دانشکده مهندسی کامپیوتر

پیاده‌سازی و کاربست یک الگوریتم جاسازی گراف برای جاسازی شبکه‌ی اندرکنش ناهمگن دارو ـ پروتئینی برای پیش‌بینی دارو ـ هدف

پروژه کارشناسی مهندسی کامپیوتر گرایش مهندسی نرم افزار

فاطمه فتاحی

استاد راهنما

دکتر مینایی

شهریور ۱۳۹۸


بسم الله الرحمن الرحيم

تأییدیه‌ی هیأت داوران جلسه‌ی دفاع از پروژه

نام دانشکده: دانشکده مهندسی کامپیوتر

نام دانشجو: فاطمه فتاحی

عنوان پروژه : پیاده‌سازی و کاربست یک الگوریتم جاسازی گراف برای جاسازی شبکه‌ی اندرکنش ناهمگن دارو_ پروتئینی برای پیش‌بینی دارو_ هدف

تاریخ دفاع: شهریور ۱۳۹۸

رشته: مهندسی کامپیوتر

گرایش: مهندسی نرم افزار

ردیف	سمت	نام و نام خانوادگی	مرتبه دانشگاهی	دانشگاه یا مؤسسه	امضـــــــــا
۱	استاد راهنما				
۲	استاد مشاور				
۳	استاد مدعو خارجی				
۴	استاد خارجی				
۵	استاد مدعو داخلی				
۶	استاد مدعو داخلی				
۷	استاد مدعو داخلی				

تأییدیه‌ی صحت و اصالت نتایج

باسمه تعالی

اینجانب فاطمه فتاحی به شماره دانشجویی **۹۴۵۲۱۱۳۵** دانشجوی رشته مهندسی کامپیوتر مقطع تحصیلی کارشناسی تأیید می‌نمایم که کلیه‌ی نتایج این پروژه حاصل کار اینجانب و بدون هرگونه دخل و تصرف است و موارد نسخه‌برداری‌شده از آثار دیگران را با ذکر کامل مشخصات منبع ذکر کرده‌ام. درصورت اثبات خلاف مندرجات فوق، به تشخیص دانشگاه مطابق با ضوابط و مقررات حاکم (قانون حمایت از حقوق مؤلفان و مصنفان و قانون ترجمه و تکثیر کتب و نشریات و آثار صوتی، ضوابط و مقررات آموزشی، پژوهشی و انضباطی ...) با اینجانب رفتار خواهد شد و حق هرگونه اعتراض درخصوص احقاق حقوق مکتسب و تشخیص و تعیین تخلف و مجازات را از خویش سلب می‌نمایم. در ضمن، مسؤولیت هرگونه پاسخگویی به اشخاص اعم از حقیقی و حقوقی و مراجع ذی‌صلاح (اعم از اداری و قضایی) به عهده‌ی اینجانب خواهد بود و دانشگاه هیچ‌گونه مسؤولیتی در این خصوص نخواهد داشت.

نام و نام خانوادگی: فاطمه فتاحی

تاریخ و امضا:

مجوز بهره‌برداری از پایان‌نامه

بهره‌برداری از این پایان‌نامه در چهارچوب مقررات کتابخانه و با توجه به محدودیتی که توسط استاد راهنما به شرح زیر تعیین می‌شود، بلامانع است:

☐ بهره‌برداری از این پایان‌نامه برای همگان بلامانع است.

☐ بهره‌برداری از این پایان‌نامه با اخذ مجوز از استاد راهنما، بلامانع است.

☐ بهره‌برداری از این پایان‌نامه تا تاریخ ..... ممنوع است.

استاد راهنما: دکتر مینایی

تاریخ:

امضا:

فاطمه فتاحی

شهریور ۱۳۹۸

چکیده

پیش بینی ارتباطات دارو_هدف به عنوان یک مرحله بنیادی در فرایند کشف دارو و استفاده مجدد از آن محسوب می شود. روش های آزمایشگاهی در این حوزه، علی رغم داشتن نتایج دقیق و سودمند بسیار هزینه بر و زمان گیر هستند. در چند دهه ی اخیر، روش های محاسباتی جدیدی به منظور حل این مسئله مطرح شده اند که علاوه بر داشتن نتایج بسیار سودمند، به طور قابل توجهی کم هزینه بوده و زمان کمتری را نسبت به روش های آزمایشگاهی در بر می گیرد. از جمله ی این روش ها که نتایج حاصل از آن بسیار قابل توجه ارزیابی شده است، پیش بینی روابط بر پایه ی شبکه [1] در شبکه های زیست پزشکی می باشد.
امروزه، روش های مبتنی بر شبکه متعددی برای پیش بینی انواع روابط میان داروها و پروتئین ها پیشنهاد گردیده است. با وجود عملکرد بسیار ارزنده آن ها، این راهکارها تنها برای شبکه های همگن [2] به خوبی کار می کنند. این روش ها تنها اطلاعات مربوط به گره [3] های شبکه را در اختیار ما می گذارند و مفاهیم یال [4] ها را در بر نمی گیرند. در این پروژه، ما با استفاده از روش ادج تو وک [5]، به جاسازی نودها در شبکه ناهمگن [6] دارو_پروتئینی پرداخته ایم و سپس با به کارگیری الگوریتم ماشین بردار پشتیبان [7]، ارتباطات دارو_هدف را پیش بینی نموده ایم. نتایج به دست آمده حاصل از این روش نسبت به سایر روش های موجود مطرح شده در این حوزه به طور چشمگیری بالاتر است.

واژگان کلیدی: پیش بینی ارتباطات دارو_هدف، جاسازی نودها، اج تو وک، ماشین بردار پشتیبان، شبکه ناهمگن

---

[1] Network
[2] Homogeneous
[3] Node
[4] Edge
[5] Edge2vec
[6] Heterogeneous
[7] Support Vector Machine

فهرست مطالب

فهرست تصاویر چ

فهرست جداول ح

فهرست الگوریتم‌ها خ

فهرست علائم اختصاری د

فصل ۱: مقدمه ۱

۱ـ۱ مقدمه ۲

فصل ۲: مروری بر منابع ۴

۲ـ۱ مقدمه ۵

۲ـ۲ تعاریف، اصول و مبانی نظری ۵

۲ـ۲ـ۱ شبکه ۵

۲ـ۳ مروری بر ادبیات موضوع ۶

۲ـ۳ـ۱ جاسازی نودها در شبکه ۶

۲ـ۴ نتیجه گیری ۹

فصل ۳: روش تحقیق ۱۱

۳ـ۱ مقدمه ۱۲

۳ـ۲ تشریح کامل روش تحقیق ۱۲

۳ـ۲ـ۱ الگوریتم ادج تو وک ۱۳
۳ـ۲ـ۲ روش ماشین بردار پشتیبان ۱۴

فصل ۴: نتایج و تفسیر آنها ۱۵
۴ـ۱ مقدمه ۱۶
۴ـ۲ محتوا ۱۶
۴ـ۲ـ۱ مجموعه داده ها ۱۶
۴ـ۲ـ۲ ارزیابی نتایج ۱۶

