

تشخیص اجتماعات ترکیبی در شبکه‌های اجتماعی

حسین علیزاده^۱

رسول حسین‌زاده^۲

اسلام ناظمی^۳

۱- دانش‌آموخته دکتری- دانشکده مهندسی کامپیوتر - دانشگاه علم و صنعت ایران - تهران - ایران

halizadeh@iust.ac.ir

۲- دانش‌آموخته کارشناسی ارشد- دانشکده مهندسی برق و کامپیوتر - دانشگاه شهید بهشتی - تهران - ایران

rasoulhossienzadeh@gmail.com

۳- استادیار- دانشکده مهندسی برق و کامپیوتر - دانشگاه شهید بهشتی - تهران - ایران

nazemi@sbu.ac.ir

چکیده: یکی از چالش‌های مهم در تحلیل شبکه‌های اجتماعی، تشخیص اجتماعات است. اجتماع مجموعه افراد یا سازمان‌هایی هستند که چگالی ارتباط آن‌ها با هم بیشتر از سایر موجودیت‌های شبکه است. خوشه‌بندی یا تشخیص اجتماعات، ساختار گروه‌ها در شبکه‌های اجتماعی و ارتباطات پنهان بین مؤلفه‌های آنها را آشکار خواهد نمود. اکثر روش‌های رایج تشخیص اجتماعات موجود قطعی نیستند و نتایج آن‌ها به مقادیر اولیه‌ای که در اکثر مواقع به صورت تصادفی انتخاب می‌شود بستگی دارد. اما خوشه‌بندی ترکیبی، بدون توجه به مقادیر اولیه تصادفی هر کدام از الگوریتم‌های پایه‌اش، با ترکیب آنها، نتایج مستحکم و پایداری تولید می‌کند. در این مقاله یک روش ترکیبی تشخیص اجتماعات با الهام از خوشه‌بندی ترکیبی پیشنهاد شده است. از مشخصه‌های روش پیشنهادی تشخیص اجتماعات ترکیبی، توانایی ترکیب با روش‌های دیگر است به گونه‌ای که می‌توان از الگوریتم‌های دقیق‌تر نیز در چهارچوب پیشنهادی استفاده کرد. نتایج تجربی در این مقاله نشان می‌دهند که روش ترکیبی پیشنهادی نسبت به متوسط روش‌های تشخیص اجتماعات پایه‌ای مورد استفاده در آن از کارایی بسیار بالاتری برخوردار بوده و در اکثر موارد حتی از بهترین الگوریتم پایه‌ای خود نیز بهتر عمل کرده است. نتایج این مقاله می‌تواند در بسیاری از مسائل از جمله تشخیص دقیق‌تر اجتماعات، بازاریابی، تبلیغات، درک شبکه و بهبود موثرهای جستجو مورد استفاده قرار گیرد.

کلمات کلیدی: خوشه‌بندی، خوشه‌بندی ترکیبی، تشخیص اجتماعات، تشخیص اجتماعات ترکیبی، تحلیل شبکه‌های اجتماعی.

تاریخ ارسال مقاله: ۱۳۹۱/۱۱/۱۷

تاریخ پذیرش مشروط مقاله: ۱۳۹۲/۰۸/۲۷

تاریخ پذیرش مقاله: ۱۳۹۲/۱۰/۱۶

نام نویسنده‌ی مسئول: حسین علیزاده

نشانی نویسنده‌ی مسئول: ایران - تهران - نارمک - خ هنگام - دانشگاه علم و صنعت ایران - دانشکده مهندسی کامپیوتر - آزمایشگاه
داده‌کاوی



۱- مقدمه

۲- تعاریف و مفاهیم

در این قسمت به بررسی تعاریف و مفاهیمی که در این مقاله مورد استفاده قرار گرفته است می‌پردازیم.

۳- شبکه‌های اجتماعی

شبکه‌های اجتماعی ساختاری اجتماعی است و از افراد با سازمان-های تشکیل شده است که گره‌های شبکه را تشکیل می‌دهند. گره‌ها توسط یک یا چند نوع خاص از وابستگی به هم متصل هستند، برای مثال تبادلات مالی، دوستی‌ها، خویشاوندی، تجارت، لینک‌های وب، سرویس بیماری‌ها(ایپیدمولوژی) یا مسیرهای هوایی نمونه‌هایی از ارتباط هستند. اما ساختارهای حاصل از این شبکه‌ها اغلب بسیار گسترده و پیچیده هستند. تحلیل شبکه اجتماعی^۴ عبارت است از نگاشت و اندازه‌گیری روابط و همکاری‌ها در بین افراد، گروه‌ها، سازمان‌ها و هر موجودیتی که قابلیت پردازش اطلاعات و دانش را داشته باشد. برای نمایش و تحلیل شبکه‌های اجتماعی معمولاً از تئوری گراف^۵ استفاده می‌شود. مولفه‌های موجود در تئوری گراف گره^۶ و لبه^۷ است[۱].

یک شبکه N را می‌توان به صورت یک گراف به شکل $G = \langle V, E \rangle$ تعریف نمود، که در آن گراف G شامل V مجموعه گره‌های گراف و E مجموعه لبه‌ها است. بعد از بدست آوردن گراف فوق می‌توان از روی گراف، ماتریس مجاورتی متناظر با آن را بدست آورد. ماتریس A را ماتریس مجاورتی گراف G گویند، این ماتریس یک ماتریس مربعی به ابعاد تعداد گره‌های گراف متناظر با آن بوده و در صورت وجود ارتباط بین دو گره به آن ماتریس عدد یک و در غیر این صورت عدد صفر منظور می‌گردد.

۴- تشخیص اجتماعات

در شبکه‌های اجتماعی برخی گره‌ها در مقایسه با کل گره‌های شبکه، ارتباط بیشتری با هم دارند که به آن‌ها اجتماع گفته می‌شود[۳]. هدف از تشخیص اجتماعات، جدا کردن گروه‌ها یا اجتماعاتی است که ارتباط بیشتری با هم دارند[۲]. در واقع تشخیص اجتماعات، تقسیم‌بندی‌های موجود در شبکه را نشان می‌دهد و اجتماعات یک گراف را از هم مجزا می‌کند. تشخیص اجتماعات به ما کمک می‌کند تا دید بهتری نسبت به ساختار شبکه پیدا کنیم[۴].

در [۳] تشخیص اجتماعات را بدین صورت زیر تعریف می‌نماید: اگر A را ماتریس مجاورتی یک گراف درنظر بگیریم، درجه یک k_i در گراف را با k_i نشان می‌دهند که برابر است با: $\sum_{j} A_{ij}$ [۴].

امروزه اینترنت و خدمات وب به سرعت در حال گسترش هستند و هم‌زمان شبکه‌های اجتماعی مجازی نقش مهمی در زندگی واقعی افراد دارد. در واقع شبکه‌های اجتماعی، شبکه‌های تعاملی هستند که از اینترنت به عنوان رسانه‌ای برای ایجاد ارتباط بین افراد استفاده می‌کنند. با افزایش سریع کاربران شبکه‌های اجتماعی، کاوش در مقیاس بالای داده‌ها می‌تواند کارایی بهتر و مؤثرتری از پتانسیل پنهان این شبکه‌ها فراهم کند[۱].

تشخیص اجتماعات^۱ یا خوشبندی^۲ یکی از مراحل اصلی در داده کاوی است که وظیفه کاوش الگوهای پنهان در داده‌های بدون برچسب را بر عهده دارد. با توجه به پیچیدگی مسئله خوشبندی و ضعف روش‌های پایه‌ای آن، امروزه برخی مطالعات در خصوص خوشبندی بندی به سمت روش‌های خوشبندی ترکیبی^۳ هدایت شده است. همان‌گونه که در بالا اشاره شد، تشخیص اجتماعات و خوشبندی شباهت‌های بسیار زیادی با هم دارند و منابع زیادی این دو را یکسان می‌دانند[۱][۲][۳][۴].

رویکرد خوشبندی ترکیبی بدین صورت می‌باشد که با استفاده از روش‌های پایه‌ای و یا هر روش ممکن موجود، با ترکیب تک‌تک روش‌ها با یکدیگر در صدد پیدا نمودن یک جواب پایدار از نتایج متفاوت می‌باشد. به طوری که جواب نهایی دارای دقت، پایداری، صحت و از اجماع نتایج روش‌های مختلف بدست می‌آید، که دارای اطمینان بالاتری می‌باشد. با توجه به قابلیت ادغام خوشبندی ترکیبی و تشخیص اجتماعات در این مقاله برای بهبود نتایج نهایی در تشخیص اجتماعات از «تشخیص اجتماعات ترکیبی» استفاده شده است. دقت و صحت اجتماعات نهایی در تشخیص اجتماعات مسئله مهمی می‌باشد، چرا که الگوریتمی در تشخیص اجتماعات وجود ندارد که در تمام حالات و داده‌های گوناگون جواب کاملاً دقیق را داشته باشد.

