

تشخیص اجتماعات با رویکرد ترکیبی در شبکه های اجتماعی

رسول حسین زاده^۱، حسین علیزاده^۲ و اسلام ناظمی^۳

^۱دانشگاه شهید بهشتی، r.hossienzadeh@sbu.ac.ir

^۲دانشگاه علم و صنعت، halizadeh@iust.ac.ir

^۳دانشگاه شهید بهشتی، nazemi@sbu.ac.ir

چکیده - تشخیص اجتماعات در شبکه های پیچیده یا شبکه های اجتماعی یکی از مهم ترین مشکلات در زمینه های علمی می باشد. خوشه بندی یا تشخیص اجتماعات، ساختار گروه ها در شبکه های اجتماعی و ارتباطات پنهان بین مولفه های آن را آشکار خواهد نمود. با در نظر گرفتن افزایش پایگاه داده های مربوط به شبکه های اجتماعی، به الگوریتم های مقیاس پذیری برای تجزیه تحلیل آن ها نیاز داریم. در این مقاله ما درصدد پیدا نمودن اجتماعات صحیح و دقیق با استفاده از خوشه بندی ترکیبی هستیم. اکثر روش های رایج تشخیص اجتماعات موجود قطعی نیستند و نتایج آن ها به مقادیر اولیه ای که در اکثر مواقع به صورت تصادفی انتخاب می شود بستگی دارد. خوشه بندی ترکیبی در تحلیل داده ها برای رسیدن به نتایج ثابت (پایدار) بدون توجه به مقادیر اولیه تصادفی می باشد. در این مقاله نشان خواهیم داد که خوشه بندی ترکیبی توانایی ترکیب با هر روش دیگری را خواهد داشت به گونه ای که دقت نتایج اجتماعات را افزایش می دهد. نتایج این بررسی ها می تواند در مسائل بسیاری از جمله تشخیص دقیق تر اجتماعات، بازاریابی، تبلیغات، درک شبکه و بهبود موتورهای جستجو مورد استفاده قرار گیرد.

کلید واژه - خوشه بندی، خوشه بندی ترکیبی، تشخیص اجتماعات، شبکه های اجتماعی.

۱- مقدمه

رویکرد خوشه بندی ترکیبی بدین صورت می باشد که با استفاده از روش های پایه ای و یا هر روش ممکن موجود، با ترکیب تک- تک این روش ها با یکدیگر درصدد پیدا نمودن یک جواب پایدار از نتایج متفاوت می باشد. به طوری که جواب نهایی دارای دقت، پایداری، صحت و از اجماع نتایج روش های مختلف بدست می آید، که دارای اطمینان بالاتری می باشد. با توجه به قابلیت ادغام خوشه بندی ترکیبی و تشخیص اجتماعات در این مقاله برای بهبود نتایج نهایی در تشخیص اجتماعات از «خوشه بندی ترکیبی» استفاده شده است. دقت و صحت اجتماعات نهایی در تشخیص اجتماعات مسئله مهمی می باشد، چرا که الگوریتمی در تشخیص اجتماعات وجود ندارد که در تمام حالات و داده های گوناگون جواب دقیق و حتی نزدیک به آن را داشته باشد.

در این مقاله سعی داریم که با استفاده از خوشه بندی ترکیبی در شبکه های اجتماعی به بررسی اجتماعات در این شبکه ها بپردازیم. بدین صورت که با اعمال روش های مختلف تشخیص اجتماعات و بدست آوردن نتایج متفاوت و در نهایت ترکیب این

امروزه اینترنت و خدمات وب به سرعت در حال گسترش هستند و همزمان شبکه های اجتماعی مجازی نقش مهمی در زندگی واقعی افراد دارد. در واقع شبکه های اجتماعی، شبکه های تعاملی هستند که از اینترنت به عنوان رسانه ای برای ایجاد ارتباط بین افراد استفاده می کنند. با افزایش سریع کاربران شبکه های اجتماعی، کاوش در مقیاس بالای داده ها می تواند کارایی بهتر و موثرتری از پتانسیل پنهان این شبکه ها فراهم کند [۱].

تشخیص اجتماعات^۱ یا خوشه بندی^۲ یکی از مراحل اصلی در داده کاوی است که وظیفه کاوش الگوهای پنهان در داده های بدون برچسب را بر عهده دارد. با توجه به پیچیدگی مسئله خوشه بندی و ضعف روش های پایه ای آن، امروزه اکثر مطالعات در خصوص خوشه بندی به سمت روش های خوشه بندی ترکیبی^۳ هدایت شده است. همان گونه که در بالا اشاره شد، تشخیص اجتماعات و خوشه بندی شباهت های بسیار زیادی با هم دارند و منابع زیادی این دو را یکسان می دانند [۱][۲][۳][۴].

گراف G گویند، این ماتریس یک ماتریس مربعی به ابعاد تعداد گره های 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 می گردد.

