نشریه علمی- ترویجی محاسبات نرم
سال چهارم/ شمارهٔ هشتم/ پاییز و زمستان ۹۴/ صفحهٔ ۲ـ۱۷

دریافت مقاله: ۹۳/۳/۱۰
پذیرش مقاله: ۹۴/۲/۷

پیشگویی پیوند در شبکه‌های اجتماعی با استفاده از ترکیب دسته‌بندی‌کننده‌ها

اعظم کی‌پور[۱*]، مرتضی براری[۲]، حسین شیرازی[۳]

[۱] دانشجوی کارشناسی ارشد هوش مصنوعی، مجتمع فناوری ارتباطات و اطلاعات، دانشگاه صنعتی مالک اشتر، تهران، ایران
a.keypour@chmail.ir
[۲] استادیار، مجتمع فناوری ارتباطات و اطلاعات، دانشگاه صنعتی مالک اشتر، تهران، ایران
barari@mut.ac.ir
[۳] دانشیار، مجتمع فناوری ارتباطات و اطلاعات، دانشگاه صنعتی مالک اشتر، تهران، ایران
shirazi@mut.ac.ir

چکیده

پیشگویی پیوند در شبکه‌های اجتماعی یکی از فعالیت‌های مهم در تحلیل شبکه‌های اجتماعی است. اهمیت پیشگویی پیوند در شبکه‌های اجتماعی به‌دلیل طبیعت دینامیک آن‌هاست. اعضا و پیوندهای ارتباطی بین آن‌ها در این شبکه‌ها مدام در حال افزایش است و این پیوندها ممکن است به دلایل گوناگون از دست برود؛ لذا با پیشگویی این پیوندها، امکان گسترش و تکمیل این شبکه‌ها و بازیابی اطلاعات و موارد از دست رفته را می‌توان به‌دست آورد. برای کشف و پیشگویی این پیوندها نیاز به اطلاعاتی است که غالباً از ساختار گرافی شبکه استخراج می‌شوند و به‌عنوان معیارهایی برای پیشگویی مورد استفاده قرار می‌گیرند. در این مقاله پس از معرفی دو معیار جدید که کارایی مؤثری در پیشگویی و نیز ارائهٔ پیشنهادات از خود نشان داده‌اند، روش جدیدی ارائه شده است که با ترکیب چند دسته‌بندی‌کننده و با بهره‌گیری از روش‌های تکاملی (الگوریتم ژنتیک و الگوریتم رقابت استعماری)، عمل پیشگویی پیوند را به انجام می‌رساند. برای اثبات کار از دو مجموعه داده واقعی Facebook و Epinions استفاده شده است. ما نشان داده‌ایم که روش پیشنهادی می‌تواند کارایی و دقت پیشگویی را افزایش دهد.

واژه‌های کلیدی: پیشگویی پیوند، شبکه اجتماعی، ترکیب دسته‌بندی‌کننده‌ها، روش‌های تکاملی.

۱. مقدمه

بسیاری مجموعه داده‌های واقعی موجود در جهان می‌توانند به شکل شبکه‌هایی ارائه شوند؛ این شبکه‌ها رأس‌هایی نشان‌دهندهٔ اشیا و یال‌هایی نشان‌دهندهٔ ارتباطات بین اشیا دارند [۱]. مثالی از این نوع مجموعه داده‌ها، داده‌های مربوط به شبکه‌های اجتماعی هستند که می‌توان آن‌ها را در ساختار یک گراف شبکه به نمایش درآورد و برای تحلیل و کاوش در این شبکه‌ها، از ساختار گراف موجود بهره گرفت. یکی از فعالیت های مهم و پرکاربرد در عرصهٔ تحلیل شبکه های اجتماعی و البته تحلیل گراف‌های شبکه ای مختلف، مسئلهٔ پیشگویی پیوند است. پیشگویی پیوند به معنی پیشگویی احتمال برقراری یک ارتباط بین دو رأس است که با توجه به اطلاعات موجود در شبکه و نیز با دانستن اطلاعاتی دربارهٔ ارتباطاتی که قبلاً ایجاد شده است و البته با دانستن این مسئله که در حال حاضر ارتباطی بین این دو رأس وجود ندارد، انجام می‌شود [۲].

شبکه‌های اجتماعی شبکه‌هایی دینامیک هستند که مدام در حال افزایش اعضا و ارتباطات و پیوندهای بین آن‌ها هستند. متأسفانه این پیوندها ممکن است به‌دلیل فرایند ایجاد ناقص و یا به‌سبب اینکه این پیوندها هنوز در این شبکه‌ها انعکاس نیافته‌اند (برای مثال دوستان دنیای واقعی که یک ارتباط اجتماعی مجازی ایجاد نکرده اند)، از دست برود[۳]. اما نکته اینجاست که می‌توان با پیشگویی این پیوندها امکان گسترش و تکمیل این شبکه‌ها و بازیابی اطلاعات و موارد از دست رفته را به انجام رساند. و بر همین اساس یکی از مسائل مهم در شبکه‌های اجتماعی، مسئلهٔ پیشگویی پیوند[1] است.

پیشگویی پیوند که یک امر مهم برای تحلیل شبکه های اجتماعی است، کاربردهایی در حوزه‌های دیگر مثل بازیابی اطلاعات، بیوانفورماتیک و تجارت الکترونیک دارد. در حوزهٔ علم وب و اینترنت می‌تواند در کارهایی از قبیل ایجاد ابرپیوند وب اتوماتیک و پیش‌بینی ابرپیوند سایت‌های وب کاربرد داشته باشد [۴].

روش‌های موجود برای پیشگویی پیوند می‌تواند در سه مقوله گنجانده شود [۴]: اولی، رویکردهای احتمالاتی است که احتمال الحاق موجودیت ها را در یک شبکه اجتماعی، با استفاده از مدل های گرافیکی بیزین مدل می‌کند. رویکرد دوم، رویکرد جبرخطی است، که شباهت بین رأس ها را در یک شبکه، با استفاده از ماتریس های شباهت کاهش رتبه محاسبه می‌کند. اما روش سوم که رایج ترین روش است، شامل مدل‌های سنتی است که مجموعه‌ای از ویژگی‌ها را برای آموزش یک مدل دسته‌بندی دودویی استخراج می‌کند. در این حالت هر نقطه داده، به یک جفت رأس در گراف شبکهٔ اجتماعی مربوط می‌شود. در اینجا برای هر جفت رأس تصمیم می‌گیریم که برای آن پیوند وجود داشته باشد یا نداشته باشد. ازآنجایی‌که پیشگویی پیوند در این حالت تبدیل به یک دسته‌بندی دودویی شده است، از همه ابزار رایج دسته‌بندی بانظارت می‌توان استفاده کرد. مسئلهٔ اصلی در این رویکرد، تعیین ویژگی‌های مناسب برای دسته‌بندی و تعیین روش دسته‌بندی کاراست. در این روش برای محاسبهٔ ویژگی های مورد نیاز، تعدادی معیار مجاورت یا شباهت بین دو رأس در شبکه در نظر گرفته و محاسبه می‌شود. سپس این ویژگی‌ها، در دسته‌بندی مورد استفاده قرار می‌گیرد. لیبن ناول و کلین برگ [۵]، اولین کسانی بودند که در سال ۲۰۰۳ این رویکرد را اعمال کردند و شباهت بین یک جفت رأس را با استفاده از معیارهای مختلف شباهت مبتنی بر گراف، استخراج کردند. در سال ۲۰۰۶، حسن و دیگران در [۶] این کار را گسترش دادند و ویژگی‌های بیشتری را به کار اضافه کردند. بعد از آن محققان دیگری در [۷] و [۸] بر روی مجموعه داده‌های فهرست کتب و نویسندگان کار کردند. در ۲۰۰۹، سانگ و دیگران در [۹] از فاکتورگیری ماتریس برای محاسبهٔ شباهت بین رئوس در شبکه‌های اجتماعی واقعی استفاده کردند. پاپادیمیتریو و دیگران در [۱۰] با ترکیب روش‌های سراسری و محلی جست‌وجوی گراف شبکه اجتماعی، اقدام به تعریف ویژگی جدیدی در پیشگویی پیوند کردند. کوکیرسکی [۱۱] و دیگران در ۲۰۱۱، چندین ویژگی جداگانه از گراف استخراج

1. Link Prediction

کردند و با استفاده از دسته‌بندی‌کننده Random Forest به نتایج خوبی در پیشگویی پیوند در رابطه با داده‌های شبکه اجتماعی فلیکر دست یافتند. پس از آن، فایر و دیگران [۳] با تعریف یک ویژگی جدید و با استفاده از روش ترکیب دسته‌بندی‌کننده‌ها، اقدام به پیشگویی پیوند در شبکه‌های اجتماعی کردند. فیزا و همکاران در [۱۲] با استفاده از یک شبکهٔ عصبی پس انتشار، برای پیشگویی وجود یا ضرورت وجود یک پیوند بین یک جفت رأس با استفاده از مشخصه‌های زوج رأس از قبیل تقابل، انتقال‌پذیری و همسایه‌های مشترک استفاده کردند. اما مسئلهٔ پیشگویی پیوند در شبکه‌های اجتماعی، به‌دلیل اهمیت لزوم دقت در پیشگویی به‌خصوص در مواردی چون کشف گروه‌های مجرم و تبهکار، همچنان یک مسئلهٔ باز است و محققان همواره به دنبال کشف راه‌های جدید برای افزایش دقت و کارایی روش‌ها هستند و در این مقاله نیز تلاش شده است این هدف را محقق سازند.