در این مقاله سعی داریم که با استفاده از تشخیص اجتماعات ترکیبی در شبکه‌های اجتماعی به بررسی اجتماعات در این شبکه‌ها بپردازیم. بدین صورت که با اعمال روش‌های مختلف تشخیص اجتماعات و بدست آوردن نتایج متفاوت و در نهایت ترکیب این روش‌ها تعدادی اجتماع که دارای دقت بالاتر، نتایج مطمئن‌تر و پایداری بیشتری هستند را پیدا نماییم. نتایج این بررسی‌ها می‌تواند در مسائل بسیاری از جمله تشخیص دقیق‌تر اجتماعات، بازاریابی، تبلیغات، درک شبکه و بهبود موتورهای جستجو مورد استفاده قرار گیرد.

قسمت‌های موجود در این مقاله بدین شرح می‌باشد که بعد از شرح تعاریف و مفاهیم و نیز بررسی کارهای انجام شده موجود، یک روش پیشنهادی در خصوص پیدا نمودن اجتماعات مطرح و در نهایت این روش را مورد ارزیابی قرار می‌دهیم. استفاده از این روش دارای مزایای و معایبی نیز می‌باشد که در قسمت نتیجه‌گیری به آن اشاره شده است.

می باشد. با در نظر گرفتن یک زیر گراف $S \subseteq G$ به طوری که

(1)

$$S \subseteq G \Rightarrow k_i(S) = k_i^{in}(S) + k_i^{out}(S)$$

در رابطه (1) نیز درجه هر گره را برای اجتماعات داخلی و

$$k_i^{out}(S) = \sum_{j \notin S} A_{ij} \text{ و } k_i^{in}(S) = \sum_{j \in S} A_{ij}$$

تعریف می کنیم. بدین معنی که درجه یک گره به نام A که عضو اجتماع Z می باشد را به عنوان مجموع اجتماعات داخلی بدین صورت

$$k_i^{in}(S) = \sum_{j \in S} A_{ij} \text{ و درجه یک گره به نام } A \text{ که}$$

عضو اجتماع Z نمی باشد را به عنوان مجموع اجتماعات خارجی

$$k_i^{out}(S) = \sum_{j \notin S} A_{ij} \text{ تعریف می کنیم.}$$

اجتماعاتی که مجموع گره های داخلی آن خیلی بیشتر از مجموع گره های خارجی آن باشد را به عنوان یک زیر گراف از کل گراف ورودی در نظر می گیریم و به انجام این کار تشخیص اجتماعات گویند و در صدد پیدا نمودن بهینه ترین و دقیق ترین اجتماعات هستیم که به صورت رابطه (2) تعریف می شود.

حال به تعریف مسئله اصلی که همان تشخیص اجتماعات می باشد می پردازیم.

$$\sum_{i \in S} k_i^{in}(S) >> \sum_{i \notin S} k_i^{out}(S) \quad (2)$$

با توجه به رابطه (2)، $\sum_{i \in S} k_i^{in}(S)$ مجموع درجه های تمامی

گره هایی که در اجتماع یا زیر گراف S وجود دارد می باشد و $\sum_{i \notin S} k_i^{out}(S)$ مجموع درجه های تمامی گره هایی که در اجتماع یا

زیر گراف S وجود ندارد می باشد. بنابراین مسئله اصلی در زمینه تشخیص اجتماعات این است که بدانیم چگونه به بهترین حالت شبکه را به گروه های اصلی آن تقسیم کنیم. در شبکه ها واقعی هیچ اطلاعاتی درباره تعداد اجتماعات وجود ندارد. در بعضی روش های تشخیص اجتماعات فرض بر آن است که تعداد اجتماعات شبکه را از قبل می دانیم. در حالی که در بسیاری از شبکه ها، هیچ دانش اولیه ای در مورد اجتماعات شبکه وجود ندارد و روش های جدید بدنیال بر طرف کردن این نقیصه هستند.

روش های تشخیص اجتماعات معمولاً با استفاده از گراف ایجاد شده از ارتباطات بین افراد در شبکه های اجتماعی ماتریس همسایگی⁸ متناظر با آن را محاسبه می کنند و بعد از آن با اعمال الگوریتم های تشخیص اجتماعات بر روی این ماتریس می توان نتایج گوناگون بر روی داده های متفاوت را بدست آورد. هر یک از روش ها با توجه به روش انتخابی و نیز نوع و حجم داده ورودی، اجتماعات متفاوتی را پیدا می

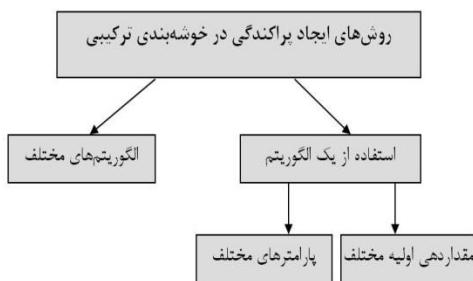
۳-۲- خوشبندی ترکیبی

خوشبندی یکی از مراحل اصلی در داده کاوی است که در جهت کاوش الگوهای پنهان در داده های بدون برچسب مورد استفاده قرار می گیرد. روش های خوشبندی عادی بدین صورت عمل می نمایند که با تعریف یک الگوریتم سعی در برطرف نمودن مشکل خوشبندی دارند. درصورتی که اعمال هر یک از این الگوریتم برروی داده های یکسان دارای نتایج گوناگونی خواهد.

در واقع هدف اصلی خوشبندی ترکیبی جستجوی نتایج بهتر و مستحکم تر، با استفاده از ترکیب اطلاعات و نتایج حاصل از چندین خوشبندی اولیه است. پراکندگی در نتایج اولیه یکی از مهم ترین عواملی است که می تواند در کیفیت نتایج نهایی خوشبندی ترکیبی اثر گذار باشد. همچنین، کیفیت نتایج اولیه نیز عامل دیگری است که در کیفیت نتایج حاصل از ترکیب موثر است. به طور خلاصه خوشبندی ترکیبی شامل دو مرحله اصلی زیر می باشد [5]:

- تولید نتایج متفاوت از خوشبندی ها، به عنوان نتایج خوشبندی اولیه بر اساس اعمال روش های مختلف که این مرحله را، مرحله ایجاد تنوع یا پراکندگی^{*} می نامند.
- ترکیب نتایج به دست آمده از خوشبندی های متفاوت اولیه برای تولید خوش نهایی؛ که این کار توسط تابع توافقی[†] (الگوریتم ترکیب کننده) انجام می شود.

برای ایجاد نتایج مختلف در این مقاله از دو روش که در شکل (1) نشان داده شده است استفاده کردہ ایم.



شکل (1): ایجاد نتایج اولیه مختلف

همان طوری که در شکل (1) مشاهده می شود، برای ایجاد نتایج مختلف می توان از یک الگوریتم یا از الگوریتم های متفاوت استفاده نمود. بدین صورت که با مقداردهی اولیه متفاوت و تغییر در پارامترهای آن می توان نتایج مختلفی را ایجاد نمود.

پس از اینکه نتایج اولیه مختلف ایجاد شد، معمولاً با استفاده از یک تابع ترکیب کننده این نتایج ترکیب می شوند. یکی از متدائل ترین

کل لبه‌های گراف است. به عبارت دیگر کسری از تعداد لبه‌ها که دو سر آن در اجتماع i قرار دارد به کل لبه‌های گراف و a_i نسبت تعداد لبه‌ایی که حداقل یک گره آن در اجتماع i باشد به کل لبه‌های گراف است. در واقع رابطه فوق سهم هر اجتماع i را در کل شبکه بیان می‌کند. هرچه این عدد بزرگتر باشد، این اجتماع سهم بیشتری در شبکه دارد و در واقع اجتماع قوی‌تر است. بدین معنی که ارتباطات داخل اجتماع در آن بیشتر از ارتباط آن با سایر اجتماعات شبکه است. اگر کل گراف شامل تنها یک اجتماع باشد یا گره‌ها بصورت تصادفی در بین اجتماعات قرار گرفته باشند $Q=0$ خواهد بود. هرچه مقدار Q به یک نزدیک‌تر باشد اجتماعات بهتر جدا شده‌اند ولی هیچ‌گاه یک نمی‌شود. در عمل $Q > 0.3$ ساختار گروهی مناسبی را نشان می‌دهد. مازولاریتی می‌تواند مقادیر منفی را نیز پیذیرد [9]. مقدار مازولاریتی ممکن است برای یک شبکه در بهترین حالت زهندی از عدد خاص بیشتر نشود. این بدان معنا نیست تشخیص اجتماع به درستی انجام نشده است، بلکه ممکن است بیشترین مقداری باشد که مازولاریتی برای آن شبکه می‌تواند داشته باشد.

