افزایش کارائی دسته‌بندی متون بر اساس بهبود روش انتخاب خصیصه

سعید جلیلی

استادیار گروه کامپیوتر - دانشکده فنی و مهندسی - دانشگاه تربیت مدرس

Sjalili@modares.ac.ir

مهدی بیطرفان

فارغ التحصیل کارشناسی ارشد کامپیوتر – دانشکده فنی و مهندسی – دانشگاه تربیت مدرس

mehdib@modares.ac.ir

(تاریخ دریافت ۸۲/۶/۲۹ ، تاریخ دریافت روایت اصلاح شده ۸۴/۱۰/۲۴، تاریخ تصویب۸۵/۲/۹)

چکیده

در دسته بندی متون معمولا از کلمات به عنوان خصیصه استفاده می شود. از آنجا که در هر مجموعه متون، تعداد بسیار زیادی خصیصه وجود دارد، در این مقاله روش‌هائی به منظور کاهش تعداد خصیصه‌ها و انتخاب خصیصه‌های مرتبط، پیشنهاد شده‌است. به طور کلی روش‌های فیلتری انتخاب خصیصه از چهار همبستگی مثبت و منفی بین خصیصه‌های سندها و دسته‌ها در ضابطه انتخاب خصیصه استفاده می‌کنند. در روش‌های پیشنهادی این مقاله ابتدا کلیه همبستگی های مثبت و منفی را با اثر مثبت و سپس همبستگی های مثبت را با اثر مثبت و همبستگی های منفی را با اثر منفی در ضابطه انتخاب خصیصه‌ها در نظر گرفته‌ایم. نتایج آزمایشات نشان دهنده تاثیر بیشتر همبستگی های مثبت، نسبت به همبستگی های منفی می‌باشد. دیگر روش پیشنهادی، ترکیبی از یک روش فیلتری و یک روش روکشی است که ابتدا با استفاده از روش فیلتری Relief-F تعدادی از خصیصه‌ها با هزینه محاسباتی کمی انتخاب می‌شوند و سپس از خصیصه‌های انتخاب شده با استفاده از روش روکشی SFS یا SBS به صورت دقیقتری با کمک دسته‌بند متون، خصیصه‌های نامرتبط حذف می‌شوند. ارزیابی‌های انجام گرفته نشان دهنده کارائی بهتر روش ترکیبی نسبت به روش‌های فیلتری است.

واژه های کلیدی: دسته‌بندی متون، انتخاب خصیصه، ترکیب فیلتری-روکشی، همبستگی مثبت و منفی بین خصیصه‌های سند و دسته

مقدمه

مدیریت اطلاعات متنی انبوه به صورتی که به ساده‌ترین روش قابل دسترسی باشند، یکی از روش‌های بازیابی اطلاعات است. در برخورد و مواجه با انبوه اطلاعات متنی، روش‌های مختلفی جهت مدیریت و جستجوی این اطلاعات وجود دارد. در روش‌های سنتی معمولا تعدادی کلمه کلیدی که بیان کننده محتوای آن متن می‌باشند به صورت دستی به هر سند متنی نسبت داده می‌شود و با توجه به این کلمات، متون مورد جستجو قرار می‌گیرند. در صورتی‌که لزوما کلمات کلیدی نسبت داده شده، بیانگر دقیقی از محتوای متن نیستند. بنابراین این روش‌ها برای بازیابی متون مناسب نمی باشند.

در روش‌های نوین بازیابی اطلاعات متنی، متون با توجه به محتوای آنها بازیابی می‌شوند. در ده سال گذشته فرایند مدیریت سندها به صورت خودکار با توجه به محتوای آنها، در بخش سیستم‌های اطلاعاتی با توجه به گسترش روزافزون اطلاعات الکترونیکی، بسیار مطرح بوده است.

یکی از روش‌هائی که به منظور مدیریت مناسب سندهای متنی مطرح است، دسته‌بندی(طبقه‌بندی) متون است. در اصل دسته‌بندی متون، فرایندی است که متون زبان طبیعی را به تعدادی دسته موضوعی از پیش تعریف شده برچسب می‌زند.

دسته‌بندی خودکار متون، سابقه طولانی مدتی دارد و ریشه‌های آن به قبل از دهه ۶۰ میلادی باز می‌گردد اما تا قبل از دهه ۹۰ میلادی زیاد مورد توجه قرار نگرفته بود و زیاد مورد علاقه جامعه محققین نبوده است، تا اینکه با رشد سخت افزارهای قدرتمند کامپیوتری و استفاده از آنها در زندگی روزمره انسان‌ها، مجددا در سیستم‌های اطلاعاتی بطور قابل ملاحظه‌ای مورد استفاده قرار گرفت [۱]. در روش‌های دسته‌بندی، به منظور یادگیری این روش‌ها، سندهای متنی به دو دسته سندهای آموزشی و سندهای آزمون تقسیم می‌شوند. از سندهای آموزشی برای یادگیری روش دسته‌بندی و از سندهای آزمون برای ارزیابی میزان دقت و کامل بودن روش دسته‌بندی استفاده

می‌شود. تاکنون روش‌های زیادی از جمله روش‌های آماری برای دسته‌بندی خودکار سندهای متنی تدوین شده‌است. از جمله آنها می‌توان به روش دسته‌بندی نزدیکترین همسایه[۱]، روش بیزین[۳]، روش SVM [۴] ، درخت های تصمیم[۵]و شبکه‌های عصبی[۳] اشاره نمود.

به منظور اعمال روش‌های دسته‌بندی، بایستی سندهای متنی به روشی که نشان دهنده محتوای آنها باشد، نمایش داده شوند. معمولا سندها را با استفاده از تک کلمات درون سندها، با درنظر نگرفتن کلمات بی ارزش (مانند از، را، چرا و...) نمایش می‌دهند، اما ما در این تحقیق سایر ترکیب‌های کلمات، علی‌الخصوص ترکیب‌های دوتائی کلمات را به همراه تک کلمه نیز در نظر گرفته‌ایم و تاثیر استفاده از آنها را روی کارائی روش‌های دسته‌بندی بررسی نموده‌ایم. با توجه به اینکه در مسائل مربوط به دسته‌بندی متون، یکی از مشکلات، فضای خصیصه‌ها با ابعاد بسیار بالا است [۶] و استفاده از ترکیب‌های دوتائی کلمات باعث افزایش این فضا می‌گردد، لذا در این مقاله با استفاده از روش‌های متداول انتخاب کلمات، کلیه کلمات نامرتبط و کلماتی که فاقد ارزش اطلاعاتی هستند از فضای کلمات حذف شده و حتی با بهبود کارائی دسته‌بند SVM Light توانستیم تا ۹۴٪ از کلمات را حذف نمائیم و فضای کلمات را کاهش دهیم.

تعداد زیاد خصیصه ها، مانع بزرگی برای اکثر روش‌های یادگیری است. کاهش فضای خصیصه‌ها، بدون از دست دادن دقت دسته‌بندی، تا حدود زیادی مشکل فوق‌الذکر را مرتفع می‌کند. علاوه بر آن کاهش خصیصه‌ها بصورت خودکار (بدون نیاز به تعریف دستی خصیصه‌ها) از اهداف اصلی روش‌های انتخاب خصیصه است. تعداد زیادی روش انتخاب خصیصه وجود دارد، اما تعداد کمی از آنها برای مسائل دسته‌بندی متون حجیم استفاده شده است که نشان دهنده این واقعیت است که بسیاری از این روش‌ها، برای فضای خصیصه‌ها با ابعاد بسیار بالا مناسب نمی‌باشند. روش‌های مختلفی جهت انتخاب خصیصه‌های متنی و درنتیجه کاهش فضای آنها وجود دارد. Yang [۶] به پنج روش آماری مختلف انتخاب خصیصه اشاره کرده است و تاثیر آنها را روی دو دسته‌بند نشان داده‌است. در[۷] نیز به هشت روش آماری انتخاب خصیصه در ناحیه متون اشاره شده است. در مجموع دو همبستگی مثبت و دو همبستگی منفی بین خصیصه‌ها و دسته‌ها را می‌توان در محاسبات لحاظ نمود که اکثر روش‌های آماری انتخاب خصیصه با توجه به ترکیب‌های مختلف این همبستگی‌ها و اثردهی مثبت و منفی به آنها، ضابطه‌ای را جهت وزن‌دهی به خصیصه‌ها تعریف می‌کنند. در این تحقیق ترکیب‌های مختلفی از این همبستگی‌ها در نظر گرفته شده است و تاثیر هر یک از آن ها به صورت جداگانه و به صورت ترکیب‌های مختلف، بررسی شده است. در [۸] روش‌های انتخاب خصیصه به دو گروه مختلف تقسیم بندی شده‌است: روش‌های فیلتری[۲] و روش‌های روکشی[۳]. روش‌های فیلتری (به عبارت دیگر روش‌های آماری) مستقل از روش یادگیری هستند، درحالی‌که روش‌های روکشی از روش یادگیری بعنوان تابع ارزیابی استفاده می کنند. روش‌های فیلتری بدون درنظر گرفتن روش یادگیری به حذف خصیصه‌ها می‌پردازند و دارای پیچیدگی زمانی پائین بوده و دقت آنها قابل پیش‌بینی نیست، در حالی‌که روش‌های روکشی دارای پیچیدگی زمانی بالا و دقت بسیار بالائی هستند. به دلیل دقت بسیار بالای روش‌های روکشی، مناسب است که از این روش‌ها در انتخاب خصیصه‌ها استفاده نمائیم[۱۱]، اما هزینه محاسباتی بسیار زیاد این روش‌ها باعث می‌شود که گاهاً اعمال این روش‌ها روی مجموعه داده‌های بسیار بزرگ، غیرممکن باشد. بنابراین استفاده ترکیبی آنها با روش‌های فیلتری و ایجاد تعادلی بین این دو گروه از روش‌ها می‌تواند زمینه مناسبی را جهت استفاده از مزایای هردوی آنها ایجاد کند. در این تحقیق، ترکیبی از یک روش فیلتری و یک روش روکشی پیشنهاد شده که ابتداء با استفاده از روش فیلتری Relief-F (که نتایج بسیار خوبی در حذف خصیصه‌ها در دسته بندی سندها از خود نشان داده است[۹] و اولین مرتبه است که در دامنه متن استفاده می‌شود) تعدادی از خصیصه‌ها با هزینه محاسباتی کمی انتخاب می‌شوند و سپس خصیصه‌های انتخاب شده با استفاده از روش روکشی SFS[۴] یا SBS[۵] به صورت دقیقتری با استفاده از دسته‌بند متون، کاهش می‌یابند.

جهت انجام آزمایشات از مجموعه داده‌های رویتر[۱۰] و نیگام [۱۹] استفاده شده است که ۱۰ زیرمجموعه داده مختلف از این دو مجموعه داده مورد استفاده قرار گرفته است و همچنین جهت انجام دسته‌بندی‌ها از دو دسته‌بند معروف SVM Light (که گونه‌ای از دسته‌بند معروف SVM است) و دسته‌بند بیزین ساده استفاده شده‌است.

