ارائه‌ی یک مدل جهت دسته‌بندی متون فارسی با استفاده از ترکیب روش‌های دسته‌بندی

ایمان جمالی[۱]، سید جواد میرعابدینی[۲*] ، علی هارون آبادی[۳]

۱: گروه کامپیوتر، دانشگاه آزاد اسلامی واحد علوم و تحقیقات بوشهر، بوشهر، ایران imanjamali25@gmail.com
۲*: عضو هیئت علمی دانشگاه آزاد اسلامی واحد تهران مرکزی، تهران، ایران jvd.2205@yahoo.com
۳: عضو هیئت علمی دانشگاه آزاد اسلامی واحد تهران مرکزی، تهران، ایران a.harounabadi@gmail.com

چکیده

برای دسته‌بندی متن از تکنیک‌های استخراج اطلاعات، پردازش زبان طبیعی و یادگیری ماشین به طور وسیع استفاده می شود به طور کلی هدف یک دسته بند متون، دسته‌بندی اسناد در قالب تعداد معینی از دسته‌های از پیش تعیین شده می‌باشد. هر سند می تواند در یک، چند و یا هیچ دسته ای قرار بگیرد. در مورد هر سند به این سؤال پاسخ داده خواهد شد که این سند در کدام یک از دسته‌ها قرار می گیرد. این موضوع می تواند در قالب یک یادگیری خودکار قرار گیرد تا با استفاده از آن بتوان هر سند را به طور خودکار به دسته‌ای نسبت داد. در این مقاله، بعد از انتخاب مجموعه داده و پاک‌سازی متون به کمک روش نرمال شده فرکانس کلمه- معکوس فرکانس سند (norm TF-IDF) به ویژگی‌ها وزن داده می‌شود و در طی دو مرحله ویژگی‌ها با استفاده از روش‌های فرکانس سند (DF) و مربع چی(SChi) انتخاب می‌شوند و بعد با استفاده از روش تحلیل مؤلفه اصلی (PCA) ابعاد ویژگی‌ها کاهش داده می‌شود و در مرحله بعد با استفاده از ترکیب ۲۱ ماشین بردار پشتیبان (SVM) به پیاده سازی مدل پیشنهادی می‌پردازیم و در نهایت صحت مدل را با روش اعتبار سنجی ۱۰ مرحله‌ای ارزیابی می‌کنیم نتایج تجربی نشان می‌دهد که این مدل می‌تواند عمل دسته‌بندی متون را برای هفت دسته با صحت ۹۱,۸۶ انجام دهد که نسبت به کارهای پیشین انجام گرفته صحت بالاتری دارد.

واژه های کلیدی: ماشین بردار پشتیبان، دسته‌بندی متون، انتخاب ویژگی، تحلیل مولفه اصلی.

۱- مقدمه

بسیاری از اطلاعات موجود به صورت متن، مستندات الکترونیکی و دیگر صورت‌های متنی دیجیتال ذخیره شده‌اند. چنین اسنادی دارای اطلاعات زیادی هستند که به آسانی قابل دسترسی نیستند. اگر از دید یک کارشناس کامپیوتر به این اطلاعات نگاه کنیم، طبیعت همه آنها بدون ساختار است، جستجو در مجموعه سندهای بزرگ نیازمند سازماندهی مناسب است، برای استخراج دانش از این اطلاعات باید ابتدا آن را درک کرده و سپس پردازش نمود تا بتوان معانی و مفاهیم داخل آن را درک کرد و ارتباط میان مفاهیم را کشف نمود. در عصر حاضر که عصر اطلاعات نامیده می شود، با وجود این حجم بالای اطلاعات، پردازش به صورت دستی، کاری پر هزینه، زمان گیر و تقریبا نشدنی می باشد. متن کاوی عنوانی است که برای انجام این فرآیند خودکار انتخاب شده است. امروزه با توجه به گسترش روزافزون متون نیاز به متن کاوی بیش از هر زمان دیگری احساس می شود. از جمله کاربردهای متن کاوی می توان به رده بندی، خوشه بندی، خلاصه سازی و یافتن روابط میان مفاهیم در متون اشاره کرد. مسئله رده بندی متون را می توان به عنوان مهمترین کاربرد متن کاوی دانست. رده بندی متون نیزدارای کاربردهای زیادی است، از جمله این کاربردها، رده بندی صفحات وب به صورت سلسله مراتبی، رده بندی مقالات روزنامه ها به وسیله موضوع، رده بندی و کدگذاری سوابق ذخیره شده بیماران، رده‌بندی و فیلتر کردن E-Mail ها و رده‌بندی رخدادهای خبری بر اساس موضوع هستند[۱].

بخش دوم مقاله به ارائه کارهایی می پردازد که تاکنون در زمینه دسته‌بندی متون فارسی انجام شده است. در بخش سوم، به ارائه مدل پیشنهادی و پیاده سازی آن با نرم افزار Rapid Miner5.3 و ارزیابی نتایج آن می پردازد. در بخش چهارم مقاله،به مقایسه مدل

پیشنهادی با کارهای پیشین از لحاظ دقت می پردازد و در نهایت در بخش آخر مقاله به نتیجه گیری حاصل از این مقاله پرداخته می شود .

