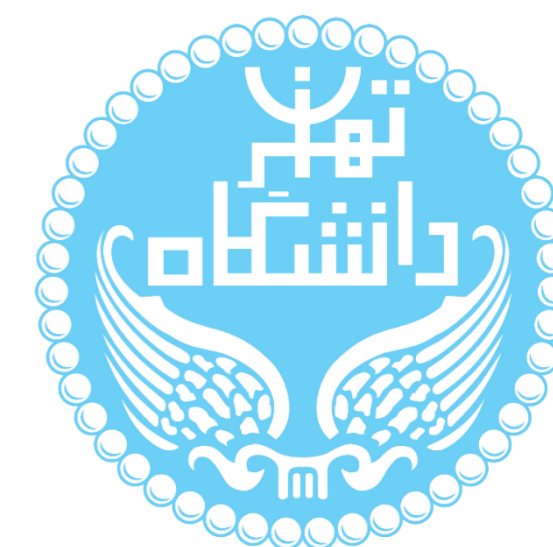


استخراج قواعد هیوریستیک به کمک سیستم هوشمند داده محور



دانشجو: رومینا امیدی
استاد راهنما: دکتر بابک نجار اعرابی
دانشکده مهندسی برق و کامپیوتر، دانشگاه تهران



نتایج

کارایی روش پیشنهادی این پژوهش، در کاربرد پزشکی که زمینه مناسبی برای اهمیت استفاده از هیوریستیک است مورد بررسی قرار گرفته و به همین منظور از دو مجموعه داده LungCancer و ThroatCancer استفاده شده است. این دو مجموعه داده علیرغم تفاوت‌های فراوان، با داشتن ویژگی‌هایی با خواستگاه یکسان، برای اهدافی همچون تشخیص سرطان استفاده می‌شوند. ابتدا بر روی دادگان LungCancer مدل ارائه شده اجرا و قواعد ابتدایی استخراج می‌شوند. ویژگی‌های مشترکی مانند سیگار کشیدن، سرفه، مشکلات ژنتیکی، و همچنین تشخیص سرطان، در هر دو مجموعه داده مشاهده می‌شود که با استفاده از این شباهت‌ها، قواعد هیوریستیک تعمیم یافته استخراج می‌شوند. در ادامه به مقایسه عملکرد این قواعد با روش‌های متداول دسته بندی می‌پردازیم:

Classifier Model	Lung Cancer Dataset	Throat Cancer Dataset
Smoking -> cancer -> death Genetics -> cancer -> death Coughing -> ~cancer ~cancer	Ecological rationality = 81.02%	Ecological rationality = 76.93%
Smoking -> cancer -> death Smoking -> cancer -> death Fatigue -> ~cancer ~cancer	Ecological rationality = 81.02%	Ecological rationality = 76.93%
Smoking -> cancer -> death Coughing -> cancer -> death ~cancer	Ecological rationality = 74.29%	Ecological rationality = 64.37%
PCA algorithm	Accuracy = 89.32%	Accuracy = 47.28%
TSNE	Accuracy = 86.32%	Accuracy = 34.72%
CNN	Accuracy = 94.15%	Accuracy = 26.53%

جدول 1: مقایسه نتایج میزان دقت قوانین هیوریستیک استخراج شده با دقت مدل‌های دیگر

با توجه به نتایج، هیوریستیک‌های بدست آمده بر اساس روش‌های متداول سنتی یادگیری ماشین تمایز معنادار نتایج را نشان می‌دهد. همچنین این روش نسبت به رقبای مبتنی بر شبکه عصبی نیز با داشتن تعمیم‌پذیری بالاتر برتری دارد. این قواعد در هر دو مجموعه داده قابل اجرا بوده و با دقت مناسبی در تشخیص سرطان عمل می‌کنند. همچنین مدل پیشنهادی با توجه به هوشمندی در جمع‌آوری، غربال و انتخاب هیوریستیک‌ها نسبت به مدل‌های پیشین که مبتنی بر exhaustive search هستند، از پیچیدگی محاسباتی کمتری برخوردار است.

جمع بندی

در این پژوهش به بررسی چالش‌های استخراج روابط مستتر میان داده با در نظر گرفتن تعمیم و تطبیق پذیری مبتنی بر رویکرد ابتکاری پرداخته شده است. هدف، ارائه روشی نوآورانه برای استخراج قوانین هیوریستیک است که قابلیت تعمیم به محیط‌های مختلف را داشته باشند. این روش، از الگوریتم‌های پیشرفته مبتنی بر مدل‌های گراف بهره می‌برد و به ما این امکان را می‌دهد که روابط علی بین ویژگی‌ها و خروجی‌ها را شناسایی کنیم. در این مدل، از تکنیک‌های پردازش زبان طبیعی مانند جایابی واژه برای استخراج شباهت‌ها و ویژگی‌های مشترک استفاده شده است. همچنین، از گراف دانش برای تعمیم قواعد به محیط‌های جدید بهره گرفته شده است. نتایج نشان می‌دهند که روش پیشنهادی قادر به سازگاری با گستره وسیعی از محیط‌ها می‌باشد و قواعد استخراج شده، عملکرد مناسبی را بر روی دادگان مختلف ارائه می‌دهند. علاوه بر کاربردهای پزشکی، کاربردهای دیگری همچون اقتصاد و بازارهای مالی، می‌توانند از بسترهای مناسب استفاده از این مدل باشند.

