ارائه یک روش دسته بندی متون با تکنیک کاهش ویژگی فیلتری و یادگیری ماشین

پیمان جلالی*[۱]، سید محسن هاشمی[۲]،عارف سیاحی[۳] ،عادل حیدری[۴]

۱- آموزشکده فنی و حرفه ای سما،دانشگاه آزاد اسلامی، واحد سوسنگرد، سوسنگرد، ایران

۲- آموزشکده فنی و حرفه ای سما،دانشگاه آزاد اسلامی، واحد سوسنگرد، سوسنگرد، ایران

۳- آموزشکده فنی و حرفه ای سما،دانشگاه آزاد اسلامی، واحد سوسنگرد، سوسنگرد، ایران

۴- آموزشکده فنی و حرفه ای سما،دانشگاه آزاد اسلامی، واحد سوسنگرد، سوسنگرد، ایران

خلاصه

پایگاه داده‌های متنی با توجه به افزایش میزان اطلاعات موجود در فرم الکترونیکی، مانند نشریات الکترونیکی، انواع مختلف مدارک الکترونیکی، پست الکترونیکی و شبکه جهانی وب به سرعت در حال رشد می‌باشند. یکی از مهم‌ترین تکنیک‌های متن‌کاوی، دسته‌بندی متون است. متن کاوی را می‌توان به عنوان متدها و الگوریتم‌هایی از فیلدهای یادگیری ماشین و آماری برای متن‌ها با هدف پیدا کردن الگوهای مفید در نظر گرفت. در این تحقیق الگوریتم انتخاب ویژگی فیلتری بهره اطلاعات مورد استفاده قرار گرفته است. این مقاله روشی برای دسته بندی متون پیشنهاد شده است که در آن ابتدا یک دسته بند ماشین بردار پشتیبان ، بیز ساده و درخت تصمیم با تعداد کمی نمونه برچسب دار ساخته می شود، سپس با استفاده از یادگیری فعال و بکارگیری روش نمونه گیری بر اساس عدم اطمینان به همراه ایده جدید مشابهت و انتخاب گروهی نمونه ها ، به صورت هدفمند نمونه های مفید را برای برچسب گذاری به کاربر می دهد تا در آموزش دسته بند از آنها استفاده کند.

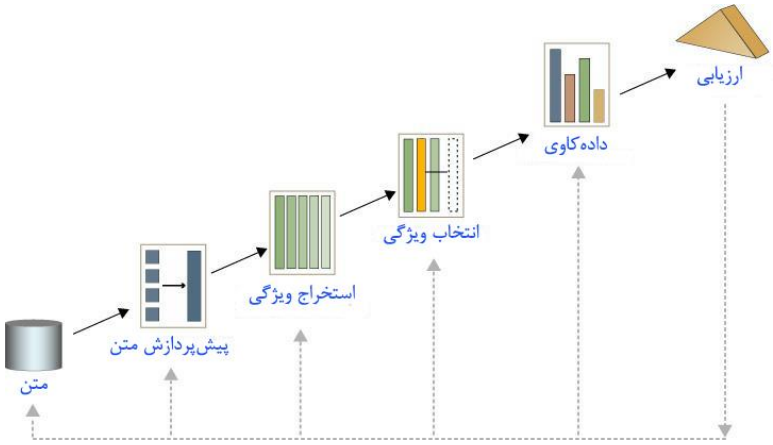
کلمات کلیدی: متن کاوی، داده کاوی، درخت تصمیم، بیزساده، یادگیری ماشین

۱. مقدمه:

امروزه با توجه به رشد روزافزون دسترسی به اسناد الکترونیکی، دسته‌بندی خودکار اهمیت ویژه‌ای یافته است. دسته‌بندی متون به عمل برچسب‌گذاری موضوعی متون زبان طبیعی برمبنای یک مجموعه از پیش تعیین‌شده می‌باشد. در طول سالیان اخیر، طبقه‌بندی‌های مختلف و با نگرش‌های مختلفی برای این کار مطرح‌شده است. نکته قابل‌توجه، امکان خودکارسازی این طبقه‌بندی‌ها به ازای متون جدید است[۹]. با افزایش روزافزون حجم اطلاعات، وجود سیستمی برای دسته‌بندی خودکار متون ضروری به نظر می‌رسد. ازاین‌رو در این پژوهش سعی خواهد شد با استفاده از شیوه کار متن‌کاوی موجود و نیز تکنیک‌های یادگیری ماشین، یک مدل با دقت و کار آیی مناسب برای دسته‌بندی خودکار متون فارسی ارائه شود تا بتوان در مواردی همچون دسته‌بندی متون در بایگانی اسناد، دسته‌بندی صفحات وب، یادگیری

