دسته بندی اسناد متنی بر اساس ماشین های دارای بردار پشتیبان مربع حداقل با تجزیه مقدار واحد

چکیده

به دلیل رشد سریع اطلاعات آنلاین، طبقه بندی متون به یکی از تکنیک های کلیدی در مدیریت و سازمان‌دهی داده های متنی تبدیل شده است. یکی از دلایل طبقه بندی اسناد، ساده‌تر کردن کشف اسناد مرتبط، پالایش و فیلتر کردن محتوا و پیگیری عناوین است.

LS-SVM یک نوع طبقه بندی است که در این مقاله برای طبقه بندی مناسب اسناد متنی از آن استفاده شده است. داده‌های متنی معمولاً مشخصاتی چند بعدی هستند که کاهش ابعاد آن‌ها را هم می‌توان با SVM انجام داد. در این مقاله، با استفاده از ماشین های دارای بردار پشتیبان مربع حداقل و تجزیه مقدار واحد، دقت طبقه‌بندی را افزایش و ابعاد داده های متنی بزرگ را کاهش می دهیم.

کلمات کلیدی: طبقه بندی متن، ماشین های دارای بردار پشتیبان مربع حداقل، تجزیه مقدار واحد

1. مقدمه

طبقه بندی متن شامل قالب های اصلی اسناد بوده و با قراردهی اسناد در مجموعه عناوین از قبل مشخص شده انجام می‌گیرد. مثالی از این قبیل، برچسب زنی خودکار اسناد ورودی با نام‌هایی چون "ورزش"، "سیاست"، "آموزش" و "تجارت" است. مهم نیست روش خاص مورد نظر از چه چیزی استفاده می کند؛ آغاز طبقه بندی داده کاوی با مجموعه آموزشی $D= (d1, \ldots , dn)$ از اسنادی است که با برچسب دسته $L \in \mathbf{L}$ (مثلاً ورزش، سیاست) مشخص شده اند. پس از آن هدف، تعیین مدل طبقه بندی است.

$$f: D \rightarrow \mathbf{L} f(d) = L \ (5)$$

که قادر به تعیین دسته صحیح سند جدید d از دامنه است.

مجموعه های بزرگی از اسناد رفته رفته رایج می شوند. اولین نکته در استفاده از روش های طبقه‌بندی در اسناد متنی بدون ساختار خاص، ایجاد قالب دارای ساختار به نام مدل فضای برداری است.

در مدل فضای برداری، هر سند را با برداری در داخل فضای برداری نمایش می دهیم. هر بعد بردار متناظر با یک کلمه بوده و مقدار هر جزء برابر تعداد تکرار نسبی آن کلمه در داخل سند است. در نتیجه، ماتریس کلمه-سند m در n به نام A حاصل می شود که m تعداد کلمات منحصر به فرد در مجموعه متون و n تعداد اسناد موجود در آن مجموعه است. در این ماتریس، Ai,j برابر تعداد تکرارiامین کلمه در jامین سند است.

در فرایند ساخت مدل فضایی برداری از روند های پیش پردازش اساسی هم چون حذف کلمات توقفی، قطع ریشه کلمات برای استخراج آن ها از متن، کلمات محتوایی منحصر به فرد یا کلمات کلیدی مجموعه ای از اسناد استفاده می شود. [1] تمام این کلمات کلیدی را می توان در ساخت مدل فضای برداری که به طور ذهنی با برداری از کلمات کلیدی حاصل از اسناد نمایش داده می شود، به کار برد.

مدل فضای برداری از ماتریس کلمه-سند که سطر های آن کلمات و ستون هایش اسناد هستند، تشکیل یافته و در کنار آن ها می توان وزن های اهمیت کلمات کلیدی را در سند و نیز مجموعه کل اسناد مشاهده کرد. [15] بنابراین در مجموعه اسناد بزرگ، هم ابعاد سطر ها و هم ستون ها، بسیار بزرگ و نیز پراکنده هستند. با توجه به اندازه VSM، مهم ترین مشکل در طبقه بندی متن، ابعاد بزرگ است. ابعاد بزرگ، چالشی برای طبقه بندی اسناد است.

