# تشخیص جوامع در شبکه های اجتماعی بدون جهت- بدون وزن با ترکیب الگوریتم ژنتیک و معیار شباهت رئوس

**فاطمه حمزه ئیان[۱]، ابراهیم صحافی زاده[۲]، طالب خفائی[*۳]**

۱- **دانشجوی کارشناسی ارشد مهندسی کامپیوتر ، نرم افزار ، دانشگاه آزاد اسلامی ، بوشهر، ایران**
**f.hamzeheian@gmail.com**

۲- **عضو هیات علمی دانشگاه پیام نور ، مرکز بوشهر، ایران**
**sahafizadeh@gmail.com**

۳- **عضو هیات علمی دانشگاه آزاد اسلامی ، مرکز بوشهر، ایران**
**Taleb.khafaie@srbiaub.ac.ir**

## چکیده

شبکه اجتماعی ساختاری متشکل از افراد، اشیا، سازمان ها و یا هر نوع موجودیتی است که به واسطه نوع خاصی از وابستگی به یکدیگر متصل شده اند و تشکیل اجتماعی را می دهند. تحلیل شبکه های اجتماعی، روابط اجتماعی را با بیان نظریه گرافها و اصطلاحات رأس و یال می نگرد. راس ها در این نوع گراف را افراد یا اشیا بازی می کنند و یال نمایانگر هر نوع ارتباطی میان راس ها می تواند باشد. ما برای تشخیص و استخراج جوامع ازالگوریتم ژنتیک بهره برده ایم.این الگوریتم از دسته الگوریتم های اکتشافی است که با داشتن مجموعه جوابهای اولیه سعی در بهینه سازی جواب ها و یافتن بهترین جواب می کند، که جهت پایین آوردن پیچیدگی زمانی،از معیار دیگری به نام شباهت رئوس نیز بهره برده ایم.الگوریتم پیشنهادی بر روی شبکه هایی بدون وزن و بدون جهت موردارزیابی قرار گرفته است. نتایج نشان می دهد الگوریتم پیشنهادی از لحاظ معیار پیچیدگی زمانی و زمان اجرا به صرفه تر از زمانی است که ازالگوریتم ژنتیک به صورت تک بعدی استفاده می شود.

**کلمات کلیدی: شبکه های اجتماعی ، تشخیص جوامع ، الگوریتم ژنتیک ، معیار شباهت رئوس**

## ۱- مقدمه

تحلیل شبکه های اجتماعی که گاهی به اختصار به آن SNA[1] هم گفته می‌شود به معنای فرایند بررسی و ارزیابی ساختارهای یک شبکه اجتماعی به عنوان یک گراف از ابزارها یا انسانهاست که با خطوط ارتباطی به یکدیگر متصل هستند. در شبکه های اجتماعی اشیا افراد هستنند و ارتباط بین آنها نمایانگر ارتباط اجتماعی بین آنهاست.شبکه ها در سیستم های کامپیوتری توسط گراف ها نمایش داده می شوند[۱]. قابلیت تشخیص پارتیشن بندی شبکه به خوشه ها می تواند اطلاعات مهم و مفیدی از قبیل نوع ارتباط کاربران ،نحوه انتقال اطلاعات را بدهد[۲].الگوریتم ژنتیک به طور گسترده توسط محققان برای مسایل تشخیص جوامع استفاده شده است[۳]. عملگرهای متنوع این الگوریتم به کاهش فضای راه حلهای ممکن منجر می شود بنابراین همگرایی را بهبود می بخشد[۴]. در اکثر الگوریتم ها و روشهای بررسی شده جهت تشخیص جوامع، الگوریتم نیاز به اطلاعات و دانش از پیش دارد، مثلا تعداد جوامع، سرشاخه های اصلی جامعه، اما الگوریتم ژنتیک با داشتن تنها تعداد ارتباطات،سعی در رسیدن به بهترین جواب یا بهترین جامعه بندی شبکه می کند.این الگوریتم کار خود را با ایجاد

