تاریخ دریافت مقاله: ۹۷/۱۲/۱۰
تاریخ پذیرش مقاله: ۹۸/۰۸/۰۳

انتخاب نمونه‌های آموزشی بهینه براساس معیارهای فاصله برای آموزش رده‌بندی احساسات

شیوا نوری سرای
دانشجوی کارشناسی ارشد، دانشکده مهندسی فناوری اطلاعات و کامپیوتر ـ دانشگاه صنعتی ارومیه ـ ارومیه ـ ایران
پست الکترونیکی: shivanoorisaray@it.uut.ac.ir

جعفر طهمورث نژاد*
استادیار دانشکده مهندسی فناوری اطلاعات و کامپیوتر ـ دانشگاه صنعتی ارومیه ـ ارومیه ـ ایران
پست الکترونیکی: j.tahmores@it.uut.ac.ir

چکیده

افزایش چشمگیر دسترس‌پذیری به نظرها و توصیه‌های برخط باعث می‌شود رده‌بندی احساسات در متون کوتاه یکی از موضوع‌های جالب توجه در تحقیقات علمی و صنعتی باشد. در زمینه رده‌بندی احساسات، اصطلاحات به کار برده شده در دامنه‌های مختلف ممکن است متفاوت باشند. درنتیجه مدلی که با داده‌های برچسب‌دار آموزشی (دامنهٔ منبع) یادگیری می‌شود ممکن است عملکرد خوبی در برچسب‌گذاری داده‌های آزمایشی (دامنهٔ هدف) نداشته باشد. یادگیری انتقالی[1] و انطباق دامنه[2] دو راه حل مفید برای مواجهه با این مشکل هستند. یادگیری انتقالی و انطباق دامنه، توزیع‌های دامنه منبع و هدف را به هم نزدیک می‌کنند تا عملکرد رده‌بند در دامنهٔ هدف بهبود یابد اما نکتهٔ حائز اهمیت این است که کدام دامنه از مجموعه دامنه‌های نامزد به‌عنوان دامنه منبع انتخاب شود. روش پیشنهادی در این مقاله یک ترکیب خطی از معیارهای فاصله بین توزیع دامنه‌های منبع و هدف است که بهترین دامنهٔ منبع را برای یادگیری رده‌بند انتخاب می‌کند. روش پیشنهادی بر روی مجموعه داده‌های همگن[3] و ناهمگن[4] ارزیابی شده است. همان‌طورکه نتایج نشان می‌دهد، مدل پیشنهادی، در مجموعه داده همگن با احتمال ۴۷٫۱ درصد (۵٫۹ درصد در مدل تصادفی) و در مجموعه داده ناهمگن با احتمال ۲۳٫۱ درصد (۸٫۳ درصد در مدل تصادفی) می‌تواند دامنه منبع صحیح را انتخاب کند که حاکی از بهبود چشمگیر عملکرد مدل پیشنهادی نسبت به مدل تصادفی در انتخاب دامنه منبع صحیح است.

واژه‌های کلیدی: رده‌بندی احساسات، معیار فاصله، یادگیری انتقالی، انتقال دانش

۱- مقدمه

در رده‌بندی احساسات، مدلی بر روی یک دامنه منبع

* sentiment classification

* نویسندهٔ مسئول

1- transfer learning
2- domain adaptation
3- homogeneous
4- heterogeneous

با داده‌های برچسب‌دار آموزش داده می‌شود، سپس مدل برای برچسب‌گذاری داده‌های دامنه هدف به کار می‌رود. در رده‌بندی احساسات، داده‌های دامنه، متن‌های کوتاه هستند که با توجه به محتوای متن‌ها، رده‌بند، برچسب مثبت یا منفی به متن‌ها می‌دهد. رده‌بندی احساسات در زمینه‌های مختلف مانند تجارت الکترونیکی، بازیابی اطلاعات، سیستم‌های پیش‌بینی بازار بورس، پردازش گفتار، سیستم‌های توصیه‌گر به کار می‌رود. یکی از مسائل بحث برانگیز در رده‌بندی احساسات، اختلاف توزیع بین داده‌های دامنه منبع و هدف، یعنی نظراتی که مدل با آن‌ها ساخته می‌شود و نظراتی که مدل با آن‌ها آزموده می‌شود است که عملکرد مدل روی دامنه هدف را تحت تاثیر قرار می‌دهد.

روش‌های یادگیری انتقالی زیادی از جمله انطباق دامنهٔ بصری از طریق یادگیری ویژگی انتقالی [۱] و استخراج ویژگی‌های مشترک برای دامنه‌های چند منبعی در یادگیری انتقالی [۲] ارائه شده است تا افت عملکرد رده‌بند که ناشی از اختلاف توزیع دامنه‌ها است را کاهش دهد.

اختلاف توزیع دامنه‌های منبع و هدف، ناشی از اختلاف توزیع حاشیه‌ای[5] و توزیع شرطی[6] بین دامنه‌های منبع و هدف است. اختلاف توزیع حاشیه‌ای، به‌صورت اختلاف در احتمال وقوع هر یک از ویژگی‌ها در دامنه‌های منبع و هدف با ویژگی‌های یکسان و اختلاف توزیع شرطی، به‌صورت اختلاف در احتمال وقوع برچسب‌های متفاوت به ازای داده‌های یکسان از دو دامنهٔ منبع و هدف تعریف می‌شود.

روش‌های یادگیری انتقالی، به دو گروه نیمه‌نظارت‌شده[7] و بدون نظارت[8] تقسیم می‌شوند. روش‌های یادگیری انتقالی، در صورتی که تمام داده‌های دامنهٔ منبع و قسمت کوچکی از داده‌های دامنهٔ هدف برچسب داشته باشند، به‌صورت نیمه‌نظارت‌شده و اگر فقط داده‌های دامنهٔ منبع برچسب داشته باشند به‌صورت بدون نظارت تعریف می‌شوند.

5- marginal distribution
6- conditional distribution
7- semi-supervised
8- unsupervised

روش‌های یادگیری انتقالی، بر کاهش اختلاف توزیع دامنه‌های منبع و هدف و انتقال دانش از دامنه منبع به دامنه هدف تمرکز دارند در حالی که به انتخاب دامنه مناسب به‌عنوان دامنه منبع، توجهی ندارند. برای انتخاب دامنه منبع می‌توان عواملی مانند موضوع دامنه، هدف از تشکیل دامنه و یا چارچوبی که دامنه در آن تعریف شده است را در نظر گرفت ولی توجه به این عوامل الزاماً بهترین دامنه منبع را به دست نمی‌دهد. در اکثر دیدگاه‌های انتخاب دامنه منبع از معیارهای شباهت دامنه استفاده می‌شود. از جمله معیارهای شباهت، تابع فاصلهٔ حداکثر اختلاف میانگین (MMD)[9] است [۳]. این تابع از کرنل‌های مختلفی برای محاسبه فاصله بین دو دامنه منبع و هدف استفاده می‌کند ولی از معایب این معیار این است که محاسبه ماتریس کرنل برای مجموعه داده‌های بزرگ بسیار هزینه‌بر است.

