دانشکده ریاضی
گروه مهندسی علوم کامپیوتر

پایان‌نامه برای دریافت درجه کارشناسی‌ارشد
گرایش محاسبات نرم و هوش مصنوعی

عنوان

استفاده از یک رویکرد تکاملی جهت انتخاب ویژگی در پردازش متن

پژوهشگر
علی جلالی

استاد راهنما
دکتر حبیب ایزدخواه

پاییز ۱۴۰۱


بسم الله الرحمن الرحيم

تقدیم به:

ماحصل آموخته هایم را تقدیم می کنم به آنان که مهر آسمانی شان آرام بخش آلام زمینی ام است

به استوارترین تکیه گاهم،دستان پرمهر پدرم

به سبزترین نگاه زندگیم،چشمان سبز مادرم

که هرچه آموختم در مکتب عشق شما آموختم و هرچه بکوشم قطره ای از دریای بی کران مهربانیتان را سپاس نتوانم بگویم.

امروز هستی ام به امید شماست و فردا کلید باغ بهشتم رضای شما

ره آوردی گران سنگ تر از این ارزان نداشتم تا به خاک پایتان نثار کنم، باشد که حاصل تلاشم نسیم گونه غبار خستگیتان را بزداید.

سپاسگزاری:

نخستین سپاس و ستایش از آن خداوندی است که بنده کوچکش را در دریای بیکران اندیشه، قطره‌ای ساخت تا وسعت آن را از دریچه اندیشه‌های ناب آموزگارانی بزرگ به تماشا نشیند. لذا اکنون که در سایه‌سار بنده نوازی‌هایش پایان‌نامه حاضر به انجام رسیده است، بر خود الزم می‌دانم تا مراتب سپاس را از بزرگوارانی به جا آورم که اگر دست یاریگرشان نبود، هرگز این پایان‌نامه به انجام نمی‌رسیدابتدا از استاد گرانقدرم جناب آقای دکتر ایزدخواه که زحمت راهنمایی این پایان‌نامه را بر عهده داشتند، کمال سپاس را دارم. سپاس آخر را به مهربانترین همراهان زندگیم، به پدر و مادرعزیزم تقدیم می‌کنم که حضورشان در فضای زندگیم مصداق بی ریای سخاوت بوده است.

نام و نام خانوادگی: علی جلالی

عنوان پایان نامه: استفاده از یک رویکرد تکاملی جهت انتخاب ویژگی در پردازش متن

مقطع / رشته / گرایش / دانشکده: کارشناسی ارشد/ علوم کامپیوتر

تاریخ فارغ التحصیلی:

استاد راهنمای اول: **استاد راهنمای دوم:**

استاد مشاور اول: **استاد مشاور دوم:**

کلیدواژه‌ها : مستندات، انتخاب ویژگی، بیوه سیاه، کاهش بعد، دقت.

چکیده:

امروزه پیشرفت امکانات نرم‌افزاری و سخت‌افزاری، موجب آسانی ذخیره شدن مقادیر زیادی داده شده است. تعداد مستندات متنی روز به روز در حال افزایش است، نامه‌های الکترونیکی، صفحات وب، متون خبری و مقالات تنها بخشی از این گستره رو به افزایش هستند. بنابراین نیاز به تکنیک‌های متن‌کاوی همانند روش‌های خودکار برای رده‌بندی متون احساس می‌شود. در امر رده-بندی خودکار متون، انتخاب ویژگی از درون متن جزء مهمترین مراحل می‌باشد. انتخاب ویژگی برای کاهش ابعاد فضای ویژگی استفاده می‌شود، چراکه فضای ویژگی برای متون شامل ده‌ها هزار کلمه خواهد بود که پردازش‌های بعدی سیستم را امکان ناپذیر می‌کند. تاکنون روش‌های مختلفی برای انتخاب ویژگی برای داده‌های متنی طراحی شده‌اند که هر یک دارای معایب و مزایایی هستند. انتخاب ویژگی در داده‌های با ابعاد بالا همانند پردازش متون و طبقه‌بندی متون کاربرد فراوانی دارد. ازآنجاییکه انتخاب ویژگی دارای پیچیدگی زمانی نمایی است، استفاده از روش‌های کلاسیک موجب افزایش زمان اجرا می‌شود. در بیشتر مواقع این روش‌ها قادر به یافتن راه‌حل بهینه نمی‌باشند. یکی از اهداف انتخاب ویژگی، یافتن راه‌حل‌های بهینه و یا نیمه بهینه است. در این پایان‌نامه روشی جدید بر مبنای الگوریتم بهینه‌سازی بیوه سیاه برای حل مسئله انتخاب ویژگی ارائه شده است. نتایج مقایسات با برخی از روش‌های دیگر انتخاب ویژگی، نشان می‌دهد که الگوریتم پیشنهادی در تکرار کمتر و سریع‌تر جواب به قابل قبول را پیدا می‌کند و درنتیجه کارایی بالاتری نسبت به دیگر روش‌های دارد؛ در واقع این روش با حذف ویژگی‌های نامرتبط منجر به افزایش دقت و سرعت یادگیری طبقه‌بندی کننده می‌شود.

فهرست مطالب

عنوان **صفحه**

فهرست شکل‌ها

عنوان **صفحه**

فهرست جداول

عنوان **صفحه**

فصل اول

کلیات تحقیق

۱- ۱) مقدمه

به طور تقریبی بیش از ۹۰ درصد از دانش امروزی به صورت متن، مستندات و سایر صورت‌های رسانه‌ای نظیر صوت، تصویر و ویدیو نگهداری می‌شود. اگر از منظر علوم کامپیوتری به این مستندات نگاه شود اکثر آنها به نحوی غیر ساخت یافته ذخیره شده‌اند. با این حال با رشد سریع اینترنت، طبیعی است که از متون نه به صورت کاغذی بلکه به صورت اطلاعات الکترونیکی و برخط استفاده شود. امروزه می‌توان کتابها و اخبار را به صورت الکترونیکی جستجو کرد. تقریباً همه شرکتها، ادارات و سازمانها دارای صفحات وب هستند و اطلاعات خود را در این صفحات ارائه می‌کنند. در نتیجه از طریق اینترنت بسیاری از اطلاعات در دسترس عموم قرار می‌گیرند. با افزایش بیش از حد متون و در اختیار قرار گرفتن اطلاعات زیاد، استفاده از مفهوم هوش تجاری در متن، تبدیل به جزئی ضروری شده است.

۱-۲) بیان مساله

امروزه با توجه به رشد شبکه‌های اجتماعی، فضای مجازی، گسترش تولید اطلاعات و اخبار در قالب متن و ناتوانی هوش انسان در پردازش سریع و تشخیص متن باعث شده تا تلاش برای واگذاری این مسئولیت به ماشین‌ها افزایش یابد. با در نظر گرفتن منابع برچسب‌گذاری شده در مراکز پخش اطلاعاتی و خبری، پژوهش در این موضوع در رده مسائل دسته‌بندی قرار می‌گیرد. یکی از اصلی‌ترین چالش‌ها در برخورد با این مسئله، وجود ویژگی‌های فراوان در فاز یادگیری است، چرا که در این نوع داده‌ها، کلمات بیانگر ویژگی‌ها هستند. تحقیقات نشان داده است که بررسی همه ویژگی‌های استخراج شده از متن نه تنها موجب بهبود نتایج نمی‌شود، بلکه میزان

خطای ماشین را تا حد زیادی افزایش می‌دهد. تاکنون پژوهش‌های بسیاری در این زمینه انجام شده است، که هر کدام با دیدگاه متفاوتی به این چالش پرداخته و در نهایت با استفاده از مدل یادگیری ماشین سعی در یادگیری و تست مدل و دستیابی به نتایج بهینه را داشته‌اند، از جمله این روشها می‌توان به آنالیز مولفه اصلی و روش‌های مبتنی بر آمار اشاره کرد. در این پایان‌نامه سعی خواهد شد با استفاده از الگوریتم تکاملی به انتخاب ویژگی‌های متن پرداخته و دقت مدل به دست آمده را افزایش داد.