ویژگی‌های مورد استفاده در تعیین میزان شباهت بین دو رأس، در دو گروه قرار می‌گیرند [۴]: گروه اول، ویژگی‌های مبتنی بر مشخصات رأس و گروه دوم، ویژگی های مبتنی بر ساختار گراف است. این ویژگی‌ها، شامل دو دسته ویژگی‌های محلی و سراسری است. ویژگی‌های محلی دارای مزیت سرعت و ویژگی‌های سراسری دارای مزیت دقت هستند.

در این مقاله، ضمن معرفی دو ویژگی و معیار جدید که توسط نگارندگان مقاله تعریف شده و مورد آزمایش قرار گرفته، روشی ارائه شده است که براساس رویکرد سوم پیشگویی پیوند، با یک دسته‌بندی مناسب به پیشگویی‌های خوبی دست می‌یابد. در این مقاله با ترکیب چند دسته‌بندی‌کنندهٔ قوی و با استفاده از روش رأی‌گیری حداکثری پیشگویی‌ای با کارایی و دقت بالا به‌دست آمده است.

این مقاله به این صورت ساماندهی شده است که در فصل دو، روش پیشنهادی این مقاله به‌طور کامل تشریح شده است. در بخش ابتدایی این فصل، ویژگی‌های مورد نیاز برای پیشگویی پیوند بیان شده و در بخش بعد ترکیب دسته‌بندی‌کننده‌ها و نحوهٔ استفاده از روش های تکاملی در روش پیشنهادی بیان شده است. در فصل سه، اعمال روش بر روی مجموعه داده‌های واقعی و کسب نتایج تجربی انجام شده است. در نهایت در فصل آخر، نتیجه‌گیری بیان شده است.

۲. روش مورد استفاده در پیشگویی پیوند

رویکرد مورد توجه برای پیشگویی پیوند در این مقاله، دسته‌بندی دودویی یال‌هایی است که هنوز تشکیل نشده و رأس‌های مورد نظر هنوز ارتباطی را ترتیب نداده‌اند. روشی که در این مقاله برای دسته‌بندی یال‌ها پیشنهاد شده، استفاده از ترکیب دسته‌بندی‌کننده‌هاست. به این ترتیب که پس از انتخاب‌های چندگانه‌ای که برای بهبود کار در مرحلهٔ انتخاب ویژگی‌های مناسب در دسته‌بندی، انتخاب دسته‌بندی‌کننده‌های مناسب کار و انتخاب روش ترکیب دسته‌بندی‌کننده‌ها صورت می‌گیرد، یال‌هایی که هنوز در گراف شبکهٔ اجتماعی ایجاد نشده‌اند (در واقع ارتباطاتی که هنوز بین اعضا شکل نگرفته است)، به دو دسته مثبت و منفی تقسیم می‌شوند. یال‌های مثبت یال‌هایی هستند که احتمال تشکیل آن‌ها در آینده وجود دارد. در ادامه پس از شرح کاملی از مقدمات کار و ارائه دو ویژگی پیشنهادی، روش پیشنهادی و نتایج تجربی توضیح داده شده است.

۲.۱. مفاهیم بنیادین

برای دسته‌بندی، در درجهٔ اول نیاز به ویژگی‌هایی داریم که براساس آن دسته‌بندی و درنتیجه پیشگویی را انجام دهیم. پیش از این بیان شد که استخراج ویژگی‌ها از گراف شبکه، با دو رویکرد محلی و سراسری صورت می‌گیرد [۴]. روش‌های محلی برای محاسبهٔ ویژگی‌های مورد نیاز مسیرهای با طول ۲ را در نظر می‌گیرند، درحالی‌که روش‌های سراسری، برای یافتن میزان شباهت بین دو فرد، کل گراف شبکه را پیمایش می‌کنند. علاوه بر این دو رویکرد روش‌هایی وجود دارند که حالتی میانه را در نظر می‌گیرند و مسیرهایی بیشتر از طول ۲ را با درنظرگرفتن یک حد آستانه مورد بررسی قرار می‌دهند. در این مقاله به‌دلیل زمان محاسباتی و پیچیدگی بالای روش‌های سراسری، از دو نوع دیگر ویژگی‌ها استفاده شده است و در

دسته‌بندی‌ها شرکت داده شده‌اند؛ البته دلیل دیگر عدم استفاده از ویژگی‌های سراسری، منطق موجود در «نظریهٔ دنیای کوچک» [۱۳] است که براساس آن، هر دو فرد به‌طور میانگین در یک رابطه با ۶ واسطه یا کمتر به هم مربوط می‌شوند. بنابراین بررسی همهٔ مسیرهای ممکن موجود بین دو رأس که در روش‌های سراسری صورت می‌گیرد، نتایج بیشتری به دست ما نمی‌دهد. ویژگی‌هایی که در این مقاله برای دسته‌بندی استفاده شده‌اند، در ادامه شرح داده شده است.

۲. ۱. ۱. استخراج ویژگی

این بخش اختصاص دارد به ویژگی‌هایی که در این کار، از گراف شبکه اجتماعی استخراج شده و در دسته‌بندی مورد استفاده قرار گرفته‌اند. اگر G=(V,E) گرافی باشد که ساختار توپولوژیک شبکهٔ اجتماعی را نشان می‌دهد، هر یال گراف با $e=(u,v)\in E$ نمایش داده می‌شود که $u,v\in V$ هستند. هدف ما ایجاد یک دسته‌بندی‌کنندهٔ مناسب، با استفاده از تکنیک‌های یادگیری ماشین است، به‌گونه‌ای که بتواند برای هر دو رأس u و v پیشگویی کند که آیا اتصال بین این دو رأس دارای احتمال بالایی هست یا خیر؛ لذا برای هر یال کاندید برای دسته‌بندی، یک مجموعه از ویژگی‌های برگرفته از ساختار توپولوژیک شبکه استخراج شده و مورد استفاده قرار گرفته است. این ویژگی‌ها به شرح زیر تعریف می‌شوند:

«ویژگی‌های درجه رأس»[۴]: درجه، در شبکه‌های اجتماعی به معنی تعداد دوستانی است، که یک فرد(رأس) دارد؛ و در یک گراف به معنی تعداد همسایه‌های یک رأس است.

$$d(v)=|\Gamma(v)| \qquad (۱)$$

که در آن

$$\Gamma(v)=\{u\,|\,(u,v)\in E \quad or \quad (u,v)\in E\} \qquad (۲)$$

به‌عنوان همسایگی یک رأس در نظر گرفته می‌شود. این ویژگی برای هریک از دو رأس یال در قالب دو ویژگی مجزا در نظر گرفته می‌شود و به همراه نسبت درجه‌های دو رأس، مجموعاً شامل سه ویژگی می‌شود.

«ویژگی‌های زیرگراف رأس» [۳]: اگر زیرگراف همسایگی را به‌صورت زیر تعریف کنیم:

$$nh_subgraph(v)=\{(x,y)\in E\,|\,x,y\in\Gamma(v)\} \qquad (۳)$$

ویژگی‌های مربوط به زیرگراف رأس می‌تواند تعداد یال‌هایی که در هر زیرگراف وجود دارد و چگالی مربوط به آن‌ها را شامل شود.

$$graph_Edge_Number(v)=|nh_subgraph(v)| \qquad (۴)$$

چگالی هر رأس برابر است با درجه رأس تقسیم بر تعداد کل ارتباطاتی که در زیرگراف مربوط به آن رأس وجود دارد.

$$Density_nh_subgraph(v)=\frac{d(v)}{|nh_subgraph(v)|} \qquad (۵)$$

هریک از ویژگی‌های زیرگراف رأس نیز مانند ویژگی‌های درجه رأس، شامل سه ویژگی مجزا خواهد بود:

«ویژگی دوستان مشترک»[1] [۱۴]: که همان معیار دوست دوست[2] مورد استفاده در پیشگویی دوست در شبکه‌های اجتماعی از قبیل facebook است. این ویژگی به‌صورت زیر محاسبه می‌شود که به معنی تعداد همسایه‌های مشترک u و v است:

$$common-friends(u,v)=|\Gamma(u)\cap\Gamma(v)| \qquad (۶)$$

«شاخص آدامیک/آدار»[3] [۱۵]: این معیار از معیار شباهتی استفاده می‌کند که مربوط به تعداد ویژگی‌های مشترکی است که دو کاربر به اشتراک گذاشته‌اند. z همسایهٔ مشترک دو کاربر است و $|\Gamma(z)|$ تعداد همسایه‌های z و در واقع درجهٔ z است. معادله این ویژگی به این صورت است:

$$adamic-adar(u,v)=\sum_{z\in\Gamma(u)\cap\Gamma(v)}\frac{1}{\log|\Gamma(z)|} \qquad (۷)$$

ایدهٔ موجود در این معیار این است که اگر همسایه مشترک دو کاربر، دوستان کمتری داشته باشد (درجه کمتر)، معلوم است گزیده‌تر انتخاب می‌کند و احتمال ایجاد ارتباط قوی‌تری وجود دارد.