فصل ۵: جمع بندی و پیشنهادها ۲۰
۵ـ۱ مقدمه ۲۱
۵ـ۲ جمع بندی ۲۱
۵ـ۳ نوآوری ۲۱
۵ـ۴ پیشنهادها ۲۱

مراجع ۲۳

واژه‌نامه فارسی به انگلیسی ۲۵

واژه‌نامه انگلیسی به فارسی ۲۷

فهرست تصاویر

۲ـ۱ جاسازی نودها در شبکه ۶

۲ـ۲ استراتژی پیاده روی تصادفی ۸

۲ـ۳ چگونگی تعریف متاپت برای یک شبکه زیست پزشکی ۱۰

۳ـ۱ مراحل روش پشنهادی ۱۳

۴ـ۱ نمودارهای مربوط به تنظیم پارامتر های الگوریتم ادج تو وک ۱۷

۴ـ۲ مقادیر بازخوانی، دقت، امتیاز اف ۱ و خطای همینگ به ازای مقادیر مختلف دو پارامتر پی و آر ۱۸

۴ـ۳ مقایسه میزان عملکرد میان روش های موجود با روش پیشنهادی ۱۹

فهرست جداول

فهرست الگوریتم‌ها

فهرست علائم اختصاری

فصل ۱

مقدمه

۱ـ۱ مقدمه

پیش بینی ارتباطات دارو_هدف پروتئینی به عنوان یک مرحله ضروری و مهم در حوزه پیدایش دارو به شمار می رود. در حقیقت این امر موجب ظهور دارو ها و هم چنین اهداف جدید برای دارو های موجود می شود. در چند دهه ی اخیر، روش های آزمایشگاهی بسیار متعددی در این موضوع مطرح گردیده است. اما اکثر این روش ها بسیار هزینه بر بوده و از لحاظ زمانی نیز وقت به خصوصی را می گیرد. آمار و ارقام نشان دهنده این است که برای تولید یک داروی جدید به طور متوسط، نیاز به ۱۰ تا ۱۵ سال زمان و حدود ۸۰۰ میلیون دلار سرمایه است[۳].

با در نظر گرفتن این مسئله، روش های محاسباتی می توانند جایگزین مناسبی برای آن ها باشند. این روش ها علاوه بر پشت سر گذاشتن مشکلات ذکر شده، دقت و عملکرد بسیار خوبی نیز ارائه می کنند.

امروزه با افزایش تنوع و ناهمگونی دارو ها و روابط میان شان با سایر عناصر، استفاده از این روش های محاسباتی بسیار ارزشمند تلقی می شود. در چند سال اخیر، تلاش های بسیاری به منظور تشخیص این چنین روابطی با استفاده از این روش ها انجام شده است. تقریبا در تمام این روش ها فرض بر آن است که دارو های با ویژگی های مشابه، هدف های مشابهی نیز دارند.

از آن جا که دانش زیست پزشکی مدام در حال رشد و پیچیدگی است و ما با تعداد بسیار زیادی مدل و روابط بینشان روبه رو هستیم، باید از یک مدل داده ای که بتواند به خوبی آن ها را توصیف نماید استفاده نماییم که شبکه نام دارد. در این مدل ما با انواع گره ها و یال ها که هر کدام اطلاعات مخصوص به خود را نگاه می دارند در ارتباطیم.

تقریبا اغلب روش های مطرح شده برای پیش بینی ارتباطات دارو_هدف برای شبکه های همگن و مدل های دو بخشی [1] کاربرد دارند. به تازگی نیز روش هایی برای شبکه های ناهمگن نیز مطرح گردیده است که در این گونه روش ها علاوه بر در نظر گرفتن اطلاعات نودها، مفاهیم مربوط به یال ها را نیز در اختیار داریم.

در این پروژه، روشی برای پیش بینی ارتباطات دارو_هدف و گرفتن دقت و بازخوانی بسیار بالا بر روی یک شبکه متشکل از انواع متنوعی از روابط و موجودیت ها ارائه شده است. هم چنین به تنظیم پارامترهای موجود در الگوریتم ادج تو وک نیز پرداخته ایم و بالاترین نتایج را ذکر نموده ایم.

در بخش های آتی، ابتدا به بررسی کارهای پیشین در این حوزه و سپس به ارائه دقیق روش ارائه شده می

[1] bipartite models

پردازیم. در انتها نتایج را بررسی و به جمع بندی و ارائه ی پیشنهادات می پردازیم.

فصل ۲

مروری بر منابع

۲ـ۱ مقدمه

در این فصل ابتدا به بررسی مفاهیم شبکه و اصول آن و سپس، به بررسی موضوع جاسازی نود ها در شبکه و انواع روش های تعیین شباهت نود می پردازیم. در بخش انواع روش های تعیین شباهت نود، در حقیقت همان روش های پیشین برای پیش بینی ارتباطات دارو_هدف که مبتنی بر شبکه هستند پرداخته خواهد شد.

۲ـ۲ تعاریف، اصول و مبانی نظری

در این بخش ابتدا به بررسی شبکه و کاربرد آن و هم چنین انواع آن می پردازیم.