۱- ۳) فرضیات

۱- روش پیشنهادی به دلیل استفاده از الگوریتم فراابتکاری در انتخاب ویژگی دقت تشخیص را افزایش می‌دهد.

۲- روش پیشنهادی سرعت تشخیص را افزایش می‌دهد.

۳- روش پیشنهادی با استفاده از ماشین بردار پشتیبان در دسته بندی ویژگی‌ها دقت دسته‌بندی را افزایش می‌دهد.

۱- ٤) اهداف

۱- افزایش دقت تشخیص متن.

۲- افزایش سرعت تشخیص

۳- افزایش دقت دسته‌بندی.

۱- ٥) ضرورت انجام تحقیق

ورود جامعه به دنیای الکترونیک و دیجیتال باعث گردیده است تمامی جنبه‌های زندگی بشری تحت تأثیر این گونه فناوری‌ها قرار بگیرد. از جمله این موارد می‌توان به الکترونیکی شدن متون و نیاز به مدیریت بهینه آنها اشاره نمود. امروزه با توجه به حجم و رشد روزافزون متون فارسی، دسته‌بندی خودکار اسناد و متون از ارزش بزرگ

عملی برخوردار و به طور فزاینده، زمینه‌ی مهمی برای تحقیق است. با استفاده از تکنیک‌های کاوش عنوان متون به راحتی می‌توان متن مدنظر را جستجو نموده و به راحتی و در کمترین زمان متون، اسناد و ... مرتبط را بدست آورد. در گذشته این کار به صورت دستی صورت می‌گرفته است، ولی امروزه با توجه به حجم گسترده داده‌های متنی، نبود زمان کافی و نیاز بشر به پردازش سریع این داده‌های متنی از روش‌های خودکار استفاده می‌شود. روشهای زیادی برای دسته‌بندی متون فارسی وجود دارد که هر کدام دارای مزایا و معایب خود می‌باشند. از جمله روشهای موفق و پیشرو در زمینه تشخیص موضوع متون استفاده از هوش مصنوعی و یادگیری ماشین در این زمینه است. که با توجه به حساسیت و اهمیت موضوع، ضرورت ایجاد روشهایی هوشمند به منظور پردازش متون قابل توجه بوده و در این تحقیق نیز به این موضوع پرداخته شده است.

۱- ٦) ساختار پایان نامه

پژوهش حاضر در پنج فصل تنظیم شده است که فصل اول حاوی مقدمه‌ای از موضوع مورد بحث وکلیات موضوع مورد بررسی بوده در ادامه فصل دوم دربردارنده‌ی مباحث نظری و مفاهیم موجود درباره‌ی موضوع مورد بحث و مروری بر ادبیات پیشین می‌باشد در فصل سوم شامل روش پیشنهادی و ارائه راهکار پیشنهادی می‌باشد، فصل چهارم به ارزیابی و ارائه نتایج حاصل می‌پردازد و در نهایت فصل پنجم به ارائه‌ی نتیجه‌گیری کلی از روش پیشنهاد شده در این پایان‌نامه پرداخته است.

فصل دوم

مبانی نظری و پیشینه‌ی تحقیق

۲- ۱) مقدمه

در دنیای امروز روزانه حجم زیادی(حجمی معادل چند ترابایت یا پتابایت) از داده‌های گوناگون در هر منظری از زندگی روزانه بشر، در حال ذخیره و جمع‌آوری هستند. در واقع رشد اطلاعات در سطح جهان روندی کاملاً تصاعدی و شگفت‌انگیز را طی می‌کند. آنالیز، تجزیه وتحلیل اینچنین داده‌ها و اطلاعات یک نیاز مهم و اساسی برای دنیای امروز است که به صورت خودکار از میان حجم عظیمی از داده‌ها و اطلاعات خام، اطلاعات با ارزش را استخراج و کشف کند و نیز آنها را به دانش سازمان یافته تبدیل کند. نیاز امروز بشر به تولید و کشف دانش از میان انبوهی از داده‌های خام منجر به تولد داده‌کاوی شده است. در واقع داده‌کاوی به استخراج اطلاعات و دانش و کشف الگوهای پنهان مفید از مجموعه داده‌های بزرگ می‌پردازد. هرچه حجم داده‌ها بیشتر و روابط میان آنها پیچیده‌تر باشد، دسترسی به اطلاعات نهفته در آن نیز مشکل‌تر می‌شود، لذا نقش داده‌کاوی به عنوان یکی از روشهای کشف دانش، روشن می‌گردد. در دنیای واقعی، علم داده‌کاوی با مجموعه داده‌های با ابعاد بالا سروکار دارد که این داده‌ها دارای تعداد زیادی اطلاعات غیرمرتبط هستند. در بیشتر مواقع تمام ویژگی‌های داده‌ها برای یافتن دانشی که در داده نهفته است، مهم نیستند؛ این موضوع خود باعث ایجاد چالش محاسباتی می‌شود. به همین دلیل در بسیاری از زمینه‌ها کاهش ابعاد داده یکی از مباحث مورد پژوهش است چراکه در کاهش ابعاد داده اغلب داده‌های غیرمهم از بین می‌روند. بنابراین ابعاد زیاد داده‌ها را می‌توان با استفاده از روش‌های کارآمد کاهش

داد. یکی از این روشها، روش استخراج و انتخاب ویژگی است که این روش با انتخاب ویژگی‌های کارآمد و مهم ابعاد زیاد داده‌ها را کاهش می‌دهد.

۲- ۲) توصیف مفاهیم اصلی

در این قسمت به بررسی و توصیف مفاهیم به کار رفته در این تحقیق پرداخته شده است، توصیف این مفاهیم به درک بهتر مطالب توسط خوانندگان محترم کمک می‌کند.

۲- ۲- ۱) متن‌کاوی

امروزه بخش وسیعی از دانش به‌صورت متن، مستندات و دیگر صورت‌های رسانه‌ای نگهداری می‌شوند که همه آن‌ها به‌صورت غیر ساختاریافته هستند. یکی از کاربردهای داده‌کاوی، متن‌کاوی است. بـرای دریافت دانش از اطلاعات یک متن، لازم است ابتدا آن را درک کـرد، سـپس پـردازش کرد تا فهمید چه معانی و مفاهیمی در آن موجود است؛ چه ارتباطی میان مفاهیم وجود دارد و از میان این مفاهیم کدام جدید و کدام قدیمی است؛ ازاین‌رو در عصر فنـاوری، هر چیزی باید بتواند به‌صورت خودکار، انجام شود. درک معنی متون نیز از این جمله کارها محسوب می‌شود. متن‌کاوی، کـاوش داده‌های متنـی و یا کشف دانش در متن از نام‌های مورد قبول در این زمینه هستند. مفهوم متن‌کاوی که به دریافت تمام اطلاعات موردنیاز از داده‌های متنی اشاره می‌کند، تقریباً عمری برابر با خود بازیابی اطلاعات دارد. به‌هرحال، متن‌کاوی دارای ویژگی‌های منحصربه‌فرد و اساسی است که باعث شده بین آن و بازیابی اطلاعات تمیز قائل شوند. متن‌کاوی در به دست آوردن اطلاعات مفیدی از داده‌های متنی که ذاتاً ساختار نیافته، غیرمتشکل و نامنظم هستند، کمک می‌کند.