یکی از روش‌های تشخیص اجتماعات روش تقسیم‌کننده می‌باشد. در شبکه‌ها، هر چه تعداد مسیرهایی که از گره یا لبه خاصی عبور می‌کنند بیشتر باشد آن گره یا لبه مهم‌تر است. بنابراین با فرض اینکه کوتاه‌ترین مسیر بین دو گره محاسبه‌پذیر باشد، اهمیت یک گره یا لبه را می‌توان اندازه‌گیری نمود، که به آن مرکزیت مابینی گفته می‌شود. یکی از روش‌های بدست آوردن اجتماعات که [10] ارائه شده است مرکزیت مابینی می‌باشد. رابطه مرکزیت مابینی به صورت رابطه (۴) تعریف می‌شود:

$$B_u = \sum_{ij} \frac{\sigma(i, u, j)}{\sigma(i, j)} \quad (4)$$

که در آن (i, u, j) تعداد کوتاه‌ترین مسیرها بین دو اجتماع i و j است که از گره یا لبه u عبور می‌کند. $\sigma(i, j)$ تعداد کل مسیرها بین دو گره i و j است. در این روش با پیدا نمودن لبه‌ای که بیشترین مرکزیت یا ارتباط با دیگر گره‌ها را دارد و حذف بازگشتی آن لبه در نهایت اجتماعات جدا شده بدست می‌آید. روش‌های دیگری نیز مانند ضریب خوش‌بندی لبه نیز وجود دارد که از این ایده استفاده نموده‌اند ولی در رابطه مربوطه لبه‌ایی که اجتماعات را به یکدیگر مرتبط می‌کنند ضریب کوچکی دارند که در هر مرحله حذف می‌شوند [9].

روش‌های دیگر که از جمله روش‌های طیفی است با ایجاد یک برنامه بر روی گراف به تقسیم نمودن گراف می‌پردازد. این معیار بر این اساس است که یک گروه‌بندی خوب تعداد لبه‌های بین اجتماعات را کمینه می‌کند، در عین حال لبه‌های داخل اجتماع را نگه می‌دارد [11]. این معیار به صورت زیر تعریف می‌شود:

گراف $G = \langle V, E \rangle$ را در نظر بگیرید که در آن V مجموعه گره‌های گراف و E مجموعه تمام لبه‌های گراف است. گره-

روش‌های ترکیب نتایج استفاده از ماتریس همبستگی^{۱۱} است. روش خوش‌بندی ترکیبی انباست مدارک (EAC) که مبتنی بر ماتریس همبستگی است اولین بار توسط فرد و جین در [6] مطرح شد و خیلی زود به صورت یک روش متداول درآمد. امروزه روش‌های دیگری نیز مبتنی بر ماتریس همبستگی ارایه شده‌اند [7]. بنابراین به صورت کلی این دو مرحله در اصل یک چارچوب برای هر روش ترکیبی خواهد بود. در ابتدا نتایج مختلف را از اجرای روش‌های پایه بدست می‌آوریم و سپس با اجرای یکتابع توافقی که در بخش بعدی به آن پرداخته می‌شود نتایج مختلف را با هم ترکیب می‌کنیم تا یک نتیجه دقیق و مستحکم و از همه مهم‌تر پایدار را ایجاد نماییم.

۳- روش‌های موجود

در این مقاله از روش‌های مختلف خوش‌بندی ترکیبی در جهت بهبود تشخیص اجتماعات استفاده شده است. در این بخش به بررسی این دو مورد می‌پردازیم.

۳-۱- تشخیص اجتماعات

روش‌های مختلفی برای بدست آوردن اجتماعات در شبکه وجود دارد. بررسی تک‌تک گره‌ها و قرار دادن آن‌ها در اجتماعات مختلف و در نهایت ارزیابی آن دارای هزینه زمانی و محاسباتی بالای است و این رویکرد به طور عملی ممکن‌پذیر نیست. برای غلبه بر این مشکل روش‌های مکانشفه‌ای^{۱۲} به کمک آمده‌اند. که از جمله این روش‌ها می‌توان به روش‌های طیفی^{۱۳}، تقسیم‌کننده^{۱۴}، تجمعی^{۱۵}، روش‌های مبتنی بر بهینه‌سازی مازولاریتی^{۱۶}، تکنیک‌های حریصانه^{۱۷}، گداختگی شبیه‌سازی شده^{۱۸}، الگوریتم‌های تکاملی^{۱۹} اشاره نمود. هر کدام از این روش‌ها به نوعی در بهبود عملکرد تشخیص اجتماعات موثر است [4][3]. مازولاریتی از پرکاربردترین معیارهای مورد استفاده در روش‌های مختلف است. این معیار کمیتی از گروه‌بندی که از کل گراف بدست آمده است، ارائه می‌کند و نقش بسزایی در تعیین صحت گروه‌بندی دارد. معمولاً برای ارزیابی هر روش تشخیص اجتماعات در شبکه، مقدار مازولاریتی گروه‌بندی پیشنهادی آن روش را برای شبکه‌ها و گراف‌های مختلف محاسبه می‌کنند. هرچه مازولاریتی به دست آمده بیشتر باشد، دقت روش موردنظر بهتر بوده است. معیار مازولاریتی مهم‌ترین محک در ارزیابی روش‌های تشخیص اجتماعات در شبکه‌های اجتماعی است. مازولاریتی به صورت رابطه (۳) زیر تعریف می‌شود [9][8][3]:

$$Q = \sum_i (e_{ii} - a_i^2) \quad (3)$$

$$a_i = \sum_j e_{ij}$$

در رابطه فوق i و j اندیس‌های اجتماع است و e_{ii} نسبت تعداد لبه‌هایی که گره‌های داخل اجتماع i را به هم متصل می‌کند به

در این روش هر کروموزوم کل گره‌های موجود در شبکه را شامل می‌شود. در ابتدا به هر گره به صورت تصادفی یک اجتماع نسبت داده می‌شود و سپس crossover که تغییراتی روی آن صورت گرفته، انجام می‌شود. گاهی نیز mutation گره‌های دو اجتماع را جابجا می‌کند اعمال می‌شود و در نهایت هر کروموزوم توسط معیار مازولاریتی سنجیده می‌شود.

روش‌های دیگری نیز در خصوص تشخیص اجتماعات وجود دارد که در اینجا به این چند روش اشاره کردیم که در این مقاله نیز تا حدودی استفاده شده است. در [18] به الگوریتم‌های دیگری برای تشخیص اجتماعات اشاره شده است.

۲-۳- خوشبندی ترکیبی

در این بخش روش‌های مختلف خوشبندی ترکیبی با استفاده از ماتریس همبستگی را مورد بررسی قرار می‌دهیم.

در [6] یک روش برای ترکیب خوشبندی ایجاد شده از مرحله یکم برای بدست آوردن خوشبندی نهایی ارائه شده است. ایده اصلی در این روش که اນباشت مدارک نام دارد، بدین صورت می‌باشد که شباهت بین نتایج مختلف بدست آمده در مرحله اول از خوشبندی ترکیبی را مورد بررسی قرار می‌دهد. روش کار بدین صورت می‌باشد که نتایج m الگوریتم خوشبندی توسطتابع توافقی بررسی و در یک ماتریس همبستگی $n \times n$ ذخیره می‌شود. تابع توافقی بدین صورت رابطه (8) تعريف می‌گردد:

$$C(i, j) = \frac{n_{i,j}}{m} \quad (8)$$

که $n_{i,j}$ تعداد دفعاتی است که جفت نمونه‌های i و j با هم در یک خوشبندی شده‌اند و m تعداد الگوریتم‌های استفاده شده یا نتایج m الگوریتم خوشبندی می‌باشد. C نیز ماتریس همبستگی حاصله از ترکیب نتایج می‌باشد، پس از ساخت ماتریس همبستگی می‌توان با استفاده از یکی از الگوریتم‌های سلسله مراتبی نظری اتصال منفرد یا اتصال میانگین^{۱۱}، خوشبندی نهایی را استخراج کرد. این نکته قابل یادآوری است که در اینجا تنها یک ماتریس استخراج می‌شود و خوشبندی نهایی نیز به سادگی با اعمال یکی از الگوریتم‌های سلسله مراتبی از این ماتریس به دست می‌آیند.

در [7] نیز سه روش دیگر بر مبنای ماتریس همبستگی ارائه شده است که عبارتند از^{۱۲} CTS^{۱۳}, SRS^{۱۴}, ASRS^{۱۵}. در هر سه الگوریتم ارائه شده یک ماتریس همبستگی متفاوت ایجاد می‌شود که هر یک از این ماتریس‌ها شباهت‌های میان روش‌های خوشبندی اولیه را نشان می‌دهد.

