Journal of Applied And Basic Machine Intelligence Research (ABMIR)

پژوهش‌های نظری و کاربردی هوش ماشینی

سال اول، شماره اول، بهار و تابستان ۱۴۰۲، صفحات ۱۰۱-۸۸

روشی نوین برای خوشه‌بندی نیمه نظارتی شبکه‌های پیچیده مبتنی بر معیار پیمانگی

محمد قدیریان [۱]، نوشین بیگدلی [۲*]

[۱] گروه مهندسی برق کنترل- دانشکده فنی و مهندسی- دانشگاه بین‌المللی امام خمینی (ره)- قزوین- ایران

[۲] گروه مهندسی برق کنترل- دانشکده فنی و مهندسی- دانشگاه بین‌المللی امام خمینی (ره)- قزوین- ایران

چکیده

مقاله پژوهشی

خوشه‌بندی، ابزاری پرکاربرد جهت تحلیل اطلاعات شبکه‌های پیچیده است که برای مدل‌سازی سامانه‌های پیچیده بکار می‌رود. پیمانگی، معیاری پایه و فراگیر جهت ارزیابی و صحت‌سنجی خوشه‌بندی شبکه‌ها است که دارای چالش‌هایی چون ان‌پی-سخت بودن مسئله و عدم امکان استفاده از دانش اولیه در خوشه‌بندی است. لذا، خوشه‌بندی مبتنی بر معیار پیمانگی، قابلیت تعمیم به خوشه‌بندی‌های نیمه نظارتی را ندارد. از طرفی، یکی از روش‌های خوشه‌بندی نیمه نظارتی، روش خوشه‌بندی مبتنی بر تجزیه نامنفی ماتریسی (NMF) است. اما این روش، ویژگی‌های خاص شبکه‌ها را در نظر نمی‌گیرد. در این مقاله، برای غلبه بر چالش‌های نام‌برده و با ارائهٔ اثباتی جدید، برای خوشه‌بندی مبتنی بر معیار پیمانگی، ساختاری مشابه با خوشه‌بندی مبتنی بر تجزیه نامنفی ماتریسی نامتقارن ارائه می‌شود که در آن، امکان بهره‌گیری از دانش اولیه و حل به روش تکراری میسر می‌گردد. سپس، روش خوشه‌بندی نیمه نظارتی نوینی به نام تجزیه نیمه نظارتی نامنفی ماتریس‌های متقارن مبتنی بر معیار پیمانگی (SSNMF-Q) با بهره‌گیری از مزیت دانش اولیه و روش حل تکراری، به‌جای حل مسئله ان‌پی-سخت ارائه می‌گردد. برای ارزیابی روش پیشنهادی، از پنج مجموعه داده واقعی استفاده‌شده که نتایج، بیانگر عملکرد بهتر SSNMF-Q در مقایسه با سایر خوشه‌بندی‌های نیمه نظارتی مبتنی بر NMF است.

تاریخ دریافت:
۱۴۰۱/۰۸/۲۴

تاریخ پذیرش:
۱۴۰۱/۱۰/۱۹

کلیدواژه‌ها:
تجزیه نامنفی ماتریسی، خوشه‌بندی گراف، خوشه‌بندی نیمه نظارتی، معیار پیمانگی

نویسنده مسئول:
n.bigdeli@eng.ikiu.ac.ir

doi : 10.22034/ABMIR.2023.19231.1018

۱- مقدمه

در دنیای کنونی سامانه‌های پیچیده توسط شبکه‌هایی همانند شبکه‌های اجتماعی، شبکه‌های ترافیکی و شبکه‌های بیولوژیکی مدل‌سازی می‌شوند [۱-۳]. با استفاده از ابزار تئوری گراف، می‌توان شبکه‌ها را توسط مجموعه‌ای از گره‌ها و یال‌ها نمایش داد. به‌طور مثال، شبکه‌های اجتماعی یکی از سامانه‌های پیچیده در دنیای کنونی هستند که برای هر ویژگی این شبکه گراف‌هایی ترسیم می‌شود. گراف ارتباطی اشخاص از طریق دنبال کننده و دنبال شونده، گرافی جهت‌دار برای یک مجموعه انسانی را تشکیل خواهد داد که هر حساب شبکه اجتماعی به‌عنوان یک گره و ارتباط دنبال کننده و دنبال شونده میان این اعضا به‌عنوان یال‌های جهت‌دار این گراف در نظر گرفته خواهند شد. با تحلیل گراف‌ها می‌توان اطلاعات بسیار مفیدی از ساختار این شبکه‌های پیچیده به دست آورد. در این زمینه، یکی از ابزارهای متداول و مؤثر جهت بررسی ساختار شبکه‌های پیچیده، خوشه‌بندی گراف‌ها است [۲]. در مثال گراف ارتباطی شبکه اجتماعی، این تحلیل به ما نشان می‌دهد که کدام حساب‌ها به دلیل نوع ارتباط در شبکه به همدیگر نزدیک‌تر هستند و در هر خوشه قرار خواهند گرفت. ازاین‌رو، با توجه به [۳]در سال‌های گذشته، روش‌های متنوعی چون خوشه‌بندی مبتنی بر پیمانگی [٤، ٥]، خوشه‌بندی مبتنی بر تجزیه نامنفی ماتریسی [٦]، خوشه‌بندی مبتنی بر انتشار برچسب [۷، ۸]، خوشه‌بندی مبتنی بر پیاده‌روی تصادفی [۹، ۱۰] و خوشه‌بندی بر اساس بهینه‌سازی تکاملی [۱۱] جهت خوشه‌بندی گراف‌ها ارائه‌شده است.

مطابق با [٤]، معیار پیمانگی مشخصاً به‌عنوان معیار صحت‌سنجی خوشه‌بندی گراف‌ها معرفی‌شده است. این معیار، چگالی ارتباط میان گره‌ها در داخل هر گروه و ارتباط میان گروهی را به‌عنوان شاخصی جهت بررسی صحت خوشه‌بندی‌ها در نظر می‌گیرد. ازاین‌رو پیمانگی، معیاری محبوب و متداول جهت خوشه‌بندی شبکه‌های گرافی است. به‌عبارت‌دیگر، این معیار مختص خوشه‌بندی شبکه‌های گرافی است. ازاین‌رو در این سال‌ها به دلیل فراگیر شدن این معیار، روش‌های مبتنی بر معیار پیمانگی متنوعی جهت خوشه‌بندی گراف‌های پیچیده ایجادشده است. ازجمله این روش‌ها می‌توان به روش‌های چون روش حریصانه[1] [۱۲]، روش حل بهینه‌سازی تکاملی [۱۳] و روش طیفی[2] [٤] اشاره نمود. اگرچه خوشه‌بندی مبتنی بر معیار پیمانگی روشی فراگیر جهت خوشه‌بندی است اما محدودیت‌هایی چون وابسته بودن به مجموع یال‌ها [١٤] و عدم استفاده از دانش اولیه خوشه‌بندی رادارند. همچنین مطابق با [٤]روش حل خوشه‌بندی مبتنی بر پیمانگی مسئله ان‌پی-سخت[3] است.

خوشه‌بندی نیمه نظارتی، به روش‌های خوشه‌بندی اطلاق می‌شود که در آن‌ها، دانش اولیه از اعضای هر گروه و یا تعداد آن‌ها در فرآیند تشخیص گروه‌ها دخالت داده‌شده و از این دانش اولیه در خوشه‌بندی استفاده می‌شود. روش‌های نیمه نظارتی مبتنی بر تجزیه نامنفی ماتریس، ازجمله روش‌های متداول خوشه‌بندی در راستای بهره‌گیری از دانش اولیه اعضای هر گروه هستند. در روش‌های خوشه‌بندی، ماتریس‌های مشابهت گراف (مانند ماتریس مجاورتی[4]) به چندین ماتریس کاهش بعد داده‌شده تبدیل می‌شوند که با استفاده از آن‌ها می‌توان خوشه‌بندی و خوشه‌های مدنظر را به دست آورد [۱۵، ٦]. در این روش‌ها برای استخراج گروه‌ها، از مدل‌ها و توابع متفاوتی همچون مدل تجزیه نامنفی ماتریسی متقارن [١٦]، مدل مقاوم تجزیه نامنفی ماتریس[۱۸،۱۷]، مدل تجزیه نامنفی ماتریسی سه عاملی]۱۹]، مدل تجزیه نامنفی ماتریسی عمیق[۲۰-۲۲] و مدل گرافی تجزیه نامنفی ماتریسی منظم شده[۲٤،۲۳] استفاده‌شده است. مطابق با[۱۷، ۱۸]مدل مقاوم تجزیه نامنفی ماتریسی توانسته است با استفاده از نرم l_2 عدم قطعیت‌هایی همچون واردکردن دانش‌های اولیه اشتباهی و خطاهای ترکیب انواع نوع داده را کم‌رنگ‌تر کنند. در مدل تجزیه نامنفی ماتریسی سه عاملی[۱۹]، معیار پیمانگی با مدل سه عاملی، ترکیبی جدید تشکیل داده است که منتج به بهبود عملکرد در خوشه‌بندی نهایی شده است. در مدل‌های تجزیه نامنفی ماتریسی عمیق[۲۰-۲۲]، خطای حل به روش تکراری به‌شدت کاهش پیداکرده است، همچنین به کمک معیار پیمانگی، بهبود قابل‌توجهی در مدل‌سازی‌ها قابل‌مشاهده است.