متن‌کاوی و یا کشف دانش از متن، اشاره به فرآیندی می‌کند که باعث به دست آوردن الگوهای غیربدیهی، جالب و باکیفیت بالا و همچنین اطلاعات و دانش از اسناد متنی ساختار نیافته می‌شود. متن‌کاوی که به‌عنوان کشف دانش از متن نیز شناخته می‌شود با داده‌کاوی تفاوت دارد، به این معنا که متن‌کاوی به جستجو در میان داده‌های متنی برای استخراج کردن اطلاعات مفید می‌پردازد که معمولاً طبیعتی ساختار نیافته دارند، درحالی‌که داده‌کاوی سعی در کشف دانش از پایگاه داده‌های ساختاریافته دارد، در بیانی بسیار ساده، متن‌کاوی روند کشف و استخراج الگوهای معنادار و روابط از مجموعه متن است. پردازش زبان طبیعی تلاش می‌کند همان‌طوری که مفاهیم زبان طبیعی به‌وسیله‌ی انسان تجزیه‌وتحلیل می‌شود، برای کامپیوتر هم قابل‌فهم باشد. این حوزه، تمام فعالیت‌هایی که به‌نوعی به دنبال کسب دانش از متن هستند را شامل می‌شود. تحلیل داده‌های متنی توسط نفون یادگیری ماشین، بازیابی اطلاعات هوشمند، پردازش زبان طبیعی یا روش‌های مرتبط دیگر، همگی در زمره مقوله یادگیری متن قرار می‌گیرند. یکی از روش‌هایی که ذکر شد، استفاده از فنون یادگیری ماشین درزمینه پردازش متن است. مسئله قابل تأمل این است که این روش‌ها، در ابتدا در مورد داده‌های ساختاریافته به کار گرفته شدند و علمی به نام داده‌کاوی را به وجود آوردند. داده‌های ساختاریافته به داده‌هایی گفته می‌شود که به‌طور کاملاً مستقل از همدیگر ولی یکسان از لحاظ ساختاری در یک محل گردآوری شده‌اند.

در واقع می‌توان گفت یکی از مهمترین قسمت‌های علم داده‌کاوی، متن‌کاوی است. متن کاوی هنر و علم استخراج اطلاعات و دانش از متن است. متن کاوی فرآیند ترجمه و سازماندهی مجموعه‌ای از سندهای متنی بزرگ است تا اطلاعات مورد نیاز تصمیم گیرندگان را فراهم کند و روابط، موجودیت‌ها و دانش نهفته در متن را کشف نماید. به عبارتی دقیق‌تر متن‌کاوی دانش اکتشاف و استخراج اطلاعات مفید از متن بدون ساختار است. تکنیک متن‌کاوی شامل بازیابی اطلاعات از متن، رده‌بندی و دسته‌بندی متون و استخراج روابط، موجودیت‌ها و رخدادها

است. در مقابل پردازش زبان طبیعی تلاشی برای استخراج نمایش معنایی کاملتری از متن است. رده‌بندی و دسته‌بندی متون از مهمترین مسائل در متن‌کاوی هستند که هم به تنهایی دارای کاربرد می‌باشند و هم به عنوان بخشی از مسائل کاوش متن به کار می‌روند. به طور کلی همواره اطلاعات زیاد، نیازمند رده‌بندی می‌باشند. مهمترین موضوع در بررسی یک سری داده متنی، امر رده‌بندی و دستهبندی آنها محسوب می‌شود.

۲- ۲- ۲) گام‌ها و مراحل متن‌کاوی

۱- انتخاب متن

۲- پردازش متن

۳- تبدیل متن به صفات خاصه

٤- انتخاب صفات خاصه از متن

٥- داده‌کاوی بر روی متن (کشف دانش از متن)

٦- تفسیر و ارزیابی خروجی متن کاوی

- انتخاب متن

در این قدم مجموعه اسنادی که قصد کاوش در بین آن‌ها وجود دارد، به‌صورت متن موجود هستند. در این گام از مراحل متن کاوی باید اسناد متنی یا داده های متنی که ارزش تحلیل را دارند. گرد آوری شود.

- پردازش متن

در این قدم فرآیندهایی همچون فرمت، ساخت توکن، پاک‌سازی متن انجام می‌شود. در طی فرآیند جمع‌آوری متون، ممکن است که آن‌ها به‌خوبی سازمان‌یافته نباشند در این صورت به‌عنوان اطلاعات ازدست‌رفته یا یکپارچگی

متون غیرعقلانی تفسیر می‌شوند. اگر متون، به‌درستی بررسی نشوند آنگاه متن‌کاوی ممکن است منجر به پدیده «ایجاد خروجی غلط توسط ورودی بی‌کیفیت و ناصحیح شود. در فاز پیش‌پردازش، مستندات به تعداد ثابتی از رده‌بندی‌های از پیش تعریف‌شده سازمان‌دهی می‌شوند. پیش‌پردازش، پیاده‌سازی موفقیت‌آمیز تحلیل متن را تضمین می‌کند اما ممکن است که زمان پردازش قابل‌توجهی را مصرف کند خروجی فاز پیش‌پردازش به دو صورت زیر است.

۱- مبتنی بر سند

در این حالت نمایش درست مستندات اهمیت دارد. برای مثال تبدیل اسناد به یک فرمت میانی و نیمه ساخت‌یافته، یا به کار بردن یک نمایه بر روی آن‌ها یا هر نوع نمایش دیگری که کار کردن با اسناد را مؤثر می‌کند. هر موجودیت در این نمایش درنهایت بازهم یک سند خواهد بود.

۲- مبتنی بر مفهوم

در این حالت نمایش اسناد بهبود بخشیده می‌شود، مفاهیم و معانی موجود در سند و ارتباط میان آن‌ها و هر نوع اطلاعات مفهومی دیگری که قابل‌استخراج است، از متن استخراج می‌شود. در این حالت نه با خود موجودیت بلکه با مفاهیمی که از این مستندات استخراج‌شده‌اند، مواجه هستیم.

- تبدیل متن به صفات خاصه

در این قدم از متون پردازش‌شده صفات خاصه استخراج می‌شود. فرایند استخراج ویژگی شامل مراحل زیر است

تجزیه‌وتحلیل مورفولوژیک: این روش با تک‌تک کلمات موجود در یک سند متنی سروکار دارد و شامل مراحل زیر است:

توکن بندی: در این مرحله سند از طریق حذف فضاهای خالی، کاما و کلیه علائم نگارشی به دنباله‌ای از رشته لغات تبدیل می‌شود.

حذف لغات توقف: در این مرحله لغات بازدارنده مانندa، Theو یا or از متون حذف می‌شوند. این مرحله از طریق کاهش تعداد لغات موجب افزایش اثربخشی و کارایی می‌شود.

ریشه‌یابی: این مرحله تکنیک نرمال‌سازی زبان‌شناسی است و برای تبدیل لغت به فرم ریشه به کار می‌رود. مثلاً لغت honesty به لغت honest و یا walking به walk تبدیل می‌شود.