۲- کارهای پیشین

سرایی و شاهقلیان، روش پیشنهادی برای دسته‌بندی متون فارسی بر اساس روش یادگیری ماشین ارائه دادند. آزمایش‌های انجام شده توسط ایشان دقت ۸۰,۸۵٪ را برای هفت کلاس متون فارسی نشان می‌دهد[۲]. الهی‌منش و مینایی، روشی با استفاده از روش رده‌بندی K نزدیک‌ترین همسایه و دو معیار فاصله متون، آزمایش‌های خود را انجام دادند. یکی از این دو معیار، الگو گرفته از نوعی رده‌بندی متون زبان عربی بوده و دیگری، معیار ترکیبی تولید شده خود است. نتایج نشان می‌دهد که این روش می‌تواند با دقت ۸۹٪ عمل دسته‌بندی را برای هفت دسته انجام دهد[۳]. مهدی پور و همکارانش، پژوهشی به نام سیستم خلاصه ساز خودکار متن فارسی با استفاده از الگوریتم ترکیبی SA-GA را ارائه کرده اند که این سیستم پس از ریشه یابی کلمات با استفاده از ترکیب روش های مبتنی بر گراف و TF-IDF با ترکیب الگوریتم SA-GA عمل می کند نتایج حاصل از ارزیابی متن خلاصه شده توسط سیستم ققنوس ۶۴,۳۵٪ برای شش دسته بوده است[۴]. هاشمی و همکارانش، پژوهشی به نام استفاده از تکنیک های متن کاوی برای دسته بندی متون فارسی با مجموعه داده همشهری انجام داده اند که با به کارگیری الگوریتم های یادگیری بیز کارایی دسته بندی را با مجموعه داده همشهری با ۷۰ درصد آموزش و ۳۰ درصد آزمایش مورد بررسی قرار داده اند نتایج کارایی به ۸۵,۴۸ درصد و خطای ۱۴,۵۲ درصد با ۸ دسته بوده است[۵]. ابراهیمی و همکارانش، پژوهشی به نام رده بندی اسناد و متون فارسی با استفاده از ماشین بردار پشتیبان SVM انجام داده اند. در این مقاله ابتدا با استفاده از روش TFCRF به کلمات وزن داده می شود و سپس از روش های PCAو الگوریتم ژنتیک استفاده شده است. نتایج تجربی نشان می دهد، که این روش می تواند با دقت ۸۹,۸۴٪ عمل دسته بندی را برای پنج دسته انجام دهد[۱]. باقری و همکارانش، پژوهشی به نام ارائه یک روش انتخاب ویژگی ترکیبی برای دسته بندی متون فارسی به نام PSA ارائه داده اند که در پژوهش خود با استفاده از روش پیشنهادی انتخاب ویژگی و روش بیز به دسته بندی هفت دسته خبری با روش اعتبار سنجی پنج مرحله ای پرداخته است که توانسته روش خود را با دقت ۸۸,۲ ٪ پیاده سازی کنند[۱۰].

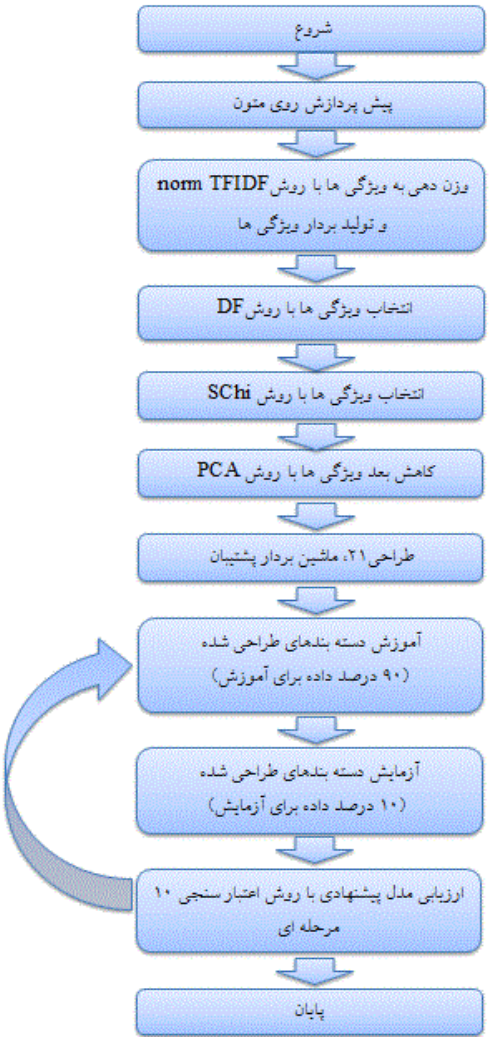
۳- مدل پیشنهادی

در این قسمت، مدلی جهت دسته‌بندی متون فارسی ارائه می‌شود. و سپس به ارزیابی کارایی آن پرداخته می شود. مراحل انجام مدل پیشنهادی بصورت شکل ۱ می باشد.

۳-۱- انتخاب مجموعه داده و پیش پردازش روی متون

در این مقاله از ۱۴۰۰ فایل متنی در قالب ۷ دسته از مجموعه داده همشهری، برای فرایند دسته بندی متون استفاده شده است. مجموعه داده همشهری با خزش وب سایت همشهری و چندین مرحله پیش پردازش و برچسب گذاری در قالب ۳۱۸ هزار سند از سال ۱۳۷۵ تا ۱۳۸۶ حاصل آمده است. نسخه ۱ این مجموعه نمونه ای است که در همایشهای CLEF در سال های ۲۰۰۸ و ۲۰۰۹، برای ارزیابی سامانه های بازیابی اطلاعات تک منظوره مورد استفاده قرار گرفته است. نسخه ۲، آخرین نسخه مجموعه است که نسبت به نسخه ۱ بزرگتر و جامعتر می باشد[۱۱]. لیست دسته‌ها در جدول ۱ قابل مشاهده است.

بعد از انتخاب مجموعه داده به پردازش متون می پردازیم که قبل از اعمال فیلترهای پردازش متون (حذف کلمات با طول کمتر از سه حرف ، تبدیل «ی» و «ک» عربی به «ی» و «ک» فارسی، حذف حائل واژگان، حذف نویسه‌های نالازم، حذف کلمات با محتوای غیرالفبایی فارسی، حذف کلمات بی ارزش فارسی) به ۲۲۱۹۴ ویژگی دست می یابیم که این تعداد ویژگی بعد از اعمال فیلترهای پردازش متون، به ۲۰۳۶۹ عدد می رسد.