راه کاهش ابعاد بزرگ مدل فضای برداری یا ماتریس کلمه-سند ، تجزیه مقدار واحد (SVD) است که در آن ماتریس کلمه-سند به سه ماتریس کوچک تر تجزیه می‌شود. این سه ماتریس را برای کاهش ابعاد در نظر گیرید. از ماتریس سند-سند برای طبقه بندی بیش تر استفاده می کنیم. ماشین دارای بردار پشتیبان، روندی است که در این مقاله برای طبقه بندی مناسب سند از آن استفاده شده است.

در این مقاله، هدف ما کاهش ابعاد ماتریس کلمه-سند به کمک SVD برای بهبود کیفیت طبقه‌بندی است.

قسمت های بعدی این مقاله را به صورت زیر سازمان دهی کرده ایم. در بخش 2 راه کار پیش پردازش را در تحلیل داده متنی مرور می کنیم. بخش 3 راه کار تجزیه مقدار واحد، بخش 4 طبقه‌بندی و بخش های 5 و 6 نتایج آزمایش راه کارمان را پوشش می دهند. بخش 7 درباره نتیجه گیری و تحقیقات آینده است.

2. پیش پردازش

برای دریافت تمام کلمات مورد استفاده در یک متن دلخواه، نیاز به فرایند نشاندار کردن وجود دارد؛ یعنی با حذف تمام علایم نگارشی و تعویض جدول‌بندی و سایر مشخصه ها به جز متن توسط نیم‌فاصله ها سند متنی را به جریانی کلمات تقسیم می‌کنیم. سپس از این نمایش نشان‌دار برای پردازش بیشتر استفاده می کنیم. جهت کاهش ابعاد مجموعه کلمات، سند مورد نظر را می شود با پالایش و قطع ریشه کلمات کوچک تر کرد.

در این بخش روش پیش پردازش ارایه شده را برای ایجاد مدل بهینه فضای برداری معرفی می‌کنیم. روش پیش پردازش ارایه شده منجر به ایجاد بهینه مدل فضای برداری در کم ترین پیچیدگی زمان می شود.

در راه کار پیش پردازش، به جمع آوری تمام کلمات وقفه می پردازیم که معمولاً موجودند. از مقادیر و کد های ASCII هر حرف بدون در نظر گرفتن کوچکی یا بزرگی آن ها استفاده کرده و با جمع کردن مقدار ASCII متناظر هر حرف به تولید یک کلمه می رسیم. به کلمه متناظر یک شماره اختصاص داده و آن ها را مرتب می کنیم.

مثال کلمه "and" را در نظر گیرید که مقادیر ASCII متناظر حرف هایش به ترتیب برابر a=97، n=111 و d=101 است. در نتیجه مقدار کلی برای کلمه "and" برابر 309 است. به طور مشابه برای کلمه "to" برابر 127+122=249 است. اما در این راه کار احتمال دارد جمع اسکی مقادیر دو کلمه همان طور که در زیر آمده با هم برابر باشند؛ مثلاً در کلمه "ask" برابر 319=107+115+97 و در کلمه "her" برابر =111+101+104 319 است.

راه حل مشکل فوق این است که در حالت مقایسه، می توان مقایسه را در جمع مقادیر اسکی انجام داده و در آرایه متناظر، رشته ای از کلمات وقفه را در نظر گرفت. بنا براین می‌توان با استفاده از این رشته مقایسه را انجام داده و مطمئن شویم هیچ کلمه ای از بین نمی رود. در ضمن باید زیر مجموعه‌ای از رشته ها را ایجاد کنیم که دارای جمع مقادیر اسکی و فقط برای مقایسه با آن زیر مجموعه کافی باشند.

