به نام یزدان پاک

دانشگاه کردستان

دانشکده فنّی و مهندسی

گروه مهندسی کامپیوتر و فنّاوری اطلاعات

عنوان:

راهکارترکیبی برای انتخاب ویژگی در داده‌های ابعاد بالا

پژوهشگر:

محمدحسین دشتبان

استاد راهنما:

دکتر پرهام مرادی

استاد مشاور:

دکتر هادی زارع

پایان نامه کارشناسی ارشد رشته مهندسی کامپیوتر گرایش هوش مصنوعی

اسفند ماه ۱۳۹۱

تعهد نامه

اینجانب محمدحسین دشتبان دانشجوی کارشناسی ارشد رشته مهندسی کامپیوتر گرایش هوش مصنوعی دانشگاه کردستان، دانشکده فنی و مهندسی گروه مهندسی کامپیوتر و فناوری اطلاعات تعهد می نماییم که محتوای این پایان نامه نتیجه تلاش و تحقیقات خود بوده و از جایی کپی برداری نشده و به پایان رسانیدن آن نتیجه تلاش و مطالعات مستمر اینجانب و راهنمایی استاد بزرگوارم بوده است.

با تقدیم احترام

محمدحسین دشتبان

۱۳۹۱/۱۲/۰۱

تقدیم با احترام

به قهرمان جاویدان ایران

کورش کبیر

دانشگاه کردستان

دانشکده فنّی و مهندسی

گروه مهندسی کامپیوتر و فنّاوری اطلاعات

پایان نامه کارشناسی ارشد رشته مهندسی کامپیوتر گرایش هوش مصنوعی

عنوان:

راهکار ترکیبی برای انتخاب ویژگی در داده‌های ابعادبالا

پژوهشگر:

محمدحسین دشتبان

در تاریخ.....توسط کمیته تخصصی و هیات داوران زیر مورد بررسی قرار گرفت و با نمره ..... و درجه ..... به تصویب رسید.

هیات داوران	نام و نام خانوادگی	مرتبه علمی	امضاء
۱- استاد راهنما	دکتر پرهام مرادی	استادیار	
۲- استاد مشاور	دکتر هادی زارع	استادیار	
۳- استاد داور خارجی	دکترحمیدفرورش	استادیار	
۴- استاد داور داخلی	دکتر فردین اخلاقیان	استادیار	

مهر و امضاء گروه

مهر و امضاء معاون آموزشی و تحصیلات تکمیلی دانشکده

چکیده

با پیشرفت روزافزون تکنولوژی در زمینه داده‌کاوی در حوزه‌های علمی مختلف، مجموعه‌داده‌های با ابعاد بسیار بالا در حال افزایش است که منجر به کاهش کارایی الگوریتمهای دسته بندی می‌شود. لذا نیاز به کاهش حجم این مجموعه داده‌ها امری ضروری است. در مجموعه‌داده‌ها با ابعادبالا، تعداد زیادی ویژگی برای هرنمونه وجود دارد که بسیاری از آنها نامرتبط و زاید می‌باشند. در این پایان‌نامه برروی انتخاب ویژگی بر روی مجموعه‌داده‌های ابعادبالای دو حوزه مختلف علم، بیوانفورماتیک و متن، کارشده است. برای هریک از این حوزه‌ها راهکارهای انتخاب ویژگی متفاوتی توسط محققان ارایه شده است که این راهکارها وابسته به ماهیت ویژگی‌های حوزه مورد نظر می‌باشد. مثلا ویژگی‌های داده‌های میکروآرایه مقدار "بیان ژن‌ها" می‌باشند که عددی حقیقی می‌باشد در حالی که در متن، ویژگی‌ها واژه‌ها بوده که الگوریتمهای ارائه شده در این حوزه بر روی خصوصیت آماری آنها که ماهیتی گسسته دارد تمرکز دارد.

