# ارایه یک روش بر مبنای الگوریتمهای تکاملی برای خوشه بندی ترکیبی

مهدی محمدی[۱]، جواد عظیمی[۲]،رضا داودی[۳]، عادل رحمانی[۴]
دانشکده مهندسی کامپیوتر-دانشگاه علم و صنعت ایران

## چکیده

**خوشه بندی ترکیبی روشی جدید در خوشه بندی می باشد که از ترکیب نتایج روشهای خوشه بندی مختلف بدست می آید. در مقابل روشهای کلاسیک و متداول خوشه بندی که از پایداری اندکی برخوردارند ، خوشه بندی ترکیبی دارای پایداری بسیار بالایی می باشدکه از آن یک روش قدرتمند در خوشه بندی اطلاعات می سازد. دو مرحله مهم در خوشه بندی ترکیبی عبارتند از: الگوریتمهای ابتدائی خوشه بندی که خوشه بندی اولیه را انجام می دهد و سپس جمع بندی نتایج این الگوریتمهای اولیه برای بدست آوردن نتیجه نهائی.در مقاله حاضر از الگوریتم ژنتیک به عنوان مرحله دوم خوشه بندی ترکیبی استفاده شده است.الگوریتم های ژنتیک بر اساس توانایی بالایی که در جستجوی فضای حالت مسئله دارند، همواره می توانند جواب بهینه یا نزدیک بهینه را در فضای حالت مسئله بدست آورند.برای افزایش کارایی الگوریتم ژنتیک ارایه شده و بهبود در صد خطای حاصل از خوشه بندی، یک عملگر جهش خاص منظوره، ارایه و بکار گیری شده است . نتایج بکار گیری الگوریتم ژنتیک ارایه شده برروی چند پایگاه داده متداول نشانگر آن است که روش ما در بیشتر مواقع از روشهای پیشینی که برای خوشه بندی ترکیبی استفاده می شد، خطای کمتری تولید می کند.**

## کلمات کلیدی

**الگوریتهای ژنتیک ، جهش هوشمند ، خوشه بندی ترکیبی ،ماتریس همبستگی.**

# A Genetic Based Clustering Ensemble Method

Mehdi Mohammadi, Javad Azimi, Reza Davoodi, Adel Rahmani

Computer Engineering Department-Iran University of Science and Technology

## Abstract

Clustering ensemble is a new approach in clustering algorithms. Clustering ensemble combines the output of different clustering algorithms to generate final clusters. Clustering ensemble algorithms are more stable and robust than any single clustering algorithm. There are two different steps in clustering ensemble: first initial clustering algorithm and second, consensus function which obtains the final partitions from the result of step one. In this paper we use the GA (genetic algorithm) as a consensus function. Since genetic algorithms have a good search space ability, they can obtain optimal or near optimal solutions per each complex and large search space. A special mutation operation is proposed to increase the ability of proposed method in generating the final partitions and skipping the local optimal. Experimental results demonstrate that the proposed GA based clustering method usually obtains acceptable error rate on common datasets. Comparison between our method and previous clustering ensemble methods shows that our algorithm has more ability to find the correct final partitions.

## Keywords

Genetic algorithm, intelligence mutation, clustering ensemble, Co-association matrix

[1] Mh_mohammadi@comp.iust.ac.ir
[2] ja_azimi@comp.iust.ac.ir
[3] sr_davodi@comp.iust.ac.ir
[4] rahmani@iust.ac.ir

## ۱- مقدمه

ایده اصلی خوشه بندی اطلاعات، جداکردن نمونه ها از یکدیگر و قرار دادن آنها در گروههای شبیه به هم می باشد، به این معنی که نمونه های شبیه به هم باید در یک گروه قرار بگیرند و نمونه های هر گروه دارای خصوصیات متفاوتی نسبت به گروهای دیگر باشند. هر یک از الگوریتمهای خوشه بندی، با توجه به توانایی های مختلفی که در خوشه بندی دارا هستند ، داده ها را به صورتهای متفاوتی دسته بندی می نمایند. به همین دلیل با بکار گیری روشهای مختلف خوشه بندی و ترکیب آنها با یکدیگر این انتظار را داریم که با گرفتن نقاط قوت هر یک از روشهای خوشه بندی، به ساختار خوشه های بهینه دست پیدا کنیم. در واقع هدف اصلی خوشه بندی ترکیبی جستجوی بهترین خوشه ها با استفاده از ترکیب نتایج الگوریتمهای دیگر است. خوشه بندی ترکیبی می تواند جوابهای مناسب تری چه از لحاظ دقت و چه از لحاظ پایداری، نسبت به روشهای عادی خوشه بندی ، تولید کند[۱].