«معیار همه دوستان» [۳]: این ویژگی تعداد کل دوستان دو رأس u و v را در نظر می‌گیرد و به شکل فرمول (۸) محاسبه

1. Common Friends
2. FOAF
3. Adamic/Adar Index

می‌شود:

$$total_friend(u,v) = |\Gamma(u) \cup \Gamma(v)| \quad (۸)$$

«ضریب جاکارد» [۱۶]: این ضریب که روشی مشهور برای محاسبهٔ شباهت بین رئوس است، به‌عنوان تعداد همسایه‌های مشترک دو رأس تقسیم بر کل دوستان این دو رأس، تعریف و محاسبه می‌شود و نشان‌دهندهٔ این است که آیا تعداد دوستان مشترک دو رأس، نسبت به کل دوستان این دو، قابل توجه است یا خیر.

$$jaccard_coefficient(u,v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|} \quad (۹)$$

«اندیس ترفیع‌هاب»[1] (HPI) [۱۷]: این اندیس برای تعیین کیفیت هم‌پوشانی توپولوژیک جفت لایه‌ها، در یک شبکهٔ دگرگون‌شونده به‌کار می‌رود و به‌صورت زیر تعریف می‌شود:

$$HPI(u,v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{\min\{|\Gamma(u)|,|\Gamma(v)|\}} \quad (۱۰)$$

«اندیس فشرده‌هاب»[2] (HDI) [۱۸]: مشابه اندیس ترفیع‌هاب است با این تفاوت که مقدار حداکثر درجه‌ها را در نظر می‌گیرد:

$$HDI(u,v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{\max\{|\Gamma(u)|,|\Gamma(v)|\}} \quad (۱۱)$$

«شاخص سالتون»[3] [۱۹]: این ویژگی به‌صورت فرمول (۱۲) تعریف می‌شود و برابر است با نسبت تعداد همسایگان مشترک به میانگین هندسی درجه دو رأس. نام دیگر این ویژگی cosine است.

$$Salton_index(u,v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{\sqrt{|\Gamma(u)| \times |\Gamma(v)|}} \quad (۱۲)$$

«شاخص سورنسون»[4] [۲۰]: این شاخص، تعداد همسایه‌های مشترک را نسبت به میانگین ریاضی آن‌ها اندازه‌گیری می‌کند. این شاخص مشابه شاخص جاکارد است، با این تفاوت که تعداد رئوس را در همسایگان مشترک به‌صورت یکتا شمارش می‌کند. این شاخص در اصل برای ایجاد گروه‌های با بزرگی یکسان براساس شباهت قسمت‌ها در جامعه‌شناسی گیاهان استفاده می‌شود. شاخص سورنسون به‌صورت زیر تعریف می‌شود:

$$sorenson_index(u,v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u)| + |\Gamma(v)|} \quad (۱۳)$$

«شاخص LHN»[5] [۲۱]: این شاخص، شباهت بالا را به جفت رأس‌هایی اختصاص می‌دهد که همسایگان مشترک زیادی در مقایسه با تعداد حداکثر ممکن همسایگان مشترک دارند، شاخص LHN به این صورت تعریف می‌شود:

$$LHN(u,v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u)| \times |\Gamma(v)|} \quad (۱۴)$$

«شاخص RA»[6] [۲۲،۱۸]: انگیزهٔ این شاخص، از اختصاص دینامیک منبع در شبکه‌های پیچیده نشئت گرفته شده است و به شکل زیر تعریف می‌شود:

$$RA_index(u,v) = \sum_{z \in \Gamma(v) \cap \Gamma(u)} \frac{1}{\Gamma(z)} \quad (۱۵)$$

در واقع در این شاخص، تعداد منابعی را که یک رأس در اختیار دارد، در نظر گرفته می‌شود و فرض می‌شود که رأس منابع خود را به‌صورت مساوی بین همهٔ همسایگانش تقسیم می‌کند.

«امتیاز الحاق ترجیحی»[7] [۲۳]: یک مفهوم مشهور در شبکه‌های اجتماعی وجود دارد که کاربران با دوستان زیاد تمایل به ایجاد ارتباطات بیشتر در آینده دارند. این ویژگی به این صورت تعریف می‌شود:

$$preferential_attachment(u,v) = |\Gamma(u)| . |\Gamma(v)| \quad (۱۶)$$

«معیار دوستان» [۳]: معیار دوستان که به معنی تعداد اتصالات بین همسایه‌های u و v است برابر خواهد بود با:

$$friend_measure(u,v) = \sum_{x \in \Gamma(u)} \sum_{y \in \Gamma(v)} \delta(x,y) \quad (۱۷)$$

$$\delta(x,y) = \begin{cases} 1 & if \quad x = y \quad or \quad (x,y) \in E \; or \quad (y,x) \in E \\ 0 & otherwise \end{cases}$$

در این معیار، $\delta(x,y)$ زمانی برابر یک خواهد بود که x همان y و همسایه مشترک دو رأس u و v باشد یا اتصالی بین

1. Hub Promoted Index
2. Hub Depressed Index
3. Salton Index
4. Sorensen Index
5. Leicht–Holme–Newman Index
6. Resource Allocation
7. Preferential Attachment Score

x به‌عنوان همسایه u و y به‌عنوان همسایه v وجود داشته باشد. طبیعتاً هرچه تعداد این اتصالات بیشتر باشد، احتمال ارتباط محکم‌تری را بین دو رأس خواهیم داشت.

«معیار وزن یال»[1] [۱۱]: این معیار ابتدا به‌صورت دو ویژگی جداگانه برای هریک از دو سرِ یال محاسبه می‌شود:

$$w(v)=\frac{1}{\sqrt{1+|\Gamma(v)|}} \quad w_{و}(u)=\frac{1}{\sqrt{1+|\Gamma(u)|}} \qquad (۱۸)$$

حال وزن یال بین دو رأس v و u می‌تواند از طرق مختلف محاسبه شده و مورد بهره‌برداری قرار گیرد. دو طریق محاسبهٔ آن در زیر بیان شده و مورد استفاده قرار گرفته است.

«جمع وزن‌ها» [۱۱]: برابر است با جمع دو وزن تعریف‌شده در فرمول شمارهٔ (۱۸) برای هریک از رئوس u و v:

$$w(u,v)=w(u)+w(v) \qquad (۱۹)$$

«ضرب وزن‌ها» [۱۱]: برابر است با ضرب دو وزن تعریف‌شده در فرمول شمارهٔ (۱۸) برای هریک از رئوس u و v:

$$w(u,v)=w(u)\times w(v) \qquad (۲۰)$$

«ضریب خوشه‌بندی»[2] [۱۲]: این معیار برابر است با نسبت تعداد همسایگان مشترک بین دو رأس بر مقدار حداقل درجه این دو رأس و مقدار آن به‌صورت زیر محاسبه می‌شود:

$$clustering-coefficient(u,v)=\frac{\sum_{z\neq u,v} a_{uz}a_{zv}}{\min\{|\Gamma(u)|,|\Gamma(u)|\}} \qquad (۲۱)$$

که در آن، a_{uz} درایه uzام ماتریس مجاورت گراف شبکه است که در صورت وجود ارتباط بین دو رأس u و z مقدار آن یک و در غیر این صورت مقدار آن صفر است.

«همبستگی درجه»[3] [۱۲]: این مقدار که متناسب با ضریب همبستگی پیرسون است، به شکل زیر تعریف می‌شود:

$$\deg ree-correlation= \frac{4.|\Gamma(u)|.|\Gamma(v)|-|\Gamma(u)|-|\Gamma(v)|}{2.|\Gamma(u)|^2+2.|\Gamma(v)|^2-|\Gamma(u)|-|\Gamma(v)|} \qquad (۲۲)$$

«روش FriendLink» [۱۰]: این روش با درنظرگرفتن یک کران یا آستانه، همهٔ مسیرهای بین دو رأس را که طول کوتاه‌تر از این کران دارند، بررسی می‌کند و از این طریق از مزیت سرعت روش‌های محلی و دقت روش‌های سراسری بهره می‌برد. در واقع این کران، بیشترین طول مسیری را که باید بین دو رأس بررسی شود، تعیین می‌کند. معادلهٔ مورد استفاده در تعیین میزان شباهت دو رأس، در این روش به‌صورت زیر است:

$$friendLink(u,v)=\sum_{i=2}^{l}\frac{1}{(i-1)}.\frac{|path_{u,v}^{i}|}{\prod_{j=2}^{i}(n-j)} \qquad (۲۳)$$

که در آن، $|path_{u,v}^{i}|$ تعداد مسیرهای با طول i بین دو رأس u و v است. l در این فرمول، کران مورد نظر است و i طول مسیرهای مختلف را در خود دارد. n تعداد کل رئوس گراف شبکه است و مخرج کسر برای نرمالیزه کردن مقادیر شباهت به‌کار می‌رود تا بالاترین شباهت دارای مقدار یک باشد.

در ادامه، دو معیار جدید که توسط نگارندگان این مقاله پیشنهاد داده شده است، بیان می‌شود:

«معیار جدید ۱»: ایدهٔ اصلی این معیار از دو معیار آدامیک/آدار و FriendLink و با درنظرگرفتن مزایا و معایب آن‌ها گرفته شده است. معیار آدامیک-آدار ایدهٔ مناسبی در تعیین میزان شباهت کاربران شبکه‌های اجتماعی ارائه می‌دهد و توانسته است پیشگویی‌های خوبی را به انجام رساند. اما اشکال این روش این است که تعیین میزان شباهت در مسیرهای حداکثر با طول ۲ انجام می‌گیرد و با رأس‌های دورتر کاری ندارد و در واقع بخشی از اطلاعات شبکه را از دست می‌دهد. بنابراین یک فرد می‌تواند با دیگر افراد، از راه‌های مختلفی با طول مسیرهای متفاوت در ارتباط باشد و اگر دو فرد با مسیرهای یکتای زیاد با طول‌های مختلف به یکدیگر متصل باشند، به تناسب طول مسیرها، احتمال زیادی وجود دارد که یکدیگر را بشناسند [۱۰]. درنتیجه، بررسی مسیرهای با طول‌های بیشتر از ۲ می‌تواند اطلاعات دقیق‌تری در پیشگویی پیوند در اختیار ما قرار دهد. البته در مقابل روش‌هایی مثل آدامیک/آدار، روش‌های سراسری هستند که همهٔ اطلاعات و ارتباطات موجود در شبکه را برای هر دو نفر

1. Edge Weight
2. Clustering Co-efficient
3. Degree Correlation

در نظر می‌گیرند و علی‌رغم پیچیدگی بالای محاسباتی و صرف زمان و حافظهٔ بالا، کارایی بالایی ندارند. دلیل این امر همان است که به‌عنوان نظریهٔ «شش درجه جدایی» که مقدمهٔ نظریهٔ «دنیای کوچک» [۱۳] است، شناخته می‌شود. این نظریه می‌گوید هر دو انسان ساکن بر روی کره زمین، به‌طور میانگین در یک رابطه با ۶ واسطه یا کمتر به هم مربوط می‌شوند، یعنی حداکثر توسط پنج واسطه به یکدیگر متصل می‌گردند. پس ما می‌توانیم با بررسی مسیرهای با طول بیشتر از ۲ و کمتر از یک حد آستانه، نتایج قابل قبول‌تری به‌دست آوریم.

