مساله‌ی پیش‌بینی لینک در گراف‌های دوبخشی

محمد مهدی سمیعی پاقلعه، ایمان تبریزیان

تابستان ۹۷

دانشگاه صنعتی امیرکبیر

دانشکده مهندسی کامپیوتر و فناوری اطلاعات

مقدمه

ما در این‌جا به توضیح و مطالعه‌ی پیش‌بینی لینک در گراف‌های دو بخشی از طریق مخصوص‌سازی روش‌های پیش‌بینی لینک به گراف‌های دوبخشی می‌پردازیم. در یک گراف، تابع پیش‌بینی لینک بین دو گره میزان مشابهت یا نزدیکی آن دو گره را بیان می‌کند. روش‌های پیش‌بینی یال برای گراف‌های معمولی، معمولا از تعداد و طول مسیرهای بین دو نقطه استفاده می‌کند. از آن‌جایی که در گراف‌های دوبخشی بین گره‌های بین دو مجموعه تنها مسیر به طول فرد وجود دارد در نتیجه این تابع‌هایی که برای گراف‌های عمومی به کار می‌رفتند به اندازه‌ی کافی کارا نیستند. در عوض می‌توانیم با استفاده از گراف کرنل‌هایی که مخصوص گراف‌های دو بخشی طراحی شده‌اند، مساله‌ی پیش‌بینی لینک را برای گراف‌های دو بخشی حل کنیم. ما در این مقاله چگونگی پیش‌بینی لینک براساس گراف کرنل‌ها را توضیح داده و در نهایت روی مجموعه داده‌ای مختلف آن را ارزیابی می‌کنیم.

مقدمه‌ای بر روش‌های پیش‌بینی لینک

روش‌های پیش‌بینی لینک به دو دسته‌ی محلی و سراسری تقسیم می‌شوند. در روش‌های پیش‌بینی لینک محلی معیارهایی مانند: تعداد همسایه‌های مشترک، Adamic یا Jacard سعی دارند تا با بررسی گره‌های واقع در حوالی راس‌های موردنظر به مساله پاسخ دهند. در روش‌های پیش‌بینی لینک سراسری تمرکز برروی استفاده از تعداد و طول مسیرهای بین دو راس موردنظر است.

پیش‌بینی به روش‌های جبری

در این‌جا ما با استفاده از قضایای جبرخطی که برای گراف‌ها وجود دارند سعی می‌کنیم تا یک روش پیش‌بینی لینک ارائه دهیم. ویژگی اساسی این روش‌ها قابل یادگیری بودن آن‌هاست. به این معنا که برای پیش‌بینی لینک مدلی به همراه پارامترها ارائه می‌شود و روش مقاله سعی در یادگیری پارامترها با استفاده از یک راه مشخص دارید.

ما در اینجا ماتریس مجاورت گراف G را با A نشان می‌دهیم. از آن‌جایی که A ماتریس مربعی است، می‌توانیم بنویسیم:

$$A = U\Lambda U^T$$

که U ماتریس بردارهای ویژه‌ی A ویژگی orthogonal بودن را دارد. Λ ماتریس قطری است که در درایه‌های روی قطر آن مقادیر ویژه ماتریس A هستند. به منظور پیش‌بینی لینک، تبدیل طیفی f را باید برروی A اعمال کنیم.

$$F(A) = UF(\Lambda)U^T$$

$F(\Lambda)$ به این معنی است که F روی هر یک از درایه‌های قطری ماتریس مقادیر ویژه اعمال شود.

کاهش مساله‌ی پیش‌بینی لینک به مساله‌ی curve fitting

ما در این بخش به توضیح چگونگی کاهش مساله‌ی پیش‌بینی لینک به مساله‌ی curve fitting می‌پردازیم. هدف مساله‌ی پیش‌بینی لینک با استفاده از گراف کرنل‌ها حداقل‌سازی مقدار زیر می‌باشد:

$$\begin{array}{ll} \min\limits_{F} & \| F(\mathbf{A}) - \mathbf{B} \|_F \\ s.t. & F \in \mathcal{S} \end{array}$$

که در آن A ماتریس آموزش و B ماتریس تست می‌باشد (در ادامه چگونگی تقسیم مجموعه یال‌ها به این ماتریس‌ها را توضیح خواهیم داد)

از بخش‌های قبل می‌دانیم که ماتریس A را می‌توان به صورت زیر به بردارها و مقادیر ویژه تجزیه کرد:

$$A = U\Lambda U^T$$

حال با جایگذاری عبارت فوق در رابطه‌ی حداقل‌سازی به عبارت زیر می‌رسیم:

$$\begin{aligned} & \| F(\mathbf{A}) - \mathbf{B} \|_F \\ = & \| \mathbf{U}F(\mathbf{\Lambda})\mathbf{U}^T - \mathbf{B} \|_F \\ = & \| F(\mathbf{\Lambda}) - \mathbf{U}^T\mathbf{B}\mathbf{U} \|_F \end{aligned}$$