[1] Greedy algorithm
[2] Spectral algorithm
[3] NP hard
[4] Adjacency matrix

نکته قابل‌توجه آن است که همهٔ مدل‌های خوشه‌بندی مبتنی بر تجزیه نامنفی ماتریسی از جنس حل مسئلهٔ بهینه‌سازی می‌باشند؛ لذا می‌توان دانش اولیه را در ماتریس‌های مشابهت گراف ترکیب کرد [۲۵، ۲۶] و یا آن را به‌صورت عبارت جداگانه‌ای به مسئله بهینه‌سازی اضافه نمود [۲۷–۲۹]. لازم به ذکر است که تمامی این روش‌ها عموماً برای پیدا کردن خوشه‌بندی بهتر، از ماتریس مجاورتی استفاده می‌کنند.

مزیت اصلی خوشه‌بندی مبتنی بر تجزیه نامنفی ماتریسی، کاربرد آن برای انواع داده‌ها ازجمله داده‌های صوتی، تصویری و متنی و گرافی است. تاکنون از این خوشه‌بندی برای مسائلی چون جداسازی منابع صوتی [۳۰]، اطلاعات چندوجهی[۳۱]، پردازش تصویر[۳۲]، استخراج کلمات کلیدی و عنوان اسناد[۳۳] استفاده‌شده است. هرچند که این عمومیت و گستردگیِ کاربرد برای انواع داده‌های از جنس متفاوت، یکی از مزایای مهم تجزیه نامنفی ماتریسی است، اما این روش قابلیت بررسی و استخراج ویژگی‌های نوع خاصی از داده را به‌صورت تخصصی ندارد. به عبارت دیگر نوع داده (داده صوتی، شبکه‌های گرافی، متنی، تصویر) در خوشه‌بندی مبتنی بر تجزیه نامنفی ماتریسی اثری نخواهد داشت. در راستای بهبود این محدودیت، به‌عنوان نمونه در[۱۹، ۲۲]معیار پیمانگی به روش خوشه‌بندی مبتنی بر تجزیه نامنفی ماتریسی به‌صورت خطی اضافه شده است تا کیفیت خوشه‌بندی را برای شبکه‌های گرافی افزایش دهد.

با توجه به مطالب فوق و مزیت روش خوشه‌بندی مبتنی بر معیار پیمانگی یعنی مختص خوشه‌بندی شبکه‌های گرافی بودن و محدودیت‌های این روش چون عدم استفاده از دانش اولیه و ان‌پی–سخت بودن مسئله، و همچنین مزایای خوشه‌بندی بر پایه تجزیه نامنفی ماتریسی همچون توانایی استفاده از دانش اولیه و حل تکراری و محدودیت عمومیت برای تمامی داده‌ها و عدم تأثیر نوع داده، ما را بر این داشت در این مقاله راه‌حل نوینی با بهره‌گیری از مزایای تجزیه نامنفی ماتریسی برای خوشه‌بندی مبتنی بر معیار پیمانگی ارائه می‌شود. بدین منظور روش خوشه‌بندی مبتنی بر معیار پیمانگی دارای ساختاری مشابه با خوشه‌بندی مبتنی بر تجزیه نامنفی ماتریسی است. درنتیجه، معیار پیمانگی می‌تواند جایگزین ماتریس مجاورتی در خوشه‌بندی مبتنی بر تجزیه نامنفی ماتریسی شود. این مزیت و نوآوری به ما کمک خواهد کرد تا از مزایای روش خوشه‌بندی بر پایه تجزیه نامنفی ماتریسی همچون روش حل تکراری و تعمیم به خوشه‌بندی نیمه نظارتی را برای ساختار جدید تعریف‌شدهٔ مبتنی بر معیار پیمانگی استفاده شود. ازاین‌رو در این مقاله مدل جدیدی برای خوشه‌بندی نیمه نظارتی به نام SSNMF-Q[1] پیشنهادشده است تا بتواند خوشه‌بندی مبتنی بر معیار پیمانگی را بهبود بخشد. لذا، نوآوری‌ها و مراحل رسیدن به این هدف به شرح زیر است.

۱. اثبات خواهد شد که خوشه‌بندی مبتنی بر پیمانگی دارای ساختاری مشابه با خوشه‌بندی مبتنی بر تجزیه نامنفی ماتریسی است. این مشابه‌سازی کمک خواهد کرد تا ترکیبی از دانش اولیهٔ خوشه‌بندی گراف‌ها در این خوشه‌بندی نوین ارائه شود.

۲. ساختاری مشابه‌سازی شده، به‌عنوان تابع اصلی روش خوشه‌بندی مبتنی بر تجزیه نامنفی ماتریسی انتخاب می‌شود. سپس بخش‌های موردنیاز از دانش اولیه به این تابع اضافه شده و مدل نهایی نیمه نظارتی SSNMF-Q تولید می‌شود.

۳. نشان داده می‌شود که مدل SSNMF-Q می‌تواند دانش اولیه را با روش خوشه‌بندی مبتنی بر معیار پیمانگی ترکیب کند و کیفیت خوشه‌بندی را با افزایش دانش اولیه بهبود بخشد.

٤. برای الگوریتم SSNMF-Q، روش حل تکراری طراحی‌شده و نتایج خوشه‌بندی بر روی پنج مجموعه داده واقعی نمایش داده می‌شود. درنهایت نیز بر اساس نتایج به‌دست‌آمده، تأثیر مدل ارائه‌شده بر روی کیفیت پاسخ بررسی می‌شود.

در ادامه ساختار این مقاله این‌گونه تنظیم‌شده است که در بخش ۲ به معرفی مختصری از خوشه‌بندی مبتنی بر معیار پیمانگی و بررسی روش‌های پیشین خوشه‌بندی نیمه نظارتی مبتنی بر تجزیه نامنفی ماتریسی پرداخته‌شده است. در بخش ۳، روش حل تکراری الگوریتم نیمه نظارتی پیشنهادی بررسی‌شده است و در بخش ٤، معیارهای ارزیابی و مجموعه داده‌هایی جهت آزمایش معرفی‌شده‌اند. سپس پارامترهای موردنیاز استخراج‌شده و نتایج بر

[1] Semi-Supervised Symmetric Nonnegative Matrix Factorization based on modularity (*Q*)

روی مجموعه داده‌های واقعی نمایش داده‌شده است و در انتها در بخش ۵، نتیجه‌گیری کلی و خلاصه‌ای از کار بیان‌شده است.

۲- خوشه‌بندی گراف‌ها

در این بخش، ابتدا تعریف مختصری از معیار پیمانگی به‌عنوان معیار صحت‌سنجی در خوشه‌بندی ارائه می‌گردد. سپس به بررسی روش‌های خوشه‌بندی مبتنی بر تجزیه نامنفی ماتریسی پرداخته می‌شود و درنهایت خوشه‌بندی گراف‌ها از دیدگاه‌های مختلف نیمه نظارتی موردبررسی قرار می‌گیرد.

۲-۱ خوشه‌بندی مبتنی بر معیار پیمانگی

گراف $G = (V, E)$ در نظر گرفته‌شده است که در آن V مجموعه گره‌ها با مجموع تعداد n گره و E مجموعه یال‌ها میان دو گره با مجموع تعداد m یال مجزا است. در سال‌های گذشته، برای تبدیل گراف‌ها به روابط ریاضی، ماتریس‌های مشابهت گراف همانند ماتریس لاپلاسین[1](L) و ماتریس مجاورتی (A) معرفی‌شده است. همچنین برای بررسی خوشه‌بندی، معیار پیمانگی (Q) بر اساس چگالی یال‌های موجود در گروه‌ها و ارتباطات میان گروهی، صحت‌سنجی گروه‌بندی‌ها را مشخص می‌کند. به‌طور مثال این معیار به‌صورت $Q = \frac{1}{2m}\sum_{ij}\left\{A_{ij} - \frac{k_i k_j}{2m}\right\}\delta(x_i, x_j)$ برای صحت‌سنجی خوشه‌بندی تعریف می‌شود. که k_i و k_j درجه مربوطه گره i ام و j ام و $\delta(x_i, x_j)$ معیار هم‌گروه بودن یا نبودن دو گره i ام و j ام است. همچنین بدیهی است که هر چه این مقدار بزرگ‌تر باشد کیفیت خوشه‌بندی بهتر خواهد بود. همچنین این معیار به‌صورت مستقل جهت تعیین خوشه‌های شبکه‌های پیچیده نیز استفاده می‌شود [٤، ١٩]. ازاین‌رو به‌طورکلی خوشه‌بندی مبتنی بر معیار پیمانگی به شکل مسئله بهینه‌سازی بر روی معیار پیمانگی و شرایط ذیل قابل بازنویسی است [١٩]:

$$\max_H Q = \max_H \frac{1}{2m} tr(H^T B H), \quad B = A - B_1 \quad S.t. H^T H = I \qquad (۱)$$

که B ماتریس معرفی معیار پیمانگی، A ماتریس مجاورتی گراف، $H \in R^{n\times k}$ ماتریس اعضای گروه برای هر گروه، $(B_1)_{ij} = \frac{k_i k_j}{2m}$ و k تعداد خوشه‌ها در شبکه است. همچنین k_i و k_j درجه مربوط به گره i ام و j ام برای محاسبه ماتریس B_1 می‌باشند.

