اولین همایش فناوری
اطلاعات، حال، آینده

خوشه بندی توافقی وزنی

حامد زجاجی

دانشجوی کارشناسی ارشد نرم افزار دانشگاه آزاد اسلامی واحد مشهد

hzojaji@gmail.com

حسین دلداری

دانشیار گروه کامپیوتر دانشگاه فردوسی مشهد

hd@ferdowsi.um.ac.ir

چکیده – خوشه بندی را می توان یکی از مهمترین مراحل در تحلیل داده ها بر شمرد. روش های خوشه بندی بسیاری تاکنون توسعه داده شده اند. یکی از این روش ها که در مطالعات اخیر مطرح شده و مورد توجه قرار گرفته است، روش خوشه بندی توافقی می باشد. هدف خوشه بندی توافقی ترکیب چند خوشه بندی و بدست آوردن یک خوشه بندی نهایی است به گونه ای که در آن خوشه ها از کیفیت و پایداری بالاتری، نسبت به خوشه ها در خوشه بندی های اولیه، برخوردار باشند. ما در این مقاله فرآیندی را جهت خوشه بندی توافقی ارائه می کنیم که در آن به هر یک از خوشه بندی های اولیه، وزنی اختصاص می یابد. در نهایت الگوریتم ارائه شده در این مقاله را با یکی از الگوریتم های کارا و مطرح در زمینه خوشه بندی توافقی مورد مقایسه قرار خواهیم داد. نتایج ارزیابی هایی که انجام داده ایم نشان دهنده این موضوع است که الگوریتم پیشنهاد شده اغلب نتایج بهتری را نسبت به الگوریتم دیگر بدست می آورد.

کلید واژه- خوشه بندی توافقی، داده کاوی، ماتریس تناظر، معیار ارزیابی خوشه بندی

۱- مقدمه

خوشه بندی اغلب به عنوان اولین و یکی از مهمترین گام‌ها در تحلیل داده ها به شمار می آید. تاکنون الگوریتم های خوشه بندی بسیاری توسعه یافته اند [۱، ۲]، نظیر خوشه بندی سلسله مراتبی (Heirarchical clustering)، خوشه بندی افرازی (Partitioning clustering)، خوشه بندی بر مبنای تراکم (Density-based clustering) و روش هایی که بر روی اجتماعی از خوشه بندی ها عمل می کنند. یکی از روش های خوشه‌بندی که در تحقیقات اخیر مورد توجه قرار گرفته است، روش خوشه بندی توافقی (Consensus clustering) می باشد. این نوع خوشه بندی در مقالات با نام های دیگری نظیر گروه های خوشه بندی (Clustering ensemble)، اجتماع خوشه بندی ها (Clustering aggregation) نیز شناخته می شود. در روش خوشه بندی توافقی نتایج حاصل از چندین خوشه‌بندی با هم ترکیب شده تا در نهایت به یک خوشه‌بندی واحد دست یابیم.

الگوریتم های خوشه بندی توافقی اغلب: خوشه بندی بهتری تولید می کنند؛ خوشه بندی ترکیب شده ای را می یابند که به تنهایی توسط هر الگوریتم خوشه بندی دیگری قابل تولید نمی باشد؛ حساسیت کمتری نسبت به نویز دارند؛ و قادر به یکپارچه سازی نتایج از منابع توزیع شده می‌باشند [۳].

اغلب روش هایی که نتایج حاصل از چندین خوشه بندی را با هم ترکیب می کنند [۳، ۴، ۵، ۶، ۷] به مسئله تناظر بین خوشه ها در خوشه بندی های مختلف توجهی ندارند. البته روش هایی وجود دارند [۸، ۹] که از برچسب گذاری مجدد جهت حل این مسئله استفاده می کنند.