بطور کلی دو مساله مهم در خوشه بندی ترکیبی وجود دارد: اول بکار بردن چند باره الگوریتمهای عادی خوشه بندی[2](مانند K-Means..) و بدست آوردن ماتریس همبستگی از روی آنها و دوم بدست آوردن خوشه های نهایی از روی ماتریس همبستگی تولید شده در مرحله قبل.

الگوریتهای ژنتیک همواره به عنوان یک روش جستجوی عمومی در بسیاری از مسایل بهینه سازی بکار رفته اند[3] .توانایی اصلی این الگوریتم های ، جستجوی فضاهای پیچیده و بدست آوردن جواب نزدیک بهینه می باشد. این الگوریتها توانایی خود را در بسیاری از مسایل مختلف نشان داده اند. از جمله این مسایل می توان به بکار گیری الگوریتهای ژنتیک در مسایل کلاسبندی نمونه ها ،آموزش شبکه های عصبی ، آموزش سیستمهای تشخیص گفتار و ... اشاره کرد[4,5,6,7]. در تمامی این مسایل با تعریف مناسبی از الگوریتم های ژنتیک ، توانسته اند به نتایج مناسبی نسبت به روشهای معمول دست یابند.

در مقاله حاضر یک روش جدید برای بدست آوردن خوشه های نهایی از روی ماتریس همبستگی، بر اساس الگوریتهای ژنتیک ارایه شده است.برای این منظور ابتدا ماتریس همبستگی را بوسیله اجرای چند باره روش K-Means بدست می آوریم[8] و سپس الگوریتم ژنتیک ارایه شده را برروی ماتریس همبستگی بکار می بریم. ساختار ژنی و همچنین نوع عملگر های مورد نیاز برای الگوریتم ژنتیک در بخشهای بعد توضیح داده خواهند شد. از آنجایی که الگوریتهای ژنتیک همواره با مشکل بهینه های محلی مواجه هستند یک عملگر جهش خاص منظوره برای بهبود کارایی الگوریتم و رهایی از بهینه های محلی طراحی و پیاده سازی کرده ایم و در پایان نیز الگوریتم ارایه شده را بروی چند پایگاه داده تست و نتایج آن را با روشهای مختلف خوشه بندی ترکیبی مقایسه نموده ایم. نتایج گزارش شده نما یانگر توانایی الگوریتم ژنتیک ارایه شده ، چه از لحاظ پایداری و چه از لحاظ دقت، برروی پایگاه داده های مختلف می باشد.

در ادامه مقاله در بخش دوم چگونگی محاسبه ماتریس همبستگی که به عنوان مرحله اول در خوشه بندی ترکیبی مطرح است، شرح داده می شود.در قسمت سوم الگوریتم ژنتیکی که برای بدست آوردن خوشه های نهایی از روی ماتریس همبستگی ارایه شده است ، مورد بحث قرار می گیرد. در بخش چهارم نتایج آزمایشات و ارزیابی الگوریتم ژنتیک ارایه شده است و مقایسه کا ملی بین روش ژنتیک مطرح شده در این مقاله و روشهای متدوال برای خوشه بندی ترکیبی ، انجام خواهد شد. و در نهایت در بخش پنجم نتیجه گیری نهایی از الگوریتم ژنتیک ارایه شده در این مقاله خواهیم داشت.

## ۲- ماتریس همبستگی

فرض کنید مجموعه داده $D$ شامل $N$ نقطه (نمونه) در فضای $d$ بعدی است. داده های ورودی را می توان بصورت یک ماتریس الگوی $N \times d$ و یا یک ماتریس عدم تشابه $N \times N$ در نظر گرفت. فرض کنید $X=\{X1\ldots\ XB1\}$ مجموعهٔ زیر مجموعه نمونه های ماست که از نمونه های اولیه استخراج شده اند. هر یک از الگوریتمهای انتخابی هنگامیکه بر روی زیر مجموعه نمونه های موجود در $X$ اجرا شوند نتایج $P = \{P1,\ldots, PB1\}$ را تولید می کنند. هر $P_i$ مجموعه ای از خوشه هاست یا به عبارت دیگر $P_i = \{C_1^i, C_2^i,\ldots C_{k(i)}^i\}$ و $X_i = C_1^i \cup C_2^i \ldots \cup C_{k(i)}^i$ بطوریکه k(i) تعداد خوشه ها در i امین خوشه بندی است.