۲-۲ خوشه‌بندی مبتنی بر پایه تجزیه نامنفی ماتریسی

درروش عمومی خوشه‌بندی مبتنی بر تجزیه نامنفی ماتریسی، ماتریس مشابهت گراف به دو ماتریس کاهش بعد یافته شده تبدیل می‌شود و سپس از روی ماتریس‌های تجزیه‌شده، خوشه‌بندی نهایی به دست می‌آید. در به دست آوردن این ماتریس‌های تجزیه‌شده سعی می‌شود که ویژگی‌های پنهان ماتریس اولیه در ماتریس‌های کاهش بعد یافته حفظ شود. به‌عبارت‌دیگر با فرض تعداد خوشه‌بندی k، ماتریس مشابهت $X \in R^{n\times n}$ به دو ماتریس جدید $H \in R^{n\times k}$ و $U \in R^{n\times k}$ تجزیه می‌شود ($X \simeq HU^T$) که ماتریس H بیانگر ماتریس ارتباط جامعه[2] و U ماتریس اعضای هر گروه[3] است. برای به دست آوردن ماتریس‌های کاهش بعد داده‌شده، مسئله بهینه‌سازی با تابع هزینهٔ ذیل معرفی می‌شود:

$$\min_{H,U} J_{nmf}(H, U) = \|X - HU^T\|_F^2 \qquad (۲)$$

که در آن $\|.\|_F$ نرم فروبینوس است. این تابع بهینه‌سازی از مدل تابع بهینه‌سازی غیرمحدب است که در پژوهش‌های متعددی همچون [١٥]، برای به دست آوردن دو ماتریس H و U، روش‌های حل تکراری مبتنی بر توابع لاگرانژین پیشنهادشده است.

یکی دیگر از مدل‌ها، تابع تجزیه نامنفی ماتریسی متقارن[4] است که با کاهش متغیرها توانسته است خطای مدل‌سازی را کاهش دهد. با توجه به [١٦]، تابع بهینه‌سازی به‌صورت رابطه (۳) است.

$$\min_H J_{SNMF}(H) = \|X - HH^T\|_F^2 \qquad (۳)$$

که ماتریس $H \in R^{n\times k}$، بیانگر اعضای هر گروه خواهد بود. با مقایسه روابط (۲) و (۳) مشاهده می‌شود که مزیت این روش، گنجاندن اطلاعات ارتباطات جامعه در ماتریس H است.

[1] Laplacian matrix
[2] Community relation matrix
[3] Community membership matrix
[4] Symmetric NMF (SNMF)

۲-۳ خوشه‌بندی نیمه نظارتی مبتنی بر تجزیه نامنفی ماتریسی

روش‌های خوشه‌بندی نیمه نظارتی، به روش‌هایی اطلاق می‌گردد که اعمال دانش اولیهٔ از اعضای برخی از گروه‌ها باعث بهبود خوشه‌بندی شود. تاکنون دانش اولیه از اعضای خوشه، به دو صورتِ عضو بودن در گروه و عضو نبودن در گروه‌های دیگر مشخص‌شده است. همچنین میزان دانش اولیه از اعضای هر گروه به‌صورت درصدی از دانش اعضای مشخص می‌شود. یکی از متداول‌ترین و فراگیرترین روش‌های خوشه‌بندی نیمه نظارتی، روش‌های خوشه‌بندی مبتنی بر تجزیه نامنفی ماتریسی است. با توجه به سیر زمانی تکامل روش‌های نیمه نظارتی، این روش‌ها معمولاً دو رویکرد کلی را دنبال می‌کنند. رویکرد اول تغییر در ماتریس مشابهت خوشه‌بندی و رویکرد دوم تغییر در تابع بهینه‌سازی تجزیه نامنفی ماتریسی است. با توجه به امکان تغییرات در ترکیب خطی مدل خوشه‌بندی مبتنی بر تجزیه نامنفی ماتریسی و مدل دانش اولیه، رویکرد دوم رویکرد به‌روزتری در سال‌های اخیر بوده است. در ادامه سعی شده است که برای این دو رویکرد توضیحات تکمیلی ارائه گردد.

رویکرد اول، تغییر در ماتریس مشابهت خوشه‌بندی مبتنی بر تجزیه نامنفی ماتریسی است. ژانگ در [۲۵] دانش اولیهٔ اطلاعات برچسب گره‌ها را با ساختار اولیه گراف ترکیب نموده است. به‌صورتی که ماتریس مجاورتی تعمیم‌یافته $\bar{A}$ را جایگزین ماتریس مجاورتی اصلی کرده است. این ماتریس تعمیم‌یافته به‌صورت ذیل تعریف شده است:

$$\bar{A} = \begin{cases} \alpha & \text{اگر } x_i, x_j \text{ دارای برچسب یکسان باشند} \\ 0 & \text{اگر } x_i, x_j \text{ دارای برچسب متفاوت باشند} \\ A_{ij} + I_{ij} & \text{حالات دیگر} \end{cases} \quad (٤)$$

که x_i یا x_j بیانگر برچسب اعضای خوشه برای گره i ام و j ام است. و همچنین در رابطه (٤)، α مقدار مثبت برابر با ۲ و I ماتریس یکه است. همان‌طور که از رابطه (٤) مشخص است این روش با در نظر گرفتن دانش از اعضای هر گروه، ماتریس مشابهت جدیدی برای گراف طراحی می‌کند تا بتواند خوشه‌بندی بهتری با در نظر دانش اولیه اعضای هر گروه ایجاد کند.

در تحقیق دیگر [۲۶]، ماتریس مجاورتی گراف بادانش اولیه جفت شدن گره‌ها در هر گروه و تفاوت گروه‌های میان گره‌های ماتریس مجاورتی بازنویسی شده است که الگوریتم SNMF-SS[1] پدید آمده است ماتریس ورودی خوشه‌بندی مبتنی بر تجزیه نامنفی ماتریسی به‌صورت ذیل تعریف می‌شود:

$$\bar{X} = X - \alpha W_{ML} + \beta W_{CL} \quad (٥)$$

که در آن W_{ML} ماتریسی تشکیل‌شده از رابطه میان اعضای هر گروه و W_{CL} ماتریس تشکیل‌شده از رابطه میان اعضای گروه‌های متفاوت را نشان می‌دهد.

در رویکرد دوم برای استفاده از دانش اولیه، هدف اضافه کردن عبارت جدید به تابع بهینه‌سازی خوشه‌بندی مبتنی بر تجزیه نامنفی ماتریسی است؛ بنابراین تابع هزینه خوشه‌بندی از مجموع بخش خوشه‌بندی مبتنی بر تجزیه نامنفی ماتریسی و بخش مربوط به دانش اولیه خوشه‌بندی تشکیل‌شده است. به‌طور معادل رابطه (٦) تابع هزینه کلی برای این رویکرد را نشان می‌دهد.

$$J_{SNMF} = J_{clustering} + J_{pairs} \quad (٦)$$

که در آن تابع J_{pairs} تابع بهینه‌سازی مربوط به دانش اولیه و $J_{clustering}$ تابع مربوط به خوشه‌بندی است. در ادامه سعی شده است که پژوهشی‌های مرتبط با این رویکرد به‌طور مختصر موردبررسی قرار گیرد.

به‌طور مثال، یانگ در [۲۷]، روش PCNMF[2] را به‌نحوی‌که دانش اولیه به‌صورت رابطه‌ای جدید وارد بهینه‌سازی تجزیه نامنفی ماتریس شود، مطرح کرد. در این رابطه جدید، تابع بهینه‌سازی به‌صورت ذیل بازنویسی می‌شود:

$$J_{PCNMF} = \|X - HU^T\|_F^2 + \lambda tr(U^T L U) \quad (٧)$$

که ماتریس X همان ماتریس مجاورتی است. همچنین $L = D - A_1$ است که در آن ماتریس قطری D با اعضای $D_{ii} = \sum_{j=1}^{n}(A_1)_{ij}$ است و ماتریس A_1 با اطلاعات از دانش اولیه به‌صورت ذیل بازنویسی می‌شود:

[1] Semi-Supervised Nonnegative Matrix -based semi-supervised clustering

[2] Pairwise Constraints-guided Nonnegative Matrix Factorization

$$(A_1)_{ij} = \begin{cases} \alpha & \text{اگر } x_i, x_j \text{ دارای برچسب یکسان باشند} \\ -(1-\alpha) & \text{اگر } x_i, x_j \text{ دارای برچسب متفاوت باشند} \\ 0 & \text{حالات دیگر} \end{cases} \quad (۸)$$

در رابطه فوق α مقدار مثبت و برابر با ۲ است.

در پژوهشی دیگر [۲۸]، روش PCSNMF[1] پیشنهادشده است در این روش دانش اولیه اعضا و ارتباط‌های هر گره در هر گروه و در گروه‌های مختلف با روش تجزیه نامنفی ماتریسی ترکیب‌شده است. تابع بهینه‌سازی مطرح‌شده برای روش PCSNMF به‌صورت ذیل است:

$$J_{PCSNMF} = \|X - HH^T\|_F^2 + \alpha\big(tr(H^T M H B_2) + tr(H^T C H)\big) \quad (۹)$$

که در آن $B_2 = \begin{bmatrix} 0 & 1 & \dots & 1 \\ 1 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \dots & 1 & 0 \end{bmatrix}$ و X همان ماتریس مجاورتی تعریف می‌شود و دانش اولیه در ماتریس‌های ذیل خلاصه‌شده است:

$$M_{ij} = \begin{cases} 1 & \text{اگر } x_i, x_j \text{ دارای برچسب یکسان باشند} \\ 0 & \text{حالات دیگر} \end{cases}$$
$$C_{ij} = \begin{cases} 1 & \text{اگر } x_i, x_j \text{ دارای برچسب متفاوت باشند} \\ 0 & \text{حالات دیگر} \end{cases} \quad (۱۰)$$

علاوه بر آن مطابق با [۱۷] با توجه به خطاهای انسانی در انتخاب اولیه خوشه‌بندی گره‌ها، روش RSSNMF[2] به‌عنوان روشی مقاوم جهت بهبود کارایی در خطاهای انسانی مطرح‌شده است که این روش توانسته است روش PCSNMF را بهبود ببخشد. تابع هزینهٔ بهینه‌سازی بهبود داده‌شده آن به‌صورت ذیل بازنویسی شده است:

$$J_{RSSNMF} = \|X - HU^T\|_2 + \alpha tr(U^T M U B_2) + \beta tr(U^T C U) \quad (۱۱)$$

که در آن $\|.\|_2$ نرم مرتبه دوم و X همان ماتریس مجاورتی است. به دلیل توانایی‌های ذاتی نرم مرتبه دوم در حذف نویز و عدم قطعیت‌ها، تابع بهینه‌سازی مطابق با رابطه (۱۱) در مقابل خطا در دانش اولیه از اعضای هر گروه مقاوم شده است و خوشه‌بندی بهتری نسبت به حضور این عدم قطعیت‌ها به وجود آورده است.

