یادگیری عمیق روی جداول رابطه‌ای

چکیده

بسیاری از ارزشمندترین داده‌های جهان در پایگاه‌های اطلاعاتی رابطه‌ای و انبارهای داده ذخیره می‌شوند. بنابراین، ساخت مدل‌های یادگیری ماشین با استفاده از این داده‌ها اهمیت دارد. مشکل اصلی این است که هیچ روش یادگیری ماشین قادر به یادگیری روی جداول متعدد نیست. روش‌های فعلی فقط می‌توانند از یک جدول یاد بگیرند. بنابراین داده‌ها ابتدا باید به صورت دستی به هم متصل شوند و در یک جدول آموزشی واحد جمع شوند (فرآیندی که به عنوان مهندسی ویژگی شناخته می‌شود). مهندسی ویژگی کُند، مستعد خطا و منجر به مدل‌های غیربهینه می‌شود.

در یادگیری ماشین، رویکرد یادگیری عمیق منجر به حذف مهندسی ویژگی می‌شود. در این رویکرد، ویژگی‌ها به صورت خودکار استخراج می‌شوند و مورد استفاده قرار می‌گیرند. کاربردهای این رویکرد در حوزه تصویر، متن و ویدئو بسیار دیده می‌شود. در سال‌های اخیر، رویکرد یادگیری عمیق کاربردهایی روی گراف نیز داشته است. در این کاربردها، مسایل یادگیری ماشین در سطح نود، یال یا گراف تعریف می‌شوند؛ مثلا رده بندی نودها (سطح نود)، پیش‌بینی برقراری ارتباط میان یک جفت نود (سطح یال) یا تخمین یک مقدار برای یک گراف (سطح گراف). برای حل این مسایل، رویکرد یادگیری عمیق، که با نام کلی شبکه‌های عصبی گرافی نیز شناخته می‌شوند، به دنبال استخراج خودکار ویژگی (یا بازنمایی) برای نود، یال و/یا گراف است. این بازنمایی‌های استخراج شده، به عنوان ورودی روشهای یادگیری ماشین استفاده می‌شوند که می‌خواهند مسئله تعریف شده در سطح نود/یال/گراف را حل کنند. به طور کلی، در رویکرد یادگیری عمیق، استخراج ویژگی از داده به صورت خودکار انجام می‌شود و عملاً مهندسی ویژگی حذف می‌شود.

حال فرض کنید که داده در داخل چند جدول رابطه‌ای ذخیره شده باشد. سوال مهم این است که آیا می‌توان از رویکرد یادگیری عمیق استفاده کرد؟ به عبارت دیگر، چگونه رویکرد یادگیری عمیق بازنمایی برای داده داخل پایگاه داده رابطه‌ای را استخراج می‌کند؟ این سوال مهم، محور اصلی این طرح پیشنهادی است. در راستای پاسخ به این سوال و در این طرح، رویکرد یادگیری عمیق مبتنی بر شبکه عصبی گرافی، برای یادگیری مستقیم روی چند جدول دنبال می‌شود.

ایده اصلی این است که پایگاه‌های داده رابطه‌ای به‌عنوان یک گراف زمانی و ناهمگن در نظر گرفته شود؛ یک گره برای هر ردیف در هر جدول و یالها ارتباط میان داده را نشان می‌دهد (ارتباط کلید اولیه و خارجی). پس از آن، شبکه‌های عصبی گرافی می‌تواند به صورت خودکار بازنمایی را برای نودها استخراج کند. این بازنمایی‌ها، به عنوان ورودی به روشهای یادگیری ماشین هستند که مسایل پیش‌بینی را درسطح نود/یال/گراف حل می‌کنند. به طور کلی، در این طرح

یک حوزه تحقیقاتی جدید تعریف می‌شود که یادگیری ماشین گرافی روی داده‌های رابطه‌ای تعمیم داده می‌شود و منجر به کاربردهای جالبی می‌شود.