[1] Social Network Analysis

جمعیت اولیه ای از نودهای شـبکه شـروع می کند بنابراین مسلم است که هر چه تعداد اعضای شبکه بیشتر شود و ارتباطات بین آنها نیز زیاد باشـد تعداد جمعیت اولیه نیز بیشـتر می شـود و مدت زمانی که الگوریتم ژنتیک صرف یافتن بهترین جواب می کند بالا می رورد بصـورتی که جهت جامعه بندی شـبکه هایی با تعداد نودها و ارتباطات زیاد الگوریتم ژنتیک عملا نمی تواند بهینه باشـد. ما با ترکیب الگوریتم ژنتیک و معیار شـباهت رئوس و دو مرحله ای کردن الگوریتم یافتن جامعه ، سعی در بهبود پیچیدگی زمانی و زمان اجرای الگوریتم ژنتیک کرده ایم.

## ۲–شناسایی جوامع

جوامع بخشی از گراف هستند که با بقیه گراف ارتباط محدودی دارند. در شبکه ها جوامع گروهی از رأسها هستند که در شبکه دارای یالی ارتباطی بین آنها وجود دارد،بدین منظور جهت شناسایی و استخراج جوامع مبتنی بر ساختار شبکه، داشتن مجموعه داده ای [1]ازارتباط میان نودها(گره ها در شبکه) کافیست از جمله داشتن لیستی از همسایگان هر کدام از نودها ، درجه اتصال هر کدام از نودها[2] و موارد دیگری. فرض کنید گراف بدون وزن و بدون جهت G(V,E) داده شده است که V تعداد رئوس در گراف و E تعداد یالها است بنابراین رابطه(۱) برقراراست. اگر شبکه از N گره تشکیل شده باشد گراف بوسیله ماتریس مجاورت [3] N*N نشان داده می شود.

$$V = \{v_i \mid i = 1,2,\dots,n\},\quad E = \{e_i \mid i = 1,2,\dots,m\} \qquad (۱)$$

معیار ماژولاریتی Q برای گراف Gدررابطه(۲)بیان می شود.این تابع تناسب توسط نیومن[۵] در مقاله ای با نام یافتن و ارزیابی ساختار جوامع درشبکه ها معرفی شده است .

$$Q = \frac{1}{2*m}\sum i,j\ \left(Adj_{(i,j)} - \frac{k_i * k_j}{2*m}\right) * \partial(c_i, c_j) \qquad (۲)$$

**Adj(i,j)** نشان دهنده درایه متناظر در ماتریس مجاورت گراف G است. m نشان دهنده تعداد یالها(ارتباطات) و توسط رابطه(۳) محاسبه می شود. $k_i$ نشاندهنده درجه i امین گره و $k_j$ نشاندهنده درجه jامین گره است و به عنوان نمونه $k_i$ می تواند بوسیله رابطه (۴)حساب شود. دلتا یک تابع است که نشان می دهد دو گره در اجتماع یکسانی (مشابهی) وجود دارند یا خیر و توسط رابطه (۵) حساب می شود.

$$m = \frac{1}{2}\sum_{ij} Adj(i,j) \qquad (۳)$$

$$k_i = \sum_j Adj(i,j) \qquad (۴)$$

$$\partial = \begin{cases} 1 & if \quad c_i = c_j \\ 0 & if \quad c_{i\neq} \quad c_j \end{cases} \qquad (۵)$$

تا کنون تعریف دقیقی از جامعه در دست نیست اما بصورت کلی میتوان گفت جوامع بایستی شرایط رابطه(۶)راارضا کنند [۶]:

$$(۶)\ d_{i(m)}^{in} \geq d_{i(m_k)}^{out}\ \&\ d_{m_1}^{Total\,(d_{in})} \geq d_{m_k}^{Total(d_{out})}$$

اگر $d_i^{in}$ تعداد یالهای درون ماژولی مربوط به گره i در ماژول $m_1$ باشد.$d_i^{out}$ تعداد یالهای بیرون ماژولی همان گره درماژول $m_1$ باشد. تعداد $k$ جامعه داشته باشیم.