۳- الگوریتم خوشه‌بندی پیشنهادی

با توجه به بررسی روش‌های مختلف و نیازمندی‌های مطرح‌شده اعم از ارائه روش حل تکراری برای خوشه‌بندی مبتنی بر معیار پیمانگی و استفاده از دانش اولیه اعضای گروه در این روش خوشه‌بندی، در این بخش ابتدا در قضیه ۱ مشابهت روش‌های خوشه‌بندی مبتنی بر معیار پیمانگی و خوشه‌بندی مبتنی بر تجزیه نامنفی ماتریسی اثبات می‌شود و سپس، الگوریتم خوشه‌بندی پیشنهادی معرفی می‌گردد.

۳-۱ ارتباط خوشه‌بندی مبتنی بر پیمانگی و خوشه‌بندی مبتنی بر تجزیه نامنفی ماتریسی

در این بخش، جهت تلفیق معیار پیمانگی با خوشه‌بندی مبتنی بر تجزیه نامنفی ماتریسی،، مشابهت میان دو خوشه‌بندی مبتنی بر معیار پیمانگی و تجزیه نامنفی ماتریسی در قالب قضیه ۱ و با توجه به مطالب بخش‌های ۲-۲ و ۲-۱ بیان و اثبات می‌شود.

قضیه ۱: خوشه‌بندی مبتنی بر پیمانگی در شبکه‌های پیچیده دارای ساختاری مشابه با خوشه‌بندی مبتنی بر تجزیه نامنفی ماتریسی است. یا به عبارتی خواهیم داشت.

$$\max_H Q = \min_H \|B - HH^T\|_F^2 \quad (۱۲)$$

اثبات: مطابق با اثبات مشابهت میان دو خوشه‌بندی مبتنی بر معیار چگالی پیمانگی[3] و تجزیه نامنفی ماتریسی در [۳٤]، تابع بهینه‌سازی معیار پیمانگی با توجه به رابطه (۱) به‌صورت ذیل بازنویسی می‌شود:

$$\max_H \frac{1}{2m} tr(H^T B H) \propto -\frac{1}{2m} \min_H tr(H^T B H) \propto -\min_H tr(H^T B H) \quad (۱۳)$$

اگر شرط $H^T H = I$ و H ماتریس ثابت در نظر گرفته شود، خواهیم داشت:

[1] Pairwisely Constrained Symmetric Nonnegative Matrix Factorization

[2] Robust Semi-Supervised Nonnegative Matrix Factorization

[3] Modularity density

$$\max_H \frac{1}{2m} tr(H^T BH) \propto \min_H \big(tr(H^T HH^T H) - 2tr(H^T BH) + tr(HH^T)\big) \quad (۱۴)$$

با توجه به ویژگی‌های ترانهاده‌های ماتریس چون $tr(H^T BH) = tr(BHH^T)$ و $tr(H^T HH^T H) = tr(HH^T HH^T)$، رابطه (۱۴) به‌صورت ذیل بازنویسی می‌شود:

$$\max_H \frac{1}{2m} tr(H^T BH) \propto \min_H tr(HH^T HH^T - 2BHH^T + BB^T) \propto \min_H \|B - HH^T\|_F^2 \quad (۱۵)$$

با توجه به رابطه (۱)، رابطه (۳) و رابطه اثبات‌شده (۱۵)، تابع هزینه بهینه‌سازی ماتریس پیمانگی دارای ساختاری مشابه با تابع هزینهٔ بهینه‌سازی تجزیه نامنفی ماتریسی است. به‌عبارت‌دیگر خواهیم داشت:

$$\max_H Q = \max_H \frac{1}{2m} tr(H^T BH) \propto \min_H \|B - HH^T\|_F^2 \quad (۱۶)$$

ازاین‌رو می‌توان اثبات کرد که خوشه‌بندی مبتنی بر معیار پیمانگی دارای ساختاری مشابه با خوشه‌بندی مبتنی بر تجزیه نامنفی ماتریسی است.

۳-۲ الگوریتم SSNMF-Q

با توجه به قضیه ۱، نشان داده شد که خوشه‌بندی مبتنی بر معیار پیمانگی از جنس خوشه‌بندی مبتنی بر تجزیه نامنفی ماتریسی با فرض ماتریس مشابهت پیمانگی است. مزیت این مشابه‌سازی بهره‌گیری از معیار پیمانگی درروش خوشه‌بندی مبتنی بر تجزیه نامنفی ماتریسی و استفاده از ساختاری جدید برای خوشه‌بندی مبتنی بر معیار پیمانگی است. درنتیجه با در نظر گرفتن این مشابه‌سازی و با توجه به حل ان‌پی–سخت بودن و عدم ارائه راه‌حل نیمه نظارتی روش حل خوشه‌بندی مبتنی بر پیمانگی، در این مقاله روش نوین خوشه‌بندی نیمه نظارتی مبتنی بر ماتریس مشابهت پیمانگی پیشنهاد می‌گردد. این ترکیب علاوه بر استفاده از دانش خوشه‌بندی گراف (پیمانگی)، از دانش‌های اولیه برای بهبود خوشه‌بندی مبتنی بر پیمانگی و تجزیه نامنفی ماتریسی استفاده خواهد کرد تا راه‌حل تکراری و روش خوشه‌بندی نیمه نظارتی مبتنی بر پیمانگی ارائه گردد. بدین منظور با ترکیب بخش خوشه‌بندی و بخش دانش اولیه از برخی از اعضای هر گروه، می‌توان تابع بهینه‌سازی (۱۷) را پیشنهاد داد که این دانش به‌صورت ترکیبی از مجموع اعضای اولیه هر گروه و اعضای دیگر گروه‌ها خواهد بود و بخش خوشه‌بندی از روش خوشه‌بندی قضیه ۱ استفاده خواهد کرد. درنتیجه خوشه‌بندی نیمه نظارتی پیشنهادی به‌صورت ذیل است:

$$\min_H J_{\text{SSNMF-Q}} = \|B - HH^T\|_F^2 + \alpha tr(HMH^T B_2) + \beta tr(HCH^T) \quad s.t. H > 0, \sum_{r=1}^{k} H_{ir} = 1 \quad (۱۷)$$

که در آن ماتریس‌های B_2، M و C با توجه به رابطه‌های (۹) و (۱۰) تعریف خواهند شد.

بدین ترتیب، با در نظر گرفتن پارامترهای داخلی متغیر، بهبود کارایی در بهینه‌سازی الگوریتم SSNMF-Q نسبت به الگوریتم PCSNMF حاصل‌شده است و با در نظر گرفتن قضیه ۱، ماتریس اختصاصی خوشه‌بندی همچون معیار پیمانگی به‌عنوان ماتریس مشابهت الگوریتم خوشه‌بندی مبتنی بر تجزیه نامنفی ماتریسی معرفی‌شده است. همچنین با توجه به رابطه (۱۷) و قضیه ۱، الگوریتم ارائه‌شده برای خوشه‌بندی مبتنی بر معیار پیمانگی یک الگوریتم نیمه نظارتی است.

۳-۳ روش حل تکراری تابع بهینه‌سازی SSNMF-Q

ازآنجاکه در رابطه (۱۷) ماتریس H ماتریس نامنفی است درنتیجه خوشه‌بندی نهایی از طریق روش حل مبتنی بر لاگرانژین قابل‌استخراج است. بر این اساس رابطه (۱۷) به‌صورت فرم ترانهاده آن یعنی رابطه (۱۸) بازنویسی می‌شود.

$$\min_H J_{\text{SSNMF-Q}} = \big(tr(BB^T) - 2tr(BHH^T) + tr(HH^T HH^T)\big) + \alpha tr(HMH^T B_2) + \beta tr(HCH^T) \quad s.t. H > 0, \sum_{r=1}^{k} H_{ir} = 1 \quad (۱۸)$$

با فرض در نظر گرفتن ماتریس φ به‌عنوان ماتریس لاگرانژین، تابع لاگرانژین به‌صورت ذیل تعریف می‌شود:

$$L(H) = J_{\text{SSNMF-Q}} + tr(\varphi H) \quad (۱۹)$$

مشتقات جزئی تابع لاگرانژین نسبت به متغیرهای مسئله به‌صورت زیر خواهد بود:

$$\frac{\partial L_{\text{SSNMF-Q}}}{\partial H} = \varphi - 4\,BH + 4HH^TH + 2\alpha B_2HM + 2\beta C \quad (۲۰)$$