روش پیشنهادی این مقاله، یک روش هوشمند انتخاب دامنهٔ بدون نظارت است که یک یا چند دامنه منبع مناسب را به‌صورت بدون نظارت از مجموع دامنه‌های در دسترس برای یادگیری رده‌بند انتخاب می‌کند در حالی که از برچسب‌های دامنه هدف استفاده نمی‌کند.

در ادامه، مقاله به‌صورت زیر سازماندهی شده است. دربخش دوم مقاله، مروری بر روی کارهای مربوطه انجام شده است. در بخش سوم مقاله، مسئله و روش پیشنهادی به تفصیل شرح داده شده است. در بخش چهارم، نتایج ارزیابی عملکرد الگوریتم پیشنهادی در مقایسه با روش‌های دیگر انتخاب دامنه منبع گزارش شده است. در انتها، مقاله با نتیجه‌گیری و ارائه پیشنهادهایی برای ادامه کار در آینده به اتمام رسیده است.

۲. مروری بر کارهای مربوطه

یادگیری انتقالی به دو دسته یادگیری انتقالی همگن و یادگیری انتقالی ناهمگن تقسیم می‌شود [٤]. در یادگیری انتقالی همگن فضای ویژگی دامنه‌ها با هم برابراست و به چهار زیرگروه که به شرح زیر می‌باشند، تقسیم می‌شود:

9- Maximum Mean Discrepancy

● یادگیـری انتقالی مبتنی بـر نمونه: در ایـن دیدگاه به منظـور کاهش اختـلاف توزیع‌های بیـن دامنه‌های منبع و هـدف، نمونه‌هایی از دامنه منبع کـه در کاهش اختلاف توزیع داده‌های منبع و هدف موثر هسـتند، انتخاب شده و یا وزن‌دهی مجدد می‌شـوند. روش انتخاب مرزنما[10] [۵]، از جمله روش‌های مبتنی بر نمونه اسـت. به نمونه‌هایی از دامنه منبع که کمترین اختلاف توزیع را با نمونه‌های دامنه هدف دارند، مرزنما گفته می‌شـود. در این روش، از معیار حداکثر اختلاف میانگین (MMD) برای محاسـبه اختلاف توزیـع نمونه‌های دامنه منبـع با نمونه‌هـای دامنه هدف بهره گرفته می‌شـود و نمونه‌هایی از دامنه منبع که دارای کمترین اختلاف توزیع با نمونه‌های دامنه هدف باشند، وزن بیشتری می‌گیرند. ایراد اصلی در روش انتخاب مرزنما این است که به ویژگی‌های مختص هر دامنه توجهی نمی‌شود.

● یادگیـری انتقالی مبتنـی بر ویژگی: در ایـن دیدگاه به منظور کاهـش اختلاف توزیع بین دامنه‌های منبع و هدف، داده‌ها به فضای ویژگی مشـترک با اختلاف توزیع کمتر نگاشـت می‌شـوند. JDA [۶] [11] و GFK [۷] [12] را از جملـه روش‌هـای این دیـدگاه می‌توان نام بـرد. در روش JDA، هدف یافتن ماتریس نگاشـتی اسـت که نمونه‌های منبع و هدف را به زیرفضای مشـترکی بـا حداقل اختلاف توزیع شرطی و حاشیه‌ای بین دامنه‌ها نگاشت کند. مشکل اصلی روش JDA عدم حفظ ساختار داده‌ها می‌باشد. روش GFK با استخراج ویژگی‌های پنهان مشترک بین دامنه‌های منبع و هدف، اختلاف توزیع حاشـیه‌ای بین دو دامنه را کاهش می‌دهد. روش GFK، برای کاهش اختلاف توزیع حاشیه‌ای بین دامنه‌های منبـع و هدف، ابعاد فضای اصلی داده‌ها را کاهش می‌دهد که به علت کاهش بیش از حد ابعاد، ساختار اصلی داده‌ها حفظ نمی‌شود.

● یادگیـری انتقالی مبتنی بر پارامتر/ مدل: در این دیدگاه، پارامترهـای مـدل آمـوزش دیـده روی نمونه‌های منبع با سـاختار نمونه‌هـای دامنه هدف تطبیق داده می‌شـود.

10- landmark
11- Joint Domain Adaptation
12- Geodesic Flow Kernel

تنظیم‌پذیری تطبیقی برای یادگیری انتقالی[13] [۸] یک روش تطبیق دامنه بدون نظارت است که از طریق کاهش اختلاف توزیع هندسـی و آمـاری بین دامنه‌های منبـع و هدف و همچنین کاهش خطای رده‌بنـد در دامنه منبع، یک رده‌بند تطبیق یافته بین دامنه‌های منبع و هدف ایجاد می‌کند.

به طور کلی، مدل‌های مختلفی بـرای یادگیری انتقالی اسـتفاده می‌شـود، از جمله این مدل‌ها می‌توان به ماشین بردار پشـتیبان (SVM)[14] [۹]، شـبکه‌های عصبی احتمالی (PNN) [15] [۱۰]، مـدل مخلوطی گوسـی (GMM)[16] [۱۱] و مدل حداکثر آنتروپی (ME)[17] [۱۲] اشاره کرد. مدل ماشین بردار پشـتیبان، نتایج خیلی خوبی در رده‌بندی داده‌ها به دست می‌دهد. در مدل ماشین بردار پشتیبان سعی بر پیدا کـردن بهترین ابرصفحه بین رده‌ها بـرای رده‌بندی بهتر داده‌ها اسـت. این مدل بیشـتر برای رده‌بندی احساسات دربارهٔ فیلم‌ها استفاده می‌شود. مدل شبکه عصبی احتمالی برای رده‌بندی داده‌های با بعد زیاد مفید است. در مدل‌های مخروطی گوسی، مولفه‌های گوسی برخی از ویژگی‌هایی که در برچسب‌گذاری مفیدتر هسـتند را نشان می‌دهند و می‌توانـد انواع مختلف توزیع‌ها را مـدل کند. مدل حداکثر آنتروپی، مدل مبتنی بر ویژگی اسـت که توزیع هر رده را پیدا می‌کند و سـبد کلمات[18] دوتایـی و عبارات را می‌توان به‌عنوان ویژگی به مـدل حداکثر آنتروپی داد بدون این‌که نگـران همپوشـانی ویژگی‌ها باشـیم. با این حـال، برای رده‌بندی می‌توان ترکیبی از رده‌بندهای فوق را نیز استفاده کرد.