[1] Dataset
[2] Connectivity Degree
[3] Adjacency Matrix

## ۳- روش پیشنهادی

در این مقاله ما برای حل مساله تشخیص جامعه الگوریتم پیشنهادی CDSGA[1] که بر پایه الگوریتم ژنتیک و معیار شباهت است، ارائه داده ایم. در ادامه پس از معرفی الگوریتم ژنتیک در تشخیص جوامع به معرفی معیار شباهت رئوس پرداخته و در بخش نتایج نشان می دهیم الگوریتم پیشنهادی از لحاظ پیچیدگی زمانی و زمان اجرا نسبت به الگوریتم ژنتیک تطبیقی [۱] به صرفه تر است .

### ۳-۱ – شباهت

معیارهای مختلفی برای بدست آوردن شباهت میان موجودیتها هست ، دراین مقاله ماااز شباهت همینگ [2]بهره برده ایم[3]. برای بدست آوردن این شباهت میان دو موجودیت، ابتدا بایستی فاصله میان آن دو را محاسبه کرد.

#### ۳-۱-۱- فاصله [4]

برای بدست آوردن فاصله میان ۲ موجودیت [5]، بردار ویژگی آنها مورد نیاز است، این بردار به صورت باینری می باشد و طبق رابطه(۷) قابل محاسبه است .[6]

$$d_{ij} = \frac{q+r}{t} \quad (۷)$$

$$t=p+q+r+s \quad (۸)$$

### ۳-۲ – معیار شباهت بر مبنای محاسبه فاصله(شباهت همینگ) :

شباهت همینگ همواره مقداری در بازه ۰ تا۱ داردو از طریق رابطه (۹) بدست می آید .

$$S_{ij} = 1 - \delta_{ij} \quad (۹)$$

رابطه(۹) بدین معنی است برای بدست آوردن شباهت میان دو شی،فاصله نرمالیزه شده آن را که از رابطه (۷) بدست می آید از عدد یک کم می کنیم.

## ۴- نمایش ژنتیک

در این بخش ژنتیک تطبیقی [7]را که می خواهیم با روش پیشنهادی خودمان مورد مقایسه قرار دهیم شرح می دهیم. فرض کنید شبکه ای شامل ۸ گره را داریم و نمایش گرافی آن را در شکل(۱) ملاحظه می کنید. هر کدام از ژنها ۲ نمونه از اطلاعات را نگهداری می کنند، شناسه جمعیت و شناسه اجتماع (شکل ۱- ب).

---

[1] Community Detection based on Node's Similarity using the Genetic Algorithm

[2] Hamming Similarity

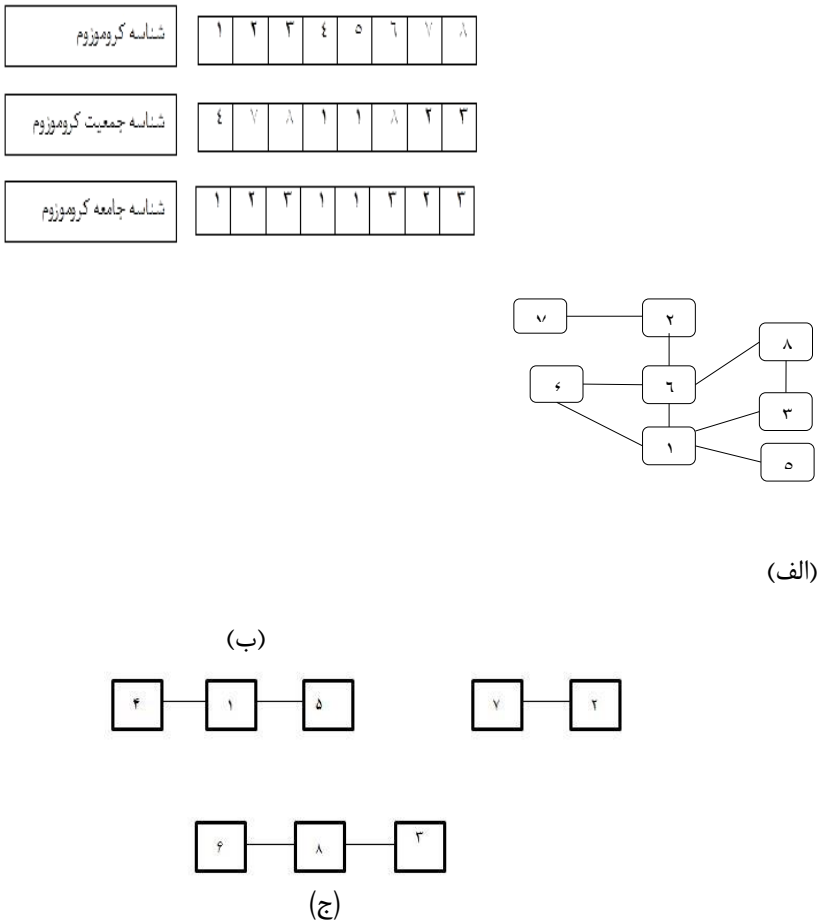
[3] شباهت همینگ از بردارهای فاصله که به طول یکسان هستند جهت بدست آوردن میزان شباهت ۲ موجودیت استفاده می کند .