در ادامه این مقاله، بخش دوم، به معرفی فضای خصیصه‌ها جهت نمایش سندهای متنی می‌پردازد. در بخش سوم معماری اطلاعات به منظور یادگیری دسته‌بندهای متنی ارائه می شود. در بخش چهارم روش‌های انتخاب خصیصه موجود بررسی می‌شوند. بخش پنجم به بررسی همبستگی بین خصیصه‌ها و دسته‌ها در ضابطه‌های روش‌های انتخاب خصیصه می‌پردازد. در بخش ششم روش‌های پیشنهادی انتخاب خصیصه معرفی می‌شوند. در بخش هفتم محیط ارزیابی آزمایشات معرفی می‌شود. در بخش‌های هشتم و نهم به ترتیب به ارزیابی روش نمایش ترکیبی و ارزیابی روش‌های انتخاب خصیصه پیشنهادی پرداخته می‌شود. در نهایت، در بخش دهم، نتیجه‌گیری بعمل آمده و پیشنهاداتی در رابطه با توسعه هر یک از روش‌ها ارائه می‌شود.

فضای خصیصه‌ها در نمایش سندهای متنی

به منظور اعمال روش‌های دسته‌بندی متون و نیز اعمال روش‌های انتخاب خصیصه‌های متون، بایستی ساختاری مناسب جهت نمایش سندها در نظر گرفته شود. ساده‌ترین و عمومی‌ترین روش نمایش متون، ایجاد یک فضای خصیصه از تمام کلمات بوده که در سندها وجود دارد. در این فضای خصیصه‌ها، پس از حذف کلمات خاص[۶] و گاهاً ریشه‌یابی[۷]، لیستی از تمامی کلمات در متون ایجاد شده و هر سند با توجه به اینکه دربر دارنده چه کلماتی از این لیست بوده و با چه وزنی این کلمات در سند اتفاق افتاده‌اند، به روش‌های مختلف نمایش داده می‌شود. جدول (۱) روش نمایش سندهای متنی را بصورت برداری نشان می‌دهد.

جدول ۱ : روش نمایش برداری سندهای متنی.

	فضای کلمات			
	F_1	F_2	...	F_n
D_1	W_{11}	W_{12}	...	W_{1n}
D_2	W_{21}	W_{22}	...	W_{2n}
...	...	...	...	...
D_s	W_{s1}	W_{s2}	...	W_{sn}

همانطور که در جدول (۱) نشان داده شده‌است مجموعه $\{F_{1,...,}F_n\}$ بیانگرفضای خصیصه‌ها است و F_i کلمه‌ای را نشان می‌دهد که حداقل یکبار در یک سند بکاررفته‌است. همچنین مجموعه $\{D_{1,...,}D_s\}$ مجموعه سندهائی را نشان می‌دهد که در این روش نمایش برداری، نمایش داده شده‌اند. هر یک از مقادیر W_{ij} نشان دهنده وزن کلمه F_j در سند D_i می‌باشد، که البته معیارهای متفاوتی جهت وزن‌دهی به کلمات وجود دارد[۲].

در اکثر مطالعات در زمینه دسته‌بندی متون، از چنین فضای خصیصه‌ای جهت نمایش سندها استفاده شده‌است. این روش نمایش، در کنار مزیت‌هائی چون سادگی نمایش و عمومی بودن برای تمامی متون، به کلمات موجود در سندها اهمیت می‌دهد و در اصل هر سند را به صورت تعدادی کلمه غیر وابسته به هم نمایش می‌دهد. درحالی‌که در زبان گاهاً از کنار هم قرار گرفتن کلمات، مفاهیم جدیدی بوجود می آید، مانند ساختارهای موصوف صفت و یا مضاف مضاف الیه و... . چنین مفاهیمی در روش اول نمایش سندها قابل بیان نیستند، درحالی‌که این مفاهیم برای مشخص کردن محتوای سندها و ارتباط سندها با دسته‌های مربوطه گاهی بسیار مهم و ضروری می‌باشد. در این مقاله جهت بررسی تاثیر این گونه از روش‌های نمایش روی کارائی دسته‌بندی، به دو صورت، ترکیب‌های دوتائی از کلمات جهت نمایش سندها مورد استفاده قرار گرفته است. یکی در نظر گرفتن تمامی ترکیب‌های دوتائی کلماتی که در کنار هم قرار می‌گیرند، پس از حذف کلمات خاص و بدون در نظر گرفتن علائم نوشتاری جملات ، مانند نقطه، کاما، حروف ربط و اضافه و... و دیگری در نظر گرفتن این‌گونه علائم و ایجاد ترکیب‌های دوتائی معنی دارتر نسبت به حالت قبل.

بعنوان مثال جمله زیر را در نظر بگیرید :

"The automated categorization (or classification) of texts into topical categories has a long history, dating back at least to the early '60s."

چنانچه علائم نوشتاری در جمله فوق درنظر گرفته نشوند، پس از حذف کلمات خاص، هر ترکیب دوتائی کلمات باقیمانده در نظر گرفته شود، مجموعه ترکیب‌های دوتائی بصورت زیر خواهد بود:

{"automated categorization", "categorization classification", "classification texts", "texts topical", "topical categories", "categories long", "long history", "history dating", "dating back", "back least", "least early", "early '60s"}

در این مجموعه، همانطور که ملاحظه می‌شود، تعداد زیادی از ترکیب‌های بی‌معنی وجود دارد درحالی‌که با درنظر گرفتن علائم نوشتاری و در نظر داشتن این واقعیت که بین هیچ ترکیب معنا داری از کلمات، حروف اضافه و

علائم نوشتاری قرار نمی‌گیرد، مجموعه بصورت زیر تبدیل شده که در آن اکثر ترکیب‌های دوتائی، معنی‌دار می‌باشند:

{“automated categorization”, “topical categories”, “long history”, “dating back”, “early ‘60s”}

روش دیگر نمایش سندها، استفاده از هر دو فضای خصیصه تک کلمه ای و ترکیب دوتائی کلمات در کنار هم می‌باشد. در این روش در بدترین حالت فضای خصیصه‌ها، به دو برابر فضای خصیصه اولیه افزایش می یابد که این مسأله باعث بالا رفتن پردازش‌ها و محاسبات می‌گردد، اما با استفاده از روش‌های انتخاب خصیصه می توان از این فضا، خصیصه‌هائی را انتخاب کرد که مرتبط‌تر با مفاهیم دسته‌بندی هستند و به این روش هم فضای خصیصه‌ها را کاهش داده و هم اطلاع راجع به ترکیب‌های معنی دار دوتائی از کلمات را که در نشان دادن محتوای یک متن، جهت دسته‌بندی آن موثر بوده، به نمایش سندها اضافه نمود. البته ترکیب‌های سه‌تائی و بالاتر را نیز می‌توان جهت نمایش متون بکار گرفت که سهامی در[۱] نشان می‌دهد اضافه کردن این ترکیب‌ها باعث پیچیدگی بیشتر و هزینه زمانی بالاتر روش‌های دسته‌بندی شده و عملاً باعث بهبودی در کارائی نمی‌شود.

معماری اطلاعات

همانطور که در بخش اول اشاره شد، در روش‌های دسته‌بندی، به منظور یادگیری دسته‌بندهای متنی، سندها به دو دسته سندهای آموزشی و سندهای آزمون تقسیم می‌شوند. از سندهای آموزشی برای یادگیری روش دسته‌بندی و از سندهای آزمون برای ارزیابی میزان دقت و کامل بودن روش دسته‌بندی استفاده می‌شود. جدول (۲) معماری اطلاعات را در روش‌های یادگیری دسته‌بندی متون نشان می‌دهد. همانطور که در جدول (۲) ملاحظه می‌شود، تعداد m دسته از پیش تعریف شده وجود دارد که سندهای مختلف هر کدام در یک یا چند دسته می‌توانند قرار داشته باشند. از s سندی که به منظور ایجاد دسته‌بند استفاده می‌شود، تعداد g سند برای آموزش و تعداد s-g سند برای آزمون صحت دسته‌بند یادگیری شده استفاده می‌شود. هر یک از مقادیر ca_{ij} می‌توانند مقدار صفر یا یک باشند. مقدار صفر نشان دهنده اینست که سند D_j در دسته C_i قرار نداشته و مقدار یک نشان دهنده وجود سند D_j در دسته C_i می‌باشد. پس از یادگیری دسته‌بند، می‌توان از آن به منظور دسته‌بندی خودکار سندهای جدید استفاده نمود[۲].

جدول ۲ : معماری اطلاعات در روش‌های دسته‌بندی متون.

دسته‌ها	مجموعه آموزش				مجموعه آزمون			
	D_1	…	…	D_g	D_{g+1}	…	…	Ds
C_1	ca_{11}	…	…	ca_{1g}	$ca_{1(g+1)}$	…	…	ca_{1s}
…	…	…	…	…	…	…	…	…
C_i	ca_{i1}	…	…	ca_{ig}	$ca_{i(g+1)}$	…	…	ca_{is}
…	…	…	…	…	…	…	…	…
C_m	ca_{m1}	…	…	ca_{mg}	$ca_{m(g+1)}$	…	…	ca_{ms}

روش‌های انتخاب خصیصه

در این بخش، روش‌های انتخاب خصیصه معرفی می‌شوند. در ادامه نیز توانائی این روش‌ها با توانائی روش‌های پیشنهادی در انتخاب خصیصه‌ها روی مجموعه‌های مختلفی از داده‌ها آزمون شده و نتایج حاصل مورد ارزیابی قرار می‌گیرد. همانطور که اشاره شد، روش‌های انتخاب خصیصه به دو گروه روش‌های فیلتری (به عبارت دیگر آماری) و روش‌های روکشی تقسیم‌بندی می‌شوند. روش‌های فیلتری با استفاده از یک ضابطه به وزن‌دهی خصیصه‌ها پرداخته، سپس به انتخاب خصیصه‌های با وزن بیشتر می‌پردازند. این روش‌ها، دارای هزینه محاسباتی پائین بوده و با توجه به کیفیت روش وزن‌دهی، دارای دقت‌های متفاوتی می‌باشند. در روش‌های روکشی، خصیصه‌ها با توجه به تاثیرشان روی بهبود دقت دسته‌بندی، انتخاب می‌شوند. این روش‌ها دارای دقت بسیار بالا و در مقابل، هزینه محاسباتی بسیار زیادی هستند. در ادامه به معرفی روش‌های متداول فیلتری و روکشی انتخاب خصیصه‌ها می‌پردازیم.