۱-۱- تشخیص اجتماعات

در شبکه های اجتماعی برخی گره ها در مقایسه با کل گره های شبکه، ارتباط بیشتری با هم دارند که به آن ها اجتماع گفته می شود [۳]. هدف از تشخیص اجتماعات، جدا کردن گروه ها یا اجتماعاتی است که ارتباط بیشتری با هم دارند [۲]. در واقع تشخیص اجتماعات، تقسیم بندی های موجود در شبکه را نشان می دهد و اجتماعات یک گراف را از هم مجزا می کند. تشخیص اجتماعات به ما کمک می کند تا دید بهتری نسبت به ساختار شبکه پیدا کنیم [۴].

در [۴] تشخیص اجتماعات را بدین صورت تعریف می نماید: اگر A را ماتریس مجاورتی یک گراف در نظر بگیریم، درجه یک گره در گراف را با k_i نشان می دهند که برابر است با: $k_i = \sum_j A_{ij}$ می باشد. با در نظر گرفتن یک زیر گراف S به-طوری که $S \subseteq G$ باشد.

$$S \subseteq G \Rightarrow k_i(S) = k_i^{in}(S) + k_i^{out}(S) \quad (1)$$

در رابطه (۱) نیز درجه هر گره را برای اجتماعات داخلی و خارجی به صورت $k_i^{in}(S) = \sum_{j \in S} A_{ij}$ و $k_i^{out}(S) = \sum_{j \notin S} A_{ij}$ تعریف می کنیم.

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

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

با توجه به رابطه (۲)، $\sum_{i \in S} k_i^{in}(S)$ مجموع درجه های تمامی گره هایی که در اجتماع یا زیرگراف S وجود دارد می باشد و $\sum_{i \in S} k_i^{out}(S)$ مجموع درجه های تمامی گره هایی که در اجتماع یا زیرگراف S وجود ندارد می باشد. بنابراین مسئله اصلی در

روش ها و بدست آوردن یکسری اجتماعات که در مقایسه با الگوریتم های پایه در تشخیص اجتماعات دارای دقت بالاتر، نتایج مطمئن تر و پایداری بیشتری هستند. نتایج این بررسی ها می تواند در مسائل بسیاری از جمله تشخیص دقیق تر اجتماعات، بازاریابی، تبلیغات، درک شبکه و بهبود موتورهای جستجو مورد استفاده قرار گیرد. قسمت های موجود در این مقاله بدین شرح می باشد 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 می دهیم. استفاده از این روش دارای مزایای و معایبی نیز می باشد که در قسمت نتیجه گیری به آن اشاره شده است.

۲- تعاریف و مفاهیم

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

۲-۱- شبکه های اجتماعی

شبکه های اجتماعی ساختاری اجتماعی است و از افراد یا سازمان هایی تشکیل شده است که گره های شبکه را تشکیل می دهند. گره ها توسط یک یا چند نوع خاص از وابستگی به هم متصل هستند، برای مثال تبادلات مالی، دوستی ها، خویشاوندی، تجارت، لینک های وب، سرایت بیماری ها (اپیدمولوژی) یا مسیرهای هواپیمایی نمونه هایی از ارتباط هستند. اما ساختارهای حاصل از این شبکه ها اغلب بسیار گسترده و پیچیده هستند. تحلیل شبکه اجتماعی عبارت است از نگاشت و اندازه گیری روابط و همکاری ها در بین افراد، گروه ها، سازمان ها و هر موجودیتی که قابلیت پردازش اطلاعات و دانش را داشته باشد. برای نمایش و تحلیل شبکه های اجتماعی معمولاً از تئوری گراف^۵ استفاده می شود. مولفه های موجود در تئوری گراف^۶ و لبه^۷ است [۱].

یک شبکه N را می توان به صورت یک گراف به شکل $G = \langle V, E \rangle$ تعریف نمود، که در آن گراف G شامل V مجموعه گره های گراف و E مجموعه لبه ها است. بعد از بدست آوردن گراف فوق می توان از روی گراف، ماتریس مجاورتی متناظر با آن را بدست آورد. ماتریس A را ماتریس مجاورتی

ترکیبی استفاده شده است، تا جوابها از مجموع جوابها حاصل گردد. دقت جوابها در این روش بسیار بهتر و مطمئن تر و پایدارتر خواهد بود. چرا که در این روش تنها از یک الگوریتم برای بدست آوردن نتیجه بر روی داده های مختلف استفاده نموده ایم و از اجماع تمامی روشها با قطعیت بیشتری در مورد نتیجه صحبت می نماییم.

خوشه بندی ترکیبی درصدد ترکیب نتایج به منظور پیدا نمودن نتایج مستحکم تر و دقیق تر از نتایج بدست آمده قبلی دارد. با ترکیب نتایج می توان یک خوشه بندی یا اجتماعات دقیق تری داشته باشیم [۵].