CTS: ایده این روش بر پایه تخمین تعداد مثلث‌های مرتبط با هم می‌باشد. در این روش با محاسبه تعداد مثلث‌های ایجاد شده بین

های گراف را به دو مجموعه مجزای B, A تقسیم می‌کنیم به طوریکه داشته باشیم، $B = V - A$ باشد. مقدار این رابطه کسری از اتصالات بین B, A با توجه به اتصالات جداگانه A, B می‌باشد.

مقدار cut بین B, A به صورت رابطه (5) است:

$$cut(A, B) = \sum_{i \in A, j \in B} W(i, j) \quad (5)$$

و مقدار association بدین صورت است:

$$assoc(A, V) = \sum_{i \in A, v \in V} W(i, v) \quad (6)$$

$W(i, v)$ وزن بین گره i و v است. رابطه (6) برای نرمال کردن اندازه خوشبندی‌ها (اجتمعات) بکار می‌رود. این روش را معمولاً برای گراف‌های وزن‌دار استفاده می‌کنند که البته قابل تعمیم به گراف‌های بدون وزن نیز هست (با در نظر گرفتن وزن یک در صورت وجود لبه و صفر در صورت عدم وجود لبه).

حال از دو رابطه اخیر برای تعریف برش نرمال شده استفاده می‌کنیم.

$$N_{cut}(A, B) = \frac{cut(A, B)}{assoc(A, V)} + \frac{cut(A, B)}{assoc(B, V)} \quad (7)$$

همان‌طور که در رابطه (7) می‌بینیم این رابطه فقط شامل دو اجتماع می‌باشد، که نیاز به تعمیم برای پیدا نمودن اجتماعات دیگر دارد. بدین صورت که همین روال را بر روی اجتماع کوچکتر ایجاد شده اجرا نماییم تا زمانی که به بهترین مقدار مازولاریتی دست پیدا نماییم.

یکی از مهم‌ترین روش‌ها، روش تجمعی می‌باشد. این روش بر این حقیقت بنا شده که گره‌های یک اجتماع ویژگی‌های مشترکی دارند و می‌توان از این ویژگی‌های مشترک برای گروه‌بندی استفاده کرد. در برابر روش‌های تقسیم‌کننده، روش تجمعی در ابتدا همه گره‌ها را جدا از هم و غیر متصل در نظر می‌گیرد و آن‌ها را بر اساس ویژگی‌های مشترک به هم متصل می‌کند تا به اجتماعات برسد. در واقع نحوه عملکرد این رویکرد بدین صورت است که لبه‌های بین اجتماعات خود به خود حذف شده و تنها لبه‌های داخل اجتماع باقی می‌مانند^[12]. از جمله روش‌های دیگر روش بیشینه‌سازی مازولاریتی می‌باشد، جستجو برای یافتن مازولاریتی بینه (یا بیشینه) از نوع مسائل بسیار سخت می‌باشد. چرا که فضای تقسیم‌بندی‌هایی که از گراف امکان پذیر است با بزرگ شدن اندازه گراف به سرعت در حال افزایش است. به همین دلیل رویکردهای جستجوی مکاشفه‌ای برای محدود کردن فضای جستجو الزامی است. در [16][13][14][15][17] روش‌هایی بر پایه ادغام کردن اجتماعات ارائه شده است که بیشتر توسط نیومون بوده است، بطوری که مازولاریتی بیشینه شود.

از جمله روش‌های دیگر در این زمینه روش‌های مبتنی بر الگوریتم‌های تکاملی هستند. روش‌های مختلفی در حوزه الگوریتم‌های تکاملی بر روی تشخیص اجتماعات انجام شده است. تارگین و بینگل^{۱۶} [17] روشی ارائه کرده که از مازولاریتی برای سنجش استفاده می‌کند.

خوشبندی برای بدست آوردن ماتریس همبستگی نهایی استفاده می-شود.

در این مقاله از این روش‌ها به عنوان تابع توافقی برای اجرای بر روی نتایج روش‌های تشخیص اجتماعات استفاده می‌نماییم.

۴- روش پیشنهادی

با توجه به اینکه اکثر روش‌های تشخیص اجتماعات پایه روی جنبه‌های خاصی از داده‌ها تمرکز می‌کنند، در نتیجه 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 نماید [5].

بنابراین تشخیص اجتماعات ترکیبی با ترکیب نتایج بدست آمده از الگوریتم‌های پایه در صدد پیدا نمودن نتایج دقیق‌تر، مطمئن‌تر و پایدارتر از نتایج بدست آمده قبلی می‌باشد. همان‌طوری که در شکل (۳) مشاهده می‌شود، انباره داده نشان داده شده در این شکل در اصل همان گراف ما خواهد بود، که از ارتباطات موجود در شبکه اجتماعی مورد بررسی ایجاد شده است. مجموعه این ارتباطات در نهایت یک گراف و از آن یک ماتریس همسایگی ایجاد می‌گردد که در صورت وجود ارتباط بین گره‌ها یک و در غیر این صورت صفر منظور می‌گردد. با اعمال الگوریتم‌های مختلف بر روی این ماتریس همسایگی و گرفتن نتایج متفاوت و در بعضی موارد یکسان می‌توان از روش ترکیبی استفاده نمود.

همان‌طوری که در روش‌های خوشبندی ترکیبی ارائه شد برای ترکیب نتایج از توابع توافقی استفاده می‌شود. توابع توافقی استفاده شده در این مقاله همان توابع EAC، SRS، CTS، ASRS می‌باشد. بدین‌صورت که اگر گره‌ای در چند نتیجه مختلف جزء یک اجتماع باشد احتمال قرارگیری آن گره در آن اجتماع بیشتر خواهد بود، در نهایت با ترکیب نتایج اولیه یک ماتریس همبستگی ایجاد خواهد شد که یک گراف وزن‌دار از نتایج ما می‌باشد، از مجموع نتایج کمک می-گیریم تا به یک نتیجه بهتر، دقیق‌تر، پایدارتری برسیم.

CD: Community Detection

Input: A(Adjacency matrix)

for Number of CD method do

```
    while (Can change initialization parameter) do
        ECD+:=run(CD method)
    end while
    if(Can change final state CD method) then
        ECD+:=run(CD method)
    end if
end for
```

ECD: Ensemble members of Community Detection methods
/*Each column show one result of community detection method*/

Co.assosiation matrix:=run(Consensus function)
C:=run(Hierarchical clustering method)

Output:R(Final CD result)

شکل (۲): شبکه کد روش تشخیص اجتماعات ترکیبی

لبه‌های دو خوشه که با هم در ارتباط هستند به ایجاد ماتریس همبستگی می‌پردازد. بدین صورت که تعداد مثلث‌های مرتبط با هم در بین دو خوشه به مجموع تعداد مثلث‌های ممکن، ماتریس مشابه را تشکیل می‌دهد. مثلث بدین معنی می‌باشد که دارای سه راس می‌باشد، اگر تمام این سه راس درون دو خوشه i و j باشد آن را به عنوان CT_{ij} معرفی می‌کنیم. بدین صورت که تعداد کل مثلث‌های ممکن بدین شکل را محاسبه می‌کنیم و تعداد کل مثلث‌های ممکن بدین صورت که حداقل یک راس آن در خوشه i و j باشد را CT_{\max} می‌گوییم. با محاسبه CT_{ij} و CT_{\max} می‌توان شباهت بین دو خوشه i و j را طبق رابطه (۹) محاسبه نمود.

$$SimCT(i, j) = \frac{CT_{ij}}{CT_{\max}} \quad (9)$$

همان‌طوری که در رابطه (۹) مشاهده می‌شود CT_{ij} تعداد مثلث‌های مرتبط با هم که همه رؤوس آن بین دو خوشه i و j می‌باشد و CT_{\max} تعداد کل مثلث‌های مرتبط با هم که حداقل یک راس آن بین دو خوشه i و j می‌باشد.

SRS: ایده این روش بدین صورت است که همسایه‌ها، در یک ماتریس شبیه به هم خواهند بود اگر به اندازه کافی درجه شباهت به هم داشته باشند. به عبارت دیگر شباهت بین دو گره با استفاده از لبه‌هایی که به طور مستقیم به این دو گره متصل هستند بدست می‌آید. بدین معنی که هرچه گره دارای درجه بالاتری باشد و به گره مدنظر نیز متصل باشد، شباهت آن گره با استفاده از گره‌های همسایه بدست می‌آید. با محاسبه تعداد لبه‌های متصل به دو گره مختلف می‌توان شباهت بین آن‌ها در خوشبندی‌های مختلف در یک ماتریس همبستگی بدست آورد.

$$SRS(i, j) = \frac{SR(i, j)}{SR_{\max}} \times DC \quad (10)$$

در رابطه (۱۰) صورت کسر $SR(i, j)$ درجه دو گره i و j خواهد بود، به شرطی که این دو گره در یک خوشبندی نباشند و مخرج آن SR_{\max} بیشترین مقداری می‌باشد که $SR(i, j)$ بدون توجه به هم خوشه بودن می‌تواند داشته باشد است. مقدار $DC = [0.1]$ می‌باشد که نشان دهنده درجه اعتمادی می‌باشد که دو گره‌ای که در یک خوشبندی نباشند شباهت داشته باشند یا خیر. معمولاً مقدار DC را با توجه به گراف مورد بررسی مقدار ثابتی در نظر می-گیریم.