[4] Distance

[5] فرض می کنیم دو موجودیتم به نامهای A,B داشته باشیم ، بنابراین برداروِیژگی به همین نام برای آن دو در نظر میگیریم .

[6] p: شامل عددی است که نشاندهنده این است مواردی که در هر دو بردار یک است.q : شامل عددی است که نشاندهنده مواردی که در بردار A ، یک است اما در بردار B صفر است، r : شامل عددی است که نشاندهنده این است مواردی که در بردار B ، یک است اما در بردار A ، صفر است ، s : شامل عددی است که نشاندهنده این است مواردی که در هر دو بردار صفر است . بنابراین t برابر است با طول بردار ویژگی ها .

[7] A New Adaptive Genetic Algorithm for Community Structure Detection [1]




**شکل ۱ . برگرفته از پژوهش اتای و کوداز [۲] . الف : مثالی از شبکه ای با ۸ نود ، ب : نمونه ای از کروموزوم بدست آمده ، ج : جوامع بدست آمده از کروموزوم (ب) ، از کروموزوم موجود ۳ جامعه بدست آمده است که این تعداد جوامع برای کروموزوم های متفاوت متغیر است .**

درالگوریتم ژنتیک تابع برازندگی مشخص کننده برازنده بودن جوابهای بدست آمده است. در الگوریتم ژنتیک تطبیقی تابع برازندگی را معیار ماژولاریتی که دررابطه(۲) در بخش(۲) قابل مشاهده است در نظر گرفته است. تابع برازندگی در بهترین خوشه، ماکزیمم مقدار Q را بدست می آورد. مقدار Q همواره در بازه ۱- الی ۱ است.این الگوریتم دارای عملگرهایی نظیر ادغام، جهش و نخبه گرایی می باشد که نرخ های هر کدام در بخش نتایج قابل مشاهده است .

## ۵ – نتایج آزمایشات

این مقاله بر روی ۳ شبکه واقعی آزمایش شده است، شبکه کلاب کاراته زاکاری، شبکه اجتماعی دلفین، دانشگاه فوتبال آمریکا[۷].اطلاعات مورد نیاز از این مجموعه های داده در جدول (۱) آمده است.

**جدول ۱ . مجموعه داده های مورد استفاده در پژوهش**

| نام شبکه | تعداد نودها | تعداد یالها | اندازه جمعیت اولیه | نام مورد استفاده در این پژوهش |
|---|---|---|---|---|
| Zachary's karate club | ۳٤ | ۷۸ | ۲۰ | Z |
| Dolphin Social Network | ٦۲ | ۱٥۹ | ۳۰ | D |
| American College Football | ۱۱٥ | ٦۱۳ | ۱۰۰ | A |

مراحل کار روش پیشنهادی در الگوریتم های ۱ و۲ آورده ایم .