راهکارهای ارائه شده برای انتخاب ویژگی به دو دسته کلی باناظر و بی‌ناظر تقسیم بندی می‌شوند. راهکارهای باناظر از برچسب کلاس‌ها در انتخاب ویژگی کمک می‌گیرند، در حالی که در حالت بی‌ناظر تنها از مقادیر ویژگی‌ها استفاده می‌شود. تحلیل واریانس از راهکارهای بی‌ناظر می‌باشد که از دیرباز مورد توجه محققان بوده است. در قسمت اول این پایان‌نامه، روش‌های انتخاب ویژگی بی‌ناظر و باناظر با تکیه بر استخراج ویژگی، تحلیل واریانس و خوشه‌بندی پیشنهاد شده است. روش ارایه شده بر روی شش مجموعه داده بزرگ بیوانفورماتیک که ویژگی های آن ژن‌ها می‌باشند، اعمال شده است. آزمایشات و بررسی‌های مختلف انجام گرفته نشان می‌دهند که روش بی‌ناظر و باناظر پیشنهادی در مجموعه داده‌های مختلف کارایی قابل قبولی را کسب نموده‌است.

در راهکار پیشنهادی دوم پایان نامه، روش انتخاب ویژگی مبتنی بر فیلتر با تکیه بر عامل‌های احتمالاتی تاثیرگذار در دسته‌بندی متن که در روش‌های انتخاب ویژگی احتمالاتی پرکاربرد به کار رفته، ارایه می‌شود. روش ارایه شده از جنبه‌های مختلف مورد تحلیل قرارگرفته و کارایی ویژگی‌های انتخابی آن در دسته‌بندی متن با روش‌های دیگر انتخاب ویژگی مبتنی بر فیلتر مقایسه شده است. آزمایشات متعدد، روش‌های فیلتر را از جنبه‌های مختلف همانند: میزان اشتراک ویژگی‌های برتر انتخاب شده، بررسی واریانس ویژگی‌ها، کارایی ویژگی‌های انتخاب شده بر اساس معیارهای مختلف، رفتار کارایی آنها با افزایش تعداد ویژگی‌ها و میزان دقت و بازیابی روش‌ها نسبت به یکدیگر، به طور عملی مورد مطالعه قرار می‌دهند. سه مجموعه داده استاندارد: Reuter-R8، 20Newsgroup و WebKB دراین مطالعه استفاده شده است. آزمایشات مختلف نشان دهنده این است که روش پیشنهادی در هر سه مجموعه داده توانایی رقابت با روش‌های موفق انتخاب ویژگی مبتنی بر فیلتر را داراست به طوریکه در برخی موارد اختلاف قابل توجهی را ایجاد نموده است.

کلمات کلیدی: انتخاب ویژگی، استخراج ویژگی، تحلیل واریانس، طبقه بندی متن، روش های مبتنی بر فیلتر، انتخاب ژن، رتبه‌دهی ژن‌ها، تحلیل میکروآرایه.

