱- مجموعه داده‌ها

روش پیشنهادی بر روی ۱۴ مجموعه داده استاندارد مورد آزمایش قرار گرفته است. برای انجام آزمایش‌ها سعی شده است که مجموعه داده‌ها از لحاظ تعداد کلاس‌ها، تعداد ویژگی‌ها و همچنین تعداد نمونه‌ها از حداکثر تنوع برخوردار باشند در نتیجه نتایج آزمایش‌ها تا حد ممکن مستحکم و قابل تعمیم خواهد بود. جدول (۱) اطلاعات مختصری از این مجموعه داده‌ها در اختیار می‌گذارد. برای اطلاعات بیشتر در مورد هر کدام از این مجموعه داده‌ها به [22] رجوع کنید. نتایج آزمایش‌ها بر روی ویژگی‌های نرمال شده از این مجموعه داده‌ها گزارش شده است. به عبارت دیگر هر کدام از ویژگی‌های این مجموعه داده‌ها با میانگین صفر و واریانس یک، N(0,1)، نرمال شده‌اند.

جدول 1: مجموعه داده

No.	Name	Feature	Class	Sample
1	Half Ring	2	2	400
2	Iris	4	3	150
3	Balance Scale	4	3	625
4	Breast Cancer	9	2	683
5	Bupa	6	2	345
6	Galaxy	4	7	323
7	Glass	9	6	214
8	Ionosphere	34	2	351
9	SA Heart	9	2	462
10	Wine	13	2	178
11	Yeast	8	10	1484
12	Pendigits	16	10	10992
13	Statlog	36	7	6435
14	Optdigits	62	10	5620

جدول 2: مقایسه روش های مختلف خوشه بندی

	روش های خوشه بندی پایه					روش های خوشه بندی ترکیبی						
	Kmeans	FCM	Subtract	Single Linkage	EAC	MAX	CSPA	HGPA	MCLA	روش پیشنهادی مقاله		
										iT	*dT*	نتیجه
Half Ring	75.75	78	86	75.75	77.17	78.48	74.5	50	74.5	0.2	0.06	**87.2**
Iris	65.3	82.66	55.3	68	**96**	72.89	85.34	48.66	89.34	0.2	0.06	**96**
Balance Scale	40.32	44	45.32	46.4	52	52.1	51.84	41.28	51.36	0.23	0.063	**54.88**
Breast Cancer	93.7	94.43	65	65.15	95.02	75.72	80.97	50.37	96.05	0.18	0.02	**96.92**
Bupa	54.49	50.1	**57.97**	57.68	55.18	56.17	56.23	50.32	55.36	0.21	0.04	57.42
Galaxy	30.03	34.98	29.72	25.07	31.95	32.78	29.41	31.27	28.48	0.2	0.05	**35.88**
Glass	42.05	47.19	36.44	36.44	45.93	44.17	38.78	41.12	51.4	0.19	0.06	**51.82**
Ionosphere	69.51	67.8	**71.5**	64.38	70.48	64.48	67.8	58.4	70.22	0.3	0.1	70.52
SA Heart	64.51	63.41	67.26	65.15	65.19	63.96	58.42	50.93	62.54	0.65	0.8	**68.7**
Yeast	31.19	29.98	31.2	31.73	31.74	32.4	14	15.23	17.56	0.5	0.5	**34.76**
Wine	65.73	**71.34**	67.23	37.64	70.56	69.17	67.41	62.36	70.22	0.2	0.05	**71.34**
Pendigits	46.97	36.77	10.4	10.46	10.47	57.02	58.32	11.14	58.62	0.02	0.12	**58.68**
Optdigit	52.52	38.33	47.72	10.28	20	76.11	75.21	64.77	77.15	0.01	0.1	**77.16**
Statlog	50.93	49.91	23.8	23.8	23.9	54.23	54.23	52.94	55.71	0.01	0.1	**55.77**