برای جستجوی مقادیر ASCII از تک حروف هایی استفاده کردیم که در روش جستجوی هم پوشانی کلمات برای یافتن سریع مقدار متناظر کاربرد داشتند.

ارایه فوق از الگوریتم قطع ریشه کلمات حاملی استفاده می‌کند که در قطع ریشه کلمات برای پردازش بهتر سند کاربرد دارد. قطه کننده ریشه کلمات حامل به پنج مرحله تقسیم می شود که مرحله 1 پسوند های i و مراحل 2 تا 4 پسوند های d را حذف می کنند. پسوند های مرکب d به پسوند های واحد d در هر لحظه تبدیل می شوند.

بنابراین مثلاً اگر کلمه ای با icational پایان یابد، مرحله 2 آن را به icate و مرحله 3 به ic کاهش می دهد. در انگلیسی سه مرحله اول لازم و ضروری هستنند. مرحله 5 کار پیچیده تری انجام می دهد.

2.1 روش انتخاب کلمات کلیدی

پس از پالایش و قطع ریشه کلمات، باید تعداد کلمات را باز هم کاهش داد و برای این کار از شاخص‌دهی و یا روند انتخاب کلمات کلیدی می توان استفاده کرد. در این حالت، فقط از کلمات کلیدی برای توصیف سند استفاده می شود. [1] یک روش ساده برای انتخاب کلمات کلیدی، استخراج کلمات از متن بر اساس آنتروپی و تعداد تکرار آن هاست. مثلاً برای هر کلمه t در واژگان کلی، می‌توان آنتروپی را به صورت زیر حساب کرد:

$$w(t) = 1 + \frac{1}{log_2|D|}\sum_{d\in D} p(d,t) log_2\ P(d,t) - (1)$$

$$\text{where } P(d,t) = \frac{tf(d,t)}{\sum_{i=1}^{n} tf(d,t)}$$

معادله (1) آنتروپی کلمه ای دلخواه را به دست می دهد. در این جا آنتروپی، سنجشی از نحوه جای‌گیری یک کلمه در اسناد مختلف با کمک جستجوی کلمات است. برای مثال، کلماتی که در اسناد مرتباً تکرار شوند، آنتروپی پایین تری خواهند داشت. آنتروپی را می توان سنجشی از اهمیت یک کلمه در محتوای سند مورد نظر دانست. [3] کلمات شاخص، تعدادی کلمات هستند که نسبت به تعداد تکرار کلی خود با آنتروپی بالا انتخاب می شوند و در کلمات با تکرار برابر، آن‌هایی که آنتروپی بالاتری دارند در مدل فضای برداری کاربرد خواهند داشت.

2.2 مدل فضای برداری

مدل فضای برداری، مدلی استاندارد در نمایش اسناد در بازیابی اطلاعات است. ایده اساسی آن نمایش هر سند به عنوان برداری از تکرار کلمات کلیدی مشخص است. مدل فضای برداری، اسناد را با بردار هایی در فضای m بعدی نمایش می دهد؛ به این معنی که هر سند d با بردار مشخص عددی (w(d)=(x(d,t1), . . . , x(d,tm نشان داده می شود.

کار اصلی نمایش فضای برداری اسناد یافتنروش کد کردن مناسب برای بردار مشخص شده است. هر عنصر بردار معمولاً یک کلمه (یا گروهی از کلمات) از مجموعه اسناد را نشان می دهد؛ به این معنی که اندازه بردار توسط تعداد کلمه (یا گروه کلمات) از مجموعه اسناد کامل تعریف می شود. ساده ترین راه کد کردن اسناد استفاده از

بردار های کلمه دودویی است؛ یعنی اگر کلمه متناظر با عناصر بردار در سند استفاده شود، این عناصر برابر یک و در غیر این صورت برابر صفر خواهند بود.