● یادگیـری انتقالی مبتنی بـر رابطه: این دیدگاه شـامل روش‌های یادگیری الگوهای سـاختار گرامری و جمله‌ای دامنه‌هـا و یافتن رابطه بین دامنه‌های منبع و هدف اسـت. RAP [19] [۱۳] یک روش انطباق دامنه تحت نظارت اسـت که برای اسـتخراج دانش از تحلیل احساسات استفاده می‌کند

13- Adaptation Regularization for Transfer Learning
14- Support Vector Machine
15- Probability Neural Network
16- Gaussian Mixture Model
17- Maximum Entropy
18- Bag of Words
19- Relational Adaptive bootstraPping

و به این صورت عمل می‌کند که ابتدا کلمات احساسی و مبحث دامنه هدف را مشخص می‌کند سپس آن‌ها را توسعه می‌دهد و رابطهٔ بین مبحث و احساسات را تعیین می‌کند.

در یادگیری انتقالی ناهمگن فضای ویژگی دامنه‌های منبع و هدف متفاوت هستند. دو دیدگاه اصلی برای حل مسئله فضای ویژگی ناهمگن وجود دارد که به شرح زیر است:

یادگیری انتقالی مبتنی بر ویژگی متقارن: در این دیدگاه، دامنه‌های منبع و هدف به‌صورت جداگانه به فضای ویژگی پنهان انتقال داده می‌شوند. روش DAMA[۱۴] [۲۰] از این دیدگاه استفاده می‌کند. در این روش به منظور ایجاد فضای پنهان و اتصال فضای ویژگی دامنه‌ها از تابع نگاشت متفاوت برای هر دامنه استفاده می‌شود. همچنین در روش DAMA، از برچسب نمونه‌ها به جای ویژگی‌های معادل نمونه‌ها برای انطباق هندسی استفاده می‌شود.

- یادگیری انتقالی مبتنی بر ویژگی نامتقارن: در این دیدگاه فضای ویژگی دامنه منبع به فضای ویژگی دامنه هدف انتقال داده می‌شود. روش HDP [۱۵] [۲۱]، یک روش تشخیص نقص بر اساس تطبیق معیار با استفاده از تحلیل آماری است. انطباق معیار برای HDP به توزیع داده‌های به اندازه کافی بزرگ نیاز دارد. مشکل عمده این روش این است که حداقل ۵۰ نمونه برای انتقال دانش و تشخیص عیب مورد نیاز است.

انتخاب دامنهٔ منبع بهینه برای آموزش رده‌بند، می‌تواند عملکرد را در روش‌های انطباق دامنه و یادگیری انتقالی به طور چشمگیری بهبود دهد. در واقع دیدگاه‌های یادگیری انتقالی با روش‌های مختلف، توزیع دامنه‌های منبع و هدف را به هم نزدیک می‌کنند تا رده‌بند روی دامنه‌ای نزدیک به دامنهٔ هدف آموزش داده شود. بدین ترتیب اگر از ابتدا یک دامنهٔ آموزشی بهینه برای یادگیری رده‌بند انتخاب کنیم و سپس با استفاده از روش‌های یادگیری انتقالی و انطباق دامنه اختلاف توزیع دامنهٔ منبع بهینه و دامنه هدف را کم کنیم بهبود چشمگیری در عملکرد رده‌بند در رده‌بندی داده‌های دامنهٔ هدف به وجود خواهد آمد. بنابراین انتخاب دامنه منبع مناسب به‌عنوان پیش مرحله روش‌های یادگیری انتقالی، می‌تواند اتلاف تطبیق[۲۲] را بهبود دهد.

در این مقاله، روش CCEM [۲۳] ایده جدیدی برای انتخاب دامنه منبع پیشنهاد می‌کند که نیاز به برچسب داده‌های دامنهٔ هدف ندارد. در واقع این روش، ترکیب خطی از معیارهای فاصله است که وزن‌ها در ترکیب خطی، با یادگیری روی دامنه‌های منبع در دسترس به دست می‌آیند و به این صورت مدل انتخاب دامنه منبع یاد گرفته می‌شود.

۳. بیان مسئله

در رده‌بندی احساسات، داده‌های دامنه، اسنادی هستند که هر سند با یک متغیر تصادفی X و برچسب آن سند با متغیر تصادفی Y نشان داده می‌شود که P و $\overline{P}$ دو توزیع روی X×Y هستند و توزیع داده‌های دامنه‌های منبع و هدف را نشان می‌دهند و توزیع حاشیه‌ای دامنه‌های منبع و هدف روی X به ترتیب با P_x و $\overline{P}_x$ مشخص می‌شود. یک رده‌بند برای هر کدام از متغیرهای تصادفی یک برچسب تعیین می‌کند و بهترین رده‌بند آن است که کمترین خطا را در تعیین برچسب داشته باشد که با بیان ریاضی به‌صورت رابطه (۱) تعریف می‌شود:

$$h_p^* := \arg\min_h Prob(h(x)\neq y | (x,y)\sim P) \quad (۱)$$

در رابطهٔ (۱)، h_p^*، بهترین رده‌بند برای رده‌بندی داده‌های دامنهٔ هدف است که از بین طیف وسیعی از رده‌بندهای $h(x)$ انتخاب می‌شود.

دومفهوم خطای رده‌بندی داخلی دامنه $\xi(P)$ و خطای رده‌بندی بین دامنه‌ای $\xi(P,\overline{P})$ به این صورت تعریف می‌شود که خطای رده‌بندی داخلی دامنه، احتمال رده‌بندی اشتباه در دامنه‌ای است که رده‌بند روی آن دامنه آموزش دیده است در حالی که خطای رده‌بندی بین دامنه‌ای به‌صورت احتمال رده‌بندی اشتباه در دامنهٔ هدف با استفاده

20- Domain Adaptation Manifold Alignment
21- Heterogeneous Defect Prediction
22- adaptation loss
23- linear combination of (Chi2),(CMD),(EMD),(MKL)

از رده‌بند یاد گرفته شـــده روی دامنه منبع تعیین می‌شود. رابطه‌های (۲) و (۳) به ترتیب خطای رده‌بندی داخلی دامنه و خطای رده‌بندی بین دامنه‌ای را نشان می‌دهند:

$$\xi(\mathrm{P}) := Prob\left(h_p^* \neq y \middle| (x,y \sim \mathrm{P}\right) \qquad (۲)$$

$$\xi(\mathrm{P},\overline{\mathrm{P}}) := Prob\left(h_p^* \neq y \middle| (x,y \sim \overline{\mathrm{P}}\right) \qquad (۳)$$