ازآنجاکه $B = A - B_1$ است و شرایط KKT[1] ($\varphi_{ij}H_{ij} = 0$) برقرار است. رابطه (۲۰) به‌صورت رابطه (۲۲) بازنویسی می‌شود:

$$-4(AH)_{ij}H_{ij} + 4(B_1H)_{ij}H_{ij} + 4(HH^TH)_{ij}H_{ij} + 2\alpha(B_2HM)_{ij}H_{ij} + 2\beta(CH)_{ij}H_{ij} = 0 \quad (۲۱)$$

درنتیجه قانون به‌روزرسانی تکراری به‌صورت ذیل طراحی می‌شود.

$$H_{ij} = H_{ij}\frac{4(AH)_{ij}}{4(B_1H)_{ij} + 4(HH^TH)_{ij} + 2\alpha(B_2HM)_{ij} + 2\beta(CH)_{ij}} \quad (۲۲)$$

با توجه به [۱۹]، شرط $\sum_{r=1}^{k} H_{ir} = 1$ به شکل زیر قابل‌اعمال است.

$$H_{ir} := \frac{H_{ir}}{\sum_{r=1}^{c} H_{ir}} \quad (۲۳)$$

درنهایت الگوریتم SSNMF-Q در ادامه تعریف می‌شود:

الگوریتم خوشه‌بندی: مدل SSNMF-Q

ورودی:

ماتریس مجاورتی A

تعداد خوشه‌بندی k

تعداد انجام حلقه I_t

دانش اولیه از خوشه‌بندی

خروجی:

برچسب گروه برای هر گره

مراحل الگوریتم:

۱: مقداردهی به ماتریس H

۲: محاسبه ماتریس‌های B_1, M, CB, با توجه به رابطه (۱۷) و (۱)

۳: شروع حلقه با فرض I_t مرتبه

۴: $H_{ij} = H_{ij}\frac{4(AH)_{ij}}{4(B_1H)_{ij}+4(HH^TH)_{ij}+2\alpha(\mathrm{B_2HM})_{ij}+2\beta(CH)_{ij}}$

۵: $H_{ir} := \frac{H_{ir}}{\sum_{r=1}^{c} H_{ir}}$

۶: انتهای حلقه

۷: برچسب گره‌ها با رابطه $(v_i, I_i) = argmax_{r\le k} H^*_{ir}$

۴- نتایج و شبیه‌سازی

جهت بررسی و ارزیابی روش‌ها خوشه‌بندی نیمه نظارتی SSNMF-Q با دیگر روش‌ها، ابتدا علاوه بر معیار پیمانگی معیار اطلاعات متقابل (NMI[2]) معرفی می‌شوند. سپس مجموعه شبکه متداول در این راستا معرفی می‌گردد و در انتها با توجه به شرایط مختلف دانش اولیه، ارزیابی و صحت‌سنجی الگوریتم‌های خوشه‌بندی بر روی این دادگان انجام می‌گیرد.

۴-۱ معیار اعتبار سنجی

در این مقاله، معیارهای پیمانگی (Q) و اطلاعات متقابل نرمالیزه به‌عنوان شاخص، جهت بررسی و ارزیابی روش‌های خوشه‌بندی در نظر گرفته‌شده‌اند. بر این اساس علاوه بر معرفی معیار پیمانگی در رابطه (۱)، در معیار اطلاعات متقابل از دانش برچسب‌های خوشه‌بندی واقعی[3] هر گره، برای ارزیابی روش‌ها استفاده می‌شود. همچنین این معیار به دلیل استفاده از دانش برچسب‌های خوشه‌بندی واقعی، معیاری فراگیر جهت ارزیابی روش‌های خوشه‌بندی گراف است. درنتیجه معیار اطلاعات متقابل می‌تواند به‌صورت ذیل بازنویسی شود:

$$NMI(C,C') = \frac{-2\sum_{i=1}^{|C|}\sum_{j=1}^{|C'|} n_{C_i\cap C'_j} \times \log\left(\frac{n \times n_{C_i\cap C'_j}}{n_{C_i} \times n_{C'_j}}\right)}{\sum_{i=1}^{|C|} n_{C_i} \times \log\left(\frac{n_{C_i}}{n}\right) + \sum_{j=1}^{|C'|} n_{C'_j} \times \log\left(\frac{n_{C'_j}}{n}\right)} \quad (۲۴)$$

که C و C' بیانگر دو خوشه‌بندی مختلف از اعضای شبکه است که در آن n_{C_i} تعداد اعضا در گروه C_i و $|C|$ تعداد کل گروه‌ها است. همچنین این معیار زمانی برابر با بیشترین مقدار خود یعنی یک خواهد شد که خوشه‌بندی C به‌طور کامل مشابه با خوشه‌بندی C' باشد و همچنین اگر C و C' به‌طور کامل هیچ شباهتی در اعضای هر خوشه با همدیگر نداشته باشند این معیار برابر با کمترین مقدار خود یعنی صفر خواهد بود.

[1] Karush-Kuhn-Tucker
[2] Normalized Mutual Information
[3] Ground truth partition labels

٤-٢ بررسی و ارزیابی الگوریتم‌های خوشه‌بندی نیمه نظارتی

در این بخش، ابتدا گراف‌های پیچیده و ساده برای ارزیابی معرفی خواهند شد. سپس الگوریتم SSNMF-Q با سایر الگوریتم‌های خوشه‌بندی نیمه نظارتی مبتنی بر تجزیه نامنفی ماتریسی با در نظر گرفتن رویکرد اضافه شدن عبارت دانش اولیه (رویکرد دوم در بخش ۲-۳) مقایسه خواهد شد.

یکی از بخش‌های اساسی و مهم برای بهره‌گیری از دانش اولیه خوشه‌بندی برخی از گره‌ها است. بدین منظور، ابتدا می‌توان جهت پیاده‌سازی و استفاده از دانش اولیه برای هر گراف غیرمستقیم، $\frac{n(n-1)}{2}$ جفت گره را خوشه‌بندی کرد. این گره‌ها، به‌طور تصادفی با توجه به درصدهای تعیین‌شده برای واردکردن دانش اولیه (۱٪، ۵٪، ۱۰٪، ۲۰٪ و ۳۰٪) انتخاب می‌شوند. سپس، با توجه به دانش برچسب‌های خوشه‌بندی واقعی، جفت گره‌های انتخاب‌شده برحسب عضو بودن در هر گروه یا حضور در گروه‌های متفاوت خوشه‌بندی می‌شوند و ماتریس‌های تولیدشده بر اساس دانش اولیه گراف‌ها برای هر الگوریتم تشکیل می‌دهند.

٤-٢-١ مجموعه شبکه‌ها

در این مقاله، پنج مجموعه داده برای ارزیابی و صحت‌سنجی الگوریتم‌های خوشه‌بندی در نظر گرفته‌شده است. مجموعه داده انتخاب‌شده به‌گونه‌ای است که هر دو نوع داده ساده (شبکه کاراته) و پیچیده (سایر شبکه‌ها) مدنظر قرارگرفته شوند تا بتوان کارایی الگوریتم جدید پیشنهادی را برای هر دو نوع داده بررسی نمود. اطلاعات تکمیلی این پنج مجموعه داده در جدول ۱ نمایش و توضیح داده‌شده است.

در این جدول پارامترهای m و n به ترتیب، تعداد یال‌ها و گره‌های هر شبکه هستند که دانسته فرض می‌شوند. همان‌گونه که دیده می‌شود، شبکه‌های پیچیده‌تر دارای گره‌ها و یال‌های بیشتری هستند. لذا، با انتخاب این شبکه‌ها، می‌توان چگونگی عملکرد الگوریتم پیشنهادی را برای شبکه‌هایی با ابعاد مختلف نیز بررسی نمود.

جدول ۱: اطلاعات مربوط به مجموعه شبکه‌های واقعی

شبکه‌ها	n (تعداد گره)	m (تعداد یال)	توضیحات
کاراته	۳٤	۷۸	شبکه‌ای از مجموعه باشگاه‌های کاراته [۳۵، ۱۷]
فوتبال	۱۱۵	٦۱۳	شبکه‌ای از مجموعه فوتبالی دانشگاه آمریکا [۳٦، ۱۷]
دلفین	٦۲	۱۵۹	شبکه‌ای میان دلفین‌ها [۳۷، ۱۹]
کتاب‌های سیاسی	۱۰۵	٤٤۱	شبکه‌ای از مجموعه کتاب‌های سیاسی در آمریکا [۳۸، ۱۹]
بلاگ‌های سیاسی	۱۲۲٤	۳۳٤۳۰	شبکه‌ای از مجموعه بلاگ‌های سیاسی در آمریکا [۳۹، ۲٤]

٤-٢-٢ تنظیم پارامترهای داخلی

در روش‌های خوشه‌بندی مبتنی بر تجزیه نامنفی ماتریسی که وابسته به انتخاب پارامترهای داخلی مدلشان هستند، تنظیم و انتخاب پارامتر یکی از بخش‌های اساسی این الگوریتم‌ها است [۱۷، ۱۹، ۲۸]. از متداول‌ترین روش‌ها جهت تنظیم پارامترها، استفاده از روش سعی و خطا و انتخاب بهترین خوشه‌بندی با استفاده از معیارهای مانند چگالی پیمانگی[۱۷]، پیمانگی[۲۸]و اطلاعات متقابل نرمالیزه[۱۹]هست. در مدل پیشنهادی SSNMF-Q، $\boldsymbol{\alpha}$ و $\boldsymbol{\beta}$ پارامترهای وزن دهی جهت کنترل و بهبود اثر دانش اولیه هم‌گروهی بودن و یا هم‌گروه نبودن گره‌ها است. همچنین این پارامترها می‌توانند در بهبود تابع بهینه‌سازی و رسیدن به بهترین پاسخ تأثیر بسزایی داشته باشند. با توجه به این موضوع، درصورتی‌که $\boldsymbol{\alpha}$ و $\boldsymbol{\beta}$ مقادیر بسیار بزرگی داشته باشند، در مدل پیشنهادی SSNMF-Q بخش خوشه‌بندی مبتنی بر تجزیه نامنفی ماتریس بی‌اثر خواهد شد. ازاین‌رو بهترین خوشه‌بندی حاصل نخواهد شد. از طرفی درصورتی‌که $\boldsymbol{\alpha}$ و $\boldsymbol{\beta}$ مقادیر بسیار کوچک