تجزیه وتحلیل نحوی: این قسمت بر روی ساختار یک‌زبان که اغلب نحو نامیده می‌شود تأکید دارد. به‌عنوان‌مثال زبان انگلیسی شامل اسم، فعل، قید، نقطه‌گذاری و دیگر بخش‌های گفتاری می‌شود.

برچسب‌گذاری اجزای واژگانی کلام: این نشانه‌گذاری معمولاً برای اضافه کردن دانش دستوری به یک لغت از یک جمله به کار می‌رود. اگر کلاس واژگانی کلمه شناخته‌شده باشد، آنگاه انجام تجزیه‌وتحلیل زبانی راحت‌تر است.

پارسینگ: تکنیکی است که برای بررسی ساختار گرامی یک جمله به کار می‌رود. جملات در یک ساختاری شبیه درخت نمایش داده می‌شوند که اصطلاحاً به آن درخت پارس گفته می‌شود که در اصل برای تجزیه‌وتحلیل درخواست‌های دستور زبانی صحیح در یک جمله به کار می‌رود. درخت پارس می‌توان با دو رویکرد بالا به پایین و یا پایین به بالا ساخته شود.

تجزیه‌وتحلیل معنایی: بر یافتن ارتباط معنادار بین واژگان تأکید دارد .چگونه معنای یک جمله به معنای عبارات، کلمات و تک‌واژه‌های تشکیل‌دهنده آن مربوط می‌شود.

- انتخاب صفات خاصه از متن

در این قدم تعدادی از صفات خاصه برای انجام کاوش انتخاب می‌شوند، زیرا همه صفات خاصه برای انجام کاوش مفید واقع نیستند. انتخاب ویژگی شامل ۳ تکنیک زیر است.

۱- تکنیک انتخاب ویژگی مبتنی بر تکرار

هدف اصلی از انتخاب ویژگی، از بین بردن اطلاعات نامربوط و مزاحم از متن موردنظر است. در این قسمت مهم‌ترین ویژگی‌ها را از طریق امتیاز لغات انتخاب می‌کند. اهمیت لغت در سند توسط نمره اختصاص داده‌شده به آن مشخص می‌شود. سند متن به‌عنوان یک مدل فضای برداری ارائه می‌شود. در این مدل هر بعد نشان‌دهنده یک اصطلاح مجزا، از یک کلمه، کلمه کلیدی یا یک عبارت است. ماتریس سند توسط n سند و m اصطلاح نشان داده می‌شود. مقادیر غیر صفر در این ماتریس نشان‌دهنده حضور اصطلاح در سند است.

۲- تکنیک اندیس گذاری معناییِ نهفته

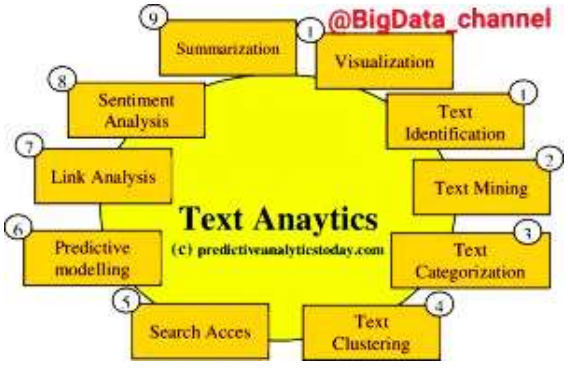
در روش‌های قبلی متون به‌تنهایی و بدون در نظر گرفتن کل مجموعه پردازش می‌شدند و اگر تصمیمی مبنی بر جواب بودن یک متن گرفته می‌شد، آن تصمیم کاملاً متکی به همان متن و مستقل از متون دیگر گرفته‌شده و هیچ توجهی به وابستگی موجود بین متون مختلف و ارتباط بین آن‌ها نمی‌شد که این مسئله یکی از عوامل پایین بودن دقت جستجوها و ناکارآمدی آن‌ها به شمار می‌رفت. این روش بر پایه تحلیل معنایی نهفته بناشده است؛ که گامی را به مجموعه مراحل موجود در پروسه اندیس گذاری اضافه می‌کرد. این روش بجای آنکه در اندیس گذاری تنها یک متن را در نظر بگیرد، کل مجموعه اسناد را باهم و در کنار یکدیگر در نظر می‌گرفت تا ببیند که چه اسنادی لغات مشابه با لغات موجود در سند موردبررسی رادارند.

۳- تکنیک نگاشت تصادفی

هنگامی‌که بردارهای داده‌ای دارای ابعاد بسیار بالایی هستند، استفاده از الگوریتم‌های تشخیص الگو و یا تحلیل داده که مکرراً مشابهات و یا فاصله فضای داده‌های اصلی را محاسبه می‌کنند غیرممکن است. LSI تطبیق واژگانی را از طریق اتخاذ یک رویکرد معنایی بهبود می‌بخشد، درحالی‌که تکنیک نگاشت تصادفی یک نقشه از محتوای یک مجموعه سند بزرگ ایجاد می‌کند. هر منطقه انتخاب‌شده در یک نقشه بیشتر می‌تواند برای استخراج اسناد جدید در موضوعات مشابه استفاده شود. تکنیک تصادفی شامل ماتریس تصادفی است که از ضرب بردارهای اصلی به دست می‌آید و یک بردار کاهش‌یافته ایجاد می‌کند[۱۸].

۵- داده‌کاوی بر روی متن (کشف دانش از متن)

با توجه به صفات خاصه‌ی انتخاب‌شده در قدم قبل، در این قدم بر روی این صفات خاصه برای استخراج الگوهای مناسب، کاوش انجام می‌شود. همچنین موارد زیر به عنوان انواع روش های داده کاوی بر روی متن یا به عبارتی متن کاوی متصور است.



شکل ۲- ۱) روش‌های متن کاوی

٦- تفسیر و ارزیابی خروجی متن کاوی

درنهایت نتایج به‌دست‌آمده مورد ارزیابی قرارگرفته و برای موارد مختلف تفسیر و استفاده می‌شود. در ارزیابی معمولاً معیارهای زیر متصور است.

۲- ۲- ۳) مسئله رده‌بندی متون

یکی از مسائل مهم در بررسی داده‌های متنی، امر رده‌بندی و دسته‌بندی آنها محسوب می‌شود. رده‌بندی به این مفهوم است که یک سری داده فراهم شده را به چند رده از قبل تعریف شده اختصاص دهد. زمانی که با مجموعه‌ای از متون کار می‌کنیم، بحث رده‌بندی به صورت طبقه‌بندی متون و یا رده‌بندی متون تعریف می‌شود. رده‌بندی متون به این صورت تعریف می‌شود که مجموعه‌ای از رده‌ها (موضوعات متن‌ها) به همراه مجموعه داده متون تهیه می‌شود و هدف یافتن رده و یا به عبارتی موضوع صحیح برای هر متن است. مسئله رده‌بندی متون در حالت کلی به صورت برچسب زدن به اسناد بر اساس تعدادی گروه از پیش تعیین شده می‌باشد[۹].

