Journal of Applied And Basic Machine Intelligence Research (ABMIR)

پژوهش‌های نظری و کاربردی هوش ماشینی

سال اول، شماره اول، بهار و تابستان ۱۴۰۲، صفحات ۸۷-۷۳

مروری بر روش‌های انتخاب ویژگی نیمه نظارتی تُنک مبتنی بر گراف

راضیه شیخ‌پور [1*]

گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه اردکان، اردکان، ایران

چکیده **مقاله پژوهشی**

در برخی از کاربردهای دنیای واقعی، داده‌هایی با ابعاد بالا وجود دارند که چالش‌های محاسباتی زیادی را ایجاد کرده‌اند. یکی از تکنیک‌های مؤثر برای کاهش ابعاد داده‌ها، انتخاب ویژگی است که با انتخاب زیرمجموعه مناسبی از ویژگی‌ها باعث سادگی مدل و بهبود کارایی آن می‌شود. در بسیاری از این کاربردها، برچسب زدن داده‌ها امری زمان‌بر و پرهزینه است که باعث می‌شود داده‌های برچسب‌دار کمی وجود داشته باشند و حجم عظیمی از داده‌های بدون برچسب در دسترس باشند. در چنین کاربردهایی، روش‌های انتخاب ویژگی نیمه نظارتی می‌توانند با استفاده از اطلاعات برچسب داده‌های برچسب‌دار و اطلاعات توزیع و ساختار هندسی داده‌های برچسب‌دار و بدون برچسب، فرایند انتخاب ویژگی را انجام دهند. در اکثر روش‌های انتخاب ویژگی نیمه نظارتی، با ایجاد یک گراف همسایگی، ویژگی‌های مناسب از طریق بررسی توانایی آن‌ها در حفظ ساختار هندسی گراف ارزیابی می‌شوند. در روش‌های کلاسیک انتخاب ویژگی نیمه نظارتی مبتنی بر گراف، ویژگی‌ها به‌صورت جداگانه ارزیابی می‌شوند و همبستگی بین ویژگی‌ها در هنگام انتخاب ویژگی در نظر گرفته نمی‌شود. روش‌های انتخاب ویژگی تُنک با در نظر گرفتن همبستگی بین ویژگی‌ها، ماتریس انتقال بهینه تُنک برای انتخاب ویژگی را محاسبه می‌نمایند. در این مقاله با بررسی روش‌های یادگیری نیمه نظارتی، مروری بر روش‌های انتخاب ویژگی نیمه نظارتی تُنک مبتنی بر گراف انجام می‌شود که با استفاده از عبارت تنظیم مبتنی بر مدل‌های تُنک و با ایجاد گراف همسایگی، ویژگی‌های مناسب را انتخاب می‌کنند. این روش‌ها ضمن برطرف کردن مشکل روش‌های انتخاب ویژگی کلاسیک، با ایجاد یک گراف همسایگی از داده‌ها ماتریس انتقال بهینه تُنک برای انتخاب ویژگی را محاسبه می‌نمایند.

تاریخ دریافت:
۱۴۰۱/۰۴/۰۵

تاریخ پذیرش:
۱۴۰۱ /۰۷ /۰۳

کلیدواژه‌ها:
انتخاب ویژگی نیمه نظارتی، یادگیری نیمه نظارتی، مدل‌های تُنک، گراف، لاپلاسین گراف

نویسنده مسئول:
rsheikhpour@ardakan.ac.ir

doi : 10.22034/ABMIR.2022.2899

۱- مقدمه

یکی از چالش‌های یادگیری در بسیاری از کاربردهای دنیای واقعی نظیر پردازش تصویر و بیوانفورماتیک، ابعاد بالا برای داده‌ها است که باید بتوان با استفاده از روش‌های انتخاب ویژگی، مناسب‌ترین ویژگی‌ها را برای دستیابی به مدل‌های یادگیری با توانایی پیش‌بینی بالا انتخاب کرد. انتخاب ویژگی باعث بهبود کارایی و سادگی مدل می‌شود و نقش مهمی در درک و تجسم داده‌ها ایفا می‌کند [۱,۲].

از دیگر چالش‌های موجود در این کاربردها، وجود تعداد اندک داده‌های برچسب‌گذاری شده است که با صرف زمان و هزینه زیادی به دست می‌آیند، درحالی‌که داده‌های بدون برچسب به‌آسانی در دسترس هستند. در چنین مسائلی، روش‌های یادگیری نظارتی ممکن است با مشکل بد تعریف شدن یادگیری روبه‌رو می‌شوند. بد تعریف شدن یادگیری به معنای آن است که برای یک فضای فرضیه مشخص، ممکن است توابع زیادی با داده‌های برچسب‌دار سازگار باشند. در این حالت امکان تصمیم‌گیری برای انتخاب تابع برچسب دهی مناسب دشوار خواهد بود. برای حل این مشکل، در یادگیری نیمه نظارتی علاوه بر داده‌های برچسب‌دار، از داده‌های بدون برچسب نیز در فرایند یادگیری استفاده می‌شود [۳,٤]. استفاده از داده‌های بدون برچسب در حقیقت معادل با یادگیری توزیع داده‌ها است و هر فرایند یادگیری برای همگرا شدن نیازمند یک دانش پیشین است. یکی از مهم‌ترین و کاربردی‌ترین فرضیات نیمه نظارتی، فرض منیفلد است. یادگیری نیمه نظارتی با فرض منیفلد، با درنظرگرفتن هندسه داده‌ها که از طریق داده‌های بدون برچسب به‌دست می‌آید، می‌تواند پاسخ‌گوی کاربردهای مختلفی باشد که در آن‌ها داده‌های برچسب‌دار کم است یا به‌سختی به دست می‌آید [۵]. ازآنجایی‌که یادگیری نیمه نظارتی نیازمند تلاش کمتری برای جمع‌آوری نمونه‌های برچسب‌دار است و باوجود تعداد محدودی نمونه برچسب‌دار به‌دقت بالایی دست می‌یابد، به همین دلیل، هم در تئوری و هم در عمل بسیار موردتوجه قرارگرفته است [٦,۷]. با توجه به پیشرفت‌های به وجود آمده در یادگیری نیمه نظارتی، روش‌های انتخاب ویژگی نیمه نظارتی نیز ارائه‌شده‌اند تا بتوانند با استفاده از داده‌های برچسب‌دار و بدون برچسب، ویژگی‌های مناسب‌تری را تشخیص دهند. در روش‌های انتخاب ویژگی نیمه نظارتی، از برچسب داده‌های برچسب‌دار و اطلاعات توزیع و ساختار هندسی داده‌های برچسب‌دار و بدون برچسب برای انتخاب ویژگی‌ها استفاده می‌شود [۸].

شیخ‌پور و همکاران در سال ۲۰۱۷ [۸]، به بررسی جامع انواع روش‌های انتخاب ویژگی نیمه نظارتی پرداختند و این روش‌ها را بر اساس دو دیدگاه تقسیم‌بندی عمومی روش‌های انتخاب ویژگی و انواع روش‌های یادگیری نیمه نظارتی تقسیم‌بندی کردند. اکثر روش‌های انتخاب ویژگی نیمه نظارتی از دیدگاه اول در دسته‌بندی روش‌های فیلتر و از دیدگاه دوم در دسته‌بندی روش‌های مبتنی بر گراف قرار می‌گیرند. روش‌های فیلتر بدون نیاز به الگوریتم یادگیری، بر اساس معیاری خاص به رتبه‌بندی ویژگی‌ها می‌پردازند و پیچیدگی زمانی بالایی ندارند. در روش‌های مبتنی بر گراف، با استفاده از داده‌های برچسب‌دار و بدون برچسب یک گراف همسایگی ساخته می‌شود و ویژگی‌ها از طریق توانایی‌شان در حفظ ساختار گراف ارزیابی می‌شوند [۸].

روش‌های انتخاب ویژگی نیمه نظارتی کلاسیک مبتنی بر گراف نظیر امتیاز لاپلاسین نیمه نظارتی [۹–۱۱]، اهمیت ویژگی‌ها را به‌صورت جداگانه ارزیابی می‌کنند و اطلاعات مفید همبستگی بین ویژگی‌ها را در فرایند انتخاب ویژگی در نظر نمی‌گیرند. برای حل این مسئله، روش‌های انتخاب ویژگی تُنک [۱۲–۱٦] ارائه‌شده‌اند که با در نظر گرفتن اطلاعات مفید همبستگی بین ویژگی‌ها، ماتریس انتقال بهینه تُنک برای انتخاب ویژگی را محاسبه می‌نمایند.

با توجه به اهمیت روش‌های انتخاب ویژگی نیمه نظارتی تُنک مبتنی بر گراف، پژوهش‌های مختلفی این روش‌ها را ارائه داده‌اند ولی تاکنون بررسی و مرور جامعی در خصوص این روش‌ها انجام‌نشده است. در پژوهش انجام‌شده توسط شیخ‌پور و همکاران [۸]، انواع روش‌های انتخاب ویژگی نیمه نظارتی موردبررسی قرار گرفتند و بر اساس ویژگی‌ها و خصوصیات روش‌ها، دسته‌بندی‌های مختلفی ارائه شد و هر روش دریکی از دسته‌بندی‌ها قرار گرفت. یکی از دسته‌بندی‌های ارائه‌شده در آن پژوهش، روش‌های انتخاب ویژگی نیمه نظارتی تُنک مبتنی بر گراف بود که خصوصیات کلی این روش‌ها بیان شد ولی انواع روش‌های ارائه‌شده در این حوزه به‌طور کامل موردبررسی قرار نگرفت. در این مقاله، ضمن بررسی

روش‌های یادگیری نیمه نظارتی و مفاهیم کلی انتخاب ویژگی نیمه نظارتی، مروری بر روش‌های انتخاب ویژگی نیمه نظارتی تُنک مبتنی بر گراف انجام می‌شود و خصوصیات، مزایا و معایب هر یک از آن‌ها بیان می‌شود. در این روش‌ها با استفاده از یک عبارت تنظیم مبتنی بر مدل‌های تُنک، همبستگی بین ویژگی‌ها در نظر گرفته می‌شود و با ساخت یک گراف همسایگی با استفاده از داده‌های برچسب‌دار و بدون برچسب و محاسبه لاپلاسین گراف، ماتریس انتقال تُنک بهینه به دست می‌آید که برای انتخاب ویژگی استفاده می‌شود.

