Graph Classification and Graph Regression

Sadaf Fatollahy

۱۱ تیر ۱۴۰۲

نام استاد: خانم دکتر طاهری
نام درس : گراف کاوی

فهرست مطالب

۱.۰ Graph classification

طبقه بندی گراف نوعی کار یادگیری ماشینی است که در آن یک مدل آموزش می بیند تا نمودارها را بر اساس ویژگی های ساختاری آنها طبقه بندی کند. برخلاف وظایف یادگیری ماشین سنتی که بر روی داده‌های با ابعاد ثابت مانند تصاویر یا متن عمل می‌کنند، طبقه‌بندی گراف شامل کار با داده‌هایی است که به صورت گراف نشان داده می‌شوند، که مجموعه‌ای از راس و یال‌هایی است که گره‌ها را به هم متصل می‌کنند.

در طبقه بندی گراف، هر گراف معمولاً به عنوان یک ماتریس مجاورت نشان داده می شود که روابط زوجی بین گره ها را در گراف نشان می دهد. هدف این مدل یادگیری تابعی است که نمایش گراف را به یک برچسب طبقه بندی نگاشت می کند و کلاسی را که گراف به آن تعلق دارد نشان می دهد.

طبقه بندی نمودار دارای طیف گسترده ای از کاربردها، از جمله طبقه بندی ترکیبات شیمیایی، پیش بینی عملکرد پروتئین، تجزیه و تحلیل شبکه های اجتماعی، و غیره است. به دلیل پیچیدگی داده های گراف و دشواری استخراج ویژگی های معنی دار از ساختار گراف، این یک کار چالش برانگیز است. با این حال، پیشرفت‌های اخیر در یادگیری عمیق و شبکه‌های عصبی گراف منجر به پیشرفت‌های قابل‌توجهی در عملکرد طبقه‌بندی گراف شده است و آن را به یک حوزه فعال تحقیقاتی در یادگیری ماشین و هوش مصنوعی تبدیل کرده است.

۲.۰ Chemical compound classification

طبقه بندی ترکیبات شیمیایی یکی از کاربردهای مهم طبقه بندی گراف است که در آن مدل های یادگیری ماشین برای پیش بینی خواص ترکیبات شیمیایی بر اساس ساختار مولکولی آنها آموزش می بینند. در این زمینه، هر ترکیب شیمیایی به عنوان یک گراف نشان داده می شود، که در آن گره ها نشان دهنده اتم ها و یال ها نشان دهنده پیوندهای شیمیایی بین اتم ها هستند.

ساختار مولکولی یک ترکیب شیمیایی می تواند تأثیر قابل توجهی بر خواص فیزیکی و شیمیایی آن مانند حلالیت، واکنش پذیری و سمیت داشته باشد. با آموزش یک مدل یادگیری ماشینی برای طبقه بندی ترکیبات شیمیایی بر اساس ساختار مولکولی آنها، می توان خواص آنها را پیش بینی کرد و ترکیباتی را با ویژگی های مطلوب برای کاربردهای خاص، مانند کشف دارو یا علم مواد شناسایی کرد.

شبکه‌های عصبی گرافی به دلیل توانایی آنها در مدل‌سازی روابط پیچیده بین اتم‌ها در یک مولکول، ثابت کرده‌اند که یک ابزار قدرتمند برای طبقه‌بندی ترکیبات شیمیایی هستند. GNN ها می توانند یاد بگیرند که ویژگی های مربوطه را از ساختار نمودار استخراج کنند، مانند وجود زیرساخت های خاص یا طول و نوع پیوندهای شیمیایی، و از این ویژگی ها برای پیش بینی خواص ترکیب استفاده کنند.

۱.۲.۰ Dataset

در این پروژه ما دیتاست BBBP را مورد مطالعه قرار دادیم. این دیتاست که مخففBlood–brain barrier penetration است به معنای نفوذ پذیری سد خونی – مغزی میباشد که از یک مطالعه اخیر در مورد مدل‌سازی و پیش‌بینی نفوذپذیری سد بدست امده است. این مجموعه داده، نفوذپذیری یک ترکیب به سد خونی مغزی را ثبت می کند و در حوزه مسئله های روانشناسی است. شامل 2039 گراف یا مولکول به همراه 127 ویژگی های راسی ، 12 ویژگی یالی و 200 ویژگی مولکولی میباشد که در زیر هر کدام توضیح داده شده است.