ما در این مقاله فرآیندی جهت خوشه بندی توافقی ارائه می دهیم که در آن ابتدا با استفاده از الگوریتمی نوین،

تناظر یک به یک بین خوشه ها در خوشه بندی های مختلف بدست می آید و سپس با استفاده از یک الگوریتم بهبود یافته، خوشه بندی های موجود با یکدیگر ترکیب می شوند. ساختار ادامه مقاله به این صورت می‌باشد: در بخش ۲ به بررسی کارهای مرتبط در زمینه خوشه بندی توافقی می پردازیم؛ در بخش ۳ روش کار را تشریح خواهیم کرد؛ در بخش ۴ پیچیدگی الگوریتم تحلیل می‌شود؛ در بخش ۵ مجموعه های داده ای که در ارزیابی‌ها استفاده شده است معرفی می گردد؛ در بخش ۶ به بررسی نتایج اجرای روش، بر روی مجموعه های داده‌ای خواهیم پرداخت؛ در بخش ۷ نیز مطالب مطرح شده در مقاله را نتیجه گیری و جمع بندی خواهیم نمود.

۲- کارهای مرتبط

در این بخش ابتدا به چند الگوریتم مطرح در زمینه خوشه‌بندی توافقی اشاره خواهیم نمود و سپس به بررسی کلی الگوریتم IVC که در [۳] آمده است، می پردازیم.

در [۴] سه روش خوشه بندی توافقی پیشنهاد شده است: CSPA، HGPA و MCLA. این سه الگوریتم قبل از انجام خوشه بندی، اجتماع خوشه بندی ها را به یک گراف تبدیل می‌کنند. در [۵] یک تابع هدف، به عنوان اطلاعات دوجانبه بین خوشه بندی نهایی و گروه های اولیه خوشه‌بندی فرموله می‌شود. روش های دیگری نیز وجود دارند که بر اساس شباهت دو به دو بین اشیاء داده‌ای عمل می‌کنند. روش های FC [۷] و HAC [۱۰، ۱۱] را می‌توان از این نوع برشمرد.

الگوریتم IVC در [۳] مطرح شده است. در این روش تعداد خوشه ها در خوشه بندی های مختلف و خوشه‌بندی نهایی باید یکسان باشد. همچنین مشخصه‌های داده ای - مانند اغلب این نوع روش ها - در انجام خوشه‌بندی شرکت داده نمی شوند. در IVC از بردار $Y=\{y_1,y_2,...,y_N\}$ استفاده می شود به طوری که $y_i=\langle\pi_1(x_i),\pi_2(x_i),...,\pi_C(x_i)\rangle$ است. در اینجا تعداد داده ها N و تعداد خوشه بندی ها C می باشد. $\pi_m(x_i)$ نیز مشخص می کند x_i در خوشه بندی شماره m در خوشه شماره چند قرار دارد. هر تکرار در الگوریتم IVC شامل دو گام می شود: در گام اول، مرکز خوشه ها در خوشه بندی نهایی محاسبه می گردد و در گام دوم، داده ها (بردار های y_i) به نزدیکترین خوشه در خوشه‌بندی نهایی نسبت داده می شوند. روش کلی کار در IVC مشابه روش kmeans می باشد. نتایج مقایسه این الگوریتم با یازده الگوریتم مطرح که در [۳] آمده است، نشان می دهد این الگوریتم از کارایی بالاتری نسبت به آنها، برخوردار است.

ما در این مقاله الگوریتم IVC را از دو جنبه بهبود داده‌ایم. اول اینکه، الگوریتم IVC مسئله تناظر بین خوشه ها در خوشه بندی های مختلف را حل نمی کند، که این مسئله در کیفیت خوشه بندی تأثیر گذار است. بنابراین ما در فرآیند پیشنهادی تناظر یک به یک بین خوشه ها را با الگوریتمی نوین تشخیص می دهیم. دوم اینکه، در روش IVC جهت محاسبه فاصله بین بردارهای y_i تا مرکز خوشه ها از فاصله همینگ استفاده می شود، در این تعیین فاصله تمام خوشه بندی های اولیه به طور یکسانی بر روی خوشه بندی نهایی تأثیر می‌گذارند، که در این حالت، کیفیت پائین برخی از خوشه‌بندی ها می‌تواند باعث کاهش کیفیت خوشه بندی نهایی گردد. ما در الگوریتم ارائه شده از شاخص Davies-Bouldin، که برای هر یک از خوشه بندی های اولیه محاسبه می‌شود، به عنوان وزن در تعیین فاصله بردارهای y_i تا مرکز خوشه ها استفاده می کنیم.