با توجـــه به اختـــلاف توزیع‌های دامنـــه منبع و هدف همواره $\xi(\mathrm{P},\overline{\mathrm{P}})$ بزرگتر از $\xi(\mathrm{P})$ اســـت [۱۶]. به اختلاف بیـــن خطای رده‌بندی داخلی دامنـــه و خطای رده‌بندی بین دامنـــه‌ای، اتلاف تطبیق گفته می‌شـــود. بهترین رده‌بند آن اســـت که خطای رده‌بندی بین دامنه‌ای کمی داشته باشد و در واقع هدف، انتخاب دامنه منبعی (P^*) اســـت که خطای رده‌بنـــدی بین دامنه‌ای را کم کند. بدیـــن ترتیب اگر $\mathcal{P}$ را مجموعـــه دامنه‌های منبع در دســـترس تعریف کنیم هدف مقاله به‌صورت رابطه (٤) تعریف می‌شود:

$$\mathrm{P}^* := \arg\min_{\mathrm{P}\epsilon\mathcal{P}} \xi(\mathrm{P},\overline{\mathrm{P}}) \qquad (٤)$$

مسئله‌ای که در پیدا کردن بهترین دامنه منبع با استفاده از رابطه (٤) به وجود می‌آید این اســـت که برای پیدا کردن خطای رده‌بندی بین دامنه‌ای به داده‌های برچسب‌دار دامنه هدف نیاز اســـت در حالی که دامنهٔ هـــدف، داده‌های بدون برچســـب داشته و فقط توزیع حاشـــیه‌ای دامنه هدف در دسترس است.

۱،۳. روش پیشنهادی CCEM

در این مقاله برای حل مشـــکل عدم وجود برچسب در دامنه هدف روشی به نام CCEM معرفی می‌شود که ترکیب خطی از چهار معیار فاصله و خطای رده‌بندی داخلی دامنه $\xi(\mathrm{P})$ است که معیارهای فاصله، فاصله آماری بین توزیع حاشیه‌ای دامنه‌های منبع و هدف را محاسبه می‌کند. برای یافتـــن دامنهٔ منبع بهینه فرض می‌شـــود دامنـــه منبعی که کمترین خطای رده‌بندی داخلی دامنه و کمترین فاصله را با دامنـــه هدف دارد $\hat{d}(\mathrm{P},\overline{\mathrm{P}}_{\mathrm{x}})$، کمترین خطای رده‌بندی بین دامنه‌ای $\xi(\mathrm{P},\overline{\mathrm{P}})$ را نیـــز دارد و در واقع خطای رده‌بندی بین دامنه‌ای را می‌توان با خطای رده‌بندی داخلی دامنهٔ منبع و فاصلـــهٔ بین دامنهٔ منبع و هـــدف تقریب زد که به‌صورت رابطهٔ (۵) نشـــان داده می‌شود و دامنه، با خطای رده‌بندی داخلـــی و فاصله تا دامنهٔ هدف کمتر، برای آموزش رده‌بند انتخاب می‌شود.

$$\xi(\mathrm{P},\overline{\mathrm{P}}) \approx \hat{d}(\mathrm{P},\overline{\mathrm{P}}_{\mathrm{x}}) + \xi(\mathrm{P}) \qquad (۵)$$

فرض می‌کنیم $\mathcal{D}$ مجموعه‌ای از توابع فاصله اســـت و $\overline{\mathrm{P}}$ مجموعـــه‌ای از دامنه‌ها اســـت و به دلیـــل این‌که در آزمایش‌های دامنهٔ هدف جداگانـــه نداریم، در هر تکرار از آزمایش یکی از دامنه‌ها به‌عنوان دامنهٔ هدف و بقیهٔ دامنه‌ها به‌عنوان دامنه‌هـــای منبع در مجموعهٔ نامزد در نظر گرفته می‌شـــوند. برای انتخاب بهترین معیار فاصله از رابطهٔ (٦) استفاده می‌کنیم:

$$\hat{d} := \arg\min_{d\,\in\,\mathcal{D}} \sum_{\mathrm{P}\in\overline{\mathrm{P}}} |\xi(\mathrm{P},\overline{\mathrm{P}}) - \hat{d}(\mathrm{P},\overline{\mathrm{P}}_{\mathrm{x}})| \qquad (٦)$$

برای انتخاب مجموعه‌ای از معیارهای فاصله، از بردار M بـــا n معیار فاصله اســـتفاده می‌کنیم که هـــر کدام از M_iهـــا یـــک معیار فاصله بـــا ضریب $\alpha_i \epsilon R^+$ اســـت. خطای رده‌بندی دامنهٔ منبع نشـــان‌دهندهٔ مناسب بودن منبع برای یادگیری رده‌بند اســـت. اگر خطای دامنه زیاد باشـــد دامنه ارزش کمتـــری بـــرای انتخاب دارد به همیـــن دلیل خطای رده‌بندی داخلی هر دامنه را محاســـبه می‌کنیم و به ترکیب خطـــی معیارهای فاصلـــه اضافه می‌کنیـــم. یافتن ضرایب بهینه بـــرای معیارهای فاصله به‌صورت رابطه (۷) تعریف می‌شود.

$$\hat{d} := \arg\min_{\alpha\,\geq\,0} \sum_{\mathrm{P}\in\overline{\mathrm{P}}} |\xi(\mathrm{P},\overline{\mathrm{P}}) - \alpha_{\mathrm{M}}(\mathrm{P},\overline{\mathrm{P}}_{\mathrm{x}})| \qquad (۷)$$

بنابراین بدون نیاز به داده‌های برچسب‌دار دامنه هدف می‌توان بهترین دامنه منبع را برای آموزش رده‌بند انتخاب کرد زیـــرا برای محاســـبه معیارهای فاصلـــه فقط توزیع حاشیه‌ای دامنه‌های منبع و هدف استفاده می‌شود بنابراین تعریف رابطـــه (٤) به‌صورت رابطـــه (۸) تغییر می‌کند که هدف مقاله است:

$$\mathrm{P}^* := \arg\min_{\mathrm{P}\epsilon\mathcal{P}} \hat{\mathrm{d}}(\mathrm{P},\overline{\mathrm{P}}_{\mathrm{x}}) \qquad (۸)$$

۲،۳. معیارهای فاصله

از آن‌جایی که توزیـــع دامنه‌های منبع و هدف اختلاف دارند در بیشتر رویکردها برای انتخاب دامنه منبع بهینه، که

اختلاف توزیع کمتری با دامنه هدف دارد، شباهت دامنه‌ها اندازه‌گیری می‌شـود. در این مقالـه از چهار معیار فاصله برای محاسـبه میزان شباهت توزیع حاشیه‌ای دامنه‌ها در مدل‌های مختلف اسـتفاده می‌شـود. چهـار معیار فاصله عبارتنـداز آزمون آماری پیرسـون[۲۴] [۱۷]، فاصله حرکت زمیـن[۲۵] [۱۸]، رگرسـیون کولبک-لیبلر متوسـط[۲۶] [۱۹]، اختلاف گشـتاور مرکزی[۲۷] [۲۰] کـه در ادامه به‌صورت اجمالی به آن‌ها پرداخته می‌شود.