[1] p.jalali05@ yahoo.com

خودکار علایق مطالعاتی و پژوهشی کاربران، فیلتر کردن خودکار پست الکترونیکی بر اساس محتوا و غیره از آن استفاده کرد.شکل (۱) فرایند کلی استخراج دانش از متن را نشان می‌دهد[۱۱].



شکل (۱): فرایند استخراج دانش از متن [۱]

۲. مسئله دسته‌بندی متون

دسته‌بندی متن شامل نسبت دادن اسناد به یکی از چند گروه از پیش تعیین شده است. برای نایل شدن به این هدف اسناد ورودی توسط یک مجموعه از مشخصات که معمولاً خصوصیات نامیده می‌شود توصیف می‌شوند. برخلاف خوشه‌بندی که شامل آموزش بدون نظارت است، در دسته‌بندی یک مجموعه آموزشی از داده‌ها با برچسب‌گذاری قبلی نیاز است . هدف دسته‌بندی تحلیل داده های ورودی و ایجاد یک مدل دقیق برای هر دسته با استفاده از این خصوصیات است. اسناد جدید در داخل یکی از این دسته‌ها دسته‌بندی می‌شوند. در مسأله دسته‌بندی متن، مشخصات کلماتی هستند که درون اسناد متنی قرار دارند. در بسیاری از موارد قبل از یادگیری ماشین انتخاب خصوصیتصورت می‌گیرد تا فضای خصوصیات را کاهش دهد[۹]. در دسته بندی هدف بدست آوردن یک تابع نگاشت بین اسناد و مجموعه کلاس ها یا گروه ها به وسیله یک سری سند که به آنها مجموعه آموزشی[۱۰]می‌گویند، می‌باشد. که از این تابع نگاشت برای تعیین خودکار گروه سند جدید استفاده می‌شود. توجه شود که در هنگام تولید تابع نگاشت مجموعه اسناد برچسب خورده هستند یعنی گروه هر یک از اسناد مشخص شده است و به وسیله این اسناد برای هر گروه یک سری ویژگی وصفات منحصر به فرد استخراج می‌کنیم.

۳. کاربردهای دسته‌بندی متون

دسته‌بندی متون کاربردهای فراوانی می‌تواند داشته باشد که ازجمله آن می‌توان به این موارد اشاره نمود [۲]. دسته‌بندی گفتاری که ترکیبی از دسته‌بندی متون و تشخیص گفتار است، دسته‌بندی متون چندرسانه‌ای از طریق عنوان‌های متنی، تشخیص نویسنده برای متون نامشخص یا موردبحث، تشخیص زبان برای متونی که زبان آن‌ها نامشخص است، تشخیص خودکار جنس متن، بایگانی اسناد، دسته‌بندی صفحات وب، یادگیری خودکار علایق مطالعاتی و پژوهشی کاربران، فیلتر کردن خودکار پست الکترونیکی بر اساس محتوا و غیره.