روش‌های فیلتری

تمامی این نوع روش‌های انتخاب خصیصه، از یک ضابطه آستانه[۸] کیفیت کلمه، جهت دستیابی به میزان حذف خصیصه‌های خواسته شده از کل خصیصه‌ها، استفاده می‌کنند که حد این آستانه با توجه به تاثیر روی پارامترهای ارزیابی دسته بندی (دقت و بازخوانی) تعیین می گردد. در مجموع دو همبستگی مثبت[۹] و دو همبستگی منفی[۱۰] بین خصیصه‌ها و دسته‌ها را می‌توان

در محاسبات منظور کرد که اکثر روش‌های انتخاب خصیصه فیلتری، با توجه به ترکیب‌های مختلف این همبستگی‌ها و وزن‌دهی مثبت و منفی به آنها ضابطه‌ای را جهت وزن‌دهی به خصیصه‌ها تعریف کرده‌اند. این روش‌ها عبارتند از: روش فرکانس سند(DF[11])، روش بهره اطلاعاتی(IG) [12]، روش اطلاعات متقابل (MI[13])، روش CHI[14]، روش ضریب همبستگی، روش SCHI[15] و روش Relief-F که در ادامه هر یک از آنها بطور مختصر معرفی می گردد.

روش فرکانس سند (DF)

فرکانس سند برای هر کلمه، برابر با تعداد سندهائی است که آن کلمه در آنها ظاهر شده است[۶] . برای هر کلمه واحد در مجموعه سندهای آموزش این ضابطه محاسبه شده و نهایتاً خصیصه‌هائی(بعبارت دیگر کلماتی) که بالاترین مقدار آستانه را دارند، انتخاب می‌شوند. این روش، روشی مقیاس‌پذیر برای حتی مجموعه داده‌های بسیار بزرگ می‌باشد و دارای پیچیدگی خطی متناسب با تعداد سندهای آموزش است.

روش بهره اطلاعاتی (IG)

بهره اطلاعاتی بعنوان ضابطه کیفیت خصیصه، بطور متواتر در روش‌های یادگیری ماشین استفاده شده‌است[۶] و دارای ضابطه‌ای به صورت زیر می‌باشد:

$$IG(t,c) = P(t,c) * Log \frac{P(t,c)}{P(t) * P(c)} + P(\bar{t},c) * Log \frac{P(\bar{t},c)}{P(\bar{t}) * P(c)}$$

(۱)

که در آن P(t, c) مقدار احتمالی است که در یک سند X از مجموعه آموزش، خصیصه t ظاهر شده و سند X متعلق به دسته c باشد. و همچنین $P(\bar{t},c)$ نشان دهنده مقدار احتمالی است که در یک سند X از مجموعه آموزش، خصیصه t ظاهر نشود و سند X متعلق به دسته c باشد. مقادیر $P(t)$، $P(\bar{t})$ و $P(c)$ به ترتیب نشان دهنده احتمال وقوع خصیصه، احتمال عدم وقوع خصیصه و احتمال وقوع دسته در سند X می باشد. برای هر کلمه t در هر دسته c، این مقدار محاسبه شده و نهایتاً ماکزیمم و یا میانگین آن مقادیر، بعنوان بهره اطلاعاتی آن کلمه منظور شده و بالاترین آنها انتخاب می‌گردد.

روش اطلاعات متقابل (MI)

اطلاعات متقابل ضابطه‌ای است که بطور عمومی در مدل کردن آماری بکار می رود[۱۲] و دارای ضابطه‌ای به صورت زیر می‌باشد:

$$MI(t,c) = Log \frac{P(t,c)}{P(t) * P(c)}$$

(۲)

چنانچه t و c مستقل از هم باشند، مقدار این ضابطه صفر می باشد. برای هر کلمه t در هر دسته c این مقدار محاسبه شده و نهایتا ماکزیمم و یا میانگین آن مقادیر، بعنوان اطلاعات متقابل آن کلمه منظور شده و بالاترین آنها انتخاب می‌گردد.

روش CHI

این روش علاوه بر استفاده از اطلاعات همبستگی مثبت از اطلاعات همبستگی منفی نیز در وزن دهی به خصیصه‌ها استفاده می‌کند[۷] و دارای ضابطه‌ای به صورت زیر است:

$$CHI(t,c) = \frac{g * [P(t,c) * P(\bar{t},\bar{c}) - P(\bar{t},c) * P(t,\bar{c})]^2}{P(t) * P(c) * P(\bar{t}) * P(\bar{c})}$$

(۳)

که در آن g تعداد سندها را نشان می دهد و $P(\bar{t},\bar{c})$ نشان دهنده احتمال آنست که در یک سند X متعلق به مجموعه آموزش، خصیصه t ظاهر نشود 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 شده و بالاترین آنها انتخاب می‌گردد.

روش ضریب همبستگی[16]

این روش از جذر ضابطه روش CHI، برای محاسبات خود استفاده می‌کند. با این کار نقش همبستگی‌های مثبت میان خصیصه‌ها و دسته‌ها را نسبت به همبستگی‌های منفی، اهمیت بیشتری می‌دهد[۱۳].
بنابراین ضابطه به صورت ساده شده زیر تبدیل می‌گردد:

$$CC(t,c) = \frac{\sqrt{g} * [P(t,c) * P(\bar{t},\bar{c}) - P(\bar{t},c) * P(t,\bar{c})]}{\sqrt{P(t) * P(c) * P(\bar{t}) * P(\bar{c})}}$$

(۴)

که در آن g تعداد سندها را نشان می‌دهد. برای هر ترم t در هر کلاس، این مقدار محاسبه شده و نهایتا ماکزیمم و یا میانگین آن مقادیر، بعنوان ضابطه ضریب همبستگی آن ترم منظور شده و بالاترین آنها انتخاب می‌گردد.

روش SCHI

ضابطه روش SCHI همان ضابطه ساده شده روش CHI است که از جذر ضابطه روش CHI برای محاسبات خود استفاده می‌کند. بعبارتی نقش همبستگی مثبت میان کلمه و سند را اهمیت بیشتری می‌دهد و همبستگی منفی را اهمیت کمتری می‌دهد، بعلاوه مقادیر $\sqrt{g}$ و عوامل قرار گرفته در مخرج کسر را حذف می‌کند[۷] . بنابراین ضابطه به صورت ساده شده زیر تبدیل می‌گردد:

$$SCHI(t,c) = P(t,c) * P(\bar{t},\bar{c}) - P(\bar{t},c) * P(t,\bar{c})$$

(۵)

روش Relief-F

روش انتخاب خصیصه Relief (شکل ۱) از دو مفهوم near-miss و near-hit استفاده می‌کند و با توجه به آنها وزن هر یک از خصیصه‌ها را به روز کرده و با استفاده از آنها، خصیصه‌هائی را که وزن بیشتری دارند انتخاب می‌کند. در واقع این روش، یک روش آماری را برای انتخاب خصیصه‌های مرتبط بکار می‌گیرد. این روش کارایی بسیار خوبی در حذف خصیصه‌ها در اسناد غیر متنی از خود نشان داده است [۹] .

```
Set all weights W[A] = 0.0;
For I = 1 to m do Begin
   Randomly select an instance R;
   Find nearest hit H and nearest miss M;
   For A = 1 to all attributes do
     W[A] = W[A] – diff(A, R, H)/m +
            diff(A, R, M)/m;
End;
```

شکل ۱ : روش انتخاب خصیصه فیلتری Relief .

که در آن m، تعداد نمونه‌هائی را نشان می‌دهد که به منظور بروزرسانی وزن‌ها مورد استفاده قرار می‌گیرد. همانطور که در شکل (۱) دیده می‌شود این روش برای هر سند انتخابی، مقادیر near-hit و near-miss از آن سند را پیدا کرده و فاصله اقلیدسی بین آن سند و سندهای پیدا شده را محاسبه کرده و مقدار فاصله با سند near-hit را با علامت منفی (یعنی diff(A, R, H)) و مقدار فاصله با سند near-miss را با علامت مثبت (یعنی diff(A, R, M)) لحاظ می‌کند.

روش Relief-F برای مسائل با چندین دسته، توسعه‌ای از روش Relief است. در این روش برای هر دسته c به غیر از دسته مربوط به سند انتخابی، یک مقدار M(c) محاسبه شده و میانگین این مقادیر بعنوان near-miss جهت بروزرسانی وزن استفاده می‌شود[۱۳] .

$$W[A] = W[A] - diff(A, R, H)/m \sum_{c \neq class(R)} [P(c).diff(A, R, M(c)]/m +$$

از میان خصیصه‌های وزن داده شده، تعدادی با بیشترین وزن، انتخاب می‌گردند.

روش‌های روکشی

این گروه از روش‌ها، از دسته‌بند بعنوان تابع ارزیابی استفاده کرده و به بررسی تاثیر هر یک از خصیصه‌ها در دقت دسته‌بندی می‌پردازند. روش‌های زیادی در این گروه وجود دارد [۸] ، شامل روش‌های انتخاب خصیصه روبه جلو SFS ، انتخاب خصیصه روبه عقب SBS ، DTM ، RC و PRESET که مهمترین آنها، روش‌های روبه جلو SFS و روبه عقب SBS می‌باشد که در ادامه بطور مختصر توضیح داده می‌شوند.

روش روبه جلوSFS

روش SFS از یک مجموعه با تعداد صفر خصیصه شروع کرده، بعد تمامی زیرمجموعه‌ها را با دقیقاً یک خصیصه ارزیابی می‌کند و یکی از زیرمجموعه‌ها را با بهترین کارائی دسته‌بندی انتخاب می‌کند، سپس به این مجموعه خصیصه‌ای را اضافه می‌کند که باعث بهترین کارائی دسته‌بندی گردد. این چرخه تا زمانی‌که هیچ‌گونه بهبودی از گسترش مجموعه خصیصه‌ها حاصل نگردد، تکرار می‌شود.

روش روبه عقب SBS

روش SBS با تمامی خصیصه‌ها شروع کرده و مکرراً یک خصیصه را حذف می‌کند، بطوریکه بیشترین بهبود

کارائی دسته‌بندی را داشته باشد. . این چرخه تا زمانی‌که هیچ‌گونه بهبودی از کاهش مجموعه خصیصه‌ها حاصل نشود، تکرار می‌گردد.

همبستگی بین خصیصه‌ها و دسته‌ها

همانطور که در بخش اول اشاره شد، بمنظور دسته‌بندی سندها بطور خودکار در دسته‌های از پیش تعییین شده، سندها به دو دسته سندهای آموزشی و سندهای آزمون تقسیم می‌شوند. از سندهای آموزشی برای یادگیری دسته‌بند و از سندهای آزمون برای ارزیابی دقت و بازخوانی و بالاخره قدرت آن دسته‌بند استفاده می‌شود. فرض کنید، f خصیصه‌ای از فضای خصیصه‌ها، x سندی از مجموعه سندهای آموزش و c نیز نشان دهنده یک دسته خاص در مجموعه دسته‌ها باشد. کلیه همبستگی‌هائی را که می‌توان بین خصیصه f و دسته c درنظر گرفت به چهار گونه تقسیم می‌شوند که در جدول (۳) نشان داده شده‌است.

