



دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

پایان‌نامه‌ی کارشناسی ارشد
گرایش هوش مصنوعی

عنوان:

یادگیری بدون برد با شبکه‌های عمیق

نگارش:

سیدمحسن شجاعی

استاد راهنما:

دکتر مهدیه سلیمانی

تابستان ۱۳۹۵

صلى الله عليه وسلم

سپاس

سپاس گزارم.

چکیده

کلیدواژه‌ها: زمان‌بندی کارکنان، زمان‌بندی مدرسه، جستجوی خلاق، برنامه درسی.

فهرست مطالب

۱	۱ مقدمه
۲	۲ روش‌های پیشین
۳	۱-۲ پیش‌بینی ویژگی
۴	۱-۱-۲ پیش‌بینی ویژگی مستقیم و غیر مستقیم
۵	۲-۲ نگاشت‌های دو خطی
۹	۳ روش پیشنهادی
۱۰	۴ نتایج
۱۱	۵ جمع‌بندی
۱۱	۱-۵ جمع‌بندی
۱۱	۲-۵ کارهای آینده

فهرست شکل ها

فصل ۱

مقدمه

فصل ۲

روش‌های پیشین

در این فصل ابتدا یک چارچوب کلی برای روش‌های مورد استفاده در یادگیری بدون برد توصیف می‌شود. سپس روش‌های موجود طبق این چارچوب دسته‌بندی شده و مرور خواهند شد. پیش از تعریف و بیان رسمی مسئله یادگیری بدون برد، استفاده از اشتراک و تمایز برخی ویژگی‌ها میان دسته‌های مختلف در بینایی ماشین مورد بررسی قرار گرفته است [؟، ؟، ؟]. اما این روش‌ها به شناسایی دسته‌های کاملاً جدید از روی این ویژگی‌ها توجه نشان نداده‌اند. مسئله یادگیری تک‌ضرب^۱ هم یک مسئله نزدیک به یادگیری بدون برد است که پیش‌تر مورد بررسی بوده است [؟]. در حقیقت می‌توان یادگیری تک‌ضرب را حالت خاصی از یادگیری از صفر در نظر گرفت که در آن توصیف دسته‌های دیده نشده به صورت یک نمونه از آن دسته ارائه شده است [؟].

پدیده شروع سرد^۲ در سامانه‌های توصیه‌گر^۳ را نیز می‌توان از حالت‌های خاص یادگیری بدون برد در نظر گرفت که در آن برای یک کاربر یا مورد جدید پیشنهاد صورت می‌گیرد.

بیان مسئله یادگیری بدون برد به طور رسمی برای اولین بار در [؟] صورت گرفت. در آن‌جا دو رویکرد کلی برای حل مسئله یادگیری از صفر بیان می‌شود. یک روش که رویکرد فضای ورودی^۴ نامیده می‌شود، سعی در مدل کردن نگاشتی با دو ورودی دارد. یکی نمونه‌ها و دیگری توصیف دسته‌ها. این نگاشت برای نمونه‌ها و توصیف‌های مربوط به یک

^۱One-shot Learning

^۲cold start

^۳Recommender Systems

^۴input space view

دسته امتیاز بالا و برای نمونه‌ها و توصیفاتی که متعلق به دسته‌ی یکسانی نیستند مقادیر کوچکی تولید می‌کند. با تخمین زدن چنین نگاشتی روی داده‌های آموزش، دسته‌بندی نمونه‌های آزمون در دسته‌هایی که تا کنون نمونه‌ای نداشته‌اند ممکن خواهد شد. به این صورت که هر نمونه با توصیف دسته‌های مختلف به این تابع داده شده و متعلق به دسته‌ای که امتیاز بیشتری بگیرد، پیش‌بینی خواهد شد. در روش دیگر که رویکرد فضای مدل^۵ نام دارد، مدل مربوط به هر دسته (برای مثال پارامترهای دسته‌بند مربوط به آن)، به عنوان تابعی از توصیف آن دسته در نظر گرفته می‌شود.

ما در این فصل از دسته‌بندی دیگری برای مرور روش‌های پیشین استفاده می‌کنیم. برای این کار ابتدا معرفی یک چارچوب کلی برای انجام یادگیری بدون برد لازم است. دو رویکرد فوق نیز در این چارچوب قابل بیان هستند، این موضوع در بخش؟؟ که مثال‌هایی از این رویکردها مرور می‌شود، روشن‌تر خواهد شد.