● اولین معیار، فاصله آزمون آماری پیرسون (Chi2) است که شـباهت بین دو توزیـع P_x و $\bar{P}_x$ را به‌صورت رابطهٔ (۹) محاسبه می‌کند که $p(w)$ و $\bar{p}(w)$ احتمال کلمه w به ترتیب در توزیع‌های P_x و $\bar{P}_x$ هستند و N تعداد کلمه‌های با بیشترین تکرار است.

$$chi2(P_x, \bar{P}_x) := \sum_{i=1}^{N} \frac{((P(w_i)-\bar{P}(w_i))^2}{P(w_i)+\bar{P}(w_i)} \qquad (۹)$$

دومین معیار، تابع فاصلهٔ حرکت زمین (EMD) است که مقدار جرم احتمالی و مسافتی که جرم احتمالی باید جابه‌جا شود تا دو توزیع دامنه منبع و هدف به هم نزدیک‌تر شوند را محاسبه می‌کند و به‌صورت رابطه (۱۰) تعریف می‌شود که $f_{i,j}$ جرم احتمالی و $d_{i,j}$ مسافت بین خوشه‌های i و j از نمایش خوشه‌ای b و $\bar{b}$ است که هرکدام از b و $\bar{b}$ شامل m خوشه هستند.

$$\mathrm{EMD}(P_x, \bar{P}_x) := \min_{F} \sum_{i=1}^{m}\sum_{j=1}^{m} d_{i,j} f_{i,j} \qquad (۱۰)$$
$$\text{Subject to: } f_{i,j} \geq 0,\ \sum_{j=1}^{m} f_{i,j} = \bar{p}_j,\ \sum_{i=1}^{m} f_{i,j} = p_i$$

سومین معیار نسـخهٔ اصلاح شدهٔ رگرسیون کولبک-لیبلر(MKLD) است که با نام آنتروپی رابطه‌ای نیز شناخته می‌شـود و به‌صـورت (۱۱) تعریـف می‌شـود. m تعداد خوشه‌های موجود در نمایش خوشه‌ای است.

$$\mathrm{MKL}(p_x, \bar{p}_x) = \sum_{i=1}^{m} E(p_x)_i \log\frac{E(p_x)_i}{E(\bar{p}_x)_i} + E(\bar{p}_x)_i \log\frac{E(\bar{p}_x)_i}{E(p_x)_i} \qquad (۱۱)$$

● چهارمین معیار، اختلاف گشتاور مرکزی (CMD) است و به‌صـورت رابطه‌هـای (۱۲) و (۱۳) تعریف می‌شـود.

P_x و $\bar{P}_x$ توزیع‌های حاشیه‌ای دامنه‌های منبع و هدف در فاصله فشـرده $[a,b]^N$ هسـتند و $E(p_x)$ و $c_k(p_x)$ به ترتیب امید ریاضی و بردار گشـتاور مرکزی از درجه k هستند. $\|\cdot\|_2$ نرم مربع است.

$$CMD(p_x, \bar{p}_x) = \frac{1}{|b-a|}\|E(p_x) - E(\bar{p}_x)\|_2 + \sum_{k=2}^{\infty}\frac{1}{|b-a|}\|c_k(p_x) - c_k(\bar{p}_x)\|_2 \qquad (۱۲)$$

$$c_k(p_x) = \left(E\left(\prod_{i=1}^{N}((p_x)_i - E(p_x)_i)^{r_i}\right)\right) \qquad (۱۳)$$
$$\text{Subject to: } r_1 + \cdots + r_N = k,\ r_1, \ldots, r_N \geq 0$$

مدل پیشـنهادی از ترکیب خطی چهار معیار ذکر شده، خطای انحراف و خطای رده‌بندی داخلی دامنه منبع تشکیل شده است که به‌صورت رابطه (۱٤) تعریف می‌شود.

$$d(P_x, \bar{P}_x) = \alpha_1 chi2(P_x, \bar{P}_x) + \alpha_2 CMD(P_x, \bar{P}_x) + \qquad (۱٤)$$
$$+ \alpha_3 \mathrm{EMD}(P_x, \bar{P}_x) + \alpha_4 \mathrm{MKL}(P_x, \bar{P}_x) + \alpha_5 \xi(\mathrm{P}) + \alpha_0$$

ضرایـب α از طریق حداقل کردن تابع اتلاف، که چهار معیـار فاصله، بـرای دامنه‌های منبع و خطـای رده‌بندی داخلی دامنه‌های منبع را به‌عنوان ورودی می‌گیرد، به دست می‌آید. با جایگذاری این ضرایب در رابطهٔ (۱٤)، برای هریک از دامنه‌های منبع و دامنهٔ هدف، فاصلهٔ هر یک از دامنه‌های منبع و دامنهٔ هدف به دست می‌آید. دامنهٔ منبعی که کمترین فاصله با دامنهٔ هدف $d(P_x, \bar{P}_x)$ را داشته باشد، به‌عنوان دامنهٔ منبع بهینه انتخاب می‌شود.

۴. ارزیابی عملکرد

در این مقاله برای کاهـش پیچیدگی دامنه‌ها، اطلاعات اضافـی با حذف علامت‌گذاری‌هـا و تبدیل همه حروف به حروف کوچک تقلیل شـده است و برای نمایش دامنه‌ها از سبد کلمات یکتایی و دوتایی استفاده شده است که کلمه‌ها فضای ویژگی‌ها را تشکیل می‌دهند و تعداد تکرار هر کلمه به‌عنوان مقدار آن ویژگی تلقی می‌شود. هر سند با مجموع بردارهای ویژگی‌هایش نشـان داده می‌شود. در این مقاله از مـدل وزن‌دهی TF-IDF[۲۸] بـرای وزن‌دهی به ویژگی‌ها

24- statistic of Pearson's χ2 test
25- Earth Mover's Distance
26- Mean Kullback-Leibler
27- Central Moment Discrepancy
28- Term Frequency-Inverse Document Frequency