الگوریتم پایه مورد استفاده الگوریتم k-means است. در اولین گام، الگوریتم k-means را بر روی $X=\{X1\ldots\ XB1\}$ اجرا می کنیم تا بتوانیم با استفاده از $P_i$ های تولید شده ماتریس همبستگی را بصورت زیر بدست آوریم:

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

$$\varphi(a,b)=\begin{cases}0 & if \quad a \neq b\\ 1 & if \quad a = b\end{cases} \qquad (۱)$$

تابع $\varphi(P_i(a),P_i(b))$ در صورتیکه دو عنصر a و b در خوشه بندی $P_i$ در یک خوشه قرار گرفته باشند مقدار ۱ و در غیر اینصورت مقدار ۰ بر می گرداند. در واقع ماتریس همبستگی نشان می دهد که هر دو نمونه در $P_i$ های مختلف چند بار در یک خوشه قرار گرفته اند.[1]

## ۳- خوشه بندی ترکیبی با کمک الگوریتمهای ژنتیک

الگوریتهای ژنتیک روشهایی هستند که بر اساس جستجوی فضای حالت بصورت تصادفی، سعی در یافتن بهترین جواب موجود در فضای حل مسئله را دارند. اما این جستجوی تصادفی توسط عملگرهایی (بر گرفته از طبیعت) به سمت پیدا کردن بهترین جواب راهنمایی می شوند.بنا براین نمی توان این الگوریتم ها را کاملا تصادفی دانست. بر اساس این خصوصیت ها الگوریتهای ژنتیک توانایی بسیار زیادی در جستجو در فضاهای حالت بسیار پیچیده و بزرگ را دارا می باشند.[8]
شکل (۱) حلقه اجرای یک الگوریتم ژنتیک به همراه عملگر های مختلف آن و چگونگی ترتیب اجرای هر یک از این عملگرها را نشان می دهد.[8]

```
Begin
1.  t=0 ;
2.  initialize population P(t);
3.  compute fitness P(t);
4.  t=t+1;
5.  if termination criterion achieved go to step 10
    ;
6.  select P(t) from P(t-1);
7.  crossover P(t);
8.  mutation  P(t);
9.  go to step 3;
10. Output best and stop
End
```

**شکل(۱):حلقه اجرای الگوریتم ژنتیک**

قسمتهای مختلف الگوریتم ارایه شده در ادامه توضیح داده می شود.

### ۳-۱- ساختار کروموزوم ها

اولین مرحله در تعریف یک الگوریتم ژنتیک ، تعریف ساختار کروموزوم آن می باشد. در حقیقت ساختار کروموزوم ارتباط بین مسئله دنیای واقعی و فضای حل مسئله ایجاد می کند و به عنوان پلی بین دنیای خارج و دنیای قابل پیاده سازی توسط الگوریتم ژنتیک ایفای نقش می کند. در الگوریتم ارایه شده، ساختار هر کروموزوم دارای دو قسمت است:

۳-۱-۱- **قسمت نمونه ها**: قسمت نمونه ها شامل تمامی اعضای پایگاه داده می باشد و هر کدام از ژنهای یک کروموزوم ارایه کننده یکی از اعضای پایگاه داده می باشد.ترتیب قرار گرفتن هر یک از اعضا در یک کروموزوم به صورت تصادفی می باشد که در مرحله مقدار دهی اولیه ، مشخص می شود. شکل (۲) یک مثال از قسمت نمونه ها را نشان می دهد. در این شکل ، پایگاه داده شامل ۱۵ عضو می باشد که با ترتیب تصادفی در این کروموزوم جای گرفته اند.