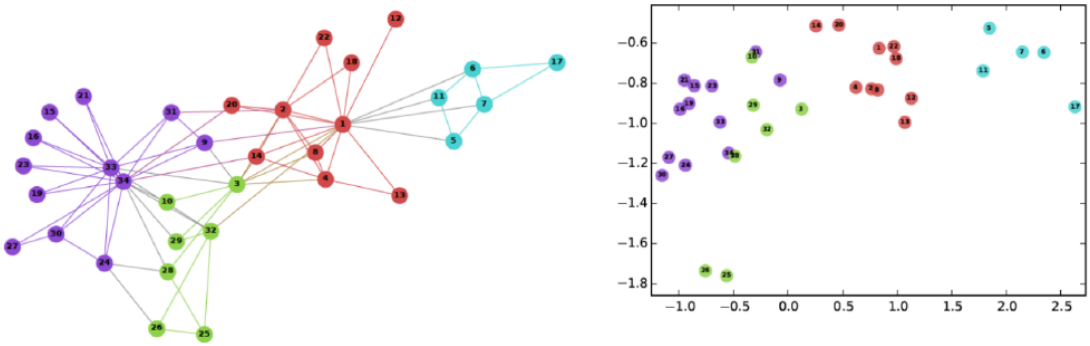
۲ـ۲ـ۱ شبکه

[۱] در بسیاری از مسائل از جمله سیستم های اجتماعی و اقتصادی برای درک و توصیف رفتار و روابط موجود در سیستم از مدل داده ای گراف برای ساده سازی شهود استفاده می گردد. در حقیقت، شبکه ها یک زبان عمومی برای توصیف و مدل سازی سیستم های پیچیده هستند. در علم زیست پزشکی نیز به دلیل پیچیدگی و تنوع روابط و موجودیت ها از این نوع ساختار داده به منظور درک بهتر استفاده می شود.

انواع شبکه

همان طور که می دانیم، ما در بسیاری از شبکه ها تنها با یک نوع نود و رابطه سر و کار داریم مثل رابطه پزشک با بیمار؛ در واقع، به این شبکه ها که در آن ما کاری به نوع یال نداریم شبکه های همگن [۲] می نامیم. اما دسته ی دیگری از شبکه ها وجود دارند که ما در دنیای امروز اغلب با آن ها سر وکار داریم. در این نوع شبکه ها، انواع متنوعی از روابط وجود دارد که ما آن ها را در نظر می گیریم. این دسته را شبکه های ناهمگن [۳] می خوانند.

[1] Network
[2] Homogeneous
[3] Heterogeneous



شکل ۲ـ۱: جاسازی نودها در شبکه

۲ـ۳ مروری بر ادبیات موضوع

در این بخش ابتدا به بررسی مفهوم جاسازی نود ها در گراف و اهمیت آن می پردازیم. سپس به روش های جاسازی نودها در گراف که تاکنون مطرح شده است، اشاره خواهیم کرد.

۲ـ۳ـ۱ جاسازی نودها در شبکه

[۴] همان طور که در بخش قبل توضیح داده شد، در شبکه های زیست پزشکی به دلیل پیچیدگی بسیار زیاد روابط و انواع زیاد نودها نمی توان اطلاعات مهم و مورد نیاز را از آن استخراج نمود. به همین منظور، روشی تحت عنوان جاسازی نود ها در گراف مطرح گردیده است که نود ها را در یک فضای با ابعاد پایین تر تعبیه می کند؛ به گونه ای که اطلاعات مربوط به آن ها از دست نرود. به این ترتیب گره های مشابه در گراف، در فضای جاسازی نیز نزدیک به هم هستند. هدف رمزگذاری گره ها به این گونه است که شباهت در فضای تعبیه شده، شباهت در شبکه ی اصلی را نشان می دهد. در شکل ۲ـ۳ـ۱ یک شبکه مربوط به یک مجموعه داده خاص، در یک فضای جاسازی شده قرار گرفته است.

در این حوزه، روش های بسیاری برای تعیین شباهت نود ها در گراف ایجاد شده است. به طور کلی این روش ها به دسته زیر تقسیم بندی می شوند:

[4] Node embedding

- شباهت مبتنی بر وابستگی[۵]
- شباهت چند گامی[۶]
- روش های پیاده روی تصادفی [۷]

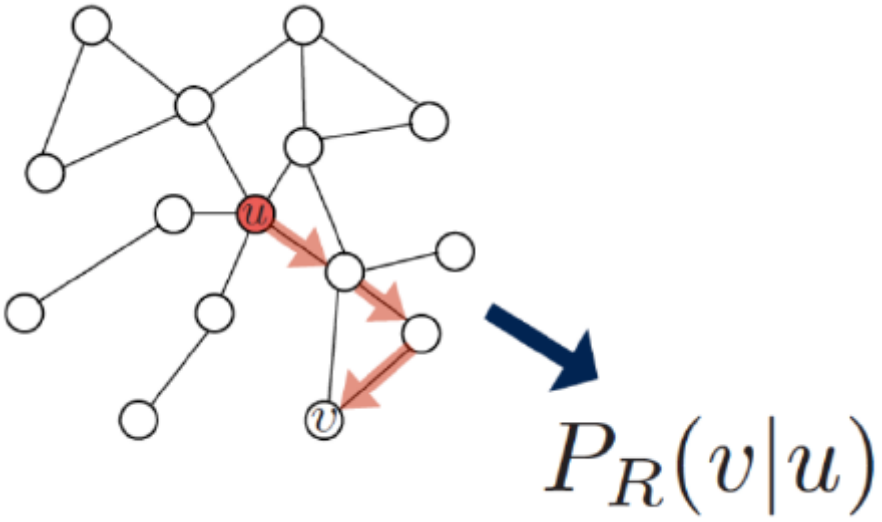
روش های پیاده روی تصادفی به دلیل بهینه بودن از لحاظ پیچیدگی زمانی و سرعت بیشتر آن ها نسبت به دو روش دیگر از توجه بسیاری برخوردارند و دو روش دیگر به ندرت به کار گرفته می شوند.
در روش های پیاده روی تصادفی، احتمال ملاقات یک گره دلخواه در یک پیاده روی تصادفی که از گره مشخصی با استفاده از یک استراتژی [۸] مشخص آغاز می شود، محاسبه می گردد.
مزیت های این روش به صورت زیر می باشد:

- انعطاف پذیری[۹] : علاوه بر اطلاعات همسایه های محلی شامل اطلاعات همسایه های مرتبه های بالاتر نیز می شود.
- کارایی: به دلیل در نظر نگرفتن همه ی نود ها در فرآیند آموزش [۱۰]، پیچیدگی کمتری دارد.