۳- روش کار

در این بخش ابتدا روش یافتن خوشه های متناظر در خوشه بندی های مختلف مورد بررسی قرار می‌گیرد. سپس الگوریتم خوشه بندی توافقی بهبود یافته مطرح خواهد شد.

۳-۱- تشخیص تناظر بین خوشه ها

اغلب الگوریتم های خوشه بندی توافقی بر روی مسئله تشخیص تناظر خوشه ها متمرکز نمی شوند. روش های برچسب گذاری مجدد نیز که این مسئله را حل می‌کنند، بر روی خوشه بندی هایی که تعداد خوشه های متفاوتی دارند قابل استفاده نمی باشند. ما در [۱۲] روشی ارائه داده ایم که علاوه بر حل نمودن مسئله مذکور می تواند

این محدودیت را نیز برطرف کند. این روش نیز نوعی روش برچسب گذاری مجدد به شمار می‌آید. در این مقاله فقط الگوریتمی که در حالت تعداد خوشه های برابر تناظر را تشخیص می دهد، آورده خواهد شد.

در این الگوریتم هر یک از خوشه ها به صورت یک الگوی بیتی نشان داده می شوند. هر بیت در این الگو متناظر با یکی از داده ها در مجموعه داده ای می باشد. در صورت وجود آن داده در خوشه، بیت متناظر آن یک و در غیر اینصورت صفر در نظر گرفته می شود. الگوی بیتی برای هر خوشه به صورت رابطه (۱) می باشد.

$$B_{im} = b_1 b_2 ... b_N \quad i \in [1..C], m \in [1..P] \qquad (۱)$$

در رابطه (۱)، i شماره خوشه بندی، m شماره خوشه، N تعداد اشیاء داده ای، C تعداد خوشه بندی ها و P تعداد خوشه در هر خوشه بندی می باشد.

جهت تشخیص تناظر بین خوشه ها ابتدا یکی از خوشه‌بندی‌ها به عنوان مرجع در نظر گرفته می شود و سپس با استفاده از الگوریتم CCD (Correspondence Clusters Detection) تناظر هر کدام از خوشه ها در خوشه بندی های دیگر با خوشه‌های خوشه بندی مرجع بدست می آید.

جهت تشخیص تناظر بین خوشه ها ابتدا فاصله بین هر جفت خوشه در دو خوشه بندی با استفاده از فاصله همینگ محاسبه می گردد، سپس جفت خوشه هایی که کمترین فاصله را با یکدیگر دارند (تا زمانی که برای هر خوشه یک متناظر پیدا شود) به عنوان جفت خوشه های متناظر در دو خوشه بندی، انتخاب می گردند. این کار به صورت تکراری بین خوشه بندی مرجع با دیگر خوشه‌بندی ها انجام می شود. شبه کد الگوریتم CCD در الگوریتم ۱ آورده شده است.

۳-۲- خوشه بندی توافقی وزنی

در الگوریتم بهبود یافته جهت خوشه بندی توافقی، به ازاء هر داده در مجموعه داده ای، برداری به صورت $y_i = \langle \pi_1(x_i), \pi_2(x_i), ..., \pi_C(x_i) \rangle$ خواهیم داشت. جهت انجام خوشه بندی توافقی ابتدا P خوشه ایجاد می‌گردد و مرکز هر یک از خوشه ها را از رابطه (۲) بدست می‌آوریم.

$$cntr_m = \langle m, m, ..., m \rangle \quad |cntr_m| = C, m \in [1..P] \qquad (۲)$$

