استفاده از دسته‌بند چندگانه به‌منظور پیش‌بینی پیوند بین موجودیت‌های یک شبکه اجتماعی

شمسی یزدی[۱]، کمال میرزایی[۲]، محمدتقی موسی‌زاده میبدی[۳]

[۱] دانشجوی کارشناسی ارشد، دانشگاه آزاد اسلامی، واحد میبد، گروه کامپیوتر، میبد، ایران
Computer.yazdi@gmail.com

[۲] عضو هیئت‌علمی، دانشگاه آزاد اسلامی، واحد میبد، گروه کامپیوتر، میبد، ایران
K.mirzaie@maybodiau.ac.ir

[۳] عضو هیئت‌علمی، دانشگاه علم و هنر، واحد یزد، گروه برق، یزد، ایران
Mousazade@sau.ac.ir

Archive of SID

چکیده

یک شبکه اجتماعی، ساختاری است شامل مجموعه‌ای از بازیگران که تعاملات میان آن‌ها به‌وسیله تعدادی پیوند نشان داده می‌شود. با گسترش روزافزون شبکه‌های اجتماعی در دنیا، لزوم استفاده از روش‌های دقیق‌تر جهت پیش‌بینی پیوند در شبکه‌های اجتماعی بیش‌ازپیش احساس می‌شود. پیش‌بینی پیوند فرایندی است که با در نظر گرفتن تصویر لحظه‌ای از شبکه، تعاملات احتمالی بین اعضا را می‌یابد. بسیاری از راه‌حل‌های موجود، خصوصاً روش‌هایی که کل گراف را پیمایش می‌کنند مدیریت مکانیسم‌های استنتاج را دچار مشکل می‌کنند. با توجه به این مسئله، شناخت و انتخاب ویژگی‌ها و نیز انتخاب روشی که بتواند در بهبود پیش‌بینی‌ها مؤثر باشد و با هزینه کمتر، پاسخ بهتری را در پیش‌بینی‌ها ارائه کند بسیار بااهمیت جلوه می‌کند. این مقاله بر اساس پیوندهای موجود در شبکه و ویژگی گره‌ها، با استفاده از چند دسته‌بند مختلف و سپس ترکیب نتایج خروجی دسته‌بندها تحت عنوان «سیستم‌های دسته‌بند چندگانه »احتمال ایجاد پیوند جدید بین دو گره را تخمین می‌زند. آزمایش‌ها، روی گراف دو شبکهٔ اجتماعی Hi5 و Facebook نشان می‌دهد که روش پیشنهادی نسبت به دسته‌بندهای منفرد و پایه از کارایی خوبی برخوردار بوده و دقت پیش‌بینی را افزایش می‌دهد.

کلمات کلیدی

پیش‌بینی پیوند، شبکه اجتماعی، دسته‌بند چندگانه، یادگیری جمعی، تئوری گراف

۱- مقدمه

یکی از فعالیت‌های مهم در عرصهٔ تحلیل شبکه‌های اجتماعی، مسأله پیش‌بینی پیوند است. پیش‌بینی پیوند، به معنای وجود یا عدم وجود یک پیوند (ارتباط) در آینده بین دو گره یک شبکه اجتماعی است و یک ابزار مهم جهت تحلیل شبکه‌های اجتماعی به شمار می‌رود. کشف ارتباط میان پروتئین‌ها در شبکه‌های زیستی نظیر شبکه‌های تعامل پروتئین-پروتئین[۱] و شبکه‌های متابولیکی، نیازمند صرف وقت و هزینه بسیار زیاد برای انجام تحقیقات در محیط آزمایشگاه می‌باشد لذا می‌توان با هزینه و تعداد آزمایش‌های کمتر به کمک پیش‌بینی پیوند، ارتباط میان پروتئین‌ها را کشف کرد [۱]. گاهی در برخی از شبکه‌ها، تعدادی پیوند به خاطر وجود خطا در شبکه به صورت اتفاقی به وجود می‌آیند. این پیوندهای نادرست می‌توانند ساختار شبکه را مختل کرده و مطالعه آن را با مشکل مواجه کنند که به کمک پیش بینی پیوند می‌توان این پیوندها را شناسایی و از شبکه حذف کرد [۲]. مدل تکاملی شبکه را می‌توان با پیش‌بینی پیوندهای آتی بدست آورد [۳]. از طرف دیگر، در حوزه پزشکی پیش‌بینی نحوه شیوع یک بیماری خاص، توسط پیش‌بینی پیوند امکان‌پذیر می‌باشد. در شبکه‌های اجتماعی می‌توان دوستان جدیدی را به

www.SID.ir

کاربران شبکه پیشنهاد داده و باعث افزایش وفاداری کاربران به شبکه اجتماعی شد [۴]. کابردهای دیگری نظیر بازیابی اطلاعات، بایوانفورماتیک، تجارت الکترونیک و حوزه‌های امنیتی دارد [۵].