فهرست مطالب

عنوان **صفحه**

فهرست شکل‌ها

فهرست جدول‌ها

مقدمه

۱-۱- سابقه و انگیزه تحقیق

امروزه، وجود دستگاه‌ها ، وسایل و حسگر های قوی در کاربردهای مختلف برای نمونه‌برداری و استخراج مشخصات و ویژگی‌های نمونه‌های داده‌ای باعث به وجود آمدن مجموعه‌هایی با تعداد ویژگی‌های بسیار زیاد[1] شده است. از طرفی سیستم‌های کاربردی مختلف در دنیای واقعی نیازمند سرعت و دقت بالایی می‌باشند که‌با این حجم عظیم داده‌ها کارایی خود را ازدست می‌دهند. ازاینرو یادگیری ساختار داده و انتخاب ویژگی‌ها و مشخصاتی که نمایانگر این ساختار باشند بسیار حایز اهمیت بوده و خصوصا در دهه اخیر به یکی از حوزه‌های فعال تحقیقاتی در نقاط مختلف دنیا تبدیل شده است. مثلا با رشد روزافزون تکنولوژی اینترنت شاهد به وجود آمدن مجموعه داده‌های بسیار بزرگ از سندهای الکترونیکی هستیم. طبقه‌بندی این سندهای الکترونیکی که شامل هزاراه واژه به زبان های مختلف می‌باشند از چالش‌های مهم سالهای اخیر می‌باشد که در این راستا روش‌های زیادی برای انتخاب واژه‌های مناسب به منظور کاهش فضای ویژگی و طبقه‌بندی مناسب‌تر انجام گرفته است [1]. انتخاب ویژگی در حوزه‌های پزشکی که شامل انتخاب ژن[2] های موثر در تشخیص بیماری‌ها می‌باشند نمونه ای دیگر از کاربردهای جدید انتخاب ویژگی در داده‌های ابعاد بالا در دهه اخیر می‌باشند [3] ,[2]. در سرویس های وب، انتخاب ویژگی برای شناسایی نیاز کاربران، تشخیص هرزنامه‌ها، بهبود جستجو و غیره مورد استفاده قرار می‌گیرد (مانند، [5] ,[4]. لذا حجم بالای تحقیقات انجام شده در سالهای اخیر بر روی انتخاب ویژگی در حوزه‌های مختلف نشان‌دهنده اهمیت هرچه بیشتر این موضوع می‌باشد.

1-High Dimensional Data
[2] Gene

برای انتخاب ویژگی در حوزه‌های مختلف خصوصا در زمینه اسناد الکترونیکی و داده‌های بیوانفورماتیک که شامل مجموعه‌هایی با ابعاد بسیار بالا می‌باشند راهکار های مختلفی ارایه شده است. که این راهکارها بر اساس ماهیت داده‌های مورد بررسی تعریف می‌شوند. در مجموعه داده‌های متنی، ویژگی‌ها معمولا خصوصیات آماری می‌باشند که از واژه‌های موجود در اسناد بدست می‌آیند. بنابراین ویژگی‌ها در این حوزه ماهیتی گسسته دارند. روش‌های بهره اطلاعات[1] [6]، شاخص بهبودیافته جینی[2] [7] و اطلاعات متقابل[3] [8] از روش‌های پرکاربرد انتخاب ویژگی در این حوزه می‌باشند. بر خلاف مجموعه داده‌های متنی، در مجموعه داده‌های میکروآرایه (بیوانفورماتیک) ویژگی‌ها که مقدار بیان ژنها می‌باشند مقداری حقیقی داشته به طوری که فضای پیوسته ای را به وجود می‌آورند. الگوریتم‌های انتخاب ویژگی مختلفی در این حوزه ارایه شده است که برخی از آنها مانند آزمایش خی‌دو[4] [6] و الگوریتم رتبه‌دهی لاپلاسین[5] [۹] می‌باشند. در این پایان‌نامه الگوریتم‌هایی برای انتخاب ویژگی هردوحوزه ارایه می‌شود.

انتخاب ویژگی در حوزه‌های مختلف می‌تواند با استخراج ویژگی نیز همراه باشد. استخراج ویژگی، تولید یک فضای ویژگی جدید بر اساس ویژگی‌های اصلی می‌باشد [۱۰]. یکی از روش‌های استخراج ویژگی روش تحلیل مولفه‌های سازنده[6] [12] ,[11] میاشد. به طور کلی، انتخاب و استخراج ویژگی و یا کاهش ابعاد داده‌ای (کاهش تعداد ویژگی‌ها) می‌تواند به عنوان پیش پردازش برای طبقه‌بندی، خوشه‌بندی و بازیابی اطلاعات، به منظور سرعت بخشیدن و افزایش دقت این عملیات مورد استفاده قرار گیرد. روش‌های انتخاب ویژگی از یک دیدگاه‌کلی قابل دسته‌بندی به روش‌های باناظر[7] و بی‌ناظر[8] می‌باشند. در روش‌های باناظر معمولا اهمیت ویژگی با توجه‌به‌همبستگی[9] که ویژگی به برچسب کلاس[10] خود دارد، مورد ارزیابی قرار می‌گیرد. برخی از روش‌های کلاسیک انتخاب ویژگی، روش Lasso توسط [۴]، روش Relief توسط [۱۳] می‌باشد. در روش Lasso با استفاده از یک تابع