الگوریتم ۱ – تشخیص تناظر خوشه ها

```
Input:  BP[1..P, 1..C]: bitArray
        BClus: int

Output: Corr[1..P, 1..C]: int

Dist: sortedList(Distance, cIdx_BClus, cIdx_y)
Sel_BClus, Sel_y: intList

foreach π_y ∈ {π_1...π_C} − {π_BClus}
  for i ∈ [1..P]
    for i ∈ [1..P]
      Dist.Add(HamDist(BP[i,BClus], BP[j,y]), i, j)
    end for
  end for
  repeat
    e = SelectNext(Dist)
    if(e.cIdx_BClus ∉ Sel_BClus and e.cIdx_y ∉ Sel_y)
      Sel_BClus.Add(e.cIdx_BClus)
      Sel_y.Add(e.cIdx_y)
      Corr[e.cIdx_y, y] = e.cIdx_BClus
    end if
  until | Sel_BClus |= P
  clear Dist, Sel_BClus, Sel_y
end foreach
```

لازم به ذکر است که رابطه (۲) تنها در زمان ایجاد اولیه خوشه ها استفاده می‌گردد.

پس از آن در یک فرآیند تکراری ابتدا داده ها را به نزدیکترین خوشه نسبت داده و سپس مرکز خوشه ها را دوباره محاسبه می کنیم. جهت تعیین فاصله y_i تا $cntr_m$ از ترکیب فاصله همینگ و وزن اختصاص داده شده به هر یک از خوشه بندی ها استفاده می کنیم. شبه کد الگوریتم WHD (Weighted Hamming Distance) جهت تعیین فاصله دو بردار با استفاده از وزن مشخص، در الگوریتم ۲ آورده شده است.

الگوریتم ۲ – فاصله همینگ وزنی

```
Input:  y, cntr_m: clusteringVector
        Weight[1..C]: float

Output: Dist: float

Dist = 0
```

```
for i ∈ [1..C]
  if(y[i] != cntr_m[i])
    Dist = Dist + Weight[i]
  end if
end for
```

وزن در نظر گرفته شده برای هر یک از خوشه بندی ها عددی بین ۰,۰۰۱ تا ۰,۹۹۹ می باشد.

در روش پیشنهادی از شاخص Davies-Bouldin جهت تعیین وزن هر یک از خوشه ها استفاده می شود. مقادیر کوچک برای این شاخص میزان فشردگی و تفکیک‌شدگی خوب خوشه ها را نشان می دهد [۱۳]. روش محاسبه این شاخص در [۱۳] آمده است و به دلیل عدم وجود فضای کافی از ذکر روابط لازم جهت محاسبه شاخص Davies-Bouldin خودداری می کنیم. این شاخص را در مقاله از این پس با DB نشان می دهیم. با استفاده از نرمال سازی Min-Max که در رابطه (۳) آمده است، مقدار هر کدام از DB ها را به گونه ای تغییر می دهیم که بین ۰,۰۰۱ تا ۰,۹۹۹ قرار گیرد. سپس با استفاده از رابطه (۴) وزن هر خوشه را بدست می آوریم.

$$MinMaxNorm(DB_i) = \frac{DB_i - Min\{DB\}}{Max\{DB\} - Min\{DB\}}(0.999 - 0.001) + 0.001 \quad (۳)$$

$$W_i = 1 - MinMaxNorm(DB_i) \quad i \in [1..P] \quad (۴)$$

در نهایت نیز با استفاده از الگوریتم خوشه بندی توافقی وزنی که نام WCC (Weighted Consensus Clustering) را برای آن در نظر گرفته ایم، عمل ترکیب خوشه بندی ها و بدست آوردن خوشه بندی نهایی انجام می شود. شبه کد الگوریتم WCC در الگوریتم ۳ آورده شده است.

الگوریتم ۳ – خوشه بندی توافقی وزنی

Input: $Y[1..N]$: **clusteringVectorArray**
$W[1..P]$: **float**

Output: π^* : Final Clustering with P clusters

Intialize π^* // Create clusters center
repeat
$S_i = \{y \mid \pi^*(y) = i\}$ $i \in [1..P]$ // i^{th} cluster
// Compute the center of each cluster
for $m \in [1..P]$

لازم به ذکر است که شماره خوشه ها در خوشه بندی ها باید قبل از اجرای الگوریتم WCC، با استفاده الگوریتم CCD و ماتریس تناظر بدست آمده توسط آن، تغییر کند.