در ادامه در بخش ۲ پیشینه پژوهش ارائه می‌شود، در بخش ۳ سیستم‌های دسته‌بند چندگانه معرفی می‌شود و در بخش ۴ روش پیشنهادی و معیارهای شباهت گره مبتنی بر ساختار گراف مطرح می‌شود و در بخش ۵ نحوه ارزیابی و نتایج حاصل از ارزیابی دسته‌بند چندگانه بیان می‌گردد و سپس در انتها در بخش ۶ نتیجه‌گیری نهایی ارائه می‌شود.

۲- پیشینه پژوهش

در سال ۲۰۰۵ اولین بار به طور رسمی در مقاله [۶] مبحثی به نام پیوندکاوی[۲] مطرح شد. روش‌های پیش‌بینی پیوند را می‌توان به دودسته با ناظر[۳] و بدون ناظر[۴] تقسیم‌بندی نمود [۷]. در روش‌های با نظارت بیشتر بر روی الگوریتم‌های دسته‌بندی و استفاده از ویژگی‌هایی همچون همسایگان مشترک و طول مسیرو ... تمرکز دارند. روش‌های بدون نظارت بر اساس الگوریتم‌های خوشه‌بندی بوده و گره‌ها را با توجه به شباهت آن‌ها به خوشه‌هایی تقسیم می‌کنند [۸]. اگرچه روش‌های با ناظر به‌عنوان متدهای پیشرفته‌تری شناخته می‌شوند، اما معمولاً از مشکل انتخاب ویژگی[۵] و کلاس‌های نامتعادل[۶] رنج می‌برند [۹]. در مقابل، در روش‌های بدون ناظر به دنبال استخراج معیاری برای میزان شباهت گره‌های شبکه هستند. مجموعه وسیعی از روش‌های اندازه‌گیری شباهت میان گره‌های یک شبکه، مبتنی بر روش قدم‌زنی تصادفی[۷] هستند [۳]. از معایب روش‌های بدون ناظر آن است که این روش‌ها از شبکه‌ای به شبکه دیگر و از گرافی به گراف دیگر، عملکرد ناپایدار خواهند داشت [۱۰]. فایر و همکاران در مقاله [۱۱] تأثیر محاسبه هریک از ویژگی‌های ساختاری را در پیش‌بینی پیوند، برای ۵ مجموعه داده[۸] themarker و youtube، flicer، facebook، academia سنجیده‌اند. احسان شرکت و همکاران در مقاله [۱۲] به یک روش پیش‌بینی بر اساس الگوریتم کلونی مورچگان (ACO) با استفاده از راهکار بدون ناظر اشاره‌کرده‌اند. روش ارائه‌شده به‌نوعی یک روش تقریبی بوده و هدف اصلی آن بهینه‌سازی زمان اجرای الگوریتم بر روی مجموعه دادگان بزرگ می‌باشد.

۳- سیستم‌های دسته‌بند چندگانه

سیستم‌های دسته‌بند چندگانه[۸] که در منابع یادگیری ماشینی با نام‌های مختلفی مانند شورای دسته‌بندها[۹] نیز خوانده می‌شود [۱۳]. در حقیقت استفاده از سیستم‌های شورایی در حوزه[۱۰] یادگیری ماشینی، از طبیعت انسان در استفاده ازنظرات مختلف برای تصمیم‌گیری‌های مهم الهام گرفته است. بنابراین یکی از روش‌های مناسب برای بهبود صحّت دسته‌بندی، استفاده از چند دسته‌بند مختلف و سپس ترکیب نتایج خروجی آن‌ها که اغلب تحت عنوان «سیستم‌های دسته‌بند چندگانه» می‌باشد.

۴- روش پیشنهادی

شبکه اجتماعی ساختاری است شامل مجموعه‌ای از بازیگران که تعاملات میان آن‌ها به وسیله تعدادی پیوند نشان داده می‌شود. یک بازیگر، موجودیت اجتماعی می‌باشد که ممکن است یک شخص، یک گروه و یا یک شرکت را شامل شود [۱۴].

اگر G=(V,E) گرافی باشد که ساختار توپولوژیک شبکه اجتماعی نشان دهد، هر لبه گراف با e مطابق رابطه (۱) نمایش داده می‌شود:

$$\begin{aligned} & e = (u,v) \\ & e \in E \\ & u,v \in V \end{aligned} \qquad (۱)$$

ساده‌ترین چارچوب روش‌های پیش‌بینی پیوند، استفاده از الگوریتم‌های مبتنی بر شباهت است که در آن به هر جفت گره u و v یک مقدار Score نسبت داده می‌شود. شباهت گره می‌تواند با استفاده از ویژگی‌های اساسی گره تعریف شود و در آن دو گره در صورتیکه ویژگی‌های مشترک زیادی داشته باشند، شبیه به هم در نظر گرفته می‌شوند؛ اما این ویژگی‌ها معمولاً پنهان هستند و باید محاسبه شوند.

