پردازش تصویر با استفاده از کدگذاری تنک و طبقه‌بندی انطباقی

فریماه شرافتی[۱]، دانشجوی کارشناسی ارشد؛ جعفر طهمورث نژاد[۲]، استادیار

۱- دانشکده مهندسی فناوری اطلاعات و کامپیوتر- دانشگاه صنعتی ارومیه- ارومیه- ایران- farimah.sherafati@it.uut.ac.ir
۲- دانشکده مهندسی فناوری اطلاعات و کامپیوتر- دانشگاه صنعتی ارومیه- ارومیه- ایران-j.tahmores@it.uut.ac.ir

چکیده: به‌دلیل افزایش حجم تصاویر تولیدشده توسط دوربین‌ها و دستگاه‌های مختلف، پردازش تصویر در بسیاری از کاربردها ازجمله پزشکی، امنیتی و رانندگی اهمیت و جایگاه بالایی یافته است. بااین‌حال بیشتر مدل‌های ایجادشده در حوزه پردازش تصویر کارایی چندانی نداشته و میزان خطای آن‌ها در برخی کاربردها تأثیرگذار است. علت اصلی ناکامی بیشتر مدل‌های ساخته‌شده، اختلاف توزیع بین داده‌های آموزشی (دامنه منبع) و داده‌های تست (دامنه هدف) می‌باشد. درواقع، مدل ساخته‌شده، قابلیت تعمیم‌دهی به داده‌هایی با خصوصیات و توزیع‌های متفاوت از داده‌های آموزشی را ندارد، به‌همین دلیل در مواجهه با داده‌های جدید دچار اُفت شدیدی می‌شود. در این مقاله ما یک روش جدید با نام کدگذاری تنک و طبقه‌بندی انطباقی (SADA) پیشنهاد می‌دهیم که یک مدل پردازش تصویری ایجاد می‌کند که در مقابل تغییرات داده‌ای مقاوم می‌باشد. مدل پیشنهادی با ایجاد یک زیر فضای مشترک بین دامنه‌های منبع و هدف اختلاف توزیع آن‌ها را به حداقل رسانده و موجب بهبود کارایی می‌شود. همچنین SADA با انتخاب نمونه‌هایی از دامنه منبع که با دامنه هدف مرتبط می‌باشند اختلاف توزیع بین دامنه‌ها را کاهش می‌دهد. علاوه‌بر آن، SADA با تطبیق پارامترهای مدل ایجادشده، یک مدل تطبیق‌پذیر برای مواجهه با شیفت داده‌ها ایجاد می‌کند. نتایج به‌دست‌آمده از آزمایش‌های متنوع، نشان می‌دهد که روش پیشنهادی ما، برتری قابل‌ملاحظه‌ای نسبت به تمام روش‌های تطبیق دامنه جدید دارد.

واژه‌های کلیدی: پردازش تصویر، تطبیق دامنه‌های بصری، کدگذاری تنک، وزن‌دهی مجدد نمونه، طبقه‌بندی انطباقی.

Image Processing via Sparse Coding and Adaptive Classification

F. Sherafati[1], MSc Student; J. Tahmoresnezhad[2], Assistant Professor

1- Faculty of IT & Computer Engineering, Urmia University of Technology, Urmia, Iran, farimah.sherafati@it.uut.ac.ir
2- Faculty of IT & Computer Engineering, Urmia University of Technology, Urmia, Iran, j.tahmores@it.uut.ac.ir

Abstrac: Due to the growing increase of generated images via cameras and various instruments, image processing has found an important role in most of practical usages including medical, security and driving. However, most of the available models has no considerable performance and in some usages the amount of error is very effective. The main cause of this failure in most of available models is the distribution mismatch across the source and target domains. In fact, the made model has no generalization to test data with different properties and distribution compared to the source data, and its performance degrades dramatically to face with new data. In this paper, we propose a novel approach entitled Sparse coding and ADAptive classification (SADA) which is robust against data drift across domains. The proposed model reduces the distribution difference across domains via generating a common subspace between the source and target domains and increases the performance of model. Also, SADA reduces the distribution mismatch across domains via the selection of the source samples which are related to target samples. Moreover, SADA adapts the model parameters to build an adaptive model to encounter with data drift. Our variety of experiments demonstrate that the proposed approach outperforms all stat-of-the-art domain adaptation methods.

Keywords: Image processing, visual domains adaptation, sparse coding, sample reweighting, adaptive classification.

تاریخ ارسال مقاله: ۱۳۹۶/۰۷/۰۵
تاریخ اصلاح مقاله:۱۳۹۶/۰۹/۰۵
تاریخ پذیرش مقاله: ۱۳۹۶/۱۰/۱۲
نام نویسنده مسئول: جعفر طهمورث نژاد
نشانی نویسنده مسئول: ایران - ارومیه - دانشگاه صنعتی ارومیه - دانشکده مهندسی فناوری اطلاعات و کامپیوتر.

۱- مقدمه

یادگیری ماشین، از شاخه‌های پرکاربرد در هوش مصنوعی می‌باشد که با ایجاد الگوریتم و برنامه‌های خودکار به دنبال هوشمندسازی ماشین‌ها و مدل‌ها می‌باشد. در الگوریتم‌های طبقه‌بندی یادگیری ماشین، یک مدل با استفاده از داده‌های برچسب‌دار موجود آموزش داده شده و بر روی داده‌های تست، جهت برچسب‌گذاری نمونه‌های جدید اعمال می‌شود. یک فرض مشترک در تمام الگوریتم‌های طبقه‌بندی وجود دارد، که توزیع و فضای خصیصه‌ای نمونه‌های آموزشی و تست را یکسان در نظر می‌گیرد. اما در کاربردهای دنیای واقعی، این فرضیه همیشه برقرار نیست. به‌عبارت دیگر، داده‌های آموزشی و تست ممکن است از دامنه‌های کاملاً متفاوتی باشند. در این حالت، مدل ساخته‌شده روی داده‌های آموزشی، کارایی پایینی روی دامنه تست خواهد داشت.

اختلاف توزیع بین نمونه‌های آموزشی و تست از آنجا نشأت می‌گیرد که در برخی سیستم‌ها به تعداد کافی داده آموزشی برچسب‌دار، جهت آموزش مدل وجود ندارد، در این حالت از دامنه‌های برچسب‌دار مرتبط، جهت ایجاد یک مدل، برای برچسب‌گذاری داده‌های تست استفاده می‌شود [۱، ۲]. با این اوصاف داده‌های تست برچسب‌گذاری می‌شوند ولی در بررسی‌های اولیه مشخص می‌شود که اغلب برچسب‌های نسبت داده شده دقیق نبوده و مدل از خطای بالایی رنج می‌برد. به چنین مشکلی، شیفت دامنه‌ها[۱] گفته می‌شود که در آن دامنه‌های آموزشی و تست از توزیع متفاوتی برخوردارند.

یادگیری انتقالی[۲] یا تطبیق دامنه‌ها[۳] یک بهبود یادگیری در یک حوزه جدید از طریق انتقال دانش از یک حوزه مرتبط می‌با شد که قبلاً یاد گرفته شده است. درواقع روش‌های یادگیری انتقالی، یک بهبود بر روی روش‌های یادگیری ماشین هستند، که آن‌ها را جهت مواجهه با مسائلی که در آن‌ها شیفت دامنه‌ها و اختلاف توزیع وجود دارد مقاوم می سازند. یادگیری انتقالی بر اساس اطلاعات برچسب دامنه هدف، به دو دسته تقسیم می شود: ۱- یادگیری انتقالی نیمه‌نظارت شده که در آن تمام نمونه‌های دامنه منبع دارای برچسب بوده، و تعداد کمی از نمونه‌های دامنه هدف دارای برچسب می‌باشند، ۲- یادگیری انتقالی بدون نظارت که در آن تمام نمونه‌های دامنه هدف بدون برچسب هستند.

وجود اختلاف توزیع بین دامنه‌ها موجب می‌شود مدلی که روی داده‌های آموزشی ایجاد شده است، عملکرد خوبی روی داده‌های تست نداشته باشد. برای سنجش اختلاف توزیع حاشیه‌ای بین دو دامنه از یک معیار غیرپارامتری به نام MMD[۴] استفاده می‌شود. MMD داده‌های منبع و هدف را به یک فضای RKHS[۵] نگاشت می‌کند و با محاسبه میانگین عناصر[۶] هرکدام از دامنه‌ها در فضای مزبور، اختلاف دو دامنه را به‌دست می‌آورد. اثبات شده است [۳]، میزان فاصله میانگین عناصر دو دامنه در فضای RKHS معادل فاصله دو دامنه در فضای اصلی است. بدین ترتیب درصورتی‌که فاصله میانگین عناصر داده‌های منبع و هدف در فضای RKHS کاهش یابد، داده‌ها با یکدیگر تطبیق شده و مدل قوی‌تر و تواناتری ایجاد می‌شود.

کدگذاری تنک[۷] ازجمله روش‌های بدون نظارت[۸] است که برای نمایش داده‌ها به‌صورت کارا مورد استفاده قرار می‌گیرد [۴، ۵]. هدف کدگذاری تنک، یافتن یک مجموعه از بردارهای اصلی است که یک بردار ورودی را بتوان بر اساس یک ترکیب خطی از بردارهای اصلی نمایش داد. مزیت استفاده از بردارهای اصلی در کدگذاری تنک این است که بردارهای اصلی به‌خوبی توانایی حفظ ساختارها و الگوهای داده ورودی را دارند، بدین ترتیب پردازش بر روی داده کدگذاری شده معادل پردازش بر روی داده‌های ورودی است.

از دیگر روش‌های مقابله با مسئله شیفت دامنه‌ها استفاده از طبقه‌بند انطباقی می‌باشد که با افزایش تطبیق‌پذیری بین دامنه‌های منبع و هدف خطای پیش‌بینی را به حداقل می‌رساند. درواقع طبقه‌بند انطباقی، ابعاد تفکیک‌کننده کلاس‌ها بین دامنه‌های منبع و هدف را با یکدیگر سازگار می‌نماید؛ به‌عبارت دیگر با بهره‌گیری از توزیع هند سی داده‌ها در ایجاد طبقه‌بند انطباقی، مدل طبقه‌بند ایجادشده بر روی نمونه‌های دامنه منبع با ساختار هندسی دامنه هدف، سازگاری بیشتری خواهد یافت.

روش پیشنهادی در این مقاله با عنوان SADA[۹]، یک روش نیمه‌نظارت‌شده با بهره‌گیری از کدگذاری تنک، تطبیق خصوصیات و طبقه‌بند انطباقی[۱۰] است. SADA با استفاده از ماتریس تنک و به کارگیری نرم $l_{2,1}$[۱۱] نمونه‌هایی از دامنه منبع که مرتبط و نزدیک به دامنه هدف می‌باشند را انتخاب می‌نماید. سپس با استفاده از تطبیق خصوصیات یک زیر فضای جدید به‌دست می‌آورد که در آن خصوصیات مشترک دو دامنه منبع و هدف نگهداری می‌شود به‌طوری‌که اختلاف توزیع حاشیه‌ای دامنه‌ها به حداقل رسانیده می‌شود. در انتها SADA از یک طبقه‌بند انطباقی با هدف حداقل‌کردن خطای پیش‌بینی مدل روی داده‌های منبع و حداکثرکردن سازگاری بین ساختار منیفلد و مدل پیش‌بینی استفاده می‌نماید.

ادامه این مقاله به‌صورت زیر سازماندهی شده است. مروری بر کارهای گذشته در بخش دوم مقاله آورده شده است. تعاریف اولیه و مفاهیم مقدماتی در بخش سوم قرار داده شده است. روش پیشنهادی در بخش چهارم معرفی شده است. پایگاه‌داده‌های مورد آزمایش و نتایج ارزیابی عملکرد روش پیشنهادی در مقایسه با دیگر روش‌های یادگیری ماشین و یادگیری انتقالی در بخش پنجم گزارش شده است. در انتها، نتیجه‌گیری و کارهای پیشنهادی آتی گنجانده شده است.

۲-کارهای پیشین

برای حل مشکل اختلاف توزیع بین دامنه‌ها، روش‌های تطبیق دامنه زیادی پیشنهاد شده است. به‌طور کلی کارهایی که در حوزه یادگیری انتقالی انجام می‌شوند به سه دسته تقسیم می‌شوند: رویکردهای مبتنی

بر نمونه [۴]، رویکردهای مبتنی بر خصوصیت [۶] و رویکردهای مبتنی بر تطبیق طبقه‌بند [۷].