۵-۲- نتایج آزمایش

روش پیشنهادی در محیط MATLAB (ver 7.1) پیاده سازی و مورد آزمایش قرار گرفته و نتایج آزمایش‌ها روی میانگین ۱۰ بار اجرای مستقل برنامه گزارش شده است. جهت ایجاد نتایج اولیه در این مقاله از K-means، FCM، الگوریتم های Spectral، انواع الگوریتم های Linkage (Single, Complete, Average) و Ward (هر کدام فقط یک بار) و Subtract استفاده شده است. عملکرد روش‌های مختلف خوشه‌بندی با استفاده از فرایند بازبرچسب‌گذاری[۲۱] بین خوشه‌های به دست آمده و کلاس‌های واقعی و مقایسه آنها محاسبه شده است. جدول (۲) عملکرد روش‌های مختلف را در مقایسه با روش پیشنهادی این مقاله نشان می‌دهد. همانطور که می بینید نتایج در اکثر موارد بهتر از سایر روش ها بوده و در برخی از داده ها همانند (Bupa و Ionosphere) با اینکه شرایط کاملا بهبود نیافته ولی اختلاف آن با بهترین روش مقدار کمی (کمتر از یک درصد) است. همان طور که در جدول (۲) نشان داده شده است روش Subtract بهترین نتیجه را روی این دو داده ایجاد کرده است و همانطور که پیشتر به آن اشاره شد روش‌های پایه خوشه بندی فقط روی بعضی از ویژگی های داده خوب کار می کنند که در این مسئله روش Subtract به خاطر ویژگی‌های خاص این دو داده، روی آنها نتایج مطلوبی ایجاد می کند ولی روی بقیه داده ها این الگوریتم نمی‌تواند نتایج خوبی ایجاد کند. این مهمترین دلیل عدم بهبود نتایج در این دو داده خاص می باشد. از دیگر دلایل این مشکل می توان به کم بودن مقادیر آستانه به علت رعایت شرط زمانی (حداکثر ۳۰ دقیقه) و پیچیدگی خاص این داده‌ها به نحوی که مجموعه الگوریتم های انتخاب شده و روش ترکیب نتوانسته نتیجه‌ی مطلوبی ایجاد کند، را نیز اشاره کرد. ولی باید توجه داشت چون اختلاف نتایج روش مقاله با بهترین روش در این دو داده بسیار کم است (کمتر از یک درصد) نتایج تقریبا قابل قبول می‌باشد.

۶- نتیجه گیری

در این مقاله یک روش جدید مبتنی بر انتخاب مجموعه ای از نتایج اولیه برای خوشه‌بندی ترکیبی پیشنهاد شده است. از آن جایی که کیفیت، پایداری و نو بودن خوشه‌های حاصل از الگوریتم‌های پایه برابر نیست و حتی حضور تعدادی از آنها می‌تواند منجر به بدتر شدن نتیجه خوشه‌بندی ترکیبی شود، این مقاله روشی برای انتخاب زیرمجموعه بهینه‌تر و موثرتر از خوشه‌های اولیه برای شرکت در ترکیب نهایی بر اساس معیارهای استقلال و پراکندگی توسعه داده که در آن روش مکاشفه ای جهت اندازه گیری درجه استقلال دو خوشه بندی مشابه معرفی شده است و با توسعه روش APMM (که روشی برای سنجش پراکندگی یک خوشه است)، روشی جهت اندازه گیری پراکندگی نتایج دو خوشه بندی ارائه شده است. نتایج تجربی روش پیشنهادی خوشه‌بندی ترکیبی بر روی ۱۴ مجموعه داده مختلف و متنوع نشان می‌دهد که این روش نسبت به روش‌های متداول و همچنین سایر روش‌های ترکیبی، برتری قابل ملاحظه‌ای دارد. همچنین، بررسی‌ها نشان می‌دهند که اگرچه روش پیشنهادی از زیرمجموعه کوچکی از نتایج خوشه بندی های اولیه استفاده می‌کند، اما به خاطر موثر بودن این زیرمجموعه و همچنین حذف خوشه‌ها با کیفیت پایین و تکراری که تاثیر منفی روی میزان همبستگی واقعی نمونه‌ها می‌گذارند، نتایج نهایی حتی از ترکیب کامل (EAC) هم بهتر می‌شود.