ادامه این مقاله به‌صورت زیر سازماندهی می‌شود. در بخش دوم یادگیری نیمه نظارتی شرح داده‌شده و انواع روش‌های یادگیری نیمه نظارتی بیان می‌شوند. در بخش سوم، انتخاب ویژگی نیمه نظارتی توصیف می‌شود و در بخش چهارم، انتخاب ویژگی نیمه نظارتی تُنک مبتنی بر گراف شرح داده می‌شود. در بخش پنجم، بحثی در خصوص هر یک از روش‌ها، مزایا و معایب آن‌ها ارائه می‌شود. در پایان، در بخش پنجم، نتیجه‌گیری مقاله بیان می‌شود.

۲ – یادگیری نیمه نظارتی

مسائل یادگیری را می‌توان به سه دسته نظارتی، غیر نظارتی و نیمه نظارتی تقسیم نمود. در یادگیری نظارتی، آموزش فقط به کمک داده‌های برچسب‌دار انجام می‌شود. در یادگیری غیر نظارتی، در زمان یادگیری برچسب هیچ داده‌ای مشخص نیست و صرفاً از بردارهای ویژگی داده‌ها برای یادگیری استفاده می‌شود، مانند مسئله خوشه‌بندی که در آن هدف افراز داده‌ها به تعدادی خوشه است به‌گونه‌ای که داده‌ها در هر خوشه با یکدیگر مشابه و در خوشه‌های متفاوت از یکدیگر متفاوت باشند. در برخی از کاربردها، تعیین برچسب داده‌ها هزینه‌بر بوده و مستلزم صرف زمان زیادی است، درحالی‌که داده‌های بدون برچسب به‌آسانی در دسترس هستند. بنابراین، استفاده از روش‌های نیمه نظارتی که بتوانند از داده‌های بدون برچسب حداکثر استفاده را نمایند و کمبود داده‌های برچسب‌دار را جبران نمایند، در کاربردهای تئوری و عملی بسیار ارزشمند است [۵,۱۷,۱۸]. در یادگیری نیمه نظارتی، در زمان آموزش علاوه‌بر بردارهای ویژگی داده‌های برچسب‌دار و برچسب آن‌ها، از بردارهای ویژگی داده‌های بدون برچسب هم استفاده می‌شود. این داده‌ها در تشخیص نحوه پراکندگی نقاط در فضای ویژگی به ما کمک می‌کنند [۸].

یکی از روش‌های حل مسئله نیمه نظارتی، درنظرگرفتن مسئله به‌صورت غیر نظارتی به انضمام تعدادی شرط یا قید بر روی‌داده‌های برچسب‌دار است. درروش دیگر حل مسئله نیمه نظارتی، به مسئله به‌صورت نظارتی نگریسته می‌شود و در ضمن اطلاعات اضافی در مورد توزیع داده‌های ورودی که متناظر با داده‌های بدون برچسب است به آن اضافه می‌شود. معمولاً از روش دوم برای حل مسئله نیمه نظارتی استفاده می‌شود [۸].

مسائل یادگیری نیمه نظارتی را می‌توان به دو صورت استقرایی[1] و ورارسانی[2] تقسیم نمود. هدف یادگیری نیمه نظارتی استقرایی این است که از روی یک مجموعه آموزشی شامل نمونه‌های برچسب‌دار و بدون برچسب، مدلی ساخته شود که بتواند برچسب داده‌های جدید را تعیین نماید. اما در یادگیری نیمه نظارتی ورارسانی، هدف این است که تنها برچسب داده‌های بدون برچسب که در فرایند آموزش حضورداشته‌اند، تعیین شود [۱۷,۱۸].

۲-۱ فرض همواری نیمه نظارتی

برای استفاده از روش‌های نیمه نظارتی، باید فرض‌های به‌خصوصی برقرار باشد که این فرض‌ها همان دانش پیشین هستند. رایج‌ترین فرضی که در یادگیری نیمه نظارتی وجود دارد، فرض همواری است که بیان می‌کند اگر دونقطه x_1 و x_2 در یک ناحیه با چگالی بالا نزدیک به هم باشند، برچسب‌های متناظر آن‌ها هم باید نزدیک به هم باشند [۵]. فرض همواری نیمه نظارتی به‌طور خاص در دو فرض خوشه[3] و فرض منیفلد[4] نمود پیدا می‌کند [۵,۱۹].

فرض خوشه بیان می‌کند که داده‌های موجود در یک خوشه، احتمالاً برچسب‌های مشابهی دارند. اگر خوشه‌ها نقاط به هم متصل و همبند در نواحی چگال تعریف شوند، آنگاه این فرض حالت خاصی از فرض همواری می‌شود.

[1] Inductive
[2] Transductive
[3] Cluster assumption
[4] Manifold assumption

فرض منیفلد بیان می‌کند (در فضای ورودی با بُعد بالا)، داده‌ها (تقریباً) روی یک منیفلد با بعد پایین‌تر قرار دارند و تابع جداساز روی منیفلد داده‌ها هموار است. این فرض حالت خاصی از فرض همواری نیمه نظارتی است. اگر منیفلد داده‌ها را ناحیه چگال فضا در نظر بگیریم، دونقطه که فاصله ژئودزیک آن‌ها روی منیفلد کم باشد، احتمالاً برچسب‌های مشابهی خواهند داشت. برای استفاده از فرض منیفلد نیاز است تا به‌نوعی منیفلد داده‌ها تخمین زده شود. بدین منظور روش‌های مبتنی بر گراف مطرح‌شده‌اند که در سال‌های اخیر، یکی از فعال‌ترین زمینه‌ها در حوزه یادگیری نیمه نظارتی بوده‌اند. در این حالت، منیفلد داده‌ها با یک گراف مدل می‌شود. در بخش ۲-۳ به‌صورت مفصل‌تری، یادگیری نیمه نظارتی با فرض منیفلد بیان می‌شود.

۲-۲ روش‌های یادگیری نیمه نظارتی

به‌طورکلی روش‌های یادگیری نیمه نظارتی را می‌توان به پنج دسته روش‌های مولد، روش‌های خودآموز، روش‌های آموزش همراه با همکاری، ماشین‌های بردار پشتیبان نیمه نظارتی (S^3VMs) و روش‌های مبتنی بر گراف تقسیم کرد [۸,۱۷]. در ادامه این روش‌ها شرح داده می‌شوند.

روش‌های مولد

در روش‌های مولد، ابتدا یک مدل پارامتری برای تابع توزیع نقاط (مثلاً توزیع گاوسی) انتخاب می‌شود که آن را با $P(x|y,\theta)$ نشان می‌دهیم. x نمونه آموزشی، y مقدار برچسب و θ مدل را بیان می‌کنند. سپس $P(y)$ از روی‌داده‌های برچسب‌دار تخمین زده می‌شود. احتمال وقوع نقاط با توجه به تابع توزیع هر دسته برحسب پارامترهای مدل، به‌صورت تحلیلی محاسبه می‌شود. سپس با اعمال قانون بیز می‌توان تابع توزیع برچسب در هر نقطه را محاسبه کرد. در روش‌های مولد، معمولاً هدف بیشینه کردن این احتمال وقوع یا به‌طور معادل بیشینه کردن درست‌نمایی آن‌ها نسبت به پارامترهای مدل است. از روش‌های مختلف می‌توان برای بهینه کردن پارامترهای مدل نسبت به میزان درست‌نمایی استفاده کرد [۵,۸].

روش‌های خودآموز

در روش‌های خودآموز، از پیش‌بینی سیستم یادگیر برای فرایند یادگیری استفاده می‌شود. ایده اصلی به این صورت است که یک یادگیر اولیه از روی‌داده‌های آموزشی برچسب‌دار ساخته می‌شود. سپس یک زیرمجموعه از داده‌های بدون برچسب به همراه برچسبی که برای آن‌ها پیش‌بینی‌شده است، انتخاب می‌شود. تعداد اعضای این زیرمجموعه بسیار کم است و ازاین‌بین، مطمئن‌ترین پیش‌بینی‌ها انتخاب می‌شوند. این زیرمجموعه به مجموعه داده‌های برچسب‌دار افزوده می‌شود و فرایند یادگیری دوباره انجام می‌شود و الگوریتم به همین صورت تکرار می‌شود [۵,۸].

روش‌های آموزش همراه با همکاری

در این روش، ابتدا دو دسته‌بند با استفاده از دو مجموعه مستقل از ویژگی‌ها (دیدها) آموزش می‌بینند. سپس به‌طور تکراری، داده‌های بدون برچسب جهت دسته‌بندی به دو دسته‌بند داده می‌شوند. داده‌هایی که توسط هرکدام از این دسته‌بندها با اطمینان بالایی دسته‌بندی‌شده‌اند، از مجموعه داده‌های بدون برچسب خارج‌شده و به دسته‌بند دیگر داده می‌شوند تا خود را به‌روز نماید. به‌روز شدن دسته‌بندها می‌تواند تا دسته‌بندی همه نمونه‌ها ادامه پیدا کند.

در بسیاری از کاربردهای دنیای واقعی، دو مجموعه مستقل از ویژگی‌ها وجود ندارد. بنابراین محققان زیادی برای حل این مسائل، روش آموزش همراه با همکاری را توسعه دادند. روش‌های توسعه‌یافته آموزش همراه با همکاری مبتنی بر یادگیری تجمیعی هستند. در روش‌های یادگیری تجمیعی، تعدادی دسته‌بند پایه وجود دارد که می‌توان با ترکیب نتایج آن‌ها به‌دقت بالاتری رسید. متداول‌ترین روش‌ها برای ترکیب نتایج دسته‌بندها، میانگین‌گیری و استفاده از رأی اکثریت هستند [۲۰,۲۱].

روش‌های ماشین بردار پشتیبان نیمه نظارتی

هدف روش‌های دسته‌بندی ماشین بردار پشتیبان نیمه نظارتی، استفاده از داده‌های برچسب‌دار و بدون برچسب برای ساخت تابع تصمیم‌گیری $W^TX- b$ است [۱۷]. ماشین بردار پشتیبان نیمه نظارتی سعی می‌کند حاشیه بین داده‌های دودسته (اعم از برچسب‌دار و

بدون برچسب) بیشینه شود. دانش پیشین یا فرض مدل در ماشین‌های بردار پشتیبان نیمه نظارتی به‌صورت زیر بیان می‌شود: "داده‌های بدون برچسب که دودسته مختلف را به وجود می‌آورند، با حاشیه‌های بزرگی از هم جدا می‌شوند."