در رویکردهای مبتنی بر نمونه، نمونه‌های منبع به‌گونه‌ای وزن‌دهی مجدد می‌شوند تا توزیع نزدیک‌تری با نمونه‌های دامنه هدف داشته باشند. یکی از جدیدترین روش‌ها در این رویکرد، روش لندمارک [۸] می‌باشد. در روش لندمارک، آن نمونه‌هایی از دامنه منبع انتخاب می‌شوند که از لحاظ توزیع دارای بیشترین شباهت با نمونه‌های دامنه هدف باشند. برای به‌دست‌آوردن اختلاف توزیع بین دو دامنه از MMD استفاده می‌شود و به نمونه‌هایی از دامنه منبع که اختلاف توزیع کمتری با نمونه‌های دامنه هدف دارند، وزن بیشتری تعلق می‌گیرد. روش TJM[۱۲] [۹] یک روش ترکیبی از روش‌های مبتنی بر نمونه و مبتنی بر خصوصیت است که برای مسائل با اختلاف توزیع بالا به‌کار گرفته می‌شود. TJM ابتدا اختلاف توزیع حاشیه‌ای بین دامنه‌ها را کاهش داده، سپس نمونه‌ها را وزن‌دهی مجدد می‌کند و نمونه‌های منبع که توزیع نزدیک‌تری با نمونه‌های هدف دارند را انتخاب می‌کند. TJM، دارای یک کرنل نسبتاً پیچیده می‌باشد و حل تابع هدف آن دشوار می‌باشد. همچنین TJM به‌دلیل حذف بخشی از داده ورودی، ساختار داده‌ها را حفظ نمی‌کند. روش SSSC[۱۳] [۴]، ابتدا نمونه‌های غیرمرتبط دامنه منبع را حذف کرده، سپس با یافتن یک ماتریس انتخاب، اختلاف توزیع بین داده‌های منبع و هدف را کاهش می‌دهد. SSSC از کدگذاری تنک برای انتخاب نمونه‌های دامنه منبع استفاده کرده و هر نمونه دامنه هدف را با حداقل تعداد نمونه دامنه منبع نمایش می‌دهد.

در رویکردهای مبتنی بر خصوصیت، هدف یادگیری یک نمایش جدید از ویژگی‌ها است. GFK[۱۴] [۶] یک روش مبتنی بر خصوصیت است که داده‌های منبع و هدف را بر روی منیفلد گرسمن[۱۵] نگاشت می‌کند تا اختلاف توزیع حاشیه‌ای آن‌ها کاهش یابد. GFK بین زیرفضاهای منبع و هدف یک مسیر کوتاه[۱۶] بر روی منیفلد ایجاد می‌کند تا تغییرات دامنه از منبع تا هدف را مورد ارزیابی قرار دهد. از آن‌جایی که حرکت بر روی مسیر کوتاه بایستی به‌صورت روان و پیوسته باشد GFK اندازه زیرفضا را کوچک در نظر می‌گیرد که موجب از بین رفتن بخشی از داده‌های ورودی می‌شود. VDA[۱۷] [۱۰] برای کاهش اختلاف توزیع حاشیه‌ای و شرطی بین دامنه‌ها، یک نمایش کم بعد از دامنه‌های منبع و هدف ایجاد می‌کند. همچنین، VDA از خوشه‌بندی مستقل از دامنه[۱۸] جهت حفظ ساختار هندسی و آماری بین دامنه‌ها در فضای جدید استفاده می‌کند.

رویکردهای مبتنی بر تطبیق طبقه‌بند، با ایجاد طبقه‌بندهای انطباقی، مدل‌های مقاوم‌تری در مقابل اختلاف توزیع بین دامنه‌های منبع و هدف ایجاد می‌کنند. درواقع طبقه‌بند انطباقی بدون ایجاد تغییر در فضای ویژگی، پارامترهای مدل طبقه‌بند ایجادشده بر روی داده‌های دامنه منبع را با دامنه هدف سازگار می‌سازد. بااین‌حال، عمده روش‌های مبتنی بر مدل بر روی حل مسائل تطبیق دامنه نیمه‌نظارت‌شده تمرکز دارند. ARTL[۱۹] [۱۱] یک روش تطبیق طبقه‌بند است که به ایجاد یک مدل انطباقی از طریق ایجاد تطبیق توزیع مشترک، حداقل‌کردن خطای پیش‌بینی و حداکثرکردن سازگاری هندسی بین دامنه‌های منبع و هدف به‌طور هم‌زمان تطبیق دامنه‌ها را انجام می‌دهد.

در این مقاله، مسئله شیفت دامنه‌ها با استفاده از یک راه‌حل نیمه‌نظارت‌شده مورد حل قرار می‌گیرد. درواقع، با ایجاد یک نمایش جدید مبتنی بر خصوصیت به کمک نمایش تنک، فاصله دامنه‌های منبع و هدف کاهش یافته و با افزودن یک طبقه‌بند انطباقی پارامترهای مدل برای مواجهه با اختلاف توزیع‌های بالا قدرتمندتر و مقاوم‌تر می‌شوند.

۳- تعاریف اولیه و مفاهیم مقدماتی

در این بخش، ابتدا تعریف مسئله ارائه شده و سپس مفاهیم مقدماتی برای ارائه روش پیشنهادی گنجانده شده است.

۳-۱- هدف تحقیق

در حوزه یادگیری ماشین، تطبیق دامنه‌ها به یک ضرورت اجتناب‌ناپذیر تبدیل شده است، زیرا دیگر، الگوریتم‌های کلاسیک یادگیری ماشین پاسخگوی مسائل موجود در دنیای واقعی نیستند. در این مقاله، ما به دنبال پیشنهاد روشی هستیم که بتواند بر محدودیت‌های الگوریتم‌های استاندارد یادگیری ماشین غلبه کرده و با انتقال دانش از یک دامنه به دامنه دیگر، بازدهی روش‌های موجود را افزایش دهد. به‌طورکلی ما در این مقاله به‌دنبال برآورده‌کردن اهداف ذیل هستیم: (۱) انتقال دانش از یک دامنه به دامنه دیگر، (۲) کاهش اختلاف توزیع داده‌های آموزشی و تست، و (۳) مقاوم‌کردن مدل‌ها در برابر تغییرات احتمالی داده‌ها. در ادامه تعریف مسئله به‌صورت ریاضی بیان شده است.

۳-۲- تعریف مسئله

دامنه[۲۰]- هر دامنه D دارای دو عنصر فضای خصیصه‌ای X و توزیع احتمال حاشیه‌ای $P(X)$ می‌باشد؛ بنابراین دامنه D به‌صورت زیر نشان داده می‌شود: $D=\{X,P(X)\}$. بدین ترتیب، دامنه‌های منبع و هدف زمانی با یکدیگر تفاوت دارند که هرکدام یا از فضای خصیصه‌ای مختلفی برخوردار بوده و یا دارای توزیع احتمال حاشیه‌ای متفاوتی باشند.

وظیفه[۲۱]- برای هر دامنه D یک وظیفه T وجود دارد که از دو مجموعهٔ برچسب Y و تابع پیش‌بینی $f(x)$ تشکیل می‌شود؛ بنابراین وظیفه T به‌صورت زیر نشان داده می‌شود: $T=\{Y,f(x)\}$. مجموعه برچسب Y برای نمونه‌های ورودی X توسط تابع $f(x)$ پیش‌بینی می‌شود و احتمال شرطی آن به‌صورت $P(Y\mid x)$ تعریف می‌شود. بدین ترتیب، زمانی دو وظیفه متفاوت هستند که یا مجموعه برچسب و یا تابع توزیع احتمال شرطی متفاوتی داشته باشند.

اگر $P_s(x)$ و $P_t(x)$ به‌ترتیب نشان‌دهنده توزیع احتمال حاشیه‌ای داده‌های منبع و هدف با شند، در صورت بروز شیفت دامنه‌ها خواهیم دا شت: $P_s(x) \neq P_T(x)$. بدین ترتیب، این مقاله به‌دنبال یک نمایش جدید و کارا برای داده‌های منبع و هدف اســت که پس از نگاشــت نمونه‌ها به زیر ف ضای جدید از توزیع م شابهی برخوردار با شند، یعنی: $P_s(x) \approx P_t(x)$. در ادا مه م تدولوژی های مورد ن یاز ج هت ای جاد زیرفضای جدید به تفکیک معرفی شده‌اند.

۳-۳- کدگذاری تنک

در تطبیق دامنه نیمه‌نظارت‌شــده، تمام نمونه‌های دامنه منبع دارای برچ سب بوده اما از بین نمونه‌های دامنه هدف فقط تعداد کمی از آن‌ها برچســب داشــته که به‌تن‌هایی قادر به ای جاد یک مدل ج هت برچســب‌گذاری نمونه‌های هدف نمی‌باشــند. اگر $X_s \in R^{d_s * n_s}$ و $X_t \in R^{d_t * n_t}$ که n_s و n_t به‌ترتیب تعداد نمونه‌های منبع و هدف و $d_s = d_t$ ابعاد ف ضای خ صو صیات با شد، هدف، یافتن نمونه‌های منبع مرتبط با نمونه‌های هدف ا ست تا بتوان طبقه‌بندی ساخت که به‌کمک آن به نمونه‌های بدون برچسب دامنه هدف برچسب مناسب نسبت داد. برای انتخاب نمونه‌ها از دامنه منبع از کدگذاری تنک استفاده می‌شود.

با استفاده از نرم ۱ می‌توان هر نمونه هدف را با حداقل تعداد نمونه منبع نمایش داد. رابطه زیر ماتریس انتخاب $A \in R^{n_s * n_t}$ را به‌دســت می‌آورد که اختلاف توزیع بین داده‌های دامنه منبع و هدف را حداقل می‌سازد.

$$\min_{A} \| X_t - X_s A \|_F^2 + \rho \| A \|_1 \qquad (۱)$$

که $\|.\|_F$ نرم فروبنیوس[۲۲] می‌باشــد و ماتریس A یک ماتریس تنک است، که با استفاده از مسئله بهینه‌سازی فوق به‌دست می‌آید و با وزن‌دهی مجدد به نمونه‌های منبع، فاصــله بین دامنه‌های منبع و هدف را کاهش می‌دهد. ρ پارامتر تنظیم پراکندگی[۲۳] اســت. با به‌کارگیری نرم ۱ تعداد ضــرایب غیر صــفر حداقل می‌شــود ولی نرم ۱ منجر به انتخاب نمونه نمی شود زیرا دقیقاً م شخص نی ست هر نمونه منبع دقیقاً با کدام‌یک از نمونه‌های هدف مرتبط ا ست. با ا ستفاده از رابطه ۲ از نرم $l_{2,1}$ به‌جای نرم ۱ اســتفاده می‌شــود که می‌تواند هر ســطر ماتریس را هم‌زمان صفر یا غیر صفر کند. سطرهای غیر صفر هر نمونه هدف را با حداقل تعداد نمونه منبع نشان می‌دهند، بنابراین سطرهای غیر صفر به عنوان نمونه‌های منبع مرتبط انتخاب می‌شوند.

$$\min_{A} \| X_t - X_s A \|_F^2 + \rho \| A \|_{2,1} \qquad (۲)$$

بدین ترتیب با استفاده از نرم $l_{2,1}$ یک مجموعه محدود از نمونه‌های منبع به‌دست می‌آید که می‌تواند همه نمونه‌های هدف را به‌خوبی توصیف کند.

۳-۴- کاهش بعد

در روش های کاهش بعد، داده‌ها از فضــای اصــلی به یک فضــای نگاشــت‌شــده منتقل می‌شــوند به‌شــرطی که هزینه بازگرداندن آن‌ها به فضای اصلی[۲۴] حداقل باشد. PCA[۲۵] [۱۲] یکی از مرسوم‌ترین روش‌ها برای انتقال داده‌ها به یک ف ضای کم بعد می‌با شد، به‌طوری‌که ساختار اصلی داده‌ها را در فضای جدید حفظ می‌کند. درواقع PCA داده‌ها را بر روی اجزای اصلی خود نگاشت کرده به‌طوری‌که داده‌ها بر روی این اجزا دارای حداکثر واریانس باشند. اجزای اصلی به‌دست‌آمده، جهت نگاشت داده‌ها به فضــای کم‌بعد مورد اســتفاده قرار می‌گیرند. اگر p ابعاد زیرفضــای جدید و $U_s \in R^{d_s * p}$ و $U_t \in R^{d_t * p}$ ماتریس‌های نگاشــت به فضای کم‌بعد باشند، رابطه ۲ به‌صورت زیر بازنویسی می‌شود:

$$\min_{A, U_t, U_s} \| U_t^T X_t - U_s^T X_s A \|_F^2 + \rho \| A \|_{2,1} \qquad (۳)$$

با اســتفاده از رابطه ۳ یک ماتریس انتخاب A و دو ماتریس نگاشت U_s و U_t به‌دست می‌آید که فاصله بین داده‌های منبع و هدف را در زیرفضــای جدید کاهش می‌دهد. همچنین PCA برای حفظ واریانس داده‌های منبع و هدف از شــرط زیر در کنار رابطه ۳ اســتفاده می‌کند که در آن V ماتریس هم‌مرکزنمودن است که برای هر دو دامنه با اســتفاده از رابطه $V = I - \frac{1}{n_s} \vec{1}\vec{1}^T$ به‌دســت می‌آید و از پراکندگی داده‌ها در زیرف ضای تعبیه شده جلوگیری می‌کند که در آن I ماتریس همانی و $\vec{1}$ بردار یک‌ها است.

$$\max_{U} \; tr(U^T X V X^T U) \qquad (۴)$$

که $U = [U_s, U_t]$ ماتریس نگاشــت منبع و هدف و $X = [X_s, X_t]$ مجموعه داده‌های منبع و هدف می‌باشد.