۴. مروری بر کارهای پیشین

در مرجع[۳] یک الگوریتم انتخاب ویژگی ناهموار مؤثر بادید دانه‌بندی برای مجموعه داده‌ها در مقیاس بزرگ ارائه دادند که در آن‌یک مجموعه داده در مقیاس بزرگ داده‌شده است، این الگوریتم ابتدا دانه‌های کوچک مختلف را انتخاب می‌کند و سپس بر روی هر دانه کوچک، میزان کاهش را از مجموعه داده اصلی برآورد می‌کند.
در مرجع [۴] یک نمونه جدید به‌طور همزمان تکاملی و الگوریتم انتخاب ویژگی ارائه داده‌اند که برای میلیون‌ها مورد از هزاران ویژگی مقیاس‌پذیر است. این پیشنهاد بر اساس اصل تقسیم و غلبه همراه با ساماندهی است، اصل تقسیم و غلبه موجب اجرای الگوریتم در زمان خطی می‌شود. همچنین این روش در محیط‌های موازی به‌راحتی قابل‌اجرا است و می‌تواند بدون بارگذاری کل مجموعه داده در حافظه کار کند.
در مرجع [۵] یک تابع معیار کلی در مورد اطلاعات متقابل در انتخاب ویژگی ارائه‌شده است، که می‌تواند بسیاری از اندازه‌گیری‌های اطلاعاتی در الگوریتم‌های پیشین را به یکدیگر نزدیک کند. در انتخاب‌گرهای سنتی، اطلاعات متقابل در کل فضای نمونه برآورد می‌شوند که نمی‌تواند دقیقاً نشان‌دهنده ارتباط میان ویژگی‌ها باشد. برای مقابله با این مشکل، هدف دوم این مقاله پیشنهاد یک الگوریتم انتخاب ویژگی جدید بر اساس اطلاعات متقابل پویا است، که تنها در موارد بدون برچسب تخمین زده‌شده است.
در مرجع [۶] یک روش نیمه نظارت بیزی برای انتخاب ویژگی طبقه‌بندی‌شده را ارائه داده‌اند. با توجه به اینکه در برنامه‌های کاربردی دنیای واقعی، دریافت کردن نمونه‌های بدون برچسب معمولاً آسان است، اما به دست آوردن برچسب‌های دقیق، مربوط به نمونه‌ها گران است، این امر منجر به زباله‌های بالقوه از اطلاعات طبقه‌بندی‌شده ارزشمند، در نمونه‌های بدون برچسب می‌شود که دور ریخته می‌شوند. درنتیجه این روش نمونه‌های فاقد برچسب را در مشکل انتخاب ویژگی‌های طبقه‌بندی‌شده، حل می‌کند.
در مرجع[۷] یک روش انتخاب ویژگی دومرحله‌ای برای طبقه‌بندی متن را ارائه دادند، که مشکل فضای موردنیاز برای اسناد با ابعاد بالا را برای نشان دادن اسناد، حل می‌کند و کارش به این صورت است که برای بالا رفتن دقت و کارایی طبقه‌بندی از انتخاب ویژگی دومرحله‌ای استفاده می‌کند که در مرحله اول، یک روش انتخاب ویژگی جدید برای کاهش اصطلاحات کم‌اهمیت اعمال می‌کند و در مرحله دوم یک فضای معنایی جدید، بین واژگان، بر اساس روش نمایه‌سازی، معنایی نهان ایجاد می‌کند.

۳. روش تحقیق:

برای ارائه یک مدل برای دسته‌بندی خودکار متون لاتین ابتدا روش های پیش پردازش انجام گردیده است و ار روش انتخاب ویژگی بهره اطلاعات که در زیر توضیح داده خواهد شد، استفاده می شود. با تجزیه‌وتحلیل دقیق داده ها با ارزش، روش جدیدی ارائه می‌شود که بتواند نقاط ضعف روش‌های پیشین را پوشش دهد.

۳.۱. مجموعه داده:

اسناد ورودی یا مجموعه داده انتخابی، مجموعه داده اخبار رویترز-۲۱۵۷۸ می‌باشد[۱۲]. این مجموعه داده در سال ۱۹۸۷ گردآوری و توسط گروهی از کارکنان خبرگزاری رویترز تهیه و شاخص‌گذاری شد.. با توجه به اینکه در این تحقیق، روش تک برچسبی - چند دسته مورد توجه است، از زیرمجموعه R (8) استفاده می‌شود. این زیرمجموعه شامل ۸ دسته اصلی و دارای ۷۶۷۴ سند متنی می‌باشد. هر سند متنی تنها به یکی از دسته‌های موجود تعلق دارد. با توجه به روش جداسازی اسناد آموزشی از اسناد آزمایشی ModApte Split، مجموعه اسناد آموزشی شامل ۵۴۸۵ سند و مجموعه اسناد آزمایشی شامل ۲۱۸۹ سند می‌باشد. جزئیات زیرمجموعه R (8) اخبار رویترز در جدول (۱) نشان داده شده است.