نزدیک به صفر را داشته باشند عملاً اثر دانش اولیه از گروه برخی از گره‌ها وارد الگوریتم خوشه‌بندی نخواهد شد. درنتیجه بررسی و ارزیابی بهترین مقدار برای α و β می‌تواند منتج به بهترین خوشه‌بندی گره‌های شبکه‌های پیچیده شود. درنتیجه برای تنظیم پارامترهای مدل پیشنهادی SSNMF-Q، مطابق با روش سعی و خطا در [۱۷، ۱۹، ۲۸]مقادیر مختلف از α و β را در بازه‌های مختلف قرار می‌دهیم سپس با توجه به معیار صحت‌سنجی، بهترین خوشه‌بندی را به‌عنوان خوشه‌های نهایی در نظر می‌گیریم. ازاین‌رو، برای تمامی مجموعه داده‌های جدول ۱، ابتدا آزمایش‌های متعددی با توجه به مقادیر مختلف α و β در بازه ۰.۵ تا ۱۰ با فرض افزایش گام به مقدار ۰.۵ انجام‌شده و در هر مرحله مقادیر معیار اطلاعات متقابل و معیار پیمانگی، محاسبه و اندازه‌گیری شده است. سپس، یا انتخاب معیار اطلاعات متقابل به‌عنوان معیار اصلی جهت انتخاب بهترین مقدار پارامترهای α و β، این پارامترها مطابق با جدول ۲ برای انواع شبکه و درصدهای دانش اولیه به دست می‌آیند.

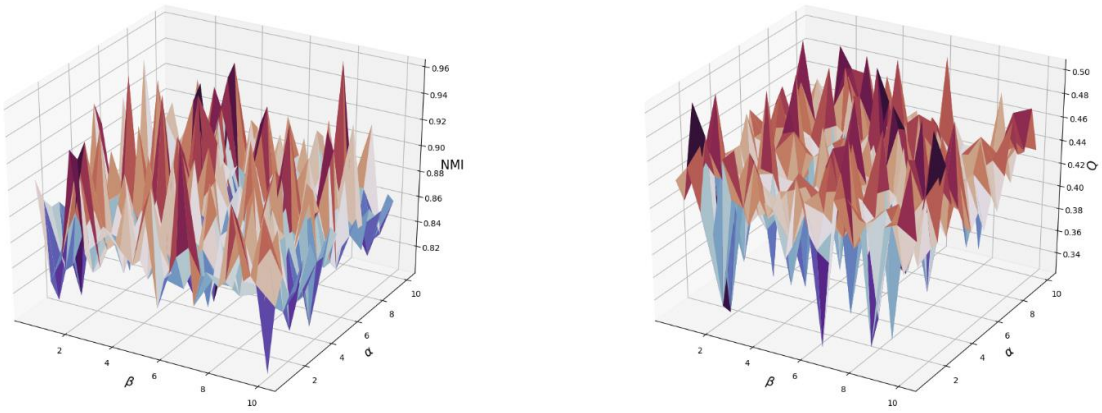
جدول ۲: اطلاعات مربوط به مجموعه شبکه‌های واقعی

شبکه‌ها	پارامتر	درصد دانش‌های اولیه متفاوت				
		٪۱	٪۵	٪۱۰	٪۲۰	٪۳۰
کاراته	α	۹.۵	٤.۵	۱	۳	۰.۵
	β	۷	٦	۷.۵	۳.۵	٤.۵
فوتبال	α	۲	۳.۵	۵.۵	۱.۵	۵
	β	۳	۹	۱.۵	۹	۱۰
دلفین	α	٦.۵	۲.۵	۹.۵	۲.۵	۱۰
	β	۷	۹	۱.۵	۰.۵	۷.۵
کتاب‌های سیاسی	α	٦	۳.۵	۷.۵	۸	۲
	β	۹.۵	۱.۵	۲.۵	۸.۵	۸.۵
بلاگ‌های سیاسی	α	۸.۵	۱.۵	٤	۵.۵	۷.۵
	β	۳.۵	۲.۵	۹.۵	۳.۵	٦.۵

به‌طور مثال مطابق شکل (۱)، نمونه‌ای از مقادیر مختلف معیارهای اندازه‌گیری برای مجموعه داده کتاب‌های سیاسی با فرض ٪۲۰ دانش اولیه در نظر گرفته‌شده است. همان‌طور که در شکل (۱) مشاهده می‌شود مقادیر متفاوت از پارامتر α و β در کیفیت پاسخ تأثیر خواهند داشت. ازاین‌رو بهترین مقادیر پارامترهای α و β برای مجموعه داده جدول ۱ با درصد دانش‌های اولیه متفاوت جهت استفاده در انجام آزمایش‌های نهایی در جدول ۲ ثبت‌شده است.

٤-۲-۳ بررسی و ارزیابی الگوریتم‌های خوشه‌بندی

برای بررسی و ارزیابی الگوریتم‌های مختلف خوشه‌بندی، از الگوریتم‌های خوشه‌بندی PCNMF [۲۷]، PCSNMF [۲۸]و RSSNMF [۱۷]، به دلیل ترکیب خطی بخش دانش اولیه و بخش خوشه‌بندی تجزیه نامنفی ماتریسی استفاده‌شده است. روش PCNMF، روشی مبنا برای ترکیب خطی دو ترم دانش اولیه و تجزیه نامنفی ماتریسی، روش PCSNMF روشی پایه جهت جداسازی انواع دانش اولیه (اعم از دانش هم‌گروه بودن یا هم‌گروه نبودن) و روش RSSNMF روشی جدید و مقاوم برای انواع دانش اولیه است. در میان الگوریتم‌های مختلف نیمه نظارتی، این روش‌ها جزو کارآمدترین آن‌ها هستند. ازاین‌رو مقایسه این روش‌ها با روش نوین SSNMF-Q، مقایسه با بهترین روش‌های نیمه نظارتی دوره زمانی‌های مختلف خواهد بود. برای پیاده‌سازی این الگوریتم‌ها، چنانچه الگوریتم برای شبکه‌های انتخاب‌شده در مرجع موردنظر شبیه‌سازی‌شده باشد، شبیه‌سازی با همان پارامترها تکرار شده است. در غیر این صورت، پارامترهای نامعلوم هر الگوریتم متناسب با مقادیر معین‌شده و با استفاده از روش‌های مطرح‌شده در مرجع موردنظر انتخاب‌شده است. از طرفی برای الگوریتم SSNMF-Q برای تنظیم پارامترها، متناسب با هر داده، پارامترهای موردنیاز مطابق با جدول ۲، تنظیم‌شده‌اند.



شکل (۱): نمایشی از تغییرات α و β با توجه به معیار پیمانگی (سمت چپ) و معیار اطلاعات متقابل (سمت راست) روی‌داده‌های کتاب‌های سیاسی

لذا، با در نظر گرفتن مقادیر مشخص‌شده برای این پارامترها و با توجه به جدول‌های ۳ تا ۷، نتایج زیر به دست می‌آید.

- در تمامی الگوریتم‌ها در جدول‌های ٤ تا ۷ با افزودن دانش اولیه با درصدهای متفاوت، خوشه‌بندی مناسب‌تری ازلحاظ دستیابی به خوشه‌بندی واقعی[1] یا همان بهتر شدن معیار اطلاعات متقابل اتفاق افتاده است؛ اما در برخی موارد مانند جدول ٥، ٦ و ۷، خوشه‌بندی نسبت به معیار پیمانگی افت داشته است. علت این امر آن است که خوشه‌بندی‌های نیمه نظارتی مبتنی بر تجزیه نامنفی ماتریسی با واردکردن دانش اولیه هر گره سعی در بهبود و رسیدن به خوشه‌بندی واقعی خواهند داشت اما، نوع و ویژگی‌های گراف را در نظر نمی‌گیرند. پس بدیهی است که خوشه‌بندی که بر اساس معیاری مبتنی بر خوشه‌بندی واقعی هستند (مانند معیار اطلاعات متقابل) بهبود کیفیت داشته باشد و هم‌چنین ممکن است که خوشه‌بندی که بر اساس معیارهای ویژگی گراف می‌باشند (مانند معیار پیمانگی) دچار افت کیفیت شوند.
- الگوریتم ارائه‌شده در این مقاله برای مجموعه داده ساده مانند کاراته در جدول ۳ توانسته است بدون استفاده از هیچ دانش اولیه‌ای همانند روش‌های دیگر بهترین عملکرد را از خود ارائه دهد.
- از مقایسه نتایج می‌توان دریافت که برای داده‌های پیچیده شامل جدول‌های ٤ تا ۷، الگوریتم پیشنهادی در این مقاله کارایی بهتر از سه الگوریتم دیگر دارد.
- همان‌طور که مشاهده می‌شود، الگوریتم SSNMF-Q الگوریتمی است که با بهره‌ای کمتر از دانش اولیه توانسته است به حداکثر مقدار معیار اطلاعات متقابل برسد. درواقع می‌توان ادعا کرد که به‌صورت میانگین الگوریتم ما با بهره کمتر از دانش اولیه توانسته است خوشه‌بندی بهتری نسبت به دیگر الگوریتم‌های به‌روز دنیا ارائه دهد. همچنین الگوریتم SSNMF-Q نشان داده است که افزایش دانش اولیه باعث بهبود خوشه‌بندی بر پایه معیار پیمانگی می‌شود.
- انتخاب ماتریس پیمانگی به‌عنوان ماتریس مشابهت گراف و انتخاب بهترین پارامتر α و β باعث بهبود پاسخ خوشه‌بندی نسبت به الگوریتم‌های چون RSSNMF و PCSNMF شده است.