۳-۵- حفظ توپولوژی

به اثبات رسـیده اسـت که حفظ سـاختار و توپولوژی داده‌ها در فضـای نگا شت شده می‌تواند کارایی مدل را به صورت قابل‌توجهی افزایش دهد [۱۰]. به‌منظور حفظ ســاختار توپولوژی هر دامنه، از فرضــیات منیفلد استفاده می‌شود. به‌همین دلیل در نگاشت داده‌ها با استفاده از ماتریس U به فضای جدید توپولوژی و ساختار داده‌های نگاشت‌شده با استفاده از فرضــیه منیفلد حفظ می‌شــود. مطابق با فرضــیه منیفلد، اگر دو داده X_i و X_j در فضای اصلی در همسایگی هم قرار داشته باشند، بایستی در زیرفضای جدید نیز هم‌سایگی آن‌ها حفظ شود. با ا ستفاده از رابطه $S(X_i, X_j) = e^{\frac{-\|x_i - x_j\|_F^2}{\delta^2}}$ که در آن δ میانگین فاصــله‌های داده‌ها اســت، می‌توان میزان همســایگی هر دو نمونه را به‌دســت آورد. از روی ماتریس همســایگی ماتریس لاپلاســین به‌دســت می‌آید که فاصــله هر نمونه از تمام نمونه‌ها به‌جز خودش را محاســبه می‌کند. بدین ترتیب با حداقل‌کردن رابطه زیر، می‌توان نگاشــت U را به‌گونه‌ای به‌دســت آورد

که همسایگی نمونه‌های منبع و هدف در فضای نگاشت‌شده حفظ شده باشد:

$$tr\left(\mathbf{U}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{U}\right)\ \min_{\mathbf{U}} \qquad (۵)$$

که در آن $\mathbf{L}$ ماتریس لاپلاسین نمونه‌های منبع و هدف است.

۴- روش پیشنهادی

این مقاله به‌دنبال پیشنهاد روشی است که بتواند یک نمایش با بعد کم برای داده‌های منبع و هدف ایجاد کند که در آن فاصله دو دامنه به حداقل مقدار ممکن برسد. همچنین نمونه‌هایی از دامنه منبع که به دامنه هدف مرتبط نمی‌باشند، حذف شوند. برای این اهداف، از نمایش تنک برای انتخاب نمونه‌ها و PCA جهت کاهش بعد داده‌ها استفاده شده است. به‌علاوه، برای کاهش فاصله بین دامنه‌های منبع و هدف با استفاده از یک روش نوین، توزیع حاشیه‌ای بین دو دامنه را به حداقل می‌رساند.

اختلاف توزیع حاشیه‌ای بین دامنه‌های منبع و هدف با استفاده از یک روش غیر پارامتری به نام MMD کاهش می‌یابد. در این روش داده‌ها به فضای RKHS برده شده و در آن‌جا اختلاف بین میانگین نمونه‌های دامنه‌های منبع و هدف محاسبه می‌شود. اگر $\mathbf{A}$ تابع نگاشت به فضای جدید در نظر گرفته شود، اختلاف توزیع حاشیه‌ای دامنه‌های منبع و هدف به‌صورت رابطه زیر تعریف می‌شود [۱۱، ۱۲]:

$$Mrg\left(\mathbf{X_s},\mathbf{X_t}\right)=\|\frac{1}{n_s}\sum_{i=1}^{n_s}\mathbf{A}\mathbf{X_i}-\frac{1}{n_t}\sum_{j=n_s+1}^{n_s+n_t}\mathbf{A}\mathbf{X_i}\|_H^2 \qquad (۶)$$

برای حل ساده‌تر رابطه (۶) می‌توان آن را به‌فرم بسته زیر نوشت:

$$Mrg\left(\mathbf{X_s},\mathbf{X_t}\right)=tr\left(\mathbf{A}^T\mathbf{X}\mathbf{M}_0\mathbf{X}^T\mathbf{A}\right) \qquad (۷)$$

که ماتریس $\mathbf{M}_0\in R^{(n_s+n_t)*(n_s+n_t)}$ همان ماتریس MMD است و اگر $\mathrm{x_i},\mathrm{x_j}\in\mathbf{X_s}$ باشد آنگاه $(\mathbf{M}_0)_{ij}=\frac{1}{n_s n_s}$، اگر $\mathrm{x_i},\mathrm{x_j}\in\mathrm{X_t}$ باشد آنگاه $(\mathbf{M}_0)_{ij}=\frac{1}{n_t n_t}$ در غیر این‌صورت $(\mathbf{M}_0)_{ij}=\frac{-1}{n_s n_t}$ است.

۴-۱- فرموله‌سازی روش SADA

روش پیشنهادی SADA ترکیبی از رویکردهای مبتنی بر نمونه و رویکردهای مبتنی بر خصوصیت است. به‌همین دلیل از روابط ۳ و ۴ جهت انتخاب نمونه و یافتن فضای خصوصیات استفاده می‌شود. همچنین با توجه به این‌که در نگاشت داده‌ها به فضای جدید باید ستی توپولوژی داده‌ها حذف شود و اختلاف دو دامنه نیز کاهش یابد، ماتریس نگاشت $\mathbf{A}$ به‌گونه‌ای حاصل می‌شود که اختلاف حاشیه‌ای دامنه‌ها با استفاده از روابط ۵ و ۶ حداقل شود. بدین ترتیب تابع هدف SADA به‌صورت زیر به‌دست می‌آید:

$$\min_{\mathbf{A},\mathbf{U_t},\mathbf{U_s}}\ \|\mathbf{U_t}^T\mathbf{X_t}-\mathbf{U_s}^T\mathbf{X_s}\mathbf{A}\|_F^2+\rho\|\mathbf{A}\|_{2,1}+\mu tr\left(\mathbf{U}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{U}\right)+\|\mathbf{U}\|_2^2$$
$$s.t.\ \mathbf{U}^T\mathbf{X}\mathbf{V}\mathbf{X}^T\mathbf{U}=\mathbf{I} \qquad (۸)$$

قسمت اول رابطه (۸) یک انتخاب نمونه انجام می‌دهد. قسمت دوم رابطه (۸) یک محدودیت روی کدگذاری تنک است که بتواند هر نمونه هدف را با حداقل تعداد نمونه منبع توصیف کند. قسمت سوم رابطه (۸) به‌منظور حفظ هندسه اصلی نمونه‌ها در فضای نگاشت‌شده است و قسمت چهارم رابطه (۸) برای جلوگیری از جواب‌های بدیهی است و محدودیت رابطه (۸) برای ماکزیمم‌کردن واریانس داده‌ها است تا کلاس‌ها به‌خوبی از یکدیگر متمایز شوند.

روش پیشنهادی SADA با استفاده از رابطه ۸ دو ماتریس نگاشت $\mathbf{U}_s$ و $\mathbf{U}_t$ را با استفاده از PCA به‌دست می‌آورد و ماتریس نگاشت $\mathbf{A}$ را با حل معادله جهت نگاشت داده‌های منبع و هدف به زیرفضای جدید ایجاد می‌کند. بااین‌حال، در بیشتر مسائل دنیای واقعی اختلاف توزیع دامنه‌های منبع و هدف به‌اندازه‌ای است که صرفاً با نگاشت داده‌ها به فضای جدید، توزیع داده‌ها یکسان نشده و کارایی مدل بهبود داده نمی‌شود. از این رو، در مرحله دوم الگوریتم از یک طبقه‌بند انطباقی جهت مقاوم‌کردن مدل در مقابل تغییرات توزیع داده‌ها بهره برده می‌شود.

۴-۲- طبقه‌بند انطباقی

هدف از ایجاد طبقه‌بند انطباقی رسیدن به دو هدف زیر است: (۱) حداقل‌کردن خطای پیش‌بینی مدل روی داده‌های منبع و (۲) حداکثرکردن سازگاری بین ساختار منیفلد و مدل پیش‌بینی. درواقع، طبقه‌بند انطباقی پارامترهای مدل را به‌گونه‌ای به‌دست می‌آورد که مقاومت مدل در مقابل تغییرات داده بین دو دامنه افزایش یابد و کارایی آن با شیفت داده‌ها متأثر نشود. از همین رو، تابع خطای مدل به‌صورت زیر تعریف می‌شود که در آن f طبقه‌بند انطباقی و $g\left(x\right)$ داده‌های نگاشت‌شده توسط ماتریس نگاشت $\mathbf{A}$ می‌باشند.

$$\min_{f\in H_k}\sum_{i=1}^{n_s}l\left(f\left(g\left(x_i\right)\right),y_i\right)+\sigma f_k^2+\gamma M_f\left(\mathbf{P_s}\left(g\left(x\right)\right),\mathbf{P_t}\left(g\left(x\right)\right)\right) \qquad (۹)$$

که در آن $\mathrm{H_k}$ مجموعه طبقه‌بندها و σ و γ پارامترهای تنظیم هستند. بخش اول رابطه (۹) پارامترهای مدل را به‌گونه‌ای تعیین می‌کند که خطای پیش‌بینی مدل حداقل شود. بخش دوم رابطه (۹) شامل مربع طبقه‌بند انطباقی جهت جلوگیری از جواب‌های بدیهی است. بخش سوم رابطه (۹) جهت حفظ ساختار و توپولوژی داده‌ها در قالب طبقه‌بند از لاپلاسین داده‌ها استفاده می‌کند.

جهت به‌دست‌آوردن مقادیر بهینه پارامترهای مدل f از رابطه (۹) مشتق گرفته و برابر صفر قرار داده شده است تا پارامترهای انطباقی طبقه‌بند به‌صورت زیر به‌دست آیند:

$$\alpha=(\sigma\mathbf{I}+\left(\mathbf{R}+\gamma L\right)\mathbf{X})^{-1}\mathbf{R}Y^T \qquad (۱۰)$$

که در آن $\mathbf{I}$ ماتریس همانی است. $\mathbf{R}$ یک ماتریس قطری است که نشان‌دهنده نمونه‌های منبع است و Y برچسب‌های نمونه‌های منبع است.

با به‌دست‌آوردن پارامترهای طبقه‌بند انطباقی با استفاده از رابطه (۱۰) می‌توان مدل را با استفاده از یک طبقه‌بند استاندارد (نظیر SVM) ایجاد کرد.

۴-۳- الگوریتم روش SADA

برای بهینه سازی پارامترهای مدل، به‌دلیل اینکه نمی‌توان هم‌زمان چند پارامتر را با هم بهینه کرد به‌صورت یک‌به‌یک بهینه‌سازی پارامترها انجام می‌شود. با استفاده از یک الگوریتم بهینه‌سازی متناوب در دو مرحله می‌توان مسئله را حل کرد. در مرحله اول ماتریس تنک $\mathbf{A}$ را با ثابت نگه‌دا شتن $\mathbf{U_s}$ و $\mathbf{U_t}$ می‌توان محا سبه کرد. سپس در مرحله دوم ماتریس‌های نگاشت $\mathbf{U_s}$ و $\mathbf{U_t}$ با ثابت نگه‌داشتن $\mathbf{A}$ محاسبه می‌شوند. با مشتق‌گیری از رابطه (۸) نسبت به $\mathbf{A}$ و مساوی قراردادن آن با صفر، مقدار $\mathbf{A}$ بهینه با استفاده از رابطه زیر به‌دست می‌آید:

$$\mathbf{A}=(\rho\mathbf{N}+\mathbf{X_s^T U_s U_s^T X_s})^{-1}\mathbf{X_s^T U_s U_t^T X_t} \qquad (۱۱)$$

که در رابطه فوق، $\mathbf{N}=\dfrac{1}{2\parallel \mathbf{A}^i \parallel_2}$ یک ماتریس قطری و وابسته به $\mathbf{A}$ است و $\mathbf{A}$ با یک روش تکراری به‌دست می‌آید. نحوه به‌دست آوردن $\mathbf{A}$ در شکل ۱ ارائه شده است.

Algorithm 1. An algorithm for solving selection matrix A

1: Input: Source data X_s,Target data X_t, Projection matrix U_s , U_t

2: Parameter: Regularizaton parameters ρ

3: Output: Selection matrix A.

5: Compute selection matrix A via solving Eq. (11).

6: Compute diagonal matrix $\mathbf{N}=\dfrac{1}{2\parallel A^i \parallel_2}$ based on A.

7: Repeat 5 and 6 until the maximum number of iterations is reached or optimum values of Eq. (8) is found.