رویکرد مورد توجه جهت پیش‌بینی پیوند، استفاده از روش‌های باناظر و تکنیک دسته‌بندی می‌باشد. با استفاده از تکنیک دسته‌بندی و با انتخاب تصادفی پیوندهای مثبت و منفی و حذف نمونه‌های کلاس حداکثری، معیارهای شباهت هر جفت کاربر (پیوند) محاسبه شده و در مرحله یادگیری با استفاده از داده‌های آموزش مدلی ساخته می‌شود. به منظور بهبود صحّت دسته‌بندی، از یادگیری گروهی استفاده می‌شود. روش پیشنهادی در چند مرحله شامل انتخاب ویژگی‌ها، انتخاب دسته‌بندها و نهایتاً چگونگی ترکیب دسته‌بند مطرح می‌باشد. در مرحله آزمایش، لبه‌هایی که هنوز تشکیل نشده‌اند و گره‌هایی که هنوز ارتباطی را تشکیل نداده‌اند را به دسته‌بند داده تا به دو گروه مثبت و منفی دسته‌بندی شوند. پیوندهای مثبت، پیوندهایی هستند که احتمال تشکیل آن‌ها در آینده وجود دارد و پیوندهای منفی، پیوندهایی هستند که بر اساس پیش‌بینی صورت گرفته، احتمال ایجاد آن‌ها وجود ندارد.

از ابزار Matlab جهت محاسبه میزان شباهت بین جفت گره‌های یک شبکه و از ابزار RapidMiner و امکانات آن نیز جهت دسته‌بندی و تحلیل داده‌ها استفاده می‌شود.

۴-۱- انتخاب ویژگی

یکی از چالش‌های پیش‌بینی پیوند در دسته‌بندی، انتخاب ویژگی در داده‌ی آموزش است که می‌تواند باعث کاهش کارایی پیش‌بینی پیوند در مرحله آزمون شود. از آنجا که ویژگی‌های با رویکرد سراسری پیچیدگی زمانی و محاسباتی زیادی دارند لذا از ویژگی‌های سراسری استفاده نمی‌شود. در این تحقیق تلاش شده برای انجام دسته‌بندی از ویژگی‌ها و معیارهای شباهت محلی مبتنی بر گره‌های یک گراف شبکه اجتماعی استفاده شود. ویژگی‌ها و معیارهایی که در حوزه مسئله پیش‌بینی پیوند مطرح هستند، هر یک به‌عنوان یکی از الگوریتم‌های پیش‌بینی پیوند محسوب می‌شوند، در این تحقیق از این معیارها به‌عنوان معیار شباهت جفت گره‌های یک شبکه اجتماعی در دسته‌بندها استفاده می‌شود. تعداد ویژگی‌هایی که در این تحقیق استفاده شده حدود ۱۱ ویژگی است که در ادامه تشریح می‌شود.

Archive of SID

www.SID.ir

۴-۱-۱- معیارهای محلی و مبتنی بر همسایگی گره :

گراف شبکهٔ اجتماعی، گراف‌هایی هستند که کاربران شبکه و روابط بین آن‌ها را نشان می‌دهند. به‌طور رسمی یک گراف $G=(V,E)$ شامل یک مجموعه گره V و یک مجموعه از پیوند E است. تعداد عناصر V و E به ترتیب با $n=|V|$ و $m=|E|$ مشخص می‌شود که n تعداد گره‌ها و m تعداد پیوندهای گراف است.

این دسته از روش‌ها بر این ایده استوارند که بین دو گره x و y در آینده با احتمال بیشتری پیوند ایجاد خواهد شد اگر مجموعه همسایه‌های آن‌ها دارای اشتراک زیادی باشند. بر اساس این ایده تعدادی معیار شباهت تعریف‌شده‌اند که عبارت‌اند از:

دوستان مشترک: معیار دوستان مشترک[۱۰] [۱۵] یا دوست‌ِدوست[۱۱] [۵] [۱۶] مستقیم‌ترین پیاده‌سازی ایده بالا خواهد بود که در شبکه‌هایی نظیر فیس‌بوک استفاده می‌شود. در این روش گره‌هایی که با طول مسیر ۲ به هم مرتبط هستند، کاندایدای پیشنهاد دوست در نظر گرفته شده و پیشنهاد می‌شوند و هرچه گره‌ها تعداد دوستان مشترک بیشتری داشته باشند، شباهت آن دو بیشتر خواهد بود ودر نتیجه با احتمال بیشتری در آینده پیوند برقرار می‌کنند. Γ (x) در رابطه (۲) نشان‌دهنده دوستان گره x است.

$$Score(x,y) = |\Gamma(x) \cap \Gamma(y)| \quad (۲)$$

ضریب جاکارد[۱۲]: ضریب جاکارد [۵, ۱۵, ۱۶] یکی از پرکاربردترین معیارها در حوزه بازیابی اطلاعات است و به‌صورت نسبت تعداد دوستان مشترک دو گره به مجموع همسایه‌های آن دو تعریف می‌شود. این معیار که از معیار رابطه (۳) محاسبه می‌شود، هرچه به یک نزدیک‌تر باشد، نشان‌دهنده شباهت بیشتر دو گره موردبررسی است.

$$Score(x,y) = |\Gamma(x) \cap \Gamma(y)|/|\Gamma(x) \cup \Gamma(y)| \quad (۳)$$

معیار آدامیک و آدار[۱۳]: معیار آدامیک آدار [۱۵-۱۷] برای یافتن ارتباطات قوی بین صفحات وب به‌کار می‌رود و مربوط به تعداد ویژگی‌های مشترکی است که دو صفحه به اشتراک گذاشته‌اند که در مسئله پیش‌بینی پیوند، این ویژگی مشترک همان دوستان مشترک دو گره است. میزان شباهت دو گره در این معیار از رابطه (۴) محاسبه می‌شود:

$$Score(x,y) = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{log|\Gamma(z)|} \quad (۴)$$

$|\Gamma(z)|$ درجه گره z است و z دوست مشترک x و y است.