در مقابل، روش FriendLink که به‌صورت مناسبی مزایای روش‌های سراسری و محلی را ترکیب کرده و با درنظرگرفتن یک کران یا آستانه، طول مسیر بیش از ۲ را دخالت داده است، دارای این اشکال است که درجه رأس‌های میانی را که در مسیر بین دو رأس مورد نظر قرار دارند، در نظر نمی‌گیرد؛ و درنتیجه بخشی از اطلاعات مهمی را که نقش بالایی در تعیین شباهت رئوس دارند، از دست می‌دهد. چراکه درجه رئوس میانی، نشان‌دهندهٔ تعداد دوستانی است که این رئوس دارند و افزایش این تعداد نشان‌دهندهٔ این است که فرد به احتمال زیادی بدون هیچ دلیل یا معیار مشخصی، با افراد زیادی طرح دوستی می‌بندد و به همین دلیل نمی‌تواند رابط مطمئنی در ایجاد دوستی‌های جدید یا یافتن دوستان قدیمی باشد.

بنا بر مطالب بیان‌شده، یک معیار تعیین شباهت بین جفت رأس‌های گراف شبکه پیشنهاد شده است که با درنظرگرفتن درجه رأس‌های میانی بین دو رأس مورد نظر، همهٔ مسیرهای با طول کمتر از l را بررسی کرده و این میزان شباهت را با توجه به معادلهٔ زیر بیان می‌کند:

$$new_measure1(u,v) = \sum_{i=2}^{l} \frac{1}{(i-1)} \left[\sum_{k=1}^{|path_{u,v}^{i}|} \frac{1}{\prod_{j \in path_{u,v}^{i}} |\Gamma(j)|} \right] \quad (۲۴)$$

در این معادله، $|path_{u,v}^{i}|$ تعداد مسیرهای با طول i بین دو رأس u و v است. j رأسی است که در مسیر بین u و v با طول i قرار دارد و $|\Gamma(j)|$ درجه رأس j است. در این معادله درجه رأس‌هایی که به‌صورت سری در یک مسیر واحد قرار دارند، با یکدیگر ضرب شده و معکوس درجات هر مسیر به طول i که بین دو رأس u و v وجود دارند، با یکدیگر جمع می‌شود.

درجه رأس‌های میانی با میزان اعتمادی که به این رأس‌ها برای رابط شدن داریم، نسبت عکس دارند. به همین دلیل معکوس درجه‌ها را با یکدیگر جمع کرده‌ایم.

$\frac{1}{(i-1)}$ فاکتوری است که به مسیرها با توجه به طول آن‌ها وزن می‌دهد. هرچه طول مسیر بیشتر باشد، وزن کمتری به آن داده می‌شود. مسیری که طول ۲ داشته باشد، بیشترین ارزش را دارد و وزن آن یک خواهد بود.

«معیار جدید ۲» [۲۴]: ایدهٔ موجود در معیار آدامیک/آدار این است که اگر همسایهٔ مشترک دو کاربر دوستان کمتری داشته باشد (درجه کمتر)، معلوم است گزیده‌تر انتخاب می‌کند و احتمال ایجاد ارتباط قوی‌تری بین دو رأس مورد نظر وجود دارد. بنابراین به درجهٔ هریک از همسایگان مشترک، وزن معکوسی داده است تا درصورتی‌که همسایهٔ مشترک دارای درجه بالاتری بود، وجود این همسایهٔ واسطه اثر کمتری روی نتیجهٔ پیشنهادات داشته باشد.

نکتهٔ قابل توجه در این روش این است که ممکن است رأس مشترکی که دوستان زیادی دارد، در بسیاری از این دوستان با دو سر ارتباط (دو کاربری که در حال بررسی میزان شباهت آن‌ها هستیم) اشتراک داشته باشد. به عبارت دیگر، اگرچه طبق معیار آدامیک/آدار، داشتن دوستان زیاد برای یک رأس واسطه یک امتیاز منفی محسوب می‌شود، اما اگر این دوستان زیاد خود دوستان مشترک یکی از دو کاربری باشند که قصد پیشنهاد دوست به آن‌ها را داریم، این تعداد دوستان باید امتیاز مثبت تلقی شود؛ چراکه اگر فردی دوستان مشترک زیادی با یک رأس داشته باشد، احتمال اینکه سایر دوستانش نیز، رأس مورد نظر را بشناسند یا تمایل به دوستی با او را داشته باشند، بالا می‌رود.

در واقع در اینجا سعی کرده‌ایم با درنظرگرفتن مسیرهای با طول ۲ که حجم محاسباتی پایینی دارد، اطلاعات بیشتری را

نسبت به دو معیار دوستان مشترک و آدامیک/آدار، در محاسبات دخیل کنیم و نتایج پیشگویی پیوند را دقیق‌تر نماییم.

بنابراین با درنظرگرفتن درجه رأس‌های میانی و نیز دوستان مشترک رأس میانی با دو رأس اصلی، این میزان شباهت با توجه به معادلهٔ زیر بیان شده است:

$$new-measure2(u,v)=\sum_{z\in\Gamma(u)\cap\Gamma(v)}\frac{|\Gamma(u)\cap\Gamma(z)|+|\Gamma(z)\cap\Gamma(v)|}{|\Gamma(z)|} \quad (۲۵)$$

که در آن، $|\Gamma(u)\cap\Gamma(z)|$، تعداد همسایه‌های مشترک u و z است. معادلهٔ بالا برای تعداد بالای دوستان مشترک رأس میانی با هریک از دو رأس اصلی، امتیاز مثبت قائل است و به همین دلیل، مجموع این تعداد در صورت معادله قرار داده شده است تا میزان شباهت را افزایش دهد. از طرف دیگر برای درجه رأس میانی در حالت عادی، امتیاز منفی در نظر گرفته شده و در مخرج معادله جای می‌گیرد.

۲. ۱. ۲. دسته‌بندی برای پیشگویی پیوند

هدف در پیشگویی پیوند، دسته‌بندی یال‌هایی که هنوز تشکیل نشده‌اند، به دو گروه مثبت و منفی است. یال‌های مثبت یال‌هایی هستند که احتمال تشکیل آن‌ها در آینده وجود دارد و یال‌های منفی یال‌هایی هستند که براساس پیشگویی صورت‌گرفته، احتمال ایجاد آن‌ها وجود ندارد. در این مقاله برای دسته‌بندی یال‌ها به دو گروه مثبت و منفی، از ترکیب دسته‌بندی‌کننده‌ها استفاده شده است. برای تکمیل کار و افزایش کارایی و بالا بردن دقت پیشگویی از روش‌های تکاملی بهره گرفته شده است. دلیل استفاده از ترکیب دسته‌بندی‌کننده‌ها این است که هنگام طراحی یک سیستم یادگیر و ازجمله یک دسته‌بندی‌کننده، انتخاب‌های فراوانی وجود دارد. این تنوع انتخاب باعث می‌شود که نوعی از واریانس در عملکرد سیستم وجود داشته باشد. درنتیجه اگر سیستم‌های مختلفی داشته و از نتایج آن‌ها استفاده شود، این امکان وجود دارد که توزیع خطا حول هدف متمرکز شده و با افزایش نمونه‌گیری از این توزیع به نتیجهٔ بهتری برسیم. در واقع به بیان دیگر باید گفت هیچ الگوریتمی وجود ندارد که برای تمامی مسائل و در تمامی زمان‌ها بهترین (دقیق‌ترین) یادگیر را به‌وجود آورد. بنابراین اگر تعدادی یادگیر پایه داشته باشیم، می‌توان با ترکیب نتایج آن‌ها به دقت بالاتری رسید. [۲۵].

تحقیقات گسترده درخصوص ترکیب دسته‌بندی‌کننده‌ها در سال‌های اخیر معلوم کرده است که کارایی ترکیب قویاً مرتبط با انتخاب دسته‌بندی‌کننده‌هایی است که قرار است با یکدیگر ترکیب شوند. کارایی دسته‌بندی‌کننده نیز به انتخاب دقیق ویژگی‌ها وابسته است [۲۶]. از طرف دیگر، تعدادی تکنیک برای ترکیب دسته‌بندی‌کننده‌ها، در دسترس هستند و انتخاب مناسب‌ترین روش به انتخاب دسته‌بندی‌کننده، ویژگی‌ها و فضای داده‌ها برمی‌گردد. در همهٔ این انتخاب‌های چندبعدی، روش‌های تکاملی یکی از مناسب‌ترین تکنیک‌ها برای فراهم کردن بالانس منطقی بین پیچیدگی جست‌وجو و کارایی راه حل‌هاست؛ زیرا این روش‌ها متناسب با شرایط مسئله پارامترهای خود را تغییر می‌دهند و براساس شرایط موجود، می‌توانند خود را آموزش دهند.