می‌توان گفت که هر روش برای یادگیری بدون برد از سه قسمت تشکیل شده است که ممکن است به صورت مستقل یا همزمان انجام شوند؛ این سه قسمت عبارتند از:

۱. یادگرفتن نگاشتی از فضای تصاویر به فضای مشترک که آن را با ψ نشان می‌دهیم.

۲. نگاشت توصیف‌ها به فضای مشترک که آن را با ϕ نشان می‌دهیم.

۳. اختصاص برچسب به تصاویر

در ادامه به بررسی روش‌های ارائه شده برای مسئله یادگیری بدون برد با استفاده از چارچوب ارائه شده خواهیم پرداخت.

۱-۲ پیش‌بینی ویژگی

این دسته از روش‌ها عموماً به حالتی از مسئله یادگیری از صفر تعلق دارند که توصیف دسته‌ها از نوع بردار ویژگی باشد. در این حالت فضای مشترک همان فضای ویژگی‌ها در نظر گرفته می‌شود. به عبارت دیگر نگاشت ψ نگاشت همانی فرض شده و یادگرفته نخواهد شد. روش‌های اولیه ارائه شده برای یادگیری بدون برد از نوع پیش‌بینی ویژگی^۶ بوده‌اند و پس از آن هم قسمت قابل توجهی از روش‌ها در این دسته جای می‌گیرند که در ادامه آن‌ها را به تفصیل مرور می‌کنیم.

^۵model space view

^۶Attribute Prediction

۲-۱-۱ پیش‌بینی ویژگی مستقیم و غیر مستقیم

در [۴] با فرض این که ویژگی‌ها به صورت مستقل از یکدیگر قابل پیش‌بینی هستند دو رویکرد برای این کار ارائه می‌کند. پیش‌بینی ویژگی مستقیم^۷ و پیش‌بینی ویژگی غیر مستقیم^۸. مدل گرافی مورد استفاده در این دو رویکرد در تصویر؟؟ آمده است. در پیش‌بینی ویژگی مستقیم برچسب‌ها به شرط دانستن ویژگی‌های درون تصویر، از تصویر مستقل هستند. در این روش برای هر یک ویژگی‌ها یک دسته‌بند یاد گرفته می‌شود. با توجه به این که ویژگی‌ها برای تصاویر آزمون معین هستند این کار با استفاده از یک دسته‌بند احتمالی برای هر ویژگی قابل انجام است. در نهایت احتمال تعلق هر یک از برچسب‌های $u \in \mathcal{U}$ با استفاده از رابطه زیر بدست خواهد آمد.

$$P(z_u|x) = \sum_{c \in \{0,1\}^a} P(u|c)p(c|x) \quad (۱-۲)$$

از با توجه به فرض استقلال ویژگی داریم $P(c|x) = \prod_{n=1}^a P(c_n|x)$. برای محاسبه جمله $P(z_u|a)$ از قانون بیز استفاده می‌کنیم:

$$P(u|c) = \frac{P(u)P(c|u)}{P(a^u)} = \frac{P(u)\mathbb{1}(c = c^u)}{P(c^u)}$$

با جایگذاری در رابطه (؟؟) خواهیم داشت:

$$P(u|x) = \frac{P(u)}{P(c^u)} \prod_{n=1}^a P(a_n^u|x) \quad (۲-۲)$$

در نهایت برچسبی که احتمال فوق را بیشینه کند، پیش‌بینی مربوط به تصویر x خواهد بود.

در روش پیش‌بینی ویژگی غیر مستقیم، IAP تخمین $P(c_i|x)$ تغییر داده می‌شود؛ به این صورت که ابتدا یک دسته‌بند چند دسته‌ای یعنی $P(y_k|x)$ روی داده‌ها یاد گرفته می‌شود و سپس رابطه ویژگی‌ها و برچسب‌ها به صورت قطعی مدل می‌شود:

$$P(c_i|x) = \sum_{k=1}^s P(y_k|x)\mathbb{1}(c_i = c_i^{y_k}) \quad (۳-۲)$$

در نهایت در هر دو روش برچسب نهایی با تخمین MAP^۹ از رابطه زیر تعیین می‌شود:

$$\hat{y} = \arg \max_{u \in \mathcal{U}} P(u|x) = \arg \max_{u \in \mathcal{U}} \prod_{i=1}^a \frac{P(c_i^u|x)}{P(c_i^u)} \quad (۴-۲)$$