استفاده شده اســت. برای ارزیابی روش پیشنهادی از دو مجموعه داده همگن و ناهمگن اســتفاده شده است این دو مجموعه داده در چارچوب، موضوع، تعداد سندهای موجود در دامنه و طول سندها با همدیگر متفاوت هستند. مجموعه داده همگن، شامل ۱۷ دسته پیکره[۲۹] از DRANZIERA است [۲۱] که از هر پیکره ۵۰۰۰ ســند مثبت و ۵۰۰۰ سند منفی انتخاب شــده اســت و طول ســندها بین ٦٤ تا ۱۲۳ کلمه متغیر است. هر پیکره شــامل احساسات کاربران آمازون روی محصولات از دسته‌های مختلف است. در هر مرحله از آزمایــش ۱۷ مرحله‌ای مجموعه داده همگن، یک دســته برای دامنهٔ هدف و ۱٦ دســته دیگر به‌عنوان دامنهٔ منبع در مجموعهٔ نامزد، در نظر گرفته می‌شود تا برای هر دسته از ۱۷ دســته پیکره به‌عنوان دامنهٔ هدف، یک دامنهٔ منبع بهینه از مجموعهٔ نامزد انتخاب شــود. مجموعه داده ناهمگن نیز از ۱۳ دسته پیکره تشکیل شده است که متوسط تعداد سند در پیکره‌ها ٦۳۲٦ اســت و طول سندها بین ۱۰ تا ۱۵ کلمه متغیر اســت. در مجموعه داده ناهمگن، احساسات دربارهٔ فیلم‌ها [۲۲]، محصولات مصرفی [۲۳]، مجموعه توییت‌ها با موضوع‌های مختلف [۲٤]، پیکره‌ها را تشکیل می‌دهند. در هر مرحله از آزمایش ۱۳ مرحله‌ای مجموعه داده ناهمگن، یک دسته برای دامنهٔ هدف و ۱۲ دسته دیگر به‌عنوان دامنهٔ منبع در مجموعهٔ نامزد، در نظر گرفته می‌شود تا برای هر دسته از ۱۳ دسته پیکره به‌عنوان دامنهٔ هدف، یک دامنهٔ منبع بهینه از مجموعهٔ نامزد انتخاب شود. داده‌های آموزشی، از ترکیب چهار معیار فاصله بین دامنه‌های مجموعهٔ نامزد و خطای داخلی دامنه‌های مجموعهٔ نامزد و داده‌های آزمون، از ترکیــب چهار معیار فاصله بین دامنهٔ هدف و دامنه‌های مجموعهٔ نامزد و خطــای داخلی دامنه‌های مجموعهٔ نامزد تشکیل می‌شود.

برای تابع اتلاف از رگرسیون خطی استفاده شده است که به‌صورت رابطهٔ (۱۵) تعریف می‌شود که با یافتن حداقل مقدار تابــع اتلاف، بهترین ضرایب α بــرای ترکیب خطی CCEM در رابطهٔ (۱٤) به دست می‌آید.

29- corpora

$$l(x,y)_{LR}=\sum_{i=1}^{n_d}\log(1+e^{-(2y_i-1)x_i\alpha+\alpha_0})+\mu\frac{1}{2}||\alpha||^2 \quad (۱۵)$$

که در مــدل CCEM برای مجموعــه داده‌های همگن و ناهمگن ضرایب بهینه α به‌صورت زیر به دست می‌آید:

$$\alpha_{\text{همگن}}=[0.01,0.002,0.04,0.04,1.61,0.14]$$
$$\alpha_{\text{ناهمگن}}=[0.021,0.015,0.58,0.007,1.28,0.24]$$

پس از یافتن ضرایــب α در رابطه (۱۵)، از رابطه (۱٦) بــرای رده‌بندی دودویــی داده‌های دامنه هدف اســتفاده می‌شود.

$$h(x)=\begin{cases}1 & if\ \alpha x+\alpha_0>0\\ 0 & else\end{cases} \quad (۱٦)$$

بــرای پیدا کردن بهترین دامنــه منبع با کمترین خطای رده‌بندی بین دامنه‌ای $\xi(\mathrm{P},\bar{P})$ جهــت آموزش رده‌بند، از خطای رده‌بندی بین دامنهٔ رابطه‌ای (۱۷) استفاده می‌شود.

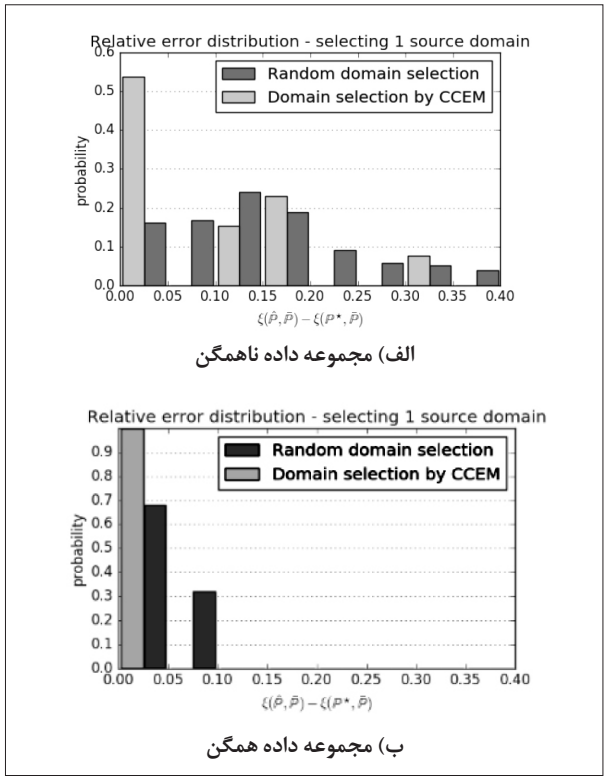
$$\xi_{related}(\hat{p},\bar{p})=\xi(\hat{p},\bar{p})-\xi(p^*,\bar{p}) \quad (۱۷)$$

در رابطــه (۱۷) $\hat{p}$ ، $\bar{p}$ و p^* بــه ترتیــب توزیــع دامنه انتخابی منبع توسط روش پیشنهادی، توزیع دامنه هدف و توزیع دامنه منبع بهینه است. می‌دانیم p^* بهترین دامنه منبع بــا کمترین خطای رده‌بندی بین دامنه‌ای اســت و اگر مدل پیشنهادی بهترین دامنه منبع را انتخاب کند خطای رابطه‌ای صفر می‌شود.

در شــکل (۱) توزیع خطای رابطه‌ای مدل پیشنهادی با مدل تصادفی برای انتخاب یک دامنه منبع مقایســه شــده اســت. محور افقی خطــای رده‌بند بین دامنــه‌ای رابطه‌ای اســت که از رابطه (۱۷) به دســت می‌آید که نشــان‌دهندهٔ خطــای انتخاب دامنه صحیح به‌عنوان دامنهٔ منبع اســت و محور عمودی، احتمال انتخاب دامنه با خطای رده‌بندی بین دامنه‌ای رابطه‌ای مشخص به‌عنوان دامنه منبع است.

همان‌طور که در شکل (۱) دیده می‌شود مدل CCEM، در مجموعــه داده ناهمگن با احتمال ۰٫۵۳ و در مجموعه داده همگــن با احتمال ۱ دامنه منبعی با کمترین خطای رابطه‌ای انتخاب می‌کند که نشان از قدرتمند بودن روش پیشنهادی در انتخاب دامنه منبع بهینه، برای آموزش رده‌بند است.