در ماشین بردار پشتیبان نیمه نظارتی، اگر تعداد داده‌های بدون برچسب u باشد، $2u$ برچسب دهی مختلف برای داده‌ها وجود دارد که معادل با ساخت همین تعداد ماشین بردار پشتیبان است. در این ماشین‌ها، ماشینی که بزرگ‌ترین حاشیه را ایجاد می‌کند ماشین بردار پشتیبان نیمه نظارتی خواهد بود. پیداکردن یک راه‌حل برای ماشین بردار پشتیبان نیمه نظارتی کاری دشوار است و هسته اصلی تحقیقات بر روی ماشین بردار پشتیبان نیمه نظارتی را تشکیل می‌دهد.

روش‌های مبتنی بر گراف

در روش‌های مبتنی بر گراف، ابتدا گرافی متشکل از داده‌های برچسب‌دار و بدون برچسب ساخته می‌شود. رأس‌های این گراف نمونه‌ها هستند و یال‌های آن‌که می‌توانند وزن‌دار باشند، نمایانگر شباهت نمونه‌ها هستند. معمولاً یال‌های گراف بدون جهت هستند. پس از ساخت گراف، از یکی از روش‌های استنتاج برچسب برای پیش‌بینی برچسب نقاط استفاده می‌شود [۱۷،۲۲].

۲-۳- یادگیری نیمه نظارتی با فرض منیفلد

در بخش ۲-۱ بیان شد که برای استفاده از یادگیری نیمه نظارتی باید پیش‌فرض‌هایی ازجمله فرض منیفلد وجود داشته باشد. منیفلد زیر فضای خمیده‌ای است که داده‌های یادگیری در فضای ویژگی‌ها روی آن قرار می‌گیرند. در بسیاری از مسائل یادگیری ماشین، داده‌های یادگیری روی یک منیفلد با ابعاد بسیار کوچک‌تر از فضای ویژگی‌ها قرار می‌گیرند. در بیشتر این مسائل، برچسب داده‌ها یک تابع هموار روی منیفلد داده‌ها است. به این فرض، اصطلاحاً فرض منیفلد گفته می‌شود.

چون معمولاً در مسائل یادگیری ماشین، تعداد محدودی داده آموزشی و آزمون وجود دارد، محاسبه منیفلد داده‌ها امکان‌پذیر نیست. به همین دلیل اکثر روش‌هایی که در این حوزه گسترش‌یافته‌اند، از گراف همسایگی به‌عنوان تخمینی برای منیفلد استفاده می‌کنند که روش ساخت این گراف، نقش اساسی در بهبود کیفیت روش‌های جداسازی نیمه نظارتی بر مبنای فرض منیفلد دارد [۲۲–۲۴].

بیان رسمی فرض منیفلد

فرض کنید که مجموعه داده‌های برچسب‌دار با l نمونه به‌صورت $X_l = [x_1, \dots, x_l]$ با برچسب‌های $Y_l = [y_1, \dots, y_l]$ و داده‌های بدون برچسب به‌صورت $X_u = [x_{l+1}, \dots, x_{l+u}]$ باشند به‌طوری‌که $x_i \in R^d\ (1 \le i \le n)$ و $y_i \in \{-1,1\}$ و $X = X_l + X_u$. تابع برچسب برای داده i ام به‌صورت f_i درنظرگرفته می‌شود. در این‌صورت فرض منیفلد به‌صورت رابطه زیر بیان می‌شود [۲۳،۲۴]:

$$\hat{S}(f) = \sum_{i,j=1}^{l+u} S_{ij}(f_i - f_j)^2 = F^T L F \qquad (۱)$$

که $F = (f_1, \dots, f_{l+u})$ تابع برچسب و S ماتریس وزن گراف است که به‌صورت زیر تعریف می‌شود:

$$S_{ij} = \begin{cases} e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}} & \text{if } x_i \in KNN(x_j) \text{ or } x_j \in KNN(x_i) \\ 0 & \text{otherwise} \end{cases} \qquad (۲)$$

در این گراف، گره i ام متناظر با نمونه (x_i) ام است. اگر x_i و x_j نزدیک باشند (x_i در میان یکی از k نزدیک‌ترین همسایه‌های x_j است یا x_j در میان یکی از k نزدیک‌ترین همسایه‌های x_i است)، یک یال وزن‌دار بین گره i و j قرار می‌گیرد.

در رابطه (۱)، $L=D-S$ لاپلاسین گراف است که D ماتریس قطری درجه رئوس گراف همسایگی است به گونه‌ای‌که $D_{ii} = \sum_{j=1}^{l+u} S_{ij}$.

در روش‌های مبتنی بر گراف، تابع برچسب F باید خصوصیات زیر را برآورده سازد [۱۷]:

- تقریب تابع برچسب $F(x)$ بر روی‌داده‌های برچسب‌دار، نزدیک به برچسب واقعی آنها باشد.
- تابع برچسب F باید روی کل گراف هموار باشد.

که شرایط فوق در چارچوب تنظیم بیان می‌شوند. شرط اول توسط تابع زیان و شرط دوم توسط نوع خاصی از تنظیم مبتنی بر گراف بیان می‌شود. روش‌های متنوعی جهت تخمین برچسب F وجود

دارند که تفاوتشان در انتخاب تابع زیان و همچنین روش تنظیم است.

بنابراین برای حل مسئله یادگیری تابع برچسب، باید مسئله زیر حل شود:

$$\min_{F} \sum_{i=1}^{l} (f_i - y_i)^2 + \gamma F^T L F \qquad (۳)$$

۳- انتخاب ویژگی نیمه نظارتی

امروزه با توجه به پیشرفت در فناوری، در بسیاری از کاربردها داده‌هایی با ابعاد بالا و تعداد زیاد ویژگی به وجود آمده‌اند که باعث به وجود آمدن چالش‌های محاسباتی زیادی شده‌اند. انتخاب ویژگی یکی از روش‌های مؤثر برای کاهش ابعاد داده‌ها است که زیرمجموعه مفیدی از ویژگی‌ها را انتخاب کرده و ویژگی‌های نامناسب را حذف می‌کند و باعث اجتناب از بیش‌برازش در هنگام ساخت مدل، بهبود کارایی و سادگی مدل می‌شود [۷، ۸].

با توجه به برچسب کلاس‌ها، روش‌های انتخاب ویژگی به سه دسته نظارتی، غیر نظارتی و نیمه نظارتی تقسیم می‌شوند. در روش‌های انتخاب ویژگی نظارتی، با استفاده از داده‌های برچسب‌دار اهمیت ویژگی‌ها ارزیابی می‌شود. روش‌های انتخاب ویژگی غیر نظارتی، بدون در نظر گرفتن برچسب داده‌ها به ارزیابی ویژگی می‌پردازند. روش‌های انتخاب ویژگی نظارتی با استفاده از تعداد زیادی داده برچسب‌دار، کارایی بالاتری نسبت به روش‌های انتخاب ویژگی غیر نظارتی دارند. ازآنجاکه در بسیاری از کاربردهای دنیای واقعی، به دلیل زمان‌بر و پرهزینه بودن، داده‌های برچسب‌دار زیادی وجود ندارد، روش‌های انتخاب ویژگی نظارتی نمی‌توانند کارایی قابل قبولی داشته باشند. راه‌حل مقابله با این مسئله، استفاده از روش‌های انتخاب ویژگی نیمه نظارتی است که با استفاده از برچسب داده‌های برچسب‌دار و خصوصیات توزیع و هندسی داده‌های برچسب‌دار و بدون برچسب مناسب‌ترین ویژگی‌ها را انتخاب می‌کنند [۸،۲۷].

روش‌های انتخاب ویژگی نیمه نظارتی از دو دیدگاه تقسیم‌بندی می‌شوند. از دیدگاه تقسیم‌بندی عمومی روش‌های انتخاب ویژگی، این روش‌ها به سه دسته فیلتر، یادگیرمبنا و تعبیه‌شده تقسیم می‌شوند. از دیدگاه انواع روش‌های یادگیری نیمه نظارتی، روش‌های انتخاب ویژگی نیمه نظارتی را می‌توان به روش‌های انتخاب ویژگی نیمه نظارتی مبتنی بر گراف، روش‌های انتخاب ویژگی نیمه نظارتی مبتنی بر یادگیری خودآموز، روش‌های انتخاب ویژگی نیمه نظارتی مبتنی بر آموزش همراه با همکاری، روش‌های انتخاب ویژگی نیمه نظارتی مبتنی بر ماشین بردار پشتیبان و سایر روش‌ها تقسیم‌بندی نمود [۸].

اکثر روش‌های انتخاب ویژگی نیمه نظارتی به‌صورت فیلتر و مبتنی بر گراف هستند. روش‌های مبتنی بر گراف دارای کارایی قابل قبولی در انتخاب ویژگی نیمه نظارتی هستند. این روش‌ها با شناسایی ساختار تفکیکی و هندسی داده‌ها، فرایند انتخاب ویژگی را انجام می‌دهند. درواقع، در روش‌های مبتنی بر گراف با ساخت گراف همسایگی و محاسبه لاپلاسین گراف، فرایند انتخاب ویژگی از طریق اندازه‌گیری توانایی ویژگی‌ها در حفظ ساختار هندسی گراف انجام می‌شود [۱۰].

روش‌های انتخاب ویژگی کلاسیک نیمه نظارتی مبتنی بر گراف نظیر امتیاز لاپلاسین نیمه نظارتی [۱۷، ۱۸]، اهمیت ویژگی‌ها را به‌صورت جداگانه ارزیابی کرده و اطلاعات مفید همبستگی میان ویژگی‌ها را نادیده می‌گیرند. برای مواجهه با این مشکل، روش‌های انتخاب ویژگی تُنک ارائه‌شده‌اند تا در فرایند انتخاب ویژگی همبستگی بین ویژگی‌ها را در نظر بگیرند و نمایش تُنکی از داده‌ها ارائه دهند [۱۲،۱۶،۲۸].