مدل فضای برداری بسیار کاربرد دارد چون نمایشی مناسب و مقداری از هر سند است.VSM شامل تعداد دفعات تکرار کلمه j در سند iراf_{ji}و تعداد اسنادی را که دارای کلمه j هستند d_j می نامد. راه کار معمولی از این روش حل $f_{ji}Xd_j$ استفاده می کند. با توجه به این شمارش ها، [9] می توان سند iام را با بردار w بعدی x_i به صورت زیر نشان داد. برای $1\leq j\leq w$جزء jام از x_iرا حاصل ضرب سه عبارت زیر قرار دهید:

$$x_{ji} = t_{ji} \cdot g_j \cdot s_i \text{-- (2)}$$

که t_{ji}جزء وزنی کلمه بوده و فقط به f_{ji} بستگی دارد درحالی که g_j جزء وزنی کلی بوده و به d_j وابسته و s_i جزء نرمالیزه کردن در x_i است. t_{ji} اهمیت نسبی کلمه را در سند در اختیار داشته و g_jاهمیت کلی کلمه را در کل مجموعه اسناد بیان می کند.

هدف این نوع قالب ها و طرح های وزنی، افزایش تمایز بین بردار اسناد مختلف برای مناسب بودن بازیابی است.

در این مقاله ازکلمه تکرار سند *معکوس تکرار* استفاده می کنیم. این قالب از موارد زیر استفاده می‌کند:

$$t_{ji} = f_{ji}, \quad g_j = \log(d/d_j) \text{ and}$$

$$s_i = \sum^{w}_{j=1}(t_{ji}\,(t_{ji}g_j)^2)^{-\frac{1}{2}}. \text{ –(3)}$$

توجه کنید که در این نرمالیزه کردن، $||x_i||=1$است؛ یعنی هر بردار سند روی سطح کره واحد R^w قرار دارد. هدف نرمالیزه کردن، مقایسه کلمات تکرار شده در سند است.

بر این اساس می توان اطمینان یافت که اسناد دارای مطالب با موضوع مشابه (با کلمات مشابه) هستند اما در طول بردار سند با هم تفاوت دارند.

3. تجزیه مقدار واحد

اسناد با مدل فضای برداری نمایش داده می شوند. اما مشاهده کردیم که VSM پراکنده و دارای ابعاد بزرگی است. منبع [4] داده با ابعاد بزرگ نتایج خوبی را در گروه بندی ارایه نمی کند. برای کاهش ابعاد بزرگ می‌توان از راه کار تجزیه مقدار واحد استفاده کرد.

[6] تجزیه مقدار واحد روشی ریاضی است که بیان می کند ماتریس مستطیلی $A_{m\times n}$ ماتریس سند-کلمه ای با مقادیر ورودی مثبت و واقعی است. تجزیه مقدار واحد با درجه کاهش یافته در ماتریس انجام می شود تا الگوی روابط بین کلمه و مفاهیم داخل متن مشخص شود.

3.1 روش تجزیه به کمک SVD

در این روش ماتریس سند-کلمه $A_{m\times n}$ به سه ماتریس با درجه کم تر و به روش زیر تجزیه می شود.

محاسبه U :

مرحله اول: $A.A^T$را حساب کنید.

مرحله دوم: مقادیر مشخصه و ویژه AA^T به دست می آیند.

مرحله سوم: بردار های مشخصه برای مقادیر مشخصه متناظر به دست می آیند.

مرحله چهارم: این بردار های مشخصه را در ماتریس قرار دهید.این ماتریس U نام دارد.

محاسبه V :

مرحله اول: $A.A^T$را حساب کنید.

مرحله دوم: مقادیر مشخصه و ویژه AA^T به دست می آیند.

مرحله سوم: بردار های مشخصه برای مقادیر مشخصه متناظر به دست می آیند.

مرحله چهارم: این بردار های مشخصه را در ماتریس قرار دهید.این ماتریس V نام دارد.