**Algorithm 1:** Similarity (pre Modulate)

**Data** : A Dataset of Nodes with their Relation ship
**Result** : Communities of Similar Nodes

N← Total Number of Nodes ;
Adj← Make N * N Adjacency Matrix ;
Neighbors list ←-Search Adj for find Neighbor of every Node ;
Similarity Matrix ←$\forall$ i,j $\in$ N , Compute Hamming Similarity ;
Pre Modulate← Module each Node with Nodes with the highest Similarity in the Similarity Matrix .

**الگوریتم ۱ .ایجاد ماژولهایی از نودهای مشابه ، داده های ورودی مجموعه های داده با ارتباطات آنها و خروجی شامل اجتماعاتی از نودهای مشابه است .**

**Algorithm 2:** Genetic Algorithm ( Modulate )

**Data : Result of Algorithm 1**
**Result :** Communities of Network

Make Initial Population ( Communities of Similar Nodes are Genomes) ;
**Repeat** X Times **{**
Make Communities of Chromosomes ;
Send Elitism Chromosome to Next Population ; // based on Elitism Rate
Crossover with Crossover Rate; // based on Crossover Rate
Mutation with Mutation Rate; // based on Mutation Rate
Computing Modularity Value for every Chromosome;
Make New Population with Chromosomes which Have High Modularity Value ;
}
**Return** : Communities with High Modularity Value ;

**الگوریتم ۲ . ایجاد ماژولهای نهایی ، ورودی برابر است با خروجی الگوریتم ۱ ، خروجی ماژولهایی که دارای بیشترین مقدار Q هستند .**

نرخ آمیزش برابر با مقدار ۰٫۸ است و نرخ انتخاب آمیزش[1] ۰٫۵ در نظرگرفته شده است ، همچنین نرخ جهش برابر با مقدار ۰٫۲ و نرخ نخبه برابر ۰٫۰۵ در نظر گرفته شده است، قابل ذکر است این مقادیر با آزمون و خطا توسط [۱] بدست آمده است .تمام آزمایشات بر روی سیستم کامپیوتری با مشخصات زیر انجام شده است:

مشخصات پردازنده :Intel (R) Core(TM) Duo CPU ، ویندوز هفت، محیط سیستم عامل ۶۴ بیت،حافظه ۴ گیگا بایت.

## ۵–۱ – مقایسه تعداد ژنوم ها و جمعیت در الگوریتم ژنتیک تطبیقی و روش پیشنهادی

در الگوریتم ژنتیک تطبیق جهت یافتن جوامع ، همواره تعداد ژنومها با تعداد موجودیت های شبکه برابر است،اما در روش پیشنهادی تعداد ژنوم ها مشخص نیست بنابراین طبق رابطه (۱۰) که از سعی و خطا بدست آورده ایم میزان تعداد ژنوم ها به صورت تقریبی بدست می آوریم.

$$\boldsymbol{Round}\left(\frac{\boldsymbol{Number\ of\ Nodes}}{\mathbf{4}}\right) \qquad (۱۰)$$

[1] Crossover Choice

با مطالعه ای که بر پژوهش سایر پژوهشگران انجام شد،جهت بدست آوردن تعداد جمعیت اولیه راه حلی ذکر نشده و بهترین تعداد صرفا از طریق آزمایشات بدست آمده، لذا ما با بررسی های انجام شده به رابطه(۱۱)جهت تخمین هرچه بهتر تعداد جمعیت اولیه رسیده ایم.

$$Number\ of\ Population\ = \frac{Number\ of\ Edges}{Round(\ ln\ (\ Number\ of\ Edges\ ))} \quad (۱۱)$$

از آنجاییکه در روش پیشنهادی یافتن تعداد یالها برابر است با بررسی ارتباطات تمام نودهای شبکه ( نودهایی که در روش پیشنهادی بدست می آیند برابر با ماژولهای اولیه هستند )است لذا جهت یافتن تعداد یالها از رابطه(۱۲)استفاده می کنیم. n ازرابطه(۱۰) بدست آمده است.همچنین می توانیم با استفاده از رابطه (۱۳) نیز تعداد یالها را بدست اوریم.

$$E = \sum_{i=1}^{n}(n-i) \quad (۱۲)$$

$$E = \frac{n^2-n}{2} \quad (۱۳)$$

بنابراین جدول(۲) را داریم.