شکل ۱: مراحل مدل پیشنهادی برای دسته بندی

جدول ۱: لیست دسته ها

ردیف	نام دسته	تعداد سند
۱	اجتماعی	۲۰۰
۲	ادب و هنر	۲۰۰
۳	اقتصاد	۲۰۰
۴	سیاسی	۲۰۰
۵	علمی فرهنگی.علمی.ارتباطات و فناوری اطلاعات	۲۰۰
۶	گوناگون.حوادث	۲۰۰
۷	ورزش	۲۰۰

۳-۲- روش وزن دهی به ویژگی‌ها در مدل پیشنهادی

۳-۲-۱-روش norm TFIDF

برای اطمینان از اینکه همه مستندات با طول های مختلف شانس برابری برای بازیابی شدن داشته باشند روش TFIDF به صورت نرمال، و به صورت رابطه ۱ ارائه می شود[۶].

$$w_{ki} = \text{norm TFIDF}(t_k , d_i) = \frac{tfidf(t_k , d_i)}{\sqrt{\sum_k (tfidf(t_k , d_i))^2}} \quad (۱)$$

که در رابطه ۱،TFIDF با توجه به رابطه ۲ به دست می آید.

$$TFIDF(t_k , d_i) = TF(t_k , d_i) * IDF(t_k , d_i) = TF(t_k , d_i) * \log(|D|/|D(t_k)|) \quad (۲)$$

که در آن $TF(t_k , d_i)$ برابر تکرار کلمه در یک سند می باشد و D تعداد کل مستندات مجموعه و $D(t_k)$ تعداد مستنداتی از مجموعه D می باشد که ویژگی t_k در آنها حداقل یک بار رخ داده باشد.

۳-۳- روش های انتخاب ویژگی‌ها در مدل پیشنهادی

۳-۳-۱- انتخاب ویژگی ها با روش DF

فرکانس سند برای هر کلمه، برابر با تعداد سندهائی است که کلمه در آن ظاهر شده است .برای هر کلمه واحد در مجموعه سندهای آموزش این ضابطه محاسبه شده است. این روش، روشی مقیاس پذیر برای حتی مجموعه داده های بسیار بزرگ می باشد و دارای پیچیدگی خطی متناسب با تعداد سندهای آموزش است[۷].

در این مقاله، ما فرض را بر این نهادیم که کلماتی که در حداکثر دو درصد سندها ظاهر شده اند کلمات خوبی نیستند و نمی توانند به عنوان یک ویژگی خوب به حساب آیند و همچنین کلماتی که در بیش از ۳۳ درصد سندها ظاهر شوند. پس ما یک حدی بین ۳ تا ۳۳ درصد را برای انتخاب ویژگی ها انتخاب کردیم که ما توانستیم ویژگی ها را در این مرحله به ۴۵۹ عدد برسانیم و بالاترین دقت مدل را در همین حد بدست آوریم.

۳-۳-۲- انتخاب ویژگی ها با روش SChi

ضابطه روش SChi همان ضابطه ساده شده روش Chi برای محاسبات Chi است که از جذر ضابطه روش خود استفاده می کند. بعبارتی نقش همبستگی مثبت میان همبستگی سند را اهمیت بیشتری می دهد و کلمه منفی را اهمیت کمتری می دهد، ضابطه ساده شده آن به صورت زیر است[۷].

$$SChi(t , c) = P(t , c) * P(t^- , c^-) - P(t^- , c) * P(t , c^-) \quad (۳)$$

$P(t, c)$ نشان دهنده احتمال آنست که در یک سند X ، ویژگی t ظاهر شود و سند متعلق به دسته c نیز باشد. $P(t^- , c^-)$ نشان دهنده احتمال آنست که در یک سند X ، ویژگی t ظاهر نشود و سند متعلق به دسته c نیز نباشد. $P(t^- , c)$ نشان دهنده احتمال آنست که در یک سند X ، ویژگی t ظاهر نشود و سند متعلق به دسته c نیز باشد. $P(t , c^-)$ نشان دهنده احتمال آنست که در یک سند X ، ویژگی t ظاهر شود و سند متعلق به دسته c نیز نباشد. برای هر کلمه t در هر دسته c این مقدار محاسبه شده ، و نهایتا ماکزیمم و یا میانگین آن مقادیر، بعنوان ضابطه Chi آن کلمه منظور می شود[۷].

ویژگی های خروجی از روش DF را به روش SChi دادیم، در این روش ما ویژگی هایی که وزن آنها بیشتر از ۰،۰۴ بود به عنوان ویژگی های جدید انتخاب کردیم و توانستیم ویژگی ها را به ۳۶۷ عدد برسانیم و بالاترین دقت مدل را بدست آوریم.