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 6 | 7 | 12 | 13 | 11 | 2 | 15 | 9 | 1 | 4 | 3 | 8 | 10 | 14 |

**شکل(۲): یک نمونه از قسمت نمونه های یک کروموزوم**

۳-۱-۲- **قسمت کرانه ها**:مرز هر یک از خوشه ها در یک کروموزوم در این قسمت مشخص می شود.طول این قسمت از کروموزوم برابر تعداد خوشه ها +۱ می باشد.شکل (۳) یک نمونه از قسمت کرانه ها را نشان می دهد. این مثال پایگاه داده شامل ۳ خوشه می باشد که در قسمت کرانه ها مرز هر یک از خوشه ها مشخص شده است.خوشه اول شامل عناصر قسمت نمونه ها از ۰ تا ۵ ، خوشه دوم از ۶ تا ۱۱ و خوشه سوم از ۱۲ تا ۱۵ می باشد. مقادیر قسمت کرانه ها برای هر کروموزوم بصورت تصادفی در مرحله مقدار دهی اولیه مشخص می شود. اولین و آخرین ژن در قسمت کرانه ها همواره ثابت می باشند(اولین ژن ۰ و آخرین ژن تعداد نمونه ها )

| | | | |
|---|---|---|---|
| 0 | 5 | 11 | 15 |

**شکل(۳): یک نمونه از قسمت کرانه ها**

### ۳-۲- عملگر همبری

نقش عملگر همبری تولید کروموزومهای جدید از روی کروموزوم های قدیمی می باشد.عملگر همبری یک فرایند احتمالی است که اطلاعات را بین دو کروموزوم جابجا کرده و بر این اساس دو کروموزوم جدید تولید می کند. در الگوریتم ژنتیک ارایه شده ، ما از عملگر همبری ای به نام Cut and Crossfill استفاده کرده ایم و این عملگر را فقط برروی قسمت نمونه ها بکار برده ایم[9].Cut and Crossfill یک عدد صحیح را بصورت تصادفی در بازه $[1,l\text{-}1]$ ($l$ طول قسمت نمونه ها می باشد) با نام $k$ تولید کرده و قسمتهایی از دو کروموزوم را که در قسمت سمت چپ نقطه $k$ قرار بگیرند را با هم عوض می کند و اطلاعات به این طریق بین دو کروموزوم عوض می کند و دو کروموزوم جدید تولید می کند.اگر حین این عمل در کروموزوم ژن تکراری تولید شود ، Cut and Crossfill آنها را حذف کرده و نمونه هایی از پایگاه داده را که در کروموزوم ظاهر نشده اند را به جای آنها قرار می دهد. به عبارت دیگر تضمین می کند که کروموزومهای تولید شده معتبر هستند.

### ۳-۳-عملگر جهش

عملکر دیگری که برای تولید کروموزوم جدید در هر الگوریتم ژنتیک استفاده می شود عملگر جهش است.تفاوت عملگر جهش با عملگر همبری این است که عملگر جهش فقط برروی یک کروموزوم اعمال شده و یک کروموزوم جدید تولید می کند. در حقیقت عملگر جهش نمونه هایی جدیدی را تولید می کند که درجمیعت حاضر موجود نیستند بنا براین توانایی جستجوی الگوریتم ژنتیک با عملگر جهش افزایش می یابد. برای بالا بردن کارایی الگوریتم ژنتیک ارایه شده از ۳ نوع عملگر جهش استفاده کرده ایم

۳-۳-۱-**Swap Mutation**: این نوع عملگر جهش جای دو ژن را در یک کروموزوم عوض می کند و با این تغییر کوچک ، یک کروموزوم جدید تولید می کند. ما این عملگر جهش را فقط برروی قسمت نمونه ها اعمال کرده ایم.[9]

۳-۳-۲-**Creep Mutation**: این عملگر جهش یک ژن را در یک کروموزوم به صورت تصادفی انتخاب کرده و یک مقدار تصادفی $r$ به آن اضافه یا از آن کم می کند( بصورت تصادفی). این نوع عملگر جهش را فقط برروی قسمت کرانه ها اعمال کرده ایم و با کمک آن محدوده خوشه ها را می توانیم عوض کنیم.[9]

**۳-۳-۳-عملگر جهش خاص منظوره**: این عملگر جهش را بطور مخصوص برای این مسئله خاص طراحی کرده ایم و آن را بعد از گذشتن چندیدن اجرا از حلقه ژنتیک.( در اجراهای نهایی حلقه ژنتیک ، وقتی که مسئله جواب تقریبی خود را یافته باشد) برروی قسمت نمونه ها بکار می بریم. نقش اصلی این عملگر جهش علاوه بر افزایش کارایی الگوریتم ، فرار از بهینه های محلی نیز می باشد زیرا این عملگر جهش همواره بهترین ژن را برای اعمال جهش انتخاب می کند. بخاطر این خصوصیت می توان آن را عملگر جهش هوشمند نیز نام گذاری کرد.مراحل اجرای این عملگر جهش بصورت زیر است.