مراجع

[1] A. Jain, M. N. Murty, and P. Flynn, "Data clustering: A review. ACM Computing Surveys", 31(3):264–323, 1999.
[2] A. Fred and A. K. Jain, "Data Clustering Using Evidence Accumulation", Proc. of the 16th Intl. Conf. on Pattern Recognition, ICPR02, Quebec City, pp. 276 – 280, 2002.
[3] A. Strehl and J. Ghosh, "Cluster ensembles - a knowledge reuse framework for combining multiple partitions", Journal of Machine Learning Research, 3(Dec):583–617, 2002.
[4] A. Fred and A. Lourenco, "Cluster Ensemble Methods: from Single Clusterings to Combined Solutions", Studies in Computational Intelligence (SCI), 126, 3–30, 2008.
[5] A. Fred and A. K. Jain, "Data Clustering Using Evidence Accumulation", Proc. of the 16th Intl. Conf. on Pattern Recognition, ICPR02, Quebec City, pp. 276 – 280, 2002.
[6] H. Alizadeh, B. Minaei-Bidgoli and H. Parvin, "Cluster Ensemble Selection Based on a New Cluster Stability Measure, Intelligent Data Analysis, IOS Press", ISI Expanded, in press, will be appeared in Vol 18(3), 2014.
[7] K. Faceli , C. P. Marcilio, D. Souto, "Multi-objective Clustering Ensemble", Proceedings of the Sixth International Conference on Hybrid Intelligent Systems (HIS'06), 2006.
[8] A. Topchy, A. K. Jain and W. F. Punch, "Combining Multiple Weak Clusterings", Proc. 3d IEEE Intl. Conf. on Data Mining, pp. 331-338, 2003.
[9] H. G. Ayad and M. S. Kamel, "Cumulative Voting Consensus Method for Partitions with a Variable Number of Clusters", IEEE Trans. On Pattern Analysis and Machine Intelligence, VOL. 30, NO. 1, 160-173, 2008.
[10] A. L. Fred and A. K. Jain, "Combining Multiple Clusterings Using Evidence Accumulation", IEEE Trans. on Pattern Analysis and Machine Intelligence, 27(6):835–850, 2005.
[11] H. Ayad and M. Kamel, "Cluster-based cumulative ensembles". In N. Oza and R. Polikar, editors, Proc. the 6th Intl. Workshop on Multiple Classifier Systems, pages 236–245. LNCS 3541, 2005.
[12] L. I. Kuncheva and S. Hadjitodorov, "Using diversity in cluster ensembles", In Proc. of IEEE Intl. Conference on Systems, Man and Cybernetics, pages 1214–1219, 2004.
[13] A. Fred and A. K. Jain, "Learning Pairwise Similarity for Data Clustering", In Proc. of the 18th Int. Conf. on Pattern Recognition (ICPR'06), 2006.
[14] J. Azimi, J. Maani and N. Mozayyeni, "Improved Clustering Ensembles", 11th International CSI Computer Conference (CSICC06), Tehran, Iran, 24-26 January, (In Persian), 2006.
[15] J. Azimi, M. Mohammadi and M. Analoui, "Clustering Ensembles Using Genetic Algorithm", IEEE International Workshop on Computer Architecture for Machine Perception and Sensing (CAMPS'06), 2006.
[16] J. Azimi and M. Analoui, "Distinguishing Marginal Samples to Improve Clustering Ensembles", 11th International CSI Computer Conference (CSICC06), Tehran, Iran, 24-26 January, (In Persian), 2006.
[17] T. Lange, V. Roth, M. L. Braun and J. M. Buhmann, "Stability-based validation of clustering solutions", Neural Computation, 16(6):1299–1323, 2004.
[18] A. Ben-Hur, A. Elisseeff and I. Guyon, "A stability based method for discovering structure in clustered data", in Pacific Symposium on Biocomputing, vol. 7, pp. 6-17, 2002.
[19] A. Fred and A. Lourenco, "Cluster Ensemble Methods: from Single Clusterings to Combined Solutions", Studies in Computational Intelligence (SCI), 126, 3–30, 2008.
[20] H. Alizadeh, H. Parvin and S. Parvin, "A Framework for Cluster Ensemble Based on a Max Metric as Cluster Evaluator". International Journal of Computer Science (IAENG), pp.1-39, 2012.
[21] P. Y. Mokn, H. Q. Huang, Y. L. Kwok, and J. S. Au, "A Robust Adaptive clustering analysis method for automatic identification of clusters", Pattern Recognition, Vol. 46, pp. 3017-3033, 2012.
[22] C. B. D. J. Newman, S. Hettich and C. Merz, UCI repository of machine learning databases, 1998, http://www.ics.uci.edu/˜mlearn/MLSummary.html.

زیرنویس‌ها

[1] Clustering
[2] Label
[3] Diversity
[4] Alizadeh-Parvin-Moshki-Minaei
[5] Robustness
[6] Novelty
[7] Stability
[8] Flexibility
[9] Consensus Function
[10] Partitions
[11] Robust
[12] Pairwise
[13] Full Ensemble
[14] Sum of Normalized Mutual Information
[15] Normalized Mutual Information
[16] Pattern Recognition
[17] Tacit knowledge
[18] Iterative
[19] Evidence Accumulation Clustering
[20] CPU=X9775, RAM=16GB, OS=Windows
[21] Relabeling