٤- انتخاب ویژگی نیمه نظارتی تُنک مبتنی بر گراف

تنظیم لاپلاسین متداول‌ترین روش برای در نظر گرفتن توزیع و ساختار هندسی داده‌های برچسب‌دار و بدون برچسب و نمایش منیفلد داده‌ها است و اکثر روش‌های انتخاب ویژگی نیمه نظارتی تُنک، از این تکنیک استفاده می‌نمایند [۱۶]. در این روش‌ها، با استفاده از مدل‌های تُنک، ماتریس انتقال بهینه تُنک برای انتخاب ویژگی محاسبه می‌شود. یکی از مدل‌های تُنک مشهور برای انتخاب ویژگی، نُرم-l_1 [۲۹] است و نُرم-l_p $(0 < p < 1)$ به‌عنوان توسعه‌ای از آن برای به دست آوردن نمایش تُنک‌تری از داده‌ها استفاده می‌شود [۳۰،۳۱]. این مدل‌های تُنک، اطلاعات مفید همبستگی بین

ویژگی‌ها را نادیده می‌گیرند. در روش‌های انتخاب ویژگی نیمه نظارتی تُنک، معمولاً از مدل‌های تُنک مبتنی بر نُرم-$l_{2,1}$ [۱۶,۲۸]، نُرم-$l_{2,p}$ $(0 < p < 1)$ [۱۲,۳۲] ونُرم-$l_{2,p}$ $(0 < p \leq 1)$ [۱۴,۳۳] برای محاسبه ماتریس انتقال بهینه تُنک استفاده می‌شود که این مدل‌ها همبستگی بین ویژگی‌ها را برای انتخاب ویژگی در نظر می‌گیرند.

در این روش‌ها، با اضافه نمودن یک عبارت تنظیم مبتنی بر مدل‌های تُنک به تابع هدف محاسبه ماتریس انتقال، ماتریس بهینه تُنک برای انتخاب ویژگی محاسبه می‌گردد. درواقع، تعداد کمی از سطرهای این ماتریس تُنک غیر صفر است که برای وزن دهی به ویژگی‌ها استفاده می‌شود [۱۲,۱۶,۲۸]. با استفاده از ماتریس انتقال تُنک بهینه محاسبه‌شده، می‌توان ویژگی‌ها را رتبه‌بندی نمود و همانند روش‌های انتخاب ویژگی فیلتر، ویژگی‌ها با رتبه بالاتر را انتخاب کرد. علاوه بر انتخاب ویژگی‌ها به‌صورت فیلتر، برخی از روش‌های انتقال ویژگی نیمه نظارتی تُنک مبتنی بر گراف می‌توانند همانند روش‌های انتخاب ویژگی تعبیه‌شده، فرایند انتخاب ویژگی و یادگیری مدل را به‌صورت هم‌زمان انجام دهند.

در تابع هدف روش‌های انتخاب ویژگی نیمه نظارتی تُنک، یک عبارت تنظیم مبتنی بر مدل‌های تُنک وجود دارد که باعث می‌شود حل آن سخت باشد و باید به شیوه‌ای مؤثر حل گردد و الگوریتمی برای حل آن ارائه گردد.

۴-۱ مفاهیم پایه

قبل از بررسی هریک از روش‌های انتخاب ویژگی نیمه نظارتی تُنک مبتنی گراف، نمادهایی که در این بخش برای تعریف رابطه‌ها به‌کاررفته‌اند بیان می‌شوند. مجموعه‌ای از n نمونه آموزشی $X=[x_1, \dots, x_l, x_{l+1}, \dots, x_n]^T \in R^{d\times n}$ شامل l نمونه برچسب‌دار و u نمونه بدون برچسب را در نظر بگیرید که هر نمونه با d ویژگی مشخص می‌شود. داده‌های برچسب‌دار به‌صورت $X_l = [x_1, \dots, x_l]$ با برچسب‌های $Y_l = [y_1, \dots, y_l]$ و داده‌های بدون برچسب به‌صورت $X_u = [x_{l+1}, \dots, x_{l+u}]$ مشخص می‌شوند که $x_i \in R^d \ (1 \leq i \leq n)$ امین نمونه آموزشی و $y_i \in \{1,2, \dots, c\}, \ (1 \leq i \leq l)$ برچسب iامین نمونه آموزشی و c تعداد کلاس‌ها را مشخص می‌کنند.

فرض کنید $W \in R^{d\times c}$ ماتریس انتقالی باشد که برای انتخاب ویژگی تُنک استفاده می‌شود. نُرم -$l_{2,1}$ این ماتریس به‌صورت زیر تعریف می‌شود:

$$\|W\|_{2,1} = \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{c} W_{i,j}^2} = \sum_{i=1}^{d} \|w^i\|_2 \qquad (٤)$$

نُرم-$l_{2,p}$ به‌صورت زیر تعریف می‌شود:

$$\|W\|_{2,p} = \left(\sum_{i=1}^{d} \|w^i\|_2^p\right)^{1/p} \quad p \in (0,1] \qquad (٥)$$

۴-۲ انتخاب ویژگی نیمه نظارتی تُنک مبتنی بر ماتریس پراکندگی بر اساس لاپلاسین گراف

یک روش انتخاب ویژگی نیمه نظارتی تُنک می‌تواند به‌صورت حل مسئله زیر باشد [۳۳]:

$$\min_{W^TW=I} Tr(W^TMW) + \lambda\|W\|_{2,p}^p, \quad p \in (0,1] \qquad (٦)$$

که عبارت تنظیم $\|W\|_{2,p}$ باعث می‌شود که سطرهای ماتریس W تُنک شود و برای انتخاب ویژگی مناسب باشد. همچنین این عبارت باعث می‌شود که همبستگی بین ویژگی‌ها در هنگام انتخاب ویژگی تُنک در نظر گرفته شود. در رابطه فوق، پارامتر $\lambda > 0$ اثر تنظیم را کنترل می‌کند و $M \in R^{d\times d}$ یک ماتریس پراکندگی نیمه نظارتی است که توزیع داده‌ها و اطلاعات برچسب را بیان می‌کند. ماتریس M می‌تواند به‌صورت زیر تعریف شود:

$$M = A + \mu D \qquad (٧)$$

پارامتر $\mu \ (0 \leq \mu \leq 1)$، وزن ماتریس D را کنترل می‌کند. ماتریس $A \in R^{d\times d}$ یک ماتریس پراکندگی نظارتی است که اطلاعات برچسب داده‌های آموزشی برچسب‌دار را بیان کرده و می‌تواند با استفاده از معیارهای مختلف نظیر معیار فیشر یا ماتریس نظارتی لاپلاسین گراف محاسبه شود. ماتریس $D \in R^{d\times d}$ یک ماتریس پراکندگی غیر نظارتی است که اطلاعات ساختار هندسی همه داده‌های آموزشی را بیان می‌کند. این ماتریس می‌تواند با استفاده از ماتریس لاپلاسین گراف یا ماتریس اسپلاین به دست آید. هنگامی‌که W بهینه به‌دست آید، ویژگی‌ها با توجه به مقدار $\|wi\|_2 \ (i = 1,..,d)$ به ترتیب نزولی مرتب‌شده و ویژگی‌ها با رتبه بالاتر انتخاب می‌شوند.

شیخ‌پور و همکاران [۳۳]، ماتریس پراکندگی غیر نظارتی D، را با استفاده از همه داده‌های آموزشی به‌صورت زیر محاسبه نمودند.

$$D = XLX^T \qquad (۸)$$

که $L \in \mathrm{R}^{n\times n}$ ماتریس لاپلاسین گراف است که با استفاده از همه داده‌های آموزشی به دست می‌آید و شباهت هر زوج از داده‌ها را بر اساس مقادیر بردارهای ویژگی بیان می‌کند. برای محاسبه ماتریس لاپلاسین گراف، برچسب داده‌ها نادیده گرفته‌شده و یک گراف همسایگی G با تمام داده‌های آموزشی ساخته می‌شود. در این گراف، هر گره متناظر با یک نمونه آموزشی است و گره‌های نزدیک به یکدیگر متصل می‌شوند. پس از ساخت گراف همسایگی، ماتریس لاپلاسین گراف محاسبه می‌شود.

ماتریس پراکندگی نظارتی A، توسط شیخ‌پور و همکاران [۳۳] به‌صورت زیر محاسبه شد.

$$A = X_l L^{sup} X_l^T \qquad (۹)$$

که $L^{sup} \in \mathrm{R}^{l\times l}$ ماتریس نظارتی لاپلاسین گراف است که با استفاده از اطلاعات برچسب داده‌های آموزشی برچسب‌دار به دست می‌آید. برای محاسبه لاپلاسین گراف L^{sup} با استفاده از اطلاعات برچسب داده‌های برچسب‌دار، یک گراف همسایگی G^{sup} با نمونه‌های برچسب‌دار (یک گره به ازای هر داده برچسب‌دار) ساخته می‌شود و نمونه‌های نزدیک با توجه به مقادیر خروجی به یکدیگر متصل می‌شوند. پس از ساخت گراف، ماتریس نظارتی لاپلاسین گراف به‌صورت $L^{sup} = D^{sup} - S^{sup}$ محاسبه می‌شود که $D^{sup} = diag(S^{sup}1)$.

۳-٤ انتخاب ویژگی نیمه نظارتی تُنک مبتنی بر تابع زیان و لاپلاسین گراف

متداول‌ترین چارچوب برای انتخاب ویژگی تُنک، کمینه نمودن تابع هدف زیر است:

$$\min_W loss(W) + \lambda R(W) \qquad (۱۰)$$

در رابطه فوق، $loss(.)$ تابع زیان، $R(W)$ عبارت تنظیم و λ پارامتر تنظیم است.

برای انتخاب ویژگی تُنک که همبستگی میان ویژگی‌ها را در نظر بگیرد، به‌جای عبارت تنظیم $\lambda R(W)$ در رابطه (۱۰)، از تنظیم نُرم-$l_{2,1}$ یا نُرم-$l_{2,p}$ $(0 < p < 1)$ یا نُرم-$l_{2,p}$ $(0 < p \leq 1)$ استفاده می‌شود.

فرض کنید $Y = [y_1, \dots, y_l, y_{l+1}, \dots, y_n]^T \in \{0,1\}^{n\times c}$ ماتریس برچسب داده‌های آموزشی باشد که Y_{ij}، j امین داده y_i را مشخص می‌کند، بنابراین اگر x_i در j امین کلاس باشد $Y_{ij} = 1$ و در غیراین‌صورت $Y_{ij} = 0$. اگر x_i بدون برچسب باشد، y_i یک بردار با مقادیر صفر خواهد بود.