-یک خوشه بصورت تصادفی انتخاب می شود.(C)

-عنصری از C که کمترین همبستگی را با اعضای هم خوشه خود دارد(بر اساس ماتریس همبستگی) انتخاب می شود.(X).

-میزان همبستگی بین X و تمام خوشه ها محاسبه می شود(بر اساس ماتریس همبستگی)

-X به خوشه ای که بیشترین همبستگی را با آن دارد انتقال می یابد. ممکن است این خوشه همان خوشه ابتدایی X یعنی C نیز باشد.

-در صورت لزوم محدوده خوشه ها در قسمت کرانه ها به روز می شود.

## ۳-۴-تابع ارزیابی

مهمترین قسمت در تعریف یک الگوریتم ژنتیک ،تعریف تابع ارزیابی است. در حقیقت الگوریتم های ژنتیک با بهینه کردن مقدار تابع ارزیابی ، می توانند به جواب بهینه برای مسئله دنیای واقعی نیز برسند. تابع ارزیابی میزان مناسب بودن یک کروموزم و در نهایت میزان مناسب بودن جواب مسئله در دنیای واقعی را مشخص می کند.

درالگوریتم ارایه شده، ما یک تابع ارزیابی ۳ مرحله ای تعریف کرده ایم که مراحل تعریف این تابع در ادامه آمده است.

۳-۴-۱-**ارزیابی داخلی** : در قسمت اول تابع ارزیابی ، براساس ماتریس همبستگی ، میزان همبستگی اعضای یک خوشه را با هم محاسبه می کنیم. رابطه (2) چگونگی محاسبه این مقدار را نشان می دهد. در این رابطه $N_{C_k}$ تعداد اعضای خوشه k ام و $Co_association(m(i),m(j))$ میزان همبستگی عنصر iام و عنصر j ام از یک کروموزوم را نشان می دهد. این میزان همبستگی را برای تک تک خوشه ها محاسبه کرده و بر روی این اعداد میانگین می گیریم و این مقدار را به عنوان تابع ارزیابی داخلی برای کروموزوم مورد نظر لحاظ می کنیم. رابطه (3) چگونگی محاسبه این میانگین را نشان می دهد. در این رابطه I تعداد خوشه های پایگاه داده مورد آزمایش را نشان می دهد

$$Fitness_C(k)=\frac{\sum_{i=1}^{N_{C_k}}\sum_{j=i+1}^{N_{C_k}} Co_association(m(i),m(j))}{N_{C_k}\left(N_{C_k}-1\right)/2} \qquad (۲)$$

$$Intra=\frac{\sum_{i=1}^{I} Fitness_C(i)}{I} \qquad (۳)$$

۳-۴-۲- **ارزیابی خارجی**: همانطوری که می دانیم خوشه های بهینه نهایی خوشه هایی هستند که علاوه بر اینکه اعضای داخلی هر خوشه باید همبستگی زیادی با یکدیگر داشته باشند ، خوشه های مختلف نیز باید نسبت به هم تا آنجا که ممکن است نا همبسته باشند. میزان همبستگی خوشه های مختلف به هم را در این قسمت تابع ارزیابی محاسبه می کنیم. برای این منظور میزان این همبستگی را بین هر دو خوشه حساب کرده و برروی این مقادیر میانگین می گیریم. رابطه (4) چگونگی محاسبه این مقدار را برای هر ژن نشان می دهد.. مقدار نهایی تابع ارزیابی خارجی را به عنوان یک جریمه برای کروموزوم در نظر می گیریم. اگر کروموزومی دارای خوشه های نا همبسته باشد میزان جریمه در نظر گرفته شده برای آن نیز کاهش می یابد. رابطه (5) مقدار نهایی جریمه برای هر کروموزوم را نشان می دهد.