۳- روش های موجود تشخیص اجتماعات

روش های مختلفی برای بدست آوردن اجتماعات در شبکه وجود دارد. بررسی تک تک گره ها و قرار دادن آنها در اجتماعات مختلف و در نهایت ارزیابی آن دارای هزینه زمانی و محاسباتی بالایی است و این رویکرد به طور عملی امکان پذیر نیست. برای غلبه بر این مشکل روش های مبتنی بر تجربه^۹ به کمک آمده اند. که از جمله این روشها می توان به روش های طیفی^{۱۰}، تقسیم کننده^{۱۱}، تجمعی^{۱۲}، روش های مبتنی بر بهینه سازی ماژولاریتی^{۱۳}، تکنیک های حریصانه^{۱۴}، گداختگی شبیه سازی شده^{۱۵}، الگوریتم های تکاملی^{۱۶} اشاره نمود. هر کدام از این روشها به نوعی در بهبود عملکرد تشخیص اجتماعات موثر است [۳][۴].

ماژولاریتی از پرکاربردترین معیارهای مورد استفاده در روش های مختلف است. این معیار کمیتی از گروه بندی که از کل گراف بدست آمده است، ارائه می کند و نقش بسزایی در تعیین صحت گروه بندی دارد. معمولاً برای ارزیابی هر روش تشخیص اجتماعات در شبکه، مقدار ماژولاریتی گروه بندی پیشنهادی آن روش را برای شبکه ها و گراف های مختلف محاسبه می کنند. هرچه ماژولاریتی به دست آمده بیشتر باشد، دقت روش مورد نظر بهتر بوده است. معیار ماژولاریتی مهمترین محک در ارزیابی روش های تشخیص اجتماعات در شبکه های اجتماعی است. ماژولاریتی به صورت فرمول (۱) زیر تعریف می شود [۶][۷]:

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 بدنبال برطرف کردن این نقیصه هستند. روش های تشخیص اجتماعات معمولاً با استفاده از گراف ایجاد شده از ارتباطات بین افراد در شبکه های اجتماعی ماتریس همبستگی^۸ 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 روشها با توجه به روش انتخابی و نیز نوع و حجم داده ورودی، اجتماعات متفاوتی را پیدا می نمایند، که در بعضی شرایط یک الگوریتم جواب خوب و در شرایط دیگر جواب نامناسبی ایجاد می نماید، بنابراین هر روشی برای مجموعه داده های خاص مناسب و مورد استفاده قرار می گیرد.

۱-۲- خوشه بندی ترکیبی

خوشه بندی 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 خوشه بندی عادی بدین صورت عمل می نمایند که با تعریف یک الگوریتم سعی در برطرف نمودن مشکل خوشه بندی دارند. در صورتی که اعمال هر یک از این الگوریتم بر روی داده های یکسان دارای نتایج گوناگونی خواهد بود. با توجه به پیچیدگی مسئله تشخیص اجتماعات و ضعف روش های پایه ای آن، امروزه اکثر مطالعات به سمت روش های خوشه بندی ترکیبی هدایت شده است.

بعد از اعمال الگوریتم های مختلف بر روی شبکه مورد تحلیل همان طوری که در «شکل ۱» نیز نشان داده می شود، یک سری نتایج متفاوت بدست خواهد آمد که به صورت $\Pi^* = \{\pi_1, \pi_2, \dots, \pi_m\}$ تعریف می گردد. با ترکیب صحیح روش های مختلف با استفاده از خوشه بندی ترکیبی می توان به یک نتیجه بسیار بهتر، مطمئن تر، پایدارتر و دقیق رسید. در این مقاله برای برطرف کردن این نقیصه از خوشه بندی

نمودن لبه ای که بیشترین مرکزیت یا ارتباط با دیگر گره ها را دارد و حذف بازگشتی آن لبه در نهایت اجتماعات جدا شده بدست می آید. روش های دیگری نیز مانند ضریب خوشه بندی لبه نیز وجود دارد که از این ایده استفاده نموده اند ولی در رابطه مربوطه لبه هایی که اجتماعات را به یکدیگر مرتبط می کنند ضریب کوچکی دارند که در هر مرحله حذف می شوند [۱۰].

روش های دیگر که از جمله روش های طیفی است با ایجاد یک برش نرمال بر روی گراف به تقسیم نمودن گراف می پردازد. این معیار بر این اساس است که یک گروه بندی خوب تعداد لبه های بین اجتماعات را کمینه می کند، در عین حال لبه های داخل اجتماع را نگه می دارد [۱۱]. این معیار به صورت زیر تعریف می شود:

گراف $G(V, E)$ را در نظر بگیرید که در آن V مجموعه گره های گراف و E مجموعه تمام لبه های گراف است. گره های گراف را به دو مجموعه مجزای B, A تقسیم می کنیم به طوریکه داشته باشیم، $B = V - A$ باشد. مقدار این رابطه کسری از اتصالات بین B, A با توجه به اتصالات جداگانه B, A می باشد.

مقدار cut بین B, A به صورت رابطه (۵) 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 v و i است. رابطه (۶) برای نرمال کردن اندازه خوشه ها (اجتماعات) بکار می رود. این روش را معمولاً برای گراف های وزن دار استفاده می کنند که البته قابل تعمیم به گراف های 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 می کنیم.