تنظیم منیفلد، یک روش معروف مبتنی بر لاپلاسین گراف است که الگوریتم‌های زیادی را تبدیل به الگوریتم‌های نیمه نظارتی نموده است [۱۶]. با به کار بردن تنظیم منیفلد در تابع زیان رابطه (۱۰)، رابطه زیر به دست می‌آید.

$$\arg\min_{W,b} Tr(W^T XLX^T W) + \mu loss(W) + \lambda \|W\|_{2,p}^p \qquad (۱۱)$$

که L لاپلاسین گراف است و با استفاده از گراف k-نزدیک‌ترین همسایه به‌دست می‌آید. μ و λ پارامترهای تنظیم هستند. در تابع فوق، عبارت تنظیم $\|W\|_{2,p}$ تضمین می‌کند که این مدل می‌تواند انتخاب ویژگی تُنک را انجام دهد.

ما و همکاران [۱۶] و شای و همکاران [۱۲] از تابع زیان مبتنی بر نُرم- مرتبه دوم $\|X_l^T W + 1_l b^T - Y_l\|_F^2$ در رابطه (۱۱) استفاده نمودند. $b \in R^c$ یک عبارت بایاس در تابع زیان و $1_l \in R^l$ یک بردار ستونی است که همه l عنصر آن یک است. ما و همکاران [۱۶] برای انجام انتخاب ویژگی نیمه نظارتی تُنک از تنظیم نُرم-$l_{2,1}$ و شای و همکاران از تنظیم نُرم-$l_{2,1/2}$ استفاده نمودند که تابع هدف را به ترتیب به یک مسئله محدب و یک مسئله غیر محدب تبدیل می‌کند. ازآنجاکه تابع زیان مبتنی بر نُرم- مرتبه دوم نسبت به نقاط پرت حساس است، شیخ‌پور و همکاران [۱۴] با استفاده از تابع زیان مبتنی بر نُرم ترکیبی $\|X_l^T W + 1_l b^T - l_{2,p}$ $(0 < p \leq 1)$ $Y_l\|_{2,p}^p$ که در مواجهه با نقاط پرت پایدار است، تابع هدفی نیمه نظارتی مبتنی بر کمینه‌سازی نُرم-$l_{2,p}$ بر روی تابع زیان و عبارت تنظیم ارائه دادند.

برای استفاده از برچسب همه داده‌های آموزشی در بهینه‌سازی ماتریس W در رابطه (۱۱)، ماتریس برچسب‌های پیش‌بینی‌شده $F = [f_1, f_2, \dots, f_n]^T \in \mathrm{R}^{n\times c}$ در نظر گرفته می‌شود به‌طوری‌که $f_i \in R^c \ (1 \le i \le n)$ برچسب پیش‌بینی‌شده برای نمونه x_i را مشخص می‌کند. ماتریس F می‌تواند با کمینه نمودن تابع هدف زیر به‌دست آید تا به برچسب‌های واقعی نزدیک باشد و بر روی گراف هموار باشد.

$$\arg\min_{F} \sum_{l=1}^{c} \left[\frac{1}{2}\sum_{i,j=1}^{n}(F_{il}-F_{jl})^2 S_{ij} + \sum_{i=1}^{n} U_{ii}(F_{il}-Y_{il})^2\right] \quad (۱۲)$$

که F_{il}، l امین عنصر f_i و $U \in R^{n\times n}$ یک ماتریس قطری است که ماتریس قانون تصمیم‌گیری نامیده می‌شود. این ماتریس باعث سازگاری برچسب‌های پیش‌بینی‌شده توسط F با برچسب‌های واقعی Y می‌شود.

رابطه (۱۲) می‌تواند به‌صورت رابطه زیر نوشته شود:

$$\underset{F}{arg\ min}\, Tr(F^T LF) + Tr\big((F-Y)^T U(F-Y)\big) \quad (۱۳)$$

بنابراین، تابع زیان در رابطه (۱۰) می‌تواند به‌صورت زیر باشد:

$$\arg\min_{F,W,b} Tr(F^T LF) + Tr\big((F-Y)^T U(F-Y)\big) + \mu loss(W)_F^2 \quad (۱۴)$$

به‌جای X_l و Y_l در رابطه (۱۱) از X و F استفاده می‌شود. با توجه به رابطه (۱۴)، ماتریس پیش‌بینی برچسب و مدل دسته‌بندی می‌توانند به‌صورت هم‌زمان یاد گرفته شوند.

با تجمیع انتخاب ویژگی تُنک مبتنی بر نُرم ماتریس-$l_{2,p}$ و یادگیری نیمه نظارتی مبتنی بر لاپلاسین گراف، تابع هدف زیر به‌دست می‌آید:

$$\arg\min_{F,W,b} Tr(F^T LF) + Tr\big((F-Y)^T U(F-Y)\big) + \mu loss(W) + \lambda\|W\|_{2,p}^p \quad (۱۵)$$

تابع هدف فوق شامل نُرم-$l_{2,p}$ است و به نظر می‌رسد که حل آن سخت باشد. بنابراین باید الگوریتمی مؤثر برای حل آن به کار گرفته شود.

٤-٤ انتخاب ویژگی نیمه نظارتی تُنک براساس ترکیب یادگیری ساختار تطبیقی و یادگیری مبتنی بر گراف

یک روش انتخاب ویژگی نیمه نظارتی تُنک با ترکیب یادگیری ساختار تطبیقی و یادگیری مبتنی بر گراف در یک چارچوب [۳٤] به‌صورت حل رابطه (۱٦) است. در این مدل با استفاده از عبارت تنظیم، یادگیری گراف تطبیقی توسعه می‌یابد و یک ماتریس شباهت از پیش تعریف‌شده برای محدود کردن ماتریس شباهت یادگیری تطبیقی برای بهترین استفاده از اطلاعات برچسب به کار می‌رود.

$$\arg\min_{F,S,W} \|X^T W - F\|_F^2 + \sum_{ij}^{n}\left\|W^T(X_i - X_j)\right\|_2^2 S_{ij} + \alpha S_{ij}^2 + Tr(F^T LF) + Tr\big((F-Y)^T U(F-Y)\big) + \lambda\|W\|_{2,1} \quad (۱٦)$$

$$s.t. S_i^T 1_n = 1, 0 \le S_{ij} \le 1$$

که α و λ پارامترهای توازن هستند. سپس، ماتریس شباهت با استفاده از یادگیری مدل رابطه (۱۷)، محدود می‌شود.

$$\arg\min_{S} \sum_{ij}^{n}\left\|W^T(X_i - X_j)\right\|_2^2 S_{ij} + \alpha\|S-A\|_F^2 \quad (۱۷)$$

$$s.t. S_i^T 1_n = 1, 0 \le S_{ij} \le 1$$

که A یک ماتریس از پیش تعریف‌شده است و با استفاده از دانش پیشین ایجاد می‌شود. به‌طور مثال ماتریس A می‌تواند توسط رابطه (۲) به دست آید. با جایگزینی یادگیری گراف تطبیقی در رابطه (۱٦) با مدل رابطه (۱۷)، رابطه (۱۸) حاصل می‌شود.

$$\arg\min_{F,S,W} \|X^T W - F\|_F^2 + \sum_{ij}^{n}\left\|W^T(X_i - X_j)\right\|_2^2 S_{ij} + \alpha\|S-A\|_F^2 + Tr(F^T LF) + Tr\big((F-Y)^T U(F-Y)\big) + \lambda\|W\|_{2,1} \quad (۱۸)$$

$$s.t. S_i^T 1_n = 1, 0 \le S_{ij} \le 1$$

برای بهترین استفاده از اطلاعات برچسب، تابع هدف نهایی انتخاب ویژگی نیمه نظارتی تُنک براساس ترکیب یادگیری ساختار تطبیقی و یادگیری مبتنی بر گراف به‌صورت رابطه (۱۹) تعریف می‌شود.

$$\arg\min_{F,S,W,Z} \|X^T W - F\|_F^2 + \sum_{ij}^{n}\left\|W^T(X_i - X_j)\right\|_2^2 S_{ij} + \alpha\|S-A\|_F^2 + Tr(F^T LF) + Tr\big((F-Y)^T U(F-Y)\big) + \|W^T X - W^T XZ\|_F^2 + \beta\|Z\|_{2,1} + \lambda\|W\|_{2,1} \quad (۱۹)$$

$$s.t. S_i^T\ 1_n = 1, 0 \le S_{ij} \le 1$$

٤-٥ انتخاب ویژگی نیمه نظارتی تُنک مبتنی بر یادگیری ساختار داده‌ها و لاپلاسین گراف

یکی از روش‌های انتخاب ویژگی نیمه نظارتی تُنک که از ترکیب یادگیری ساختار داده‌ها مبتنی بر توسعه یادگیری داده‌های نرم و سخت و انتخاب ویژگی تُنک مبتنی بر نُرم $l_{1,2}$- استفاده می‌کند، به‌صورت حل مسئله (۲۰) است که ساختار داده‌های نرم به مقادیر محاسبه‌شده وزن‌ها بین زوج داده‌ها و ساختار داده‌های سخت به برچسب‌های تخمینی به‌دست‌آمده از خوشه‌بندی یا یادگیری نیمه نظارتی اشاره دارد [۳۵]:

$$\min_{Z,E,Y,W} \left\{ \|Z\|_1 + \lambda_E \|E\|_1 + \lambda_Z \|\Theta \odot Z\|_1 - \frac{Tr(WM^bW^T)}{Tr(WM^tW^T)} + \gamma_W \|W\|_{1,2} \right\} \quad (۲۰)$$
$$s.t. X = XZ + E,\ diag\ (Z) = 0,\ Y = [Y_l, Y_u]$$

در این رابطه، λ_E و λ_Z پارامترهای تنظیم هستند، M^b و M^t به ترتیب، ماتریس‌های پراکندگی بین کلاسی و کل کلاس در آنالیز تفکیکی خطی هستند. $\odot$ ضرب هادامارد، $\Theta_{ij} = \left(\frac{\alpha}{2}\right) \|Wx_i - Wx_j\|^2 + \frac{1-\alpha}{2\|y_i - y_j\|^2}$ و $\alpha \in [0,1]$ پارامتر توازن است. هر نمونه می‌تواند به‌صورت خطی توسط یک ترکیب خطی از نمونه‌های دیگر توصیف شود به‌طوری‌که $X = XZ$ و diag (Z)=0، که Z ماتریس ضرایب بازسازی است. این مدل می‌تواند با استفاده از داده‌های نویزی به‌صورت $X = XZ + E$ تعریف شود که E ماتریس خطا است.