استراتژی های پیاده روی تصادفی

همان گونه که در قسمت قبل مطرح شد، روش های پیاده روی تصادفی با استفاده از یک سری استراتژی های مشخص، به تعیین شباهت نود در گراف می پردازند. در ادامه تعدادی از استراتژی های مطرح شده را توضیح خواهیم داد.
یکی از ساده ترین ایده های مطرح شده توسط پروزی و همکارانش در [۱۱]، استراتژی دیپ واک[۱۱] بود که از هر نود، پیاده روی های تصادفی با طول ثابت و بدون هیچ تأثیری [۱۲] را اجرا می کنیم.
کمی بعد از آن در سال ۲۰۱۶، گرور و لسکوک، این استراتژی را با دو پارامتر پی [۱۳] به معنای پارامتر بازگشت

[5] Adjacency-based Similarity
[6] Multi-hop Similarity
[7] Random Walk Approaches
[8] Strategy
[9] Flexibility
[10] Training
[11] DeepWalk
[12] Unbiased
[13] p



شکل ۲ـ۲: استراتژی پیاده روی تصادفی

[۱۴]و کیو [۱۵]به معنای دور شدن [۱۶]تحت تأثیر قرار دادند و روشی تحت عنوان نود تو وک [۱۷] نتایج با دقت بالاتری را کسب کردند[۶]. در واقع، این دو پارامتر باعث یک توازن [۱۸]بین دید های محلی و جهانی شبکه می شود.

در این میان روش های دیگری نیز بر همین اساس ایجاد شده اند. از جمله این روش ها می توان به [۱] اشاره نمود که بر مبنای وزن های یادگیرنده به وجود آمده است.

استراتژی دیگری که در همین دوره بعد از نود تو وک کانون توجه پژوهشگران قرار گرفت، الگوریتم لاین [۱۹] بود[۱۲].این روش، یک روش بهینه سازی جایگزین نود تو وک محسوب می شود که براساس احتمالات پیاده روی های تصادفی تک گام [۲۰] و دو گام طرح ریزی شده است. این استراتژی از آن جهت بسیار مورد توجه قرار گرفت که بسیار مقیاس پذیر بوده و برای شبکه های بی وزن [۲۱]، وزن دار[۲۲] ، بی جهت[۲۳] و جهت دار[۲۴]

[14]Return parameter
[15]q
[16]Walk away parameter
[17]Node2vec
[18]Trade off
[19]LINE
[20] One-hop
[21]Unweighted
[22]Weighted
[23]Undirected
[24]Diredcted

به خوبی کار می کند. تاکنون روش های متعددی مطرح شده است؛ اما به نظر می رسد این روش عملکرد و نتایج بسیار دقیق تری نسبت به سایر روش های موجود دارد.
مشکلی که در همه ی این روش ها به چشم می خورد این است که تنها برای شبکه های همگن به خوبی عمل می کنند. در حالی که شبکه های زیست پزشکی غالباً به صورت ناهمگن می باشند و لذا این گونه روش ها نمی توانند به خوبی عملیات جاسازی نودها در شبکه را نشان دهند.
برای حل این مشکلات، روش دیگری تحت عنوان متاپت تو وک [۲۵]معرفی شد که علاوه بر شبکه های همگن برای شبکه های ناهمگن به خوبی کار می کند[۴]. مبنای این روش، تعریف یک سری متاپت [۲۶]می باشد که بر مبنای آن ها در داخل گراف عمل پیاده روی تصادفی را انجام می دهد. این استراتژی، همبستگی بین انواع مختلف نود را در نظر می گیرد. علی رغم کارایی این روش برای شبکه های ناهمگن معایبی نیز دارد که به شرح زیر می باشد:

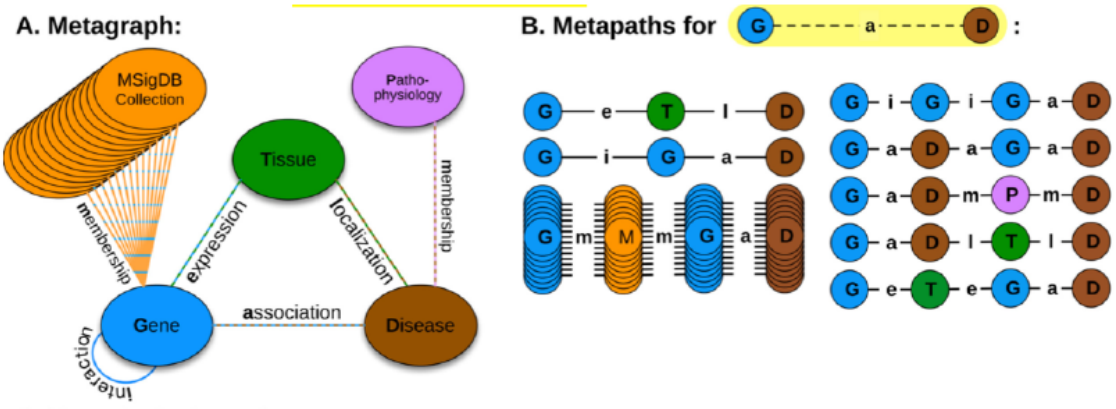
- برای تعریف هر متاپت نیاز به دانش در آن حوزه می باشیم. برای مثال باید بدانیم که هر پروتئین با چه عناصری می تواند ارتباط داشته باشد و زنجیره مان را طبق آن تعریف نماییم.
- تنها از یک متاپت در هر لحظه برای تولید یک پیاده روی تصادفی استفاده میکند.
- مفاهیم مربوط به یال ها را در نظر نمی گیرد.

همان گونه که در شکل ۲ـ۳ـ۱ مشاهده می کنید تمام ارتباطات ممکن بین ژن و بیماری را به عنوان متاپت تعریف نموده است.

۲ـ۴ نتیجه گیری

در این بخش مفاهیم و مطالعات پیشین مربوط به حوزه مدنظر بررسی شدند. همان گونه که مطالعه شد، تقریباً تمامی روش ها و استراتژی های مربوط به عمل پیاده روی تصادفی در شبکه به منظور جاسازی نود ها در فضایی با ابعاد پایین تر محدود به شبکه های همگن می شدند و روش های جدید نیز معایب زیادی دارند که به آن ها اشاره شد. در نتیجه، نیاز به یک استراتژی جدید برای تعیین شباهت نود ها در گراف احساس می شود. در مطالعه پیش رو سعی در استفاده از یک روش جدید است.

[25] Metapath2vec
[26] Metapath



شکل ۲ـ۳: چگونگی تعریف متاپت برای یک شبکه زیست پزشکی

فصل ۳

روش تحقیق

۳ـ۱ مقدمه

در این بخش به بررسی روش ارائه شده در این پروژه برای پیش بینی ارتباطات دارو_هدف در شبکه ناهمگن می پردازیم.