ASRS: این روش نیز همانند روش SRS می‌باشد با این تفاوت که در SRS شباهت بین دو گره در تمام روش‌های خوشبندی مورد بررسی قرار می‌گیرد ولی در این روش تقریبی از تمامی روش‌های

قابلیت اعتماد را نیز برای ما به همراه می‌آورد، نتایج ارزیابی‌ها دلیل برای ادعا می‌باشد.

از نظر هزینه اجرایی، هزینه اجرایی این روش بالا و مقدار مصرف حافظه آن نیز بالا می‌باشد و علت آن نیز بدلیل استفاده از روش‌های مختلف و به تعداد زیاد می‌باشد. ولی هزینه اجرایی این روش برابر با بدترین حالت اجرای یک الگوریتم از بین تمام الگوریتم‌های تشخیص اجتماعات خواهد بود، چرا که تعداد الگوریتم‌های مورد بررسی در این فرایند ترکیبی در نهایت محدود خواهد بود.

به طور کلی روش پیشنهادی تشخیص اجتماعات ترکیبی در صدد ترکیب نتایج مختلف تشخیص اجتماعات برای بدست آوردن نتایج دقیق‌تر و پایدارتر با بررسی نتایج اولیه می‌باشد.

۱-۴- ارزیابی روش پیشنهادی

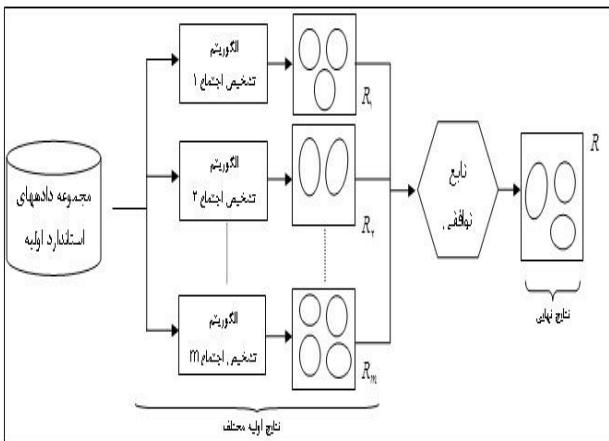
در این مقاله چهار روش مختلف تشخیص اجتماعات با استفاده از MATLAB (ver R2009a) پیاده‌سازی شده است، که در صدد آن هستیم که با ترکیب این چهار روش با استفاده از خوشه‌بندی ترکیبی به یک خوشه‌بندی یا اجتماعات دقیق‌تری نسبت به تک‌تک نتایج بررسیم. از مزایای این روش می‌توان به مقیاس‌پذیری بالای آن اشاره نمود بدین معنا که هر چه تعداد روش‌های ما برای تشخیص اجتماعات بیشتر باشد نتایج بهتری را در خوشه‌بندی ترکیبی به ارمغان می‌آورد و نیز هر روش را بدون محدودیت و با تغییر شرط اتمام آن می‌توان استفاده نمود، چرا که ما در خوشه‌بندی ترکیبی به نوع روش نگاه نمی‌کنیم بلکه نتیجه روش را مورد بررسی قرار می‌دهیم و با بررسی نتایج بدست آمده از تک‌تک روش‌ها یک خوشه‌بندی دقیق‌تری را ارائه می‌دهیم.

در اینجا به بررسی این چهار الگوریتم پیاده‌سازی شده برای تشخیص اجتماعات می‌پردازیم که تک‌تک به چه صورتی اجتماعات را بدست می‌آورند. کلیه این روش‌ها از روی ماتریس مجاورت ایجاد شده از روی گراف به طوری که اگر ارتباطی بین گره‌ها وجود داشته باشد یک و در غیر این صورت صفر خواهد بود استفاده می‌کنند و نیز کلیه این روش‌ها نیز قابل تعمیم برروی گراف‌های جهت‌دار و وزن دار می‌باشد.

الگوریتم اول که با نام Newman یا N در این مقاله معرفی شده است، یک روش برنبایی مرکزیت مابینی می‌باشد، که در روش‌های موجود به آن اشاره شد. در این روش لبه‌ایی که بیشترین مرکزیت را داشته باشند، به صورت بازگشتی حذف نموده و به این ترتیب یکسری اجتماعات جدا از هم را خواهیم داشت [9].

الگوریتم دوم که با نام Greedy یا NG در این مقاله معرفی شده است، یک روش بر مبنای روش تجمعی ارائه شده است. ایده اصلی در این روش بهینه‌سازی مأذولاتی و حریصانه می‌باشد، در این روش با استفاده از رابطه مأذولاتی در صدد پیدا نمودن اجتماعاتی با پیشنهادین حالت مأذولاتی هستند. این روش چون از انواع روش-

در این مقاله یک روش جدید برای افزایش دقت نتایج تشخیص اجتماعات ارائه شده است. این روش که در شکل(۲) نشان داده شده است، روش تشخیص اجتماعات ترکیبی می‌باشد با استفاده از روش‌های پایه‌ای برای بدست آوردن نتایج پایدارتر و دقیق‌تر مورد استفاده قرار می‌گیرد. شبه کد نشان داده شده در شکل(۲) با دریافت یک ماتریس همسایگی و اجرای روش ترکیبی با استفاده از ماتریس همبستگی در تشخیص اجتماعات نتیجه پایدارتر و دقیق‌تری را به عنوان نتیجه نهایی ارائه می‌کند.



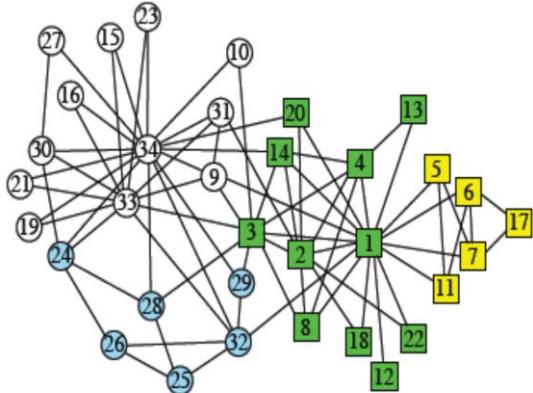
شکل(۳): الگوریتم پیشنهادی برای تشخیص اجتماعات ترکیبی

ماتریس همبستگی ایجاد شده برآیندی از تمامی روش‌های اولیه خواهد بود، در این ماتریس برای هر لبه در گراف مورد نظر عددی منظور می‌شود که این عدد مقدار وابستگی آن لبه به گره‌ای که با آن در تماس می‌باشد را نشان می‌دهد. به عبارت دیگر اجماع روش‌های مختلف در مورد اینکه این دو گره به چه میزان با هم در ارتباط هستند یا خیر را نشان خواهد داد. لبه‌های که ارتباط ضعیف دارند، در الگوریتم تشخیص اجتماعات ترکیبی سیاست حذف آن‌ها اتخاذ می‌گردد که به دقیق‌ترین نتایج نزدیک باشد. در نهایت با اعمال یکی از الگوریتم‌های سلسله مراتبی برای استخراج نتایج نهایی بر روی آن ماتریس همبستگی، تشخیص اجتماعات کامل‌تر و دقیق‌تر را خواهیم داشت.

بعد از اعمال الگوریتم‌های مختلف بر روی شبکه مورد تحلیل همان‌طوری که در شکل(۳) نیز نشان داده می‌شود، یک سری نتایج متفاوت بدست خواهد آمد که به صورت $\{R_1, R_2, \dots, R_m\}$ تعریف می‌گردد. این نتایج تنها با اجرای روش‌های تشخیص اجتماعات متفاوت بدست آمده‌اند. بدین معنی که الگوریتم تشخیص اجتماعات مختلف را اجرا می‌نماییم و نتایج مختلف را به عنوان ورودی به تابع توافقی ارسال می‌کنیم، که با ترکیب صحیح روش‌های مختلف با استفاده از توابع توافقی که در این مقاله به آن اشاره شده است، می‌توان به یک نتیجه بسیار بهتر، مطمئن‌تر، پایدارتر و دقیق‌ریسید. چون این نتیجه از اجماع یا هم‌فکری روش‌های مختلف ایجاد می‌شود این



تجمع را تشخیص داد. یکی اطراف گره ۳۳ و ۳۴ و دیگری اطراف گره ۱ قرار دارد[20]. رنگ‌های نشان داده شده در این شکل نیز تعداد گره‌های این باشگاه را به ۴ اجتماع مختلف تفکیک نموده است که هر رنگ نشان دهنده یک اجتماع می‌باشد.