۴- تحلیل پیچیدگی

تعدا تکرار ها در الگوریتم پیشنهادی WCC، متغیر و وابسته به مجموعه داده ای است. تعداد این تکرار ها در آزمایشات ما کمتر از ۶ بوده است. روش های بسیاری جهت بهبود سرعت برای kmeans توسعه داده شده است که می‌توان از آنها برای الگوریتم WCC نیز استفاده نمود.

پیچیدگی مکانی الگوریتم WCC نیز $O(NC)$ است که N تعداد داده ها و C تعداد خوشه بندی ها اولیه است.

۵- مجموعه های داده ای

ما الگوریتم پیشنهادی را بر روی چهار مجموعه داده ای ارزیابی کرده ایم. این چهار مجموعه iris، glass، vehicle و segment می باشند. iris شامل داده هایی در مورد سه نوع گل زنبق است. glass شامل داده هایی در رابطه با اجزاء شیمیایی تشکیل دهنده ۶ نوع شیشه است. vehicle شامل داده هایی در مورد ۴ نوع وسیله نقلیه است. segment نیز شامل داده هایی در مورد قطعه بندی تصاویر است. این مجموعه های داده ای از UCI Machine Learning Repository می‌باشند. مشخصات این چهار مجموعه داده ای در جدول ۱ آورده شده است.

جدول ۱ – مشخصات مجموعه های داده ای

نام	فیلدها	داده ها	کلاس ها	خوشه ها

iris	۴	۱۵۰	۳	۳
glass	۹	۲۱۴	۷	۶
vehicle	۱۸	۸۴۶	۴	۴
segment	۱۹	۲۳۱۰	۷	۷

۶- نتایج

اجتماع خوشه بندی های اولیه، برای هر مجموعه داده ای شامل ۱۰ خوشه بندی می باشد. هر یک از این ۱۰ خوشه بندی، با استفاده از یکی از سه روش خوش بندی KMeans، XMeans و EM در نرم افزار weka نسخه ۳،۵،۸ بدست آمده است. هر کدام از این سه روش را با پارامترهای اولیه متفاوت جهت تولید ۱۰ خوشه بندی مورد استفاده قرار داده ایم.

جهت ارزیابی خوشه بندی نهایی از سه معیار دقت (Accuracy)، فشردگی (Compactness) و Rand استفاده شده است.

معیار دقت با استفاده از برچسب کلاس ها میزان دقت روش خوشه بندی در انتساب داده ها به خوشه ها را نشان می دهد. لازم به ذکر است که برچسب کلاس ها تنها برای ارزیابی نتایج استفاده می شوند و در فرآیند خوشه بندی دخالتی ندارند. رابطه (۵) این معیار را نشان می دهد.

$$Accuracy(\pi)=\frac{\sum_{m=1}^{P} majority(C_m \mid L_m)}{N} \qquad (۵)$$

$majority(C_m \mid L_m)$ در رابطه (۵) تعداد داده ای هایی از یک کلاس مشخص است که دارای اکثریت در خوشه C_m می باشند.

معیار فشردگی، میانگین فاصله بین هر دو داده در خوشه‌های مشابه را اندازه گیری می کند. رابطه (۶) این معیار را نشان می دهد.

$$Compactness(\pi)=\frac{1}{N}\sum_{m=1}^{P} n_m\left(\frac{\sum_{x_i,x_j\in C_m} d(x_i,x_j)}{n_m(n_m-1)/2}\right) \qquad (۶)$$

$d(x_i,x_j)$ در رابطه (۶) فاصله بین دو داده x_i و x_j می‌باشد.

معیار Rand [۱۴] نیز تعداد تصمیمات درست و نادرست در قرار دادن جفت داده ها در خوشه ها را با استفاده از برچسب کلاس ها شمارش می کند. این معیار نیز در رابطه (۷) آورده شده است.

$$Rand=\frac{(SS+DD)}{SS+SD+DS+DD} \qquad (۷)$$

در رابطه (۷) SS تعداد جفت داده هایی است که در یک خوشه می باشند و از کلاس مشابهی هستند؛ SD تعداد جفت داده هایی است که در یک خوشه می باشند و از کلاس های متفاوتی هستند؛ DS تعداد جفت داده هایی است که در دو خوشه متفاوت می باشند و از کلاس مشابهی هستند؛ و DD تعداد داده هایی است که در دو خوشه متفاوت می باشند و از کلاس های متفاوتی هستند [۱۴].