۳ـ۲ تشریح کامل روش تحقیق

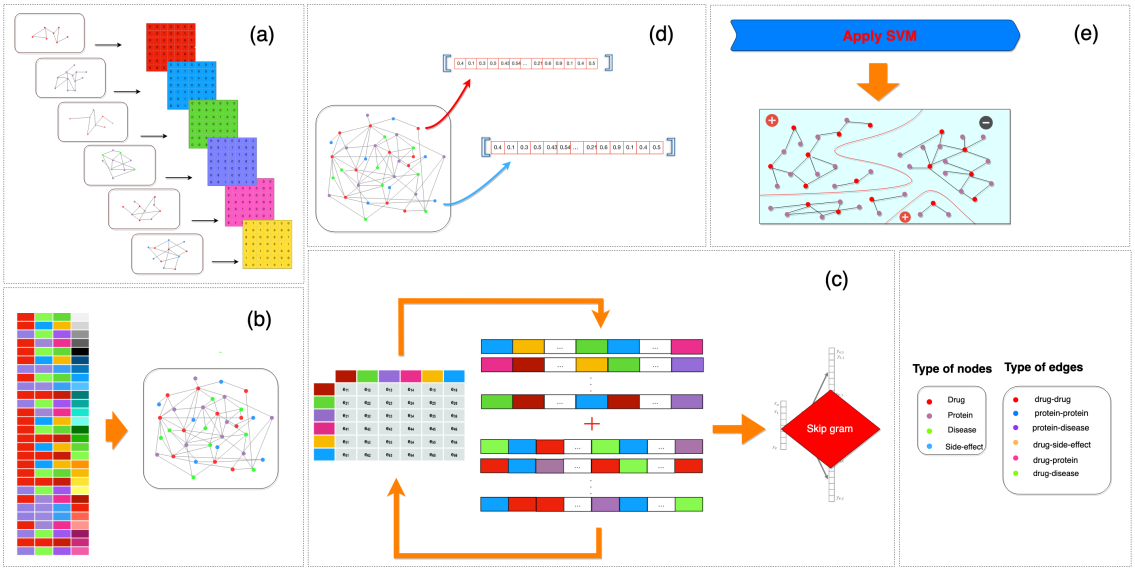
به طور کلی، روش پیشنهادی ما از چند گام که در ادامه به آن اشاره خواهد شد، تشکیل شده است. ابتدا، ما پایگاه داده ای مطابق با مقاله [۱۰] با استفاده از چهار پایگاه داده ایجاد نمودیم. این پایگاه داده که جزئیات آن در فصل بعد آمده است، از چهار نوع نود و شش نوع یال تشکیل یافته است. به عبارت دیگر شش نوع ماتریس [۱] که روابط مختلف بین نود ها را نشان می دهد داریم. عدد صفر در خانه های آرایه به منزله عدم وجود رابطه بین دو راس و نبود یال و عدد یک حاکی از وجود یک رابطه یا یال بین آن دو است.

در مرحله بعد که در شکل ۳ـ۲ به خوبی واضح می باشد، برای ساختن شبکه مربوطه مان تمام سطر هایی که نشان دهنده یک ارتباط میان دو نود می باشد را استخراج کرده و شبکه مورد نظرمان را می سازیم.

حال نوبت به آن رسیده است که بااستفاده از یکی از روش های جاسازی نود در گراف، شبکه جاسازی شده را به دست بیاوریم. همان گونه که در فصل پیشین اشاره شد، روش های بررسی شده هر کدام مزایا و معایب خود را داشتند. ما برای این پروژه، از روش جدیدی تحت عنوان ادج تو وک که در مقاله [۵] به خوبی شرح داده شده است استفاده نمودیم که با عملکرد دقیق تر و بهتری به جاسازی نود ها در شبکه بپردازیم.

پس از انجام عمل جاسازی نودها در شبکه، هر نود با استفاده از یک بردار نماینده [۲] با ابعاد [۳] ۱۲۸ بازنمایی می شود. به منظور پیش بینی و طبقه بندی یال ها باید به اندازه تمام یال های با برچسب [۴] مثبت [۵]، یال هایی با برچسب منفی [۶] تولید نماییم. منظور از برچسب منفی، یال هایی ست که در حقیقت در گراف وجود ندارند یا ارتباطی میان آن دو نود مشخص وجود ندارد. به این نوع داده ها، نمونه های منفی می گویند. حال در گام آخر، با استفاده از یک طبقه بند به پیش بینی ارتباطات دارو_هدف می پردازیم. ما در این آزمایش از طبقه بند

[1] Matrix
[2] representative Vector
[3] Dimension
[4] Label
[5] Positive
[6] Negative



شکل ۳ـ۱: مراحل روش پشنهادی

ماشین بردار پشتیبان استفاده نمودیم.

در ادامه به بررسی دقیق الگوریتم ادج تو وک و روش ماشین بردار پشتیبان می پردازیم.

۳ـ۲ـ۱ الگوریتم ادج تو وک

در شبکه های ناهمگن ما با انواع نودها و روابط سروکار داریم. ژنگ جاو و همکارانش با استفاده ازفرایند پیاده روی تصادفی، یک ماتریس انتقال نوع یال براساس وزن های انتقال میان انواع یال ها ایجاد نمودند. او با استفاده از فرمول ۳ـ۱ ، احتمال نودهای همسایه یک نوع نود مشخص را به حداکثر می رساند :

$$\arg\max_{\theta,M} \prod_{v\in V} \prod_{c\in N(v)} \mathrm{p}(c \mid v;\theta;M) \qquad (۱ـ۳)$$

در شبکه G(V،E) ، Vبه مجموعه ی نودها اشاره می کند و E مجموعه ی یال هارا مشخص می نماید. هم چنین N(V) یک گروه از همسایگان نود Vمی باشد. θ به پارامتری شدن جاسازی نودها اشاره می کند. ماتریس Mنیز همان ماتریس انتقال نوع یال می باشد.