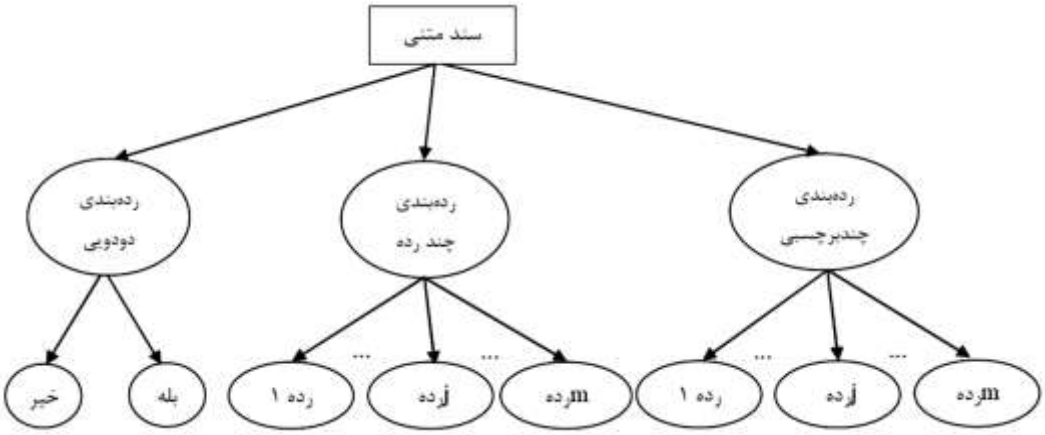
رده‌بندی و دسته‌بندی متون از مهمترین مسائل در متن‌کاوی هستند که هم به تنهایی دارای کاربرد می‌باشند و هم به عنوان بخشی از مسائل کاوش متن به کار می‌روند. به طور کلی همواره اطلاعات زیاد، نیازمند رده‌بند می‌باشند. مهمترین موضوع در بررسی یک سری داده متنی، امر رده‌بندی و دسته‌بندی آنها محسوب می‌شود . به طور خلاصه، رده‌بندی به این مفهوم است که یک سری داده فراهم شده را به چند رده از قبل تعریف شده اختصاص دهیم. زمانی که با مجموعه‌ای از اسناد متنی کار می‌کنیم بحث رده‌بندی به صورت طبقه‌بندی متون و یا رده‌بندی متون تعریف می‌شود. تعریف رده‌بندی متون به این صورت است که مجمو عه‌ای از رده‌ها (موضوعات متن‌ها) به همراه یک سری سند متنی تهیه می‌شود و هدف یافتن رده و یا به عبارتی موضوع صحیح برای هر متن است. در حال حاضر اطلاعات متنی بسیار زیادی در اینترنت موجود است که این موجب شده تا دیگر قادر به رده‌بندی به صورت دستی نباشیم. بنابراین رده‌بندی اطلاعات متنی به صورت خودکار به رده‌های مختلف به امری ضروری

تبدیل شده است. علاوه بر این، سرویس‌های جدیدی که قبلاً وجود نداشتند با رشد اطلاعات اینترنتی پدیدار شده‌اند. از جمله این سرویس‌ها می‌توان صافی‌کننده‌های پست الکترونیکی و صفحات وب را نام برد. صافی کننده‌های پست الکترونیکی باعث جلوگیری از ارسال نامه‌های تجاری ناخواسته و یا هرزنامه‌های الکترونیکی به افراد نامشخص می‌شوند. این صافی‌کننده‌ها با رده‌بندی نامه‌های الکترونیکی به نامه‌های معمولی و هرزنامه‌ها این کار را انجام می‌دهد. صافی‌کننده‌های صفحات وب به طور کلی از دسترسی افراد با سنین کم به سایتهای شامل مطالب ناخوش آیند، مانند متون دارای مضامین خشونت و یا جنسی، جلوگیری می‌کنند. صافی‌کننده‌های صفحات وب، متون را به دو رده بی زیان و مضر طبقه‌بندی می‌کنند. رده بندی متون به سه فرم کلی که شکل ۳- ۲ نمایش داده شده است، می‌باشد. این سه فرم به ترتیب رده بندی دودویی، رده بندی چند-رده و رده بندی چند- برچسبی هستند.

- در رده‌بندی دودویی یک متن، تنها متعلق به یکی از دو رده داد ه شده خواهد بود. بنابراین رده‌بند بایستی متن را به یکی از دو رده نسبت دهد.

- در رده‌بندی چند-رده یک متن نمونه تنها به یکی از چندین رده تعریف شده تعلق خواهد یافت.

-در رده‌بندی چند- برچسبی، یک متن نمونه ممکن است به چندین رده تعلق داشته باشد. به عبارتی دیگر در این نوع رده بندی رده‌ها دارای سندهای متنی مشترکی خواهند بود.



شکل ۳- ۲) انواع رده‌بندی متون

در مسئله رده‌بندی متون، رده‌بندی‌ها ممکن است از جنبه دیگری به تک - برچسبی، چند - برچسبی، سند- محور، رده- محور، سخت و نرم دسته بندی شوند.

۲- ۲- ٤) استخراج ویژگی

با پیشرفت علم حجم اسناد متنی موجود بر روی رسانه‌های دیجیتال و اینترنت، افزایش یافته است و این موضوع ضرورت استفاده از سیستم‌های خودکار تشخیص و دسته‌بندی متن را بیشتر پررنگ می‌کند. روش‌های دسته‌بندی متن جزو روش‌های یادگیری ماشین هستند و استخراج و انتخاب ویژگی مرحله‌ی بسیار مهم در رویه‌ی دسته‌بندی متون به شمار می‌رود، زیرا در این مرحله واژه‌های کلیدی انتخاب می‌شوند تا به‌عنوان بهترین نمایش‌دهنده برای سند متنی مورد استفاده قرار بگیرند. هدف روش‌های انتخاب ویژگی به دست آوردن یک مجموعه‌ی کوچک‌تر از ویژگی‌های موجود در سند می‌باشند که به طرز مؤثری محتوای سند را بیان می‌کند. الگوریتم‌های مختلفی برای دسته‌بندی متون وجود دارد. مشکلی که در دسته‌بندی متن وجود دارد، حجم زیاد ویژگی‌ها است که باعث کاهش

دقت نتایج دسته‌بندی می‌شود .برای انتخاب و برای حل این مشکل و کاهش ابعاد ویژگی‌ها از متدهای انتخاب ویژگی استفاده می‌کنند.

در واقع استخراج ویژگی فرایندی است که در آن با انجام عملیاتی بر روی داده‌ها، ویژگی‌های بارز و تعیین‌کننده آن مشخص می‌شود. هدف استخراج ویژگی این است که داده‌های خام به شکل قابل استفاده‌تری برای پردازش‌های آماری بعدی درآیند.

روشهای مختلف انتخاب ویژگی، تلاش می‌کنند تا از میان 2^N مجموعه کاندید برای یک مجموعه N عضوی، بهترین زیرمجموعه را بر اساس تابع ارزیابی پیدا کنند. در تمام این روشها بر اساس کاربرد و نوع تعریف، زیرمجموعه‌ای به عنوان جواب انتخاب می‌شود که بتواند مقدار یک تابع ارزیابی را بهینه کند. باوجود اینکه هر روشی سعی می‌کند بهترین ویژگی‌ها را انتخاب کند، اما با توجه به وسعت جواب‌های ممکن و اینکه این مجموعه جوابها به صورت توانی با N افزایش می‌یابد، پیدا کردن جواب بهینه مشکل و در N های متوسط و بزرگ بسیار پرهزینه است. به طورکلی روشهای مختلف انتخاب ویژگی را بر اساس نوع جستجو به دسته‌های مختلفی تقسیم-بندی می‌کنند. در بعضی روشها تمام فضای ممکن جستجو می‌شود. روش‌های مختلف استخراج ویژگی ممکن است یک یا چند کار زیر را انجام دهند.