**جدول (۲) ، مقایسه تعداد نودها و جمعیت اولیه در الگوریتم ژنتیک تطبیقی و روش پیشنهادی**

| شبکه / الگوریتم | Z | | D | | A | |
|---|---|---|---|---|---|---|
| | تعداد نودها | جمعیت اولیه | تعداد نودها | جمعیت اولیه | تعداد نودها | جمعیت اولیه |
| AGA | ۳۴ | ۲۰ | ۶۲ | ۳۰ | ۱۱۵ | ۱۰۰ |
| CDSGA | ۸ | ۱۰ | ۱۶ | 20 | ۲۸ | ۶۴ |

## ۵– ۲– ارزیابی روش پیشنهادی از لحاظ پیچیدگی زمانی

پیچیدگی زمانی الگوریتم ژنتیک تطبیقی توسط رابطه(۱۴) محاسبه می شود و این معیار برای الگوریتم پیشنهادی توسط رابطه(۱۵)، محاسبه می شود. با توجه به اینکه اندازه جمعیت نسبت به حالت اولیه برای هر کدام از مجموعه های داده حداقل ۱۰ تا کاهش پیدا می کند بنابراین این مقدار رااز متغیری کم کرده ایم.

$$O(N*P) = \left(N * \frac{E}{\ln(E)}\right) \quad (۱۴)$$

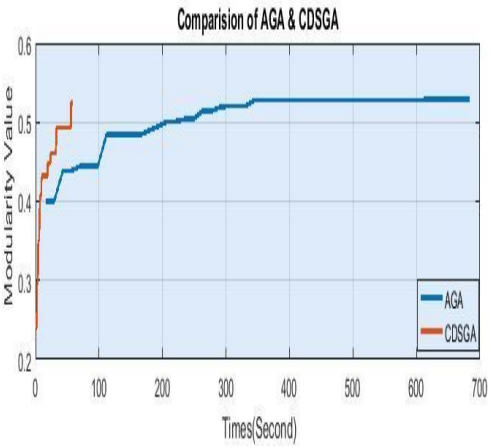
$$O(N*P) = \left(\frac{N}{4} * \left(\frac{E}{\ln(E)} - \partial\right)\right) \qquad \partial \geq 1 \quad (۱۵)$$

همانگونه که از روابط بالا مشاهده می کنیم، پیچیدگی زمانی روش پیشنهادی به نسبت ژنتیک تطبیقی کاهش داشته و با بالارفتن تعداد ارتباطات و اندازه جمعیت، روش پیشنهادی در مدت زمان قابل قبول تری می تواند به جوابی نزدیک بهینه دست یابد.

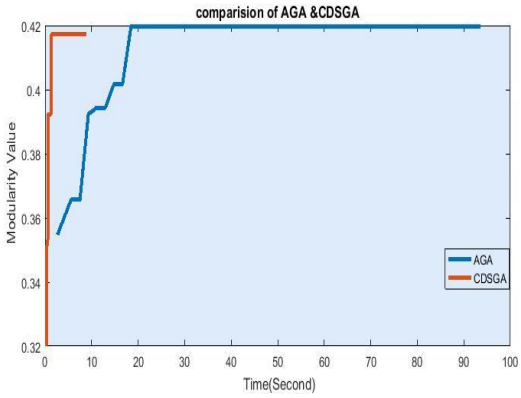
## ۵-۳- نمودار پیشرفت ماژولاریتی در واحد زمان

در شکل۲، نمودار پیشرفت ماژولاریتی در واحد زمانی را برای ۳ مجموعه داده ملاحظه می کنید. همانگونه که مشاهده می شود با اختلاف بسیار ناچیزی اما در واحد زمان قابل قبول تری می تواند مقدار بهترین مقدار ماژولاریتی را یافته و به بهینگی همگرا شود.