اتصال ترجیحی[۱۴]: در اتصال ترجیحی [۱۵, ۱۶, ۱۸] (به‌عنوان مدلی برای رشد شبکه‌ها)، احتمال ایجاد پیوند جدید به گره متناسب با Γ(x)هست. با توجه به رابطه (۵) شباهت دو گره x و y عبارت است از:

$$Score(x,y) = |\Gamma(x)| \times |\Gamma(y)| \quad (۵)$$

اندیس ترفیع‌هاب[۱۵]: اندیس ترفیع‌هاب (HPI) [۱۹] برای تعیین کیفیت هم‌پوشانی توپولوژیک جفت لایه‌ها در یک شبکه دگرگون شونده به کار می‌رود و به‌صورت رابطه (۶) تعریف می‌شود:

$$HPI(x,y) = \frac{|\Gamma(x)\cap\Gamma(y)|}{min\{|\Gamma(x)|,|\Gamma(y)|\}} \quad (۶)$$

اندیس فشرده‌ی‌هاب[۱۶]: اندیس فشرده‌ی‌هاب (HDI) [۱۹] مشابه اندیس بالاست، با این تفاوت که مقدار حداکثر درجه‌ها را در نظر می‌گیرد:

$$HDI(x,y) = \frac{|\Gamma(x)\cap\Gamma(y)|}{max\{|\Gamma(x)|,|\Gamma(y)|\}} \quad (۷)$$

معیار وزن پیوند[۱۷]: معیار وزن پیوند [۱۸] ابتدا به‌صورت دو ویژگی جداگانه برای هر یک از دو پیوند طبق رابطه (۸) و (۹) محاسبه می‌شود:

$$w(x) = \frac{1}{\sqrt{1+|\Gamma(x)|}} \quad (۸)$$

$$w(y) = \frac{1}{\sqrt{1+|\Gamma(y)|}} \quad (۹)$$

سپس وزن پیوند بین دو گره u و v می‌تواند بصورت جمع وزن‌ها یا ضرب وزن‌ها محاسبه‌شده و مورد بهره‌برداری قرار گیرد:

مجموع وزن گره‌ها: جمع وزن‌ها [۱۸] برابر است با جمع دو وزن تعریف‌شده در رابطه (۱۰):

$$w(x,y) = w(x) + w(y) \quad (۱۰)$$

حاصلضرب وزن گره‌ها: ضرب وزن‌ها [۱۸] برابر است با ضرب دو وزن تعریف‌شده در طبق رابطه (۱۱):

$$w(x,y) = w(x) \times w(y) \quad (۱۱)$$

همبستگی درجه[۱۸]**:** همبستگی درجه [۲۰] متناسب با ضریب همبستگی پیرسون است که به‌صورت رابطه (۱۲) تعریف می‌شود:

$$Degree\ Corelation(x,y) = \frac{4.|\Gamma(x)|.|\Gamma(y)|-|\Gamma(x)|-|\Gamma(y)|}{2.|\Gamma(x)|^2+2.|\Gamma(y)|^2-|\Gamma(x)|-|\Gamma(y)|} \quad (۱۲)$$

شاخص سالتون[۱۹]: شاخص سالتون [۱۵] تعداد همسایه‌های مشترک نسبت به میانگین هندسی آنها را با توجه به رابطه (۱۳) محاسبه می‌کند:

$$Score(x,y) = \frac{|\Gamma(x)\cap\Gamma(y)|}{\sqrt{|\Gamma(x)|.|\Gamma(y)|}} \quad (۱۳)$$

شاخص سورنسون[۲۰]: شاخص سورنسون [۱۵] تعداد همسایه‌های مشترک نسبت به میانگین حسابی آنها را با توجه به رابطه (۱۴) محاسبه می‌کند:

$$Score(x,y) = \frac{2|\Gamma(x)\cap\Gamma(y)|}{|\Gamma(x)|+|\Gamma(y)|} \quad (۱۴)$$

۴-۲- نمونه برداری

نمونه‌برداری یکی از روش‌های اصلی جهت انتخاب دادها است. یکی دیگر از چالش‌های پیش‌بینی پیوند در دسته‌بندی، نامتعادل بودن کلاس‌ها [۲۱] در داده‌ی آموزش است که می‌تواند باعث کاهش کارایی پیش‌بینی پیوند در مرحله آزمون شود. نامتعادل بودن کلاس به این مفهوم است که اکثریت نمونه‌های متعلق به یک کلاس با برچسب یکسان می‌باشد؛ درواقع پیوندهایی که یک گره با سایر رئوس تشکیل می‌دهد درصد بسیار کمی از کل پیوندهای موجود در شبکه را شامل می‌شود. این مسئله، آموزش و استنتاج در این شبکه‌ها را با مشکل جدی مواجه می‌سازد [۵].