شکل ۱: الگوریتم به‌دست‌آوردن ماتریس A

همچنین، با فرض اینکه متغیر A در رابطه (۸) ثابت باشد، بایستی نسبت به متغیر U مشتق گرفته شود. با توجه به اینکه محدودیت مسئله خود برحسب متغیر U می‌باشد به کمک لاگرانژ مقدار بهینه متغیر U به‌دست می‌آید. مقدار تابع لاگرانژ به‌صورت زیر محاسبه می‌شود:

$$\min_{\mathbf{U}} tr\left(\mathbf{U}^T\mathbf{X}(\mathbf{M}+\mathbf{Z})\mathbf{X}^T\mathbf{U}\right)-tr\left(\left(\mathbf{U}^T\mathbf{X}\mathbf{V}\mathbf{X}^T\mathbf{U}-\mathbf{I}\right)\mathbf{Q}\right) \qquad (۱۲)$$

که $\mathbf{Q}=\mathrm{diag}(q_1,\ldots,q_p)\in\mathbf{R^{p*p}}$ ضرایب لاگرانژ بوده و یک ماتریس قطری است. با مشتق‌گیری از رابطه (۱۲) برحسب $\mathbf{U}$ و برابر صفر قرار دادن آن، رابطه زیر به‌دست می‌آید:

$$\mathbf{X}(\mathbf{M}+\mathbf{Z})\mathbf{X}^T\mathbf{U}=\mathbf{X}\mathbf{V}\mathbf{X}^T\mathbf{U}\mathbf{Q} \qquad (۱۳)$$

که $\mathbf{M}+\mathbf{Z}=\begin{bmatrix}\mathbf{A}\mathbf{A}^T+\mu\mathbf{L_s}+\tau\mathbf{Z_{ss}} & -\mathbf{A}+\tau\mathbf{Z_{st}}\\ -\mathbf{A}^T+\tau\mathbf{Z_{ts}} & \eta\mathbf{I}+\mu\mathbf{L_t}+\tau\mathbf{Z_{tt}}\end{bmatrix}$ است و μ پارامتر تنظیم گراف است. روش SADA با جزئیات کامل، در شکل ۲ ارائه شده است.

Algorithm 2. Image processing via sparse coding and adaptive classification (SADA)

1: Input: Source data X_s,Target data X_t, Projection matrix $\mathbf{U_s^{'}}$, $\mathbf{U_t^{'}}$, subspace dimension P, Y_s source data label.

2: Parameters: Regularizaton parameters δ,γ

3: Output: Target domain labels Y_t

4: Randomly initialize $\mathbf{U}\in\mathbf{R}^{(d_s+d_t)*p}$.

6: Compute selection matrix A via Algorithm 1.

7: Compute projection matrix **U** via solving Eq. (13).

8: Repeat 6 and 7 until the maximum number of iterations is reached or optimum values of Eq. (8) is found.

// Data projection

9: $\mathbf{X_s^{'}}=\mathbf{U_s^T X_s A}$

10: $Y_s^{'}=Y_s\mathbf{A}$

11: $\mathbf{X_t^{'}}=\mathbf{U_t^T X_t}$

12: $\mathbf{X^{'}}=[\mathbf{X_s^{'}},\mathbf{X_t^{'}}]$

// Adaptive classification

13: Compute weight matrix $\mathbf{W}$ by $\mathbf{W_{ij}}=e^{-\frac{\|(x_i-x_j)^2\|}{\delta}}$ where $(x_i,x_j)\in X'$

14: Compute diagonal matrix $\mathbf{D}$ by $\mathbf{D_{ii}}=\sum_{j=1}^{n_s+n_t}\mathbf{W_{ij}}$.

15: $\mathbf{L}\leftarrow\mathbf{I}-\mathbf{D}^{\frac{-1}{2}}\mathbf{W}\mathbf{D}^{\frac{1}{2}}$

16: Compute diagonal matrix R where $R_{ii}=\begin{cases}1 & x_i\in X_s'\\ 0 & otherwise\end{cases}$

17: $\alpha\leftarrow\left(\sigma\mathbf{I}+\left(\mathbf{R}+\gamma L\right)X\right)^{-1}\mathbf{R}Y'^T$

18: Learn an adaptive classifier f.

19: Return target domain labels Y_t using classifier f.

شکل ۲: الگوریتم SADA

در الگوریتم SADA، هدف به‌دست‌آوردن ماتریس انتخاب $\mathbf{A}$ و ماتریس نگاشت $\mathbf{U}$ و درنهایت پیش‌بینی برچسب‌های دامنه هدف با استفاده از طبقه‌بند انطباقی است. SADA در ابتدا با استفاده از الگوریتم ۱ ماتریس انتخاب $\mathbf{A}$ را به‌د ست آورده و سپس با ا ستفاده از الگوریتم ۲ یک مقدار اولیه تصادفی به ماتریس نگاشت $\mathbf{U}$ می‌دهد و با استفاده از رابطه ۱۳ مقادیر دقیق ماتریس نگاشت $\mathbf{U}$ را به‌دست می‌آورد. با به‌دست‌آمدن مقادیر A و U در یک فرایند تکرارشونده، داده‌های ورودی به‌صورت $\mathbf{X_s^{'}}=\mathbf{U_s^T X_s A}$ و $\mathbf{X_t^{'}}=\mathbf{U_t^T X_t}$ به زیرفضای جدید نگاشت می‌شوند. در زیرفضای به‌دست‌آمده $\mathbf{X_s^{'}}$ و $\mathbf{X_t^{'}}$ دارای

حداقل اختلاف توزیع می‌باشند. در مرحله بعد طبقه‌بند انطباقی بر روی داده‌های نگاشت‌شده $\mathbf{x}_s'$ و $\mathbf{x}_t'$ اعمال می‌شود تا مدل مقاوم‌تری در برابر تغییرات داده‌ای ایجاد شود.

۵-آزمایش‌ها

در این بخش، روش پیشنهادی از جنبه‌های مختلف بررسی شده و تأثیر پارامترهای مختلف مدل بر روی کارایی آن بیان شده است. سپس، ضرورت هر پارامتر در الگوریتم نشان داده شده و عملکرد SADA با دیگر الگوریتم‌ها، مورد مقایسه و ارزیابی قرار گرفته است و نتایج به دست آمده به تفصیل مورد بحث قرار گرفته است.

۵-۱- تنظیمات آزمایش و مجموعه داده‌ها

۵-۱-۱- معرفی مجموعه داده‌ها

پایگاه‌داده‌های مورد استفاده در این مقاله عبارتند از: (۱) آفیس[۲۶] و کالتک[۲۷]، (۲) اعداد (USPS و MNIST)، (۳) چهره (پای[۲۸]).

مجموعه داده آفیس [۱۳] و کالتک [۱۴]، از مجموعه پایگاه‌داده‌های شناخته‌شده برای مسئله تطبیق دامنه‌های بصری است که شامل چهار دامنه آمازون[۲۹]، وبکم[۳۰]، DSLR[۳۱] و کالتک است. تصاویر موجود در دامنه آمازون از وب‌سایت‌های تجاری آنلاین دانلود شده است، تصاویر دامنه وبکم توسط دوربین‌های کم کیفیت وب گرفته شده است و تصاویر دامنه DSLR توسط دوربین‌های با کیفیت SLR دیجیتال گردآوری شده‌اند. تصاویر موجود در دامنه کالتک از وب‌سایت گوگل جمع‌آوری شده‌اند و تنها از ۱۰ کلاس مشترک بین دامنه‌ها استفاده شده است. تمام داده‌های منبع در الگوریتم استفاده شده‌اند و ۱۰ درصد از داده‌های هدف به‌صورت تصادفی به‌عنوان داده برچسب‌دار آموزشی استفاده شده است و بقیه داده‌های هدف به‌عنوان داده تست مورد استفاده قرار گرفته‌اند. بر روی این مجموعه داده، ۱۲ آزمایش طراحی شده است که در هر یک از این آزمایش‌ها یکی از مجموعه داده‌ها (به‌عنوان نمونه آمازون)، به‌عنوان دامنه منبع و یکی از سه مجموعه داده دیگر به‌عنوان دامنه هدف انتخاب می‌شوند.

مجموعه داده‌های USPS [۱۵] و MNIST [۱۶]، شامل اعداد دست‌نویس ۰ تا ۹ می‌باشند که USPS شامل تقریباً ۹۰۰۰ تصویر با اندازه ۱۶×۱۶ و MNIST شامل در حدود ۷۰۰۰۰ تصویر با اندازه ۲۸×۲۸ می‌باشد. تمام تصاویر به‌اندازه ۱۶×۱۶ تبدیل شده و دارای ۲۵۶ بعد می‌باشند. همه داده‌های منبع در الگوریتم استفاده شده‌اند و تنها ۱۰ درصد از داده‌های هدف به‌صورت تصادفی به‌عنوان داده برچسب‌دار آموزشی استفاده شده و بقیه داده‌های هدف به‌عنوان داده تست در نظر گرفته شده‌اند. این پایگاه داده شامل ۱۰ کلاس مختلف می‌باشد و به منظور آزمایش هر دو دامنه در شرایط یکسان، دامنه USPS_vs_MNIST(U_M) ایجاد شده است که به‌طور تصادفی ۱۸۰۰ نمونه از داده‌های دامنه USPS به‌عنوان داده‌های آموزشی و ۲۰۰۰ نمونه از داده‌های دامنه MNIST به‌عنوان داده‌های تست به‌کارگرفته می‌شود. با جابه‌جایی نمونه‌های آموزشی و تست دامنه MNIST_vs_USPS(M_U) برای یک آزمایش دیگر ایجاد شده است.

پایگاه داده پای [۱۷]، پایگاه‌داده معیار برای تشخیص چهره است که شامل ۴۱۳۶۸ تصویر چهره از ۶۸ فرد با اندازه تصاویر ۳۲×۳۲ است. این پایگاه داده شامل ۵ دامنه مختلف است که عبارتند از: پای ۱ (حالت چپ)، پای ۲ (حالت بالا)، پای ۳ (حالت پایین)، پای ۴ (حالت روبرو)، پای ۵ (حالت راست). در مجموع ۲۰ آزمایش بین دامنه‌ای بر روی پایگاه‌داده پای قابل طراحی است که از میان ۵ دامنه، دو دامنه متفاوت به‌عنوان دامنه‌های منبع و هدف انتخاب می‌شوند و کارایی الگوریتم پیشنهادی در شرایط متفاوت مورد بررسی قرار می‌گیرد. اطلاعات جزئی‌تر از پایگاه‌داده‌ها در جدول ۱ نشان داده شده است.

جدول ۱: معرفی پایگاه‌داده‌ها

پایگاه داده	تعداد نمونه‌ها	تعداد ابعاد	تعداد کلاس‌ها	اختصار
USPS	۱۸۰۰	۲۵۶	۱۰	U
MNIST	۲۰۰۰	۲۵۶	۱۰	M
Amazon	۹۵۸	۸۰۰	۱۰	A
Webcam	۲۹۵	۸۰۰	۱۰	W
Dslr	۱۵۷	۸۰۰	۱۰	D
Caltech256	۱۱۲۳	۸۰۰	۱۰	C
PIE1	۳۳۳۲	۱۰۲۴	۶۸	P1
PIE2	۱۶۲۹	۱۰۲۴	۶۸	P2
PIE3	۱۶۳۲	۱۰۲۴	۶۸	P3
PIE4	۳۳۲۹	۱۰۲۴	۶۸	P4
PIE5	۱۶۳۲	۱۰۲۴	۶۸	P5

۵-۱-۲- تنظیمات آزمایش

در تمام آزمایش‌ها، از طبقه‌بند NN به‌عنوان طبقه‌بند پایه استفاده شده است. τ ضریب MMD می‌باشد که به صورت تجربی برای پایگاه‌داده آفیس و کالتک ۱، اعداد ۵، پای ۰/۰۰۵ و پارامتر η، ۰/۰۰۱ در نظر گرفته شده است. پارامتر تنظیم تنکی ρ، پارامتر تأثیرگذار در انتخاب نمونه‌های دامنه منبع می‌باشد که برای همه پایگاه‌داده‌ها با ۰/۰۰۱ تنظیم شده است. به‌علت اینکه در محدوده وسیعی از مقادیر پارامتر تنظیم گراف μ ثابت می‌باشد، بنابراین برای همه پایگاه‌داده‌ها، ۰/۰۰۱ فرض شده است. همچنین، پیاده‌سازی روش پیشنهادی SADA، توسط نرم‌افزار متلب[۳۲] انجام گرفته است.

۵-۲- تأثیر پارامترها

برای به‌دست‌آوردن تنظیمات مدل، روش SADA با مقادیر مختلف پارامترها مورد ارزیابی قرار گرفته است. سه پارامتر مختلف در این روش

وجود دارد که مقادیر بهینه آن‌ها برای پایگاه‌داده‌های مختلف در جدول ۲ نشان داده شده است.