به طورکلی، الگوریتم ادج تو وک مقادیر ماتریس M را ابتدا به ۱ مقدار دهی می کند. سپس با استفاده از چهارچوب[۷] انتظار حداکثر[۸]، این ماتریس آموزش می بیند .چهارچوب انتظار حداکثر از دو مرحله انتظار و حداکثر سازی تشکیل یافته است. در مرحله انتظار، با استفاده از فرمول۳ـ۲ همبستگی[۹] میان انواع یال ها را بهینه می کند:

$$M(\mathrm{e}_i, \mathrm{e}_j) = Sigmoid(\frac{E[(\vec{v}_i - \mu(\vec{v}_i))(\vec{v}_j - \mu(\vec{v}_j))]}{\sigma(\vec{v}_i)\sigma(\vec{v}_j)}) \qquad (۳ـ۲)$$

در این فرمول، e_i و e_j همان نوع یال ها هستند و اشاره به ماتریس انتقال نوع لبه به روز شده می کند. همچنین،v_iو v_j بردارهایی هستند که e_i و e_j را بازنمایی می کنند. مقدار E[·] همان مقدار انتظار می باشد. در مرحله حداکثرسازی، آخرین نسخه به روز شده از ماتریس انتقال نوع یال را دریافت کرده و پیاده روی های تصادفی را بر مبنای این ماتریس انجام می دهد. احتمال پیاده روی های تصادفی با استفاده از فرمول زیر تعیین می شود:

$$\mathrm{p(n \mid v; u}; M) = \frac{\omega_{\mathrm{vn}} \cdot M_{T(\mathrm{u,v})T(\mathrm{v,n})} \cdot \alpha_{\mathrm{pq}}(\mathrm{n,u})}{\sum_{k \in N(\mathrm{v})} \omega_{\mathrm{vk}} \cdot M_{T(\mathrm{u,v})T(\mathrm{v,k})} \cdot \alpha_{\mathrm{pq}}(\mathrm{k,u})} \qquad (۳ـ۳)$$

۳ـ۲ـ۲ روش ماشین بردار پشتیبان

بردارهای پشتیبان به زبان ساده، مجموعه ای از نقاط در فضای n بعدی داده ها هستند که مرز دسته ها را مشخص می کنند و مرزبندی و دسته بندی داده ها براساس آنها انجام می شود و با جابجایی یکی از آنها، خروجی دسته بندی ممکن است تغییر کند .ماشین بردار پشتیبان ، یک دسته بند یا مرزی است که با معیار قرار دادن بردارهای پشتیبان ، بهترین دسته بندی و تفکیک بین داده ها را برای ما مشخص می کند. در این روش فقط داده های قرار گرفته در بردارهای پشتیبان مبنای یادگیری ماشین و ساخت مدل قرار می گیرند و این الگوریتم به سایر نقاط داده حساس نیست و هدف آن هم یافتن بهترین مرز در بین داده هاست به گونه ای که بیشترین فاصله ممکن را از تمام دسته ها (بردارهای پشتیبان آنها) داشته باشد .

[7] Framework
[8] Expectation Maximization
[9] Correlation

فصل ۴

نتایج و تفسیر آنها

۴ـ۱ مقدمه

در این بخش نتایج حاصل از روش پیشنهادی مطرح شده در فصل سوم را بر روی مجموعه داده ای که ساخته ایم، بررسی می کنیم.

۴ـ۲ محتوا

۴ـ۲ـ۱ مجموعه داده ها

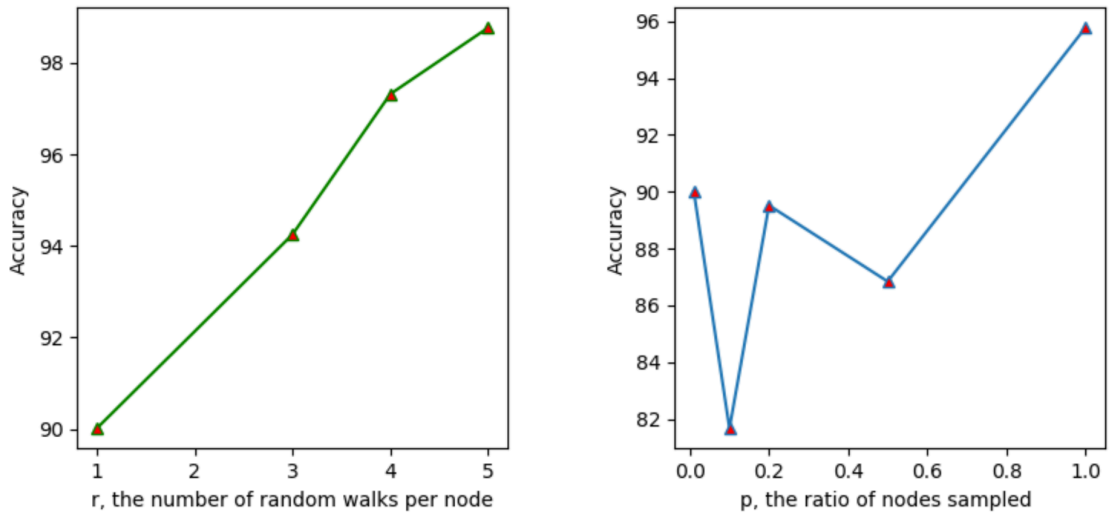
ما مجموعه داده هایمان را طبق مقاله [۱۰] که ترکیبی از چند پایگاه داده ی مرتبط با دارو است، ایجاد نمودیم. این مجموعه داده از چهار نوع نود که شامل دارو، پروتئین، اثر جانبی و بیماری و انواع یال تشکیل شده است. یال ها نیز انواع مختلفی مثل دارو_دارو، دارو_پروتئین، دارو_اثر جانبی، دارو_بیماری، پروتئین_پروتئین و پروتئین_بیماری دارند. این مجموعه داده حدود ۱۱۴۸۸ نود و ۱۸۹۵۴۳۳ یال دارد. نود های دارو و روابط دارو_هدف از مجموعه داده دراگ بانک[۱] [۸] برگرفته شده است. هم چنین نود های پروتئینی و روابط پروتئین_پروتئین از مجموعه داده اچ پی آردی [۲] [۷] گرفته شده است. علاوه بر این، نود های اثرات جانبی و روابط دارو_اثر جانبی از پایگاه داده سایدر[۳] [۹] است. در نهایت، نود های بیماری، یال های دارو_بیماری و پروتئین_بیماری از پایگاه داده [۲] استفاده شده است.

۴ـ۲ـ۲ ارزیابی نتایج

چندین آزمایش به منظور عملکرد روش پیشنهادی انجام گرفته است. ما از اعتبار سنجی متقابل کی فولد[۴] به منظور افزایش اطمینان از ارزیابی مان استفاده کرده ایم. در این مطالعه، ما روش ماشین بردار پشتیبان با تابع هسته [۵]آربی اف[۶] را برای رسیدن به بالاترین دقت به کارگرفته ایم.