Table 2: Initial atom features

Feature	Description	Size
atom type	type of atom (ex. C, N, O), by atomic number	100
atomic mass	mass of the atom, divided by 100	1
#bonds	number of bonds the atom is involved in	6
#Hs	number of bonded hydrogen atoms	5
hybridization	sp, sp2, sp3, sp3d, or sp3d2	5
formal charge	integer electronic charge assigned to atom	5
chirality	unspecified, tetrahedral CW/CCW, or other	4
aromaticity	whether this atom is part of an aromatic system	1

Table 3: Initial bond features

Feature	Description	Size
bond type	single, double, triple, or aromatic	4
conjugated	whether the bond is conjugated	1
in ring	whether the bond is part of a ring	1
stereo	none, any, E/Z or cis/trans	6

شکل ۱: node features and edge features

ویژگی های مولکولی در نهایت با ویژگی های بدست امده از GNN ادغام شده و برای طبقه بندی گراف به کار میرود. ما با استفاده از GNN ها این موضوع را بررسی کردیم و نتایج زیر حاصل شد که بهترین نتیجه برای مدل ۴ بود.

ما برای سنجش مدل از معیار AUC-ROC استفاده کردیم. منحنی AUC-ROC یک ابزار سنجش عملکرد برای مسائل طبقه بندی ست. ROC یک منحنی احتمال است و AUC نشان دهنده درجه یا معیار تفکیک پذیری است. این نشان می دهد که مدل چقدر می تواند بین کلاس ها تمایز قائل شود. هرچه این عدد به یک نزدیک تر باشد یعنی مساخت زیر نمودار به یک نزدیک است و عملکرد بهتر است.همانطورکه از جدول مشخص است مدل ۴ بهترین عملکرد را داشته است.

	GNN	Message	Aggregation	Dropout	Batch-n	Test score
Model 1	GCN (2 layer)	default GCN	default GCN	False	False	0.63
Model 2	GNN2	u-add-v	mean	False	False	0.58
Model 3	GNN2	u-mul-v	sum	False	False	0.63
Model 4	GNN2(4 layer)	u-mul-v	sum	False	False	0.69
Model 5	GNN2(4 layer)	u-mul-v	sum	False	True	0.62
Model 6	GNN2(4 layer)	u-mul-v	sum	True	True	0.61

Table ۱: result of classification method on dataset

Graph regression ۳.۰

رگرسیون گرافی نوعی تکنیک یادگیری ماشینی است که شامل پیش‌بینی یک متغیر خروجی پیوسته از ورودی ساختار یافته گراف است. در رگرسیون گراف، داده های ورودی به صورت یک گراف نمایش داده می شوند، جایی که گره ها نشان دهنده راس های گراف و یال ها نشان دهنده روابط بین آن راس ها هستند. هدف از رگرسیون گراف، یادگیری نگاشت بین گراف ورودی و یک متغیر خروجی پیوسته، مانند مقدار عددی یا بردار است.

رگرسیون گراف می تواند در کاربردهای مختلفی مانند کشف دارو، علم مواد و پیش بینی ساختار پروتئین استفاده شود. به عنوان مثال، در کشف دارو، رگرسیون گراف را می توان برای پیش بینی میل پیوندی یک مولکول داروی کاندید به پروتئین هدف، با توجه به نمایش گرافی مولکول و پروتئین استفاده کرد. GNN ها را می توان با استفاده از توابع مختلف ، مانند میانگین مربعات خطا MSE یا میانگین خطای مطلق MAE آموزش داد، و می توان با استفاده از تکنیک های پس انتشار خطا بهینه سازی کرد.

Dataset ۱.۳.۰

در این پروژه برای قسمت رگرسیون ما از دیتاست ESOL استفاده کردیم که از داده های حلال در آب برای برخی از ترکیبات تشکیل شده است و در حوزه مسئله های شیمی فیزیک است. شامل ۱۱۲۸ گراف یا مولکول به همراه 127 ویژگی های راسی ، 12 ویژگی یالی و 200 ویژگی مولکولی میباشد. هدف ما بررسی میزان حلالیت این ترکیبات در اب است. با استفاده از GNN های متختلف نتایج زیر حاصل شد :

برای سنجش مدل از معیار RMSE استفاده کردیم که اختلاف میانگین بین مقادیر پیش بینی شده مدل آماری و مقادیر واقعی را اندازه گیری می کند. از نظر ریاضی، انحراف معیار باقیمانده ها است. باقیمانده ها فاصله بین خط رگرسیون و نقاط داده را نشان می دهندو در این مسئله هرچه مقدار ان پایین تر باشد بهتر است. با توجه به جدول مدل ۲ بهترین عملکرد را داشت.

	GNN	Message	Aggregation	Dropout	Batch-n	Test score
Model 1	GCN (2 layer)	default GCN	default GCN	False	False	2.57
Model 2	GNN2	u-add-v	mean	False	False	2.27
Model 3	GNN2	u-mul-v	sum	False	False	3.12
Model 4	GNN2(4 layer)	u-mul-v	sum	False	False	2.12
Model 5	GNN2(4 layer)	u-mul-v	sum	False	True	3.26
Model 6	GNN2(4 layer)	u-mul-v	sum	True	True	3.22

Table ۲: result of regression method on dataset