[1] Information Gain
[2] GINI Index
[3] Mutual Information
[4] Chi-Square Test
[5] Laplacion Score
[6] Principal Component Analysis
[7] Supervised
[8] Unsupervised
[9] correlation
[10] Class label

بهینه سازی سعی در انتخاب ویژگی‌ها می‌شود. در این روش از ضرایب رگرسیون برای تولید یک راه‌حل تنک استفاده می‌نماید. در نهایت ویژگی‌هایی که ضریب رگرسیون صفر دارند حذف می‌شوند.

دسته دیگر از الگوریتم‌های انتخاب ویژگی، روش‌های بی‌ناظر، شامل روش‌هایی می‌شوند که در آنها "هدف" انتخاب ویژگی هایی است که زیرساخت هندسی[1] موجود در ساختار داده را به‌بهترین وجه حفظ نمایند. یکی از ساده ترین این روش‌ها، روش واریانس داده‌ای[2] می‌باشد. در این روش، داده‌ها بر روی بعدی که بیشترین واریانس را دارد نگاشت[3] شده و انتخاب ویژگی از ویژگی‌های اصلی انجام می‌گیرد. روش تحلیل مولفه‌های اصلی توسط [11] نیز از حداکثر واریانس استفاده می‌نماید اما در ادامه ویژگی‌ها را تبدیل وانتخاب ویژگی را از ویژگی‌های تبدیل‌یافته[4] انجام می‌دهد.

اخیرا الگوریتم‌های انتخاب ویژگی LapAOFS[5] با بهره گیری از مفاهیم طراحی تجربی توسط [14] ارایه شده است. این الگوریتم با حداقل سازی ماتریس کوواریانس داده‌ای سعی در بدست آوردن ویژگی‌هایی نموده اند که‌هم ساختار هندسی و هم ساختار فضایی[6] داده‌ها را حفظ نمایند. در این الگوریتم که برای انتخاب ویژگی در داده‌های ابعاد بالا می‌باشد امکان استفاده از برچسب کلاس‌ها در صورت موجود بودن وجود ندارد.

۱-۲- مسئله‌ی تحقیق و اهداف پایان‌نامه

به طور کلی در این پایان‌نامه دو روش انتخاب ویژگی مجزا برای داده‌های ابعاد بالای میکروآرایه و متن های الکترونیکی ارایه می‌شود.

روش انتخاب ویژگی LapAOFS که از آخرین روش‌های بی‌ناظر در انتخاب ویژگی می‌باشد. در این روش همانطور که در مقدمه بدان اشاره شد امکان استفاده از برچسب کلاس‌ها در صورت موجود بودن، وجود ندارد. استفاده از برچسب کلاس‌ها و ساختار داده‌ای به طور همزمان می‌تواند به مراتب الگوریتم‌های بی‌ناظر را در انتخاب ویژگی کاراتر سازد. چرا که استفاده از اطلاعات خبره علاوه بر خصوصیات نهفته در ساختار داده می‌تواند باعث هدایت انتخاب ویژگی گردد [15]. لذا در این پایان‌نامه با تکیه بر یک الگوریتم استخراج ویژگی سعی می‌شود که فضای داده‌ای جدیدی را برای

[1] underlying geometrical structure
[2] data variance
[3] project
[4] transformed
[5] Laplacian A-Optimality Feature Selection
[6] manifold

الگوریتم رتبه‌دهی بی‌ناظر LapAOFS فراهم بیـاوریم. در روش پیشـنهادی بـا کـاهش تـاثیر نقـاط داده‌ای و با در نظر گرفتن مرزهای ویژگی در هر کلاس سـعی در تـاثیر دادن برچسـب کـلاس‌هـا در انتخاب ویژگی این الگوریتم شده است.