جدول۳ : همبستگی‌های بین خصیصه‌ها و دسته‌ها/

دسته	خصیصه	
	خصیصهf درسند x وجود دارد	خصیصهf درسند x وجود ندارد
سند x متعلق به دسته c است	A	C
سند x متعلق به دسته c نمی‌باشد	B	D

جدول (۳)، در اصل نشان دهنده دو نوع همبستگی مثبت و منفی بین خصیصه‌ها و دسته‌ها می‌باشد.

A بیان کننده حالتی است که خصیصه‌ای مثلf در سند x وجود داشته باشد و سند هم متعلق به دسته c باشد. B بیان کننده حالتی است که خصیصه‌ای مثلf در سند x وجود داشته باشد اما سند متعلق به دسته c نباشد. همین‌طور C بیان کننده حالتی است که خصیصه‌ای مثلf در سند x وجود نداشته باشد اما سند متعلق به دسته c باشد و نهایتا D نشان دهنده حالتی است که خصیصه‌ای مثلf در سند x وجود نداشته باشد و سند نیز متعلق به دسته c نباشد. همبستگی‌های A, D از نوع همبستگی مثبت و همبستگی‌های B, C از نوع همبستگی منفی می باشند. هر یک از این همبستگی ها به تنهائی اطلاعی در رابطه با چگونگی دسته‌بندی از روی خصیصه‌ها را به مجموعه دانش دسته‌بندی سندها اضافه می‌نماید در نتیجه استفاده از آن‌ها در ضابطه انتخاب خصیصه مناسب خواهد بود. با نگاه دقیق‌تر به روش‌های انتخاب خصیصه، ملاحظه می‌شود که اکثر این روش‌ها از ترکیب‌های مختلف این چهار همبستگی استفاده نموده‌اند. بعنوان مثال روش MI که تنها از همبستگی مثبت A استفاده می‌کند، کارائی قابل مقایسه‌ای در مقایسه با روش‌های دیگر انتخاب خصیصه ندارد، در حالی‌که روش IG که از دو همبستگی مثبت و منفی A وC استفاده می‌کند، ضابطه بسیار خوبی جهت وزن‌دهی به خصیصه‌ها و انتخاب آنها تعریف می‌کند. نهایتا روش CHI که از تمامی چهار همبستگی در ضابطه خود استفاده می‌کند، یکی از بهترین روش‌های انتخاب خصیصه‌ها در مسائل دسته‌بندی متون است.

روش‌های پیشنهادی انتخاب خصیصه

در این بخش، سه روش پیشنهادی انتخاب خصیصه معرفی می گردد. هدف اصلی روش‌های پیشنهادی، کاهش مجموعه بزرگ خصیصه‌ها (بعبارتی هزاران خصیصه) به یک زیرمجموعه کوچکی از خصیصه‌ها (بعبارتی صدها خصیصه) بدون از دست دادن توانائی سیستم در دقت دسته‌بندی سندها و یا حداکثر با از دست دادن کمی از دقت می‌باشد. روش‌های پیشنهادی اول و دوم انتخاب خصیصه برمبنای تحلیل همبستگی بین خصیصه‌ها و دسته‌ها ارائه می شوند و روش پیشنهادی سوم انتخاب خصیصه با بهره‌گیری از ترکیب دو روش فیلتری و روکشی ارائه می شود. در بخش ارزیابی روشهای پیشنهادی، جایگاه روش‌های پیشنهادی انتخاب خصیصه در مقایسه با روش‌های موجود انتخاب خصیصه با استفاده از روش‌ها دسته‌بندی SVM Light و روش دسته‌بندی بیزین ساده ارائه شده است.

انتخاب خصیصه با استفاده از همبستگی بین خصیصه‌ها و دسته‌ها

با توجه به اینکه هر یک از همبستگی های مثبت و منفی، نقش موثری در ضابطه انتخاب خصیصه داشته و نیز اطلاعی راجع به دسته‌بندی به مجموعه اطلاعات ما اضافه می‌نماید، بنابراین مناسب است که کلیه این

همبستگی ها در ضابطه انتخاب خصیصه در نظر گرفته شوند. آزمایشات مختلف روی همبستگی‌های چهارگانه مثبت و منفی نشان می‌دهد[۱۵] که در صورت استفاده از هر یک از این همبستگی‌ها به تنهائی در ضابطه انتخاب خصیصه‌ها نتایج خوبی بدست نمی آید، در نتیجه لازم است ترکیبی از آنها با وزن‌های مختلف استفاده شود. به عنوان مثال ضابطه MI که تنها از یک همبستگی مثبت در ضابطه انتخاب خصیصه استفاده کرده است، نتایج خوبی در آزمایشات نشان نمی دهد (بخصوص در حذف تعداد زیاد خصیصه، کارائی روش بشدت افت می کند) در حالیکه روش IG با در نظر گرفتن دو همبستگی مثبت و منفی نتایج بهتری در حذف خصیصه ها دارد و نهایتا روش های CHI و SCHI که کلیه همبستگی های مثبت و منفی را در ضابطه خود لحاظ کرده اند، از بهترین روش های حذف خصیصه می باشند. بنابراین در این مقاله ضابطه کلی روش انتخاب خصیصه پیشنهادی در ضابطه FS نشان داده شده است. همانطور که در این ضابطه مشخص شده است، همبستگی های مختلف مثبت و منفی با ضرایب α_1، α_2، β_1 و β_2 آورده شده اند. این ضرایب، پارامترهای تنظیم وزن هر یک از همبستگی ها میباشد که با توجه به درجه اهمیت آنها، میتواند مقادیر مختلفی داشته باشد. به عنوان مثال در ضابطه IG ضرایب α_1 و β_1 برابر با یک و ضرایب α_2 و β_2 برابر با صفر می باشد و در ضابطه MI، ضریب α_1 برابر با یک و سه ضریب دیگر برابر با صفر هستند.

ضابطه(FS)- ضابطه کلی روش انتخاب خصیصه پیشنهادی

$$Fs(t,c)=\alpha_1 * P(t,c) * Log\frac{P(t,c)}{P(t)*P(c)}+$$
$$\alpha_2 * P(\bar{t},\bar{c}) * Log\frac{P(\bar{t},\bar{c})}{P(\bar{t})*P(\bar{c})}+$$
$$\beta_1 * P(\bar{t},c) * Log\frac{P(\bar{t},c)}{P(\bar{t})*P(c)}+$$
$$\beta_2 * P(t,\bar{c}) * Log\frac{P(t,\bar{c})}{P(t)*P(\bar{c})}$$

در ضابطه FS منظور از P(t, c) احتمال اینست که با مشاهده خصیصه t در سند x، سند متعلق به دسته c باشد(همبستگی مثبت)، $P(\bar{t},c)$ مقدار احتمالی را نشان می‌دهد که با عدم مشاهده خصیصه t در سند x، سند متعلق به دسته c باشد(همبستگی منفی)، $P(t,\bar{c})$ مقدار احتمالی را نشان می‌دهد که با مشاهده خصیصه t در سند x، سند متعلق به دسته c نباشد(همبستگی منفی) و نهایتاً ، $P(\bar{t},\bar{c})$ مقدار احتمالی را نشان می‌دهد که با عدم مشاهده خصیصه t در سند x، سند نیز متعلق به دسته c نباشد(همبستگی مثبت). در این مقاله از بین ترکیب‌های مختلف اولیه که مورد آزمون و بررسی قرار گرفته، دو ترکیب موثر از این همبستگی‌ها در قالب دو ضابطه مختلف معرفی شده است. در ضابطه اول (FS1) کلیه همبستگی های مثبت و منفی با ضریب یک و در ضابطه دوم (FS2) همبستگی های مثبت با ضریب یک و همبستگی های منفی با ضریب منفی یک اثر داده شده است. ایده درضابطه FS1 این است که از کلیه همبستگی های چهارگانه در وزن‌دهی به خصیصه‌ها با یک وزن یکسان استفاده کنیم، چرا که تمامی همبستگی ها، نقش تعیین کننده‌ای در تعیین دسته‌ها دارند. درحالی‌که در ضابطه FS2 نسبت به ضابطه اول، اهمیت همبستگی‌های مثبت حفظ شده، اما نقش همبستگی‌های منفی، وارون شده است. به عبارت دیگر با در نظر گرفتن ضرایب منفی در ضابطه FS2، در عمل یک تعادل نسبی بین همبستگی‌های مثبت و همبستگی‌های منفی برقرار شده است.

انتخاب خصیصه به روش ترکیبی R+S

همانطور که گفته شد، روش‌های انتخاب خصیصه روکشی دارای دقت بسیار زیادی هستند اما در مقابل، هزینه محاسباتی بالائی هم دارند و روش‌های انتخاب خصیصه فیلتری هزینه محاسباتی کمی دارند اما در مقابل دقت آن‌ها قابل پیش‌بینی نیست. به منظور استفاده از مزیت‌های هر دو نوع روش، روش R+S پیشنهاد شده‌است. این روش، یک روش ترکیبی است که در ادامه بطور مختصر توضیح داده می شود. ابتدا، خصیصه‌های نامرتبط[۱۷] (خصیصه‌هائی که اطلاعات مفیدی در رابطه با تعیین دسته سندها ندارند) با استفاده از روش فیلتری Relief-F ، با هزینه محاسباتی کمی حذف می‌شوند. حذف خصیصه‌ها تا موقعی که بهبود در دقت دسته‌بندی بوجود می‌آید، ادامه پیدا می‌کند. سپس، یکی از روش‌های انتخاب خصیصه روکشی SFS یا SBS روی خصیصه‌های

انتخاب شده اعمال می‌شود تا خصیصه‌های نامرتبط با دقت بیشتر دربرابر هزینه محاسباتی بالاتر، حذف گردند [۱۷] . ایده این روش این است که هر یک از گام‌های آن در واقع قصد دارد خصیصه‌ها را فیلتر کند تا نهایتاً یک زیرمجموعه کوچکی از خصیصه‌ها باقی بماند.

در [۱۴] گزارش شده‌است که SBS کارائی بهتری نسبت به روش SFS دارد و علت آن را می‌توان در ماهیت روش SBS دانست که نقش یک خصیصه داده شده را در مقایسه با تمامی خصیصه‌های دیگر بررسی می کند. متاسفانه روش SBS برای یک مجموعه سند با تعداد زیادی خصیصه، یک روش عملی و امکانپذیر نیست، بنابراین زمانی‌که تعداد خصیصه‌ها بعد از حذف خصیصه‌های نامرتبط (توسط روش Relief-F) تعداد زیادی باشد، بایستی از الگوریتم SFS استفاده گردد.