شکل(۴): گراف باشگاه کاراته [3]

مجموعه داده‌های موسیقی دانان جاز: در این پایگاهداده ۱۹۶ باند موسیقی وجود دارد که بین سال‌های ۱۹۱۲ و ۱۹۴۰ فعالیت می‌کردند که بیشتر آن‌ها در سال‌های اطراف ۱۹۲۰ مشغول بوده‌اند. این پایگاهداده موسیقی دان‌هایی را لیست می‌کند که در این باندها می‌نواختند. این گراف شامل ۱۹۶ گره و ۲۷۴۲ لبه است، اگر دو موسیقی دان در یک باند می‌نواختند لبه‌ها بین آن‌ها ترسیم می‌شود[21].

لیگ فوتبال دانشگاه امریکا: این شبکه که لیگ فوتبال دانشگاه امریکا را نشان می‌دهد، از ۱۱۵ گره تشکیل شده است. هر گره نشان دهنده یک تیم می‌باشد و در صورت بازی بین دو تیم در یک فصل یک لبه بین آن دو ایجاد می‌گردد. تعداد لبه‌ها در این شبکه ۶۱۶ لبه می‌باشد که نشان دهنده تعداد بازی‌های انجام شده بین دو تیم بوده است. این مشاهدات برای سال‌های ۲۰۰۱ تا ۲۰۰۰ بوده است[9].

شبکه متابولیک: در یک سلول یا میکرو ارگانیزم فرآیندهایی که انتقال انرژی و اطلاعات را تولید می‌کنند، به طور یکپارچه از طریق شبکه پیچیده‌ای از مولکول‌های سلولی و واکنش‌های اینشان است. با وجود این که نقش کلیدی این شبکه‌ها حفظ کارکردهای سلولی است. اساساً ساختار بزرگ آن‌ها ناشناخته است. این شبکه، یک شبکه متابولیکی کرم‌ها به نام C.Elegans است که ۴۵۳ گره دارد. گره‌ها نمایانگر متabolیک‌ها و لبه‌ها نشان دهنده واکنش‌های متabolیکی است[16].

مجموعه داده‌های ایمیل: مجموعه داده‌های ایمیل Enron توسط FERC منتشر شد و البته اشکالاتی هم داشت. پس از آن توسط افراد دیگری بررسی شد و بسیاری از اشکالات آن بر طرف شد. این داده‌ها همه نوع ایمیل‌های شخصی و اداری است. این نسخه از داده‌ها شامل ۴۳۱,۵۱۷ ایمیل از ۱۵۱ کاربر است که در ۳۵۰۰ فولدر توزیع شده است. این گراف شامل ۱۱۳۳ گره و ۵۴۵۱ لبه است این

های تجمعی می‌باشد به یک معیار مشابه نیاز داریم که در این روش مازوچیتی معیار مشابه خواهد بود[14].

الگوریتم سوم که با نام Newman & Girvan یا GN در این مقاله معرفی شده است، یک روش مشابه مرکزیت مابینی می‌باشد با این تفاوت که به جای بررسی تک‌تک گره‌ها در هر مرحله و بدست آوردن یک لبه برای حذف شدن در هر مرحله لبه‌ها با ضرب بالا را حذف نموده و در نهایت یک دندروگرام از کل گراف ایجاد می‌شود که یک نمایش سلسله مراتسی می‌باشد، با برش این دندروگرام می‌توان اجتماعات را تفکیک نمود ولی در برش می‌بایست دقت نمود که بهینه‌ترین حالت را داشته باشد، برای پیدا نمودن بهینه‌ترین حالت نیز از مازوچیتی استفاده شده است[15].

الگوریتم چهارم که با نام Radicchi & et al R در این مقاله معرفی شده است، نیز یک روش بر مبنای بهینه‌سازی مازوچیتی می‌باشد که با استفاده از مازوچیتی و تعریف متغیرهای دیگری که در آن مقادیر مازوچیتی و نیز حالات گره‌ها در آن وضعیت ذخیره می‌شود. این روش باعث افزایش سرعت در پیدا نمودن اجتماعات و نیز با داشتن یک پیش‌زنی از مراحل قبلی با دقت بالاتری به ترکیب گره‌ها می‌پردازد[19].

حال با ترکیب این چهار روش با استفاده از توابع توافقی که مبتنی بر ماتریس هم‌ستگی می‌باشد، به ترکیب نتایج می‌پردازیم، تا بتوانیم در تشخیص اجتماعات نتایج دقیق تری داشته باشیم. برای ارزیابی روش پیشنهادی از تعدادی داده استاندارد استفاده شده است که به بررسی آن‌ها می‌پردازیم.

۲-۴- داده‌های مورد استفاده

داده‌های مورد استفاده داده‌های استانداردی می‌باشند که تقریباً در تمامی روش‌های تشخیص اجتماعات از آن‌ها استفاده می‌شود. از جمله این داده‌ها می‌توان به باشگاه کاراته زاکاری^{۲۶}[20]، مجموعه داده‌های موسیقی دانان جاز^{۲۷}[21]، لیگ فوتبال دانشگاه امریکا^{۲۸}[9]، شبکه متabolیک، مجموعه داده‌های ایمیل Ernon توسط FERC^{۲۹} و مجموعه داده NetSciecne^{۳۰} اشاره نمود.

باشگاه کاراته زاکاری: همان‌طوری که در شکل (۴) مشاهده می‌شود، این مجموعه یک گراف استاندارد برای تشخیص اجتماع است و در بسیاری از مقالات از آن استفاده شده است. این گراف شامل ۳۴ گره و ۷۸ لبه است، که اعضای یک باشگاه کاراته در ایالات متحده هستند و در طول ۳ سال مشاهده شده‌اند. لبه‌ها، افرادی را که خارج از فعالیت‌های باشگاه با هم ارتباط دارند به هم متصل می‌کند. گره ۳۴ مسئول باشگاه و گره ۱ مری می‌باشد، در بعضی نقاط درگیری بین مسئول باشگاه با مری موجوب ایجاد شکاف بین اعضای باشگاه شده است و آن‌ها را به دو گروه طرفدار این دو فرد تقسیم کرده است (با دایره و مربع نشان داده شده است). با نگاه به شکل (۴) می‌توان دو

همان طوری که در جدول(۱) مشاهده می شود معیار ارزیابی برای روش پیشنهادی مازولاریتی می باشد که مقدار آن بین (۰,۱+)- می باشد. اگر مقدار مازولاریتی صفر شود یعنی تمامی گره ها را در یک اجتماع قرار داده ایم و اگر منفی شود به معنی خطای زیاد در پیدا نمودن اجتماعات می باشد. در تمامی نتایج بالا مقدار آن مثبت بوده است، روش ما در تمامی حالت دارای دقت، پایداری و صحت بالایی می باشد. بدین معنی که همواره جواب بهینه بدست آمده توسط روش های تشخیص اجتماعات معمولی را دارد و در صدد پیدا نمودن جواب بهینه تری از ترکیب نتایج می باشد. با این حال که نوع داده در حال تغییر است ولی به جواب بهینه نزدیک می باشد. اگر این روش را بر روی داده های بسیار بزرگ تست گردد این تفاوت مشهودتر خواهد بود چراکه این روش های تشخیص اجتماعات معمولی بر روی داده های بزرگ تقریباً جواب خوبی را نخواهند داشت، ولی برآیند نتایج جواب بهینه تری را خواهد داشت.

حال به بررسی روش های تشخیص اجتماعات ترکیبی و مقایسه آن با روش های پایه ای در تشخیص اجتماعات می پردازیم. روش های پایه همان چهار روش ارایه شده در روش پیشنهادی می باشند و روش های ترکیبی برای روش پیشنهادی نیز از توابع ترکیبی مبتنی بر ماتریس همبستگی استفاده می نمایند.

ایمیل ها بخش ضمیمه شده (مثل عکس، فیلم یا هر فایل دیگری) ندارد [22].

مجموعه داده NetSciecne: در این شبکه تعداد ۱۵۸۹ نویسنده وجود دارد که با هم در مورد تئوری شبکه پژوهش می کنند. این مجموعه داده با بررسی انجام شده بر روی مقالات مختلف در تئوری شبکه توسط نیومن در سال ۲۰۰۶ گردآوری شد. بدین صورت که ارتباطات این نویسندها در مقالات مختلف را به صورت یک شبکه مدل می نماید. تعداد لبه ها در این شبکه ۲۷۴۲ لبه می باشد که نشان دهنده ارتباطات بین نویسندها است. حال که پیش زمینه لازم برای روش پیشنهادی فراهم شد و داده های مورد استفاده در ارزیابی نیز مورد بررسی قرار گرفت، حال به بررسی نتایج گوناگون می رسیم.