- حذف نویز داده‌ها
- جداسازی اجزای مستقل داده‌ها
- کاهش ابعاد برای تولید بازنمایی مختصرتر
- افزایش بعد برای تولید بازنمایی جدایی‌پذیرتر

الف) پیش پردازش

اولین مرحله در دسته‌بندی متن تبدیل اسناد به صورت رشته‌ای از کاراکترها با فرمت‌های مختلف می‌باشد که برای روش‌های یادگیری و طبقه‌بندی نمایش داده می‌شود. همواره بهتر است در بازیابی اطلاعات ریشه کلمه را پیدا کرده تا بتوان آن کلمه را به صورت واحد در اسناد به کار برد و این کلمه‌ی واحد، منجر به نمایش مقدار ویژگی در متن می‌شود.

ب) توکن‌بندی

این فرآیند انجام می‌گیرد. این فرآیند به این صورت است که جریان متن به کلمه‌ها، عبارات، نشانه‌ها یا عناصر معنی‌دار شکسته می‌شود که به هر کدام از آن‌ها توکن گفته شده و به این فرآیند توکن‌بندی Tokenization می‌گویند.

پ) ریشه‌یابی

در مرحله ریشه‌یابی، ریشه کلمه‌ها به فرم اصلی در می‌آید و هر گونه پیشوند و پسوندی از ابتدا و انتهای آن حذف می‌شود. حذف کلمات ایست یا توقف کلمات ایست به کلمه‌هایی گفته می‌شود که حاوی هیچ‌گونه معنی مفیدی نیستند مانند حروف ربط و حروف اضافه. لیست کلمات توقف برای اکثر زبان ها باید استخراج شود که با استفاده از این لیست می‌توان دید که اگر این کلمه‌ها در داخل اسناد متنی وجود داشته باشد از اسناد حذف می‌شود و اگر کلمه در لیست نبود حذف نمی‌شود و با این کار از تعداد ویژگی‌ها کم می‌شود. نمونه‌ای از کلمات

ایست می‌توان به کلمه‌های و، در، به، که، از، این، را، است، با، برای و غیره اشاره کرد. عبارات تاکید یا نشانه‌گذاری هم شامل " " ? : ,] [: / ... . ! } { *) (# _ – می‌باشد.

ت) نمایش متون

برای نمایش متون می‌توان از روش فضای برداری استفاده نمود. با این نمایش می‌توان ویژگی‌ها را از داخل اسناد استخراج کرد. در مدل فضای برداری، سندها به وسیله برداری از کلمه‌ها نمایش داده می‌شوند و مجموعه سندها به وسیله ماتریس کلمه در سندA ، نمایش داده می‌شوند.

ث) انتخاب خصیصه یا ویژگی

انتخاب ویژگی یا خصیصه یک مرحله‌ی بسیار مهم در رویه‌ی دسته‌بندی به شمار می‌رود، زیرا در این مرحله واژه‌های کلیدی انتخاب می‌شوند تا به عنوان بهترین نمایش‌دهنده برای سند متنی مورد استفاده قرار بگیرند. اگر تعداد واژه‌های کلیدی انتخاب شده کم باشد صحت و کارایی سیستم تحت تاثیر قرار می‌گیرد و کاهش می‌یابد و در مقابل اگر تعداد واژه‌های کلیدی انتخاب شده زیاد باشد باعث کاهش کارایی سیستم در بعد زمان خواهد شد و سرعت آموزش در فاز آموزش پایین می‌آید[۲].

۲- ۲- ۵) روش‌های انتخاب ویژگی

روش‌های انتخاب ویژگی در دو دسته‌ی زیر طبقه‌بندی می‌شوند.

- روش‌های فیلتری یا پالایشی یا آماری Filtering method
- روش‌های روکشی یا پوششی Wrapper methods

این روش‌ها ساده‌ترین روش انتخاب ویژگی‌ها می‌باشند و اساس آن بر پایه‌ی نگه داشتن ویژگی‌هایی می‌باشد که بیشترین امتیاز را از تابعی که اهمیت نسبی یک واژه را می‌سنجد دریافت می‌کند. هر سند Di شامل تمام کلمه‌ها، فاصله‌ها، علایم و برچسب‌هایی هست که در آن سند موجود می‌باشد.

- روش فرکانس سند[1]

معیارفرکانس سند به ازای هر کلمه به صورت جداگانه محاسبه می‌شود. این معیار برای هر کلمه برابر با نسبت تعداد سندهای متنی شامل کلمه مورد نظر به کل تعداد سندها می‌باشد. پس از محاسبه معیار، ویژگی‌هایی به عنوان ویژگی‌های مفید انتخاب خواهند شد که مقدارشان از یک حد آستانه بیشتر باشد. فرض بر این است که کلمات بسیار نادر دارای اطلاعات زیادی برای پیش بینی رده نیستند و یا اینکه در کارایی سیستم رده بندی متون تاثیر ندارند. به هر حال حذف کلمات نادر باعث کاهش ابعاد فضای ویژگی نیز خواهد شد، همچنین با حذف کلمات نویزی که معمولاً تعداد تکرار کمی دارند می‌توان دقت رده‌بندی را افزایش داد

روش فرکانس سند ساده‌ترین تکنیک برای کاهش یک مجموعه واژگان است. این معیار برای مجموعه داده‌های بزرگتر با پیچیدگی محاسباتی تقریبًا خطی نسبت به داده های آموزشی، به راحتی قابل تعمیم است. با اینکه روش فرکانس سند به عنوان شیوه‌ای مناسب برای افزایش کارایی تعریف می‌شود ولی در مسئله انتخاب ویژگی روش بسیار کارایی محسوب نمی‌شود.

- روش فرکانس کلمه[2]

[1] Document Frequency(DF)
[2] Term Frequency(TF)

کاهش فضای ویژگی یکی از مهمترین گامهای رده‌بندی متن می‌باشد. حذف عباراتی که ارزش اطلاعاتی کمتری دارند یا داده نویز هستند، کارایی الگوریتم رده‌بندی را افزایش می‌دهد و سربار محاسباتی را کاهش می‌دهد. برای کاهش فضای ویژگی کافی است ارزش اطلاعاتی کلیه عبارات را پیدا کنیم و عبارات را بر اساس ارزش اطلاعاتی مرتب کنیم. برای انتخاب عبارات مناسب یک حد آستانه گذاشته و عباراتی را که ارزش اطلاعاتی کمتر از این حد داشته باشند حذف می‌کنیم. روش انحراف معیار فرکانس کلمه روشی وابسته به مجموعه‌ای از رده‌های ممکن به صورت $C = c_1 . \ldots . c_k$ داده شده است. برای هر کلمه ابتدا فرکانس و تعداد تکرار در هر رده و سپس متوسط فراوانی در کلیه ردهها را محاسبه می‌کنیم. معیار انحراف معیار فرکانس کلمه برای ویژگی f با استفاده از رابطه‌ی ذیل حاصل می‌گردد.

$$\mathrm{TFV} = \sum_{i=1}^{k} [\mathrm{tf}(f. c_i) - \mathrm{mean\_tf}(f)]^2 \qquad (۱-۲)$$

- روش فرکانس کلمه–فرکانس معکوس سندTF-IDF [1]

یکی از قالب‌های بسیار رایج مورد استفاده برای وزن دهی ویژگی‌ها است تقریباً همه روش‌های وزن‌دهی دیگر به نحوی بر گرفته از این روش هستند. که این روش به شرحی که در ادامه بیان شده است طراحی شده است.

۱- بسامد مربوط به یک کلمه در سند به عنوان وزن محلی محسوب می‌شود.