^۷Direct Attribute Prediction

^۸Indirect Attribute Prediction

^۹Maximum a Posteriori

روش ارائه شده در [۹] مشابه همین روش است با این تفاوت که احتمال مشاهده هر کدام ویژگی‌ها را هم در محاسبه دخیل می‌کند تا با وزن‌های متفاوت با توجه به اهمیتشان در دسته‌بندی نقش داشته باشند. ضعف بزرگ این روش‌ها فرض مستقل بودن ویژگی‌ها از یکدیگر است؛ چرا که این فرض در مسائل واقعی معمولاً برقرار نیست. برای مثال زمانی که ویژگی آبی بودن برای یک موجود در نظر گرفته می‌شود احتمال ویژگی پرواز کردن برای آن بسیار کاهش می‌یابد. مدل‌های گرافی برای در نظر گرفتن وابستگی‌های میان ویژگی‌ها به کار گرفته شده‌اند. نویسندگان [۹] برای در نظر گرفتن ارتباط بین ویژگی‌ها و ارتباط ویژگی‌ها با برچسب نهایی روش‌های مدل‌سازی موضوع^{۱۰} را از حوزه یادگیری در متن اقتباس می‌کنند. همچنین نویسندگان [۹] برای این کار یک چارچوب بر اساس مدل‌های گرافی احتمال معرفی می‌کنند. در این چارچوب یک شبکه بی‌زی^{۱۱} برای مدل کردن این روابط در نظر گرفته می‌شود و ساختار آن که نشان‌دهنده وابستگی یا استقلال ویژگی‌ها با هم یا با برچسب است، با کمک روش‌های یادگیری ساختار^{۱۲} شناخته می‌شود.

۲-۲ نداشت‌های دو خطی

حالت دیگری از چارچوب کلی معرفی شده در ابتدای فصل این است که نگاشت به فضای مشترک یک نگاشت دوخطی باشد. یعنی به این صورت که W نگاشتی خطی است که $x^T W$ تصویر x را به فضای توصیف‌ها نگاشته و W^C توصیف c را به فضای تصاویر می‌نگارد. در نهایت تابع مطابقت میان یک توصیف و تصویر به صورت زیر تعریف می‌شود:

$$F(x, c) = \phi(x)^T W \theta(y) \quad (5-2)$$

در این حالت، این که فضای مشترک کدام یک از فضاهای تصاویر یا توصیفات هستند، جواب روشی ندارد. نقطه‌ی قوت این روش‌ها در امکان پیچیده‌تر کردن تابع هزینه است. چرا که در حالتی که نگاشت خطی است مسائل بهینه‌سازی پیچیده‌تری نسبت به حالت غیر خطی قابل حل خواهند بود. یک انتخاب متداول برای تابع هزینه، توابع رتبه‌بند^{۱۳} هستند. با توجه به این که عموماً بعد از یادگیری این نگاشت، دسته‌ای که نزدیک‌ترین توصیف را (با معیاری مثل فاصله یا ضرب داخلی) دارد، به عنوان پیش‌بینی تولید می‌شود، چنین تابع هزینه‌ای یک انتخاب طبیعی است. چرا که مسئله‌ی نزدیکترین همسایه در اصل یک مسئله رتبه‌بندی است و استفاده از یک تابع هزینه‌ی رتبه‌بند برای یادگیری نگاشت

^{۱۰} Topic Modeling

^{۱۱} Bayesian Network

^{۱۲} Structure Learning

^{۱۳} ranking function

بهتر از مجموع مربعات است که تنها فاصله نقاط از برجسب خودشان را در نظر می‌گیرد [۴].