بیان مسئله

می‌دانیم که جداول رابطه‌ای در سیستم‌های اطلاعاتی بسیار کاربرد دارد. برنامه‌نویسان با استفاده از دستورات به زبان SQL، این جداول را تولید، نگهداری و دستکاری می‌کنند. در این طرح، ایجاد یک زبان درخواست پیش‌بینی با نام PQL نیز دنبال می‌شود. برنامه نویس با دستوراتی که به زبان PQL می‌نویسد، یک مساله پیش‌بینی در سطح نود/یال/گراف را روی تمام یا قسمتی از گراف تعریف و حل می‌کند (گراف معادل با داده داخل جداول است). نتایج این پیش‌بینی به صورت یک فیلد، مجموعه‌ای از فیلدها یا مجموعه‌ای از رکوردها در اختیار برنامه‌نویس قرار می‌گیرد؛ به طوری که می‌تواند از نتایج این پیش‌بینی در دستورات SQL استفاده کند. بدین ترتیب، استفاده از دستورات SQL و PQL منجر به تولید سیستم اطلاعاتی می‌شود که در داخل خود از روشهای یادگیری ماشین برای انجام پیش‌بینی استفاده می‌کند و از پیش‌بینی های هوشمند خود برای توسعه سیستم اطلاعاتی استفاده می‌کند.

بسیاری از ارزشمندترین داده‌های جهان در پایگاه‌های داده رابطه‌ای و انبارهای داده ذخیره می‌شوند. بسیاری از سازمانها نیز اطلاعات خود را در جداول رابطه‌ای ذخیره می‌کنند. این جداول، اساس سیستم‌های اطلاعاتی سازمان است. برنامه نویسان با استفاده از دستورات SQL با این جداول کار می‌کنند و سیستم‌های اطلاعاتی مختلفی را در سازمان توسعه می‌دهند .

اگر سازمان بخواهد که سیستم اطلاعاتی خود را هوشمند کند، استفاده از مدل‌های یادگیری ماشین در دستور کار قرار می‌گیرد. روال معمول آن است که داده‌ها ابتدا باید به صورت دستی به هم متصل شوند و در یک جدول آموزشی واحد جمع شوند. دلیل تجمیع اطلاعات در یک جدول آن است که روش‌های یادگیری ماشین قادر به یادگیری روی جداول متعدد نیستند. این روش‌ها فقط می‌توانند از یک جدول یاد بگیرند .

مهندسی ویژگی به فرآیندی گفته می‌شود که در راستای ایجاد یک جدول مشترک، فیلدهایی از جداول مختلف انتخاب شوند، تغییر یا تعویض شوند و به یک جدول واحد منتقل شوندانتقال و یا تغییر آنها در راستای تشکیل یک جدول واحد را شناخته می‌شود)بنابراین مهندسی ویژگی کُند، مستعد خطا و منجر به مدل‌های غیربهینه می‌شود.

پایگاه داده تا کنون ساخت گراف دانش از روی داده های غیرساخت یافته ی متنی بسیار مطالعه و بررسی شده است. حتی گراف های دانش عمومی نیز تولید شده است. سازمان ها نیز به دنبال ساخت گراف دانش خود هستند بدین صورت که اطلاعات موجود در سازمان خود را به گراف دانش تبدیل کرده و از آن در کاربردهای مختلف استفاده نمایند. اما در اکثر سازمان‌ها، بیشتر اطلاعات به صورت ساخت یافته و در پایگاه داده های رابطه‌ای ذخیره شده است. برای یکسان بودن نحوه عملکرد، می‌توان داده‌های ساخت یافته را نیز به صورت گراف دانش در نظر گرفت. به‌طوری که موجودیتها در پایگاه داده به صورت گره و روابط بین موجودیتها به

صورت یال لحاظ شوند. بدین ترتیب می‌توان دو گراف دانش متفاوت را در یک سازمان تصور کرد: گراف دانش غیررابطه‌ای و گراف دانش رابطه‌ای. تمرکز این طرح روی گراف دانش رابطه‌ای است که از اطلاعات موجود از جداول بدست می‌آید.

با استفاده از دستورات به زبان SQL سیستم اطلاعاتی قادر است تا داده ها را در پایگاه داده درج، ویرایش و حذف کند. همزمان با این تغییرات، گراف دانش رابطه‌ای نیز بروز می‌شود. سوال اصلی این است که این گراف چه کاربردی در توسعه هوشمند سیستم اطلاعاتی دارد؟

کاربردهای مختلف یادگیری ماشین مانند رده‌بندی، دسته‌بندی، توصیه‌گر و تخمین مقدار روی گراف قابل تعریف و توسعه هستند. با ظهور پردازش عمیق در این حوزه، یعنی شبکه عصبی گرافی، امکان پیاده‌سازی این روشها روی گراف های بزرگ نیز بوجود آمده است. بنابراین وجود گراف می‌تواند به ایجاد فابلیت یادگیری ماشین در سیستم اطلاعاتی منجر شود. به گونه‌ای که با افزایش داده (که منجر به افزایش حجم گراف می‌شود) پیش‌بینی های بهتری انجام داد .