$$Penalty_C(C_k,C_l)=\frac{\sum_{i=1}^{N_{C_k}}\sum_{j=1}^{N_{C_l}} Co_association(m(i),m(j))}{N_{C_k}*N_{C_l}} \qquad (۴)$$

$$Penalty=\frac{\sum_{i=1}^{I}\sum_{j=i+1}^{I} Penalty_C(C_i,C_j)}{I(I-1)/2} \qquad (۵)$$

۳-۴-۳- **قسمت نهایی تابع ارزیابی**:در این مرحله مقدار تابع ارزیابی خارجی برای هر کروموزم را از مقدار تابع ارزیابی داخلی کم می کنیم.در حقیقت با استفاده از این تابع ارزیابی ۳ مرحله ای کروموزم هایی در سیکل ژنتیک برنده خواهند شد که علاوه بر همبسته بودن اعضای هر خوشه ، خوشه های مختلف نیز با هم نا همبسته باشند. رابطه (6) چگونگی محاسبه این مقدار را نشان می دهد.

$$Fitness = Intra - Penalty \qquad (۶)$$

### ۳-۵-عملگر انتخاب

در حین عمل انتخاب ، کروموزم هایی که از مقدار تابع ارزیابی بیشتری برخوردار باشند، شانس بیشتری برای حضور در نسل بعدی دارا می باشند. در حقیقت عملگر انتخاب تعدادی کروموزم را بصورت تصادفی انتخاب کرده و از میان آنها هر کدام که دارای مقدار تابع ارزیابی بهتری هستند را انتخاب می کند. عملگر انتخاب وظیفه همگرا کردن جمیعت بسوی جواب نهایی را بر عهده دارد.

در این مقاله ما روش انتخاب تورنمنت را با پارامتر دو در نظر گرفته ایم. در این روش دو کروموزم بصورت تصادفی انتخاب شده و کروموزمی که از مقدار ارزیابی بهتری برخوردار باشد به عنوان کروموزم نسل بعد انتخاب می شود.[9]

## ۴- ارزیابی الگوریتم ژنتیک ارایه شده

برای ارزیابی الگوریتم ژنتیک ارایه شده از چند پایگاه داده معتبر استفاده کرده ایم و نتایج آن را با روشهای مانند CSPA , HGPA, Average linkage, K-means, مقایسه کرده ایم [10,11]که تمامی این روشها بر اساس خوشه بندی ترکیبی کار می کنند و خوشه بندی نهایی را از روی ماتریس همبستگی محاسبه می کنند.

پایگاه داده های مورد استفاده شامل ۳ پایگاه داده از مجموعه UCI شامل Iris, Wine, Soybean و یک پایگاه داده واقعی به نام O8X می باشد. جدول(۱)خلاصه ای از خصوصیات این جداول را نشان می دهد.

**جدول (۱): خصوصیات پایگاه دادهای مورد آزمایش**

| پایگاه داده | تعداد نمونه ها | تعداد خصوصیات | تعداد خوشه ها | تعداد نمونه ها در هر خوشه |
|---|---|---|---|---|
| O8X | ۴۵ | ۸ | ۳ | ۱۵-۱۵-۱۵ |
| Iris | ۱۵۰ | ۴ | ۳ | ۵۰-۵۰-۵۰ |
| Wine | ۱۷۸ | ۱۳ | ۳ | ۴۸-۷۱-۵۹ |
| Soybean | ۴۷ | ۳۵ | ۴ | ۱۰-۱۰-۱۰-۱۷ |

جدولهای ۲ تا ۵ میزان خطای بدست آمده بر روی پایگاه داده های مورد آزمایش را نشان می دهد. هر کدام از اعداد ارایه شده در این جدول ها حاصل میانگین ۵۰ اجرا مختلف بوده است.در این جدولها GACE ،روش خوشه بندی ارایه شده بر اساس الگوریتهای ژنتیک است که در این مقاله ارایه شده است ، CAL روش خوشه بندی ترکیبی بر اساسAverage Linkage ، CK روش خوشه بندی ترکیبی بر اساس K-Means می باشد. HGPA, CSPA دو روش بر اساس ماتریس همبستگی و روشهای مبتنی بر گرافها می باشد.در این جدولها H تعداد تکرار الگوریتمهای اولیه برای بدست آوردن ماتریس همبستگی می باشد.