جدول ۲: مقادیر بهینه پارامترها برای سه پایگاه داده بصری (آفیس و کالتک، اعداد و پای)، P: ابعاد فضای جدید، δ: پارامتر تنظیم در طبقه‌بند انطباقی، γ: پارامتر تنظیم در طبقه‌بند انطباقی.

پارامترهای بهینه	P	δ	γ
آفیس و کالتک	۶۰	۰/۱	۰/۰۰۱
اعداد	۲۲۰	۰/۰۱	۰/۱
پای	۱۶۰	۰/۰۰۵	۰/۰۰۵

پارامترهای γ و δ در محدوده [۰/۱، ۰/۰۰۱] و پارامتر P در محدوده [۲۰، ۲۲۰] در نظر گرفته شده‌اند. نتایج به‌دست‌آمده از روش SADA با مقادیر مختلف پارامتر P با روش‌های JDA، TJM و VDA در پایگاه داده‌های آفیس و کالتک و اعداد مقایسه شده است.

نتایج به‌دست‌آمده برای مقادیر مختلف پارامتر δ بر روی پایگاه‌داده‌های آفیس و کالتک، اعداد و پای در شکل ۳ نشان می‌دهد که برای پایگاه‌داده آفیس و کالتک، مقادیر بالای δ دقت مدل را کاهش می‌یابد. مقادیر بالای پارامتر δ پیچیدگی مدل را افزایش داده و باعث می‌شود تأثیر سایر عوامل در ایجاد طبقه‌بند انطباقی نادیده گرفته شود. در چنین حالتی از ساختار اصلی داده‌ها، در ایجاد تطبیق بین دامنه‌های منبع و هدف استفاده نشده است. محدوده بهینه در پایگاه داده آفیس و کالتک برای پارامتر δ، [۰/۱، ۰/۰۵] است. پایگاه‌داده اعداد حساسیت چندانی به مقدار پارامتر δ ندارد و در هر دو آزمایش تغییرات زیادی با مقادیر مختلف متغیر δ بر روی نتایج ایجاد نمی‌شود. بااین‌حال برای مقادیر کمتر از ۰/۱ نتایج مطلوب‌تری به‌دست می‌آید. پایگاه‌داده پای نسبت به پارامتر δ حساسیت زیادی دارد و در محدوده بالا مدل صحت کاهش قابل‌ملاحظه‌ای پیدا می‌کند، محدوده بهینه پایگاه‌داده پای برای پارامتر δ، [۰/۰۱، ۰/۰۰۵] است.

نتایج به‌دست‌آمده برای مقادیر مختلف پارامتر γ بر روی پایگاه‌داده‌های آفیس و کالتک، اعداد و پای در شکل ۴ نشان می‌دهد که برای پایگاه‌داده آفیس و کالتک، مقادیر بالای γ دقت مدل را کاهش می‌دهد. درواقع مقادیر بالای پارامتر γ، باعث نادیده‌گرفتن اطلاعات داده‌های برچسب‌دار دامنه منبع در ایجاد طبقه‌بند انطباقی می‌شود. محدوده بهینه در پایگاه داده آفیس و کالتک برای پارامتر γ، [۰/۰۱، ۰/۰۰۱] است. پایگاه‌داده اعداد نسبت به مقادیر مختلف پارامتر γ حساسیت کمی دارد و در مقادیر بالا صحت مدل افزایش می‌یابد. پایگاه‌داده پای نسبت به پارامتر γ نیز حساسیت زیادی دارد و در محدوده بالا صحت کاهش قابل‌ملاحظه‌ای پیدا می‌کند. بدین ترتیب محدوده بهینه پایگاه‌داده پای برای پارامتر γ، [۰/۱، ۰/۰۱] است.

نتایج به‌دست‌آمده بر روی پایگاه‌داده آفیس و کالتک در شکل ۵ حاکی از آن است که پایگاه‌داده‌های مختلف در آفیس و کالتک دارای رفتار متفاوتی نسبت به مقادیر مختلف پارامتر P می‌باشند. دقت به‌دست‌آمده برای پارامتر P در هر روش، نشان‌دهنده صحت روش برای بازسازی داده‌ها در فضای جدید می‌باشد. به‌عبارتی، نشان‌دهنده صحت انتقال دانش از دامنه منبع به دامنه هدف در فضای جدید می‌باشد. برای به‌دست‌آوردن مقدار محدوده بهینه پارامتر، محدوده‌ای انتخاب می‌شود که دارای بهینه‌ترین صحت بر روی تعداد بیشتری از پایگاه داده‌ها باشد. برای بیشتر پایگاه داده‌ها محدوده بهینه پارامتر P در پایگاه داده آفیس و کالتک، [۴۰، ۱۶۰] می‌باشد.

نتایج به‌دست‌آمده برای مقادیر مختلف P در پایگاه داده اعداد در شکل ۶ نشان می‌دهد که پایگاه‌داده اعداد دارای حساسیت کمی نسبت به مقادیر مختلف P می‌باشد. بهترین محدوده برای پارامتر P در پایگاه داده اعداد، [۱۴۰، ۲۲۰] می‌باشد.

ρ پارامتر تنظیم تنکی است که یکی از پارامترهای تأثیرگذار در SADA می‌باشد. ρ تنکی ضرایب ماتریس A را کنترل می‌کند و بنابراین تعداد نمونه های منبع مرتبط را انتخاب می‌کند. صحت طبقه‌بندی و تعداد نمونه‌های انتخاب‌شده با توجه به ρ در جدول ۳ گزارش شده است. به‌عنوان نمونه در پایگاه‌داده U-M هنگامی‌که ρ=۰/۵ است هیچ نمونه‌ای از دامنه منبع انتخاب نشده است؛ اما زمانی که ρ کوچک است، ماتریس A تنکی کمتری دارد و درنتیجه نمونه‌های منبعی که توزیع نزدیک‌تری با نمونه‌های هدف دارند انتخاب می‌شوند. به‌عنوان نمونه در پایگاه‌داده U-M هنگامی‌که ρ = ۰/۰۰۱ است، تمام نمونه‌های منبع انتخاب شده‌اند. در تمام آزمایش‌ها مقدار پارامتر ρ برابر با ۰/۰۱ تنظیم شده است.

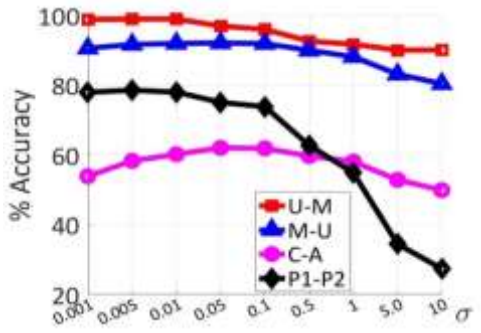
در شکل ۷ تعدادی از نمونه‌های انتخاب‌شده و انتخاب‌نشده از دامنه کالتک را نشان می‌دهد زمانی که دامنه آمازون به‌عنوان دامنه هدف در نظر گرفته شده است. به‌عنوان مثال، کوله‌پشتی از دامنه کالتک که شباهت بیشتری به کوله‌پشتی از دامنه آمازون دارند انتخاب شده و آن‌هایی که هیچ‌گونه شباهتی ندارند انتخاب نشده‌اند.

ρ	۰/۰۰۱	۰/۰۰۵	۰/۰۱	۰/۰۵	۰/۱	۰/۵
U-M	۹۸/۹۲	۹۸/۷۰	۹۸/۳۹	۹۷/۲۲	۹۷/۱۴	۹۷/۷۰
U-M #	۱۸۰۰	۱۸۰۰	۱۸۰۰	۷۸۲	۴۰۹	۰
C-A	۶۱/۸۹	۶۱/۹۹	۶۱/۳۳	۵۵/۶۹	۵۸/۰۷	۵۸/۷۴
C-A #	۱۰۶۸	۶۴۶	۴۹۴	۷۳	۲	۰

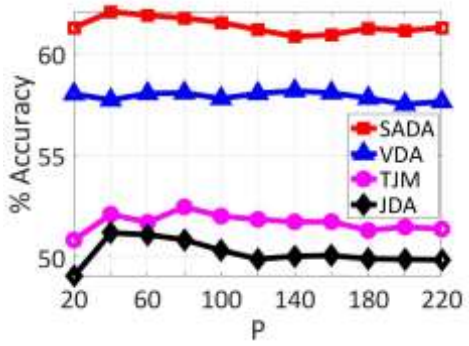
جدول ۳: تأثیر پارامتر تنظیم تنکی ρ بر تعداد نمونه‌های انتخاب شده از دامنه منبع

۵-۳- ضرورت هر پارامتر در الگوریتم

در این بخش، اهمیت وجودی برخی از پارامترها با طراحی دو آزمایش بر روی پایگاه‌داده آفیس و کالتک بررسی شده و نتایج آن در جدول ۴ نشان داده شده است.



شکل ۳: ارزیابی صحت پایگاه‌داده‌های U-M، M-U، C-A، P1-P2 با مقادیر مختلف پارامتر σ



شکل ۵: ارزیابی دقت روش‌های TJM، JDA، VDA و SADA در پایگاه‌داده C-A با مقادیر مختلف پارامتر P

۱. اســـتفاده از همه نمونه‌های منبع (Orj): برای آموزش مدل به جای انتخاب برخی از نمونه های منبع از همه نمونه‌های دامنه منبع استفاده شده است.

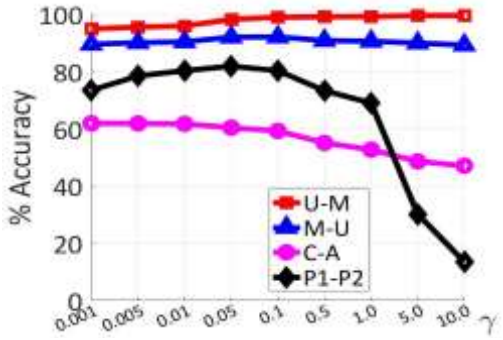
۲. استفاده از نرم l_1 به‌جای نرم $l_{2,1}$: با استفاده از نرم l_1 به‌جای نرم $l_{2,1}$ تعداد ضـرایب غیر صـفر در ماتریس $\mathbf{A}$ افزایش یافته و دقیقاً مشـــخص نمی‌شـــود که نمونه منبع دقیقاً با کدام نمونه هدف مرتبط است.

۳. حذف پارامتر تنظیم گراف ($SADA_{\mu^-}$): با حذف این پارامتر اهمیت نگهداری ساختار هندسی هر دامنه مشخص می‌شود.

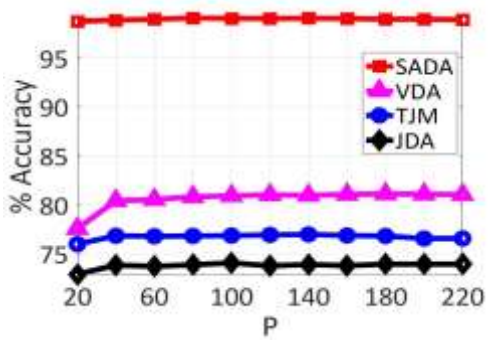
۴. اســـتفاده از طبقه‌بند SVM به‌جای طبقه‌بند انطباقی: با استفاده از یک طبقه‌بند استاندارد SVM، اهمیت استفاده از طبقه‌بند انطباقی مشخص می‌شود.

با بررســی نتایج حاصــل‌شــده از آزمایش‌های متعدد در جدول ۴ می‌توان نتیجه گرفت که اســتفاده از تمام نمونه‌های منبع در ســاخت مدل باعث کاهش ملموس کارایی مدل می‌شـــود. از این رو SADA با اســـتفاده از کدگذاری تنک و نرم $l_{2,1}$ اقدام به حذف نمونه های غیرمرتبط بین دو دامنه می‌نماید.

با بررسی تأثیر استفاده از نرم‌های مختلف می‌توان نتیجه گرفت که



شکل ۴: ارزیابی صحت پایگاه‌داده‌های U-M، M-U، C -A، P1-P2 با مقادیر مختلف پارامتر γ



شکل ۶: ارزیابی دقت روش‌های TJM، JDA، VDA و SADA در پایگاه‌داده U-M با مقادیر مختلف پارامتر P

نرم l_1 کارایی قابل‌توجهی نداشـــته و اســـتفاده از نرم $l_{2,1}$ باعث می‌شود که هر نمونه هدف با حداقل تعداد نمونه منبع به‌خوبی توصیف شود.

بررســی تأثیر پارامتر همســایگی نشــان‌دهنده این اســت که، حفظ توپولوژی و ســاختار داده‌ها در نگاشــت آن‌ها به زیرفضــای جدید تأثیر بسیار تعیین‌کننده‌ای در نتایج به‌دست‌آمده دارد. درواقع، در صورتی‌که داده‌ها بدون حفظ توپولوژی به فضــای جدید نگاشــت شــوند، ســاختار اصلی خود را از دست داده و مدل به‌دست‌آمده قابل اعتماد نمی‌باشد.