سومین عبارت در تابع هدف رابطه (۲۰)، شامل عبارت زیر است:

$$\min_{Y_u} \|\Theta \odot Z\|_1 \leftrightarrow \arg\min_{Y_u} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\|y_i - y_j\|_2^2 \quad (۲۱)$$

که این مسئله به مسئله یادگیری نیمه نظارتی مطابق رابطه (۲۲) تبدیل می‌شود.

$$\arg\min_{Y_u} trace\ (YLY^T), \quad s.t.: Y = [Y_l, Y_u] \quad (۲۲)$$

که L ماتریس لاپلاسین گراف است.

٤-٦ انتخاب ویژگی نیمه نظارتی تُنک مبتنی بر یادگیری نیمه نظارتی و محدودیت غیر وابسته و لاپلاسین گراف

یکی از روش‌های انتخاب ویژگی نیمه نظارتی مبتنی بر یادگیری نیمه نظارتی و محدودیت غیر وابسته به‌صورت زیر تعریف می‌شود [۳۶].

$$\min_{Z,F_u,\alpha} \|[X_l, X_u]^1 Z - \alpha[[F_l, F_u]]\|_F^2 + \beta Tr(F^T LF) + \gamma \|Z\|_{2,1} \quad (۲۳)$$
$$s.t. Z^T S_t Z = I$$

که $S_t = XX^T$ ماتریس پراکندگی کلی، $Z = \alpha W$، α، β و γ پارامترها و L ماتریس لاپلاسین گراف است.

محدودیت غیر وابسته $Z^T S_t Z = I$ در رابطه (۲۳) می‌تواند با محدودیت غیر وابسته عمومی $Z^T {S_t}^{(p)} Z = I$ جایگزین شود، به‌طوری‌که ${S_t}^{(p)}$ به‌عنوان $(XX^T + \gamma P)$ تعریف می‌شود که P یک ماتریس قطری $d \times d$ است و عنصر p_{ii} به‌صورت رابطه (۲٤) تعریف می‌شود.

$$p_{ii} = \frac{1}{2\sqrt{\|Z_i\|_2^2 + \varepsilon}} (\varepsilon \to 0, i = 1, 2, \ldots, d) \quad (۲٤)$$

بنابراین، تابع هدف نهایی برای انتخاب ویژگی نیمه نظارتی مبتنی بر محدودیت غیر وابسته عمومی به‌صورت زیر تعریف می‌شود.

$$\min_{Z,F_u,\alpha} \|[X_l, X_u]^1 Z - \alpha[[F_l, F_u]]\|_F^2 + \beta Tr(F^T LF) + \gamma \|Z\|_{2,1} \quad (۲۵)$$
$$s.t. Z^T {S_t}^{(p)} Z = I$$

این تابع هدف با استفاده از لاپلاسین گراف و عبارت تنظیم مبتنی بر نُرم $l_{2,1}$-، ویژگی‌های متمایز و غیر وابسته را انتخاب می‌کند. محدودیت غیر وابسته عمومی باعث می‌شود که یک راه‌حل فرم بسته به چارچوب تابع هدف اضافه شود و ساختار هندسی داده‌ها در طول فرایند بهینه‌سازی حفظ شود. سطرهای ماتریس انتقال به‌دست‌آمده با استفاده از محدودیت غیر وابسته عمومی تُنک هستند ولی ستون‌های آن دارای رتبه کامل هستند که این محدودیت تضمین می‌کند که حداقل c سطر غیرصفر وجود داشته باشد و از وقوع تُنک بودن بیش‌ازحد سطرهای ماتریس جلوگیری می‌کند. درواقع برای اضافه نمودن راه‌حل فرم بسته به چارچوب تابع هدف، باید ماتریس انتقال را محدود نمود تا رتبه کامل باشد. برای به

دست آوردن محدودیت رتبه کامل، معمولاً دو روش وجود دارد: محدودیت تعامد $W^TW=I$ و محدودیت غیر وابسته $W^TS_tW=I$ که $S_t=XX^T$ یک ماتریس پراکندگی کلی است. محدودیت غیر وابسته باعث می‌شود داده‌های غیر وابسته بیشتری استخراج شوند.

۵- بحث

در بخش ٤، روش‌های انتخاب ویژگی نیمه نظارتی تُنک مبتنی بر گراف بررسی شدند. در این بخش در جدول ۱، خصوصیات، مزایا و معایب روش‌های انتخاب ویژگی نیمه نظارتی تُنک مبتنی بر گراف بیان می‌شود.

در تمام روش‌هایی که در بخش ٤ شرح داده شد، ساختار هندسی داده‌ها با استفاده از لاپلاسین گراف حفظ می‌شود. در این روش‌ها، توزیع و ساختار هندسی داده‌ها در ارزیابی ویژگی‌ها در نظر گرفته می‌شود. همچنین در تمام این روش‌ها، یک عبارت تنظیم مبتنی بر نُرم-$l_{2,1}$[۱٦,۳٤,۳٦]، نُرم $l_{1,2}$ [۳۵]، نُرم- $l_{2,1/2}$[۱۲] یا نُرم-$l_{2,p}$ $(0 < p \leq 1)$ [۱٤,۳۳] استفاده می‌شود که این عبارت باعث می‌شود سطرهای ماتریس انتقال محاسبه‌شده تُنک باشند و این ماتریس بتواند برای انتقال ویژگی استفاده شود. همچنین این عبارت باعث می‌شود همبستگی بین ویژگی‌ها در فرایند انتخاب ویژگی در نظر گرفته شود. در تمام این روش‌ها، عبارت تنظیم موجود در تابع هدف باعث می‌شود که حل تابع هدف سخت باشد و نیاز به الگوریتمی با رویکرد تکراری برای حل تابع هدف باشد.

روش انتخاب ویژگی نیمه نظارتی تُنک مبتنی بر ماتریس پراکندگی براساس لاپلاسین گراف [۳۳]، یک روش انتخاب ویژگی فیلتر است که بدون استفاده از تابع رگرسیون و با استفاده از ماتریس پراکندگی نیمه نظارتی مبتنی بر لاپلاسین گراف، عمل می‌کند. این روش، از محدودیت تعامد استفاده می‌نماید و وابسته به روش دسته‌بندی نیست.

در این روش تنها در مرحله ساخت گراف همسایگی با استفاده از داده‌های برچسب‌دار، اطلاعات برچسب داده‌های برچسب‌دار در نظر گرفته می‌شود و در مرحله کمینه‌سازی تابع هدف، اطلاعات برچسب داده‌ها در نظر گرفته نمی‌شود.

روش‌های انتخاب ویژگی نیمه نظارتی تُنک مبتنی بر تابع زیان و لاپلاسین گراف [۱۲,۱٤,۱٦]، انتخاب ویژگی نیمه نظارتی تُنک براساس ترکیب یادگیری ساختار تطبیقی و یادگیری مبتنی بر گراف [۳٤] و انتخاب ویژگی نیمه نظارتی تُنک مبتنی بر یادگیری نیمه نظارتی و محدودیت غیر وابسته و لاپلاسین گراف [۳٦]، از ترکیب یک روش یادگیری، تنظیم منیفلد و انتخاب ویژگی تُنک استفاده می‌نمایند. در این روش‌ها از یک تابع زیان استفاده می‌شود و می‌توانند به‌عنوان انتخاب ویژگی تعبیه‌شده، یادگیری مدل و انتخاب ویژگی را به‌صورت هم‌زمان انجام دهند. همچنین، این روش‌ها می‌توانند به‌صورت انتخاب ویژگی فیلتر عمل نمایند و با استفاده از ماتریس انتقال تُنک محاسبه‌شده، ویژگی‌ها را رتبه‌بندی نمایند. درروش‌هایی که از تابع زیان استفاده می‌کنند، از تابع برچسب در تابع زیان مسئله کمینه‌سازی استفاده می‌شود که باعث می‌شود انتخاب ویژگی و یادگیری به‌صورت هم‌زمان انجام شوند و موجب افزایش کارایی این روش‌ها نسبت به روش‌های تُنک فیلتر می‌شود.

جدول (۱): خصوصیات، مزایا و معایب روش‌های انتخاب ویژگی نیمه نظارتی تُنک مبتنی بر گراف

روش انتخاب ویژگی	خصوصیات	مزایا	معایب
انتخاب ویژگی نیمه نظارتی تُنک مبتنی بر ماتریس پراکندگی براساس لاپلاسین گراف [۳۳]	• رتبه‌بندی ویژگی‌ها و انتخاب آن‌ها همانند روش‌های فیلتر • استفاده از عبارت تنظیم مبتنی بر نُرم-$l_{2,p}$ $(0 < p \leq 1)$ • استفاده از ماتریس پراکندگی نیمه نظارتی • استفاده از محدودیت تعامد	• دستیابی به نمایش تُنک‌تری از داده‌ها به علت عبارت تنظیم مبتنی بر نُرم-$l_{2,p}$ $(0 < p < 1)$	• عدم استفاده از تابع زیان و در نظر نگرفتن اطلاعات برچسب در تابع زیان در فرایند کمینه‌سازی تابع هدف • کاهش کارایی نسبت به روش‌های انتخاب ویژگی تُنک تعبیه‌شده