در این کار، از ظرفیت‌های روش‌های تکاملی الگوریتم ژنتیک[1] و الگوریتم رقابت استعماری[2] برای انتخاب در ابعاد مختلف ترکیب دسته‌بندی‌کننده‌ها استفاده کرده‌ایم. دلیل استفاده از روش‌های تکاملی این است که این روش‌ها، راه حل‌های مختلفی از یک مسئله را به‌صورت همزمان بررسی می‌کنند و خود را به راه حل‌های خاص یا معیارهای ویژه‌ای محدود نمی‌کنند. به همین دلیل می‌توانند در همهٔ شرایط خود را با مسئله وفق دهند و براساس شرایط مسئله پارامترهای خود را تنظیم کرده، پاسخ بهینه را به‌دست آورند. البته آنچه مسلم است ترکیب دسته‌بندی‌کننده‌هاست (که در واقع دسته‌بندی در چندین مرحله است)، زمان محاسباتی نسبتاً بالایی خواهد داشت و این نوع روش‌ها، در کاربردهای برخط از قبیل ارائهٔ پیشنهاد دوست و پیشنهاد کالا چندان مفید فایده نخواهد بود. اما کاربردهای غیربرخط زیادی وجود دارند که دقت و کارایی

1. Genetic Algorithm
2. Imperialist Competitve Algorithm (ICA)

در آن نسبت به زمان محاسباتی، اهمیتی بسیارتری دارد. ازجمله این کاربردها می‌توان تشخیص و شناسایی گروه‌های مجرم و تروریست را نام برد که اشتباه در تشخیص می‌تواند هزینه‌های جبران‌ناپذیری را بر جامعه تحمیل کند.

۳.۱.۲. روش‌های تکاملی مورد استفاده در دسته‌بندی

در این کار از دو الگوریتم از الگوریتم‌های روش‌های تکاملی استفاده شده است: اولی که الگوریتم ژنتیک است، برای انتخاب ویژگی‌ها مورد استفاده قرار گرفته، و دومی که الگوریتم رقابت استعماری است، برای وزن‌دهی نتایج دسته‌بندی در ترکیب دسته‌بندی‌ها به‌کار گرفته شده است.

الگوریتم ژنتیک

الگوریتم ژنتیک یک ابزار قدرتمند برای جست‌وجوی سراسری و بهینه‌سازی است. یک تکنیک جست‌وجوی برگرفته از ژنتیک طبیعی است که می‌تواند مسائل غیرخطی را با جست‌وجوی همهٔ فضای جست‌وجو و با استفاده از عملیات‌های انتخاب، بازترکیب و جهش برای به‌دست آوردن مجموعه‌ای از پارامترهای مناسب حل کند.

الگوریتم ژنتیک براساس یک جمعیت که هر عضو آن به‌طور موازی رشد می‌کند و در یک روند تکراری اصلاح می‌شود، عمل می‌کند. هر عضو جمعیت، یک راه حل مسئله را در خود دارد. در واقع الگوریتم ژنتیک روی یک جمعیت از راه حل‌های احتمالی مسئله عمل می‌کند و اساس آن بقای شایسته‌ترین‌ها برای تخمین بهترین راه حل است. در هر نسل، مجموعهٔ جدیدی از تخمین‌ها با انتخاب اعضا متناسب با میزان شایستگی آن‌ها ایجاد می‌شود و این فرایند موجب تکامل جمعیت خواهد شد [۲۷].

الگوریتم ژنتیک می‌تواند با انتخاب یک تابع شایستگی متناسب با نوع مسئله و استفاده از روش‌های مختلف بازترکیب و جهش کروموزوم‌ها، نقش مؤثری در تعیین پارامترهای مورد نیاز ایفا کند. بخش‌های اصلی این الگوریتم که متناسب با مسئله تنظیم می‌شوند، روش بازنمایی مسئله در قالب کروموزوم‌ها، روش‌های بازترکیب و جهش، روش انتخاب نسل جدید و تابع برازندگی است. برای تعیین پارامترهای مسئله می‌توان مقادیر این پارامترها را به‌عنوان ژن‌های موجود در کروموزوم در نظر گرفت و در کنار یکدیگر در یک رشته قرار داد. سپس طی عملیات بازترکیب و جهش، کروموزوم‌ها ارتقا داده می‌شوند و در نهایت، براساس تابع برازندگی، شایستگی هریک از کروموزوم‌ها تعیین شده و از طریق روش‌های انتخاب، نسل جدید مشخص می‌شوند. این کار تا آنجایی ادامه می‌یابد که یا به شایستگی مورد نظر برسیم یا تعداد تکرارهای الگوریتم به پایان برسد.

الگوریتم ژنتیک به‌عنوان یک روش بسیار خوب که انواع مختلفی از بهینه سازی پارامتر و انتخاب‌های چندبعدی را به انجام می‌رساند، شناخته شده است [۲۶].

الگوریتم رقابت استعماری

اخیراً الگوریتم جدیدی با نام الگوریتم رقابت استعماری ارائه شده که برای بهینه‌سازی، به‌جای بهره‌گیری از طبیعت از یک پدیدهٔ اجتماعی الهام گرفته است. در این الگوریتم، هر عنصر جمعیت، یک کشور نامیده می‌شود که در واقع نقاطی تصادفی درون فضای جست‌وجو هستند. سپس، چند کشور قوی‌تر (دارای برازندگی بیشتر) به‌عنوان استعمارگر انتخاب می‌شوند (در الگوریتم ICA به‌جای اصطلاح برازندگی، از اصطلاح قدرت استفاده می‌شود). به این ترتیب کشورهای قدرتمند به‌عنوان استعمارگر و کشورهای ضعیف به‌عنوان مستعمره قرار می‌گیرند. در ابتدای اجرای الگوریتم، کشورها به‌صورت تصادفی تولید و چند کشور قدرتمند به‌عنوان استعمارگر انتخاب می‌شوند. سپس سایر کشورها به‌صورت تصادفی به یکی از استعمارگران منتسب می‌شوند. تعداد مستعمرات هر استعمارگر، متناسب با قدرتش است. کشورهای استعمارگر با اعمال سیاست جذب (همگون‌سازی) در راستای محورهای مختلف، کشورهای مستعمره را به سمت خود می‌کشند. سیاست جذب و همگون‌سازی سیاستی است که برای نزدیک شدن فرهنگ و ساختار اجتماعی مستعمرات در فرهنگ حکومت مرکزی انجام می‌گیرد. این موضوع به‌وسیلهٔ حرکت تصادفی هر کشور مستعمره به سمت کشور استعمارگر خود، در فضای جست‌وجو مدل‌سازی شده است. حرکت کشور

مستعمره به سمت کشور استعمارگر به اندازهٔ x و با انحراف زاویه‌ای θ صورت می‌گیرد که این مقادیر به‌طور تصادفی تعیین می‌شوند.

ممکن است در روند حرکت کشورها در طول اجرای الگوریتم، یک کشور مستعمره قدرت بیشتری از استعمارگر نظیر خود پیدا کند. در این حالت، جای کشور مستعمره و استعمارگر عوض خواهد شد. به عبارت دیگر در مراحل بعد، تمام کشورهای مستعمره استعمارگر قبلی، به استعمارگر جدید تعلق خواهند گرفت و حرکت این مستعمرات به سمت استعمارگر جدید خواهد بود. در هر مرحله از تکرار الگوریتم، میان استعمارگران رقابتی برقرار است. در این رقابت، استعمارگری که نسبت به دیگر استعمارگران قدرت کمتری دارد، یکی از مستعمرات خود را از دست می‌دهد. در این فرایند ضعیف‌ترین مستعمره از ضعیف‌ترین استعمارگر به‌طور تصادفی به یکی از استعمارگران دیگر ملحق می‌شود. احتمال انتساب این مستعمره جدید به هریک از استعمارگران نیز متناسب با میزان قدرت آن‌هاست. اگر استعمارگری تمام مستعمرات خود را از دست بدهد، خود به‌صورت مستعمرهٔ یک استعمارگر دیگر در خواهد آمد. مراحل الگوریتم به همین ترتیب ادامه می‌یابد تا بالاخره تعداد استعمارگران به یک برسد. در این حالت تمام کشورها، مستعمرهٔ یک استعمارگرند و الگوریتم به پایان می‌رسد؛ البته شرایط توقف دیگری مانند تعداد تکراری معین نیز، می‌تواند به‌کار رود [۲۸].

در این الگوریتم نیز می‌توان با جایگذاری پارامترهای مسئله به‌عنوان کشورهای موجود، به پاسخ‌های مورد نیاز رسید.

الگوریتم رقابت استعماری یکی از قدرتمندترین الگوریتم‌های تکاملی است که استفادهٔ زیادی در حل انواع مختلف مسائل بهینه‌سازی دارد [۲۹].

۲.۲. روش پیشنهادی

به بیانی ساده، فرایند انتخاب، فرایند معتبر دانستن تعدادی از عناصر در مقابل کنار گذاشتن سایر عناصر از یک مجموعهٔ مشخص است. در رابطه با مدل‌های دسته‌بندی، گزینه‌های انتخاب می‌تواند به‌عنوان شکل‌های مختلف از تصمیمات مختلف برای دستیابی به خروجی نهایی دسته‌بندی در نظر گرفته شود [۲۶]. انتخاب در ترکیب دسته‌بندی‌کننده‌ها می‌تواند در چند مرحلهٔ انتخاب ویژگی‌ها، انتخاب دسته‌بندی‌کننده‌ها و در نهایت، انتخاب روش ترکیب دسته‌بندی‌کننده‌ها و چگونگی این ترکیب مطرح باشد.