در شــکل (۲) خطــای رده‌بندی بین دامنــه‌ای مطلق را



الف) مجموعه داده ناهمگن

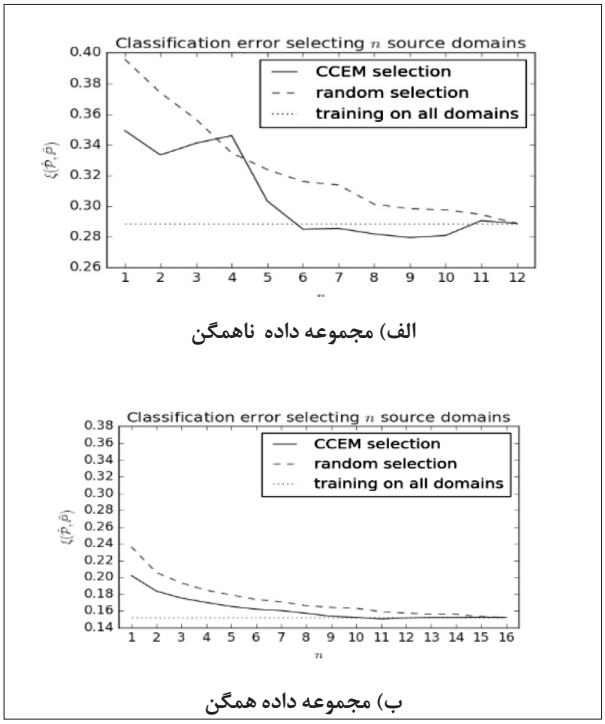
ب) مجموعه داده همگن

شکل ۱: توزیع خطای رابطه‌ای مدل CCEM و مدل تصادفی برای انتخاب یک دامنه منبع، الف) برای مجموعه داده ناهمگن، ب) برای مجموعه داده همگن

زمانی کــه مدل بیش از یک دامنه منبــع را انتخاب می‌کند نشان می‌دهد که در این شــکل مدل‌های CCEM، تصادفی و مدلــی که روی همه مجموعه داده‌ها آموزش دیده، با هم مقایســه شــده‌اند. محور افقی، تعداد دامنه منبع انتخابی و محور عمــودی، خطای رده‌بندی بین‌دامنه‌ای اســت که از رابطه (۳) به دست می‌آید.

در شــکل (۲) همان‌طور که دیده می‌شود مدل CCEM نســبت به مدل تصادفی خوب عمل می‌کند و حتی عملکرد مدل پیشنهادی در مجموعه داده ناهمگن زمانی که بین ٦ تا ۱۱ دامنه منبع انتخاب می‌شود نسبت به مدل آموزش دیده روی همهٔ مجموعه داده‌ها بهتر است.

در جدول (۱) احتمال انتخاب بهترین دامنه منبع درست، احتمــال انتخاب یکــی از ۳ دامنه منبع بــا بدترین خطای رده‌بندی بین دامنه‌ای و خطای رده‌بندی بین دامنه‌ای مطلق برای ســه مدل CCEM، تصادفی و حالت بهینه نشان داده شده است.



الف) مجموعه داده ناهمگن

ب) مجموعه داده همگن

شکل ۲: مقایسه خطای رده‌بندی بین دامنه‌ای مطلق برای مدل‌های CCEM و تصادفی با انتخاب n دامنه منبع و مدلی که رو همهٔ مجموعه داده‌ها آموزش دیده، الف) برای مجموعه داده ناهمگن، ب) برای مجموعه داده همگن

در جدول (۱) پارامتر احتمال انتخاب دامنهٔ منبع صحیح، نشــان‌دهندهٔ این اســت که دامنهٔ صحیح با چــه احتمالی به‌عنوان دامنهٔ منبع انتخاب می‌شود که مدل پیشنهادی در مجموعــه داده همگن با احتمال نزدیک به ۵۰ درصد (٤٧,١ درصد) و در مجموعه داده ناهمگن با احتمال ۲۳,۱ درصد می‌تواند دامنه منبع صحیح را انتخاب کند که نسبت به مدل تصادفی بهبود چشمگیری دارد.

پارامتر احتمال انتخاب یکی از سه دامنهٔ منبع با بدترین خطای بین دامنه‌ای، در جدول (۱) نشــان می‌دهد که روش انتخابی با چه احتمالی یکی از ســه دامنه‌ای که بیشــترین خطای رده‌بنــدی بین دامنه‌ای دارند را به‌عنوان دامنهٔ منبع انتخاب می‌کند که مــدل CCEM با احتمال صفر درصد در مجموعه داده همگن و بــا احتمال ۷,۷ درصد در مجموعه داده ناهمگن عملکرد بهتری نســبت به مدل تصادفی دارد و نشــان‌دهندهٔ این است که روش ارائه شده، دامنه منبع با خطای رده‌بندی بین دامنه‌ای بالا را انتخاب نمی‌کند.

پارامتر خطای رده‌بنــدی بین دامنه‌ای مطلق در جدول

جدول ۱: عملکرد سه مدل انتخابی حالت بهینه، CCEM و تصادفی

	متد انتخابی	مدل بهینه (٪)	مدل CCEM (٪)	مدل تصادفی (٪)
همگن	احتمال انتخاب دامنه منبع صحیح	۱۰۰	۴۷,۱	۵,۹
	احتمال انتخاب یکی از سه دامنه منبع با بدترین خطای بین دامنه‌ای	۰	۰	۷۰,۶
	خطای رده‌بندی بین دامنه‌ای مطلق	۱۹,۹	۲۰,۲	۲۴,۷
ناهمگن	احتمال انتخاب دامنه منبع صحیح	۱۰۰	۲۳,۱	۸,۳
	احتمال انتخاب یکی از سه دامنه منبع با بدترین خطای بین دامنه‌ای	۰	۷,۷	۴۱,۷
	خطای رده‌بندی بین دامنه‌ای مطلق	۲۵,۲	۳۳,۲	۴۰,۳

(۱) نشان‌دهندهٔ میانگین خطای رده‌بندی بین دامنه‌ای همهٔ دامنه‌های هدف است. به عبارت دیگر هر یک از دامنه‌های مجموعهٔ نامزد یک بار به‌عنوان دامنهٔ هدف انتخاب می‌شود و خطای رده‌بندی بین دامنه‌ای برای هر کدام از دامنه‌های هدف محاسبه شده و میانگین آن‌ها را به‌عنوان خطای رده‌بندی بین دامنه‌ای مطلق در نظر می‌گیریم. مدل CCEM با خطای ۳۳.۲ درصد نسبت به مدل تصادفی با خطای ۴۰,۳ درصد عملکرد نسبتاً خوبی دارد.