درنهایت تأثیر طبقه‌بند انطباقی بر روی مدل بررســی شــده اســت. نتایج به‌دست‌آمده، نشان می‌دهد که SADA با تطبیق پارامترهای مدل کارایی الگوریتم را به‌صـــورت قابل‌ملاحظه‌ای افزایش می‌دهد. درواقع، طبقه‌بند انطباقی مقاوم‌پذیری مدل را در را ستای تغییرات داده به‌اندازه قابل‌توجهی افزایش می‌دهد.

جدول ۴: ضرورت هر پارامتر در الگوریتم

S-T: دامنه منبع و هدف، Orj: همه نمونه‌های منبع،

$SADA_{\mu^-}$: الگوریتم با حذف پارامتر تنظیم گراف

S-T	Orj	l_1 norm	$SADA_{\mu^-}$	SVM	SADA
C-A	۶۱/۶۹	۶۱/۸۸	۵۵/۶۲	۵۷/۶	۶۱/۹۲
C-A #	۱۱۲۳	۱۰۸۷	۲۷۳	۱۰۶۹	۱۰۶۹
C-W	۶۲/۳۸	۶۲/۳۴	۴۴/۳۴	۴۸/۶	۶۳/۴۷
C-W#	۱۱۲۳	۹۷۹	۱۸۸	۸۸۰	۱۰۶۹

۵-۴- اهمیت استفاده از برچسب‌های دامنه هدف (یادگیری نیمه‌نظارت‌شده)

دقت الگوریتم پیشنهادی بر روی پایگاه‌داده آفیس و کالتک در دو حالت نیمه‌نظارت شده (SADA) و بدون نظارت ($SADA^-$)، در شکل ۸ نشان داده شده است. نتایج به‌دست‌آمده به‌وضوح نشان‌دهنده تأثیر قابل‌ملاحظه برچسب‌های دامنه هدف در افزایش دقت مدل دارد. درواقع، در یادگیری نیمه‌نظارت شده، داده‌های برچسب‌دار دامنه هدف به‌عنوان یک پل و رابط بین دو دامنه عمل می‌کنند. بدین ترتیب، در یافتن زیر فضای مشترک بین دامنه‌های منبع و هدف اطلاعات تفکیک‌کننده دو منبع نیز مورد استفاده قرار می‌گیرند.

۵-۵- مقایسه با الگوریتم‌های موجود

الگوریتم SADA با هفت الگوریتم زیر مورد مقایسه قرار گرفته است: طبقه‌بند نزدیک‌ترین همسایه (NN[۳۳])، PCA [۱۱]، TCA[۳۴] [۱]، GFK [۶]، JDA[۳۵] [۱۸]، TJM [۹] و VDA [۱۲]. به‌دلیل این که تمامی روش‌های نام‌برده‌شده (به‌جز NN)، روش‌های کاهش بعد می‌باشند، از طبقه‌بند استاندارد نزدیک‌ترین همسایه [۱۹] برای آموزش یک طبقه‌بند بر روی داده‌های دامنه منبع جهت پیش‌بینی برچسب داده‌های دامنه هدف استفاده می‌شود.

صحت طبقه‌بندی SADA و هفت روش دیگر بر روی پایگاه‌داده‌های آفیس و کالتک، اعداد و پای در جدول ۵ نشان داده شده است. همان طور که مشاهده می شود SADA کارایی بهتری نسبت به هفت روش دیگر از خود نشان داده و در بیشتر حالات عملکرد مطلوبی ارائه می دهد. نتایج گزارش‌شده شامل میانگین و انحراف از استاندارد برای همه روش‌ها می‌باشد.

نتایج به‌دست‌آمده حاکی از میانگین صحت ۷۵/۷۵٪ SADA روی ۳۴ پایگاه داده و بهبود صحت ۴/۵۶٪ نسبت به بهترین الگوریتم مورد مقایسه VDA و ۲۳/۲۹٪ نسبت به الگوریتم استاندارد NN می‌باشد و در ۲۶ آزمایش از ۳۴ آزمایش SADA عملکرد بهتری نسبت به روش‌های دیگر ارائه نموده است. در ادامه عملکرد SADA در مقایسه با هر یک از روش‌های موجود مورد بحث قرار گرفته است.

PCA یکی از روش‌های کاهش بعد است که برای کاهش ابعاد داده‌ها مورد استفاده قرار می‌گیرد. PCA اختلاف توزیع بین دامنه‌ها را چندان کاهش نمی‌دهد و عملکرد ضعیفی نسبت به دیگر الگوریتم‌های تطبیق دامنه نشان می‌دهد. نتایج حاصل از آزمایش‌ها حاکی از این است که SADA به تفکیک نسبت به PCA دارای متوسط بهبود صحت ۹/۳۹٪ در پایگاه‌داده آفیس و کالتک، ۹/۴٪ در پایگاه‌داده اعداد و ۳۹/۳۹٪ در پایگاه‌داده پای است و به‌طور کلی دارای ۲۸/۰۴٪ بهبود صحت است. عمده‌ترین دلیل برتری SADA نسبت به PCA استفاده از تطبیق خصوصیات برای کاهش اختلاف توزیع حاشیه‌ای بین دامنه‌ها می‌باشد.

TCA یک روش تطبیق دامنه بدون نظارت است و از برچسب‌های دامنه منبع برای تطبیق دامنه‌ها استفاده نمی‌کند و تنها اختلاف توزیع حاشیه‌ای بین دامنه‌ها را کاهش می‌دهد و برای حفظ ویژگی‌های اصلی داده‌های ورودی، واریانس داده‌ها را حداکثر می‌سازد. نتایج حاصل از آزمایش‌ها حاکی از این است که SADA به تفکیک نسبت به TCA دارای متوسط بهبود صحت ۴/۶۴٪ در پایگاه‌داده آفیس و کالتک، ۱۲/۱۲٪ در پایگاه‌داده اعداد و ۶۰/۹۳٪ در پایگاه‌داده پای است و به‌طورکلی دارای ۳۹/۱۹٪ بهبود است. دلیل بهبود قابل‌ملاحظه SADA نسبت به TCA این است که SADA علاوه بر استفاده از برچسب‌های دامنه منبع، از بخش کوچکی از برچسب‌های دامنه هدف نیز برای تطبیق دامنه‌ها استفاده می‌کند و با استفاده از یک طبقه‌بند انطباقی خطای پیش‌بینی مدل روی داده‌های منبع را حداقل می‌سازد.

در روش GFK به‌منظور کاهش اختلاف توزیع حاشیه‌ای، داده‌های منبع و هدف به یک زیرفضای کوچک‌تری در منیفلد گرسمن نگاشت می‌شوند، اما به‌دلیل این‌که زیر فضاهای نگاشت شده دارای ابعاد بسیار کمی می‌باشد باعث از بین‌رفتن بخشی از داده‌ها شده و نمایش خوبی از داده‌ها ایجاد نمی شود. نتایج حاصل از آزمایش‌ها حاکی از این است که متوسط بهبود صحت روش SADA به تفکیک نسبت به روش GFK در پایگاه‌داده آفیس و کالتک ۶/۲۳٪، در پایگاه‌داده اعداد ۸/۸٪ و در پایگاه‌داده پای ۳۷/۵۱٪ است و در مجموع دارای ۲۵/۷۸٪ بهبود است. دلیل برتری SADA نسبت به GFK این است که در روش SADA یک زیر فضای مشترک با حداکثر واریانس ایجاد می‌شود که ساختار اصلی داده‌ها را حفظ می‌کند و با استفاده از تطبیق خصوصیات، اختلاف توزیع حاشیه‌ای کاهش می‌یابد.

روش‌های TJM و JDA ازجمله جدیدترین روش‌های تطبیق دامنه می‌باشند. روش TJM یک روش ترکیبی از روش‌های مبتنی بر نمونه و مبتنی بر خصوصیت است که تنها اختلاف توزیع حاشیه‌ای را کاهش می‌دهد. بااین‌حال TJM به‌دلیل اینکه بخشی از داده ورودی را حذف می‌کند، ساختار اصلی داده‌ها را حفظ نمی‌کند. درحالی‌که روش JDA، یک روش بدون نظارت می‌باشد که اختلاف توزیع حاشیه‌ای و شرطی بین دامنه‌های منبع و هدف را به‌طور هم‌زمان کاهش می‌دهد و دارای عملکرد بهتری نسبت به TJM می‌باشد.

متوسط بهبود صحت SADA به تفکیک نسبت به روش TJM در پایگاه‌داده آفیس و کالتک ۵/۲۳٪، در پایگاه‌داده اعداد ۱۶/۵۴٪ و در پایگاه‌داده پای ۵۹/۸۵٪ است. همچنین متوسط بهبود صحت روش SADA به تفکیک نسبت به روش JDA در پایگاه‌داده آفیس و کالتک ۴/۱۱٪، در پایگاه‌داده اعداد ۱۲/۱۰٪ و در پایگاه‌داده پای ۶/۰۶٪ است.

دلیل برتری SADA نسبت به روش‌های TJM و JDA این است که SADA ساختار داده‌ها را حفظ کرده و از یک روش نیمه‌نظارت‌شده جهت انطباق دامنه‌ها استفاده می‌کند. همچنین SADA از یک طبقه‌بند انطباقی برای کاهش خطای پیش‌بینی مدل بر روی داده‌ها استفاده می کند.