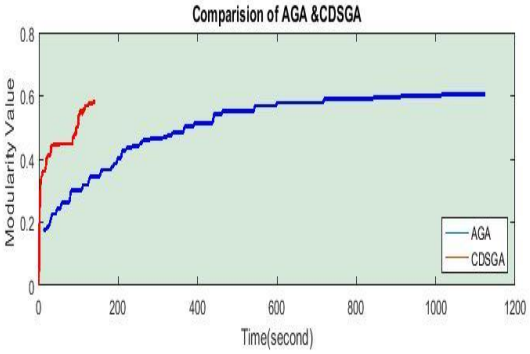
D




Z




A



**شکل ۲ . نمودار پیشرفت ماژولاریتی در واحد زمان ، الگوریتم ژنتیک تطبیقی مدت زمان زیادی را صرف می کند تا به جوابی نزدیک به بهینه برسد در حالیکه روش پیشنهادی در مدت زمانی بسیار کمتر از آن به جوابی نزدیک به بهینه می رسد . در بالای هر کدام از نمودارها نام مجموعه داده قرار گرفته شده است .**

## ۶– نتیجه گیری و پیشنهاد

با گسترش روز افزون شبکه های اجتماعی و حجم بالایی از اطلاعات، نیاز به داشتن ساختارهای معنی داری از شبکه ها بیش از پیش احساس می شود.از آنجاییکه الگوریتم ژنتیک جز دسته الگوریتم های اکتشافی می باشد و برای این دسته از مسایل بهترین پاسخ را ارائه می دهد اما دارای مشکلی از لحاظ زمان اجرا هستند که با افزایش تعداد ژنوم ها همواره جمعیت افزایش پیدا کرده و زمان اجرا نیز به صورت پلکانی افزایش می یابد، بنابراین می توان با تغییراتی در انتخاب جمعیت اولیه و یا هر کدام از عملگرهای ژنتیک،از زمان اجرا کاسته و پیچیدگی زمانی را کاهش دهیم. این روش دارای کاستی هایی نیز می باشد که می توان با ترکیب با سایر روشها و معیارهای شباهت کاستی ها را برطرف کرد.

## ۷– منابع و مراجع


۲ .ترشیزی نژاد, فاطمه؛ مهرداد جلالی و داود بهره پور، ۱۳۹۴، روشی جهت تشخیص جوامع در شبکه های اجتماعی مبتنی بر الگوریتم رقابت استعماری، دومین کنگره بین المللی فن آوری، ارتباطات و دانشICTCK2015، مشهد، دانشگاه آزاد اسلامی واحد مشهد،https://www.civilica.com/Paper-ICTCK02-ICTCK02_077.html

[1].Atay, Y., & Kodaz, H. (2016). A new adaptive genetic algorithm for community structure detection. In Intelligent and Evolutionary Systems (pp. 43-55). Springer, Cham .

[3].Meng, X., Dong, L., Li, Y., & Guo, W. W. (2017). A genetic algorithm using K-path initialization for community detection in complex networks. Cluster Computing, 20(1), 311-320.

[4].Li, K., & Xiong, L. (2015, November). Community detection based on an improved genetic algorithm. In International Symposium on Intelligence Computation and Applications (pp. 32-39). Springer, Singapore

[5].Girvan, M., Newman, M.E.J.: Community Structure in Social and Biological Networks. Proceedings of the National Academy of Sciences 99(12), 7821–7826 (2002) .

[6]. Raghavan, Usha Nandini, Réka Albert, and Soundar Kumara. "Near linear time algorithm to detect community structures in large-scale networks." Physical review E 76.3 (2007): 036106.

[7]. http://konect.uni-koblenz.de/

[8]. Li, W., Huang, C., Wang, M., & Chen, X. (2017). Stepping community detection algorithm based on label propagation and similarity. Physica A: Statistical Mechanics and its Applications, 472, 145-155.
[9]. Azaouzi, M., Rhouma, D., & Romdhane, L. B. (2019). Community detection in large-scale social networks: state-of-the-art and future directions. Social Network Analysis and Mining, 9(1), 23.
[10]. Chouchani, N., & Abed, M. (2018). Online social network analysis: detection of communities of interest. Journal of Intelligent Information Systems, 1-17.