بنابراین از یک سو باید پیش‌بینی های مختلفی را روی گراف انجام داد و از سوی دیگر این پیش‌بینی ها را در توسعه سیستم اطلاعاتی استفاده نمود. به عنوان مثال اگر محصول جدیدی به جدول محصولات اضافه شد، بر اساس گراف موجود پیش‌بینی شود که این محصول به کدام کاربر پیشنهاد شود. سیستم اطلاعاتی نیز بر اساس این پیش‌بینی خدماتی را به کاربران خود ارایه کند.

مساله اصلی این طرح در استقاده از یادگیری ماشین روی گراف دانش حاصل از داده‌های رابطه‌ای در توسعه سیستم اطلاعاتی است.

اهداف اصلی و فرعی طرح

۱. توسعه یک گراف دانش رابطه ای که ساختار یک پایگاه داده رابطه ای را منعکس می کند و امکان نمایش بهتری از روابط داده ها را فراهم می‌کند.
۲. تعریف یک زبان برای پرس وجو که به طور خاص برای استخراج پیش بینی ها از گراف دانش رابطه ای طراحی شده است، که دسترسی و کاربرد تحلیل های پیش بینی کننده را بهبود می‌دهد.
۳. استفاده از نتایج پیش بینی در پرس وجوهای پایگاه داده رابطه‌ای، که امکان استفاده از نتایج تحلیل را در توسعه سیستم اطلاعاتی می‌دهد.
۴. گسترش قابلیتهای سیستمهای اطلاعاتی با استفاده از توانایی های پیش بینی کننده، که فرآیند تصمیم گیری و کارایی عملیاتی را بهبود می‌دهد.

پیشینه تحقیق در ایران و جهان

یک گراف دانش با چهارتایی $G = (V, E, R, T)$ تعریف می‌شود که در آن V مجموعه تمام گره‌ها، E مجموعه تمام یال‌ها، R مجموعه تمام رابطه‌ها و T مجموعه تمام انواع گره‌ها است. در گراف دانش هر یال به‌صورت سه‌گانه نشان داده می‌شود: (سرآیند،

رابطه ، تالی) یا (Head, Relation, Tail). سرآیند و تالی هر کدام می‌توانند یک نود یا موجودیت باشند و Relation نیز رابطه بین این دو موجودیت است؛ این ترکیب سه‌گانه تشکیل یک حقیقت را می‌دهد.

گراف‌های دانش دارای کاربردهای فراوانی هستند که در ادامه به دو کاربرد آن به‌صورت نمونه اشاره می‌شود [۱]:

الف) ارائه اطلاعات: اگر گراف دانشی وجود داشته باشد که در آن نام فیلم‌ها و کارگردان‌های آن‌ها و همچنین بازیگران فیلم‌ها آمده باشد، این گراف می‌تواند برای پاسخ‌دهی به انواع کوئری‌ها قابل استفاده باشد، به عنوان مثال نام تمام فیلم‌های کارگردان فیلم «تایتانیک». اینگونه اطلاعات از طریق ساخت گراف دانش و تحلیل آن‌ها قابل ارائه هستند.

ب) پاسخ به سوالات: به طور سنتی، پاسخ به سؤالات با استفاده از محتواهای بدون ساختار مانند پاراگراف، متن و یا صفحه ویکی پدیا انجام می‌شود. قابلیت‌های اکتشافی چنین سیستم‌هایی محدود به محتوای داده‌شده است، به عنوان مثال، با داشتن یک صفحه ویکی‌پدیا در مورد آلبرت انیشتین یک کاربر می‌تواند با پرسیدن سوال «آلبرت انیشتین متولد کجاست؟» محل تولد انیشتین را به دست آورد. اما پرسیدن سؤالات بیشتر در مورد این شهر (به عنوان مثال، «چند نفر در یوالِام زندگی می‌کنند؟») نتایج مناسبی را برنمی‌گردانند. در این حالت، ما به پیشینه ساختاری احتیاج داریم که گراف‌های دانش برای ما به همراه می‌آورند