[1] Ground truth partition labels

جدول ۳: نتایج الگوریتم‌ها بر روی مجموعه کاراته

درصد	معیار پیمانگی				معیار اطلاعات متقابل			
	PCNMF [۲۷]	PCSNMF [۲۸]	RSSNMF [۱۷]	SSNMF-Q	PCNMF [۲۷]	PCSNMF [۲۸]	RSSNMF [۱۷]	SSNMF-Q
٪۱	۰.۳۷۱	۰.۳۷۱	۰.۳۷۱	۰.۳۷۱	۱	۱	۱	۱
٪۵	۰.۳۷۱	۰.۳۷۱	۰.۳۷۱	۰.۳۷۱	۱	۱	۱	۱
٪۱۰	۰.۳۷۱	۰.۳۷۱	۰.۳۷۱	۰.۳۷۱	۱	۱	۱	۱
٪۲۰	۰.۳۷۱	۰.۳۷۱	۰.۳۷۱	۰.۳۷۱	۱	۱	۱	۱
٪۳۰	۰.۳۷۱	۰.۳۷۱	۰.۳۷۱	۰.۳۷۱	۱	۱	۱	۱

جدول ٤: نتایج الگوریتم‌ها بر روی مجموعه فوتبال

درصد	معیار پیمانگی				معیار اطلاعات متقابل			
	PCNMF [۲۷]	PCSNMF [۲۸]	RSSNMF [۱۷]	SSNMF-Q	PCNMF [۲۷]	PCSNMF [۲۸]	RSSNMF [۱۷]	SSNMF-Q
٪۱	۰.۵۲۸	۰.۵۰۷	۰.۵۳٦	۰.۵٤۰	۰.۸۷۸	۰.۸۳۰	۰.۹۰٤	۰.۹۲۰
٪۵	۰.۵٤۰	۰.۵۲۸	۰.۵۸۰	۰.۵۸۵	۰.۹۲۰	۰۸۷۸	۰.۹۳۰	۰.۹۳۵
٪۱۰	۰.۵۸۰	۰.۵٤۷	۰.۵۸۷	۰.۵۹٤	۰.۹۳۰	۰.۹۰٤	۰.۹۵٦	۰.۹٦۵
٪۲۰	۰.۵۸۷	۰.۵۸۰	۰.٦۰۱	۰.٦۰۱	۰.۹۵٦	۰.۹۳۰	۰.۹۷۳	۰.۹۷۳
٪۳۰	۰.٦۰۱	۰.۵۹٤	۰.٦۰۱	۰.٦۰۱	۰.۹۷۳	۰.۹٦۵	۰.۹۷۳	۰.۹۷۳

جدول ۵: نتایج الگوریتم‌ها بر روی مجموعه دلفین

درصد	معیار پیمانگی				معیار اطلاعات متقابل			
	PCNMF [۲۷]	PCSNMF [۲۸]	RSSNMF [۱۷]	SSNMF-Q	PCNMF [۲۷]	PCSNMF [۲۸]	RSSNMF [۱۷]	SSNMF-Q
٪۱	۰.۳۸٤	۰.۳۸۹	۰.۳۷۸	۰.۳۷۸	۰.۹٦۷	۰.۹۵۱	۰.۹۸۳	۰.۹۸۳
٪۵	۰.۳۷۸	۰.۳۸٤	۰.۳۷۳	۰.۳۷۳	۰.۹۸۳	۰.۹٦۷	۱	۱
٪۱۰	۰.۳۷۳	۰.۳۷۸	۰.۳۷۳	۰.۳۷۳	۱	۰.۹۸۳	۱	۱
٪۲۰	۰.۳۷۳	۰.۳۷۳	۰.۳۷۳	۰.۳۷۳	۱	۱	۱	۱
٪۳۰	۰.۳۷۳	۰.۳۷۳	۰.۳۷۳	۰.۳۷۳	۱	۱	۱	۱

جدول ٦: نتایج الگوریتم‌ها بر روی مجموعه کتاب‌های سیاسی

درصد	معیار پیمانگی				معیار اطلاعات متقابل			
	PCNMF [۲۷]	PCSNMF [۲۸]	RSSNMF [۱۷]	SSNMF-Q	PCNMF [۲۷]	PCSNMF [۲۸]	RSSNMF [۱۷]	SSNMF-Q
٪۱	۰.٤۳۸	۰.۵	۰.۳۹۵	۰.٤۰۳	۰.٤۳۰	۰.٤۰٤	۰.٤٤۸	۰.٤۹۲
٪۵	۰.۳۵۹	۰.٤۸۸	۰.٤۰۵	۰.۳۳۵۷	۰.٤٤۰	۰.٤۱۷	۰.٤۷۰	۰.٤۹٦
٪۱۰	۰.۳٤٤	۰.٤۵٤	۰.۳۹۷	۰.۳۳۸	۰.٤٤۸	۰.٤۲۵	۰.٤۹۱	۰.۵۰۲
٪۲۰	۰.۳۵۷	۰.٤۰۳	۰.۳۳۱	۰.۳۳۸	۰.٤٦۵	۰.٤۳۰	۰.۵۰۲	۰.۵۱۲
٪۳۰	۰.۳۳۱	۰.۳۹۷	۰.۳۲۵	۰.۳۲۵	۰.٤۷٦	۰.٤٤۸	۰.۵۰۲	۰.۵۱۲

جدول ۷: نتایج الگوریتم‌ها بر روی مجموعه بلاگ‌های سیاسی

درصد	معیار پیمانگی				معیار اطلاعات متقابل			
	PCNMF [۲۷]	PCSNMF [۲۸]	RSSNMF [۱۷]	SSNMF-Q	PCNMF [۲۷]	PCSNMF [۲۸]	RSSNMF [۱۷]	SSNMF-Q
٪۱	۰.٤۲۲	۰.٤۲٦	۰.٤۲۳	۰.٤۱۹	۰.۹۵۹	۰.۹۳٦	۰.۹٦۸	۰.۹۷۳
٪۵	۰.٤۲۳	۰.٤۲٤	۰.٤۱۹	۰.٤۱۷	۰.۹٦٤	۰.۹٤۳	۰.۹۷۳	۰.۹۸٤
٪۱۰	۰.٤۲۳	۰.٤۲۲	۰.٤۱۷	۰.٤۱۷	۰.۹۷۲	۰.۹۵۹	۰.۹۸٤	۰.۹۸٤
٪۲۰	۰.٤۱۷	۰.٤۲۳	۰.٤۱۷	۰.٤۱۷	۰.۹۸٤	۰.۹۷۱	۰.۹۸٤	۰.۹۸٤
٪۳۰	۰.٤۱۷	۰.٤۱۷	۰.٤۱۷	۰.٤۱۷	۰.۹۸٤	۰.۹۸٤	۰.۹۸٤	۰.۹۸٤

۵- نتیجه‌گیری

خوشه‌بندی مبتنی بر معیار پیمانگی یکی از روش‌های متداول برای خوشه‌بندی گراف‌ها است. ازاین‌رو در سال‌های گذشته، روش‌های مختلفی برای استخراج گروه‌های گراف مطرح و بررسی‌شده است. محدودیت‌های روش خوشه‌بندی مبتنی بر پیمانگی، ان‌پی–سخت بودن حل مسئله و عدم استفاده از دانش اولیه از اعضای گروه‌های برخی از گره‌ها است. از طرفی، یکی از روش‌های رایج جهت خوشه‌بندی نیمه نظارتی، روش خوشه‌بندی مبتنی بر تجزیه نامنفی ماتریسی است. اما این روش، ویژگی‌های خاص شبکه‌ها را در نظر نمی‌گیرد. ازاین‌رو ما در این مقاله اثبات کرده‌ایم که خوشه‌بندی مبتنی بر معیار پیمانگی دارای ساختاری مشابه با خوشه‌بندی مبتنی بر تجزیه نامنفی ماتریسی است. بر این اساس ماتریس پیمانگی به‌عنوان ماتریس مشابهت گراف درروش خوشه‌بندی مبتنی بر تجزیه نامنفی ماتریسی در نظر گرفته‌شده است و به‌عنوان روش نیمه نظارتی برای خوشه‌بندی مبتنی بر معیار پیمانگی ارائه‌شده است. سپس با توسعه الگوریتم SSNMF-Q امکان ترکیب دانش اولیه از اعضای گروه‌ها با روش مشابه‌سازی شده ایجاد گشته و برای حل بهینه‌سازی، روش حل تکراری پیشنهاد و محاسبه‌شده است. درنهایت نتایج خروجی الگوریتم‌ها بر روی پنج مجموعه

داده‌ٔ مختلف استخراج و بر اساس نتایج آن، کارایی و تأثیرگذاری الگوریتم پیشنهادی بررسی‌شده است.

References

[1] S. Kumar and R. Hanot, " Community Detection Algorithms in Complex Networks: A Survey", Advances in Signal Processing and Intelligent Recognition Systems, Vol. 1365, no. 202, 215, 2021.