جدول ۵. صحت (%) طبقه‌بند بر روی ۳۴ پایگاه داده بصری

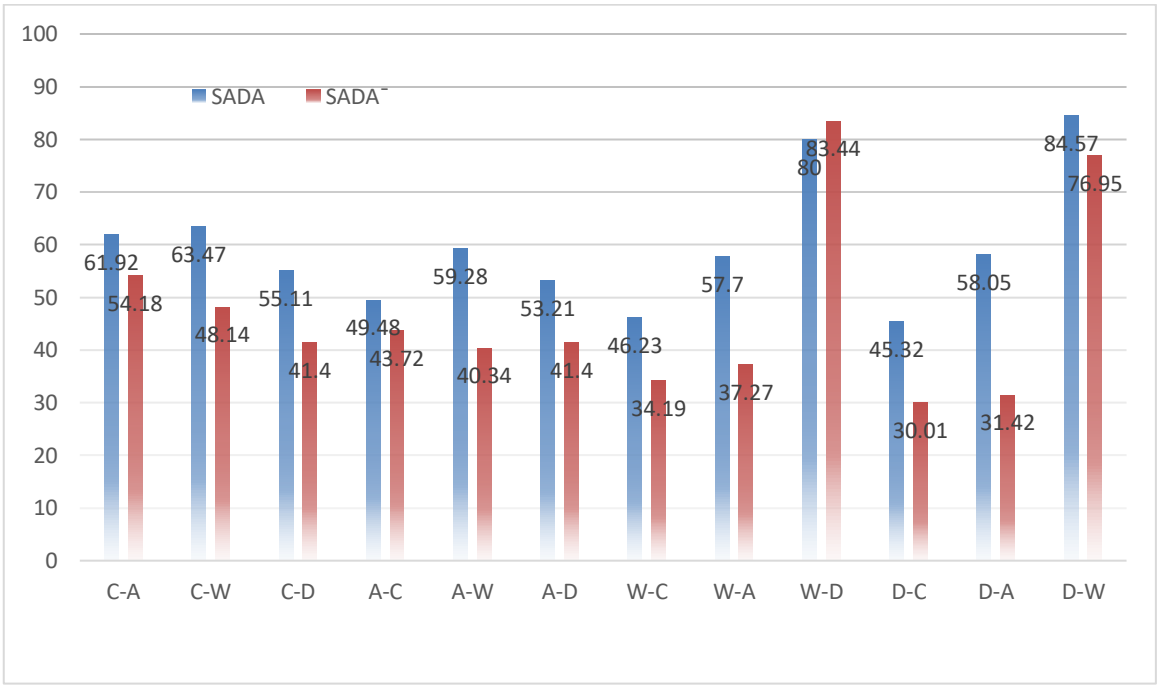
Dataset	NN	PCA	TCA	GFK	JDA	TJM	VDA	SADA
U_M	۸۲/۶۸±۰/۸۸	۸۳/۷۷±۱/۳۵	۸۰/۳۶±۱/۰۱	۸۴/۰۶±۰/۸۶	۶۷/۸۶±۱/۶۱	۷۵/۹۴±۱/۲۱	۷۳/۵۷±۱/۰۶	**۹۸/۹۲±۰/۳۰**
M_U	۸۹/۳۶±۰/۴۲	۸۸/۲۸±۰/۸۹	۸۶/۳۱±۰/۷۹	۸۹/۲۷±۰/۳۱	۷۴/۹۹±۴/۷۸	۸۱/۹۰±۲/۰۶	۸۰/۰۰±۱/۲۲	**۹۲/۰۰±۱/۷۳**
P1_P2	۴۷/۴۹±۱/۳۶	۳۰/۸۲±۱/۰۶	۱۲/۷۸±۰/۴۶	۳۴/۵۶±۱/۴۲	۷۴/۶۴±۱/۱۲	۱۳/۶۱±۰/۳۳	۷۳/۹۹±۲/۵۵	**۷۸/۵۱±۴/۰۱**
P1_P3	۴۹/۷۱±۲/۷۴	۳۳/۰۹±۰/۶۶	۱۴/۱۹±۰/۳۲	۳۶/۵۳±۰/۷۵	۷۳/۴۷±۱/۰۶	۱۴/۱۴±۰/۱۵	۷۲/۲۷±۲/۱۸	**۷۹/۲۱±۴/۷۶**
P1_P4	۵۹/۸۴±۱/۴۷	۴۲/۷۰±۰/۹۹	۱۹/۲۳±۰/۳۴	۴۴/۶۱±۱/۱۵	۸۷/۴۴±۱/۰۷	۲۲/۴۴±۰/۱۰	۸۵/۳۷±۰/۴۰	**۹۱/۱۹±۱/۲۹**
P1_P5	۴۹/۰۷±۱/۴۲	۳۲/۰۲±۱/۲۱	۱۴/۹۲±۰/۲۳	۳۴/۲۸±۱/۰۷	۶۹/۹۳±۱/۷۶	۱۳/۱۰±۰/۲۱	۶۸/۴۰±۱/۴۱	**۷۷/۰۹±۷/۸۰**
P2_P1	۶۲/۸۴±۱/۵۰	۴۴/۹۰±۱/۱۶	۱۸/۶۲±۰/۳۸	۴۵/۴۷±۰/۸۸	۷۴/۶۱±۵/۵۰	۲۳/۵۸±۰/۵۳	۸۰/۱۳±۰/۶۱	**۹۰/۲۷±۱/۱۶**
P2_P3	۵۹/۵۱±۲/۹۷	۴۲/۳۷±۱/۹۲	۲۴/۴۷±۰/۷۷	۴۴/۹۲±۰/۸۵	۷۵/۰۶±۱/۱۱	۲۵/۶۰±۰/۴۰	۷۲/۵۲±۸/۱۹	**۸۲/۳۵±۰/۹۳**
P2_P4	۷۲/۵۹±۰/۳۱	۵۸/۰۰±۰/۷۰	۳۲/۴۴±۰/۴۴	۵۸/۵۹±۰/۷۶	۸۳/۷۳±۱/۸۱	۳۱/۵۰±۰/۲۹	۸۶/۵۹±۰/۹۷	**۸۹/۴۳±۲/۲۶**
P2_P5	۵۱/۵۵±۱/۱۸	۳۴/۷۵±۱/۳۲	۱۴/۷۳±۰/۶۴	۳۵/۲۰±۱/۴۲	۶۶/۰۹±۲/۰۷	۱۶/۸۵±۰/۲۰	۶۸/۲۱±۰/۶۵	**۷۰/۵۴±۱/۰۰**
P3_P1	۶۲/۲۹±۱/۳۹	۴۴/۵۳±۱/۰۴	۱۷/۹۹±۰/۵۱	۴۴/۶۵±۱/۲۷	۷۳/۶۶±۷/۰۰	۲۲/۸۱±۰/۷۶	۷۷/۴۶±۱/۶۰	**۹۰/۱۸±۱/۲۰**
P3_P2	۵۵/۲۹±۱/۲۰	۳۹/۳۳±۰/۴۷	۲۲/۰۱±۰/۶۷	۴۲/۷۰±۰/۵۳	۷۱/۱۲±۱/۲۰	۲۹/۱۲±۰/۱۰	۷۲/۸۸±۹/۹۰	**۷۷/۳۰±۱/۳۳**
P3_P4	۷۱/۰۴±۰/۶۹	۵۷/۲۱±۰/۴۶	۳۱/۷۷±۰/۴۳	۵۷/۳۰±۰/۷۹	۸۲/۵۲±۱/۸۵	۳۳/۳۹±۰/۳۷	۸۴/۸۵±۱/۱۷	**۸۹/۶۸±۱/۵۳**
P3_P5	۵۱/۲۵±۳/۰۴	۳۴/۲۶±۱/۸۶	۱۴/۲۸±۰/۴۰	۳۶/۱۷±۲/۴۷	۶۵/۲۹±۲/۳۱	۱۵/۱۵±۰/۳۹	۶۹/۷۹±۲/۸۳	**۷۱/۹۷±۲/۷۴**
P4_P1	۶۳/۹۹±۱/۰۶	۴۴/۰۸±۱/۳۲	۱۸/۶۴±۰/۶۶	۴۶/۰۹±۰/۸۲	۸۶/۱۲±۱/۲۹	۳۰/۶۸±۰/۷۹	۸۵/۵۳±۰/۹۵	**۹۲/۸۹±۱/۲۱**
P4_P2	۷۳/۳۸±۰/۶۲	۵۶/۲۲±۰/۷۲	۳۶/۴۰±۰/۴۷	۵۹/۹۱±۰/۹۳	۸۶/۵۳±۰/۶۹	۲۹/۰۳±۰/۲۳	۸۲/۶۸±۷/۱۲	**۸۶/۹۰±۰/۷۲**
P4_P3	۸۱/۵۱±۱/۲۴	۶۴/۷۱±۱/۱۶	۴۲/۶۸±۰/۴۳	۶۷/۲۵±۰/۹۵	**۹۰/۱۵±۰/۳۳**	۳۹/۵۰±۱/۲۵	۸۸/۵۴±۱/۰۱	۸۵/۹۰±۴/۲۳
P4_P5	۵۵/۹۴±۲/۱۷	۳۶/۰۳±۱/۸۶	۱۶/۰۹±۰/۲۵	۳۹/۵۲±۱/۵۶	۷۱/۶۵±۳/۵۷	۱۵/۰۷±۰/۳۰	۷۳/۷۳±۱/۷۲	**۷۸/۶۸±۳/۲۰**
P5_P1	۶۱/۹۵±۱/۳۷	۴۳/۹۶±۱/۵۵	۱۸/۵۵±۰/۵۴	۴۴/۶۴±۱/۲۹	۷۵/۱۹±۹/۵۲	۱۹/۱۰±۰/۴۳	۷۸/۵۱±۱/۶۷	**۸۸/۱۳±۱/۵۳**
P5_P2	۴۷/۶۳±۱/۷۲	۳۰/۹۴±۰/۴۳	۱۲/۷۴±۰/۱۵	۳۲/۹۲±۰/۶۵	۶۴/۹۸±۶/۳۰	۱۲/۷۰±۰/۲۵	۷۰/۷۲±۱/۷۲	**۷۲/۹۴±۵/۱۶**
P5_P3	۵۰/۴۸±۴/۴۱	۳۴/۲۳±۰/۸۷	۱۴/۶۵±۰/۲۰	۳۶/۱۰±۰/۲۶	۶۷/۸۶±۱/۶۱	۱۳/۵۱±۰/۸۱	۷۱/۸۰±۳/۹۹	**۷۶/۲۰±۲/۱۷**
P5_P4	۶۰/۱۷±۰/۶۸	۴۴/۱۵±۰/۶۰	۲۰/۵۱±۰/۴۰	۴۴/۵۲±۰/۵۸	۷۴/۹۹±۴/۷۸	۱۸/۴۰±۰/۳۰	۷۸/۹۵±۰/۹۶	**۸۷/۶۶±۱/۹۵**
C_A	۲۸/۰۷±۱/۱۲	۴۷/۹۰±۲/۷۰	۵۰/۷۷±۱/۳۲	۴۷/۷۴±۱/۰۰	۵۰/۳۴±۱/۳۳	۵۰/۸۵±۱/۷۳	۵۳/۳۴±۳/۱۱	**۶۱/۹۲±۶/۵۷**
C_W	۲۹/۰۲±۱/۱۵	۴۶/۵۳±۹/۹۱	۵۱/۸۵±۴/۰۹	۵۴/۷۹±۷/۹۴	۵۳/۰۰±۱۰/۶۰	۵۵/۸۱±۷/۲۰	۵۷/۹۲±۷/۶	**۶۳/۴۷±۱۲/۹۴**
C_D	۲۷/۳۷±۰/۸۶	۴۲/۶۳±۶/۵۴	۵۱/۸۲±۲/۹۶	۴۹/۸۵±۲/۳۷	۵۰/۳۶±۴/۵۰	۴۹/۹۳±۴/۵۲	۵۱/۳۹±۱۰/۹۲	**۵۵/۱۱±۵/۰۰**
A_C	۲۸/۰۰±۱/۰۶	۳۹/۲۰±۰/۵۲	۴۲/۷۰±۱/۷۰	۴۲/۴۶±۰/۸۲	۴۲/۷۹±۱/۲۲	۴۲/۰۰±۰/۹۱	۴۴/۴۳±۰/۹۱	**۴۹/۴۸±۱/۶۷**
A_W	۳۴/۳۸±۴/۹۲	۴۶/۱۱±۹/۲۳	۵۰/۸۷±۷/۵۳	۵۲/۹۴±۹/۱۲	۵۲/۲۶±۶/۳۷	۵۳/۴۷±۴/۲۴	۵۹/۲۸±۸/۲۴	**۵۹/۴۳±۶/۴۹**
A_D	۲۷/۲±۵۲/۳۷	۳۷/۱۵±۶/۸۰	۴۲/۱۲±۵/۴۵	۴۴/۲۳±۴/۵۲	۴۷/۰۸±۴/۰۶	۴۹/۴۹±۲/۳۴	۵۱/۴۶±۵/۵۹	**۵۳/۲۱±۱/۰۰**
W_C	۲۵/۳±۳۵/۴۱	۳۷/۵۹±۲/۴۴	۴۰/۰۴±۲/۸۵	۳۸/۷۲±۱/۰۹	۴۰/۶۲±۱/۶۴	۳۸/۵۹±۳/۲۵	۴۱/۵۵±۲/۷	**۴۶/۲۳±۱/۳۱**
W_A	۳۶/۵۵±۴/۱۰	۴۸/۱۴±۴/۱۴	۵۲/۵۴±۳/۵۴	۴۸/۵۸±۴/۸۱	۵۴/۳۹±۶/۳۴	۴۸/۶۹±۳/۲۲	۵۶/۹۵±۲/۷۳	**۵۷/۷۰±۴/۸۵**
W_D	۶۰/۵۱±۲/۶۶	۸۶/۵۰±۱/۶۹	۹۲/۷۷±۱/۰۰	۸۸/۶۹±۷/۴۹	۹۰/۳۶±۳/۱۷	۸۸/۳۲±۳/۵۵	**۹۴/۸۲±۲/۸۹**	۸۰/۰۰±۸/۵۵
D_C	۲۵/۵۰±۵/۲۷	۳۷/۸۱±۱/۰۵	۳۹/۹۷±۱/۵۶	۳۷/۵۲±۱/۳۷	۳۹/۶۵±۳/۱۳	۳۷/۸۹±۱/۱۳	۴۰/۸۴±۴/۶۲	**۴۵/۳۲±۳/۷۸**
D_A	۳۷/۰۷±۷/۱۲	۴۷/۹۷±۶/۹۲	۵۲/۵۵±۷/۱۳	۴۶/۸۹±۷/۳۶	۵۲/۵۹±۸/۳۰	۴۸/۵۱±۳/۵۹	۵۷/۲۰±۱۰/۱۱	**۵۸/۰۵±۸/۱۳**
D_W	۶۴/۷۲±۹/۶۹	۸۴/۱۵±۰/۹۲	۹۰/۶۸±۰/۶۲	۸۷/۱۳±۷/۹۳	۹۱/۴۳±۰/۹۸	۸۸/۰۴±۵/۰۳	**۹۳/۹۶±۰/۵۱**	۸۴/۵۷±۲/۲۳
متوسط صحت	۵۲/۴۶±۲/۲۸	۴۷/۷۱±۲/۲۵	۳۶/۵۶±۱/۴۸	۴۹/۹۷±۲/۲۸	۶۸/۹۳±۳/۲۰	۳۶/۷۳±۱/۵۳	۷۱/۱۹±۲/۵۵	**۷۵/۷۵±۴/۰۷**

روش VDA علاوه‌بر تطبیق توزیع‌های حاشیه‌ای و شرطی از یک خوشه‌بندی مستقل از دامنه برای بهبود صحت طبقه‌بند استفاده می کند. بااین‌حال، به‌دلیل ویژگی‌های متفاوت داده‌های آموزشی و تست در دامنه‌های منبع و هدف، صحت بالایی در پیش‌بینی داده‌های دامنه هدف را دارا نمی‌باشد. متوسط بهبود صحت روش SADA به تفکیک نسبت به روش VDA در پایگاه‌داده اعداد ۱۸/۶۷ ٪ و در پایگاه‌داده پای ۳/۶۳٪ است. درواقع SADA با استفاده از یک طبقه‌بند انطباقی خطای پیش‌بینی مدل روی داده‌های منبع را حداقل ساخته و سازگاری بین ساختار منیفلد و مدل پیش‌بینی را حداکثر می‌سازد.