۳-۴- روش کاهش بعد ویژگی‌ها در مدل پیشنهادی

۳-۴-۱- کاهش بعد ویژگی ها با روش PCA

تکنیک PCA بهترین روش برای کاهش ابعاد داده به صورت خطی می باشد . یعنی با حذف ضرایب کم اهمیت بدست آمده از این تبدیل ، اطلاعات از دست رفته نسبت به روش های دیگر کمتر است ، در این روش محورهای مختصات جدیدی برای داده ها تعریف شده و داده ها بر اساس این محورهای مختصات جدید بیان می شوند . اولین محور باید در جهتی قرار گیرد که واریانس داده ها ماکسیمم شود (یعنی در جهتی که پراکندگی داده ها بیشتر است). دومین محور باید عمود بر محور اول به گونه ای قرار گیرد که واریانس داده ها ماکسیمم شود . به همین ترتیب محورهای بعدی عمود بر تمامی محورهای قبلی به گونه قرار می گیرند که داده ها در

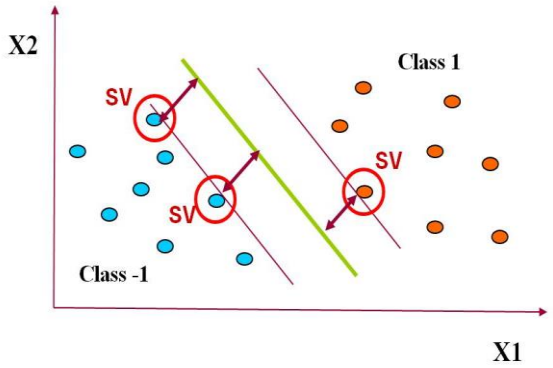
آن جهت دارای بیشترین پراکندگی باشند . این روش به عنوان تصویر متعامد داده ها درون یک فضای خطی بـا ابعـاد کمتـر اسـت بـه قسمی که واریانس بین داده های تصویر شده ، حداکثر شود ، که در این روش اگر ما N نمونه با ابعاد D داشته باشیم داده ها به فضای با ابعادM ، M<D به صورتی نگاشت داده می شود که در آن واریانس بین داده ها جدید حداکثر است. که اساس کار آن بر اساس مقدار ویژگی می باشد که بردارهای ویژگی بر اساس مقدار ویژگی مرتب شده اند[۸].

در این مقاله ما تعداد ویژگی را به صورت آزمون و خطا، بالا و پایین کردیم و توانستیم ویژگی ها را به ۷۵ عدد برسانیم و در این حالت مدل پیشنهادی بهترین دقت را داشت.

۳-۵- پیاده سازی مدل پیشنهادی با ترکیب ۲۱ ماشین بردار پشتیبان

ماشین بردار پشتیبان دسته‌بندی است، که در سال ۱۹۹۲ توسط واپنیک معرفی شده و بر پایه تئوری آموزش آماری دو کلاسه بنا گردیده است، بعدها در سال ۱۹۹۹ توسط شالکوف، و همکاران گسترش یافت[۱۲]. شهرت این روش به خاطر موفقیت آن در تشخیص حروف دست نویس است، که با شبکه‌های عصبی به دقّت تنظیم شده برابری می‌کند. در فاز آموزش دسته‌بندی کننده ماشین بردار پشتیبان، سعی می‌شود که مرز تصمیم‌گیری به گونه‌ای انتخاب گردد، که حداقل فاصله آن با هر یک از کلاس‌های مورد نظر بیشینه شود.

در مسایلی که داده‌ها به‌صورت خطی جداپذیر نباشند، به فضایی با ابعاد بیشتر نگاشت داده می‌شوند، تا بتوان آنها را در این فضای جدید به‌صورت خطی جدا نمود. این نوع انتخاب باعث می‌شود، که تصمیم‌گیری نسبت به شرایط نویز حساس نبوده و پاسخ‌دهی خوبی داشته باشد. این نحوه انتخاب مرز بر اساس نقاطی به نام بردارهای پشتیبان انجام می‌شود. در اصل ماشین بردار پشتیبان برای جداسازی دو کلاس طراحی شده است که می‌توان آن را برای جداسازی چند کلاس تعمیم داد[۱۳]. شکل ۲ طرحی شماتیک از تفکیک دو کلاس در دو بعد را نمایش می‌دهد.



شکل۲: نمایش بردارهای پشتیبان و مرز جدا کننده دو کلاس[۱۳].

تئوری ماشین بردار پشتیبان به طراحی صفحه جداکننده‌ای با بـردار نرمـالw و فاصـله از مبـدأ b مـی‌پـردازد، کـه ترکیـب خطـی از بردارهای پشتیبان است. این تابع جداساز $F(x)$ به‌صورت رابطه ۴ تعریف می‌شود:

$$F(x) = w \cdot x + b \qquad (۴)$$

نمونه‌های آموزشی برای دسته‌بند ماشین بردار پشتیبان به‌صورت $(x_1,y_1),\ldots,(x_n , y_n) \in R^N * \{\pm 1\}$ مشخص می‌شود، n تعداد داده‌های آموزشی را مشخص می‌کند. وقتی که داده‌های آموزشی به طور خطی جدا پذیر باشند، می‌توان صفحهٔ جداکننده F(x) را پیدا کرد که $y_i F(x_i) \geq 1 , i= 1, \ldots, n$

فاصله ما بین یک نمونه آموزشی و صفحهٔ جداکننده به صورت رابطه ۵ تعیین می‌شود:

$$\frac{y_i F(x_i)}{\|w\|} \qquad (۵)$$

با توجه به رابطه‌های بالا داریم:

$$\frac{y_i F(x_i)}{\|w\|} \geq \frac{1}{\|w\|} \qquad (۶)$$

همچنین اگر کمترین فاصله نمونه‌های آموزشی از هر کلاس تا صفحه جداساز را حاشیه M بنامیم، هدف ما در این الگوریتم ماکزیمم کردن این حاشیه است. پس ماکزیمم کردن حاشیه M برابر با مینیمم کردن w است. از ذکر محاسبات پیچیده ریاضی برای حصول این رابطه خودداری می‌شود، علاقه‌مندان می‌توانند به منابع ذکر شده در متن مراجعه نمایند.با توجه به شرایط بیان شده حل مسئله بالا معادل حل رابطهٔ ۸ است.

$$min\frac{1}{2}\ \|w\|^2 \qquad (۷)$$

با توجه به این شرایط که

$$y_i(w\ .\ x+b\ )-1\geq 0 \qquad i=1,\dots,p \qquad (۸)$$

برای حل این مسئله α_i را به عنوان ضرب‌کننده لاگرانژ تعریف می‌کنیم، در این صورت به حل مسئله لاگرانژین ۹ می‌پردازیم[۱۴].