در اینجا یادآوری می‌گردد، ما ابتدا از روش انتخاب خصیصه FS2 در روش ترکیبی استفاده کرده بودیم، اما با توجه به شکل (۸) که روش آماری انتخاب خصیصه Relief-F (که اولین بار است که در دامنه متون استفاده می‌شود) نقطه ماکزیمم بالاتری نسبت به روش انتخاب خصیصه FS2 داشت، در نتیجه در روش ترکیبی R+S در بخش فیلتری آن از روش Relief-F استفاده کردیم.

محیط ارزیابی

به منظور ارزیابی روش نمایش ترکیبی و روش‌های انتخاب خصیصه پیشنهادی در این بخش ابتدا روش‌های دسته‌بندی و مجموعه‌های داده‌ای مورد استفاده در آزمایشات، معرفی شده، سپس معیارهای سنجش کارائی دسته‌بندها معرفی می‌گردد.

دسته‌بندها

برای ارزیابی تاثیر روش‌های نمایش سندها و بررسی تاثیر روش‌های انتخاب خصیصه از دسته‌بندهای SVM Light و بیزین ساده استفاده شده‌است. این دسته‌بندها ازشناخته شده‌ترین دسته‌بندها در زمینه دسته‌بندی متون می‌باشند. دسته‌بند SVM Light کاراترین روش دسته‌بندی بوده که توسط Joachims ارائه شده است [۴] . دسته‌بند بیزین ساده نیز یکی از دسته‌بندهای با کارائی بالا بوده که جهت انجام آزمایشات بکار گرفته شده‌است[۱۶] . مجموعه داده‌های مختلف با استفاده از روش‌های دسته‌بندی فوق، دسته‌بندی می‌گردند و نتایج بدست آمده از آن‌ها، جهت ارزیابی روش‌های پیشنهادی در این مقاله مورد استفاده قرار می‌گیرد.

جدول ۴ : مجموعه‌های داده انتخابی از رویتر.

مجموعه‌های داده	تعداد کل سندها	تعدادسندهای آموزشی	تعدادسندهای آزمون	تعداد دسته‌ها	تعداد خصیصه تک کلمه	تعداد خصیصه دو کلمه‌ای یکپارچه	تعداد خصیصه دو کلمه‌ای منقطع
# ۱	۲۱۴۵۰	۱۴۷۰۴	۶۷۴۶	۱۳۵	۲۶۸۳۲	۲۶۸۳۱	۱۷۹۵۴
# ۲	۱۴۳۴۷	۱۰۶۶۷	۳۶۸۰	۹۳	۱۸۳۸۴	۱۸۳۸۳	۱۲۴۵۲
# ۳	۱۳۲۷۲	۹۶۱۰	۳۶۶۲	۹۲	۱۷۷۳۹	۱۷۷۳۹	۱۱۸۷۵
# ۴	۱۲۹۰۲	۹۶۰۳	۳۲۹۹	۹۰	۱۴۸۰۹	۱۴۸۰۸	۱۱۰۲۰
# ۵[۱]	۱۲۹۰۲	۹۶۰۳	۳۲۹۹	۱۰	۱۶۲۰۸	۱۶۲۰۷	۱۱۶۵۲
# ۶	۱۲۰۰۰	۹۰۰۰	۳۰۰۰	۱۳۵	۱۶۷۹۴	۱۶۷۹۳	۱۱۷۵۹
# ۷	۱۲۰۰۰	۹۰۰۰	۳۰۰۰	۸۵	۱۴۸۹۰	۱۴۸۸۹	۱۰۸۷۵

مجموعه‌های داده

از دو مجموعه داده برای ارزیابی روش‌های پیشنهادی در این مقاله دربرابر سایر روش‌های نمایش سندها و انتخاب خصیصه، استفاده شده است: مجموعه داده رویتر۲۱۵۷۸ [۱۰] و مجموعه داده 20-Newsgroups [۱۹] . مجموعه اخبار رویتر بطور گسترده‌ای جهت انجام تحقیقات دسته‌بندی متون مورد استفاده قرار گرفته است. تعداد کل سندهای دارای برچسب دسته، در مجموعه کامل داده‌های رویتر، ۲۱۴۵۰ سند می‌باشد که کمتر از نیمی از سندها توسط انسان برچسب موضوع خورده‌است. در این تحقیق، فقط از سندهائی استفاده شده‌است که حداقل یک موضوع داشته‌اند. سپس آنها را به طور تصادفی به مجموعه‌های دوتائی جهت انجام آموزش و آزمون تقسیم بندی کرده‌ایم. مجموعه داده 20-Newsgroups نیز شامل ۲۰ دسته از اطلاعات گروه‌های خبری بوده که در هر دسته تعداد ۱۰۰۰ سند در رابطه با موضوع هم نام دسته می‌باشد. تعداد ۱۰ مجموعه داده از مجموعه داده‌های رویتر و 20-Newsgroups به منظور ارزیابی کارائی هریک از روش‌ها استفاده شده‌است. جدول

[۱] ده طبقه با بیشترین تکرار سند در این مجموعه داده‌ای در نظر گرفته شده است.

(۴)، هفت مجموعه داده مختلف انتخابی از داده‌های رویتر را نشان می دهد، درحالی‌که جدول (۵) نشان دهنده سه مجموعه داده مختلف از داده‌های 20-Newsgroups می‌باشد.

جدول ۵: مجموعه‌های داده انتخابی از 20-Newsgroups .

مجموعه‌های داده	تعداد کل سندها	تعدادسندهای آموزشی	تعدادسندهای آزمون	تعداد دسته‌ها	تعداد خصیصه تک کلمه	تعداد خصیصه دو کلمه‌ای یکپارچه	تعداد خصیصه دو کلمه‌ای منقطع
#۱	۱۳۰۰۰	۱۰۰۰۰	۳۰۰۰	۲۰	۸۳۲۹	۸۳۲۸	۵۶۲۵
#۲	۱۳۰۰۰	۸۰۰۰	۵۰۰۰	۲۰	۷۰۲۸	۷۰۲۷	۵۰۷۸
#۳	۱۳۰۰۰	۱۱۰۰۰	۲۰۰۰	۲۰	۹۳۵۶	۹۳۵۵	۷۲۱۴

در هر یک از مجموعه‌های داده استفاده شده، تعداد خصیصه‌های تک کلمه ای، تعداد خصیصه‌های دوکلمه ای یکپارچه و تعداد خصیصه‌های دو کلمه ای منقطع نیز آورده شده است. در اینجا منظور از خصیصه‌های دوکلمه ای یکپارچه، خصیصه‌هائی است که بدون درنظر گرفتن علائم نوشتاری جملات در متون، بصورت ترکیب دوتائی انتخاب می‌شوند. این دسته از خصیصه‌های دو کلمه ای دارای ترکیب‌های بی‌معنی زیاد می‌باشند. همچنین منظور از خصیصه‌های دوکلمه ای منقطع، خصیصه‌هائی است که با در نظر گرفتن علائم نوشتاری جملات در متون مانند نقطه، کاما، سمیکلن و... با دقت بیشتری بصورت ترکیب دوتائی کلمات انتخاب می‌شوند. بنابراین ترکیب‌های معنی‌دارتری نسبت به حالت یکپارچه دارند.

سنجش کارائی

بطور کلی، روش‌های انتخاب خصیصه بعنوان مرحله‌ای قبل از یادگیری دسته‌بندها اعمال می‌گردند. تأثیر روش‌های انتخاب خصیصه و نیز روش‌های نمایش سندها با استفاده از سنجش کارائی هر یک از دسته‌بندها، روی سندها بدست می‌آید. به منظور سنجش کارائی از تعاریف استاندارد دقت[18] (P)، بازخوانی[19] (R)و تابع F_β استفاده می‌گردد.

$$P = \frac{\text{تعداد دسته‌های پیدا شده صحیح}}{\text{تعداد کل دسته‌های پیدا شده}}$$

$$R = \frac{\text{تعداد دسته‌های پیدا شده صحیح}}{\text{تعداد کل دسته‌های صحیح}}$$

$$F_\beta = \frac{(\beta^2+1)\times P\times R}{\beta^2\times P+R}$$

رابطه P میزان دقت در انجام دسته‌بندی‌ها را نشان می‌دهد و رابطه R میزان کامل بودن مجموعه یافت شده را نشان می‌دهد. بالا بودن هر دوی این عوامل نشان دهنده میزان کیفیت بالای روش دسته‌بندی است. هر چند بالا رفتن مقدار دقت، معمولاً با کاهش مقدار بازخوانی همراه است و بالا رفتن مقدار بازخوانی، معمولاً کاهش میزان دقت را در بر دارد.

با توجه به نوع کاربرد، گاهی بالا بودن دقت از اهمیت ویژه‌ای برخوردار است و گاهی نیز بالا بودن بازخوانی مطلوب می‌باشد. بنابراین به منظور وزن‌دهی‌های مختلف به دقت و بازخوانی می‌توان از ضابطه F_β استفاده نمود و با توجه به نوع کاربرد و اهمیت هر یک از این دو فاکتور(دقت و بازخوانی) می‌توان وزن‌های مختلفی به آنها داد. در بسیاری از پژوهش‌های دانشگاهی از ضابطه F_1 که به هر دوی این عوامل وزن یکسانی می‌دهد، استفاده می‌شود. در این مقاله نیز از این ضابطه به منظور ارزیابی‌های مختلف استفاده شده‌است.

ارزیابی روش نمایش ترکیبی

در این بخش، به منظور ارزیابی روش‌های نمایش اسناد از ۱۰ مجموعه داده که در بخش مجموعه های داده معرفی شد، استفاده شده‌است. جهت بررسی این مسأله که با در نظر گرفتن ترکیب‌های دوتایی کلمات، بهبودی در دسته‌بندی اتفاق می‌افتد یا خیر، آزمون‌های مختلفی روی مجموعه‌های داده بخش مجموعه های داده انجام گرفته که نتایج F1 حاصل از آنها در جدول (۶) ارائه شده‌است.