جدول (۱): جزئیات زیرمجموعه اخبار رویترز

دسته	تعداد اسناد آموزشی	تعداد اسناد آزمایشی	تعداد کل اسناد دسته
Acq	1596	696	**2292**
Crude	253	121	**374**
Earn	2840	1083	**3923**
Grain	41	10	**51**
Interest	190	81	**271**
Money-fx	206	87	**293**
Ship	108	36	**144**
Trade	251	75	**326**

۳.۲.پیش پردازش:

برای دسته‌بندی متون پس از فراهم کردن مجموعه داده، نوبت به مرحله پیش‌پردازش می‌رسد. در این مرحله اسناد متنی خام باید به فرمی که قابل استفاده در مرحله انتخاب ویژگی و الگوریتم یادگیری باشند، تبدیل شوند. در مرحله پیش‌پردازش روش پیشنهادی، عملیات Transform Case، Tokenization، Filter StopWords، Stemming و Generate n-gram انجام می‌شود. عملکرد هر کدام در زیر بیان شده است.

Transform Case: در این مرحله تمامی کاراکترهای موجود در متن به فرم یکسان تبدیل می‌شوند. در این مرحله تمامی کاراکترها به حروف کوچک تبدیل می‌شوند.

Tokenization: در این مرحله کل متن به صورت کلمات متوالی جدا از هم تقسیم می‌شود.

Filter StopWords: در این مرحله کلمات زائد در زبان انگلیسی حذف خواهند شد.

Stemming: در این مرحله ریشه کلمات در زبان انگلیسی با استفاده از الگوریتم ریشه‌یابی پورتر برای حذف پسوندها و پیشوندهای کلمات در جهت کاهش طول کلمات و تا زمانی که به فرم ریشه‌شان تبدیل شوند، استفاده شده است.

Generate n-gram: در این مرحله برای شاخص‌بندی و کاهش ابعاد متن از n-gram استفاده شده است. با استفاده از n-gram می‌توان متن را به صورت یک سری از کلمات متوالی به طول n نشان داد. این مدل ابتدا برای مسائل پردازش گفتار مطرح شد. ولی اکنون نگارش‌های مختلفی از این مدل برای مسائل پردازش زبان طبیعی و دسته‌بندی متون مطرح شده است[۱۷و۱۶]. با توجه به آزمایش‌های انجام‌شده بر روی مقادیر مختلف n و برای جلوگیری از افزایش پیچیدگی، از n-gram با میزان n=2 استفاده شد.پس از مراحل فوق، وزن‌دهی ویژگی‌ها انجام می‌شود. در این تحقیق از روش وزن‌دهی TF-IDF استفاده شده است که عملکرد دسته‌بندها با استفاده از این روش وزن‌دهی مورد ارزیابی قرار می‌گیرند.

۳.۳. انتخاب ویژگی با روش بهره اطلاعات (IG):

این روش یکی از پرکاربردترین و محبوب‌ترین روش‌ها می‌باشد که در بسیاری از پژوهش‌ها مورد استفاده قرارگرفته و نتایج مطلوبی داشته است. در سال‌های اخیر، سودمندی اطلاعات به عنوان معیاری برای میزان سودمندی واژه در زمینه یادگیری ماشین مطرح شده است. این معیار تعدادی قلم‌های اطلاعاتی را برای پیشگویی دسته، با کار بر روی وجود و عدم وجودیک حمله در یک دیتاست به دست می‌آورد. به عبارتی در این روش میزان سودمندی یک حمله در دیتاست تعداد حملاتی است که برای پیشگویی دسته، با توجه به وجود یا عدم وجود حمله دردیتا ست به دست می‌آید. رابطه (۱) مربوط به این روش می‌باشد [۱۳].

$$IG(t_k, c_i) = \sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_k, \bar{t}_k\}} p(\bar{t}, c) \log_2 \frac{P(t, c)}{P(\bar{t})P(c)} \qquad (1)$$

با داشتن یک مجموعه آموزشی، می‌توان بهره اطلاعاتی را برای هر واژه محاسبه نمود. حملاتی که بهره اطلاعاتی آن‌ها از یک حد آستانه کمتر است از فضای ویژگی حذف می‌شوند. طبق تعریف محاسبه شامل تخمینی از احتمال‌های شرطی از دسته تعیین‌کننده واژه و محاسبات آنتروپی است. در مرحله یادگیری از الگوریتم های یادگیری ماشین زیر استفاده می نماییم.