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

همان طور که در رابطه (۷) می بینیم این رابطه فقط شامل دو اجتماع می باشد، 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

$$Q = \sum_i (e_{ii} - a_i^2) \quad (1)$$

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

در رابطه فوق i و j اندیس های اجتماع است و e_{ij} نسبت تعداد لبه هایی که گره های داخل اجتماع i را به هم متصل می کند به کل لبه های گراف است. به عبارت دیگر کسری از تعداد لبه ها که دو سر آن در اجتماع i قرار دارد به کل لبه های گراف و a_i نسبت تعداد لبه هایی که حداقل یک گره آن در اجتماع i باشد به کل لبه های گراف است. در واقع رابطه فوق سهم هر اجتماع i را در کل شبکه بیان می کند. هرچه این عدد بزرگتر باشد، این اجتماع سهم بیشتری در شبکه دارد و در واقع اجتماع قوی تر 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 گره ها بصورت تصادفی در بین اجتماعات قرار گرفته باشند $Q=0$ خواهد بود. هرچه مقدار Q به یک نزدیک تر باشد اجتماعات بهتر جدا شده اند ولی هیچگاه یک نمی شود. در عمل $Q > 0.3$ ساختار گروهی مناسبی را نشان می دهد. ماژولاریتی می تواند مقادیر منفی را نیز بپذیرد [۸]. مقدار ماژولاریتی ممکن است برای یک شبکه در بهترین حالت زه بندی از عدد خاص بیشتر نشود. این بدان معنا نیست تشخیص اجتماع به درستی انجام نشده است، بلکه ممکن است بیشترین مقداری باشد که ماژولاریتی برای آن شبکه می تواند داشته باشد.

یکی از روش های تشخیص اجتماعات روش تقسیم کننده می باشد. در شبکه ها، هر چه تعداد مسیرهایی که از گره یا لبه خاصی عبور می کنند بیشتر باشد آن گره یا لبه مهم تر است. بنابراین با فرض اینکه کوتاه ترین مسیر بین دو گره محاسبه پذیر باشد، اهمیت یک گره یا لبه را می توان اندازه گیری نمود، که به آن مرکزیت مابینی گفته می شود. یکی از روش های بدست آوردن اجتماعات که [۹] ارائه شده است مرکزیت مابینی می باشد. رابطه مرکزیت مابینی به صورت رابطه (۴) تعریف می شود:

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

که در آن $\sigma(i, u, j)$ تعداد کوتاه ترین مسیر ها بین دو اجتماع i و j است که از گره یا لبه u عبور می کند. $\sigma(i, j)$ تعداد کل مسیر ها بین دو گره i و j است. در این روش با پیدا

۴- روش پیشنهادی

ماژولاریتی دست پیدا نماییم.

با توجه به اینکه اکثر روش های تشخیص اجتماعات پایه روی جنبه های خاصی از داده ها تمرکز می کنند، در نتیجه روی مجموعه داده های خاصی کارآمد می باشند. به همین دلیل، نیازمند روش هایی هستیم که بتواند با استفاده از ترکیب این الگوریتم ها و گرفتن نقاط قوت هر یک، نتایج بهینه تری را تولید نماید. در واقع هدف اصلی تشخیص ترکیبی اجتماعات جستجوی نتایج بهتر، مطمئن تر و مستحکم تر، با استفاده از ترکیب نتایج حاصل از چندین الگوریتم پایه می باشد [۵].

بنابراین تشخیص ترکیبی اجتماعات با ترکیب نتایج بدست آمده از الگوریتم های پایه درصدد پیدا نمودن نتایج دقیق تر، مطمئن تر و پایدارتر از نتایج بدست آمده قبلی می باشد. همان طوری که در «شکل (۱)» مشاهده می شود، 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 همبستگی ایجاد می گردد که در صورت وجود ارتباط بین گره ها یک و در غیر این صورت صفر منظور می گردد. با اعمال الگوریتم های متفاوت بر روی این ماتریس همبستگی و گرفتن نتایج متفاوت و در بعضی موارد یکسان می توان از روش ترکیبی تشخیص استفاده نمود. بدین صورت که اگر گره ای 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 دیگری ایجاد خواهد شد که یک گراف وزن دار از نتایج ما می باشد، از مجموع نتایج کمک می گیریم تا به یک نتیجه بهتر، دقیق تر، پایدارتر برسیم.