در [۴] تابع هزینه رتبه‌بند WSABIE [۴] که برای حاشیه‌نویسی تصاویر پیشنهاد شده، به مسئله یادگیری بدون برد انطباق می‌دهد. تابع هزینه WASBIE به این صورت تعریف شده است:

$$L(x_s, Y_s; W, \theta) = \frac{1}{N_s} \sum_{n=1}^{N_s} \lambda_{r_\Delta(x_n, y_n)} \sum_{y \in \mathcal{Y}} \max(\cdot, l(x_n, y_n, y)) \quad (۶-۲)$$

$$l(x_n, y_n, y) = \mathbb{1}(y \neq y_n) + \phi(x_n)^T W \theta(y) - \phi(x_n)^T W \theta(y_n) \quad (۷-۲)$$

که در آن λ_k یک تابع نزولی از k است. این تابع، پیش‌بینی اشتباه ویژگی‌ها را این گونه جریمه می‌کند که به ازای برجسب نادرستی که رتبه بالاتری از برجسب صحیح در دسته‌بندی دریافت کرده، جریمه‌ای متناسب با امتیاز برجسب ناصحیح در نظر گرفته می‌شود. ضریب نزولی λ_k میزان جریمه را برای برجسب‌های غلط در رتبه‌های بالا، بیشتر در نظر می‌گیرد. در انطباق برای یادگیری بدون برد، بهینه‌سازی تنها روی نگاشت W انجام شده و تابع θ دانسته فرض می‌شود: $\theta(y) = c_y$.

ایده‌ی بالا در [۴] ادامه داده شده است، در این حالت تابع مطابقت بین توصیف‌ها و تصاویر از رابطه (۵-۲) تعریف می‌شود. تابع هزینه ساده‌تر از حالت قبل به صورت

$$\frac{1}{N_s} \sum_{n=1}^{N_s} \max(\cdot, l(x_n, y_n, y))$$

در نظر گرفته شده که l همانند رابطه (۷-۲) است. هم‌چنین برای استفاده از چند توصیف به صورت هم‌زمان، تعریف تابع مطابقت به صورت زیر تعمیم داده می‌شود:

$$F(x, y; \{W\}_{1 \dots K}) = \sum_k \alpha_k \theta(x)^T W_k \phi_k(y) \quad (۸-۲)$$

$$s.t. \sum_k \alpha_k = 1$$

که $\phi_k(y)$ توصیف‌های مختلف از دسته‌ی y را نشان می‌دهد و W_1, \dots, W_K نگاشت‌های میان هر یک از این توصیف‌ها و فضای تصاویر را. وزن‌های α_k که میزان اهمیت یا اطمینان هر یک از توصیف‌ها را نشان می‌دهد، با اعتبارسنجی تعیین می‌شوند. روش ارائه شده با انواع اطلاعات جانبی سازگار است. اطلاعات جانبی که بر روی آن‌ها تست انجام شده

است شامل بردار ویژگی‌های دودویی یا پیوسته تعیین شده توسط انسان و نمایش برداری متون دایره‌المعارفی با روش‌های word2vec [۴] و GloVe [۵] است. هم‌چنین نویسندگان این پژوهش یک نسخه با نظارت از word2vec ارائه می‌دهند که در جریان آموزش آن از موضوع هر متن هم استفاده می‌شود.

روش فوق در

یک نحوه‌ی استفاده دیگر از نگاشت‌های دو خطی، دسته‌بندی مستقیم با این نگاشت است.

$$\underset{W \in \mathbb{R}^{d \times a}}{\text{minimize}} \|X_s^T W C_s - Y\|_{Fro}^2 + \Omega(W) \quad (9-2)$$

که در آن Ω یک جمله منظم‌سازی است. در این حالت اگر تبدیل را از فضای تصاویر به فضای ویژگی‌ها نگاه کنیم، نگاشت W باید تصاویر را به زیرفضایی عمود به تمامی بردار ویژگی‌های مربوط به برجسب‌های نادرست بنگارد. عملکرد خوب این روش، با وجود استفاده از تابع هزینه ساده مجموع مربعات خطا که در یادگیری ماشین تابع هزینه‌ی مناسبی برای دسته‌بندی به شمار نمی‌آید، به جمله منظم‌سازی آن نسبت داده می‌شود. جمله منظم‌سازی Ω به این صورت تعریف می‌شود:

$$\Omega(W) = \lambda \|W C_s\|_{Fro}^2 + \gamma \|X_s^T W\|_{Fro}^2 + \lambda \gamma \|W\|_{Fro}^2 \quad (10-2)$$