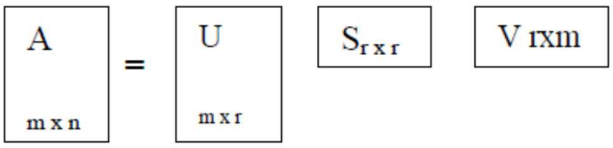
محاسبه S :

مرحله اول: مقدار مؤثر (RMS) را برای مقادیر مشخصه AA^Tیا A^TA محاسبه کنید.

مرحله دوم: این مقادیر را به صورت قطری و نزولی مرتب کنید. مقادیر باقی مانده را صفر کنید. ماتریس S هم حاصل می شود.

[10] پس از انجام محاسبات فوق می توان ماتریس سند-کلمه A را به سه ماتریس USV^T با درجه کم تر تجزیه کرد که $U^TU=I$ و $V^TV=I$ است؛ ستون های U مقادیر مشخصه عمودی AA^T ، ستون های V بردار های مشخصه عمودی A^TA ، و S ماتریس قطری شامل مقادیر RMS بردار های مشخصه از U تا V و به ترتیب نزولی است. پس

از تجزیه ماتریس سند-کلمه A که U ماتریس کلمه $m\times m$ ، S ماتریس قطری $m\times n$ و V^T ماتریس ترانهاده V با ابعاد $n\times n$ و شامل بردار های سند در سطرهایش خواهند بود که در شکل زیر به نمایش درآمده اند.



شکل 1. تجزیه ماتریس A به USD^{-1} با استفاده از SVD

در دنیای واقعی، داده ابعاد بزرگی دارد؛ یعنی ماتریس سند-کلمه A تعداد زیادی سطر و ستون دارد. از SVD در تشکیل ماتریسی با درجه کم استفاده می شود. در فرایند SVD مربوط به کاهش ابعاد، موارد مشابه شبیه‌تر و موارد نامشابه، متفاوت تر می شوند. طبقه بندی به کمک ماشین دارای بردار پشتیبان فقط بر داده‌های کوچک اعمال شده و به طور چشم گیری باعث کاهش سختی محاسبات می شود.

4. طبقه بندی

طبقه بندی سند زمینه‌ای است که با گروه بندی نظارت شده اسناد متنی و انتقال آن ها به گروه های مناسب‌تر سروکار دارد که معمولاً نماینده عناوین موجود در مجموعه اسناد هستند. [7] طبقه بندی اسناد کاربردهای بسیاری دارد؛ از جمله خوشه‌بندی و دسته بندی نتایج حاصل از موتور های جستجو، تحلیل داده، پالایش و فیلتر کردن ایمیل، هدایت نامه ها و غیره.

از لحاظ فکری، هر یک از الگوریتم های طبقه بندی دارای حالتی آزمایشی ، [8] که در آن پروفایل دسته‌ها را می توان از نمونه های بسیار بزرگ اسناد به همراه طبقه بندی آن ها (مجموعه آموزشی) آموخت و حالتی کاربردی هستند که در آن از پروفایل دسته ها برای تعیین دسته های محتمل و معمول تر اسناد استفاده می‌شود.

با وجود مجموعه اسناد D و مجموعه دسته های C ، یک طبقه بند برای دسته c برابر تابع $f_i= D = \{0,1\}$ است که تابع نامشخص $f_i:D\rightarrow\{0,1]$ را تخمین می زند که این تابع هم بیانگر ارتباط اسناد دسته c_iاست. [13] با در دست داشتن دسته های مجموعه اسناد C و سند نمونه ای برای هر دسته، طبقه بندی بسازید که با وجود سند d دسته (های) مشابه با سند d را پیدا کند.