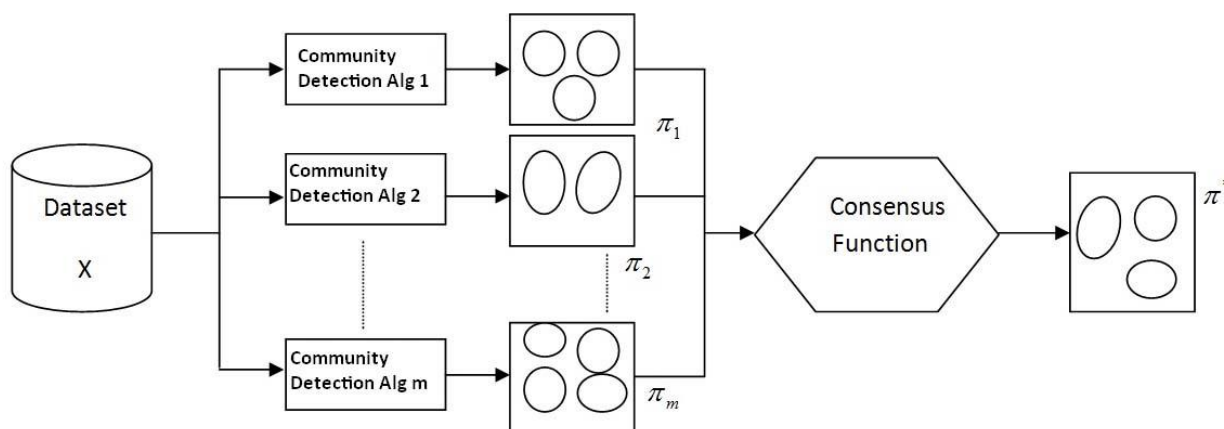
ماتریس همبستگی ایجاد شده برآیندی از تمامی روش های اولیه خواهد بود، در این ماتریس برای هر یال در گراف مورد نظر عددی منظور می شود که این عدد مقدار وابستگی آن یال به گره ای که با آن در تماس می باشد را نشان می دهد. به عبارت دیگر اجماع روش های مختلف در مورد اینکه آیا این دو نود با هم در ارتباط هستند یا خیر را نشان خواهد داد. یال های که ارتباط ضعیف دارند، در الگوریتم تشخیص ترکیبی اجتماعات سیاست

یکی از مهم ترین روش ها، روش تجمعی می باشد. این روش بر این حقیقت بنا شده که گره های یک اجتماع ویژگی های مشترکی دارند و می توان از این ویژگی های مشترک برای گروه بندی استفاده کرد. در برابر روش های تقسیم کننده، روش تجمعی در ابتدا همه گره ها را جدا از هم و غیر متصل در نظر می گیرد و آن ها را بر اساس ویژگی های مشترک به هم متصل می کند تا به اجتماعات برسد. در واقع نحوه عملکرد این رویکرد بدین صورت است که لبه های بین اجتماعات خود به خود حذف شده و تنها لبه های داخل اجتماع باقی می ماندند [۱۲].

از جمله روش های دیگر روش بیشینه سازی ماژولاریتی می باشد، جستجو برای یافتن ماژولاریتی بهینه (یا بیشینه) از نوع مسائل بسیار سخت می باشد. چرا که فضای تقسیم بندی هایی که از گراف امکان پذیر است با بزرگ شدن اندازه گراف به سرعت در حال افزایش است. به همین دلیل رویکردهای جستجوی مکاشفه ای برای محدود کردن فضای جستجو الزامی است. در [۱۴][۱۳] نیومن روشی بر پایه ادغام کردن اجتماعات ارائه کرده است بطوری که ماژولاریتی بیشینه شود.

از جمله روش های دیگر در این زمینه روش های مبتنی بر الگوریتم های تکاملی هستند. روش های مختلفی در حوزه الگوریتم های تکاملی بر روی تشخیص اجتماعات انجام شده است. تارگین و بینگل [۱۵] روشی ارائه کردند که از ماژولاریتی برای سنجش استفاده می کند. در این روش هر کروموزوم کل گره های موجود در شبکه را شامل می شود. در ابتدا به هر گره به صورت تصادفی یک اجتماع نسبت داده می شود و سپس crossover که تغییراتی روی آن صورت گرفته، انجام می شود. گاهی نیز mutation گره های دو اجتماع را جابجا می کند اعمال می شود و در نهایت هر کروموزوم توسط معیار ماژولاریتی سنجیده می شود.

روش های دیگری نیز در خصوص تشخیص اجتماعات وجود دارد که ما در اینجا به این چند روش اشاره کردیم که در این مقاله نیز تا حدودی استفاده شده است. در [۱۶] به الگوریتم های دیگری برای تشخیص اجتماعات اشاره شده است.



شکل (۱): الگوریتم پیشنهادی برای تشخیص ترکیبی اجتماعات

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 بر روی گراف‌های جهت‌دار و وزن‌دار می‌باشد.

الگوریتم اول یک روش بر مبنای مرکزیت مابینی می‌باشد، که در روش‌های موجود به آن اشاره شد. در این روش یال‌هایی که بیشترین مرکزیت را داشته باشند، به صورت بازگشتی حذف نموده و به این ترتیب یکسری اجتماعات جدا از هم را خواهیم داشت [۹].

الگوریتم دوم یک روش مشابه مرکزیت مابینی و روش ضریب خوشه‌بندی لبه می‌باشد با این تفاوت که به جای بررسی تک‌تک گره‌ها در هر مرحله و بدست آوردن یک یال برای حذف شدن در هر مرحله یال‌ها با ضریب بالا را حذف نموده و در نهایت یک دندروگرام^{۱۸} از کل گراف ایجاد می‌شود که یک نمایش سلسه مراتبی می‌باشد، با برش این دندروگرام می‌توان اجتماعات را تفکیک نمود ولی در برش می‌بایست دقت نمود که بهینه‌ترین حالت را داشته باشد [۱۰].