[2] K. D. Asim, T. Yahui and G. R. Yulia, " Community detection in complex networks: From statistical foundations to data science applications ", WIREs Computational Statistics, Vol. 14, no. 2, 2021.

[3] M.E.J. Newman, "Networks", Oxford university press, 2018.

[4] M.E.J. Newman and M. Girvan, "Finding and evaluating community structure in networks", Physical Review E, Vol. 69, no. 2, Issue: 026113, 2004.

[5] N. Fariahhag, M. Mordi, Z. J. Wang, "Community structure detection from networks with weighted modularity", Pattern Recognition Letters, Vol. 122, pp. 14-22, 2019.

[6] T. Li, X. Wang, S.H. Zhu, S.H. Zhu and C. Ding, "Community discovery using nonnegative matrix factorization", Data Min. Knowl. Discovery, Vol. 22 no. 3, pp. 493–521, 2011.

[7] C. Li, H. Chen and T. Li, "A stable community detection approach for complex network based on density peak clustering and label propagation ", Applied Intelligence, Vol. 52pp. 1188-1208, 2022.

[8] T. Wang, S. Chen, X. Wang and J. Wang, "Label propagation algorithm based on node importance", Physica A: Statistical Mechanics and its Applications, Vol. 551, no. 124137, 2020.

[9] M. Rosvall and C.T. Bergstrom, "Maps of random walks on complex networks reveal community structure", Proceedings of the National Academy of Sciences, Vol. 105, no.4, pp. 1118–1123, 2008.

[10] J. Zhou, L. Li, A. Zeng, Y. Fan and Z. Di, "Random walk on signed networks", Physica A: Statistical Mechanics and its Applications, Vol. 508, pp. 558-556, 2018.

[11] R. Shang, K. Zhao, W. Zhang, J. Feng, Y. Li and L. Jiao, " Evolutionary multiobjective overlapping community detection based on similarity matrix and node correction, Applied Soft Computing, Vol. 127 no. 109397, 2022.

[12] J. Sanchez and A. Duarte, "Iterated Greedy algorithm for performing community detection in social networks", Future Generation Computer Systems, Vol. 88, pp. 785-791, 2018.

[13] M. Guerrero, F. G. Montoya, R. Baños, A. Alcayde and C. Gil, "Adaptive community detection in complex networks using genetic algorithms", Neurocomputing, Vol. 266, pp. 101-113, 2017.

[14] S. Fortunato and M. Barthłemy, "Resolution limit in community detection", Proceedings of the National Academy of Sciences, Vol. 104, no. 1, pp. 36–41, 2007.

[15] R. Ghadirian and N. Bigdeli, " Hybrid Adaptive Modularized Tri-Factor Non-Negative Matrix Factorization for Community Detection in Complex Networks ", Scientia Iranica, Vol. 14, 2022.

[16] Z. Liu, G. Yuan and X. Luo, "Symmetry and Nonnegativity-Constrained Matrix Factorization for Community Detection," IEEE/CAA Journal of Automatica Sinica, vol. 9, no. 9, pp. 1691-1693, 2022.

[17] C. He, Q. Zheng, Y. Tang, S. Liu, J. Zheng, "Community detection method based on robust semi-supervised nonnegative matrix factorization", Physica A: Statistical Mechanics and its Applications, Vol. 523, no. 1, pp. 279 – 291, 2019.

[18] C. He, Y. Tang, K. Liu, H. Li and S. Liu, "A robust multi-view clustering method for community detection combining link and content information", Physica A: Statistical Mechanics and its Applications Vol. 514, pp. 396-411, 2018.

[19] C. Yan and Z. Chang, "Modularized tri-factor nonnegative matrix factorization for community detection enhancement", Physica A: Statistical Mechanics and its Applications, Vol. 533, no. 122050, 2019.

[20] P. M. Zheng and Z. Zhou, "Structural Deep Nonnegative Matrix Factorization for community detection", Applied Soft Computing, Vol. 97, no: B, Issue: 106846, 2020.

[21] D. Handshutter, N. Gillis and X. Seibert, "A survey on deep matrix factorizations", Computer Science Review, Vol. 42, Issue: 100423, 2021.

[22] J. Huang, T. Zhang, W. Yu, J. Zhu and E. Cai, "Community Detection Based on Modularized Deep Nonnegative Matrix Factorization", International Journal of Pattern Recognition and Articial Intelligence, Vol. 35, no. 35, Issue: 2159006, 2021.

[23] H. jin and S. Li, "Graph regularized nonnegative matrix tri-factorization for overlapping community detection", Physica A: Statistical Mechanics and its Applications,. Vol. 515, pp. 376-387, 2019.

[24] C. Chen, W.Zho and B. Peng, "Differentiated graph regularized non-negative matrix factorization for semi-supervised community detection ", Physica A: Statistical Mechanics and its Applications,.Vol. 604, no. 127692, 2022.

[25] Z.-Y. Zhang, "Community structure detection in complex networks with partial background information", Europhysics Letters, Vol. 101, no. 4, pp. 48005, 2013.

[26] X. Ma, L. Gao, X. Yong, and L. Fu, "Semi-supervised clustering algorithm for community structure detection in complex networks", Physica A: Statistical Mechanics and its Applications, Vol. 389, no. 1, pp. 187 – 197, 2010.

[27] Y. Yang and B. Hu, "Pairwise constraints-guided non-negative matrix factorization for document clustering",in Web Intelligence, IEEE/WIC/ACM International Conference on. IEEE, pp. 250– 256, 2007.

[28] X.H. Shi, H.T. Lu, Y.C. He and S. He, "Community detection in social network with pairwisely constrained symmetric non-negative matrix factorization", Proceedings of the 7th IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 541–546, 2015.

[29] X. Liu, W. Wang, D. He, P. Jiao, D. Jin and C. Vittorio, "Semi-supervised community detection based on non-negative matrix factorization with node popularity", Information Sciences, Vol. 381, no. 12, pp. 304–321, 2017.

[30] C. Févotte, E. Vincent and A. Ozerov. " Single-channel audio source separation with NMF: divergences, constraints and algorithms ", Audio Source Separation, Springer, hal-01631185f, pp. 1-24, 2018.

[31] G. A. Khan, J. Hu, T. li, B. Diallo and H. Wang,." Multi-view data clustering via non-negative matrix factorization with manifold regularization", International Journal of Machine Learning and Cybernetics, Vol. 13, pp. 677-689, 2022.

[32] S. Peng, W. Ser, B. Chen and Z. Lin, " Robust semi-supervised nonnegative matrix factorization for image clustering", Pattern Recognition, Vol. 111, no. 107683, 2021.

[33] K. Huang, X. Fu and N.D. Sidiropoulos, "Anchor-free correlated topic modeling: Identifiability and algorithm", Advances in Neural Information Processing Systems, pp. 1794-1802, 2016.

[34] X. Ma, D. Dong, Q. Wang, “Community detection in multi-layer networks using joint nonnegative matrix factorization”, IEEE Transaction on Knowledge and Data Engineering. 31 (2): 273-286, 2019.

[35] W.W. Zachary, "An information flow model for conflict and fission in small groups", Journal of anthropological research, Vol. 33, no. 4, pp. 452–473, 1977.

[36] A. Lancichinetti, S. Fortunato and F. Radicchi, "Benchmark graphs for testing community detection algorithms", Physical Review E, Vol. 78, no. 4, Isuue: 046110, 2008.

[37] D. Lusseau, K. Schneider, O.J. Boisseau, P. Haase, E. Slooten and S.M. Dawson, "The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations", Behavioral ecology and sociobiology, Vol. 54, no. 4, pp. 396–405, 2003.

[38] J. Kunegis. KONECT: "The Koblenz Network Collection", Proceedings of the 22nd international conference on World Wide Web companion, pp. 1343–1350, 2013.

[39] L. A. Adamic, N. Glance, "The political blogosphere and the 2004 US election: divided the blog", Proceedings of the 3rd workshop on Link discovery, ACM, pp. 36–43, 2005.

Journal of Applied And Basic Machine Intelligence Research (ABMIR)

Applied and Basic Machine Intelligence Research

Vol. 1, No. 1, Spring & Summer 2023, 88-101

A novel semi-supervised clustering method for complex network based on modularity

Mohammad Ghadirian [1], Nooshin Bigdeli [1*]

[1] Control Engineering Department, Imam Khomeini International University, Qazvin, Iran

Article Information

Original Research Paper

Received:
2022 November 15

Accepted:
2022 January 9

Keywords:
Non-negative matrix factorization, community detection, semi-supervised clustering, modularity criterion

Corresponding Author*:

n.bigdeli@eng.ikiu.ac.ir

Abstract

Clustering is a powerful tool for analayzing complex networks which is widely used for modeling complex systems. Modularity is a comprehensive criterion for evaluating the quality of clusters. However, it has some limitations and challenges such as being a NP-hard problem and not using prior information. So, Modularity-based community detection cannot be extended as a semi-supervised community detection method. On the other hand, one of the most common semi-supervised methods which can use prior knowledge for clustering is community detection based on non-negative matrix factorization (NMF). But, this method is not able to consider the features of the networks. Therefore, in this paper to overcome the mentioned limitations and challenges and by presenting a new proof, a structure similar to community detection based on NMF is presented for modularity-based community detection which can employ prior knowledge and iterative solution. Therefore, a novel semi-supervised community detection based on modularity (SSNMF-Q) criteria is developed by utilizing prior information and iterative solution instead of solving a NP-hard problem. To evaluate SSNMF-Q, five real world networks are used and it is shown that the SSNMF-Q had better performance compared to other semi-supervised community detection methods based on NMF.

doi : 10.22034/ABMIR.2023.19231.1018