دسته‌بندهای پایه‌ای که در این تحقیق استفاده‌شده نسبت به نامتعادل بودن داده بسیار حساس هستند. برای حل این مشکل روش‌های مختلفی پیشنهادشده [۲۲, ۲۳] به عنوان مثال با اضافه کردن نمونه به کلاس اقلیت یا کاهش مجموعه داده‌های کلاس اکثریت تعادلی میان تعداد اعضای کلاس‌های مختلف برقرار می‌شود. روشی که در این تحقیق برای رفع مشکل نامتعادل بودن کلاس‌ها استفاده می‌کند با حذف داده‌های کلاس اکثریت انجام می‌پذیرد. از میان پیوندهای مثبت (موجود) یا منفی (غیر موجود) در

Archive of SID

www.SID.ir

مجموعه داده، دودسته مساوی از پیوندهای مثبت و منفی در نظر گرفته و همه آزمایش‌ها، بر روی این مجموعه انتخابی، انجام می‌شود.

۴-۳- انتخاب دسته بند

انتخاب دسته‌بندی کننده‌ها، شاید یکی از مشهورترین بخش‌های انتخاب در ترکیب دسته‌بندی کننده‌ها باشد. از الگوریتم‌های یادگیر با ناظر مبتنی بر درخت تصمیم که بطور گسترده در مقالات بکار گرفته شده [۲۴] استفاده خواهد شد. تجربه نشان داده که الگوریتم‌های درخت تصمیم در Weka نسبت به درخت تصمیم در RapidMiner از دقت بالاتری برخوردار است. بدین ترتیب از قابلیت‌های کلیدی نرم‌افزار RapidMiner همراه با الگوریتم‌های قدرتمند مدل سازی نرم‌افزار Weka به صورت توأم استفاده می‌شود تا به دقت بالاتری دست یافت. سه دسته‌بند W-RepTree، W-Simple Cart، W-Random Forest با توجه به مقالات [۱۰, ۲۴, ۲۵] و شرایط مسئله و ویژگی‌های موجود که بهترین پاسخ‌ها و کارایی قابل قبولی نسبت به سایر دسته‌بندی کننده‌ها ارائه داده‌اند انتخاب شده است.

- W-RepTree

این مدل یک درخت تصمیم با یادگیری سریع است که با استفاده ازبهره اطلاعات[۲۱] درخت را می‌سازد [۲۶].

- W-Random Forest

این مدل مجموعه‌ای از درختان تصادفی یا یک جنگل تصادفی را یاد می‌گیرد. مدل حاصل نتیجه رأی‌گیری از تمام درختان خواهد بود [۲۴].

- W-Simple Cart

CART یک درخت تصمیم است که علاوه بر قدرت بالا، از ساختار ساده و قابل درکی برخوردار است. این درخت برای دسته‌بندی، رکوردهای اعمال شده به مدل، از بالا به پایین درخت را پیمایش می‌کند. درختان بدون استفاده از یک قانون متوقف‌کننده به رشد حداکثری خود می‌رسند و سپس اصلاح می‌شوند. هدف CART تولید تنها یک درخت نیست بلکه تولید یک سری درختان اصلاح شده است [۲۴].

۴-۴- ترکیب دسته بندها

فرایند ترکیب دسته‌بندها یکی دیگر از مراحل کار می‌باشد که از میان روش‌های موجود، روش رأی‌گیری حداکثری را به‌عنوان مبنای کار قرار گرفته است. در واقع این تکنیک جهت تخمین برچسب یک نمونه جدید، برچسب پیشنهادی هر یک از دسته‌بندهای پایه را به عنوان یک رأی انتخاب نموده و در نهایت با شمارش آراء، رأی اکثریت به عنوان برچسب کلاس نمونه موردنظر انتساب می‌کند. استفاده از این تکنیک، افزایش قابل ملاحظه در دقت مدل می‌شود و همچنین مدل در مواجهه با مجموعه داده‌های دارای نویز یا مقادیر مفقوده نیز مقاوم‌تر خواهد شد.

۵- ارزیابی

در هر دو مرحله اعم از مرحله یادگیری و مرحله ارزیابی، برای تفکیک داده‌های آموزشی و آزمایشی از روش 10-Fold Cross Validation استفاده شده است. بدیهی است که در این شیوه ارزیابی، دقت محاسبه شده برای دسته بند قابل اعتماد بوده و دانش حاصل شده جامع خواهد بود و البته افزایش زمان ارزیابی دسته بند نیز مهم‌ترین مشکل آن می‌باشد.

برای هر یک از ویژگی‌ها، با محاسبه معیار استانداردی به نام AUC[۲۲] ، کارایی روش‌های دسته‌بندی، موردبررسی قرار گرفته است. AUC معادل ناحیه زیر منحنی ROC[۲۳] است [۱۶] که هر چه مقدار این عدد مربوط به یک دسته بند بزرگ‌تر باشد کارایی نهایی دسته بند مطلوب‌تر ارزیابی می‌شود.