ویژگی‌ها

انتخاب ویژگی از دو جنبه اهمیت پیدا می‌کند [۲۶]: اول اینکه انتخاب برخی از ویژگی‌ها به‌جای همهٔ آن‌ها به‌صورت چشمگیری، هزینهٔ محاسباتی روش‌های دسته‌بندی را کاهش می‌دهد و زمان انجام دسته‌بندی را می‌کاهد. دوم اینکه در عمل بسیاری از ویژگی‌ها، کمکی به افزایش کارایی دسته‌بندی نمی‌کنند و برعکس در مواردی موجب کاهش کارایی نیز می‌شوند؛ ازاین‌رو باید با شناسایی این ویژگی‌ها و حذف آن‌ها از عملیات دسته‌بندی در جهت افزایش کارایی و دقت روش‌های دسته‌بندی حرکت کنیم.

راهکارهای مختلفی برای انتخاب ویژگی وجود دارد. یکی از راه‌های مناسب که در این مقاله مورد استفاده قرار گرفته، استفاده از روش‌های تکاملی و مشخصاً الگوریتم ژنتیک برای انتخاب ویژگی است. برای بهره‌گیری از این الگوریتم، طول کروموزوم را برابر با تعداد ویژگی‌های موجود در نظر می‌گیریم، به‌گونه‌ای که هر ژن به یکی از ویژگی‌ها تعلق گیرد. یک بودن هر ژن نشان‌دهندهٔ شرکت ویژگی مربوط در دسته‌بندی و صفر بودن آن نمایندهٔ عدم شرکت آن ویژگی است. تابع شایستگی‌ای که برای انتخاب ویژگی مورد نظر است، باید دو جنبهٔ مهم را در نظر بگیرد: یکی تعداد ویژگی‌ها و دیگری کارایی حاصل از دسته‌بندی با ویژگی‌های در نظر گرفته شده. این تابع شایستگی را به این صورت تعریف کرده‌ایم:

$$fitness = 10^4 \times performance + 0.5 \times num_features \quad (۲۶)$$

performance مقدار کارایی دسته‌بندی‌کننده به‌ازای

درنظرگرفتن کروموزوم مورد نظر است؛ که در این مقاله مقدار AUC دسته‌بندی مدنظر است. ازآنجایی‌که به دنبال بالاترین AUC هستیم، وزن بالایی را در این قسمت برای performance در نظر گرفته‌ایم. بنابراین زمانی که در ازای مقادیر یکی از کروموزوم‌ها، با AUC بالا در دسته‌بندی مواجه شویم، در این معادله مقدار شایستگی متناسب با آن کروموزوم افزایش می‌یابد و نشان‌دهندهٔ این است که دسته‌بندی با درنظرگرفتن این ویژگی‌ها (که در کروموزوم مربوط دارای مقدار یک هستند)، می‌تواند به نتایج خوبی منجر شود. num_features نیز تعداد ویژگی‌های دارای مقدار یک، در کروموزوم مربوطه است که کمتر بودن آن نشان‌دهندهٔ کم شدن هزینهٔ محاسبات، چه در زمان محاسبهٔ ویژگی و چه در هنگام انجام دسته‌بندی است. به همین دلیل ضریب وزنی کمتر از ۱ را برای آن تعیین کرده‌ایم که با افزایش تعداد ویژگی مقدار شایستگی کاهش یابد.

برای عمل بازترکیب از بازترکیب یک نقطه‌ای استفاده شده است. به این ترتیب که یک عدد تصادفی کوچک‌تر از طول کروموزوم ایجاد شده و بیت‌های بعد این عدد در دو عنصر بازترکیب شونده، با یکدیگر تعویض می‌شوند.

در عملیات جهش، از عملگر جهش معمولی استفاده شده است که در آن، یک عدد تصادفی انتخاب شده و مقدار آن خانه از صفر به یک و بالعکس تغییر می‌کند. برای انتخاب نسل بعدی نیز از روش رولت ویل استفاده شده است.

دسته‌بندی کننده‌ها

انتخاب دسته‌بندی‌کننده شاید یکی از مشهودترین بخش‌های انتخاب در ترکیب دسته‌بندی‌کننده‌ها باشد. عموماً برای انتخاب دسته‌بندی‌کننده‌ها در فرایند ترکیب، دو رویکرد وجود دارد. با توجه به رویکرد اول، ترکیب‌کننده ابتدا به‌صورت دلخواه، چند دسته‌بندی‌کننده را انتخاب می‌کند. پس از آن با روش‌هایی، دسته‌بندی‌کننده‌هایی که بهترین کارایی را نتیجه می‌دهند، انتخاب می‌شوند. در رویکرد دوم، چند دسته‌بندی‌کنندهٔ مشخص تعیین می‌شوند و ترکیب‌کننده، خودش را با دسته‌بندی‌کننده‌های انتخاب‌شده تطبیق می‌دهد تا بهترین پاسخ را به‌دست آورد [۲۶]. در این مقاله با نظر به رویکرد دوم هشت دسته‌بندی‌کننده که با توجه به شرایط مسئله و ویژگی‌های موجود بهترین پاسخ‌ها را ارائه داده‌اند، انتخاب شده و مورد استفاده قرار گرفته‌اند.

روش ترکیب دسته‌بندی‌کننده‌ها

فرایند انتخاب با انتخاب دسته‌بندی‌کننده‌ها به پایان نمی‌رسد، بلکه باید پس از تعیین تعدادی دسته‌بندی‌کنندهٔ بهینه، تعدادی از روش‌های ترکیب مورد بررسی قرار گرفته و روش دارای بهترین کارایی انتخاب شود. برای این مقاله، روش‌های مختلفی مورد آزمون قرار گرفته‌اند و از میان روش‌های موجود ترکیب، روش رأی‌گیری حداکثری به‌عنوان مبنای کار قرار گرفته که البته با بهره‌گیری از روش‌های تکاملی، نتایج ترکیب را بهبود داده شده است.

در این کار، الگوریتم رقابت استعماری به‌عنوان یکی از تکنیک‌های جست‌وجوی مبتنی بر جمعیت، در ترکیب دسته‌بندی‌کننده‌ها، به روش رأی گیری حداکثری مورد استفاده قرار گرفته است. این روش به این ترتیب انجام می‌شود که یک ترکیب خطی از خروجی دسته‌بندی‌کننده‌ها را محاسبه کرده و سپس با رأی‌گیری، نتیجهٔ نهایی را مشخص می‌کنیم. به این صورت که خروجی مربوط به کلاس iام، از طریق جمع خطی وزن دار خروجی مربوط به کلاس iام هر دسته‌بندی‌کننده j برای همهٔ دسته‌بندی‌کننده‌ها به‌دست می‌آید.

$$y_i = \sum_{j=1}^{N} w_i^j z_i^j \qquad i=\{1,\ldots,c\} \qquad (۲۷)$$

که در آن، N تعداد دسته‌بندی‌کننده‌ها و c تعداد کلاس‌هاست. z_i^j خروجی برای کلاس i از دسته‌بندی‌کنندهٔ j است و w_i^j ضریب وزنی معادل همین خروجی است. ازآنجایی‌که ما در پیشگویی پیوند با دو کلاس صفر و یک مواجهیم، دو y محاسبه می‌شود. در این روش، خروجی هر دسته‌بندی‌کننده را در z_1^j و متمم صفر یا یک آن را در z_2^j قرار می‌دهیم.

کلاس برنده برای هریک از داده‌ها (هریک از یال‌ها) کلاسی است که بیشترین مقدار y را دارد. بدین معنی که پس از محاسبهٔ y_i ها، کلاس موردنظر برابر مقدار اندیس y

بزرگ‌تر است.

$$Z = \arg\max_i y_i \quad (۲۸)$$

بنابراین مسلم است که در این قسمت، مسئلهٔ اصلی پیداکردن وزن‌های w_i^j برای شرکت در ترکیب خطی است. هریک از c دسته‌بندی‌کننده که قرار است با یکدیگر ترکیب شوند، دارای وزن‌هایی برای هریک از m کلاس خود هستند. بنابراین m×c وزن داریم که باید به طریقی مقادیر آن را بیابیم. روشی که در این مقاله برای محاسبهٔ وزن‌های مورد نیاز استفاده شده، الگوریتم رقابت استعماری است. در الگوریتم رقابت استعماری که پیش از این شرح داده شد، هر عنصر جمعیت یک کشور نامیده می‌شود که ابعادی به اندازهٔ تعداد عناصر مسئله (عناصری که به دنبال یافتن آن‌ها هستیم) دارد. بنابراین در این روش، ابعاد هریک از کشورها برابر m×c است که مقادیر هریک از آن‌ها بین ۲۵ و ۲۵- در نظر گرفته شده است. تابع شایستگی در این روش منطبق با کارایی نهایی ترکیب دسته‌بندی‌کننده‌ها تعریف می‌شود که این کارایی در این قسمت نیز مقدار AUC خروجی است. در سایر مقادیر این الگوریتم، از مقادیر پیش فرض الگوریتم استفاده کرده‌ایم.

پس از اینکه وزن‌های مربوط به هریک از دسته‌بندی‌کننده‌ها با استفاده از الگوریتم رقابت استعماری محاسبه شد، با جایگذاری این مقادیر در معادلهٔ خطی (۲۷)، می‌توان کلاس مربوط به هریک از نمونه‌های داده ورودی را تعیین کرد.

۳. فعالیت‌های تجربی

در این مقاله برای اثبات روش ارائه‌شده، از چند مجموعه دادهٔ واقعی استفاده کرده و با بررسی عملی اثبات کرده‌ایم که روش مورد بحث می‌تواند کارایی پیشگویی پیوند در شبکه‌های اجتماعی را افزایش دهد؛ ازاین‌رو برای اثبات کار و بررسی کارایی، روش 10-Fold_validation را به‌کار برده و با تقسیم مجموعه داده‌های موجود به ۱۰ قسمت، مقدار AUC را محاسبه کرده‌ایم. محاسبهٔ ویژگی‌های مورد نیاز با استفاده از زبان برنامه‌نویسی جاوا انجام شده است.