الگوریتم سوم یک روش بر مبنای بهینه‌سازی ماژولاریتی و حریصانه می‌باشد، در این روش با تعریف رابطه ماژولاریتی که در بالا به آن اشاره شد، درصد پیدا نمودن اجتماعاتی با بیشینه‌ترین حالت ماژولاریتی هستند. در این روش ماژولاریتی معیار مشابهت ما خواهد بود [۱۲].

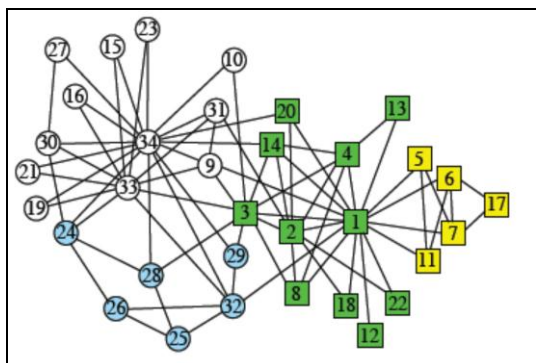
الگوریتم چهارم یک روش نیز بر مبنای بهینه‌سازی ماژولاریتی می‌باشد که با استفاده از ماژولاریتی و تعریف متغیرهای دیگری که در آن مقادیر ماژولاریتی و نیز حالات گره‌ها در آن وضعیت

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 وزن‌دار، تشخیص اجتماعات کامل‌تر و دقیق‌تری را خواهیم داشت. از نظر هزینه اجرایی، هزینه اجرایی آن نیز 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 خواهند بود.

۴-۱- ارزیابی روش پیشنهادی

در این مقاله چهار روش مختلف تشخیص اجتماعات با استفاده از زبان Matlab پیاده‌سازی شده است، که درصد آن هستیم که با ترکیب این چهار روش با استفاده از خوشه‌بندی ترکیبی به یک خوشه‌بندی یا اجتماعات دقیق‌تری نسبت به تک‌تک نتایج برسیم. 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 تک-تک 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



شکل (۲): گراف باشگاه کاراته [۱۹]

مجموعه داده‌های موسیقی دانان جاز: در این پایگاه داده ۱۹۶ باند موسیقی وجود دارد که بین سال‌های ۱۹۱۲ و ۱۹۴۰ فعالیت می‌کردند که بیشتر آن‌ها در سال‌های اطراف ۱۹۲۰ مشغول بوده‌اند. این پایگاه داده موسیقی‌دان‌هایی را لیست می‌کند که در این باندها می‌نواختند. این گراف شامل ۱۹۶ گره و ۲۷۴۲ یال است، اگر دو موسیقی‌دان در یک باند می‌نواختند لبه‌ها بین آن‌ها ترسیم می‌شود [۲۰].

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 یال است این ایمیل‌ها بخش ضمیمه شده (مثل عکس، فیلم یا هر فایل دیگری) ندارد [۲۱].

۳-۴- نتایج ارزیابی

«جدول (۱)» نتایج اعمال الگوریتم‌های مختلف را نشان می‌دهد. این جدول سه نوع داده رایج که برای ارزیابی روش‌های تشخیص اجتماعات مورد استفاده قرار می‌گیرند را نشان می‌دهد. تعداد گره‌ها این داده‌های استاندارد در جلوی آن‌ها مشخص گردیده است. یک الگوریتم خوب می‌بایست در تعداد کم و زیاد گره‌ها خوب و بهینه جواب دهد چرا که در شبکه‌های اجتماعی تعداد گره‌ها بسیار متنوع می‌باشند و در بعضی موارد تعداد آن به کل کاربران در شبکه جهانی تامین پیدا می‌کند. روش ارزیابی این الگوریتم‌ها نیز مازولاریتی می‌باشد که در روش‌های موجود به آن

ذخیره می‌شود. این روش باعث افزایش سرعت در پیدا نمودن اجتماعات و نیز با داشتن یک پیش‌زمینه از مراحل قبلی با دقت بالاتری به ترکیب گره‌ها می‌پردازد [۱۸].

حال با ترکیب این چهار روش با استفاده از خوشه‌بندی ترکیبی که در این مقاله ما از یکی از روش‌های خوشه‌بندی ترکیبی که روش مبتنی بر ماتریس همبستگی می‌باشد استفاده کرده‌ایم و به ترکیب نتایج پرداخته‌ایم، تا بتوانیم در تشخیص اجتماعات نتایج دقیق‌تری داشته باشیم. البته استفاده از روش‌های دیگر خوشه‌بندی ترکیبی نیز امکان‌پذیر می‌باشد ولی با توجه به اینکه تعداد الگوریتم‌های اولیه ما کم می‌باشد این روش بهتر جواب می‌دهد.

۲-۴- داده‌های مورد استفاده

داده‌های مورد استفاده داده‌های استاندارد می‌باشند که تقریباً در تمامی روش‌های تشخیص اجتماعات از آن‌ها استفاده می‌شود. از جمله این داده‌ها می‌توان به باشگاه کاراته زاکاری [۱۹]، مجموعه داده‌های موسیقی دانان جاز [۲۰] و مجموعه داده‌های ایمیل Enron توسط FERC [۲۱] اشاره نمود.