۳-۴ - نتایج ارزیابی

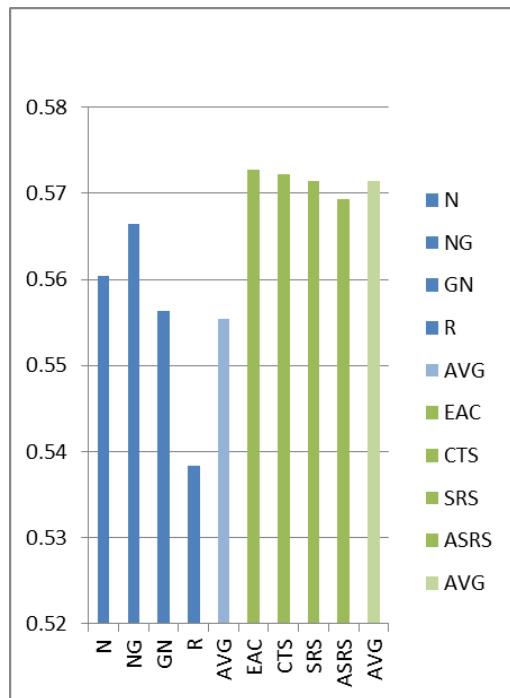
جدول(۱) نتایج اعمال الگوریتم های مختلف را نشان می دهد. این جدول شش نوع داده راچ که برای ارزیابی روش های تشخیص اجتماعات مورد استفاده قرار می گیرند را نشان می دهد. تعداد گره ها و لبه های این داده های استاندارد در جلوی آنها مشخص گردیده است. یک الگوریتم خوب می باشد که در شبکه های اجتماعی تعداد گره ها خوب و بهینه جواب دهد چرا که در شبکه های اجتماعی تعداد گره ها بسیار متنوع می باشند و در بعضی موارد تعداد آن به کل کاربران در شبکه جهانی تامین پیدا می کند. روش ارزیابی این الگوریتم ها نیز مازولاریتی می باشد که در روش های موجود به آن اشاره گردیده است، و این روش (که مقدار آن بین منفی یک و یک می باشد) مهمترین محک برای ارزیابی تشخیص اجتماعات شناخته شده است.

جدول (۱): مقایسه روش های مختلف تشخیص اجتماعات بر اساس معیار مازولاریتی

Radicchi & et al [19]		Newman & Girvan [15]		Newman Greedy [14]		Newman [9]		تعداد لبه ها	تعداد گره ها	Data Set
۳ اجتماع	0.3853	۴ اجتماع	0.4009	۴ اجتماع	0.3980	۴ اجتماع	0.4086	۷۸	۳۴	کاراته
۶ اجتماع	0.5581	۶ اجتماع	0.5714	۹ اجتماع	0.6057	۷ اجتماع	0.5788	۶۱۶	۱۱۵	فوتبال
۴ اجتماع	0.4394	۵ اجتماع	0.4397	۴ اجتماع	0.4442	۳ اجتماع	0.4389	۲۷۴۲	۱۶۸	جاز
۱۰ اجتماع	0.4019	۱۱ اجتماع	0.4334	۱۳ اجتماع	0.4269	۱۰ اجتماع	0.4332	۲۰۳۲	۴۵۳	متابولیکی
۱۱ اجتماع	0.5036	۱۰ اجتماع	0.5463	۱۰ اجتماع	0.5766	۱۱ اجتماع	0.5485	۵۴۵۱	۱۱۳۳	ایمیل
۴۰۳ اجتماع	0.9419	۴۰۵ اجتماع	0.9472	۴۰۹ اجتماع	0.9511	۴۰۷ اجتماع	0.9547	۲۷۴۲	۱۵۸۹	NetScience



همان‌طوری که مشاهده می‌شود نتایج حاصل دارای دقت بالاتر و نیز پایداری بیشتر نسبت به روش‌های پایه‌ای دارد.



شکل(۵): نمودار حاصل از نتایج مختلف پایه و پیشنهادی

همان‌طوری که در جدول(۲) ملاحظه می‌شود روش‌های N, NG و R روش‌های پایه برای تشخیص اجتماعات و روش‌های ASRS و SRS، EAC، CTS روش‌های ترکیبی تشخیص اجتماعات می‌باشند. در دو ستون AVG به ترتیب مقادیر میانگین نتایج روش پایه و ترکیبی آورده شده است. در سطر آخر از جدول(۲) میانگین اعمال نتایج مختلف یک الگوریتم یا روش خاص بر روی تمامی مجموعه داده‌ها آورده شده است که نشان دهنده نحوه اجرای و یا چگونگی نتایج آن الگوریتم یا روش مورد استفاده بر روی مجموعه داده‌های متفاوت می‌باشد.

همان‌طوری که در جدول(۲) مشاهده می‌شود اعمال روش ترکیبی نتایج دقیق‌تر و پایدارتری را به نسبت روش‌های پایه‌ای به ارمغان می‌آورند. در شکل(۵) به صورت نموداری این موضوع مشهودتر می‌باشد.

همانند جدول(۲) در زیر نمودار حاصل از شکل(۵) نیز روش‌های N, NG و R روش‌های پایه برای تشخیص اجتماعات و روش‌های ASRS و SRS، EAC، CTS روش‌های ترکیبی تشخیص اجتماعات می‌باشند. مقادیر میانگین نتایج روش پایه و ترکیبی به ترتیب با رنگ آبی کم رنگ و سبز کم رنگ آورده شده است که میانگین اعمال نتایج مختلف یک الگوریتم یا روش خاص تشخیص اجتماعات ترکیبی بر روی تمامی مجموعه داده‌ها آورده شده است.

جدول (۲): ماتریسی محاسبه شده توسط روش‌های اولیه و مقایسه با روش‌های ECD

AVG	ECD				AVG	Basic Community Detection Method				Data Set
	ASRS	SRS	CTS	EAC		R	GN	NG	N	
0.4190	0.4179	0.4197	0.4188	0.4197	0.3982	0.3853	0.4009	0.3980	0.4086	کاراته
0.6002	0.5985	0.5985	0.6019	0.6019	0.5777	0.5581	0.5714	0.6027	0.5788	فوتبال
0.4445	0.4445	0.4438	0.4447	0.4450	0.4405	0.4394	0.4397	0.4442	0.4389	جاز
0.4363	0.4334	0.4374	0.4368	0.4377	0.4238	0.4019	0.4334	0.4269	0.4332	متابولیکی
0.5745	0.5685	0.5763	0.5765	0.5769	0.5437	0.5036	0.5463	0.5766	0.5485	ایمیل
0.9539	0.9534	0.9528	0.9546	0.9551	0.9487	0.9419	0.9472	0.9511	0.9547	NetScience
0.5714	0.5693	0.5714	0.5722	0.5727	0.5554	0.5383	0.5564	0.5665	0.5604	AVG

بندی ترکیبی و ترکیب آن با روش‌های تشخیص اجتماعات روشی را با نام تشخیص اجتماعات ترکیبی را معرفی نمودیم. با پیاده‌سازی چهار روش متفاوت از روش‌های تشخیص اجتماعات و بدست آوردن نتایج آن‌ها و نیز در نهایت با استفاده از توابع توافقی در تشخیص اجتماعات ترکیبی که رویکرد اصلی این مقاله بوده است، به ترکیب نتایج بدست آمده از آن‌ها پرداختیم. برای ترکیب نتایج، یکی از روش‌های خوش‌بندی ترکیبی با نام روش مبتنی بر ماتریس همبستگی، مورد استفاده قرار گرفته است.

همان‌گونه که مشاهده گردید از مزایای این روش استفاده از روش‌های قبلی به تعداد زیاد که نشان از مقیاس‌پذیری بالای این روش و نیز دقیق‌تر و پایدارتری را برای ما خواهد داشت. از معایب این نتایج دقیق‌تر و پایدارتری آن به نسبت روش‌های دیگر و در مجموع نتایج آن که با بدترین حالت هزینه اجرایی در روش به هزینه اجرایی آن که با بدترین حالت هزینه اجرایی در الگوریتم‌های استفاده شده برابر می‌باشد و نیز حافظه زیادی را نیز مصرف می‌نماید، همچنین ممکن است در همه جا بهینه‌ترین جواب را نداشته باشیم و الگوریتم‌هایی باشند که برای داده‌های خاصی جواب دقیق‌تری را داشته باشند و نیز در اعمال این روش می‌بایست تعداد خوش‌ها نیز برای ما مشخص باشد ولی می‌توان از روش‌های دیگر خوش‌بندی ترکیبی استفاده نمود که تعداد خوش‌های در آن نامشخص باشد.