۳.۴. الگوریتم Naïve Bayes

یکی دیگر از الگوریتم‌های مطرح‌شده در زمینه دسته‌بندی[۱]، بیز ساده است. این روش به دلایل متعددی اهمیت دارد. اینکه ساخت آن بسیار ساده است و نیازی به برنامه‌های تخمین پارامتر تکرارشونده پیچیده ندارد. بدین معنی که می‌توان از آن برای مجموعه داده‌های بسیار وسیع نیز استفاده نمود. این الگوریتم یکی از سریع‌ترین الگوریتم‌ها در زمینه دسته‌بندی می‌باشد[۸].

۳.۵. درخت تصمیم

درخت‌های تصمیم‌گیری ابزار استانداردی در داده‌کاوی هستند. این الگوریتم‌ها هم در متغیرها و هم سایز مجموعه آموزش سریع و همچنین مقیاس‌پذیر[1] هستند. یکی از مشکلات درخت‌های تصمیم‌گیری برای متن‌کاوی این است که تنها به تعداد کمی از ترم‌ها وابسته است[۱۴].دسته‌کننده درخت تصمیم گیری یکی از پر استفاده ترین روشهای یادگیری با نظارت است که برای کاوش[۱۱] داده‌ها مورد استفاده قرار می‌گیرد. این دسته‌کننده برای تفسیر آسان است و می‌تواند بصورت قوانین IF-THEN-ELSEبازنمایی گردد. در این دسته‌کننده تابعی توسط مناطق ثابت تکه های تقریب زده می‌شود و به هیچ دانش قبلی از توزیع داده‌ها نیاز ندارد. این دسته‌کننده بر روی داده‌های پارازیت دار به خوبی کار می‌کند.

۳.۶. ماشین بردار پشتیبان (SVM)

یکی از روش‌های یادگیری بانظارت است که از آن برای طبقه‌بندی و رگرسیون استفاده می‌کنند.این روش از جمله روش‌های نسبتاً جدیدی است که در سال‌های اخیر کارایی خوبی نسبت به روش‌های قدیمی‌تر برای طبقه‌بندی از جمله شبکه‌های عصبی پرسپترون نشان داده است. مبنای کاری دسته‌بندی کنندهٔ SVM دسته‌بندی خطی داده‌ها است و در تقسیم خطی داده‌ها سعی می‌کنیم خطی را انتخاب کنیم که حاشیه اطمینان بیشتری داشته باشد. حل معادلهٔ پیدا کردن خط بهینه برای داده‌ها به وسیله روش‌های QP که روش‌های شناخته شده‌ای در حل مسائل محدودیت‌دار هستند صورت می‌گیرد[۱۵].

۴. نتایج:

۴.۱. ارزیابی روش بر روی دسته‌بندSVM

در این مرحله، جهت ارزیابی روش پیشنهادی در مرحله یادگیری از دسته‌بند SVM استفاده شده است. جدول ۲ نتایج ارزیابی روش پیشنهادی را با استفاده از دسته‌بند SVM نشان می‌دهد. نتایج به دست آمده نشان‌دهنده کارایی بالای روش پیشنهادی با استفاده از دسته‌بند ماشین بردار پشتیبان در میانگین دقت برابر ٪۹۹.۲۵، میانگین فراخوانی برابر ٪۹۹.۱۷، میانگین F1 برابر ٪۹۹.۲۰، معیار درستی برابر ٪۹۹.۶۳ و خطای دسته‌بندی با میزان ٪۰.۳۷ می‌باشد.