اکنون به علت قطری بودن ماتریس Λ عبارت بالا به مساله‌ی زیر تبدیل خواهد شد:

$$\min_f \quad \sum_i (f(\mathbf{\Lambda}_{ii}) - \mathbf{U}_{\cdot i}^T\mathbf{B}\mathbf{U}_{\cdot i})^2$$

که پرواضح است که یک مساله‌ی curve fitting است. اکنون هدف ما پیدا کردن یک f با پارامترهای مناسب است که مساله بهینه‌سازی فوق را حل کند.

ارائه روش پیش‌بینی یال مبتنی بر گراف کرنل برای گراف های دوبخشی

از آن‌جایی که ماتریس مجاورت یک گراف دوبخشی به صورت زیر است:

$$A = [0\ R; R^T 0]$$

که R یک ماتریس مستطیلی می‌باشد. از آنجایی که $A^{2n} = [(RR^T)^n 0; 0\ (R^T R)^n]$ در نتیجه مشاهده می‌کنیم که در مساله‌ی پیش‌بینی یال بین راس های دو مجموعه یک گراف دو بخشی نیازی به A^{2n} نداریم. و در عوض باید تمرکز خود را ماتریس‌هایی به فرم $A^{2n+1} = [0\ (RR^T)^n R; R^T (RR^T)^n 0]$ قرار دهیم. در نتیجه بایستی تابع F را بدین نحو تعریف کنیم که:

$$\begin{aligned} F(\mathbf{R}\mathbf{R}^T)\mathbf{R} &= F(\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T\mathbf{V}\mathbf{\Lambda}\mathbf{U}^T)\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T \\ &= \mathbf{U}F(\mathbf{\Lambda}^2)\mathbf{\Lambda}\mathbf{V}^T \end{aligned}$$

اکنون اگر $G(\Lambda) = F(\Lambda^2)\Lambda$ تعریف کنیم و B را نیز ماتریس مستطیلی تست تعریف کنیم، هدف ما کمینه کردن مقدار زیر خواهد شد: (اثبات این مساله مانند مطلب قبلی خواهد بود با این تفاوت که ما این جا از تجزیه SVD برای ماتریس مستطیلی R استفاده کردیم و در مطلب قبلی از تجزیه ماتریس بردارهای ویژه)

$$G(\Lambda) - U^T B V$$

آزمایش و ارزیابی

ما دیتاست گراف 100k جهت آزمایش و ارزیابی این الگوریتم استفاده کردیم. روش آزمایش ما به این صورت است که ۷۰ درصد یال‌های ماترس کلی را به عنوان مجموعه‌ی داده‌ی برای آموزش مدل‌ها استفاده می‌کنیم. خود این ۷۰ درصد را نیز به دو دسته که خود ۳۰ درصد و ۷۰ درصد می‌باشد تقسیم می‌کنیم. که این ۷۰ درصد در نقش ماتریس A و ۱۰۰ درصد نیز در نقش ماتریس B می‌باشد(ماتریس‌های استفاده شده در آموزش مدل) پس از اینکه مدل‌های ما آموزش دیده شده‌اند، عملکرد آن‌ها روی ماتریس کلی را با استفاده معیار Mean Average Precision می‌سنجیم.

ما در اینجا برای توابع چندجمله‌ای و sinh مقدار MAP را محاسبه کردیم که تقریبا برابر با اعداد ذکر شده در مقاله می‌باشد.

	مقدار MAP در مقاله	مقدار MAP بدست آمده توسط ما
sinh	0.738	0.698895540622008
polynomial	0.822	0.8246037943796815
Neuman	0.631	0.7191786425165047

علت تفاوت اندک نتایج ما با نتایج مقاله در در نظر گرفتن مجموعه‌ی تست و یادگیری دارد. مجموعه‌ی تست در پیاده‌سازی ما تمام ۷۰ درصد یال‌هاست. حال آنکه در مقاله این مجموعه برابر با ۳۰ درصد ۷۰ درصد یال‌ها می‌باشد. توضیح آنکه اگر ماتریسی را به دو ماتریسA و B با سایزهای سی درصد و ۷۰ درصد سایز ماتریس اولیه افراز کنیم، مقاله‌ی اصلی سعی در ارائه‌ی یک F از A به B داشته حال آنکه ما سعی در ارائه‌ی یک F از A به A+B داشتیم.