۵. نتیجه‌گیری

در این مقاله با توجه به بار محاسباتی آموزش رده‌بند روی همه مجموعه داده‌ها، روش CCEM پیشنهاد شد که از ترکیب خطی چهار معیار فاصله، به همراه خطای رده‌بندی داخلی دامنه و خطای انحراف[۳۰] تشکیل می‌شود و با استفاده از توزیع دامنه منبع و توزیع حاشیه‌ای دامنه هدف، می‌تواند با احتمال خوبی بهترین دامنه منبع را به دست دهد. با تعریف n به‌عنوان تعداد دامنه‌های مجموعهٔ نامزد، مدل بر روی n(n-1) جفت دامنهٔ منبع و هدف مورد آزمایش قرار می‌گیرد تا با یافتن بهترین بردار ضرایب α برای ترکیب خطی رابطهٔ (۱۴)، بتواند با حداقل کردن خطای مطلق تخمین، دامنه منبع بهینه‌ای انتخاب کند. مدل پیشنهادی بر روی داده‌های همگن و ناهمگن با مدل تصادفی و مدل آموزش دیده روی همه دامنه‌ها مقایسه شد.

برای بهبود این روش می‌توان روی روش‌های انتخاب ویژگی تمرکز کرد که ویژگی‌هایی با اطلاعات مفیدتر درباره برچسب داده‌ها انتخاب شوند. در این مقاله رده‌بند روی یک دامنه آموزش می‌بیند در حالی که اگر از روش‌هایی استفاده شود که رده‌بند روی چند دامنه یادگیری کند نتایج بهتری به دست می‌آید. برای تعیین تعداد بهینه دامنه‌ها برای یادگیری نیز می‌توان از روش‌های متفاوت استفاده کرد. همچنین اگر بتوان معیارهای اندازه‌گیری فاصله‌ای را پیدا کرد که روی همه مجموعه داده‌ها خوب کار کنند یکی از معایب این روش، که عملکرد ضعیف بعضی معیارها روی مجموعه داده‌های مختلف است، رفع خواهد شد.

30- offset error

مراجع

1. Tahmoresnezhad, J., and Hashemi, S., "Visual domain adaptation via transfer feature learning", Knowledge and Information Systems, vol.50, no. 2, pp. 585-605, 2016.
2. Tahmoresnezhad, J., and Hashemi, S., "Common feature extraction in multi-source domains for transfer learning", information and knowledge technology (IKT), vol. 7, IEEE, 2015.
3. Gretton, A., Borgwardt, K. M., Rasch, M. J., Sch¨olkopf, B., and Smola, A., "A kernel two-sample test, Journal of Machine Learning Research", vol. 13, pp. 723–773, Mar 2012.
4. Weiss, K., Khoshgoftaar, T.M. and Wang, D., A "survey of transfer learning", Journal of Big data, vol.3, no.1, pp.9, Dec 2016.
5. Gong, B., Grauman, K. and Sha, F., "Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation", In International Conference on Machine Learning, vol.28, no.1 pp. 222-230, February 2013.

6. Long, M., Wang, J., Ding, G., Sun, J., and Yu, PS., "Transfer feature learning with joint distribution adaptation", IEEE international conference on computer vision, pp. 2200–07, Dec. 2013.
7. Gong, B., Shi, Y., Sha, F., and Grauman, K., "Geodesic flow kernel for unsupervised domain adaptation", IEEE conference on computer vision and pattern recognition, pp. 2066–73, Jun. 2012.
8. 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.
9. Maulik, U., and Chakraborty, D., "Remote sensing image classification: a survey of support-vector-machine-based advanced techniques", IEEE Geoscience and Remote Sensing Magazine, vol. 5, no. 1, pp. 33-52. Mar. 2017.
10. Iounousse, J., Er-Raki, S., El Motassadeq, A. and Chehouani, H., "Using an unsupervised approach of Probabilistic Neural Network (PNN) for land use classification from multitemporal satellite images", Applied Soft Computing, vol. 30, pp. 1-13. May. 2015.
11. Gui, C., Li, W., Pan, Z., Zhang, J., Zhu, J., Cui, D., "A classifier for diagnosis of manic psychosis state based on SVM-GMM", Sydney: The 10th International Conference on Information Technology and Applications (ICITA2015), Jul. 2015.
12. Wetzel, D., Lopez, A. and Webber, B., "A maximum entropy classifier for cross-lingual pronoun prediction", In Proceedings of the Second Workshop on Discourse in Machine Translation, pp. 115-121, 2015.
13. Li, F., Pan, S. J., Jin, O., Yang, Q., and Zhu, X., "Cross-domain co-extraction of sentiment and topic lexicons", In: Proceedings of the 50th annual meeting of the association for computational linguistics long papers, vol. 1, pp. 410–19, July. 2012.
14. Wang, C., and Mahadevan, S., "Heterogeneous domain adaptation using manifold alignment", In: Proceedings of the 22nd international joint conference on artificial intelligence, vol. 2, pp. 541–46, Jun. 2011.
15. Nam, J., Fu, W., Kim, S., Menzies, T. and Tan, L., "Heterogeneous defect prediction", IEEE Transactions on Software Engineering, vol. 44, no.9, pp. 874-896, Sep. 2018.
16. Blitzer, J., Dredze, M., Pereira, F., and et al., "Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification", in ACL, vol. 7, pp. 440–447, 2007.
17. K. Pearson, "X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling", The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, vol. 50, no. 302, pp. 157–175, Jul. 1900.
18. Beecks, C., Uysal, M., and Seidl, T., "Earth mover's distance vs. quadratic form distance: An analytical and empirical comparison", in 2015 IEEE International Symposium on Multimedia (ISM), pp. 233–236, Dec. 2016.
19. Zhuang, F., Cheng, X., Luo, P., Pan, S. J., and Qing He., "Supervised representation learning: Transfer learning with deep auto encoders", In International Joint Conference on Artificial Intelligence, Jun. 2015.
20. Zellinger, W., Grubinger, T., Lughofer, E., Natschl¨ager, T., SamingerPlatz, S., "Central moment discrepancy (cmd) for domain-invariant representation learning", In International Conference on Learning Representations, 2017.
21. Dragoni, M., Tettamanzi, A. G., Pereira, C. D. C., Dranziera: an evaluation protocol for multi-domain opinion mining, in Tenth International Conference on Language Resources and Evaluation (LREC 2016), European Language Resources Association (ELRA), pp. 267–272, May. 2016.
22. University of Michigan, UMICH SI650 - Sentiment Classification, 2011.
23. Kotzias, D., Denil, M., De Freitas, N., and Smyth, P., "From group to individual labels using deep features", in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 597–606, Aug. 2015.
24. 'CrowdFlower, Data for everyone', [Online]. Available: https://www.figure-eight.com/data-for-everyone/. [Accessed: 2017].