روش دوم در این پایان‌نامه به بررسی انتخاب ویژگی و روش‌های احتمالاتی انتخاب ویژگی مبتنـی بر فیلتر در مجموعه داده‌های متنی می‌پردازد. در داده‌های ابعاد بالای مـتن مسـاله انتخـاب واژه‌هـای موثر در طبقه‌بندی اسناد با روشهای احتمالاتی مناسب از چالشهای مهم در ارایـه روش‌هـای انتخـاب ویژگی مبتنی بر فیلتر می‌باشد. روش‌های احتمالاتی مختلفـی در ایـن حـوزه مطـرح شـده‌انـد کـه از عامل‌های احتمالاتی متفاوتی استفاده می‌کنند. برای اولین بار در این پایان‌نامه عامل‌های احتمـالاتی[1] روش‌های انتخاب ویژگی احتمالاتی مبتنی بر فیلتر مورد بررسی قرار می‌گیرند به طوریکه در نهایت با بکارگیری مهمترین عامل‌های تاثیرگذار در انتخاب ویژگی‌های مناسب، در میان روش‌های احتمـالاتی مختلف، به ارایه روشی احتمالاتی پرداخته می‌شود.

۱-۳- دورنمای پایان‌نامه

این پایان‌نامه با احتساب این فصل در ۵ فصل کلی گردآوری شده است. در ذیل به طور خلاصه بـه تشریح هر فصل می‌پردازیم:

فصل اول به اهمیت موضوع، چالش ها و اهداف پایان‌نامه در کنار معرفی برخی از آخرین روش‌های ارایه شده در حوزه مورد بررسی می‌پردازد،

فصل دوم به مروری بر کارهای انجام شده در زمینه انتخاب ویژگی می‌پردازد. در این فصل تمـامی روش‌های انتخاب ویژگی که در فصل‌های بعد از آنها استفاده شده، تشریح می‌شود،

فصل سوم به ارایه روش انتخاب ویژگی بـی‌نـاظر بـرای ویژگـی‌هـایی کـه مقـادیر حقیقـی دارنـد می‌پردازد. در این فصل با تکیه بر یـک الگـوریتم اسـتخراج ویژگـی پیشـنهادی سـعی در تـاثیر دادن برچسب کلاس در الگوریتم انتخاب ویژگی LapAOFS می‌کند. سـپس روشـی بانـاظر بـه منظـور بررسی بیشتر استخرج ویژگی ارایه شده، مورد آزمایش قرار می‌گیرد. از مجموعه داده‌های ابعاد بـالای بیوانفورماتیک برای ارزیابی روش پیشنهادی استفاده می‌گردد،

[1] Probabilistic elements

فصل چهارم به ارایه روش انتخاب ویژگی احتمالاتی مبتنی بر فیلتر برای مجموعه داده‌های متنی می‌پردازد. عاملهای احتمالاتی روش‌های فیلتر مختلف مورد بررسی قرارگرفته و بر اساس تاثیرگذارترین عامل‌ها روش فیلتر احتمالاتی جدیدی ارایه می‌شود. سپس آزمایشات، مقایسه‌ها و تحلیل‌های بسیاری بر روی سه مجموعه داده متنی مشهور انجام می‌گیرد.

فصل پنجم در نهایت به جمع‌بندی و کارهای تحقیقاتی بیشتر در آینده میپردازد.

پیشینه

۲-۱- مقدمه

انتخاب ویژگی فرآیندی است که در آن زیرمجموعه‌ای از ویژگی‌های موجود بر اساس معیارهای اهمیت ویژگی انتخاب می‌شوند [۱۶]. بعد از اولین مقالات در زمینه انتخاب ویژگی‌های مرتبط توسط [۱۷], [۱۸] تا به امروز پیشرفت های زیادی در این امر حاصل شده است. این پیشرفت‌ها از یکسو حاصل تلاش محققان در زمینه‌های داده‌کاوی و از سوی دیگر حاصل پیشرفت تکنولوژی داده‌کاوی در زمینه‌های علمی مختلف می‌باشد. انتخاب ویژگی‌ها در زمینه‌های مختلف غالبا به منظور پیش‌بینی ، طبقه‌بندی، خوشه‌بندی و کاهش حجم داده‌ها انجام می‌گیرد.