علاوه براین، ما به تنظیم پارامترهای الگوریتم ادج تو وک به منظور یافتن بهترین نتیجه و میزان تغییرات

[1] DrugBank
[2] HPRD
[3] SIDER
[4] k-Fold cross validation
[5] Kernel Function
[6] RBF



شکل ۴ـ۱: نمودارهای مربوط به تنظیم پارامتر های الگوریتم ادج تو وک

پرداختیم. این روش شامل چهار پارامتر اساسی آر[۷]، دابلیو[۸]، پی[۹] و اِن[۱۰]. آر به تعداد پیاده روی های تصادفی روی هر نود اشاره می کند که مقدار پیش فرض آن برابر با یک می باشد. پارامتر دابلیو به طول هر پیاده روی اشاره کرده که معادل با ۵۰ می باشد. علاوه بر آن، پارامتر پی همان نسبت نود های نمونه گذاری شده است که مقدار پیش فرض آن ۰.۱۰ است و اِن تعداد تکرارها برای آموزش ماتریس انتقال نوع یال است که برابر ۱۰ می باشد. ما در این آزمایش مقدار دو متغیر آر و پی را تغییر داده و دو پارامتر دیگر را ثابت نگه داشتیم. شکل۴ـ۲ـ۲ این موضوع را به خوبی نشان می دهد. با افزایش آر، به درصد قابل توجهی می رسیم که این به دلیل تاثیر روابط یال های بیشتر در طول هر نود می باشد. اما با افزایش مقدار پی، دقت ما به طور نامعینی می باشد و گاهی زیاد و گاهی کم می شود. لذا از تغییر مقدار پیش فرض آن صرف نظر کرده و آن را ثابت نگاه می داریم.

معیار های ارزیابی زیادی برای اندازه گیری عملکرد مطالعه مان وجود دارد. ما چهار معیار ارزیابی امتیاز

[7] r
[8] W
[9] p
[10] N

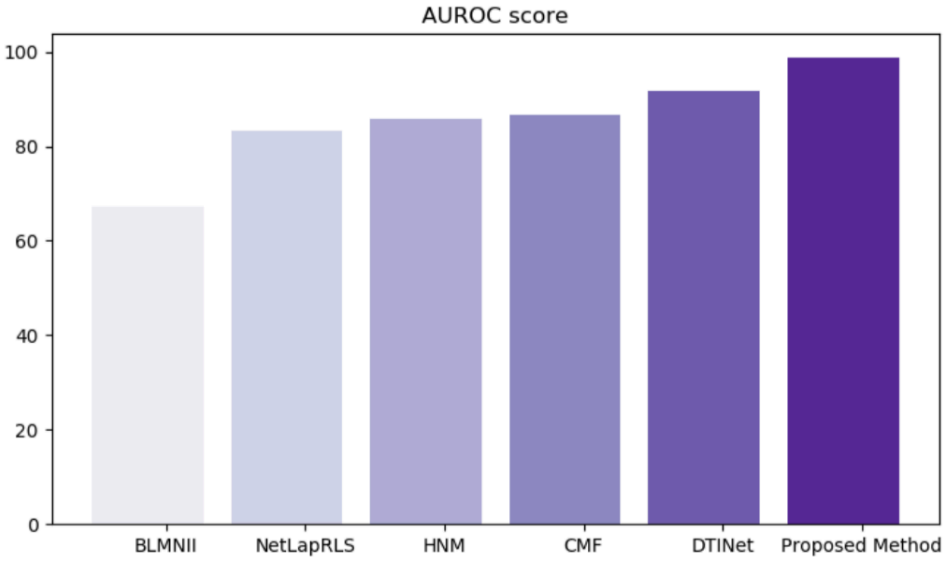
		Recall	Precision	F1 score	Hamming loss
p	0.1	0.81	0.81	0.81	0.18
	0.2	0.89	0.89	0.89	0.10
	0.5	0.86	0.86	0.86	0.13
	1	0.95	0.95	0.95	0.04
r	3	0.94	0.94	0.94	0.05
	4	0.97	0.97	0.97	0.02
	5	**0.98**	**0.98**	**0.98**	**0.01**

شکل ۴ـ۲: مقادیر بازخوانی، دقت، امتیاز اف ۱ و خطای همینگ به ازای مقادیر مختلف دو پارامتر پی و آر

اف ۱[11]، بازخوانی[12]، دقت[13] و فاصله همینگ[14] را برای هریک از مقادیر پارامترهای پی و آر را محاسبه کرده و در شکل ۴ـ۲ـ۲ یادداشت کرده ایم

در شکل ۴ـ۲ـ۲، ما میزان امتیاز آیو آر اُ سی[15] را با روش هایی که تاکنون موردتوجه قرار گرفته اند، مورد مقایسه قرار دادیم. همان گونه که مشخص است،. این روش نسبت به روش های موجود به دقت و عملکرد بسیاری دست یافته است.

[11] F1-score
[12] Recall
[13] Precision
[14] Hamming loss
[15] AUROC Score



شکل ۴ـ۳: مقایسه میزان عملکرد میان روش های موجود با روش پیشنهادی

فصل ۵

جمع بندی و پیشنهادها

۵ـ۱ مقدمه

پیش بینی ارتباطات دارو_هدف از موضوعات مطرح در حوزه کشف دارو و استفاده مجدد از آن محسوب می شود، که با الگوریتم ادج تو وک به این کار پرداختیم. در این الگوریتم، علاوه بر مفاهیم نود ها اطلاعات مربوط به یال ها نیز حفظ می شود. در این بخش به معرفی کلی مطالعه ی انجام شده و جمع بندی می پردازیم. در انتها نیز درباره ی نوآوری و پیشنهادات آتی برای این مطالعه بحث می کنیم.

۵ـ۲ جمع بندی

در این مطالعه، سعی در پیش بینی و تشخیص ارتباطات و لینک های میان دارو و هدف کردیم. برای این کار، ما ابتدا مجموعه داده ای از روابط مختلف میان عناصر را طبق مجموعه داده های معتبر به وجود آورده و سپس الگوریتم ادج تو وک که به تازگی ارائه شده است را به منظور جاسازی نود ها به کار بردیم. در انتها با استفاده از روش ماشین بردار پشتیبان به پیش بینی این ارتباطات پرداختیم. هم چنین به منظور رسیدن به دقت بالاتر، به تنظیم پارامتر های این دو الگوریتم پرداختیم. روشی که در این پژوهش پیشنهاد شده است نسبت به روش های موجود از دقت بالاتری برخوردار می باشد.

۵ـ۳ نوآوری

روشی که در این پژوهش از آن برای پیش بینی ارتباطات دارو_هدف استفاده شده است، تاکنون مورد توجه قرار نگرفته است. هم چنین این روش نسبت به روش هایی که تاکنون ارائه شده است، میزان عملکرد و دقت قابل توجهی را ایفا می کند.