انتخاب ویژگی نیمه نظارتی تُنک مبتنی بر تابع زیان و لاپلاسین گراف [۱۶]	• انجام هم‌زمان فرایند یادگیری مدل و انتخاب ویژگی (انتخاب ویژگی تعبیه‌شده) • قابلیت استفاده به‌صورت فیلتر و رتبه‌بندی ویژگی‌ها • استفاده از عبارت تنظیم مبتنی بر نُرم-$l_{2,1}$ • استفاده از تابع زیان مبتنی بر نُرم-مرتبه دوم • استفاده از تنظیم منیفلد	• ترکیب یادگیری نیمه نظارتی و انتخاب ویژگی تُنک	• حساس به نقاط پرت
انتخاب ویژگی نیمه نظارتی تُنک مبتنی بر تابع زیان و لاپلاسین گراف [۱۲]	• انجام هم‌زمان فرایند یادگیری مدل و انتخاب ویژگی (انتخاب ویژگی تعبیه‌شده) • قابلیت استفاده به‌صورت فیلتر و رتبه‌بندی ویژگی‌ها • استفاده از عبارت تنظیم مبتنی بر نُرم-$l_{2,1/2}$ • استفاده از تابع زیان مبتنی بر نُرم-مرتبه دوم • استفاده از تنظیم منیفلد	• ترکیب یادگیری نیمه نظارتی و انتخاب ویژگی تُنک • دستیابی به نمایش تُنک‌تری از داده‌ها به علت عبارت تنظیم مبتنی بر نُرم-$l_{2,1/2}$	• حساس به نقاط پرت
انتخاب ویژگی نیمه نظارتی تُنک مبتنی بر تابع زیان و لاپلاسین گراف [۱۴]	• انجام هم‌زمان فرایند یادگیری مدل و انتخاب ویژگی (انتخاب ویژگی تعبیه‌شده) • قابلیت استفاده به‌صورت فیلتر و رتبه‌بندی ویژگی‌ها • استفاده از عبارت تنظیم مبتنی بر نُرم-$l_{2,p}$ $(0 < p \leq 1)$ • استفاده از تابع زیان مبتنی بر نُرم-$l_{2,p}$ $(0 < p \leq 1)$ • استفاده از تنظیم منیفلد	• ترکیب یادگیری نیمه نظارتی و انتخاب ویژگی تُنک • دستیابی به نمایش تُنک‌تری از داده‌ها به علت عبارت تنظیم بر نُرم-$l_{2,p}$ $(0 < p \leq 1)$ • پایدار در مواجهه با نقاط پرت	• نیاز به راه‌حل پیچیده برای حل تابع هدف به دلیل استفاده هم‌زمان تابع زیان و عبارت تنظیم مبتنی بر بر نُرم-$l_{2,p}$ $(0 < p \leq 1)$
انتخاب ویژگی نیمه نظارتی تُنک براساس ترکیب یادگیری ساختار تطبیقی و یادگیری مبتنی بر گراف [۳۴]	• انجام هم‌زمان فرایند یادگیری مدل و انتخاب ویژگی (انتخاب ویژگی تعبیه‌شده) • قابلیت استفاده به‌صورت فیلتر و رتبه‌بندی ویژگی‌ها • استفاده از عبارت تنظیم مبتنی بر نُرم-$l_{2,1}$ • استفاده از تابع زیان مبتنی بر نُرم-مرتبه دوم • استفاده از ساختار منیفلد محلی و عمومی داده‌ها • استفاده از یادگیری ماتریس شباهت به‌صورت تطبیقی برای بهبود کیفیت ماتریس شباهت	• ترکیب یادگیری نیمه نظارتی و انتخاب ویژگی تُنک • استفاده هم‌زمان از ساختار منیفلد محلی و عمومی داده‌ها • یادگیری ساختار محلی منیفلد داده‌ها به‌صورت دقیق • بهبود بیشتر کیفیت ماتریس شباهت با استفاده از اطلاعات برچسب	• حساس به نقاط پرت
انتخاب ویژگی نیمه نظارتی تُنک مبتنی بر یادگیری نیمه نظارتی و محدودیت غیر وابسته و لاپلاسین گراف [۳۶]	• انجام هم‌زمان فرایند یادگیری مدل و انتخاب ویژگی (انتخاب ویژگی تعبیه‌شده) • قابلیت استفاده به‌صورت فیلتر و رتبه‌بندی ویژگی‌ها • رتبه‌بندی ویژگی‌ها و انتخاب آن‌ها همانند روش‌های فیلتر • استفاده از عبارت تنظیم مبتنی بر نُرم-$l_{2,1}$ • استفاده از تابع زیان مبتنی بر نُرم-مرتبه دوم • استفاده از تنظیم منیفلد • استفاده از محدودیت غیر وابسته عمومی	• ترکیب یادگیری نیمه نظارتی و انتخاب ویژگی تُنک • حفظ ساختار هندسی داده‌ها در طول فرایند بهینه‌سازی به دلیل استفاده از محدودیت غیر وابسته عمومی • استخراج بیشتر داده‌های غیر وابسته به دلیل استفاده از محدودیت غیر وابسته عمومی	• حساس به نقاط پرت

انتخاب ویژگی نیمه نظارتی تُنک مبتنی بر یادگیری ساختار داده‌ها و لاپلاسین گراف [۳۵]	• استفاده از عبارت تنظیم مبتنی بر نُرم-$l_{1,2}$ • ترکیب یادگیری ساختار و توزیع داده‌ها مبتنی بر توسعه یادگیری داده‌های نرم و سخت و انتخاب ویژگی تُنک	• بهبود عملکرد انتخاب ویژگی به دلیل یادگیری ساختار داده‌ها مبتنی بر توسعه یادگیری داده‌های نرم و سخت	• عدم استفاده از تابع زیان و در نظر نگرفتن اطلاعات برچسب در تابع زیان در فرایند کمینه‌سازی تابع هدف • کاهش کارایی نسبت به روش‌های انتخاب ویژگی تُنک تعبیه‌شده

روش‌های انتخاب ویژگی تُنک ارائه‌شده در پژوهش [۱۲,۱۶,۳۴,۳۶]، از تابع زیان مبتنی بر نُرم- مرتبه دوم استفاده می‌نمایند که این تابع به نقاط پرت حساس است ولی روش ارائه‌شده در پژوهش [۱۴] از تابع زیان مبتنی بر نُرم-$l_{2,p}$ استفاده می‌کند که این تابع در مواجهه با نقاط پرت پایدار است.

روش‌های ارائه‌شده در پژوهش [۱۲,۱۴,۳۳] به دلیل استفاده از عبارت تنظیم مبتنی بر بر نُرم- $l_{21/2}$و نُرم-$l_{2,p}$ $(0 < p \leq 1)$ می‌توانند نسبت به روش‌های دیگر به نمایش تُنک‌تری از داده‌ها دست یابند.

درروش ارائه‌شده در پژوهش [۳۶]، محدودیت غیر وابسته عمومی باعث حفظ ساختار هندسی داده‌ها در طول فرایند بهینه‌سازی می‌شود. این روش به دلیل استفاده از محدودیت غیر وابسته عمومی، داده‌های غیر وابسته بیشتری را نسبت به روش ارائه‌شده در پژوهش [۳۳] که مبتنی بر محدودیت تعامد است، استخراج می‌کند.

در روش‌های یادگیری مبتنی بر لاپلاسین گراف، کیفیت ماتریس شباهت نمونه‌ها براساس گراف بر عملکرد مدل یادگیری تأثیر می‌گذارد. یادگیری گراف تطبیقی، با استفاده از یادگیری ماتریس شباهت به‌صورت تطبیقی کیفیت ماتریس شباهت را بهبود می‌بخشد. بااین‌حال، اکثر روش‌های مبتنی بر یادگیری گراف تطبیقی اطلاعات برچسب را نادیده می‌گیرند، که ممکن است بر روی کیفیت ماتریس شباهت تأثیر بگذارد. از سوی دیگر، بسیاری از روش‌های انتخاب ویژگی نیمه نظارتی، تنها ساختار محلی داده‌ها را در نظر می‌گیرند و ساختار عمومی آن‌ها را نادیده می‌گیرند که می‌تواند منجر به افزونگی بالا در ویژگی‌های انتخاب‌شده شود. برای مقابله با این مشکلات، درروش ارائه‌شده در پژوهش [۳۴] یادگیری گراف تطبیقی از طریق اطلاعات برچسب داده‌ها توسعه می‌یابد که هدف آن بهبود بیشتر کیفیت ماتریس شباهت با استفاده از اطلاعات برچسب برای محدود کردن یادگیری گراف است. این روش، یادگیری ساختار تطبیقی و یادگیری مبتنی بر گراف تطبیقی را در یک چارچوب ترکیب می‌کند و با استفاده از یک ماتریس شباهت از پیش تعریف‌شده برای محدود کردن ماتریس شباهت یادگیری تطبیقی، از اطلاعات برچسب داده‌های برچسب‌دار بهترین استفاده را می‌کند. درواقع در این روش، تابع هدف روش ارائه‌شده در پژوهش [۱۶] توسعه‌یافته و یک ماتریس شباهت از پیش تعریف‌شده به کار می‌رود. درروش انتخاب ویژگی نیمه نظارتی تُنک براساس ترکیب یادگیری ساختار تطبیقی و یادگیری مبتنی بر گراف، ساختار منیفلد محلی و عمومی هم‌زمان در نظر گرفته می‌شوند. این روش با در نظر گرفتن اطلاعات برچسب در یادگیری گراف تطبیقی، یادگیری ساختار محلی منیفلد داده‌ها را دقیق‌تر انجام می‌دهد و ساختار عمومی داده‌ها را برای تسهیل در فرایند انتخاب ویژگی در نظر می‌گیرد.

روش انتخاب ویژگی نیمه نظارتی تُنک مبتنی بر یادگیری ساختار داده‌ها و لاپلاسین گراف [۳۵]، از ترکیب یادگیری ساختار و توزیع داده‌ها مبتنی بر توسعه یادگیری داده‌های نرم و سخت و انتخاب ویژگی تُنک استفاده می‌کند. ساختار داده‌های نرم به مقادیر محاسبه‌شده وزن‌ها بین زوج داده‌ها و ساختار داده‌های سخت به برچسب‌های تخمینی به‌دست‌آمده از خوشه‌بندی یا یادگیری نیمه نظارتی اشاره دارد.

۶- نتیجه‌گیری

در این مقاله، ضمن بررسی روش‌های یادگیری نیمه نظارتی و مفاهیم کلی انتخاب ویژگی نیمه نظارتی، مروری جامع بر روش‌های انتخاب ویژگی نیمه نظارتی تُنک مبتنی بر گراف انجام

شد و خصوصیات، مزایا و معایب هریک از آن‌ها بیان شد. این روش‌ها با در نظر گرفتن همبستگی بین ویژگی‌ها و با استفاده از ساخت گراف همسایگی و محاسبه لاپلاسین گراف، ویژگی‌های مناسب را انتخاب می‌کنند. در روش‌های انتخاب ویژگی نیمه نظارتی تُنک مبتنی بر گراف، ساختار هندسی داده‌ها با استفاده از لاپلاسین گراف حفظ می‌شود و یک عبارت تنظیم مبتنی بر مدل‌های تّنک به تابع هدف اضافه می‌شود. درواقع، در این روش‌ها یک ماتریس انتقال تُنک بهینه محاسبه می‌شود و به دلیل وجود عبارت تنظیم مبتنی بر مدل‌های تُنک بسیاری از سطرهای این ماتریس صفر هستند. سطرهای این ماتریس برای وزن‌دهی به ویژگی‌ها مورداستفاده قرار می‌گیرند و ویژگی‌های متناسب با سطرهای غیرصفر انتخاب می‌شوند. عبارت تنظیم در تابع هدف روش‌های انتخاب ویژگی نیمه نظارتی تُنک باعث می‌شود که حل تابع هدف سخت باشد، بنابراین نیاز است الگوریتمی با رویکرد تکراری برای حل تابع هدف در این مسائل ارائه شود.