با توجه به جدول (۶) ملاحظه می‌شود که در نظر گرفتن فضای خصیصه‌های دو کلمه ای، بدون درنظر گرفتن علائم نوشتاری جملات، بدلیل اینکه ترکیب‌های نامفهوم بسیار زیادی در این حالت وجود دارد، باعث افت F1 در دسته‌بند SVM Light می‌گردد. اما با در نظر گرفتن علائم نوشتاری جملات در نمایش متون، در شش مورد از آزمایشات ۱۰گانه، نتایج بهتری نسبت به فضای خصیصه‌های تک کلمه‌ای بوجود می‌آید، که نشان دهنده مرتبط بودن ترکیب‌های دوتائی کلمات با کارایی دسته‌بندی می‌باشد. با در نظر گرفتن فضای خصیصه‌های ترکیبی، شامل تک کلمه و ترکیب‌های دوتائی کلمات

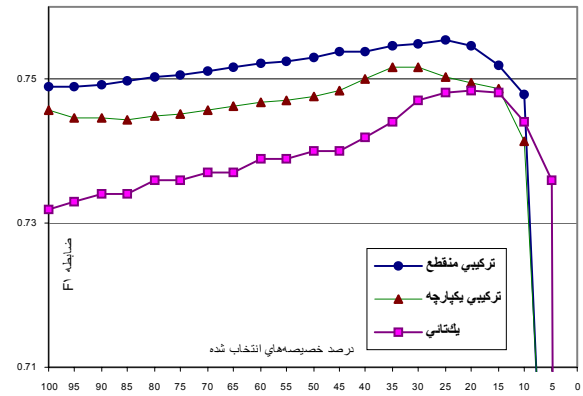
یکپارچه، نتایج به طور کلی در هفت مورد از آزمایشات بهتر از فضای خصیصه‌های تک کلمه‌ای بوده و این نشان می‌دهد که هنوز مفاهیمی نیز مانند: موصوف صفت، مضاف مضاف الیه و... در سندهای متنی وجود دارد که در دسته‌بندی، مهم بوده و با استفاده از فقط تک کلمه مدل نمی‌شود. بالاخره بهترین مورد کارائی مربوط به حالت ترکیبی است که مجموعه خصیصه‌های تک کلمه ای را به همراه خصیصه‌های ترکیب دوتائی کلمات منقطع، با درنظر گرفتن علائم نوشتاری جملات، جهت نمایش سندها، استفاده می‌کند که این تائیدی بر قدرت روش نمایش ترکیبی اسناد در بخش فضای خصیصه های مقاله است.

جدول ۶ : مقادیر F1 حاصل از دسته‌بند SVM Light روی نمایش‌های مختلف متون.

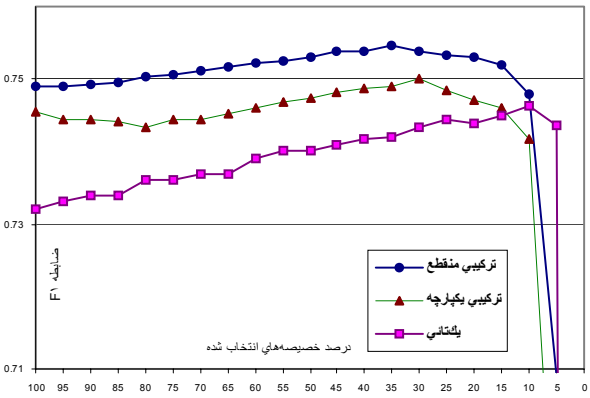
مجموعه‌های داده‌ای	فضای تک کلمه ای	فضای دو کلمه‌ای		فضای ترکیبی	
		یکپارچه	منقطع	یکپارچه	منقطع
# ۱	0.687	0.587	0.668	0.69	0.7
# ۲	0.669	0.569	0.677	0.709	0.712
# ۳	0.786	0.756	0.787	0.774	0.788
# ۴	0.697	0.58	0.641	0.677	0.686
# ۵	0.827	0.745	0.829	0.836	0.832
# ۶	0.751	0.686	0.763	0.781	0.791
# ۷	0.756	0.681	0.731	0.745	0.758
# ۸	0.822	0.725	0.805	0.834	0.832
# ۹	0.587	0.6	0.655	0.594	0.599
# ۱۰	0.743	0.733	0.788	0.791	0.797

از طرفی با توجه به جدول (۶)، ملاحظه می‌گردد که تنها درنظر گرفتن ترکیب‌های دوتائی کلمات و عدم استفاده از تک کلمه‌ها به عنوان خصیصه باعث افت ضابطه F1 می‌گردد. علت افت این مقدار شاید بدلیل اینست که در این حالت مهمترین خصیصه‌ها در سندها، که همانا "اسامی" هستند، نادیده گرفته می‌شوند. بنابراین باتوجه به نتایج بدست آمده، استفاده از خصیصه‌های تک کلمه ای و ترکیب‌های دوتائی کلمات، بطور کلی باعث بهبود دقت دسته‌بندی شده، بخصوص در حالتی‌که ترکیب‌های بامعنی با استفاده از علائم نوشتاری جملات در زبان، ایجاد گردد.

با توجه به فضای بزرگ خصیصه‌ها، بخصوص در حالت ترکیبی، برای کاهش آن‌ها می‌توان از روش‌های انتخاب خصیصه استفاده نمود. با استفاده از روش‌های انتخاب خصیصه IG و CHI همانطور که در شکل‌های (۲) و (۳) دیده می‌شود، می‌توان تعداد زیادی از خصیصه‌ها را (در حدود ۸۶٪) حتی با بهبود در مقدار F1 (کارائی دسته‌بند) حذف نمود.



شکل ۲ : تاثیر روش انتخاب خصیصه CHI روی کارائی دسته‌بند SVM-Light در نمایش‌های مختلف سندها.

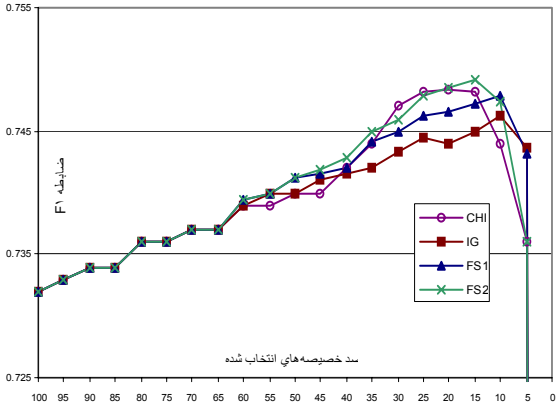


شکل ۳ : تاثیر روش انتخاب خصیصه IG روی کارائی دسته‌بند SVM-Light در نمایش‌های مختلف سندها.

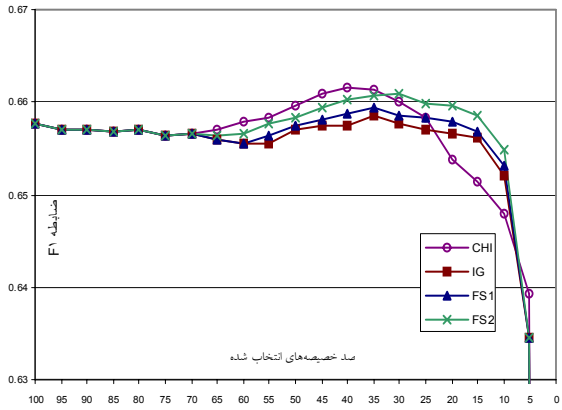
بنابراین با استفاده از یک روش مناسب انتخاب خصیصه می‌توان فضای اولیه را با بهبود در مقدار F1 به صورت چشم‌گیری کاهش داده و باعث بهبود در کارائی روش‌های دسته‌بندی شد.

ارزیابی روش‌های پیشنهادی انتخاب خصیصه

در این بخش ابتدا دو روش انتخاب خصیصه FS1 و FS2 و سپس روش انتخاب خصیصه ترکیبی R+S انتخاب خصیصه مورد ارزیابی قرار می گیرند.



شکل ۴: کارائی دسته‌بند SVM Light در تعدادی از روش‌های انتخاب خصیصه.



شکل ۵ :کارائی دسته‌بند بیزین ساده در تعدادی از روش‌های انتخاب خصیصه.

ارزیابی روش‌های FS1 و FS2

در این بخش به منظور ارزیابی روش‌های انتخاب خصیصه از ۱۰ مجموعه داده که در بخش مجموعه های داده معرفی شد، استفاده شده‌است. جهت ارزیابی روش‌های پیشنهادی انتخاب خصیصه آزمون‌های مختلفی روی ۱۰ مجموعه داده‌ای بخش مجموعه های داده به تفکیک دو دسته‌بند SVM Light و بیزین ساده انجام گرفته که نتایج F1 حاصل از آنها در شکل‌های (۴) و (۵) نشان داده شده است. در این شکل‌ها، درصد خصیصه انتخاب شده روی محور طول‌ها و میانگین F1 ده مجموعه داده روی محور عرض‌ها نشان داده شده است. با استفاده از مجموعه‌های داده جداول (۴) و (۵)، آزمون‌هائی روی تک‌تک همبستگی‌های چهارگانه مثبت و منفی بین خصیصه‌های سندها و دسته‌ها انجام گرفته [۱۸] که نتایج آزمایشات دسته‌بندی حاکی از آن است که همانند روش MI این روش‌ها، روش مناسبی جهت انتخاب خصیصه نمی‌باشند. این روش‌ها، نتایج ضعیفی را بخصوص در حذف بیشتر خصیصه‌های نامرتبط از خود نشان می‌دهند، که این مطلب نشان دهنده این است که در نظر گرفتن تنها یک همبستگی اگرچه برای تصمیم‌گیری در رابطه با دسته سندها لازم است، اما کافی نبوده و هنوز نیاز به استفاده از اطلاعات بیشتری برای تصمیم گیری خوب داریم. بنابراین اولین نتیجه‌گیری این بخش عبارتست از :

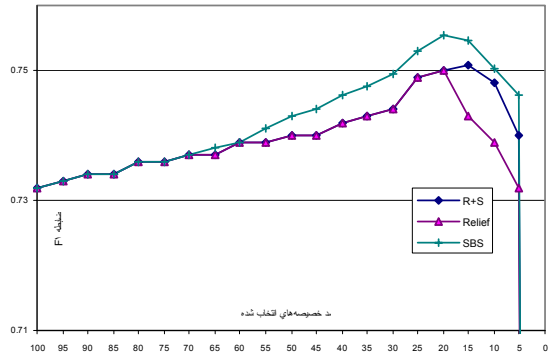
> مقادیر ضرایب همبستگی‌های مثبت و منفی (یعنی α_1، α_2، β_1 و β_2 در ضابطه FS) بایستی یک مقدار غیر صفر باشد. (نتیجه‌گیری ۱)

با توجه به نتایج بدست آمده از آزمایشات (شکل‌های ۴ و ۵)، ملاحظه می‌شود که روش‌های انتخاب خصیصه پیشنهادی (FS1 و FS2) با داشتن هزینه محاسباتی بالاتر ازIG، نتایج بهتری را نسبت به این روش در هر دو دسته‌بند، از خود نشان می‌دهند که این مجدداً مؤید (نتیجه‌گیری ۱) است. علاوه بر آن روش پیشنهادی FS2 نتایج بهتری را نسبت به روش انتخاب خصیصه CHI در دسته‌بند SVM Light (شکل ۴) نشان داده‌است. در شکل (۵) نیز ملاحظه می‌شود که روش انتخاب خصیصه FS2 دارای ماکزیمم پائین‌تری نسبت به روش انتخاب خصیصه CHI است اما در آستانه‌های حذف بالاتر خصیصه‌ها، روش انتخاب خصیصه FS2 دارای مقدار F1 بیشتری نسبت به CHI می‌باشد. این مطلب (با توجه به شکل‌های ۶ و ۸) برای روش انتخاب خصیصه FS2 در مقایسه با روش Relief-F نیز صادق است. بنابراین، روش‌های پیشنهادی، روش‌هائی قابل مقایسه با روش‌های بسیار خوب در انتخاب خصیصه‌ها می‌باشند. این موضوع نیز (نتیجه‌گیری ۱) را تأیید می‌کند.