باشگاه کاراته زاکاری: همانطوری که در «شکل (۲)» 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 یال 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 مربی می‌باشد، در بعضی نقاط درگیری بین مسئول باشگاه با مربی موجب 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 (۲)» می‌توان دو تجمع را تشخیص داد. یکی اطراف گره ۳۳ و ۳۴ و دیگری اطراف گره ۱ قرار دارد [۱۹]. 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.

جدول ۱: مقایسه روش های مختلف تشخیص اجتماعات بر اساس معیار ماژولاریتی

Our Method	Reichardt [۱۸]	Newman Greedy [۱۲]	Radicchi et al [۱۰]	Girvan & Newman [۹]	تعداد یال ها	تعداد گره ها	Data Set
۰,۴۱۵	۰,۴۰۲	۰,۴۱۷	۰,۳۹۲	۰,۳۸۱	۷۸	۳۴	کاراته
۰,۴۴۳	۰,۴۴۵	۰,۴۴۲	۰,۴۰۵	۰,۴۳۹	۲۷۴۲	۱۹۶	جاز
۰,۵۷۲	۰,۵۷۱	۰,۵۷۴	۰,۵۳۲	۰,۴۹۴	۵۴۵۱	۱۱۳۳	ایمیل

ماتریس همبستگی، مورد استفاده قرار گرفته است. البته امکان استفاده از روش های دیگر نیز وجود دارد که به علت کمبود تعداد حالات جواب خوبی را نمی داند، برای استفاده از آن روش ها به تعداد بیشتری نتایج گوناگون نیاز می باشد که می توان در کارهای بعدی به آن پرداخت.

همان گونه که مشاهده گردید از مزایای این روش استفاده از روش های قبلی به تعداد زیاد که نشان از مقیاس پذیری بالای این روش و نیز دقت بالای آن به نسبت روش های دیگر و در مجموع نتایج دقیق تر و پایدارتری را برای ما خواهد داشت. از معایب این روش به هزینه اجرایی آن که با بدترین حالت هزینه اجرایی در الگوریتم های استفاده شده برابر می باشد و نیز حافظه زیادی را نیز مصرف می نماید، همچنین ممکن است در همه جا بهینه ترین جواب را نداشته باشیم و الگوریتم هایی باشند که برای داده های خاصی جواب دقیق تری را داشته باشند و نیز در اعمال این روش می بایست تعداد خوشه ها نیز برای ما مشخص باشد ولی می توان از روش های دیگر خوشه بندی ترکیبی استفاده نمود که تعداد خوشه های در آن نامشخص باشد.

امکان استفاده از روش های دیگر خوشه بندی ترکیبی نیز وجود دارد، که با پیاده سازی های بیشتری از الگوریتم های تشخیص اجتماعات می توان از آن ها نیز استفاده نمود. همان گونه که گفته شد می توان از تشخیص اجتماعات در مسائل بسیاری از جمله تشخیص دقیق تر اجتماعات، بازاریابی، تبلیغات، درک شبکه، بهبود موتورهای جستجو، ارتباطات (از قبیل تلفنی، سیار و...) و تعاملات افراد و انجام فعالیت های امنیتی و جاسوسی مورد استفاده قرار گیرد.

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.

همان طوری که در «جدول (۱)» مشاهده می شود معیار ارزیابی برای روش پیشنهادی ماژولاریتی می باشد که مقدار آن بین (-1, +1) می باشد. اگر مقدار ماژولاریتی صفر شود یعنی تمامی گره ها را در یک اجتماع قرار داده ایم و اگر منفی شود به معنی خطا 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 با این حال که نوع داده در حال تغییر است ولی به جواب بهینه نزدیک می باشد. اگر این روش را بروی 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.

۵- نتیجه

هدف این مقاله بهبود نتایج روش های تشخیص اجتماعات می باشد. رویکرد مورد استفاده رویکرد ترکیبی بوده است، با توجه به قدرت روش های خوشه بندی ترکیبی و نتایج دقیق آن و امکان استفاده آن در روش های تشخیص اجتماعات در این مقاله از آن برای بدست آمده نتایج دقیق تر، مطمئن تر، پایدارتر و مستحکم تر استفاده شده است.

با پیاده سازی چهار روش متفاوت از روش های تشخیص اجتماعات و بدست آوردن نتایج آن ها و نیز در نهایت با استفاده از روش های خوشه بندی ترکیبی که رویکرد اصلی این مقاله بوده است، به ترکیب نتایج بدست آمده از آن ها هستیم. برای ترکیب نتایج، یکی از روش های خوشه بندی ترکیبی با نام روش مبتنی بر