مراجع اصلی

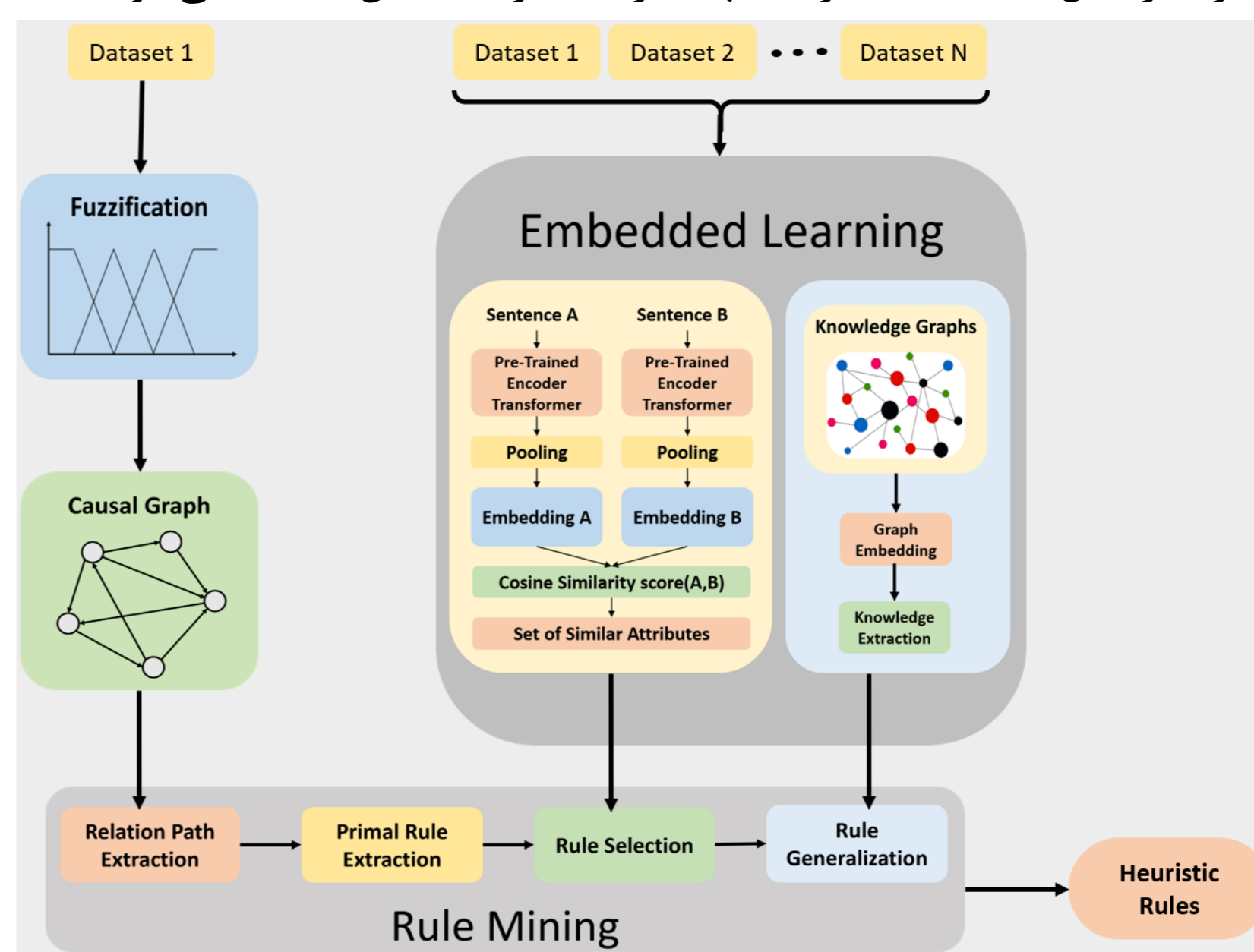
- Rich, P., Blokpoel, M., de Haan, R. et al.(2021) Naturalism, tractability and the adaptive toolbox. Synthese 198, 5749–5784.
- Simonne, L., Pernelle, N., Saïs, F., & Thomopoulos, R. (2022). Discovering Causal Rules in Knowledge Graphs using Graph Embeddings. In 2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) (pp. 95-102).
- Omran, P. G., Wang, Z., & Wang, K. (2019). Knowledge graph rule mining via transfer learning. In 2019 Advances in Knowledge Discovery and Data Mining: 23rd Pacific-Asia Conference, Part III 23 (pp. 489-500).
- Miller, G. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11), 39–41.