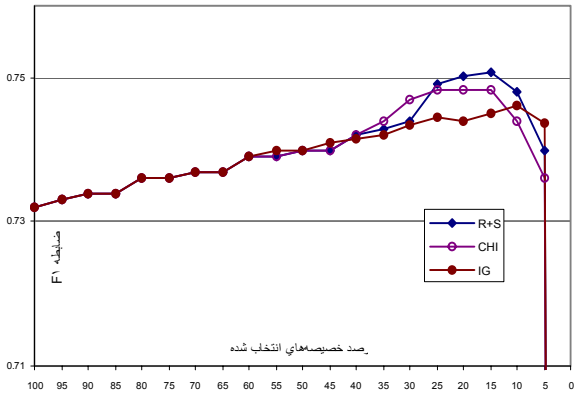
همچنین، همانطور که در شکل‌های (۴) و (۵) ملاحظه می‌شود، بالاتر بودن منحنی کارائی روش انتخاب خصیصه FS2، نسبت به کارائی روش FS1، در هر دو دسته‌بند SVM Light و بیزین ساده نشان دهنده تاثیر بهتر نقش منفی همبستگی‌های منفی در ضابطه FS2 است. در ضابطه CHI و دو گونه دیگر آن نیز هر چهار همبستگی مثبت و منفی در نظر گفته شده است و علاوه بر آن همبستگی‌های منفی یک نقش تعدیل کننده در مقادیر ارزش خصیصه‌ها دارند. و این همان چیزی است که ما نیز در ضابطه FS2 شاهد آن هستیم. اما تفاوت ضابطه FS2 با

ضابطه CHI در نحوه ترکیب این چهار همبستگی می‌باشد. بنابراین دومین نتیجه‌گیری این بخش عبارت است از :

> بهتر است همبستگی‌های مثبت یک نقش مثبت (یعنی $\alpha_2>0$ و α_1 در ضابطه FS) و همبستگی‌های منفی یک نقش منفی (یعنی $\beta_2<0$ و β_1 در ضابطه FS) به عبارت دیگر تعدیل کننده در ضابطه‌های ارزش‌گذاری خصیصه‌ها داشته باشند. (نتیجه‌گیری ۲)



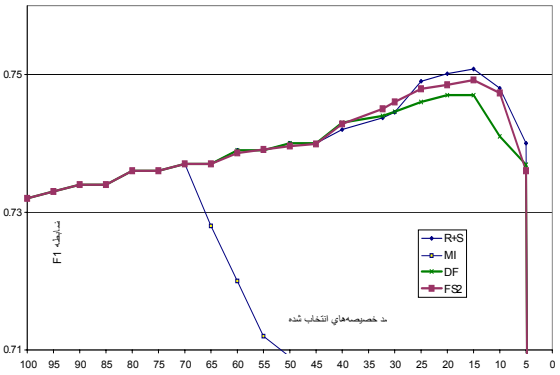
شکل ۶ : کارائی دسته‌بند SVM Light در تعدادی از روش‌های انتخاب خصیصه.



شکل ۷ : کارائی دسته‌بند SVM Light در تعدادی از روش‌های انتخاب خصیصه.

ارزیابی روش انتخاب خصیصه ترکیبی R+S

جهت ارزیابی روش انتخاب خصیصه ترکیبی R+S (در بخش ۶-۲)، آزمون‌های مختلفی روی ۱۰ مجموعه داده‌ای بخش ۷-۲ با استفاده از دسته‌بند SVM Light انجام گرفته که نتایج F1 حاصل از آن‌ها در شکل‌های (۶)، (۷) و (۸) نشان داده شده است. در این شکل‌ها، درصد خصیصه انتخاب شده روی محور طول‌ها و میانگین F1 ده مجموعه داده روی محور عرض‌ها نشان داده شده است.



شکل ۸ : کارائی دسته‌بند SVM Light در تعدادی از روش‌های انتخاب خصیصه.

با توجه به نمودار شکل‌های (۶)، (۷) و (۸) ملاحظه می‌شود که بیشترین مقدار F1 متعلق به روش انتخاب خصیصه روکشی (SBS) می‌باشد. روش انتخاب خصیصه روکشی، با داشتن هزینه محاسباتی بسیار بالا ، روشی بسیار دقیق برای حذف خصیصه‌ها می‌باشد. پس از آن، روش انتخاب خصیصه ترکیبی R+S کارائی تقریبا بهتری را نسبت به سایر روش‌های انتخاب خصیصه از خود نشان می‌دهد. روش انتخاب خصیصه Relief-F (که اولین بار است که در دامنه متون استفاده می‌شود) همانطور که در شکل (۶) ملاحظه می‌شود، کارائی بسیار خوبی در حذف خصیصه‌های متون از خود نشان می‌دهد و در نتیجه می‌تواند بعنوان روشی جهت حذف خصیصه در دامنه متون نیز مورد استفاده قرار گیرد. تمامی روش‌های انتخاب خصیصه، به غیر از روش MI، می‌توانند بالغ بر ۹۴٪ از خصیصه‌ها را ضمن بهبود کارائی اولیه دسته‌بندی (اندازه گیری شده بوسیله ضابطه F1) حذف کنند. به عنوان مثال با استفاده از روش انتخاب خصیصه Relief-F، کلمات مجموعه داده ششم از ۱۶۷۹۴ ترم، به ۱۵۷۹ ترم (حدودا ۹۴٪) کاهش یافته ضمن اینکه ضابطه F1 دسته‌بند SVM Light از ۷۵٫۱٪ به ۷۷٫۲٪ بهبود می‌یابد. در روش انتخاب خصیصه ترکیبی R+S می‌توان با توجه به اهمیت دقت دسته‌بندی و یا هزینه محاسباتی در کاربرد مورد نظر، در هر یک از قسمتهای حذف خصیصه، روش انتخاب خصیصه Relief-F را متوقف و حذف خصیصه‌ها را با دقت بیشتری توسط روش انتخاب خصیصه روکشی SBS ادامه داد. در این مقاله پس از اینکه روش انتخاب خصیصه Relief-F به مقدار بیشینه

F1 رسید، ادامه حذف خصیصه‌ها به روش انتخاب خصیصه روکشی SBS ادامه یافته است. همچنین با توجه به شکل (۷) ملاحظه می‌شود که استفاده از روش انتخاب خصیصه ترکیبیR+S باعث بهبود ضابطه F1 نسبت به روش‌های انتخاب خصیصه IG و CHI می‌شود، که بهترین روش‌های انتخاب خصیصه در دامنه متون هستند. روش انتخاب خصیصه MI قابل مقایسه با روش‌های دیگر انتخاب خصیصه نبوده و فقط تا ۳۰٪ حذف خصیصه‌ها نتایج مشابهی با سایر روش‌ها از خود نشان می‌دهد (شکل ۸)، بخصوص این روش در حذف بیشتر خصیصه‌ها با افت شدید F1 مواجه می‌گردد. همانطور که در شکل (۸) ملاحظه می‌شود، روش انتخاب خصیصه DF با هزینه محاسباتی پائین، و روش انتخاب خصیصه FS2 نیز با هزینه محاسباتی کمتر از روش انتخاب خصیصه ترکیبی R+S، کارائی نزدیکی با روش انتخاب خصیصه ترکیبی R+S داشته و می‌توانند در کاربردهائی که هزینه محاسباتی اهمیت زیادی دارد، مورد استفاده قرار گیرند. همچنین روش انتخاب خصیصه DF می‌تواند بعنوان مرحله اول از روش ترکیبی، جایگزینی برای روش انتخاب خصیصه Relief-F باشد تا با هزینه محاسباتی کمتری خصیصه‌های اولیه حذف شده و در مرحله بعدی خصیصه‌های باقیمانده با استفاده از روش روکشی بصورت دقیق‌تری انتخاب شوند. روش پیشنهادی انتخاب خصیصه ترکیبی R+S با داشتن هزینه زمانی بسیار پائین‌تر نسبت به روش انتخاب خصیصه روکشی، دقتی نزدیک به این روش داشته و می‌تواند در مواردی که دقت دسته‌بندی بسیار مهم است و از طرفی هزینه زمانی بسیار بالا را نمی‌توان پرداخت کرد، مورد استفاده قرار گیرد. از طرف دیگر اگر به دلیل هزینه نسبتاً زیاد روش انتخاب خصیصه ترکیبی R+S در مقایسه با روش‌های انتخاب خصیصه فیلتری، نخواهیم از این روش استفاده کنیم، در اینصورت اگر تعداد خصیصه‌های کمتری را با مقدار کاهش کمتری در F1 انتخاب کنیم در نتیجه روش انتخاب خصیصه FS2 راه‌کار مناسبی است، اما اگر اصل برای ما بیشترین مقدار F1 باشد در نتیجه روش انتخاب خصیصه Relief-F راه‌کار مناسب‌تر است. نکته دیگر اینکه کارائی نزدیک روش‌های انتخاب خصیصه DF وRelief-F به این نکته اشاره دارد که کلمات با فرکانس تکرار زیاد برای مسائل دسته‌بندی متون، اطلاعاتی مفید هستند. (البته این مطلب در مورد کلمات خاص صدق نمی‌کند).