جهت ارزیابی میزان توافق خوشه بندی نهایی با خوشه بندی های اولیه نیز از معیار ANMI استفاده می کنیم. این معیار در [۴] آورده شده است و ما به دلیل عدم وجود فضای کافی از ذکر جزئیات روش محاسبه آن خودداری می کنیم.

جهت مقایسه دو الگوریتم IVC و WCC، ابتدا الگوریتم IVC را بر روی ۱۰ خوشه بندی اولیه و برای هر مجموعه داده ای آزمایش کردیم. سپس الگوریتم پیشنهادی در این مقاله (WCC) را بر روی همان داده‌ها اعمال کردیم. جداول ۲ و ۳ به ترتیب نتایج حاصل از ارزیابی الگوریتم های IVC و WCC را با استفاده از معیارهای ذکر شده، نشان می دهند. در جداول ۲ و ۳ از حروف A و C به ترتیب جهت نشان دادن معیار دقت و معیار فشردگی استفاده شده است.

جدول ۲: نتایج ارزیابی الگوریتم IVC

مجموعه	A	C	Rand	ANMI
iris	۰،۸۸۷	۱،۸۴۳	۰،۸۵۵	۰،۸۱۰

glass	۰,۵۳۷	۳,۳۳۵	۰,۶۰۰	۰,۶۷۲
vehicle	۰,۳۹۴	۱۸۰,۴۵۷	۰,۵۸۵	۰,۶۲۹
segment	۰,۶۶۳	۲۰۷,۰۸۷	۰,۸۲۸	۰,۷۲۶

جدول ۳: نتایج ارزیابی الگوریتم WCC

مجموعه	A	C	Rand	ANMI
iris	**۰,۹۱۰**	۱,۸۴۵	**۰,۸۷۳**	**۰,۸۱۶**
glass	۰,۵۳۷	**۳,۲۹۹**	**۰,۶۱۷**	**۰,۶۸۳**
vehicle	**۰,۴۴۸**	**۱۵۷,۶۶۸**	**۰,۶۰۴**	۰,۵۸۲
segment	**۰,۶۷۴**	**۲۰۱,۹۷۹**	**۰,۸۳۶**	۰,۷۲۰

لازم به ذکر است که مقادیر حاصل از ارزیابی معیار دقت و Rand عددی بین صفر و یک می باشد. مقادیر بزرگتر برای این دو معیار و مقدار کوچکتر برای معیار فشردگی کیفیت بالاتر خوشه بندی را نشان می دهند.

مقادیری که در جدول شماره ۳ به صورت پر رنگ نشان داده شده است نشان دهنده این مسئله است که WCC با توجه به معیارهای ارزیابی اغلب بهتر و کاراتر عمل می‌کند.

در رابطه با معیار ANMI لازم به ذکر است که این شاخص نشان دهنده کیفیت خوشه بندی نمی باشد بلکه تنها میانگین میزان توافق بین خوشه بندی نهایی و خوشه بندی های اولیه را نشان می دهد. بنابراین کاهش این شاخص در الگوریتم WCC در برخی از موارد نسبت به الگوریتم IVC، به دلیل وزن دهی به خوشه بندی‌های اولیه می باشد که این مسئله می تواند بر روی میانگین میزان توافق نهایی بدست آمده، تأثیرگذار باشد.

۷- نتیجه گیری

ما در این مقاله سه الگوریتم CCD، WHD و WCC را مطرح کردیم. الگوریتم CCD مسئله تناظر بین خوشه‌ها در خوشه بندی های مختلف را حل می کند که این مسئله در اکثر روش های خوشه بندی توافقی لحاظ نمی شود. الگوریتم WHD با ترکیب فاصله همینگ و وزن هر خوشه بندی، فاصله بردارهای مربوط به خوشه‌بندی ها را محاسبه می کند. وزن خوشه ها نیز با استفاده از معیار Davies-Bouldin تعیین می گردد.