۵-۱- مجموعه داده‌های مورد استفاده

از مجموعه داده‌های بدون جهت دو شبکهٔ اجتماعی "Hi5" و "Facebook" به منظور بررسی کارایی الگوریتم استفاده شده است . این مجموعه داده توسط [۱۶] در دو بازه زمانی در تاریخ ۳۰ اکتبر ۲۰۰۹ و ۱۵ دسامبر ۲۰۱۰ به روش خزش روی گراف شبکه اجتماعی Facebook و Hi5 جمع‌آوری شده است. مشخصات کامل این مجموعه داده در جدول (۱) آمده است.

N : تعداد گره‌ها E : تعداد پیوندها

جدول (۱) : مشخصات مجموعه داده [۱۶]

E	N	Type	Data Set
۱۳۶۹۲	۳۶۹۴	Undirected Unsigned	Facebook 3.7 K
۸۸۲۶۱	۶۳۳۲۹	Undirected Unsigned	Hi5 63K

۵-۲- نتایج

همهٔ ۱۱ ویژگی‌هایی که پیش از این مطرح شده با بهره‌گیری از مجموعه داده Facebook و Hi5 مورد بررسی قرار داده و نتایج AUC حاصل از استفاده هر یک از این ویژگی‌ها به تنهایی در دسته‌بندی و استفاده از همه ویژگی‌ها در دسته‌بندی در جدول (۲) ارائه شده است.

جدول (۲) : نتایج AUC استفاده از هر ویژگی‌ها در دسته‌بند

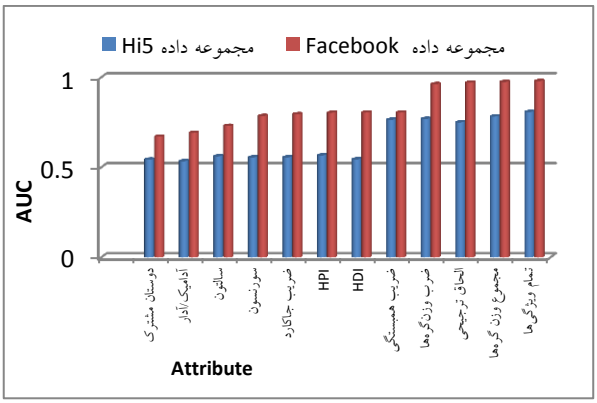
ویژگی	مجموعه داده Facebook	مجموعه داده Hi۵
دوستان مشترک	۰,۶۷۱	۰,۵۴۴
آدامیک/آدار	۰,۶۹۲	۰,۵۳۵
سالتون	۰,۷۳۱	۰,۵۶۲
سورنسون	۰,۷۸۶	۰,۵۵۶
ضریب جاکارد	۰,۷۹۶	۰,۵۵۶
HPI	۰,۸۰۴	۰,۵۶۷
HDI	۰,۸۰۵	۰,۵۴۵
ضریب همبستگی	۰,۸۰۵	۰,۷۶۵
ضرب وزن‌گره‌ها	۰,۹۶۳	۰,۷۷
الحاق ترجیحی	۰,۹۷۱	۰,۷۵
مجموع وزن گره‌ها	۰,۹۷۵	۰,۷۸۳
تمام ویژگی‌ها	**۰,۹۸**	**۰,۸۰۸**

جدول (۲) نشان می‌دهد زمانی که هر کدام از ویژگی‌ها به تنهایی در دسته‌بند مورداستفاده قرار می‌گیرند میزان AUC دسته‌بند پایین می‌باشد، در حالیکه در روش پیشنهادی با در نظر گرفتن کلیه ویژگی‌ها، علاوه بر پوشش ضعف‌های یکدیگر، دسته‌بند نیز کارایی قابل قبولی از خود نشان می‌دهد.

Archive of SID

www.SID.ir

شکل (۱) بهبود کارایی دسته‌بند چندگانه را با درنظر گرفتن کلیه ویژگی‌ها را نشان می‌دهد.



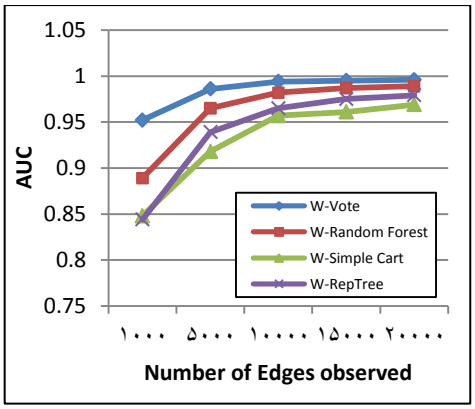
شکل(۱): نتایج AUC دسته‌بند با هر ویژگی

جدول (۳) مقایسه نتایج حاصل از هر یک دسته‌بندهای پایه و دسته‌بند چندگانه را نشان می‌دهد.