شرکت‌های بزرگ فناوری اطلاعات، گراف‌های دانش متنوعی را ارائه کرده‌اند. همچنین دانشگاه استنفورد نیز به تازگی دیتاست‌های متنوعی را برای تست و ارزیابی الگوریتم‌ها ارائه کرده است . در ادامه به چند گراف دانش اشاره می‌شود که به‌صورت عمومی ارائه‌شده است:

الف) فری‌بیس : این گراف دانش توسط شرکت گوگل و بر اساس داده‌های موجود در ویکی‌پدیا تولید شده است. اولین بار این گراف در سال ۲۰۰۷ ایجاد شده و توسعه آن در سال ۲۰۱۶ متوقف شده‌است. این گراف شامل بیش از ۵۰ میلیون گره، ۳۸ هزار نوع رابطه و بیش از ۳ میلیارد سه‌گانه است. با این حال بخش زیادی از رابطه‌ها در این گراف کامل نیستند. به عنوان مثال ۷۸ درصد از تمام اشخاصی که در این گراف ثبت شده‌اند، فاقد ویژگی «ملیت» هستند و ۹۳ درصد آن‌ها فاقد «تاریخ تولد» هستند. به همین دلیل لزوم توسعه ابزارهای تکمیل گراف، ضروری به نظر می‌رسد [۲].

ب) ویکی‌دیتا: یک گراف دانش بسیار بزرگ و زنده که توسط مجموعه ویکی‌پدیا ارائه‌شده و همچنان در حال گسترش است [۳].

ج) گراف‌ها BioKG و WikiKG2 : این دو گراف دانش زیرمجموعه تمام دیتاست‌هایی هستند که در فریم‌ورک OGB ارائه‌شده‌اند [۴]: برای اینکه یادگیری رابطه‌ای روی داده‌های زیست‌پزشکی، استانداردتر و قابل تکرارتر شود، یک گراف دانش بیولوژیکی جدید توسط دانشگاه استنفورد ارائه‌شده است که مجموعه‌ای از داده‌های رابطه‌ای انتخاب‌شده از پایگاه‌های بیولوژیکی باز را در قالبی یکپارچه با شناسه‌های مشترک و مرتبط به هم ارائه می‌دهد. مجموعه داده ogbl-biokg یک گراف دانش است که آقای لسکووک به همراه همکاران، با استفاده از داده‌های تعداد زیادی از مخازن داده‌های زیست پزشکی ایجاد کرده‌اند. این گراف، شامل ۵ نوع موجودیت است: بیماری ها (۱۰۶۸۷ گره)، پروتئین‌ها

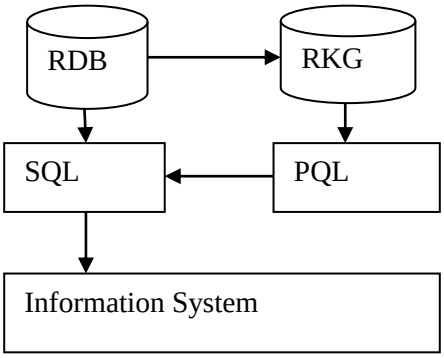
(۱۷۴۹۹)، داروها (۱۰۵۳۳ گره)، عوارض جانبی (۹۹۶۹ گره)، و عملکردهای پروتئین (۴۵۰۸۵ گره). ۵۱ نوع رابطه مستقیم وجود دارد که دو نوع موجودیت را به هم متصل می‌کند، از جمله ۳۹ نوع تداخل دارو-دارو، ۸ نوع تعامل پروتئین-پروتئین، و همچنین دارو- پروتئین، اثر جانبی دارو، دارو- پروتئین و روابط عملکرد-عملکرد. همه رابطه‌ها به عنوان لبه‌های جهت دار مدل می‌شوند، که در میان آن‌ها روابطی که انواع موجودات یکسان را به هم متصل می‌کنند (به عنوان مثال، پروتئین-پروتئین، دارو-دارو، عملکرد-عملکرد) همیشه متقارن هستند، به عنوان مثال، لبه‌ها دو جهته هستند. این مجموعه داده مربوط به هر دو تحقیقات زیست پزشکی و بنیادی ML است. در سمت اساسی ML، مجموعه داده، چالش‌هایی را در مدیریت یک KG ناقص با مشاهدات متناقض احتمالی ارائه می‌کند. این به این دلیل است که مجموعه داده ogbl-biokg شامل برهمکنش‌های ناهمگنی است که از مقیاس مولکولی (به عنوان مثال، برهمکنش‌های پروتئین-پروتئین در یک سلول) تا کل جمعیت (مثلاً گزارش‌هایی از عوارض جانبی ناخواسته تجربه شده توسط بیماران در یک کشور خاص) را شامل می‌شود. علاوه بر این، سه‌گانه‌ها در KG از منابعی با سطوح مختلف اطمینان، از جمله بازخوانی‌های تجربی، حاشیه‌نویسی‌های تنظیم‌شده توسط انسان، و ابرداده‌های استخراج‌شده به‌طور خودکار به دست می‌آیند.