۱.۳. مجموعه داده

در این مقاله، از دو مجموعه داده واقعی Epinions[1] و Facebook[2] استفاده شده است. Epinions که یک شبکهٔ اجتماعی «چه کسی به چه کسی اعتماد می‌کند[3]» است، شامل 49k کاربر و 487k یال بین جفت رأس‌هاست. مجموعه داده Facebook نیز که در ۳۰ اکتبر ۲۰۰۹ جمع‌آوری شده [۱۰]، شامل 3.7k کاربر و 13.7k یال است.

برای اثبات کار ارائه‌شده در مقاله، برای هر شبکهٔ اجتماعی یک مجموعه دادهٔ جدید با انتخاب تصادفی ۲۵۰۰۰ یال مثبت که در گراف وجود دارند و ۲۵۰۰۰ یال منفی که در گراف موجود نیستند، ایجاد شده است.

برای استخراج ویژگی‌ها از زبان برنامه‌نویسی جاوا استفاده کرده‌ایم. هر دسته‌بندی‌کننده برای انجام عمل دسته‌بندی به ویژگی‌هایی نیاز دارد تا برای هریک از داده‌ها براساس مقادیر ویژگی‌های آن‌ها معلوم کند که هر داده باید در کدام دسته قرار گیرد. بنابراین ازآنجایی‌که ما در اینجا با جفت رأس به‌عنوان داده سروکار داریم، پس از اینکه برای هر جفت رأس موجود در مجموعه دادهٔ انتخاب‌شده، همهٔ ویژگی‌ها را محاسبه کردیم، این مقادیر به‌عنوان ویژگی‌های مورد استفاده در دسته‌بندی، به دسته‌بندی‌کننده‌ها داده شدند. به این ترتیب که از مقادیر به‌دست‌آمده برای آموزش هشت دسته‌بندی‌کننده و پس از آن برای آزمایش روش ارائه‌شده استفاده کرده‌ایم. برای این کار از WEKA که برنامه‌ای است که در دانشگاه weikato به زبان جاوا نوشته شده و بسیاری از الگوریتم‌های یادگیری ماشین را در خود دارد، بهره برده‌ایم.

۲.۳. نتایج تجربی

در این مقاله، از هشت دسته‌بندی‌کنندهٔ متفاوت که هریک به تنهایی کارایی قابل قبولی در رابطه با داده‌های مورد استفاده ارائه داده‌اند، استفاده کرده‌ایم. این دسته‌بندی‌کننده‌ها عبارت‌اند از: Random Forest [30]، SVM [31]، NaiveBayes [32]،

1. http://www.trustlet.org/wiki/
2. http://delab.csd.auth.gr/~symeon
3. who-trusts-whom

adaBoostM1 [35]، Logistic [34]، PART [33]، ClassificationViaRegression [36] و LogitBoost [37]. دلیل این انتخاب آن بوده است که این هشت دسته‌بندی‌کننده نسبت به سایر دسته‌بندی‌کننده‌های مطرح، نتایج بهتری را در رابطه با مجموعه داده‌های مورد استفاده ارائه داده‌اند. برای اجرای دسته‌بندی‌ها و محاسبهٔ کارایی آن‌ها از نرم‌افزار WEKA استفاده کرده‌ایم. نتایج دسته‌بندی مربوط به هریک از این دسته‌بندی‌کننده‌ها را به‌ازای هریک از مجموعه داده‌های مورد استفاده را در جداول (۱) و (۲) نشان داده‌ایم.

برای بررسی دقت دسته‌بندی و در واقع دقت پیشگویی وجود یا عدم وجود پیوند بین هر دو رأس، از یک معیار استاندارد با نام AUC[1] [۳۸] بهره برده شده است. AUC معادل ناحیهٔ زیر منحنی ROC[2] و برابر است با احتمال اینکه یک انتخاب تصادفی یک پیوند از دست رفته (پیوندی که وجود داشته و به دلایلی از دست رفته است یا در آینده تشکیل خواهد شد) مقدار شباهت بالاتری از یک انتخاب تصادفی پیوندی که وجود ندارد و در آینده نیز به وجود نخواهد آمد، داشته باشد. از لحاظ پیاده‌سازی، این معیار به این صورت محاسبه می‌شود که از n مقایسه مستقل، اگر 'n بار تعداد پیوندهای اشتباه امتیاز بالاتری داشته باشند و "n بار پیوندهای اشتباه و پیوندهایی که وجود ندارند، دارای امتیاز یکسانی باشند، مقدار AUC به‌صورت زیر تعریف می‌شود:

$$AUC = \frac{n'+0.5n''}{n} \qquad (۲۹)$$

با توجه به روش پیشنهادی و توضیحات ارائه‌شده در مقاله، مراحل کار را می‌توان به‌صورت زیر خلاصه کرد:

- ابتدا با استفاده از گراف شبکه و داده‌های موجود، ویژگی‌های مربوط به هر یال را محاسبه می‌کنیم.
- چند دسته‌بندی‌کننده را که از نظر کیفیت قابل قبول هستند، در نظر می‌گیریم.
- با استفاده از الگوریتم ژنتیک برای هر دسته‌بندی‌کننده، ویژگی‌هایی که بهترین پاسخ را تولید می‌کنند، تعیین می‌کنیم.
- با توجه به نتایج الگوریتم ژنتیک و کروموزوم‌های به‌دست‌آمده، برای هر دسته‌بندی‌کننده، ویژگی‌های نامناسب را حذف کرده، دسته‌بندی‌ها انجام می‌دهیم.
- نتایج هریک از دسته‌بندی‌کننده‌ها و متمم آن‌ها را به‌ازای داده‌های آموزشی به‌دست می‌آوریم.
- با استفاده از الگوریتم رقابت استعماری و داده‌های به‌دست‌آمده از مرحلهٔ قبل، وزن‌های مربوط را به‌دست می‌آوریم.
- با توجه به وزن‌های به‌دست‌آمده و با استفاده از جمع وزن‌دار خطی و رأی‌گیری بین نتایج، کلاس‌های مربوط به داده‌های آزمایشی تعیین می‌شود.

نتایجی که با توجه به مراحل ذکرشده، برای دو مجموعه دادهٔ واقعی مورد استفاده در این مقاله به‌دست آمد، در جداول زیر آورده شده است. جدول (۱) نتایج مربوط به مجموعه داده FaceBook را در خود دارد.

جدول (۱): نتایج مربوط به مجموعه دادهٔ facebook

دسته‌بندی‌کننده‌ها	AUC اولیه	پس از اعمال الگوریتم ژنتیک		AUC نهایی
		AUC	تعداد ویژگی	
Random Forest	۰/۹۳۷	۰/۹۴۲	۱۱	۰/۹۶۶
SVM	۰/۸۰۷	۰/۸۰۸	۲۰	
Naïve Bayes	۰/۸۶۵	۰/۸۷۹	۱۶	
PART	۰/۹۵۱	۰/۹۵۳	۱۶	
Logistic	۰/۹۴۸	۰/۹۴۸	۳۰	
AdaBoostM1	۰/۹۴۴	۰/۹۴۵	۱۹	
classificationViaRegression	۰/۹۵۷	۰/۹۵۷	۱۷	
LogitBoost	۰/۹۴۷	۰/۹۴۷	۳۰	

در جدول (۲) نتایج مربوط به دسته‌بندی داده‌های مربوط به مجموعه دادهٔ Epinions نشان داده شده است.

1. Area Under the Curve
2. Receiver-Operating Characteristic

جدول (۲): نتایج مربوط به مجموعه دادهٔ Epinions

دسته‌بندی‌کننده‌ها	AUC اولیه	پس از اعمال الگوریتم ژنتیک		AUC نهایی
		AUC	تعداد ویژگی	
Random Forest	۰/۷۲۸	۰/۷۶۹	۱	۰/۸۰۹
SVM	۰/۶۷۰	۰/۶۷۰	۳۰	
Naïve Bayes	۰/۷۴۱	۰/۷۴۲	۱۲	
PART	۰/۷۷۲	۰/۷۷۶	۸	
Logistic	۰/۷۵۸	۰/۷۷۶	۱۵	
AdaBoostM1	۰/۷۶۸	۰/۷۶۹	۱۶	
classificationVia Regression	۰/۷۸۰	۰/۷۸۱	۲۰	
LogitBoost	۰/۷۷۲	۰/۷۷۲	۳۰	

همان‌طور که در هر دو جدول بالا مشاهده می‌شود، مقدار AUC دسته‌بندی در طی مراحل مختلف رو به بهبود می‌رود تا در آخرین مرحله به بهترین نتیجهٔ خود می‌رسد. AUC نهایی پس از اعمال الگوریتم رقابت استعماری و تعیین وزن‌های مورد استفاده در رأی گیری و انجام دسته‌بندی به‌دست می‌آید. این AUC حاصل ترکیب نتایج هشت دسته‌بندی‌کننده است که هریک به‌تنهایی نسبت به سایر دسته‌بندی‌کننده‌ها، نتایج بهتر و AUC بالاتری را نتیجه داده‌اند. بهبود نتیجهٔ نهایی نشان‌دهندهٔ کارایی مناسب روش ارائه‌شده در این مقاله است؛ چنان‌که هیچ‌یک از دسته‌بندی‌های مطرح به‌تنهایی و حتی پس از اعمال الگوریتم ژنتیک و افزایش کارایی نتوانستند به چنین کارایی‌ای دست یابند.