[1] Scalable

جدول (۲): نتایج ارزیابی روش با استفاده از دسته‌بند SVM

دسته	Precision	Recall	F1
Acq	99.58	99.73	**99.65**
Trade	99.88	100	**99.94**
Ship	100	99.60	**99.80**
Interest	97.35	98.07	**97.70**
Grain	100	98.99	**99.49**
Crude	99.78	99.83	**99.80**
Earn	99.92	99.81	**99.86**
Money-fx	97.45	97.31	**99.38**
Average precision: 99.25			
Average recall: 99.17			
Average F1: 99.20			
Accuracy: 99.63			
Classification error: 0.37			

۴.۲.ارزیابی روش بر روی دسته‌بند Naive Bayes

در این مرحله، جهت ارزیابی روش پیشنهادی در مرحله یادگیری از دسته‌بند Naive Bayes استفاده شده است. جدول (۳) نتایج ارزیابی روش پیشنهادی را با استفاده از دسته‌بند Naive Bayes نشان می‌دهد. نتایج به دست آمده روش پیشنهادی با استفاده از دسته‌بند **Naive Bayes** در میانگین دقت برابر ٪۷۶.۶۰، میانگین فراخوانی برابر ٪۸۹.۱۰، میانگین F1 برابر ٪۸۲.۳۷، معیار درستی برابر ٪۹۰.۳۷ و خطای دسته‌بندی با میزان ٪۹.۶۳ می‌باشد که نسبت به ماشین بردار پشتیبان در کلیه معیارها دارای عملکرد ضعیف‌تری می‌باشد.

جدول (۳): نتایج ارزیابی روش با استفاده از دسته‌بند Naive Bayes

دسته	Precision	Recall	F1
Acq	87.23	91.25	**89.194**
Trade	92.15	85.20	**85.54**
Ship	47.17	86.32	**61**
Interest	80.74	90.88	**85.51**
Grain	43.04	99.25	**60**
Crude	80.52	83.84	**82.14**
Earn	99.70	91.32	**95.32**
Money-fx	81.31	84.70	**82.97**
Average precision: 76.60			
Average recall: 89.10			
Average F1: 82.37			
Accuracy: 90.37			
Classification error: 9.63			

۴.۳.ارزیابی روش بر روی دسته‌بند J48

در این مرحله، جهت ارزیابی روش پیشنهادی در مرحله یادگیری از دسته‌بند J48 استفاده شده است. جدول (۴) نتایج ارزیابی روش پیشنهادی را با استفاده از دسته‌بند J48 نشان می‌دهد. نتایج به دست آمده روش پیشنهادی با استفاده از دسته‌بند J48 در میانگین دقت برابر ٪۸۳.۶۵، میانگین فراخوانی برابر ٪۷۹.۹۶، میانگین F1 برابر ٪۸۱.۷۶، معیار درستی برابر ٪۹۲.۵۸ و خطای دسته‌بندی با میزان ٪۷.۴۲ می‌باشد که نسبت به ماشین بردار پشتیبان در کلیه معیارها دارای عملکرد ضعیف‌تری می‌باشد ولی در معیارهای میانگین دقت، درستی و خطای دسته‌بندی دارای عملکرد بهتری نسبت به دسته‌بند **Naive Bayes** می‌باشد.

جدول (۴): نتایج ارزیابی روش با استفاده از دسته‌بند J48

دسته	Precision	Recall	F1
Acq	89.97	92.96	**91.44**
Trade	90.08	90.80	**90.43**
Ship	55.05	55.56	**55.30**
Interest	88.48	90.57	**89.51**
Grain	61.11	46.81	**53.01**
Crude	92.36	90.07	**91.20**
Earn	97.10	96.28	**96.68**
Money-fx	82.51	75.50	**78.84**
Average precision: 83.65			
Average recall: 79.96			
Average F1: 81.76			
Accuracy: 92.58			
Classification error: 7.42			

۵. نتیجه گیری:

با توجه به اهمیت دسته‌بندی متون، هدف در این پژوهش طراحی و پیاده‌سازی یک مدل کارآمد جهت دسته‌بندی خودکار برای متون لاتین با استفاده از تکنیک‌های داده‌کاوی و یادگیری ماشین بوده است که با استفاده از آن بتوان متون را با کارایی بالا دسته‌بندی نمود. در این قسمت، نتایج به دست آمده از مدل پیشنهادی با استفاده از الگوریتم SVM که دارای بالاترین کارایی نسبت به الگوریتم‌های بیز ساده و درخت تصمیم بر روی روش پیشنهادی است، مورد مقایسه قرار گرفت. نتایج به دست آمده نشان‌دهنده برتری روش پیشنهادی با استفاده از الگوریتم یادگیری SVM در میانگین دقت با میزان ۹۹.۲۵٪، میانگین فراخوانی با میزان ۹۹.۱۷٪، میانگین F1 با میزان ۹۹.۲۰٪، درستی با میزان ۹۹.۶۳٪ و خطای دسته‌بندی با میزان ۰.۳۷٪ درمقایسه با دسته‌بندی با استفاده از الگوریتم‌های منفرد می‌باشد.

مراجع:

1) Shahbaz, M., Ahsen, S. M., Shaheen, M., Shaheen, M., Masood, S. A., 2011. An Effective Preprocessing Methodology for Textual Data Classification, Journal of American Science, 7 (6), pp. 944-951.
2) Ruta, D., Gabrys, B., 2000. An Overview of classifier fusion Methods, University of Paisley, Computing & Information Systems, 7 (1), pp. 1-10.
3) Jiye Liang , FengWanga, Chuangyin Dangb, Yuhua Qian , 2012. An efficient rough feature selection algorithm with a multi-granulation view , In: International Journal of Approximate Reasoning 53 pp:912–926 .
4) Nicolas Garcia-Pedrajas, Aida de Haro-Garcia, Javier Perez-Rodriguez , 2013. A scalable approach to simultaneous evolutionary instance and feature selection, Information Sciences 228 pp:150–174 .
5) Huawen Liu, JiguiSun, LeiLiu, HuijieZhang , 2009. Feature selection with dynamic mutual information , Pattern Recognition 42 pp: 1330 – 1339 .
6) Ruichu Cai , ZhenjieZhang , ZhifengHao , 2011. BASSUM: A Bayesian semi-supervised method for classification feature selection , Pattern Recognition 44 pp: 811–820 .
7) Jiana Meng, Hongfei Lin , Yuhai Yu , 2011. A two-stage feature selection method for text categorization , Computers and Mathematics with Applications 62 pp: 2793–2800 .
8) Ganiz, M. C., George, C., & Pottenger, W. M. (2011). Higher order Naive Bayes: A novel non-IID approach to text classification. Knowledge and Data Engineering, IEEE Transactions on, 23(7), 1022-1034.
9) Aghdam, M. H., Aghaee, N. G., Basiri, M. E., 2009. Text feature selection using ant colony optimization, Expert Systems with Applications, 36 (3), pp. 6843-6853.

10) Al-Mubaid, H., Umair, S. A., 2006. A New Text Categorization Technique Using Distributional Clustering and Learning Logic, IEEE Transactions on Knowledge and Data Engineering, 18 (9), pp. 1-10.
11) Arturo, M. R., 2006. Automatic Text Categorization of documents in the High Energy Physics domain, Ph. D. Thesis, University of Granada.
12) Dataset, available in: http://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection, Retrived on 28 July 2013.
13) Jing, L. P., Huang, H. K., & Shi, H. B. (2002). Improved feature selection approach TFIDF in text mining. In Machine Learning and Cybernetics, 2002. Proceedings. 2002 International Conference on (Vol. 2, pp. 944-946). IEEE.
14) Bi, Y., Bell, D., Wang, H., Guo, G., Dubitzky, W., 2004-a. Classification Decision Combination for Text Categorization: An Experimental Study, Database and Expert Systems Applications, 3180, pp. 222-231.
15) Kecman, V. (2001). Learning and soft computing: support vector machines, neural networks, and fuzzy logic models. MIT press.
16) Peng, F., Schuurmans, D., Wang, S., 2003. Language and Task Idependent Text Categorization with Simple Language Models, In Proceedings of the HLT-NAACL.
17) Wei, Z., Miao, D., Hugues, J., Zhao, R., Li, W., 2009. N-grams based feature selection and text representation for Chinese Text Classification, International Journal of Computational Intelligence Systems, 2 (4), pp. 365-374.