مقدمه

در زمینه داده‌کاوی، یکی از چالش‌های اساسی، استخراج الگوها و اطلاعات مفید از داده‌ها و استفاده از آنها برای تصمیم‌گیری‌های آینده است. هنگام تصمیم‌گیری در محیط‌هایی با حضور عدم قطعیت، روش‌های احتمالاتی که برای محاسبات ریسک استفاده می‌شوند، کارایی خود را از دست می‌دهند. قوانین هیوریستیک، به عنوان قوانینی که نزدیک به فهم و هوش انسانی هستند، در محیط‌هایی با حضور عدم قطعیت، عملکرد قابل توجهی از خود نشان می‌دهند. با این حال، چالش اصلی در اینجا یافتن بهترین قوانین هیوریستیک جهت تصمیم‌گیری در محیط‌های مختلف است. در این پژوهش، هدف بررسی و ارائه چارچوبی است که بتواند به صورت هوشمند و داده‌محور، قواعد هیوریستیکی استخراج کند که قابلیت اعمال بر روی گستره وسیعی از محیط‌ها را دارا باشد و تطبیق‌پذیر و تعمیم‌پذیر نسبت به شرایط متفاوت باشد. در همین راستا، ابتدا از روابط علی حاکم بر داده بهره می‌جویم، سپس با استفاده از گراف‌های دانش، قواعد تعمیم داده خواهند شد تا به شکل قواعد هیوریستیک در بیایند. در نهایت، با استفاده از معیار ER، عملکرد این قواعد بر روی دادگان جدید ارزیابی می‌شود. نتایج نشان می‌دهند که قواعد استخراج شده از این روش، قابل اعمال بر روی گستره‌ای از محیط‌ها با بازدهی مناسب می‌باشند، که این موضوع در حوزه‌هایی که دانش پیشین در آنها محدود یا دارای هزینه بالایی است اهمیت خود را نمایان می‌سازد.

مدل پیشنهادی

قواعد هیوریستیک به عنوان یک الگو یا روش ذهنی برای حل مسائل پیچیده و تصمیم‌گیری استفاده می‌شوند. این قواعد معمولاً برای رسیدن به راه‌حل‌های سریعتر و موثرتر در شرایطی که زمان یا منابع محدود است، به کار می‌روند. هر قاعده هیوریستیک را می‌توان به صورت یک درخت تصمیم نمایش داد که می‌تواند بر روی محدوده وسیعی از دامنه‌ها اعمال شود. به منظور پیاده‌سازی یک سیستم هوشمند جهت استخراج قوانین هیوریستیک، ابتدا باید مجموعه‌ای از قوانین ابتدایی از دادگان اولیه استخراج شود. بدین منظور از سیستم فازی‌سازی برای گسسته‌سازی ویژگی‌های پیوسته استفاده می‌شود و به نمونه‌های داده، برچسب‌های فازی مرتبط اختصاص داده خواهد شد. سپس برای استخراج قوانین اولیه، لازم است روابط علی حاکم بر ویژگی‌ها و نیز نحوه ارتباط آنها با خروجی، مورد بررسی قرار گیرد که در اینجا با بهره‌گیری از مدل GOLEM گراف جهت‌دار متناظر با دادگان ساخته می‌شود.



شکل 1: معماری مدل استخراج کننده قواعد هیوریستیک به کمک سیستم هوشمند داده محور

در گام بعد، قواعد اولیه تعمیم داده می‌شوند تا به شکل قوانین هیوریستیک درآیند. برای آنکه قوانین به دادگان جدید قابل تعمیم باشند، از مدل‌های از پیش آموزش‌دیده‌ی Transformer Encoder به منظور جایابی مفاهیم و کلمات استفاده می‌شود. با مقایسه جایابی ویژگی‌ها، آنهایی که میان دادگان اولیه و دادگان هدف مشترک می‌باشند استخراج و بر اساس آنها، قوانین اولیه استخراج شده، غربال می‌شوند. در مرحله بعد، به منظور تعمیم این قواعد انتخاب شده، از جایابی گراف‌های دانش استفاده می‌کنیم. در اینجا از دو گراف دانش WordNet و ConceptNet جهت یافتن ارتباطات میان گزاره‌های موجود در قواعد مرحله قبل و سایر مفاهیم انتزاعی موجود در گراف دانش و نوع ارتباط آنها با این گزاره‌ها استفاده می‌شود. به کمک روش‌های جایابی گراف و استخراج مفاهیم انتزاعی، می‌توان فرم تعمیم داده شده قوانین را بدست آورد و بدین ترتیب مجموعه‌ای تعمیم و تطبیق‌پذیر از قواعد هیوریستیک در پایان فرآیند چارچوب پیشنهادی حاصل می‌شود.