نکتهٔ اساسی موجود در نتایج تجربی که می‌تواند کارایی معیارهای تعریف‌شدهٔ جدید را مشخص کند، این است که در انتخاب‌هایی که الگوریتم ژنتیک روی ویژگی‌ها داشته است، در مجموعه دادهٔ FaceBook در ۵ دسته‌بندی‌کننده، معیار جدید ۱ به‌عنوان یکی از معیارهای مورد انتخاب توسط الگوریتم ژنتیک مطرح شده و در مجموعه داده Epinions در ۳ دسته‌بندی‌کننده این حالت روی داده است. در رابطه با معیار جدید ۲ در مجموعه دادهٔ FaceBook، ۳ دسته‌بندی‌کننده و در مجموعه دادهٔ Epinions، ۲ دسته‌بندی‌کنندهٔ این معیار را مورد انتخاب قرار داده‌اند.

۴. نتیجه‌گیری

یکی از رویکردهای مورد توجه در پیشگویی پیوند، استفاده از یادگیری با نظارت و انجام پیشگویی در قالب دسته‌بندی دوتایی است. در این رویکرد دو مسئله قابل توجه است: یکی مجموعه ویژگی‌هایی است که در دسته‌بندی مورد استفاده قرار می‌گیرند و دیگری روش دسته‌بندی. در این مقاله، از تعدادی از ویژگی‌های مطرح در پیشگویی پیوند که از ساختار توپولوژیک گراف شبکه استخراج شده‌اند، استفاده شده است. به‌علاوه، دو ویژگی جدید نیز معرفی شده و مورد استفاده قرار گرفته است. تجربه نشان داده است که در دسته‌بندی‌ها، در بسیاری موارد می‌توان با ترکیب چند دسته‌بندی‌کننده، از مزایای همهٔ این دسته‌بندی‌کننده‌ها استفاده کرد و کارایی نهایی را افزایش داد. بنابراین در این مقاله نشان داده شده است که با ترکیب چند دسته‌بندی‌کننده که بهترین کارایی را در رابطه با داده‌های واقعی شبکهٔ اجتماعی از خود نشان داده‌اند و با بهره‌گیری از الگوریتم‌ها و روش‌های تکاملی، می‌توان کارایی دسته‌بندی را افزایش داد و درنتیجه، پیشگویی بهتری را در این شبکه‌ها انجام داد.

مراجع

[1] Getoor, L., Diehl, C., P., "*Link mining: a survey*", ACM SIGKDD Explorations Newsletter, Vol. 7, No. 2, pp. 3-12, 2005.

[2] Feng, Xu., Jichang, Zhao., Xua, Ke., *"Link prediction in complex networks: a clustering perspective"*, The European Physical Journal B-Condensed Matter and Complex Systems, Vol. 85, No. 1, pp. 1-9, 2012.

[3] Fire, Michael., Lena, Tenenboim., Ofrit, Lesser., Rami, Puzis., Lior, Rokach., Elovici, Yuval., "*Link prediction in social networks using computationally efficient topological features*", IEEE International Conference on Privacy, Security, Risk, and Trust, IEEE International Conference on Social Computing, pp. 73-80, 2011.

[4] Al Hasan, Mohammad., Zaki, Mohammed J., "*A survey of link prediction in social networks*", Social network data analytics. Springer US, pp. 243-275, 2011.

[5] Liben-Nowell, David., Kleinberg, Jon., *"The link prediction problem for social networks*", Journal of the American Society for Information Science and Technology, Vol. 58, No. 7, pp. 1019-1031, 2007.

[6] Al Hasan, M. and Zaki, M. J., "*Social network data analytics*", C. C. Aggarwal, Ed. Springer, 2011.

[7] Doppa, J. R., Yu, J., Tadepalli, P., and Getoor, L., "*Chance-constrained programs for link prediction*", Proceedings of Workshop on Analyzing Networks and Learning with Graphs at NIPS Conference, 2009.

[8] Sa, H. R. and Prudencio, R. B. C., "*Supervised learning for link prediction in weighted networks*", III International Workshop on Web and Text Intelligence, 2010.

[9] Song, H. H., Cho, T. W., Dave, V., Zhang, Y., and Qiu, L., "*Scalable proximity estimation and link prediction in online social networks*," Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference, ser. IMC '09. New York, NY, USA: ACM, pp. 322–335, 2009.

[10] Papadimitriou, A., Panagiotis, S., Manolopoulos, Y., *"Friendlink: Link prediction in social networks via bounded local path traversal*", International Conference on Computational Aspects of Social Networks (CASoN), pp. 66-71, 2011.

[11] Cukierski, W., Benjamin, H., Yang, B., "*Graph-based features for supervised link prediction*", Neural Networks (IJCNN), The IEEE 2011 International Joint Conference on, pp. 1237-1244, 2011.

[12] Feyessa, T.e, Marwan, B., Lebby, G., *"Node-pair feature extraction for link prediction*", IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing, pp. 1421-1424, 2011.

[13] Milgram, S., "*The small world problem*", Psychology today, Vol. 2, No. 1, pp. 60-67, 1967.

[14] Chen, J., Geyer, W., Dugan, C., Muller, M., and Guy, I., "*Make new friends, but keep the old: recommending people on social networking sites*", Proceedings of the 27th international conference on Human factors in computing systems, pp. 201–210, 2009.

[15] Adamic, L., Adar, E., "*How to search a social network"*, Social Networks, Vol. 27, No. 3, pp. 187–203, 2005.

[16] Jaccard, P.,*" tude comparative de la distribution florale dans une portion des alpes et des jura*", Bull. Soc. Vaud. Sci. Nat. Vol. 37, 1901.

[17] Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., Bargbasi, A.L., "*Community structure in social and biological networks"*, Science, Vol. 297, pp. 1553, 2002.

[18] Zhou, T., Lü, L., Zhang, Y.C., "*Predicting missing links via local information*", The European Physical Journal B-Condensed Matter and Complex Systems, Vol. 71, No. 4, pp. 623-630, 2009.

[19] Salton, G., McGill, M.J.,*"Introduction to modern information retrieval*", McGraw-Hill, Auckland, 1983.

[20] Sorensen, T., *"A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons*", Biol. Skr. Vol. 5, No. 1, pp.1-34, 1948.

[21] Leicht, E.A., Holme, P., Newman, M.E.J., "*Vertex similarity in networks*", Physical Review E, Vol. 73, pp. 026120, 2006.

[22] Ou, Q., Jin, Y.-D., Zhou, T., Wang, B.-H., Yin, B.-Q., "*Power-law strength-degree correlation from resource-allocation dynamics on weighted networks*", Physical Review E, Vol. 75, pp. 102-112, 2007.

[23] Barabsi, A.-L., Albert, R., *"Emergence of scaling in random networks"*, Science, Vol. 286, No. 5439, pp. 509-512, 1999.

[۲۴] کی پور، ا.، براری، م.، شیرازی، ح.، «**ارائه روشی جدید برای پیشگویی پیوند بین رأس‌های موجود در شبکه‌های اجتماعی**»، نشریه مدیریت فناوری اطلاعات، دورهٔ ۶، شمارهٔ ۳، دانشگاه تهران، ۱۳۹۳، ص۴۷۵-۴۸۶.

[25] Ethem, A., *"introduction to machine learning*", The MIT Press, 2010.

[26] Gabrys, B., Dymitr, R., "*Genetic algorithms in classifier fusion*", Elsevier, Applied Soft Computing, Vol. 6, pp. 337–347, 2006.

[27] Vizitiu, I.-C., Ioan, N., Stoica, A., Ciotîrnae, P., Radu, A., Molder, C., *"An optimal full-genetic technique used to train RBF neural networks*", IEEE, pp. 319-322, 2010.

[28] Atashpaz Gargari, E., Caro, L., *"Imperialist competitive algorithm: An algorithm for optimization inspired by imperialistic competition"*, IEEE Congress on Evolutionary Computation (CEC 2007), pp. 4661-4667, 2007.

[29] Tabealhojeh, H., Ghanbarzadeh, A., "*two steps optimization path planning algorithm for robot manipulators using imperialist competitive algorithm*," second RSI/ISM International Conference on Robotics and Mechatronics (ICROM), pp. 15-17, 2014.

[30] Breiman, L., *"Random forests"*, Machine Learning, Vol. 45, No. 1, pp. 5-32, 2001.

[31] Platt, J., "*Fast training of support vector machines using sequential minimal optimization"*, B. Schoelkopf and C. Burges and A. Smola, editors, Advances in Kernel Methods - Support Vector Learning, 1998.

[32] George, H. J., Langley, P., *"Estimating continuous distributions in bayesian classifiers"* Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, pp. 338-345, 1995.

[33] Eibe, F., Ian, H. W., *"Generating accurate rule sets without global optimization"*, Fifteenth International Conference on Machine Learning, pp. 144-151, 1998.

[34] Cessie, l, S., van, Houwelingen, J.C., *"ridge estimators in logistic regression"*, Applied Statistics, Vol. 41, No. 1, pp. 191-201, 1992.

[35] Freund, Y., Schapire, R.E., *"Experiments with a new boosting algorithm"*, Thirteenth International Conference on Machine Learning, San Francisco, pp. 148-156, 1996.

[36] Frank, E., Wang, Y., Inglis, S., Holmes, G., Witten, I.H., *"Using model trees for classification"*, Machine Learning, Vol. 32, No. 1, pp. 63-76, 1998.

[37] Friedman, J., Hastie, T., Tibshirani, R., "*Additive logistic regression: a statistical view of boosting*", The annals of statistics, Vol. 28, No. 2, pp.337-407, 2000.

[38] Hanely, J.A., McNeil, B.J., "*The meaning and use of the area under a receiver operating characteristic (ROC) curve*", Radiology, Vol. 143, No. 29, 1980.