$$L(w,b,\alpha)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{n}\alpha_i(y_i(w.x_i+b)-1) \qquad (۹)$$

با توجه به اینکه داریم:

$$\alpha_i\geq 0 \qquad i=1,\dots,n$$

رابطه L تابعی از متغیرهای w,b, α می‌باشد، با مشتق گیری از رابطه ۹ نسبت به متغیرهای b و wداریم:

$$\frac{\partial L(w,b,\alpha)}{\partial w}=w-\sum_{i=1}^{n}\alpha_i y_i x_i=0 \qquad (۱۰)$$

$$\frac{\partial L(w,b,\alpha)}{\partial w}=\sum_{i=1}^{n}\alpha_i y_i=0$$

با جایگذاری روابط خواهیم داشت:

$$max\ L(w,b,\alpha)=\sum_i\alpha_i-\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j(x_i^T x_j) \qquad (۱۱)$$
$$\sum\alpha_i y_i=0,\alpha_i\geq 0\forall i$$

پس از حل این مساله ضریب α_i تنها برای یک سری از داده‌های آموزشی غیر صفر است. نمونه‌های آموزشی که ضریب لاگرانژ آنها غیر صفر است $\alpha_i\neq 0$ را بردار پشتیبان می‌نامیم، که در حقیقت داده‌هایی هستند که کمترین فاصله را با صفحهٔ جداکننده دارند. روش‌های بهینه سازی مختلفی برای به دست آوردن ضرایب لاگرانژ طراحی شده‌اند، معروف‌ترین آنها روش‌های حداقل مربعات، برنامه‌ریزی درجه دوم و بهینه سازی کمینه متوالی می‌باشد. در این مقاله از روش بهینه‌سازی کمینه متوالی برای این منظور استفاده شده است. این روش برای حل برنامه‌ریزی درجه دومی (QP) است که در طول آموزش ماشین بردار پشتیبان ایجاد می‌شود. این الگوریتم توسط جان پلات، در سال ۱۹۹۸ در تحقیقات مایکروسافت اختراع شد[۱۵].

در بسیاری از مواقع داده‌ها به صورت خطی جدایی پذیر نیستند. در این حالت می‌توان تابع کرنل را به صورتی‌که بتواند داده‌های آموزشی را به فضایی دیگر منتقل کند، که در فضای جدید امکان تفکیک خطی کلاس‌ها وجود داشته باشد. مفهوم این توضیح در شکل ۳ نمایش داده می‌شود. تابع کرنل می‌تواند انواع مختلفی داشته باشد، که ما در این مقاله از تابع کرنل خطی استفاده نموده ایم.

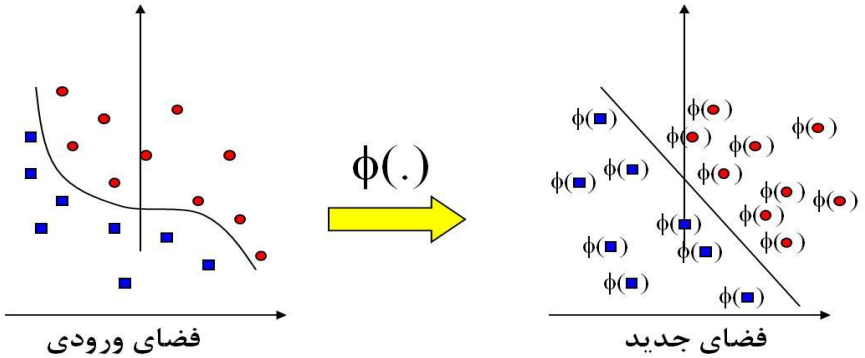
$$K(x\ ,x_i)=(x^T x_i) \qquad (۱۲)$$

در این صورت تابع جداساز با توجه به رابطه $w=\sum_{i=1}^{n}\alpha_i y_i x_i$ به فرم ۱۳ تغییر خواهد کرد.

$$F(x)=sgn(\sum\alpha_i y_i\,K(x,x_i)+b) \qquad (۱۳)$$

استفاده از نتایج چند دسته بند با عنوان یادگیری دسته جمعی، یک رویکرد مؤثر در یادگیری ماشینی است که در آن به منظور بهبود دقت یادگیری، نتایج دسته بندها با یکدیگر ترکیب شده و یک سیستم مرکب شکل می گیرد. ترکیب دسته بندها شامل دو بخش است. بخش اول شامل ایجاد دسته بندهای پایه مناسب، انتخاب نوع دسته بندها، تعداد دسته بندها و ویژگی های مناسب برای هر دسته بند است. بخش دوم شامل نحوه ترکیب خروجی دسته بندها به منظور حصول بهترین نتیجه برای دسته بندی الگوهاست[۱۶].

یکی از متداولترین روش‌های ترکیب دسته‌بندها، روش مبتنی بر رأی‌گیری است. ما در این مقاله از این روش برای انتساب متون نامشخص به یک کلاس استفاده نموده ایم.