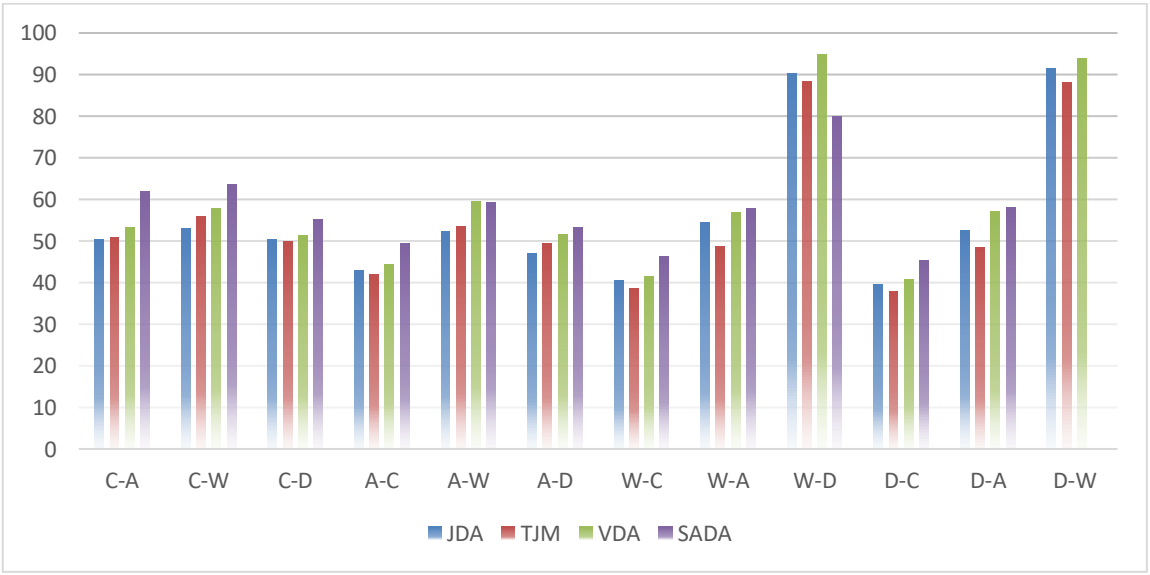
نمودارهای مربوط به مقایسه صحت طبقه بند بر روی ۳۴ پایگاه داده با روش‌های JDA، TJM، VDA و SADA در شکل‌های ۹ و ۱۰ نشان داده شده است. به‌دلیل اینکه الگوریتم SADA، یک الگوریتم تکرار شونده برای تخمین دقیق برچسب داده‌های دامنه هدف است از ۱۰ تکرار برای به‌دست‌آوردن نتایج این الگوریتم استفاده شده است.

نمونه‌ها از دامنه هدف آمازون | نمونه‌های انتخاب‌شده از دامنه منبع کالتک | نمونه‌های انتخاب‌نشده از دامنه منبع کالتک

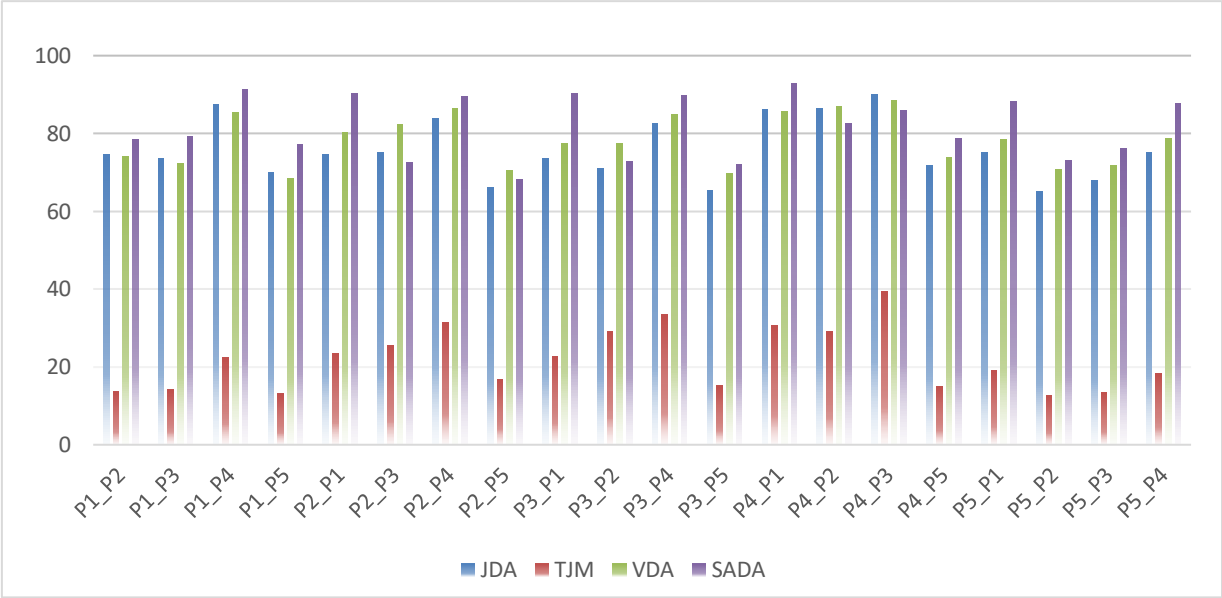
شکل ۷: نمونه‌های انتخاب‌شده از دامنه منبع کالتک برای دامنه هدف آمازون



شکل ۸: مقایسه پایگاه‌داده آفیس و کالتک در دو حالت نیمه نظارت‌شده (SADA) و بدون نظارت ($SADA^{-}$)



شکل ۹: صحت طبقه‌بندی در پایگاه داده‌های آفیس و کالتک و اعداد با استفاده از روش‌های JDA، TJM، VDA و FMM (نمایش بهتر به‌صورت رنگی)



شکل ۱۰: صحت طبقه‌بندی در پایگاه داده پای با استفاده از روش‌های JDA، TJM، VDA و FMM (نمایش بهتر به‌صورت رنگی)

۶- نتیجه‌گیری

در این مقاله، یک روش نوین جهت حل مسئله شیفت در دامنه‌ها، با عنوان کدگذاری تنک و طبقه‌بندی انطباقی برای تطبیق دامنه‌های بصری (SADA) پیشنهاد شد. SADA یک روش نیمه‌نظارت‌شده می‌باشد که در دو مرحله تلاش می‌کند اختلاف توزیع بین داده‌های آموزشی و تست را کاهش دهد. درواقع SADA مدلی ایجاد می‌کند که با به حداقل‌رساندن ناسازگاری داده‌ها، مقاومت و تحمل‌پذیری بیشتری در برابر تغییرات توزیع داده‌ها بین دامنه‌های منبع و هدف داشته باشد. SADA در گام اول با استفاده از نرم $l_{2,1}$ نقش نمونه‌هایی از داده‌های منبع را که عامل ایجاد اختلاف بین دو دامنه هستند را کم‌رنگ‌تر می‌کند و با استفاده از MMD اختلاف توزیع حاشیه‌ای دو دامنه را به حداقل می‌رساند. علاوه‌بر این، SADA با تطبیق پارامترهای مدل به کمک یک طبقه‌بند انطباقی مقاومت و تحمل‌پذیری آن در مقابل تغییرات داده‌ها را به حداکثر می‌رساند.

برای ارزیابی عملکرد SADA، آزمایش‌های مختلفی بر روی انواع پایگاه‌داده‌های واقعی ترتیب داده شده است و نتایج ارزیابی‌ها نشان از برتری قابل‌ملاحظه SADA نسبت به دیگر الگوریتم‌های تطبیق دامنه دارد. برای ادامه کار، برنامه‌ریزی جهت گسترش SADA برای کاهش هم‌زمان توزیع اختلاف‌های حاشیه‌ای و شرطی انجام شده است.

مراجع

[1] S. J. Pan, I. W. Tsang, J. T. Kwok and Q. Yang, "*Domain adaptation via transfer component analysis*", IEEE Trans. Neural Netw, vol. 22, no. 2, pp. 199–210, 2011.

[2] J. Tahmoresnezhad and S. Hashemi, "*A generalized kernel-based random k-sample sets method for transfer learning*", Iran J Sci Technol Trans Electrical Eng, vol. 39, pp. 193-207, 2015.

[3] B. Okutmuştur, "*Reproducing kernel Hilbert spaces*", 2005.

[4] X. Li, M. Fang, J. J. Zhang and J. Wu, "*Sample selection for visual domain adaptation via sparse coding*", Signal Processing: Image Communication, vol 44, pp. 92-100, 2016.

[5] طاهره زارع بیدکی و محمدتقی صادقی، «بهینه‌سازی وزن‌ها در کرنل مرکب برای طبقه‌بند مبتنی بر نمایش تنک کرنلی»، مجله مهندسی برق دانشگاه تبریز، جلد ۴۷، شماره ۳، صفحات ۱۰۵۹-۱۰۷۲، ۱۳۹۶.

[6] B. Gong, Y. Shi, F. Sha and K. Grauman, "*Geodesic flow kernel for unsupervised domain adaptation*", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2066-2073, 2012.

[7] L. Bruzzone and M. Marconcini, "*Domain adaptation problems: a DASVM classification technique and a circular validation strategy*", IEEE Trans Pattern Anal Mach Intell, vol. 32, no. 5, pp. 770–787, 2010.

[8] B. Gong, K. Grauman and F. Sha, "*Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation"*, Proceedings of the International Conference on Machine Learning, vol. 28, no. 1, pp.222-230, 2013.

[9] M. Long, J. Wang, G. Ding, J. Sun and P. S. Yu, "*Transfer joint matching for unsupervised domain adaptation*", IEEE conference on computer vision and pattern recognition, pp. 1410-1417, 2014.

[10] J. Tahmoresnezhad and S. Hashemi, "*Visual domain adaptation via transfer feature learning*", KnowlInf Syst, vol. 50, no. 2, pp. 585-605, 2016.

[11] M. Long, J. Wang, G. Ding, S. J. Pan and P. Yu, "*Adaptation regularization: a general framework for transfer learning*", IEEE Trans. Knowl. Data Eng, vol. 26, pp. 1076–1089, 2013.

[12] Jolliffe I, Principal component analysis, Wiley, vol. 2, pp. 433-459, 2002.

[13] K. Saenko, B. Kulis, M. Fritz and T. Darrell, "*Adapting visual category models to new domains*", Proceedings of the European Conference on Computer Vision, pp. 213-226, 2010.

[14] G.Griffin, A. Holub and P. Perona, "*Caltech-256 object category dataset*", Technical Report7694, 2007.

[15] J. J. Hull, "*A database for handwritten text recognition research*", IEEE Trans. Pattern Anal. Mach. Intell, vol. 16, no. 5, pp. 550–554, 1994.
[16] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, "*Gradient-based learning applied to document recognition*", Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.
[17] T. Sim, S. Baker and M. Bsat, "*The CMU pose, illumination, and expression (PIE) database*", Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition, pp. 53-58, 2002.
[18] M. Long, J. Wang, G. Ding, J. Sun and S. YuPhilip, "*Transfer feature learning with joint distribution adaptation*", IEEE international conference on computer vision, pp. 2200-2207, 2013.
[19] مهرداد حیدری ارجلو، سید قدرت اله سیف السادات و مرتضی رزاز، «*یک روش هوشمند تشخیص جزیره در شبکه توزیع دارای تولیدات پراکنده مبتنی بر تبدیل موجک و نزدیک‌ترین k-همسایگی* (kNN) »، مجله مهندسی برق دانشگاه تبریز، جلد ۴۳، شماره ۱، صفحات ۱۵-۲۶، ۱۳۹۲.

زیرنویس‌ها

[1] Domains shift
[2] Transfer learning
[3] Domains adaptation
[4] Maximum mean discrepancy (MMD)
[5] Reproducing kernel Hilbert space
[6] Element mean
[7] Sparse coding
[8] Unsupervised
[9] image processing via Sparse coding and ADAptive classification
[10] Adaptive classifer
[11] $l_{2,1}$ norm
[12] Transfer Joint Matching
[13] Sample selection sparse coding
[14] Geodesic flow kernel
[15] Grassman manifold
[16] Geodesic
[17] Visual Domain Adaptation via transfer feature learning
[18] Domain invariant clustering
[19] Adaptation regularization based transfer learning
[20] Domain
[21] Task
[22] Frobenius norm
[23] Sparsity regularization parameter
[24] Reconstruction error
[25] Principal component analysis
[26] Office
[27] Caltech
[28] PIE
[29] Amazon
[30] Webcam
[31] Digital single-lens reflex
[32] Matlab
[33] Nearest neighbor
[34] Transfer component analysis
[35] Transfer feature learning with Joint Distribution Adaptation