امکان استفاده از روش‌های دیگر توابع توافقی نیز وجود دارد، برای استفاده از آن روش‌ها دیگر به تعداد بیشتری نتایج گوناگون نیاز می‌باشد که با پیاده‌سازی‌های بیشتری از الگوریتم‌های تشخیص اجتماعات می‌توان از آن‌ها نیز استفاده نمود که می‌توان در کارهای بعدی به آن پرداخت. تشخیص اجتماعات ترکیبی می‌تواند در مسائلی از جمله تشخیص دقیق‌تر اجتماعات، بازاریابی، تبلیغات، درک شبکه، بهبود موتورهای جستجو، ارتباطات (از قبیل تلفنی، سیار و...) و تعاملات افراد و فعالیت‌های امنیتی و جاسوسی مورد استفاده قرار گیرد.

مراجع

- [1] Easley, D. and Kleinberg, J. "Networks, Crowds, and Markets: Reasoning about a Highly Connected World", Cambridge University Press, 2011.
- [2] Liu, B. "Web DataMining Exploring Hyperlinks, Contents, and Usage Data", Springer, 2007.
- [3] Fortunato, S. "Community detection in graphs", Physics Reports, vol. 486, no. 3-5, pp. 75–174, 2010.
- [4] Porter, M.A., Onnela, J.P. and Mucha, P.J. "Communities in networks", Notices of the American Mathematical Society, vol. 56, no. 9, pp. 1082–1097, 2009.
- [5] Strehl, A. and Ghosh, J. "Cluster Ensembles -A Knowledge Reuse Framework for Combining Multiple Partitions", Journal of Machine Learning Research, pp. 583-617, 2003.
- [6] Fred, A.L.N. and Jain, A.K. "Data Clustering Using Evidence Accumulation", Proc. of the 16th Intl. Conf.

در شکل(۵) نتایج مختلف را بر روی نمودار نشان می‌دهد، با استفاده از این روش پیشنهادی می‌توان با اطمینان بیشتری به نتایج حاصل از روش‌های تشخیص اجتماعات اتکا نمود. چرا که نتایج روش‌های تشخیص اجتماعات می‌تواند در تصمیمات بلند مدت و یا کوتاه مدت و سیاست‌های سازمانی و یا مواردی از این قبیل تاثیر بسزای داشته باشد.

۴-۴- چالش‌ها

با توجه به اینکه روش پیشنهادی شامل دو مرحله پشت سر هم تولید نتایج اولیه و سپس ترکیب آنهاست، برای تحلیل پیچیدگی زمانی این روش باید پیچیدگی هر دو مرحله را تحلیل کرد. برای مرحله اول، پیچیدگی بدلین الگوریتم استفاده شده در روش ترکیبی می‌باشد. بدین معنا که اگر از ۱۰ روش مختلف در روش‌های ترکیبی استفاده نماییم، به دلیل آنکه تک‌تک روش‌ها می‌بایست اجرا گردد تا نتایج حاصل شود و بعد از اجرا بتوان از ترکیب آن‌ها استفاده نمود، بنابراین پیچیدگی آن برابر با بدترین حالت در هر کدام از ۱۰ روش مورد استفاده می‌باشد. در مرحله ترکیب نیز پیچیدگی زمانی برابر است با پیچیدگی زمانی توابع توافقی مورد استفاده که از متداول‌ترین روش‌های ترکیبی می‌باشد. برای بررسی بیشتر تحلیل پیچیدگی روش‌های ترکیبی مورد استفاده می‌توان به منابع [۶] و [۷] رجوع کرد. روش پیشنهادی در نرم‌افزار مطلب شبیه‌سازی شده است و از ماتریس دو بعدی برای نمایش شبکه استفاده شده است. با توجه به این که انجام عملیات بروی ماتریس‌ها در نرم‌افزار مطلب نیاز به حجم حافظه زیادی دارد، بنابراین بررسی شبکه‌های بزرگتر از ۵۰۰۰ نod در این نرم‌افزار با مشکلات حافظه‌ای روبروست. برای رفع این مشکل و بررسی شبکه‌های با نودهای بیشتر می‌بایست اولاً عملیات بروی ماتریس‌ها را به صورتی اصلاح نمود که به حجم حافظه و عملیات کمتری نیاز داشته باشد و ثانیاً می‌توان ساختمان داده‌ای جدید تعریف کرد و آن را جایگزین ماتریس نماییم و یا اینکه از ماتریس اسپارس استفاده کنیم.

۵. نتیجه‌گیری

هدف این مقاله بهبود نتایج روش‌های تشخیص اجتماعات می‌باشد. روش مورد استفاده در این مقاله رویکرد ترکیبی می‌باشد، با توجه به قدرت روش‌های خوش‌بندی ترکیبی و نتایج دقیق روش مذکور و امکان استفاده آن در روش‌های تشخیص اجتماعات در این مقاله از تشخیص اجتماعات ترکیبی برای بدست آوردن نتایج دقیق‌تر، مطمئن‌تر، پایدارتر و مستحکم‌تر استفاده شده است. بنابراین در این مقاله یک روش جدید برای تشخیص اجتماعات ترکیبی پیشنهاد شده است. در ابتدای مقاله به بررسی شبکه‌های اجتماعی و تشخیص اجتماعات در آن پرداختیم و با معرفی خوش-



- ¹ Community Detection
- ² Clustering
- ³ Ensemble Clustering
- ⁴ Social Network Analysis(SNA)
- ⁵ Graph theory
- ⁶ Node
- ⁷ Edge
- ⁸ Adjacency Matrix
- ⁹ Diversity
- ¹⁰ Consensus Function
- ¹¹ Co-assosiation Matrix
- ¹² Evidence Accumulation Clustering
- ¹³ Heuristic
- ¹⁴ Spectral
- ¹⁵ Divisive
- ¹⁶ Agglomerative
- ¹⁷ Modularity optimization
- ¹⁸ Greedy techniques
- ¹⁹ Simulated annealing
- ²⁰ Evolutionary
- ²¹ Targin and Bingol
- ²² Single Linkage(SL) or Average Linkage(AL)
- ²³ Connected Triple based Similarity
- ²⁴ SimRank based Similarity
- ²⁵ Approximate SimRank based Similarity
- ²⁶ Zachary karate club
- ²⁷ Jazz bands
- ²⁸ American College Football League
- ²⁹ Federal Energy Regulatory Commission

- [7] on Pattern Recognition, ICPR02, Quebec City, pp. 276 – 280, 2002.
- [8] Iam-on, N. and Garrett, S. “*LinkCluE: A MATLAB Package for Link-Based Cluster Ensembles*”, In Journal of Statistical Software, Issue 9, Volume 36, 2011.
- [9] Lancichinetti, A. “*Community detection algorithms: a comparative analysis*”, Physical Review E, vol. 80, no. 5, p. 056117, 2009.
- [10] Girvan, M. and Newman, M.E.J. “*Community structure in social and biological networks*”, Proceedings of the National Academy of Sciences of the United States of America, vol. 99, no. 12, p. 7821, 2002.
- [11] Newman, M.E.J. “*A measure of betweenness centrality based on random walks*”, Social networks, vol. 27, no. 1, pp. 39–54, 2005.
- [12] Radicchi, F., Castellano, C., Cecconi, F., Loreto, V. and Parisi, D. “*Defining and identifying communities in networks*”, Proceedings of the National Academy of Sciences of the United States of America, pp. 2658–2663, 2004.
- [13] Lipczak, M. and Milios, E. “*Agglomerative genetic algorithm for clustering in social networks*”, In Proceedings of the 11th Annual conference on Genetic and evolutionary computation, pp. 1243–1250, 2009.
- [14] Zhang, X.S. and et al. “*Modularity optimization in community detection of complex networks*”, EPL (Euro Physics Letters), vol. 87, p. 38002, 2009.
- [15] Newman, M.E.J. “*Fast algorithm for detecting community structure in very large networks*”, Physical review E, vol. 69, 2004.
- [16] Newman, M.E.J. and Girvan, M. “*Finding and evaluating community structure in networks*”, Physical review E, vol. 69, no. 2, p. 26113, 2004.
- [17] Clauset, A., Newman, M.E.J. and Moore, C. “*Finding community structure in very large networks*”, Physical Review E, vol. 70, no. 6, p. 66111, 2004.
- [18] Tasgin, M. and Bingol, H. “*Community detection in complex networks using genetic algorithm*”, Arxiv preprint cond-mat/0604419, 2006.
- [19] Leskovec, J., Lang, K.J. and Mahoney, M. “*Empirical comparison of algorithms for network community detection*”, In Proceedings of the 19th International conference on World Wide Web, pp. 631–640, 2010.
- [20] Reichardt, J. and Bornholdt, S. “*Statistical mechanics of community detection*”, Physical Review E, 2006.
- [21] Zachary, W.W. “*An information flow model for conflict and fission in small groups*”, Journal of Anthropological Research, vol. 33, no. 4, pp. 452–473, 1977.
- [22] Gleiser, P. and Danon, L. “*Community structure in jazz*”, Arxiv preprint condmat/0307434, 2003.
- [23] Shetty, J. and Adibi, J. “*The Enron email dataset database schema and brief statistical report*”, Information Sciences Institute Technical Report, University of Southern California, 2004.