شکل۳: تبدیل فضای داده

در مدل پیشنهادی دسته های ورودی را به ۲۱ حالت مختلف دو دسته ای تفکیک نمودیم و آنها را به ۲۱ کلاسه‌بند ماشین بردار پشتیبان با استفاده از کرنل خطی داده ایم و در نهایت زمانی که یک متن نامشخص به هر یک از دسته بند ها داده می شود، به آن یک برچسب زده می شود، آن برچسبی که بیشترین تکرار را دارد، متن نامشخص ما به آن دسته تعلق می گیرد. لیست دسته‌بندها عبارتند از: ۱- اقتصادی و اجتماعی ۲- اقتصادی و ادب و هنر ۳- اقتصادی و سیاسی ۴- اقتصادی و علمی فرهنگی.علمی.ارتباطات و فناوری اطلاعات ۵- اقتصادی و گوناگون.حوادث ۶- اقتصادی و ورزش ۷- ادب و هنر و اجتماعی ۸- ادب و هنر و سیاسی ۹- ادب و هنر و علمی فرهنگی.علمی.ارتباطات و فناوری اطلاعات ۱۰- ادب و هنر و گوناگون.حوادث ۱۱- ادب و هنر و ورزش ۱۲- اجتماعی و سیاسی ۱۳- اجتماعی و علمی فرهنگی.علمی.ارتباطات و فناوری اطلاعات ۱۴- اجتماعی و گوناگون.حوادث ۱۵- اجتماعی و ورزش ۱۶- سیاسی و علمی فرهنگی.علمی.ارتباطات و فناوری ۱۷- سیاسی و گوناگون.حوادث ۱۸- سیاسی و ورزش ۱۹- علمی فرهنگی.علمی.ارتباطات و فناوری اطلاعات و گوناگون.حوادث ۲۰- علمی فرهنگی.علمی.ارتباطات و فناوری اطلاعات و ورزش ۲۱- گوناگون.حوادث و ورزش.

۳-۶- ارزیابی مدل پیشنهادی

در آخرین مرحله پس از پیاده سازی الگوریتم های دسته بندی باید با استفاده از متون آزمایشی، صحت، دقت، بازخوانی، معیار ارزیابی F مدل پیشنهادی را بدست می آوریم.

قبل از بیان روابط سنجش دقت دسته بندی نیاز به معرفی پیش نیازهای زیر است.ابتدا به جدول ۲ توجه کنید.

جدول۲: ماتریس درهم ریختگی برای یک مسئله دسته بندی دو دسته ای [۹].

		رکوردهای واقعی	
		دسته+	دسته−
رکوردهای تخمینی	دسته+	*TP*	*FP*
	دسته−	*FN*	*TN*

این جدول، بیانگر دقت الگوریتم دسته بندی برای تشخیص دسته ها است و حاصل ضرب کلیه حالات دسته بندی سیستم با دسته بندی یک شخص خبره است. چهار مقدار جدول بالا به شرح زیر، تعریف می گردند.

FP:تعدادی از داده ها که به غلط به عنوان دسته مثبت شناسایی می شوند.

TP: تعدادی از داده ها که به درست به عنوان دسته مثبت شناسایی می شوند.

FN: تعدادی از داده ها که به غلط به عنوان دسته منفی شناسایی می شوند.

TN: تعدادی از داده ها که به درست به عنوان دسته منفی شناسایی می شوند.

چهار عبارت موجود در چهار سلول جدول بالا مطابق روش های استاندارد کیفیت سنجی دسته بندی تعریف شده اند. به ترتیب نشان دهنده مقادیر بازخوانی دسته بند هستند. این مقادیر همواره بین(۰ و ۱) می باشند و هرچه به ۱ نزدیکتر باشند، عملکرد روش پیشنهادی بهتر خواهد بود[۹].

معیارهای ارزیابی دسته+ :

$$\textbf{بازخوانی} = \frac{TP}{FN+TP} \quad (۱۴)$$

$$\textbf{دقت} = \frac{TP}{TP+FP} \quad (۱۵)$$

$$\text{معیار F} = \frac{2.Recall.Precision}{Recall+Precision} \quad (۱۶)$$

معیارهای ارزیابی دسته − :

$$\textbf{بازخوانی} = \frac{TN}{TN+FP} \quad (۱۷)$$

$$\textbf{دقت} = \frac{TN}{TN+FN} \quad (۱۸)$$

$$\text{معیار F} = \frac{2.Recall.Precision}{Recall+Precision} \quad (۱۹)$$

معیارهای ارزیابی دسته بند :

$$\textbf{صحت} = \frac{TP+TN}{FN+TP+TN+FP} \quad (۲۰)$$

$$\textbf{خطا} = \frac{FN+FP}{TP+FP+FN+TN} \quad \textbf{صحت} \text{ -1=} \quad (۲۱)$$

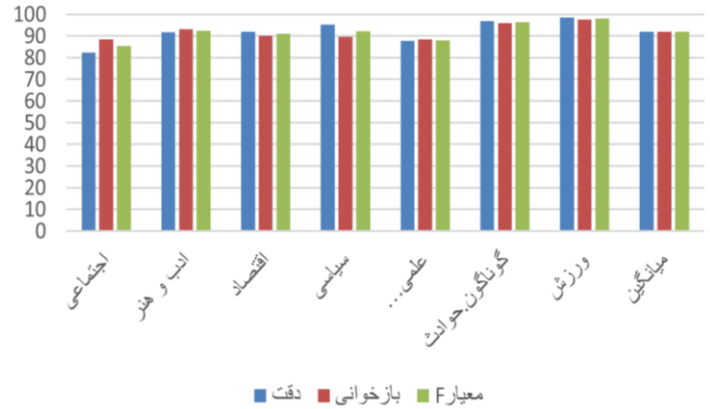
یکی از روش های ارزیابی الگوریتم های دسته‌بندی روش جامع k-Fold Cross Validation می باشد در این روش کل مجموعه داده‌ها به k قسمت مساوی تقسیم می‌شوند. از k-1 قسمت به عنوان مجموعه داده‌های آموزشی استفاده می‌شود و براساس آن مدل ساخته می‌شود و با یک قسمت باقی مانده عملیات ارزیابی انجام می‌شود. فرآیند مزبور به تعداد k مرتبه تکرار خواهد شد، به گونه ای که از هر کدام از k قسمت تنها یکبار برای ارزیابی استفاده شده و در هر مرتبه یک دقت برای مدل ساخته شده، محاسبه می‌شود.