از سوی دیگرانتخاب ویژگی با توجه به ماهیت ویژگی مانند ترتیبی، گسسته یا پیوسته بودن، با روش‌های مختلفی انجام می‌گیرد. مثلا در داده‌های بیوانفورماتیک ویژگی‌ها مقداری حقیقی داشته و از این‌رو حوزه انتخاب ویژگی به صورت پیوسته مورد بحث قرار می‌گیرد. در عین حال انتخاب ویژگی در زمینه متن‌های اینترنتی به صورت انتخاب ویژگی بر اساس اطلاعات آماری از ویژگی‌ها مانند تعداد واژه در کلاس بدست می‌آیند که مشخصا فضای انتخابی گسسته را به وجود می‌آورند.

۲-۲- روش‌های انتخاب ویژگی

طبقه‌بندی داده‌های ابعاد بالا در مجموعه داده‌های مختلف خود شامل مسائل و مشکلات زیادی است که باعث ارائه‌ی روش‌های مختلف انتخاب ویژگی در حوزه‌های متعدد شده است. انتخاب ویژگی به طور کلی قابل دسته‌بندی به دو روش باناظر و بی‌ناظر می‌باشد. در روش‌های باناظر اهمیت ویژگی‌ها با توجه به‌همبستگی آنها با برچسب کلاس‌ها مورد ارزیابی قرارمی‌گیرد. از روش‌های متداول انتخاب ویژگی باناظر می‌توان به روش‌های ReliefF توسط [۱۳]، Lasso توسط [۱۹]، SVM-RFE توسط [۲۰] و Fisher score در کتاب طبقه‌بندی الگو توسط [۲۱] اشاره داشت.

در عین‌حال استفاده از روش‌های باناظر برای بسیاری از مجموعه داده‌ها که بدون برچسب هستند ممکن نیست. ازاین‌رو محققان به ارائه‌ی روش‌های یادگیری بی‌ناظر برای طبقه‌بندی داده‌ها و انتخاب ویژگی نموده‌اند. توجه به‌این نکته حائز اهمیت است که روش انتخاب ویژگی بی‌ناظر روشی است که به برچسب کلاس توجهی ندارد، لذا این روش‌ها را می‌توان برای هر نوع مجموعه داده‌ای بکار برد. برخی از این روش‌ها عبارتند از:

روش انتخاب ویژگی بی‌ناظر ترتیبی [۱۶]، روش Laplacian score توسط [۲۲]، روش انتخاب ترکیبی بر اساس حداکثرسازی احتمال [۲۳]، روش انتخاب حداکثرآنتروپی [۲۴] و روش تحلیل مؤلفه‌های سازنده (PCA) [۲۵]. در روش تحلیل مؤلفه‌های اصلی ابتدا عمل تبدیل ویژگی‌ها به فضای دیگر انجام گرفته و در نهایت مجموعه‌ای از ویژگی‌های تبدیل‌یافته به جای ویژگی‌های اصلی معرفی می‌شوند. اخیرا روش LapAOFS توسط [۱۴] براساس رتبه‌دهی لاپلاسین [۲۶] و نسخه ارتقا یافته آن [۲۲] ارائه شده است. این روش که راه‌کاری بی‌ناظر برای رتبه‌دهی ویژگی‌ها می‌باشد، بر اساس ساختار داده‌ها در گراف سعی در انتخاب موثرترین ویژگی‌ها می‌نماید.