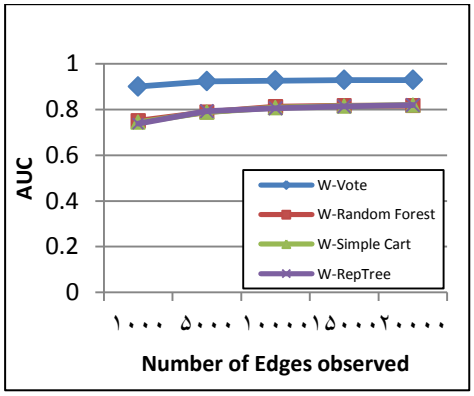
جدول (۳) : نتایج AUC دسته‌بندچندگانه با افزایش تعداد پیوند

تعداد پیوند	الگوریتم	مجموعه داده Facebook	مجموعه داده Hi5
۱۰۰۰	W-RepTree	۰,۸۴۴	۰,۷۵۱
	W-Simple Cart	۰,۸۴۸	۰,۷۴۴
	W-Random Forest	۰,۸۸۹	۰,۷۳۸
	W-Vote	**۰,۹۵۲**	**۰,۹**
۵۰۰۰	W-RepTree	۰,۹۳۹	۰,۷۸۹
	W-Simple Cart	۰,۹۱۸	۰,۷۸۸
	W-Random Forest	۰,۹۶۵	۰,۷۹۳
	W-Vote	**۰,۹۸۶**	**۰,۹۲۳**
۱۰۰۰۰	W-RepTree	۰,۹۶۵	۰,۸۱۳
	W-Simple Cart	۰,۹۵۷	۰,۸۰۸
	W-Random Forest	۰,۹۸۲	۰,۸۰۵
	W-Vote	**۰,۹۹۴**	**۰,۹۲۶**
۱۵۰۰۰	W-RepTree	۰,۹۷۵	۰,۸۱۷
	W-Simple Cart	۰,۹۶۱	۰,۸۱۵
	W-Random Forest	۰,۹۸۷	۰,۸۱۴
	W-Vote	**۰,۹۹۵**	**۰,۹۲۹**
۲۰۰۰۰	W-RepTree	۰,۹۷۹	۰,۸۱۷
	W-Simple Cart	۰,۹۶۹	۰,۸۱۹
	W-Random Forest	۰,۹۸۹	۰,۸۱۹
	W-Vote	**۰,۹۹۶**	**۰,۹۲۹**

همان‌گونه که در جدول (۳) مشاهده می‌شود، مقدار AUC حاصل از دسته‌بندچندگانه که با رأی‌گیری بین دسته‌بندهای پایه بدست آمده نسبت به هر یک از دسته‌بندهای پایه بالاتر بوده و با افزایش تعداد پیوندهای مشاهده شده از شبکه نیز بهبود یافته است به گونه ای که با ۲۰۰۰۰ پیوند مشاهده شده از شبکه، کارایی دسته‌بند چندگانه در مجموعه داده Facebook حدود ۰,۹۹۶ و در مجموعه داده Hi5 حدود ۰,۹۲۹ می‌باشد که این بهبود نتیجه نهایی پس از ترکیب نتایج ۳ دسته‌بند نشان‌دهنده کارایی مناسب روش ارائه‌شده در این تحقیق است. شکل (۳) و (۴) مصدق این مطلب است.



شکل (۲): مقایسه AUC دسته‌بندچندگانه ودسته‌بندهای پایه با مجموعه داده Facebook



شکل (۳):مقایسه AUC دسته‌بندچندگانه ودسته‌بندهای پایه با استفاده از مجموعه داده Hi5

۶- نتیجه گیری

اجرای الگوریتم‌های یادگیری دسته جمعی مبتنی بر سیستم‌های دسته‌بند چندگانه، روشی است برای آنکه بتوان تقریب بهتری از یک دسته‌بند بهینه را فراهم کرد. در این مقاله به‌منظور پیش‌بینی پیوندهای آتی در یک شبکه اجتماعی، از روش یادگیری گروهی استفاده شده که با ترکیب چند دسته بند پایه و ایجاد یک دسته‌بند چندگانه محتمل‌ترین پیوندهای آتی در شبکه به‌طور کارا پیش‌بینی می‌شود . بر اساس ارزیابی‌ای که با استفاده از مجموعه داده‌های دو شبکهٔ اجتماعیHi5 و Facebook انجام شده است، نتایج نشان می‌دهد که روش پیشنهادی با بهره‌گیری از معیارهای مشابهت محلی و با استفاده از روش‌های یاگیری جمعی می‌تواند پیوندهای بین موجودیت‌های یک شبکه اجتماعی را به گونه‌ای پیش‌بینی نماید که نسبت به دسته‌بندهای پایه از کارایی خوبی برخوردار بوده و دقت پیش‌بینی را می‌تواند افزایش دهد.

مراجع

[۱] Lü, L. and Zhou, T., "Link prediction in complex networks: A survey," *Physica A: Statistical Mechanics and its Applications*, vol. 390, pp. 1150-1170, 2011.

[۲] Huang, Z. and Lin, D. K., "The time-series link prediction problem with applications in communication surveillance," *INFORMS Journal on Computing*, vol. 21, pp. 286-303, 2009.