روش انجام تحقیق

شکل زیر روش انجام تحقیق را نشان می‌دهد. فرض کنید یک سیستم اطلاعاتی (Information System) یک پایگاه داده شامل چند جدول مختلف دارد. سیستم برای توسط دستوراتی به زبان SQL داده ها را مورد دستکاری و استفاده قرار می‌دهد. در روش پیشنهادی، از روی جداول مختلف پایگاه داده، یک گراف دانش رابطه ای (RKG) ساخته می‌شود. با کمک یک زبان درخواست پیشنهادی که در شکل با نام PQL مشخص شده است، درخواستهای پیش‌بینی روی گراف پردازش و انجام می‌شود. این درخواست ها با کمک مفاهیم یادگیری ماشین روی گراف انجام می‌شوند. نتایج این درخواستها، می‌تواند در دستورات SQL استفاده شود. بدین ترتیب، سیستم اطلاعاتی قادر خواهد بود تا بر اساس پیش‌بینی های مبتنی بر یادگیری ماشین توسعه یابد. به عبارت دیگر، سیستم اطلاعاتی هوشمند طراحی شود.

به عنوان مثال برای PQL، فرض کنید اطلاعات مربوط به مشتریان، کالاها و خریدهای کالا توسط مشتریان در داخل پایگاه داده رابطه ای ذخیره شده باشد. این اطلاعات در گراف دانش رابطه‌ای متناظر نیز وجود دارد. چون این اطلاعات در گراف وجود دارد، می‌توان تصور نمود که شبکه عصبی گرافی برای هر گره در گراف بردار بازنمایی تولید می‌کند. حال کاربر جدیدی به سیستم اضافه می‌شود. همزمان درخواست زیر می‌تواند پرسیده شود: کاربر جدید به کدام کاربرهای قبلی شبیه است؟ یا یک لیست از کالاها که این کاربر جدید به آنها علاقمند است را تولید کن!

طبق روش پبشنهادی، این درخواستها را باید با زبان پیشنهادی PQL نوشته شود. این درخواست ها با استفاده از شبکه عصبی گرافی آموزش دیده روی گراف دانش رابطه‌ای اجرا می‌شوند. نتایج در اختیار برنامه نویس است تا آنها را در توسعه سیستم اطلاعاتی یا در داخل دستورات SQL خود استفاده کند.



نوآوری ها

نوآوری های طرح پیشنهادی به شرح ذیل است:

۱. یادگیری مستقیم روی داده‌های چند جدولی: این روش امکان یادگیری مستقیم روی جداول متعدد متصل به هم با کلیدهای اصلی و خارجی را فراهم می‌کند، بدون نیاز به ترکیب دستی داده‌ها به یک جدول واحد.
۲. استفاده از شبکه‌های عصبی گراف(GNN) : داده‌های رابطه‌ای به صورت یک گراف زمانی و ناهمگن نمایش داده می‌شوند، که در آن هر سطر یک گره و ارتباطات کلید اصلی-خارجی به عنوان لبه‌ها در نظر گرفته می‌شوند. سپس از شبکه‌های عصبی گراف برای یادگیری ویژگی‌ها و انجام پیش‌بینی‌ها استفاده می‌شود.
۳. پیام‌رسانی گراف زمانی: در این روش، گره‌ها تنها می‌توانند پیام‌ها را از گره‌های دیگر با زمان‌های قبلی دریافت کنند. این موضوع باعث جلوگیری از نشت اطلاعات زمانی (time travel) می‌شود.
۴. جدول آموزشی و برچسب‌گذاری خودکار: برای هر وظیفه پیش‌بینی، یک جدول آموزشی ایجاد می‌شود که شامل برچسب‌های آموزشی محاسبه شده از داده‌های تاریخی است. این جدول آموزشی شامل کلیدهای خارجی، زمان و برچسب هدف می‌باشد.