۵ـ۴ پیشنهادها

برای کارهای آتی می توان روابط بیشتری هم چون ارتباطات پروتئین_پروتئین[1] و یا دارو_دارو[2] را با به کار گیری این روش پیش بینی نمود. هم چنین، اعمال روش پیشنهادی بر روی مجموعه داده های دیگر می تواند

[1] Protein-Protein interaction(PPI)
[2] Drug-Drug interaction(DDI)

به پیدایش و کشف روابط بیشتر دارو ـ هدف شود.

مراجع

[1] Abu-El-Haija, S., Perozzi, B., Al-Rfou, R., and Alemi, A. A. Watch your step: Learning node embeddings via graph attention. in *Advances in Neural Information Processing Systems* (2018), pp. 9180–9190.

[2] Davis, A. P., Murphy, C. G., Johnson, R., Lay, J. M., Lennon-Hopkins, K., Saraceni-Richards, C., Sciaky, D., King, B. L., Rosenstein, M. C., Wiegers, T. C., et al. The comparative toxicogenomics database: update 2013. *Nucleic acids research 41*, D1 (2012), D1104–D1114.

[3] DiMasi, J. A. New drug development in the united states from 1963 to 1999. *Clinical Pharmacology & Therapeutics 69*, 5 (2001), 286–296.

[4] Fu, G., Ding, Y., Seal, A., Chen, B., Sun, Y., and Bolton, E. Predicting drug target interactions using meta-path-based semantic network analysis. *BMC bioinformatics 17*, 1 (2016), 160.

[5] Gao, Z., Fu, G., Ouyang, C., Tsutsui, S., Liu, X., Yang, J., Gessner, C., Foote, B., Wild, D., Ding, Y., and Yu, Q. edge2vec: Representation learning using edge semantics for biomedical knowledge discovery. *BMC Bioinformatics 20*, 1 (Jun 2019), 306.

[6] Grover, A., and Leskovec, J. node2vec: Scalable feature learning for networks. in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining* (2016), ACM, pp. 855–864.

[7] Keshava Prasad, T., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., et al. Human protein reference database—2009 update. *Nucleic acids research 37*, suppl_1 (2008), D767–D772.

[8] Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., et al. Drugbank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic acids research 39*, suppl_1 (2010), D1035–D1041.

[9] Kuhn, M., Campillos, M., Letunic, I., Jensen, L. J., and Bork, P. A side effect resource to capture phenotypic effects of drugs. *Molecular systems biology 6*, 1 (2010).

[10] Luo, Y., Zhao, X., Zhou, J., Yang, J., Zhang, Y., Kuang, W., Peng, J., Chen, L., and Zeng, J. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nature communications 8*, 1 (2017), 573.

[11] Perozzi, B., Al-Rfou, R., and Skiena, S. Deepwalk: Online learning of social representations. in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (2014), ACM, pp. 701–710.

[12] Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., and Mei, Q. Line: Large-scale information network embedding. in *Proceedings of the 24th international conference on world wide web* (2015), International World Wide Web Conferences Steering Committee, pp. 1067–1077.

واژه‌نامه فارسی به انگلیسی

انعطاف پذیری ..... Flexibility
آموزش ..... Training
بدون تاثیر ..... Unbiased
پارامتر بازگشت ..... Return parameter
پارامتر دور شدن ..... Walk away parameter
ابعاد ..... Dimension
چهارچوب ..... Framework
همبستگی ..... Correlation
اعتبارسنجی متقابل ..... Cross validation
تابع هسته ..... Kernel Function
مبتنی بر شبکه ..... Network based
ماشین بردار پشتیبانی ..... Support vector machine
جاسازی نودها ..... Node embedding
مدل های دو بخشی ..... Bipartite Models
ارتباطات دارو_هدف ..... Drug-target interactions
شبکه ..... Network
توازن ..... Trade off
بی وزن ..... Unweighted
جهت دار ..... Diredcted
بردار نماینده ..... Representative Vector
همگن ..... Homogeneous
ناهمگن ..... Heterogeneous
شباهت مبتنی بر وابستگی ..... Adjacency-based Similarity
شباهت چند گامی ..... Multi-hop Similarity
روش های پیاده روی تصادفی ..... Random Walk Approaches
استراتژی ..... Strategy

واژه‌نامه انگلیسی به فارسی

جاسازی نودها ..... Node embedding
مدل های دو بخشی ..... Bipartite Models
ارتباطات دارو ـ هدف ..... Drug-target interactions
شبکه ..... Network
همگن ..... Homogeneous
ناهمگن ..... Heterogeneous
شباهت مبتنی بر وابستگی ..... Adjacency-based Similarity
شباهت چند گامی ..... Multi-hop Similarity
روش های پیاده روی تصادفی ..... Random Walk Approaches
استراتژی ..... Strategy
انعطاف پذیری ..... Flexibility
آموزش ..... Training
بدون تاثیر ..... Unbiased
پارامتر بازگشت ..... Return parameter
پارامتر دور شدن ..... Walk away parameter
توازن ..... Trade off
بی وزن ..... Unweighted
جهت دار ..... Diredcted
بردار نماینده ..... Representative Vector
ابعاد ..... Dimension
چهارچوب ..... Framework
همبستگی ..... Correlation
اعتبارسنجی متقابل ..... Cross validation
تابع هسته ..... Kernel Function
مبتنی بر شبکه ..... Network based
ماشین بردار پشتیبانی ..... Support vector machine

Abstract:

Drug-target protein interaction prediction is considered as a fundamental step in the process of drug discovery and re-purposing. The experimental methods in this domain, despite having accurate and beneficial results, are time-consuming. In the last few years, not only have these computational methods been overcome to these problems, but the results of them have been extremely efficient.One type of these techniques is based on network prediction on biomedical networks. Today, different network-based approaches have been suggested for identifying relations between drugs and targets which work on homogeneous networks well. But these approaches do not consider edge semantics in the network. In this project, we employ the edge2vec algorithm for node embedding on the heterogeneous network and then using the support vector machine method to predict drug-target interactions. Our method has obtained a significant accuracy than the existing methods.

Keywords: Drug-Target interaction prediction, Node embedding, Edge2vec, Support Vector Machine, Heterogeneous network

Iran University of Science and Technology

Computer Engineering Department

Drug-Target Interaction Prediction by Embedded Vectors of Nodes in Heterogeneous Interactions Network

Bachelor of Science Thesis in Computer Engineering

By:

Fatemeh Fattahi

Supervisor:

Dr. Minaei

September 2019