۲-۳- انتخاب ویژگی در متن

یکی از زمینه‌های بسیار پرکاربرد انتخاب ویژگی در حوزه طبقه‌بندی متن‌ها[1] می‌باشد. طبقه‌بندی متن فرآیند طبقه‌بندی یک سند بر اساس محتویات درونی آن می‌باشد [۲۷]. در واقع با پیشرفت روزافزون تکنولوژی اینترنت حجم سندهای الکترونیکی به طور چشم‌گیری در حال افزایش می‌باشد. لذا، طبقه‌بندی این سندها بر اساس ویژگی‌های منتخب به جای استفاده از حجم انبوهی از ویژگی‌ها از مهم‌ترین مسائل داده‌کاوی اخیر می‌باشد. در دهه‌ی اخیر، روش‌های زیادی برای انتخاب و طبقه‌بندی این داده‌ها توسط محققان ارائه شده است. روش‌های انتخاب ویژگی قابل دسته‌بندی به روش‌های فیلتر[2]، روپوش[3] و ترکیبی[4] و ادغام شده[5] می‌باشند [۲۸]. در بخش های زیر تشریحی مختصر از هریک ازاین روش‌ها آورده شده است.

[1] Text Classification
[2] Filter
[3] Wrapper
[4] Hybrid
[5] Embeded

۲-۳-۱- روش روپوش

در روش روپوش، یک ماشین یادگیر دلخواه برای رتبه‌دهی به زیرمجموعه‌ای از ویژگی‌ها، با توجه به توان آن ویژگی‌ها در امر پیش‌بینی به کار می‌رود. در این روش معمولا روش‌های اکتشافی[1] مانند جستجوی تابو[2]، الگوریتم ژنتیک[3]، کلونی مورچه‌ها[4] و غیره، همراه با روش‌های فیلتر مورد استفاده قرار می‌گیرند. این روش‌ها هنگامی که فضای ویژگی بسیار بزرگ باشد با جستجوی مبتنی بر طبیعت قادر به یافتن راه حل‌های بهینه و یا نزدیک به بهینه می‌باشند.

اخیرا و با افزایش حجم داده‌ها در مجموعه داده‌های بزرگ از روش‌های روپوش استفاده زیادی شده است که حجم عظیم مقالات در این زمینه گواه‌این مدعاست. در [۲۹] راهکاری کلی بر مبنای روش روپوش برای انتخاب ویژگی‌هایی که بیشترین وابستگی به کلاس داده را دارند، معرفی شده است. در [۲۳], [۳۰], [۳۱] روش‌هایی بر اساس الگوریتم ژنتیک برای دسته‌بندی متن ارائه شده است. نسخه تغییر یافته‌ای بر مبنای الگوریتم کلونی مورچه‌ها در [۳۲] استفاده شده است که سعی در حداکثر استفاده از فضای جستجوی ویژگی‌ها برای یافتن ویژگی‌های موثر نموده است.

۲-۳-۲- انتخاب ویژگی مبتنی بر فیلتر

انتخاب ویژگی مبتنی بر فیلتر نیازمند تحلیل آماری مجموعه ویژگی‌ها بدون استفاده از هر نوع مدل و یا ماشین یادگیر می‌باشد [۳۳]. روش‌های فیلتر نسبت به بقیه‌ی روش‌ها روش‌های سریع‌تری هستند اما از طرف دیگر وابستگی بین ویژگی‌ها را مورد بررسی قرار نمی‌دهند. برخی از موفق ترین روش‌های مبتنی بر فیلتر در حوزه متن عبارتید از:

توان واژه[5] [۳۴]، روش احتمالاتی DFS [۳۵]، اطلاعات متقابل[31] ، آزمایش خی‌دو [۸]، بهره‌ی اطلاعاتی [۶]، اندیس بهبود یافته‌ی جینی [۷]، معیار انحراف از توزیع پواسون [۳۶]، حداقل اختلاف کلاس[6] [۳۷]، میزان ابهام[7] [۳۸]، میزان تمایز دهندگی بین کلاس‌ها[8] [۳۹] و آزمون فرض

[1] Heuristics
[2] Tabu Search
[3] Genetic Algorithm
[4] Ant Colony
[5] Term Strength
[6] minimum class difference
[7] ambiguity measure
[8] class discriminating measure