جدول ۳: ارزیابی مدل پیشنهادی (به درصد)

ردیف	نام فارسی	دقت	بازخوانی	معیار F
۱	اجتماعی	۸۲,۳۳	۸۸,۵۰	۸۵,۳۰
۲	ادب و هنر	۹۱,۶۳	۹۳,۰۰	۹۲,۳۱
۳	اقتصاد	۹۱,۸۴	۹۰,۰۰	۹۰,۹۱
۴	سیاسی	۹۵,۲۱	۸۹,۵۰	۹۲,۲۷
۵	علمی فرهنگی.علمی.ارتباطات و فناوری اطلاعات	۸۷,۶۲	۸۸,۵۰	۸۸,۰۶
۶	گوناگون.حوادث	۹۶,۹۷	۹۶,۰۰	۹۶,۴۸
۷	ورزش	۹۸,۴۸	۹۷,۵۰	۹۷,۹۹
۹	میانگین	۹۲,۰۱	۹۱,۸۶	۹۱,۹۰
۱۰	صحت دسته بند	۹۱,۸۶		
۱۱	خطای دسته بند	۸,۱۴		

در این روش ارزیابی دقت نهایی دسته بند برابر با میانگین k دقت محاسبه شده خواهد بود .معمول‌ترین مقداری که در متون علمی برای k در نظر گرفته می‌شود برابر با ۱۰ می‌باشد. بدیهی است هر چه مقدار k بزرگتر شود، دقت محاسبه شده برای دسته بند قابل اعتمادتر بوده و دانش حاصل شده جامع‌تر خواهد بود و البته افزایش زمان ارزیابی دسته بند نیز مهمترین مشکل آن می‌باشد. حداکثر مقدار k برابر با تعداد رکوردهای مجموعه داده اولیه است که این روش ارزیابی با نام Leaving One Out شناخته می‌شود[۹].

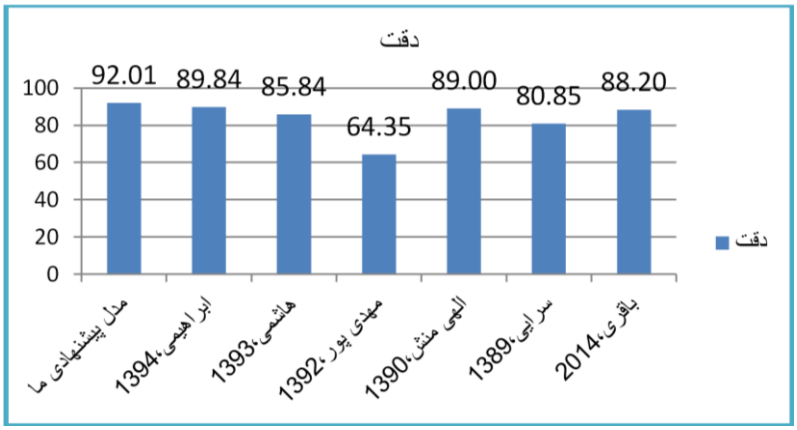
ما در مقاله خود k را برابر ۱۰ قرار داده ایم و ۱۰ بار این مدل ارزیابی را اجرا نموده و در نهایت میانگین ارزیابی محاسبه گردید. در جدول ۳ ، پنج معیار دقت ، بازخوانی ، معیار f ، صحت و خطای دسته بند، مدل پیشنهادی قابل مشاهده است. در شکل ۴ معیارهای ارزیابی برای مدل پیشنهادی در هفت دسته، قابل مشاهده است.



شکل۴: معیارهای ارزیابی مدل پیشنهادی

۴- مقایسه مدل پیشنهادی با کارهای پیشین

در شکل۵ به مقایسه مدل پیشنهادی با کارهای پیشین پرداخته ایم و به وضوح مشخص است که دقت مدل پیشنهادی نسبت به کارهای پیشین بالاتر می باشد.



شکل۵: مقایسه مدل پیشنهادی با کارهای پیشین

۵- نتیجه گیری

در این مقاله، بعد از انتخاب مجموعه داده و پاک‌سازی متون به کمک روش نرمال شده فرکانس کلمه- معکوس فرکانس سند (norm TF-IDF) به ویژگی‌ها وزن داده می‌شود و در طی دو مرحله ویژگی‌ها با استفاده از روش‌های فرکانس سند (DF) و مربع چی(SChi) انتخاب می‌شوند و بعد با استفاده از روش تحلیل مؤلفه اصلی (PCA) ابعاد ویژگی‌ها کاهش داده می‌شود و در مرحله بعد با استفاده از ترکیب ۲۱ ماشین بردار پشتیبان (SVM) دودویی به پیاده سازی مدل پیشنهادی می‌پردازیم و در نهایت صحت مدل را با

روش اعتبار سنجی ۱۰ مرحله‌ای ارزیابی می‌کنیم نتایج تجربی نشان می‌دهد که این مدل می‌تواند عمل دسته‌بندی متون را برای هفت دسته با دقت ۹۲,۰۱، فراخوانی ۹۱,۸۶، معیار F ۹۱,۹۰، صحت ۹۱,۸۶ و خطای ۸,۱۴ ،انجام دهد که نسبت به کارهای پیشین انجام گرفته کارایی بالاتری دارد، علت اصلی برتری این مقاله، انتخاب درست ویژگی ها و استفاده از ترکیب ۲۱ ماشین بردار پشتیبان دودویی که باعث می شود مدل پیشنهادی بتواند با دقت بالاتری نسبت به کارهای پیشین، به دسته بندی متون نامشخص بپردازد.در پایان می توان این مدل را برای دسته های بیشتر و یا سایر زمینه های داده کاوی از قبیل؛ وب،تصاویر و غیره نیز پیاده سازی کرد و به ارزیابی دقت و صحت مدل پرداخت.