Archive of SID

www.SID.ir

[۳] Liben‐Nowell, D. and Kleinberg, J., "The link‐prediction problem for social networks," *Journal of the American society for information science and technology*, vol. 58 ,pp. 1019-1031, 2007.

[۴] Yin, D., Hong, L., and Davison, B. D., "Structural link analysis and prediction in microblogs," in *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 1163-1168, 2011.

[۵] Al Hasan, M .and Zaki, M. J., "A survey of link prediction in social networks," in *Social network data analytics*, ed: Springer, pp. 243-275, 2011.

[۶] Getoor, L. and Diehl, C. P., "Link mining: a survey," *ACM SIGKDD Explorations Newsletter*, vol. 7, pp. 3-12, 2005.

[۷] Li, R.-H., Yu, J. X., and Liu, J., "Link prediction: the power of maximal entropy random walk," in *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 1147-1156, 2011.

[۸] Volkova, S., "LINK PREDICTION IN SOCIAL NETWORKS," 2009.

[۹] Backstrom, L. and Leskovec, J., "Supervised random walks: predicting and recommending links in social networks," in *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 635-644, 2011.

[۱۰] Lichtenwalter, R. N., Lussier, J. T., and Chawla, N. V., "New perspectives and methods in link prediction," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 243-252, 2010.

[۱۱] Fire, M., Tenenboim, L ,.Lesser, O., Puzis, R., Rokach, L., and Elovici, Y., "Link prediction in social networks using computationally efficient topological features," in *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pp. 73-80, 2011.

[۱۲] Sherkat, E., Rahgozar, M., and Asadpour, M., "Structural link prediction based on ant colony approach in social networks," *Physica A: Statistical Mechanics and its Applications*, vol. 419, pp. 80-94, 2015.

[۱۳] Dietterich, T. G., "Machine-learning research," *AI magazine*, vol. 18, p. 97, 1997.

[۱۴] Boyd, D. M. and Ellison, N. B., "Social network sites: definition, history, and scholarship," *Engineering Management Review, IEEE* ,vol. 38, pp. 16-31, 2010.

[۱۵] Bliss, C. A., Frank, M. R., Danforth, C. M., and Dodds, P. S., "An evolutionary algorithm approach to link prediction in dynamic social networks," *Journal of Computational Science*, vol. 5, pp. 750-764, 2014.

[۱۶] Papadimitriou, A., Symeonidis, P., and Manolopoulos, Y., "Fast and accurate link prediction in social networking systems," *Journal of Systems and Software*, vol. 85, pp. 2119-2132, 2012.

[۱۷] Adamic, L. and Adar, E., "How to search a social network," *Social Networks* ,vol. 27, pp. 187-203, 2005.

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

[۱۹] Zhu, Y.-X., Lü, L., Zhang ,Q.-M., and Zhou, T., "Uncovering missing links with cold ends," *Physica A: Statistical Mechanics and its Applications*, vol. 391, pp. 5769-5778, 2012.

[۲۰] Feyessa, T., Bikdash, M., and Lebby, G., "Node-pair feature extraction for link prediction," in *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pp. 1421-1424, 2011.

[۲۱] Vorraboot, P., Rasmequan, S., Chinnasarn, K., and Lursinsap, C" ,.Improving classification rate constrained to imbalanced data between overlapped and non-overlapped regions by hybrid algorithms," *Neurocomputing*, vol. 152, pp. 429-443, 2015.

[۲۲] Tsai, C.-F. and Chang, C.-W., "SVOIS: support vector oriented instance selection for text classification," *Information Systems*, vol. 38, pp. 1070-1083, 2013.

[۲۳] Yang, Z. and Gao, D., "An active under-sampling approach for imbalanced data classification," in *Computational Intelligence and Design (ISCID), 2012 Fifth International Symposium on*, 2012, pp. 270-273.

[۲۴] Gokgoz, E. and Subasi, A., "Comparison of decision tree algorithms for EMG signal classification using DWT," *Biomedical Signal Processing and Control*, vol. 18, pp. 138-144, 2015.

[۲۵] Song, H. H., Cho, T. W., Dave ,V., Zhang, Y., and Qiu, L., "Scalable proximity estimation and link prediction in online social networks," in *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, pp. 322-335, 2009.

[۲۶] Tanha, J., van Someren, M., and Afsarmanesh, H., "Semi-supervised self-training for decision tree classifiers," *International Journal of Machine Learning and Cybernetics*, pp. 1-16, 2015.

زیرنویس‌ها

1 Protein-Protein Interaction (PPI) Network
2 Link Mining
3 Supervised
4 Unsupervised
5 Feature Selection
6 Imbalanced Classes
7 Random Walk
8 Multiple Classifier System (MCS)
9 Classifier Ensembles
10 Common Friends
11 FOAF
12 Jaccard's Coefficient (JC)
13 Adamic/Adar Index (AA)
14 Preferencial Attachment
15 Hub Promoted Index
16 Hub Depressed Index
17 Edge weight
18 Degree Correlation
19 Salton index
20 Sorenson index
21 Information gain
22 Area Under Curve
23 Receive Oprating Characteristic

Archive of SID

www.SID.ir