جدول (۲): نرخ خطای(٪) بدست آمده برای Iris

| H | GACE | CAL | CK | HGPA | CSPA |
|---|---|---|---|---|---|
| ۱۰ | 10.3 | 11.3 | 25.2 | 38.2 | 10.9 |
| ۲۰ | 10.1 | 10.3 | 22.2 | 39.1 | 10.9 |
| ۳۰ | 9.8 | 10.2 | 20.4 | 43.4 | 11.3 |
| ۴۰ | 9.9 | 10.1 | 20.9 | 41.9 | 10.7 |
| ۵۰ | 9.7 | 10.1 | 18.2 | 42.1 | 10.6 |

جدول (۳): نرخ خطای(٪) بدست آمده برای Wine

| H | GACE | CAL | CK | HGPA | CSPA |
|---|---|---|---|---|---|
| ۱۰ | ۳۰,۳ | ۳۲,۳ | ۳۵,۶ | ۴۲ | ۳۰,۷ |
| ۲۰ | ۲۹,۵ | ۳۲,۶ | ۳۵,۲ | ۳۸,۹ | ۲۹,۷ |
| ۳۰ | ۲۹,۲ | ۳۱,۱ | ۳۳,۷۴ | ۳۶,۲ | ۲۹,۷ |
| ۴۰ | ۲۹,۳ | ۳۰,۸ | ۳۴,۱۲ | ۳۵,۸ | ۲۹,۵ |
| ۵۰ | ۲۹,۲ | ۳۱,۱ | ۳۳,۵۱ | ۳۳,۷ | ۲۹,۳ |

جدول (۴): نرخ خطای(٪) بدست آمده برای Soybean

| H | GACE | CAL | CK | HGPA | CSPA |
|---|---|---|---|---|---|
| ۱۰ | ۲۴,۱ | ۲۵,۵ | ۴۱,۲ | ۲۹,۳ | ۲۹,۶ |
| ۲۰ | ۲۳,۸ | ۲۴,۱ | ۳۹,۳ | ۲۸,۳ | ۲۷,۳ |
| ۳۰ | ۲۳,۴ | ۲۳,۸ | ۳۸,۷۲ | ۲۷,۹ | ۲۷,۱ |
| ۴۰ | ۲۲,۶ | ۲۴,۳ | ۳۵,۴ | ۲۸,۲ | ۲۷,۵ |
| ۵۰ | ۲۲,۸ | ۲۴,۲ | ۳۱,۵ | ۲۷,۶ | ۲۷,۴ |

جدول (۵): نرخ خطای(٪) بدست آمده برای O8X

| H | GACE | CAL | CK | HGPA | CSPA |
|---|---|---|---|---|---|
| ۱۰ | ۲۰,۱ | ۳۸,۱ | ۲۲,۵ | ۱۵,۲ | ۱۵,۲ |
| ۲۰ | ۱۹,۸ | ۳۷ | ۲۱,۷ | ۱۴,۶ | ۱۴,۸ |
| ۳۰ | ۱۸,۱ | ۳۶,۴۴ | ۲۰,۷ | ۱۴ | ۱۴,۵ |
| ۴۰ | ۱۸,۵ | ۳۶,۳۲ | ۲۱,۷ | ۱۴,۵ | ۱۴,۲ |
| ۵۰ | ۱۸,۲ | ۳۵,۷ | ۱۹,۴ | ۱۴ | ۱۴,۵ |

هما نطوری که در جدولهای بالا دیده می شود GACE برروی تمامی پایگاه دادههای مورد آزمایش نتایج مناسبی تولید کرده است که این نشانگر مقاوم بودن الگوریتم ژنتیک ارایه شده برروی پایگاه دادههای مختلف است و این همان انتظاری است که می توان از یک الگوریتم

ژنتیک به عنوان یک سیستم توانای جستجوی فضای حالت ، انتظار داشت.

مقایسه GACE با CKو CAL نشان می دهد که GACE همواره از روشهایی که با کمک فاصله ،خوشه های نهایی را محاسبه می کنند ،بهتر عمل می کند. زیرا روشهای مبتنی بر فاصله بصورت حریصانه عمل کرده و یک جستجوی جامع برروی فضای حالت انجام نمی دهند بنا براین ممکن است در برخی از اجرا ها ،جوابهای بسیار دوری از جواب بهینه را پیدا کنند(این خصوصیت الگوریتهای حریصانه می باشد).