مراجع

- [۱۴] A.Clauset, M.E.J.Newman and C.Moore, "Finding community structure in very large networks," *Physical Review E*, vol. ۷۰, no. ۶, p. ۰۶۶۱۱۱, ۲۰۰۴.
- [۱۵] M.Tasgin and H.Bingol, "Community detection in complex networks using genetic algorithm," *Arxiv preprint cond-mat/۰۶۰۴۴۱۹*, ۲۰۰۶.
- [۱۶] J.Leskevicius, K.J.Lang and M.Mahoney, "Empirical comparison of algorithms for network community detection," *In Proceedings of the ۱۹th international conference on World Wide Web*, pp. ۶۳۱-۶۴۰, ۲۰۱۰.
- [۱۷] N.Iam-on S.Garrett, "LinkCluE: A MATLAB Package for Link-Based Cluster Ensembles," *In Journal of Statistical Software*, Issue ۹, Volume ۳۶, August ۲۰۱۱.
- [۱۸] J.Reichardt, S.Bornholdt, "Statistical mechanics of community detection", *Physical Review E*, ۲۰۰۶.
- [۱۹] W.W.Zachary, "An information flow model for conflict and fission in small groups," *Journal of Anthropological Research*, vol. ۳۳, no. ۴, pp. ۴۵۲-۴۷۳, ۱۹۷۷.
- [۲۰] P.Gleiser and L.Danon, "Community structure in jazz," *Arxiv preprint condmat/۰۳۰۷۴۳۴*, ۲۰۰۳.
- [۲۱] J.Shetty and J.Adibi, "The Enron email dataset database schema and brief statistical report," *Information Sciences Institute Technical Report, University of Southern California*, ۲۰۰۴.
- [۱] D.Easley and J.Kleinberg, "Networks, Crowds, and Markets: Reasoning about a Highly Connected World," *Cambridge University Press*, June ۱۰, ۲۰۱۱.
- [۲] B.Liu, "Web DataMining Exploring Hyperlinks, Contents, and Usage Data," *Springer*, ۲۰۰۷.
- [۳] S.Fortunato, "Community detection in graphs," *Physics Reports*, vol. ۴۸۶, no. ۳-۵, pp. ۷۵-۱۷۴, ۲۰۱۰.
- [۴] M.A.Porter, J.P.Onnela, and P.J.Mucha, "Communities in networks," *Notices of the American Mathematical Society*, vol. ۵۶, no. ۹, pp. ۱۰۸۲-۱۰۹۷, ۲۰۰۹.
- [۵] A.Strehl and J.Ghosh, "Cluster Ensembles -A Knowledge Reuse Framework for Combining Multiple Partitions," *Journal of Machine Learning Research*, pp. ۵۸۳-۶۱۷, ۲۰۰۳.
- [۶] A.Lancichinetti, "Community detection algorithms: a comparative analysis," *Physical Review E*, vol. ۸۰, no. ۵, p. ۰۵۶۱۱۷, ۲۰۰۹.
- [۷] X.S.Zhang and et al., "Modularity optimization in community detection of complex networks," *EPL (Euro Physics Letters)*, vol. ۸۷, p. ۳۸۰۰۲, ۲۰۰۹.
- [۸] M.Girvan and M.E.J.Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. ۹۹, no. ۱۲, p. ۷۸۲۱, ۲۰۰۲.
- [۹] M.E.J.Newman, "A measure of betweenness centrality based on random walks," *Social networks*, vol. ۲۷, no. ۱, pp. ۳۹-۵۴, ۲۰۰۵.
- [۱۰] F.Radicchi, C.Castellano, F.Cecconi, V.Loreto, and D.Parisi, "Defining and identifying communities in networks," *Proceedings of the National Academy of Sciences of the United States of America*, pp. ۲۶۵۸-۲۶۶۳, Mar ۲, ۲۰۰۴.
- [۱۱] M.Lipczak and E.Milios, "Agglomerative genetic algorithm for clustering in social networks," *In Proceedings of the ۱۱th Annual conference on Genetic and evolutionary computation*, pp. ۱۲۴۳-۱۲۵۰, ۲۰۰۹.
- [۱۲] M.E.J.Newman, "Fast algorithm for detecting community structure in very large networks," *Physical review E*, vol. ۶۹, ۲۰۰۴.
- [۱۳] M.E.J.Newman and M.Girvan, "Finding and evaluating community structure in networks," *Physical review E*, vol. ۶۹, no. ۲, p. ۰۲۶۱۱۳, ۲۰۰۴.

زیر نویس

Community Detection	^۱
Clustering	^۲
Ensemble Clustering	^۳
Social Network Analysis(SNA)	^۴
Graph theory	^۵
Node	^۶
Edge	^۷
Co-association Matrix	^۸
Heuristic	^۹
Spectral	^{۱۰}
Divisive	^{۱۱}
Agglomerative	^{۱۲}
Modularity optimization	^{۱۳}
Greedy techniques	^{۱۴}
Simulated annealing	^{۱۵}



یازدهمین کنفرانس سیستم های هوشمند ایران
۹ و ۱۰ اسفند ۱۳۹۱
(ICIS^{۲۰۱۳})
۱1th Iranian Conference on Intelligent Systems
February ۲۶th & ۲۸th, ۲۰۱۳



Evolutionary ^{۱۶}
Targin and Bingol ^{۱۷}
Dendrogram ^{۱۸}
Zachary karate club ^{۱۹}
Jazz bands ^{۲۰}
Federal Energy Regulatory Commission ^{۲۱}