۲- معکوس بسامد سند به عنوان وزن کلی در این روش وزن‌دهی در نظر گرفته می‌شود. این مقدار معکوس موجب می‌شود کلمات و ویژگی‌هایی که تعداد تکرار زیادی در بین سندهای متنی مجموعه دارند، جریمه شوند.

۳- نرمال کردن کسینوسی که به ندرت، با تقسیم بر میانگین هندسی به کار گرفته می‌شود.

[1] TF-Inverse Document Frequency(IDF)

- روش بهره اطلاعاتی[1]

روش بهره اطلاعاتی یکی از محبوب‌ترین روش‌های انتخاب ویژگی در زمینه مباحث یادگیری ماشین است. به عنوان نمونه از این روش در الگوریتم استنتاجی درخت تصمیم گیری C4.5 استفاده شده است.

$$IG = -\sum_{C} P(C)\log P(C) + P(w)\sum_{C} P(C|w)\log P(C|w) + P(\overline{w})\sum_{C} P(C|\overline{w})\log P(C|\overline{w})) \qquad (۲-۲)$$

در این رابطه C مجموعه رده‌ها را نشان می‌دهد و w کلمه ای است که بهره اطلاعاتی برای آن محاسبه می‌شود. به طور کلی روش بهره اطلاعاتی، میانگین کاهش آنتروپی را که از رخداد یا فقدان یک کلمه بدست می‌آید محاسبه می‌کند. به عبارتی دقیق‌تر این روش معیاری است از تعداد بیت اطلاعاتی که از وجود یا عدم وجود یک کلمه برای هر سند متنی به دست می‌آید. معیار بهره اطلاعاتی در کاربرد انتخاب ویژگی به صورت زیر محاسبه خواهد شد.

$$IG(t.c) = P(t.c)\log\frac{P(t.c)}{P(t)*P(c)} + P(\bar{t}.c)\log\frac{P(\bar{t}.c)}{P(\overline{t})*P(c)} \qquad (۲-۳)$$

برای تمام ویژگی‌های در مجموعه متون و به ازای هر رده این مقدار بهره اطلاعاتی محاسبه می‌شود و سپس بیشترین مقدار و یا میانگین بین تمام این مقادیر به عنوان بهره اطلاعاتی آن ویژگی در نظر گرفته می‌شود. پس از محاسبه معیار بهره اطلاعاتی برای تمام ویژگی‌ها، کلماتی که دارای مقدار کمتر از یک حد آستانه باشند از مجموعه ویژگی‌ها حذف خواهند شد.

[1] Information Gain (IG)

- روش اطلاعات متقابل[1]

معیار اطلاعات متقابل بین یک کلمه t و یک رده c به صورتی که در رابطه‌ی ۲- ۳ نشان داده شده است، می‌باشد. مقدار P(t) احتمال رخداد یک کلمه t در یک سند است و P(t) احتمال مربوط به رده c می‌باشد. همچنین P(t,c) احتمال اشتراکی بین کلمهt ورده c است.

$$MI(t_i.c) = \sum_{t_i \in \{0.1\}} \sum_{c \in \{+.-\}} P(t_i.c) \log \frac{P(t_i.c)}{P(t_i)P(c)} \qquad (۲- ٤)$$

زمانی که t متعلق به یک رده خاص نسبت به سایر رده‌ها باشد، معیار اطلاعات متقابل زیاد خواهد شداین معیار برای تمام کلمات و تمام رده‌ها محاسبه شده و سپس بیشترین مقدار و یا میانگین بین تمام این مقادیر به عنوان مقدار اطلاعات متقابل آن ویژگی در نظر گرفته می‌شود. مسلمًا انتظار این است که ویژگی‌های با بیشترین مقادیر اطلاعات متقابل به عنوان کلمات کلیدی رده‌ها برگزیده شود.

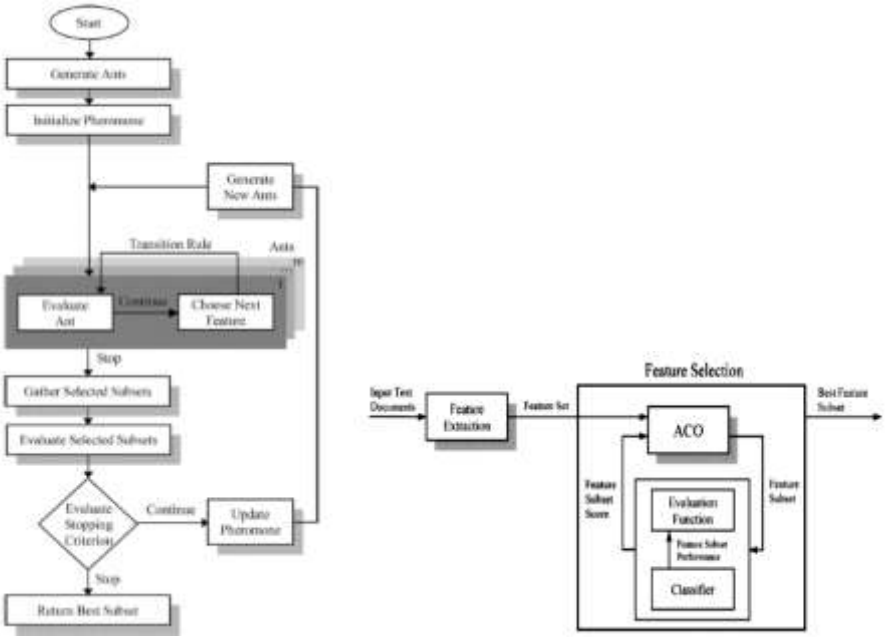
- روش ضریب همبستگی[2]

این روش از جذر روش مربع چی برای انجام محاسبات استفاده می‌کند. با توجه به اینکه این معیار نیز همانند اکثر معیارهای قبلی مقادیر را به صورت جفت ویژگی و رده محاسبه می‌کند، باید مقدار این معیار را برای هر ویژگی به تنهایی استخراج کنیم. برای انجام این عمل می‌توان میانگین و یا ماکزیمم مقادیر این معیار برای هر ویژگی در رده‌های مختلف را بررسی کرده و جواب بهتر را برگزید.

[1] Mutual Information (MI)
[2] Correlation Coefficient

۲- ۳) مرور ادبیات پیشین

حسین زاده و همکارانش در سال ۲۰۰۹ مقاله‌ای تحت عنوان انتخاب ویژگی در متن با استفاده از الگوریتم کلونی مورچه ارائه کردند که در این مقاله بیان شده است، انتخاب ویژگی و استخراج ویژگی مهمترین مراحل در سیستمهای طبقه بندی است. معمولاً انتخاب ویژگی برای کاهش ابعاد مجموعه داده‌ها با ده‌ها یا صدها هزار ویژگی مورد استفاده قرار می‌گیرد که پردازش بیشتر غیرممکن است. یکی از مسائلی که انتخاب ویژگی در آن ضروری است، طبقه بندی متن است. یک مشکل عمده طبقه بندی متن، ابعاد بالای فضای ویژگی است. بنابراین، انتخاب ویژگی مهمترین مرحله در طبقه بندی متن است. در حال حاضر روش‌های زیادی برای انتخاب ویژگی متن وجود دارد. برای بهبود عملکرد طبقه بندی متن، در این مقاله ما یک الگوریتم انتخاب ویژگی جدید ارائه می دهیم که مبتنی بر بهینه سازی کلونی مورچه است. الگوریتم بهینه سازی کلونی مورچه‌ها با مشاهده مورچه های واقعی در جستجوی کوتاهترین مسیرها به منابع غذایی الهام گرفته شده است.الگوریتم پیشنهادی به راحتی قابل اجرا است و به دلیل استفاده از یک طبقه بندی کننده ساده، پیچیدگی محاسباتی آن بسیار کم است.