مراجع

[۱] ایمان. ابراهیمی، و همکاران، "رده بندی متون فارسی با استفاده از ماشین بردار پشتیبان مبتنی بر روشهای انتخاب ویژگی PCA و الگوریتم ژنتیک،" کنفرانس ملی برق و الکترونیک، گناباد، ۲۹-۲۸ مرداد۱۳۹۴.

[۲] محمدحسین. سرایی، و آذر. شاهقلیان، "کاوش متون فارسی بر مبنای روش دسته بندی،" نشریه علمی پژوهشی انجمن کامپیوتر ایران، جلد ۸ ، شماره ۱ و۳ ، صفحه ۱۳-۸، ۱۳۸۹.

[۳] محمدحسین. الهی‌منش، و بهروز. مینایی، "رده‌بندی متون فارسی با استفاده از روش‌های آماری،" ارائه شده در سمینار فناوری‌های پردازش هوشمند متون اسلامی، ۲۹-۲۶ فروردین ۱۳۹۰، صفحه ۹۵-۹۰.

[۴] الهام. مهدی پور، و همکاران، "سیستم خلاصه ساز خودکار متن فارسی با استفاده از الگوریتم ترکیبیSA-GA ،" همایش ملی مهندسی کامپیوتر و توسعه پایدار با محوریت شبکه های کامپیوتری، مدل سازی و امنیت سیستم ها، مشهد، موسسه آموزش عالی خاوران، ۲۸ آذر۱۳۹۲.

[۵] سیدمحسن. هاشمی، و همکاران، "استفاده از تکنیک های متن کاوی برای دسته بندی متون فارسی با مجموعه داده همشهری،" *کنفرانس بین المللی مهندسی، هنر و محیط زیست، کشور لهستان*، ۲۱ آذر۱۳۹۳.

[۶] مینا. ملکی، و احمد. عبدالله زاده بارفروش، "TFCRF: روش جدید وزن دهی ویژگی مبتنی بر اطلاعات کلاس در حوزه طبقه بندی مستندات،" *دوازدهمین کنفرانس سالانه انجمن کامپیوتر ایران، تهران، دانشگاه شهید بهشتی*، ۳-۱ اسفند۱۳۸۵.

[۷] سعید. جلیلی، و مهدی. بیطرفان، "افزایش کارایی دسته بندی متون براساس بهبود روش انتخاب خصیصه،" *نشریه دانشکده فنی*، جلد ۴۰، شماره ۳، صفحه ۳۲۸-۳۱۳، ۱۳۸۵.

[۸] مهدی. برفامی، و سهیل. فاطری، "استفاده از ترکیب شبکه های عصبی جهت دسته بندی متون فارسی مبتنی بر الگوریتم های PCA , KNN, GAبرای انتخاب ویژگی،" *اولین کنفرانس رویکرد های نوین در مهندسی کامپیوتر و بازیابی اطلاعات ایران، دانشگاه آزاد اسلامی واحد رودسر و املش*، ۱۵ مهر۱۳۹۲.

[۹] محمد. صنیعی آباده، و همکاران، داده‌کاوی کاربردی، تهران: *انتشارات نیاز دانش*، ۱۳۹۱.

[10] A. Bagheri, and et al, "PSA: A Hybrid Feature Selection Approach for Persian Text Classification," *Journal of Computing and Security*, Vol. 1, No. 4, pp. 261-272, 2014.

[11] http://ece.ut.ac.ir/dbrg/hamshahri/faindex.html.

[12] B. Schölkopf, and et al, "Advances in Kernel Methods Support Vector Learning," *Cambridge, MA: MIT Press*, 1998.

[13] B. E. Boser, "A training algorithm for optimal margin classifiers," *Proceedings of the fifth annual workshop on computational learning theory*, Pittsburgh, pp. 144-152, 1992.

[14] N. Christiani, and et al, "An introduction to support vector machines," *Cambridge, MA: Cambridge University*, 2000.

[15] J. Platt, "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines," *Technical Report MSR-TR-98-14*, 1998.

[16] D. Ruta, and et al, "An Overview of Classifier Fusion Methods," *Computing and Information Systems*, Vol. 7, pp. 1-10, 2000.

Offering a Model for Persian Texts Classify by Combination of Classification Methods

Iman Jamali[1], S.Javad Mirabedini[2*], Ali Haronabadi[3]

[1]Islamic Azad University, Bushehr Branch, Bushehr, Iran
[2*]Islamic Azad University, Central Tehran Branch, Tehran, Iran
[3]Islamic Azad University, Central Tehran Branch, Tehran, Iran
1: imanjamali25@gmail.com
2*: jvd.2205@yahoo.com
3: a.harounabadi@gmail.com

ABSTRACT:

To classify text information extraction techniques, natural language processing and machine learning has been widely used general purpose of categories of documents, classified documents in the form of a certain number of categories are pre-determined. Each document can be in one, several or no category is placed. In the case of any document to this question will be placed the document on which of the categories. This can be in the form of an automatic learning to use it any document can be automatically assigned to a category. In this thesis, data collection and cleanup after you select text using the normal method of word frequency -inverse document frequency (norm TF-IDF) is the weight features and features in two stages using document frequency (DF) and Chi square (SChi) are selected, and then using principal component analysis (PCA) features reduced dimensions, and at a later stage by combining 21 support vector machine (SVM) the proposed model we have implemented, and the accuracy of the model to assess the 10-step method validation. Experimental results show that this model can text classification accuracy of 91.86 for the seven categories do, which has a higher accuracy than the earlier work done.

KEYWORDS: Support Vector Machine ,Classification of texts, feature selection, Principal Component Analysis.