4.1 ماشین های بردار پشتیبان

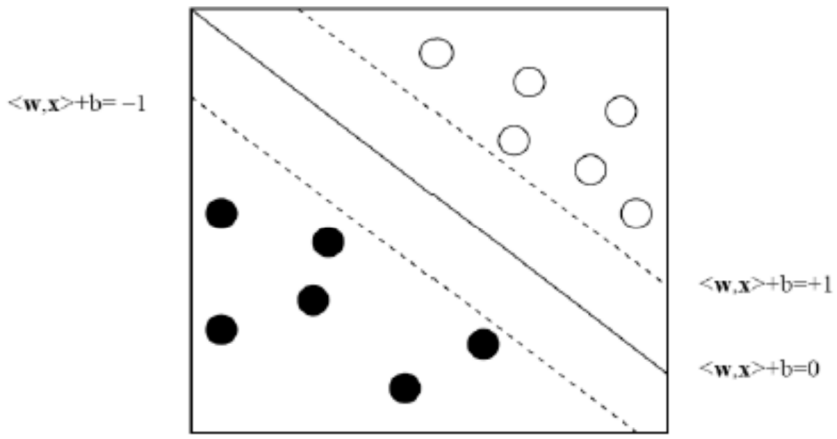
ماشین بردار تکنیک یادگیری نظارت شده ای در زمینه یدگیری ماشینی است که هم در طبقه بندی و رگرسیون کاربرد دارد. [4] SVM استاندارد مجموعه ای از داده های ورودی داشته و دو دسته ممکن برای هر داده مورد نظر را پیش بینی می کند.

[14] ماشین بردار پشتیبان سیستمی در یادگیری است که از فضای فرضی تابع خطی در فضای مشخص با ابعاد بزرگ استفاده کرده و آموزش آن توسط الگوریتم یادگیری حاصل از نظریه بهینه سازی انجام می شود که بیانگر این است که یادگیری اولیه از نظریه یادگیری آماری حاصل می گردد.

با وجود مثال های آموزشی $\{x_i,y_i\}$ که $i=1,\ldots,l$بوده و هر مثال دارای d ورودی ($x_i \varepsilon R^d$) و برچسب دسته با یکی از دو مقدار ($y_i \in \{-1,1\}$)است. [5] حال تمام ابرصفحات در R^d با استفاده از بردار (w) و ثابت (b) که در معادله آمده اند، پارامتری می شوند.

$$w.x+b=0 \quad (4)$$

که w و b پارامتر های مدل هستند.



شکل 2. طبقه بندی خطی قابل جداسازی

شکل 2 مجموعه آموزشی دو بعدی را شامل مربع های پر رنگ و کم رنگ نشان می دهد.یک محدوده تصمیم‌گیری که مثال های آموزشی را به دسته های مربوط به خود تقسیم می کند با خط پر رنگ به نمایش در آمده است. هر نمونه از محدوده تصمیم گیری باید در معادله (4) صدق کند.

[12] یک مسئله طبقه بندی دو دسته ای را به عنوان اطلاعات منحصر به فرد درباره دسته های نمایش داده شده با ترتیبی از داده های برچسب دار محدود در نظر گیرید:

$$S=\{ (x,y)/x_i=(x_i^{(1)},.....x_i^{(d)})^T \varepsilon R^d,$$

$$y_i=\{-1,1\}, i=1,N\}\text{-----}(5)$$

اولین جزء هر زوج (x_i,y_i) از S نمایشگر مثالی از دسته با برچسب y_i است.

4.1.1 حالت خطی قابل جداسازی

ترتیبی به صورت خطی قابلیت جداسازی دارد که تابع جداساز خطی $f:R^d \rightarrow R$ موجود باشد که نمونه های S را از هم جدا کند؛ به این صورت:

$$f(x)=b+w1x^{(1)}+.....+wdx^{(d)} \text{ -----}(6)$$

در هر $x= (x^{(1)}, . . . , x^{(d)}) \varepsilon R$، برای هر $(x_i,y_i) \varepsilon S$ و اگر $y_i=1$ آنگاه $f(x_i)> 0$ و اگر $y_i=1$ باشد آنگاه $f(x_i)<0$خواهد بود.