مقایسه GACE با HGPA و CSPA نشان می دهد که GACE در بیشتر موارد از روشهای مبتنی بر گرافها بهتر عمل می کند. در مواردی هم که روشهای مبتنی بر گراف بهتر عمل می کنند ، GACE نتایجی بسیار نزدیک به آنها تولید می کند. اما نکته مهم که در مقایسه GACE و روشهای مبتنی بر گرافها وجود دارد این است که، GACE بروی تمامی پایگاه داده ها خطای مناسب تولید می کند به عبارت دیگر نسبت به تغییر پایگاه داده حساس نیست ، اما روشهای مبتنی بر گراف این خصوصیت مثبت را ندارند.

## ۵- نتیجه گیری

در مقاله حاضر یک روش خوشه بندی ترکیبی بر اساس الگوریتمهای ژنتیک ارایه شد .برای افزایش کارایی الگوریتم ژنتیک ارایه شده و همچنین جلوگیری از بوجود آمدن بهینه های محلی ، عملگر جهش خاص منظوره طراحی و پیاده سازی شده است .مقایسه الگوریتم خوشه بندی ترکیبی ارایه شده با الگوریتهای متداول خوشه بندی ترکیبی، نمایانگر آن است که روش ما در ۸۰٪ موارد خطای کمتری در خوشه بندی نمونه های پایگاه داده های مختلف ایجاد می کند . خصوصیت مهم دیگر روش ارایه شده آن است که الگوریتم ما با بکار گیری توانایی الگوریتمهای ژنتیک در جستجوی فضای حالات ، نسبت به تغییر پایگاه داده وابسته نیست و همواره می تواند برروی پایگاه داده های گوناگون خطای مناسبی تولید کند ، بر خلاف روشهای خوشه بندی متداول که ممکن است برروی برخی از پایگاههای داده با خصوصیات خاص خوب عمل کنند و برروی پایگاه داده های دیگر با خصوصیات متفاوت نتایج مناسبی تولید نکنند.

## ۶- مراجع

[1] A. Strehl and J. Ghosh, "*Cluster ensembles - a knowledge reuse framework for combining multiple partitions*". Journal on Machine Learning Research, 3, pp. 583-617, 2002.
[2]A. Strehl & J. Ghosh, "Cluster ensembles—a knowledge reuse framework for combining partitionings", *National Conf. on Artificial Intelligence*, Alberta, Canada ,2002,pp. 93–98.
[3] L.J. Eshelman (Ed.), *Proceedings of the Sixth International Conference Genetic Algorithms*, Morgan Kaufmann, San Mateo, 1995.
[4]S. K. Pal and P. P.Wang, Eds., *Genetic Algorithms for Pattern Recognition*.Boca Raton, FL: CRC, 1996.
[5] Hancock P. J. B., 1992. *Genetic algorithms and permutation problems: a comparison of recombination operators for neural net structure specification. Proceedings of the Int.* Workshop on Combinations of Genetic Algorithms and Neural Networks (COGANN-92), pp. 108-122. IEEE Computer Society Press, Los Alamitos, CA.
[6] T.Yada, M.Ishikawa, H.Tanaka, and K.Asai (1994) *DNA Sequence Analysis using Hidden Markov Model and Genetic Algorithm*. In Genome Informatics Vol.5, pp.178-179.
[7] E.S. Gelsema (Ed.), *Special Issue on Genetic Algorithms, Pattern Recognition Letters*, vol. 16(8), Elsevier Sciences Inc., Amsterdam, 1995.
[8] L. Davis (Ed.), *Handbook of Genetic Algorithms*, Van Nostrand Reinhold, New York, 1991.
[9] D.E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, New York, 1989.
[10] B Minaei, A. Topchy and W. F. Punch, "*Ensembles of Partitions via Data Resampling* ", in Proc. Intl. Conf. on Information Technology, ITCC 04, Las Vegas, 2004
[11] A. Topchy, B. Minaei-Bidgoli, A.K. Jain, W. Punch ." *Adaptive Clustering ensembles*" In Proc. Intl. Conf on Pattern Recognition, ICPR'04, Cambridge, pp. 272-275, UK, 2004