در نهایت الگوریتم WCC نیز خوشه بندی توافقی وزنی را انجام می دهد. ارزیابی هایی که انجام داده ایم نشان می دهد که روش پیشنهادی در این مقاله جهت خوشه‌بندی توافقی، کاراتر از روش خوشه بندی IVC عمل می کند. لازم به ذکر است که روش IVC یکی از روش های بسیار کارا در این زمینه است که در [۳] با یازده الگوریتم مطرح دیگر مقایسه شده است.

کارهای آینده ای که می تواند در راستای این کار پژوهشی انجام شود، اول بررسی معیارهای دیگر به جای DB جهت انجام خوشه بندی توافقی وزنی و ارزیابی نتایج آن، دوم انجام خوشه بندی توافقی وزنی بر روی خوشه بندی هایی با تعداد خوشه متفاوت، می باشد.

سپاسگزاری

از جناب آقای مهندس احمد خادم زاده عضو هیئت علمی دانشگاه آزاد اسلامی واحد مشهد که ما را در تهیه این مقاله یاری نمودند، کمال تشکر را داریم.

۸- مراجع

[۱] P. Berkhin, "Survey on Clustering Data Mining Techniques", Grouping Multidimensional Data, pp. 25-71, 2006.
[۲] S. B. Kotsiantis, P. E. Pintelas, "Recent Advanced in Clustering: A Brief Survey", WSEAS Transactions on Information Science and Applications 1, pp. 73-81, 2004.
[۳] N. Nguyen, R. Caruana, "Consensus Clustering", Proceedings of the Sixth International Conference on Data Mining (ICDM), pp. 607-612, 2007
[۴] A. Strehl, J. Ghosh, "Cluster ensembles – a knowledge reuse framework for combining partitionings", The Journal of Machine Learning Research, vol. 3, pp. 583-617, 2003.
[۵] A. Topchy, A. K. Jain, W. Punch, "Combining multiple weak clusterings", In Proceedings of the Third IEEE International Conference on Data Mining (ICDM), pp 331-338, 2003.
[۶] A. Topchy, A. K. Jain, W. Punch, "A mixture model for clustering ensembles", In Proceedings of AIAM Data mining, pp 379-390, 2004.
[۷] A. Gionis, H. Mannila, P. Tsaparas, "Clustering Aggregation", In Proceedings of Twenty-fitst International Conference on Data Engineering (ICDE), pp. 341-352, 2005
[۸] S. Dudoit, J. Fridlyand, "Bagging to improve the accuracy of a clustering procedure", Bioinformatics, vol. 19, pp. 1090–1099, 2003.
[۹] B. Fischer, J. M. Buhmann, "Path-based clustering for grouping of smooth curves and texture segmentation", IEEE Transaction on Pattern Analysis and Machine. Intelligence, vol. 25, pp.

513–518, 2003.

[۱۰] X. Z. Fern, C. E. Brodley, "Random projection for high dimensional data clustering: A cluster ensemble approach", In Proceedings of the 20th ICML, pp. 168-193, 2003.

[۱۱] A. L. Fred, A. K. Jain. "Data clustering using evidence accumulation", In Proceedings of the 16th International Conference on Pattern Recognition, pp. 276-280, 2002.

[۱۲] ح. زجاجی، م. علیشاهی، ب. شاکری، ح. دلداری، "الگوریتمی جهت یافتن تناظر بین خوشه ها در خوشه بندی ها متفاوت به منظور استفاده در خوشه بندی توافقی"، دومین کنفرانس داده کاوی ایران (ICDM)، ۱۳۸۷

[۱۳] M. D. Toledo, "A Comparison in Cluster Validation Techniques", Master's thesis, University of Puerto, 2005

[۱۴] E. Amigó, J. Gonzalo, J. Artiles, F. Verdejo, "A comparison of extrinsic clustering evaluation metrics based on formal constraints", The journal of Information Retrieval, 2008