در روش‌های انتخاب ویژگی نیمه نظارتی تُنک مبتنی بر گراف که از تابع زیان استفاده می‌کنند، فرایند یادگیری مدل و انتخاب ویژگی می‌تواند به‌صورت هم‌زمان انجام شود که باعث افزایش کارایی این روش‌ها نسبت به روش‌های تُنک فیلتر می‌شود.

References

[1] Y. Hu, Y. Zhang, D. Gong, Multiobjective Particle Swarm Optimization for Feature Selection With Fuzzy Cost, IEEE TRANSACTIONS ON CYBERNETICS. 51 (2021) 874–888.
[2] G. Dhiman, D. Oliva, A. Kaur, K.K. Singh, S. Vimal, A. Sharma, K. Cengiz, BEPO: A novel binary emperor penguin optimizer for automatic feature selection, Knowledge-Based Systems. 211 (2021).
[3] Z. Feng, Q. Zhou, Q. Gu, X. Tan, G. Cheng, X. Lu, J. Shi, L. Ma, DMT: Dynamic mutual training for semi-supervised learning, Pattern Recognition. 130 (2022).
[4] T. Huynh, A. Nibali, Z. He, Semi-supervised learning for medical image classification using imbalanced training data, Computer Methods and Programs in Biomedicine. 216 (2022).
[5] O. Chapelle, B. Schölkopf, A. Zien, Semi-supervised learning, MIT press Cambridge, 2006.
[6] C. Quintero-Gull, J. Aguilar, LAMDA-HSCC: A semi-supervised learning algorithm based on the multivariate data analysis, Expert Systems with Applications. 202 (2022).
[7] M. Belkin, P. Niyogi, V. Sindhwani, Manifold regularization: A geometric framework for learning from labeled and unlabeled examples, Journal of Machine Learning Research. 7 (2006) 2399–2434.
[8] Razieh Sheikhpour; Mehdi Agha Sarrama; Sajjad Gharaghani; Mohammad Ali Zare Chahookia, A Survey on semi-supervised feature selection methods, Pattern Recognition. 64 (2017) 141–158.
[9] H. Cheng, W. Deng, C. Fu, Y. Wang, Z. Qin, Graph-based semi-supervised feature selection with application to automatic spam image identification, in: Computer Science for Environmental Engineering and EcoInformatics, Springer, 2011: pp. 259–264.
[10] J. Zhao, K. Lu, X. He, Locality sensitive semi-supervised feature selection, Neurocomputing. 71 (2008) 1842–1849.
[11] G. Doquire, M. Verleysen, A graph Laplacian based approach to semi-supervised feature selection for regression problems, Neurocomputing. 121 (2013) 5–13.
[12] C. Shi, Q. Ruan, G. An, Sparse feature selection based on graph Laplacian for web image annotation, Image and Vision Computing. 32 (2014) 189–201.
[13] Z. Zeng, X. Wang, J. Zhang, Q. Wu, Semi-supervised feature selection based on local discriminative information, Neurocomputing. 173 (2016) 102–109.
[14] R. Sheikhpour, M.A. Sarram, S. Gharaghani, M.A.Z. Chahooki, A robust graph-based semi-supervised sparse feature selection method, Information Sciences. 531 (2020) 13–30.
[15] X. Chen, G. Yuan, F. Nie, Z. Ming, Semi-supervised Feature Selection via Sparse Rescaled Linear Square Regression, IEEE Transactions on Knowledge and Data Engineering. 32 (2018) 165–176.
[16] Z. Ma, F. Nie, Y. Yang, J.R.R. Uijlings, N. Sebe, S. Member, A.G. Hauptmann, Discriminating joint feature analysis for multimedia data understanding, IEEE TRANSACTIONS ON MULTIMEDIA. 14 (2012) 1662–1672.
[17] X. Zhu, A.B. Goldberg, Introduction to semi-supervised learning, 2009.
[18] C. Leistner, Semi-supervised ensemble methods for computer vision, PhD Thesis, Graz University of Technology, 2010.

[19] L. Zuo, L. Li, C. Chen, The graph based semi-supervised algorithm with ℓ1-regularizer, Neurocomputing. 149 (2015) 966–974.
[20] Z. Zhou, M. Li, Semi-supervised regression with Co-training, in: International Joint Conference on Arti¯cial Intelligence (IJCAI'05), 2005: pp. 908–913.
[21] F. Bellal, H. Elghazel, A. Aussem, A semi-supervised feature ranking method with ensemble learning, Pattern Recognition Letters. 33 (2012) 1426–1433.
[22] Z. Song, X. Yang, Z. Xu, I. King, Graph-Based Semi-Supervised Learning: A Comprehensive Review, IEEE Transactions on Neural Networks and Learning Systems. (2022).
[23] M. Ghazvininejad, M. Mahdieh, H.R. Rabiee, P.K. Roshan, M.H. Rohban, Isograph: Neighbourhood graph construction based on geodesic distance for semi-supervised learning, in: Proceedings - IEEE International Conference on Data Mining, ICDM, 2011: pp. 191–200.
[24] N. Pourdamghani, H.R. Rabiee, F. Faghri, M.H. Rohban, Graph based semi-supervised human pose estimation: When the output space comes to help, Pattern Recognition Letters. 33 (2012) 1529–1535.
[25] W. Zhong, X. Chen, F. Nie, J. Zhexue, Adaptive discriminant analysis for semi-supervised feature selection, Information Sciences. 566 (2021) 178–194.
[26] M. Tubishat, S. Ja, M. Alswaitti, S. Mirjalili, Dynamic Salp Swarm Algorithm for Feature Selection, Expert Systems with Applications. 164 (2021) 113873.
[27] K. Benabdeslem, M. Hindawi, Constrained laplacian score for semi-supervised feature selection, in: Machine Learning and Knowledge Discovery in Databases, Springer, 2011: pp. 204–218.
[28] Y. Han, Y. Yang, Y. Yan, Z. Ma, N. Sebe, S. Member, Semisupervised feature selection via spline regression for video semantic recognition, IEEE Transactions on Neural Networks and Learning Systems, 26 (2015) 252–264.
[29] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society. Series B (Methodological). 58 (1996) 267–288.
[30] S. Foucart, M.-J. Lai, Sparsest solutions of underdetermined linear systems via ℓq-minimization for 0< q<1, Applied and Computational Harmonic Analysis. 26 (2009) 395–407.
[31] R. Chartrand, Exact reconstruction of sparse signals via nonconvex minimization, IEEE Signal Processing Letters. 14 (2007) 707–710.
[32] C. Shi, Q. Ruan, S. Member, G. An, R. Zhao, Hessian semi-supervised sparse feature selection based on L21/2-matrix norm, IEEE Transactions on Multimedia. 17 (2015) 16–28.
[33] R. Sheikhpour, M.A. Sarram, E. Sheikhpour, Semi-supervised sparse feature selection via graph Laplacian based scatter matrix for regression problems, Information Sciences. 468 (2018) 14–28.
[34] J. Lai, H. Chen, W. Li, T. Li, J. Wan, Semi-supervised feature selection via adaptive structure learning and constrained graph learning, Knowledge-Based Systems. 251 (2022) 109243.
[35] M. Fan, X. Zhang, J. Hu, N. Gu, D. Tao, Adaptive Data Structure Regularized Multiclass Discriminative Feature Selection, IEEE Transactions on Neural Networks and Learning Systems. (2021) 1–14.
[36] X. Li, Y. Zhang, R. Zhang, Semisupervised Feature Selection via Generalized Uncorrelated Constraint and Manifold Embedding, IEEE Transactions on Neural Networks and Learning Systems. (2021).

Journal of Applied And Basic Machine Intelligence Research (ABMIR)

Applied and Basic Machine Intelligence Research

Vol. 1, No. 1, Spring & Summer 2023, 73-87

A review on graph-based semi-supervised sparse feature selection methods

Razieh Sheikhpour[1*]

[1] **Department of Computer Engineering, Faculty of Engineering, Ardakan University, P.O. Box 184, Ardakan, Iran**

Article Information

Original Research Paper

Received:
26 June 2022

Accepted:
25 September 2022

Keywords:
Semi-supervised feature selection, Semi-supervised learning, Sparse models, Graph Laplacian

Corresponding Author*:
rsheikhpour@ardakan.ac.ir

Abstract

In some real-world applications, there is high-dimensional data which has led to many computational challenges. Feature selection is an effective technique for data dimensionality reduction, which simplifies the model and improves its performance by selecting the appropriate subset of features. In many of these applications, labeling of data is costly and time consuming, leaving little labeled data available and large amounts of unlabeled data available. In such applications, semi-supervised feature selection methods perform the feature selection process using the information of labeled data, and the distribution and geometric structure of labeled and unlabeled data. In most semi-supervised feature selection methods, a neighborhood graph is created and the importance of features is evaluated via their ability to maintain the geometric structure of the graph. In classical graph-based semi-supervised feature selection methods, the features are evaluated one by one and the correlation between features is not considered in feature selection process. To overcome this problem, sparse feature selection methods have been presented which consider the correlation between features, and calculate the optimal sparse transformation matrix for feature selection. In this paper, we investigate the semi-supervised learning methods, and review the graph-based semi- supervised sparse feature selection methods which select the appropriate features using the graph created by the labeled and unlabeled data, and the sparse regularization term. These methods solve the problem of classical semi-supervised methods by considering the correlation between features, create a neighborhood graph using the labeled and unlabeled data, calculate the graph Laplacian matrix, and compute the optimal sparse transformation matrix for feature selection.

: 10.22034/ABMIR.2022.2899

Journal of Applied And Basic Machine Intelligence Research (ABMIR)

Applied and Basic Machine Intelligence Research

Vol. 1, No. 1, Spring & Summer 2023, 73-87

A review on graph-based semi-supervised sparse feature selection methods