قابلیت جداسازی خطی را می‌توان به صورت وجود پارامترهای $b \varepsilon R$ و $w \varepsilon R$ نشان داد به طوری که $w^Tzi+b>0$ ، $Z_i=y_ix_i$ و $w=(w1.....wd)$ است در این حالت می گوییم ابرصفحه

$$H_{wb}: w^Tx+b=0\text{------}(7)$$

بدون خطاهای S قابل جداسازی است.

4.2 SVM مربع حداقل

[2] ماشین های دارای بردار پشتیبان مربع حداقل (LS-SVM) نوع دیگری از SVM استاندارد هستند. تابع هزینه، تابع مربعات حداقل منظم شده با محدودیت هایی در تساوی است که به سیستم های خطی Karush Kuhn-Tucker منتهی می شود. [17] با استفاده از روش های تکرارهم چون الگوریتم گرادیان مزدوج (CG) می توان راه حل را به خوبی یافت.LS-SVM ها ارتباط نزدیکی با شبکه های منظم ، فرایند های گوسی و تحلیل جداساز kernel-fisher دارند اما به آن ها افزوده و از درک اولیه-دوگانه ای استفاده می‌کنند.

به ارتباط بین نسخه های kernel الگوریتم های تشخیص الگوی قدیمی و توسعه به شبکه های مکرر، کنترل و مدل سازی قوی هم اشاره شده است. از چهارچوب مدرکی Bayesian با سه سطح از تداخل استفاده شده و امکان درک های احتمالی، انتخاب ابر پارامتر ها، مقایسه مدل ها و انتخاب ورودی هم فراهم شده است.

5. راه کار ما

بزرگی ابعاد مدل فضای برداری برای داده متنی باعث سختی محاسبات و نتایج طبقه بندی نامناسب می شود. در این تحقیق دریافتیم که کاهش ابعاد بزرگ با استفاده از تجزیه مقدار واحدو افزایش کیفیت گروه بندی سند متنی با کمک ماشین های بردار پشتیبان امکان پذیر است.

سیستم کلی طبقه بندی براساس LS-SVM است. راه کارمان را در چند مرحله زیر شرح می دهیم.

الف: اسناد آموزشی را با بردار سند مدل سازی کنید.

• اول TOKENIZATION ؛ در این فرایند، اسناد متنی توسط حذف علایم نگارشی و تعویض جدول بندی‌ها و سایر مشخصه های غیر متنی با نیم فاصله ، به جریانی از کلمات تقسیم می شوند.

• تمام کلمات وقفه معمول را حذف کنید. ما از راه کار ASCII محور، مقادیر ASCII هر حرف بدون در نظرگیری کوچک یا بزرگی آن ها، و جمع مقادیر ASCII متناظر هر کلمه برای تولید یک عدد استفاده کردیم. عددی را به کلمه متناظر اختصاص داده و آن ها را مرتب کنید.

• از کلمات کلیدی و تعداد تکرار آن ها برای تشکیل ماتریس کلمه – سند A استفاده کنید.

•SVD را در ماتریس A اعمال کرده و به شکل تجزیه شده و کوچک شده ماتریس سند برسید.

ب: بردارهای سند را داخل سیستم طبقه بندی SVM قرار دهید. پارامترهای SVM تنظیم کرده و تابع اساسی و ریشه ای (RBF) را به عنوان تابع kernel در SVM انتخاب کنید:

$$S(x, xi) = \exp\left\{\frac{-\|x-xi\|^2}{-2}\right\} \text{---(8)}$$

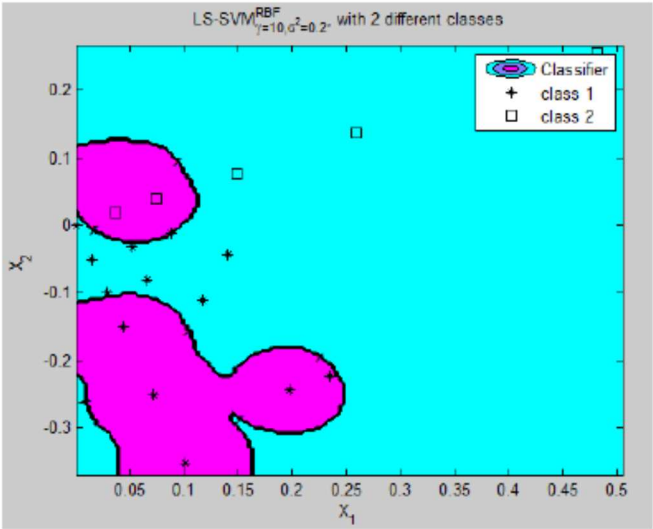
ج: از سیستم طبقه بندی حاصل از مرحله فوق که اسناد آن متعلق به دسته مربوط هستند استفاده کنید.

6. نتایج آزمایشی

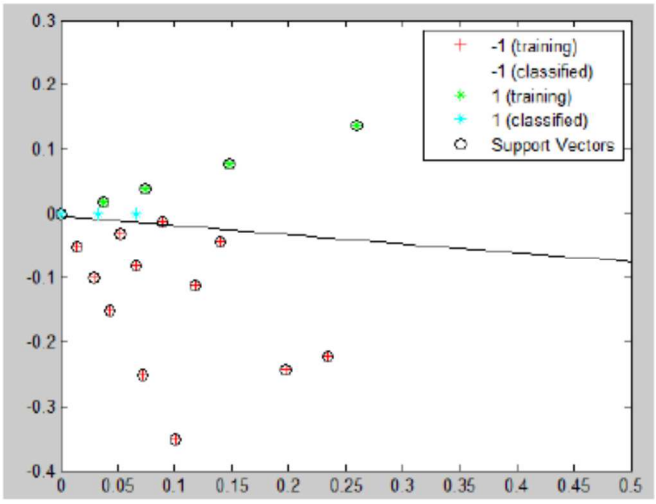
در آزمایشات از اسناد استاندارد استفاده کرده ایم که در کل شامل 15809 کلمه هستند. این کلمات را پس از حذف کلمات وقفه به 1840 کلمه ، و پس از قطع ریشه کلمات باقی مانده به 323 کلمه کاهش دادیم.

در این بخش نتایج آزمایشات خود را نمایش می دهیم. حاصل چندین آزمایشمان، پیش پردازش سند، کاهش ابعاد و طبقه بندی آن بود. برای آزمایشات از اسناد استاندارد استفاده کردیم.

پس از پیش پردازش ازتجزیه مقدار واحد استفاده کردیم.SVD ماتریس کلمه – سند را به سه ماتریس تجزیه کرد. شکل های 3 و 4 به ترتیب ، نتایج طبقه بندی اسناد آموزشی مورد نظر و طبقه بندی را نمایش می دهند.



شکل 3. طبقه بندی دو سند مختلف با استفاده از LS-SVM



شکل 4. آموزش و طبقه بندی اسناد به کمک LS-SVM

7. نتیجه گیری و تحقیقات آینده

در این مقاله از قطع ریشه کلمات به روش پیش پردازش استفاده کردیم که بر اساس ASCII بود تا کلمات وقفه را حذف کرده و کلمات کلیدی را از افعال و اسامی داخل سند بیابیم. در یافتن کلمات کلیدی از راه کار آنتروپی محور استفاده کردیم که بهترین راه برای کشف کلمات کلیدی در اسناد ورودی بود. هم چنین برای کاهش ابعاد ماتریس کلمه – سند ورودی، SVD را به کار بستیم.

این مقاله الگوریتم جدیدی را به نام LS-SVM معرفی می‌کند که ترکیبی از مزایا LSI و SVM را با هم دارد. نتایج آزمایشات هم تأیید می کنند که LS-SVM روشی بسیار کاربردی و مؤثر در طبقه بندی اسناد است. در تحقیقات آینده تمرکز خود را بر افزایش مناسب بودن و انطاف پذیری قالب های پیش پردازشی و طبقه بندی خود در اسناد دارای چند قالب ادامه خواهیم داد.