این جمله منظم‌سازی با دیدگاه نگاشت دوخطی طبیعی است. چرا که $W C_s$ را می‌توان یک دسته‌بند خطی روی فضای تصاویر در نظر گرفت و از طرفی $X_s^T W$ یک دسته‌بند روی بردارهای ویژگی است در نتیجه طبیعی است که پارامترهای این دو دسته‌بند با نرم فروبنیوس آن‌ها کنترل شود. استفاده از توابع نرم دوم برای خطا و منظم‌سازی در این روش باعث شده است که مسئله بهینه‌سازی جواب به صورت فرم بسته داشته باشد و زمان اجرا نسبت به سایر روش‌ها بسیار کمتر باشد. این روش در [۴] برای توصیفات متنی توسعه داده شده است. با توجه به ابعاد بالای داده‌های متنی و هم‌چنین نویز زیادی که در آن‌ها در مقایسه با بردارهای ویژگی وجود دارد، ماتریس تبدیل W به دو ماتریس تجزیه می‌شود:

$$W = V_x^T V_c \quad (11-2)$$

با این تجزیه از افزایش شدید تعداد پارامترها در اثر افزایش بعد بردار توصیف‌ها جلوگیری می‌شود. (دقت کنید که بعد C برابر $d \times a$ است) علاوه بر این V_c می‌تواند برای حذف قسمت‌های نویزی C به کار گرفته شود و V_x مانند W در حالت اصلی عمل کند یعنی پارامترهای یک دسته‌بند را از روی توصیف‌ها تولید کند. در نهایت تابع هزینه برای این روش

به صورت زیر تعریف می‌شود:

$$\min_{V_x, V_c} \|X_s^T + V_x^T V_c C\|_{Fro}^2 + \lambda_1 \|V_x^T V_c C\|_{Fro}^2 + \lambda_2 \|V_c^T\|_{2,1}^2 \quad (12-2)$$

که $\|M^T\|_{2,1} = \sum_i \|M_{(i)}\|_2$ و این نوع منظم‌سازی، ستون‌های ماتریس V_c را به سمت تنک بودن سوق خواهد داد. در واقع اگر λ_2 بزرگ انتخاب شود، V_c نقش یک ماتریس انتخاب ویژگی^{۱۴} را خواهد داشت. جمله‌های منظم‌سازی دیگر در (۱۰-۲) به دلیل تاثیر اندکشان در آزمایشات عملی حذف شده‌اند.

در [۹] نیز که برای اولین بار توصیف تنها نام برجسب دسته‌ها در نظر گرفته شده، از نگاشت دو خطی استفاده شده است. در این روش نام برجسب‌ها با استفاده از مدل نهان‌سازی کلمات word2vec کلمات به بردارهایی نگاشته می‌شوند. ابعاد فضای نهان‌سازی کلمات یک فرارامتر است که در این مقاله با اعتبار سنجی تعیین شده است. استخراج ویژگی از تصاویر با استفاده از شبکه عصبی کانولوشنال [۹] که روی دسته‌های دیده شده آموزش داده شده، انجام می‌شود. در نهایت یک تابع بیشترین حاشیه^{۱۵} برای یادگیری نگاشت دو خطی پیشنهاد می‌شود.

$$L((x_n, y_n); W) = \sum_{y \neq y_n} \max(0, \xi - x_n W c_{y_n} + x_n W c_y) \quad (13-2)$$

که در آن ξ حاشیه دسته‌بندی است. دسته‌بندی نمونه‌های جدید با نگاشتن x به فضای برجسب‌ها و استفاده از دسته‌بند نزدیکترین همسایه صورت می‌گیرد. در این پژوهش گزارش شده که نتایج چنین تابع هزینه‌ای بهتر از تابع هزینه مجموع مربعات بوده است، چرا که مسئله‌ی نزدیکترین همسایه در اصل یک مسئله رتبه‌بندی^{۱۶} است و استفاده از یک تابع هزینه‌ی رتبه‌بند برای یادگیری نگاشت بهتر از مجموع مربعات است که تنها فاصله نقاط از برجسب خودشان را در نظر می‌گیرد.

^{۱۴}feature selection

^{۱۵}Max margin

^{۱۶}ranking

فصل ۳

روش پیشنهادی

فصل ۴

نتایج

فصل ۵

جمع بندی

۱-۵ جمع بندی

۲-۵ کارهای آینده

Abstract

Keywords: Timetabling, School Timetabling Problem, Personnel Scheduling



Sharif University of Technology

Department of Computer Engineering

M.Sc. Thesis

Artificial Intelligence

Deep Zero-shot Learning

By:

Seyed Mohsen Shojaee

Supervisor:

Dr. Mahdaieh Soleymani

Summer 2017