نتیجه‌گیری

در این مقاله، ضمن بررسی و ارزیابی روش‌های مختلف نمایش سندها بصورت خصیصه‌های تک کلمه‌ای، ترکیب دوتائی کلمات و یا ترکیبی از آنها و همچنین در نظر گرفتن علائم نوشتاری جملات زبان، مانند: نقطه، کاما و... به منظور تشخیص معنی دار بودن ترکیب‌های دوتائی ایجاد شده، ملاحظه می‌شود که درنظر گرفتن فضای خصیصه‌ها به صورت فقط ترکیب‌های دوتائی کلمات یکپارچه، جهت نمایش سندها، باعث افت در مقدار F1 می گردد و نیز در نظر گرفتن فضای ترکیب دوتائی کلمات منقطع نیز، باعث بوجود آمدن کمی بهبود در F1 می‌شود. از طرفی در نظر گرفتن ترکیب خصیصه‌های تک کلمه ای و دو کلمه ای بطور کلی باعث بهبود دقت دسته‌بندی شده، بخصوص در حالتی که ترکیب‌های با معنی با استفاده از علائم نوشتاری جملات در زبان، ایجاد گردد، در اکثر موارد باعث بهبود در دقت دسته‌بندی می‌شود که خود نشان دهنده وجود یکسری ترکیب‌های دوتائی مهم در متون است که در حالت فضای خصیصه‌های تک کلمه ای مدل نشده‌اند. البته استفاده از فضای خصیصه‌ها به صورت ترکیبی، باعث افزایش ابعاد فضای خصیصه‌ها شده که می‌توان با استفاده از روش‌های متداول انتخاب خصیصه، تعداد زیادی از آنها را ضمن بهبود در دقت دسته‌بندی حذف نمود. بنابراین، روش نمایش سندها به صورت ترکیبی از خصیصه‌های تک‌کلمه‌ای و دوکلمه‌ای با معنی، به منظور نمایش ترکیب‌های دوتائی مهم که در روش خصیصه‌های تک‌کلمه‌ای بیان نمی‌شود، باعث بهبود در کارائی دسته‌بندی می‌شود و می‌تواند بعنوان روشی مناسب جهت نمایش سندها در مسائل مربوط به دسته‌بندی متون استفاده گردد. با توجه به موارد ذکر شده در رابطه با پیدا کردن ترکیب‌های دوتائی با معنی که با استفاده از مفاهیم ابتدائی نوشتاری جملات در زبان ایجاد شده‌است، پیشنهاد می‌شود با توجه به مفاهیم پیچیده‌تر ساختاری جملات در زبان، ترکیب‌های مختلف با معنی برای نمایش سندها در روش‌های دسته‌بندی، لحاظ گردد و تاثیر اینگونه روش‌های نمایش سندها روی کارائی دسته‌بند بررسی شود. همچنین با بررسی دقیق همبستگی‌ها در ضوابط

اکثر روش‌های انتخاب خصیصه، چهار همبستگی بین خصیصه‌های سندها و دسته‌ها را می‌توان ملاحظه نمود که هر یک از روش‌های فیلتری انتخاب خصیصه، ترکیبی از این روابط را جهت وزن‌دهی به خصیصه‌ها و ایجاد یک ضابطه جهت تصمیم‌گیری برای انتخاب خصیصه، استفاده می‌کنند. در این مقاله با در نظر گرفتن کلیه همبستگی‌های مثبت و منفی بین خصیصه‌ها و دسته‌ها به تولید ضابطه‌هائی جدید (یعنی FS1 و FS2) جهت انتخاب خصیصه‌ها پرداخته و یک مرتبه کلیه همبستگی‌ها را با علامت مثبت و مرتبه دیگر، همبستگی‌های مثبت را با علامت مثبت، و همبستگی‌های منفی را با علامت منفی در ضابطه انتخاب خصیصه لحاظ نمودیم. نتایج آزمایشات انجام گرفته روی دو دسته‌بند معروف SVM Light و بیزین ساده نشان دهنده اینست که روش‌های پیشنهادی این مقاله (یعنی FS1 و FS2) با داشتن هزینه محاسباتی بالاتر نسبت به IG، نتایج بهتری را در هر دو دسته بند از خود نشان می‌دهند. بطوریکه در آزمایش دسته‌بند SVM Light حتی نتایج بدست آمده از روش انتخاب خصیصه FS2، بهتر از نتایج بدست آمده از روش انتخاب خصیصه CHI می‌باشد. در آزمایشات انجام شده با استفاده از روش دسته‌بندی بیزین ساده، روش انتخاب خصیصه FS2 دارای ماکزیمم پائین‌تری نسبت به روش انتخاب خصیصه CHI است اما در آستانه‌های حذف بالاتر خصیصه‌ها، روش انتخاب خصیصه FS2 دارای مقدار F1 بیشتری نسبت به CHI می‌باشد. این مطلب برای روش انتخاب خصیصه FS2 در مقایسه با روش انتخاب خصیصه Relief-F نیز صادق است.

آزمایشات متعدد انجام شده در این تحقیق، بیانگر این است که اولاً لازم است ضرایب α_1، α_2، β_1 و β_2 در ضابطه FS غیر صفر باشند (به عبارت دیگر از هر چهار همبستگی مثبت و منفی در ضابطه انتخاب خصیصه استفاده شود) و ثانیاً $\alpha_2>0$ و α_1 (یک مقدار مثبت) و $\beta_2<0$ و β_1 (یک مقدار منفی) انتخاب گردد تا همبستگی‌های منفی، همبستگی‌های مثبت را تعدیل کنند. در ادامه توسعه روش‌های انتخاب خصیصه، می‌توان با تغییر در وزن دهی به همبستگی‌های مثبت و منفی(α_1، α_2، β_1 و β_2) تاثیر هر یک از این همبستگی‌ها را بصورت دقیق‌تری بررسی نمود.

نهایتاً، بررسی و ارزیابی روش‌های مختلف انتخاب خصیصه و روش انتخاب خصیصه FS2 و روش انتخاب خصیصه ترکیبی R+S نشان می دهد ضمن بهبود کارائی، می‌توانند تا ۹۴٪ خصیصه‌ها را حذف کنند. روش‌های انتخاب خصیصه FS2, Relief-F، R+S و CHI موثرترین روش‌ها در حذف خصیصه‌ها به صورت حریصانه، پس از روش روکشی بوده که بدون از دست دادن دقت دسته‌بندی و یا حداکثر با کمی از دست دادن دقت، می‌توانند استفاده شوند. همچنین روش Relief-F (که اولین بار است که در دامنه متون استفاده می‌شود) نتایج بسیار خوبی در حذف خصیصه‌های متون از خود نشان می‌دهد و می‌تواند بعنوان روشی جهت انتخاب خصیصه‌ها در ناحیه متون نیز مورد استفاده قرار گیرد.

روش ساده DF کارائی نزدیکی به روش‌های خوب انتخاب خصیصه داشته و با توجه به هزینه محاسباتی پائین خود، یک روش مناسب برای انتخاب خصیصه‌های اطلاعاتی می باشد. زمانی‌که هزینه محاسبات بسیار زیاد باشد و کم هزینه بودن محاسبات از اهمیت ویژه‌ای برخوردار باشد این روش بعنوان یک روش مقیاس‌پذیر و با هزینه محاسباتی پائین می‌تواند به جای روش‌های دیگر مورد استفاده قرار گیرد.

با درنظر گرفتن روش‌های مختلف انتخاب خصیصه و هزینه‌های محاسباتی هریک از آنها، می‌توان ترکیب‌های مختلفی از روش‌های فیلتری و روکشی را جهت انجام امور انتخاب خصیصه‌ها در دسته‌بندی متون استفاده نمود. با توجه به ترکیب‌های متفاوتی که می‌توان از روش‌های فیلتری و روکشی داشت در آینده ترکیب‌های مختلف این روش‌ها مورد بررسی قرار خواهد گرفت واز نظر هزینه محاسباتی و کارائی، با روش‌های پیشنهادی مقایسه خواهد شد.

مراجع

1 - Sahami, M. (1999). *Using Machine Learning to Improve Information Access*. Ph.D. Thesis, Computer Sccience Department, Stanford University.

2 - Lewis, D. (1994). "An introduction to information retrieval." *In Proceedings of 17th Ann. Int. ACM SIGIR Conference on Research and Development in Information Retrieval.*

3 - Yang, Y. (1999). "An evaluation of statistical approaches to text categorization." *Journal of Information Retrieval,* Vol. 1, No. 1/2, PP.67-88.

4 - Joachims, T. (1998). "Text categorization with support vector machines: learning with many relevant features." *In Proceedings of 10th European Conferece on Machine Learning (ECML-98),* PP.137-142.

5 - Cardie, C. (1993). "Using decision tree to improve case based learning." *In Proceedings of 10th Int. Conference on Machine Learning (ICML-93),* PP. 25-32.

6 - Yang, Y. and Pedersen, J. A. (1997). "A comparative study on feature selection in text categorization." *In Proceedings of 14th International Conference on Machine Learning (ICML-97),* PP.412-420.

7 - Lang, K. (1995). "NEWSWEEDER: learning to filter netnews." *In Proceedings of the 12th International Conference on Machine Learning, (ICML-95),* PP.331-339.

8 - Dash, M. and Liu, H. (1997). "Feature selection for classification." *Intelligent Data Analysis,* Vol. 1, No.3.

9 - Kononenko, I. (1994). "Estimating attributes: analysis and extention of RELLIEF." *In Proceedings of 6th European Confrence on Machine Learning (ECML-94),* PP.171-182.

10 - Lewis, D. (2000). *The Reuters-21578 Collection.*
http://www.davidlewis.com/resources/testcollections/reuters21578.

11 - Kohavi, R. and Sommerfield, D. (1995). "Feature subset selection using wrapper methods: overfitting and dynamic search space topology." *In Proceeding of the 1st international Conference on Knowledge Discovery and Data Mining,* PP.192-197.

12 - Kira, K. and Rendell, L. A. (1992). "The feature selection problem: traditional methods and a new algorithm." *In Proceedings of 10th National Conference on Artificial Intelligence,* PP.129-134.

13 - Langley, P. (1994). "Selection of relevant features in machine learning." *AAAI Fall Symposium on Relevance,* PP.140-144.

14 - Almuallim, H. and Dietterich, T. G. (1994). "Learning boolean concepts in the presence of many irrelevant features." *Artificial Intelligence 69,* No. 1-2, PP.279-306.

۱۵ - جلیلی، س. و بیطرفان، م. "بررسی اثر روابط بین خصیصه‌های سندها و دسته‌ها در ضابطه‌های روش‌های آماری انتخاب خصیصه در بهبود دسته‌بندی متون." یازدهمین کنفرانس برق، دانشگاه شیراز، (۱۳۸۲)..

16 - Aggarwal, C. and Yu, P. (1998). "Data mining techniques for associations." *Clustring and Classification. LNAI 1574*.

۱۷ - جلیلی، س. و بیطرفان، م. "انتخاب خصیصه به روش ترکیبی فیلتری-روکشی در دسته‌بندی متون." هشتمین کنفرانس کامپیوتر، دانشگاه فردوسی مشهد، (۱۳۸۱).

۱۸ - بیطرفان م. "بهبود روش انتخاب خصیصه در دسته‌بندی متون." دانشگاه تربیت مدرس، گروه کامپیوتر، پایان نامه کارشناسی ارشد، (۱۳۸۱).

19 - 20-Newsgroup Dataset. http://www.cs.cmu.edu/afs/cs.cmu.edu/projects/20-newsgroup/data.

واژه های انگلیسی به ترتیب استفاده در متن

1 - Nearest Neighbor	2 - Filtering Methods	3 - Wrapper Methids	
4 - Sequential Forward Selection	5 - Sequential Back Selection	6 - Stop Word	
7 - Stemming	8 - Thereshold	9 - Positive Correlation	
10 - Negative Correlation	11 - Document Frequency	12 - Information Gain	
13 - Mutual Information	14 - Chi Square	15 - Simplified Chi Square	
16 - Correlation Coefficient	17 - Irrelevant	18 - Precision	19 - Recall