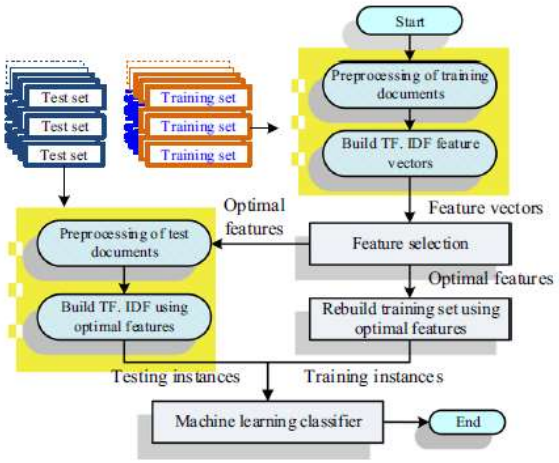
شکل ۲- ۲) الگوریتم انتخاب ویژگی پیشنهادی

این مقاله به نقص رویکردهای آماری برای انتخاب ویژگی‌ها می‌پردازد. این تکنیک‌ها معمولاً نمی‌توانند به طور کامل باعث کاهش بهینه ویژگی شوند، زیرا هیچ اکتشافی کامل نمی‌تواند بهینه بودن را تضمین کند. بنابراین، رویکردهای تصادفی سازوکار امیدوار کننده انتخاب ویژگی را فراهم می کند. ما یک روش انتخاب ویژگی بهینه جدید بر اساس بهینه‌سازی کلونی مورچه (ACO) پیشنهاد شده است. ACO توانایی همگرایی سریع را دارد. این قابلیت جستجوی قوی در فضای مسئله را دارد و می تواند به طور کارآمد زیرمجموعه حداقل ویژگی را پیدا کند. در این روش از طبقه بندی ساده (طبقه‌بندی نزدیکترین همسایه) استفاده شده است که می‌تواند بر عملکرد دسته‌بندی تأثیر بگذارد. نتایج شبیه‌سازی در مجموعه داده‌های Reuters-21578 برتری الگوریتم پیشنهادی را نشان می‌دهد[۱۱].

باهاسینی و همکارانش در سال ۲۰۱۸ مقاله‌ای تحت عنوان انتخاب ویژگی با استفاده از Chi-Square بهبود یافته برای طبقه‌بندی متن عربی ارائه کردند در این مقاله بیان شده است، در استخراج متن، انتخاب ویژگی یک روش معمول برای کاهش تعداد زیادی از ویژگی‌های فضا و بهبود دقت طبقه‌بندی است. در این مقاله، یک روش بهبود یافته برای طبقه‌بندی متن عربی پیشنهاد شده است که برای افزایش عملکرد طبقه‌بندی از انتخاب ویژگی-Chi Square استفاده می کند (که در اینجا به عنوان ImpCHI نامیده می‌شود). علاوه بر این، همچنین این روش با سه معیار انتخاب ویژگی سنتی یعنی اطلاعات متقابل، کسب اطلاعات و مجذور کای مقایسه شده است. با توجه به کار قبلی، کار فعلی را ارزیابی می‌شود، و روش را از نظر سایر روش‌های ارزیابی با استفاده از طبقه‌بندی SVM ارزیابی شده است. برای این منظور، یک مجموعه داده از ۵۰۷۰ سند عربی در شش کلاس مستقل طبقه‌بندی می‌شود. انتخاب ویژگی‌ها در کاهش داده‌های بزرگ در طبقه‌بندی متن موثر است. این کار می‌تواند روند طبقه‌بندی را بهبود بخشد. در این مقاله با استفاده از روش Chi-Square بهبود یافته استخرج ویژگی صورت می‌پذیرد بدین منظور به هر ویژگی در هر کلاس وزنی اختصاص داده می شود. سپس، تمام این وزنها با یک حداکثر وزن نهایی ترکیب می‌شوند. از نظر عملکرد، یافته های تجربی نشان می‌دهد که ترکیب روش ImpCHI و طبقه‌بندی SVM از نظر دقت، فراخوان و اندازه‌گیری از سایر ترکیبات بهتر عمل می‌کند. این ترکیب به طور قابل توجهی عملکرد مدل طبقه‌بندی متن عربی را بهبود می‌بخشد. و این روش دقتی برابر ۹۰ درصد را ایجاد می‌کند[۸].

چانتار و همکارانش در سال ۲۰۱۹ مقاله‌ای با عنوان انتخاب ویژگی با استفاده از بهینه‌ساز گرگ خاکستری باینری مبتنی بر نخبگان برای طبقه‌بندی متن عربی ارائه کردند در این مقاله به این نکته توجه شده است که، در این فرآیند، انتخاب ویژگی یک مرحله اساسی است زیرا ممکن است هزاران مجموعه ویژگی ممکن در طبقه بندی

متن در نظر گرفته شود. در این مقاله یک بهینه ساز گرگ خاکستری باینری (GWO) در یک رویکرد انتخاب ویژگی پیچیده برای حل مشکلات طبقه بندی متن عربی پیشنهاد شده است. GWO باینری پیشنهادی به منظور انتخاب ویژگی پیشنهاد شده است. عملکرد روش پیشنهادی با استفاده از مدل های مختلف یادگیری، از جمله درخت تصمیم، نزدیکترین همسایه K، طبقه‌بندی کننده‌های Naive Bayes و SVM، بررسی شده است. برای ارزیابی کارایی روشهای مختلف بسته بندی مبتنی بر BGWO از سه مجموعه داده عمومی عربی یعنیAlwatan ، Akbar-Alkhaleej و Al-jazeera-News استفاده شده است. در این مقاله روش انتخاب ویژگی یک مسئله بهینه‌سازی باینری است که در آن یک زیرمجموعه ویژگی می‌تواند به عنوان یک بردار دودویی نشان داده شود، هر عنصر در بردار نشان دهنده یک ویژگی واحد در مجموعه داده است. اگر یک عنصر مقدار ۱ داشته باشد، آنگاه انتخاب می شود و اگر مقدار آن ۰ باشد، انتخاب نمی‌شود. برای استفاده از روش GWO به منظور انتخاب ویژگی باید مقادیر واقعی بر روی یک فضای باینری نگاشت شود. هدف اصلی در این پژوهش عملکرد و نتایج بهینه‌ی طبقه‌بندی می‌باشد و انتخاب زیر مجموعه‌ی انتخابی به عنوان بردار ویژگی دارای اهمیت بسزایی می‌باشد. از این رو انتخاب تعداد ویژگی‌ها و میزان خطای طبقه‌بندی KNN که توسط الگوریتم BGWO صورت می‌پذیرد حائز اهیمت می‌باشد. هرچه میزان خطا و حداقل تعداد ویژگی های انتخاب شده کمتر باشد، زیر مجموعه ویژگی بهتر است[۹].



شکل ۲- ۳) فرایند طبقه بندی متن

مدل‌های یادگیری مورد بررسی در این روش درخت تصمیم، نزدیکترین همسایه K (KNN)، Naı¨ve Bayes (NB)، و طبقه بندی SVM می‌باشند. داده‌های مورد استفاده در این پژوهش داده عربی، Alwatan، Akbar-Alkhaleej و Al-jazeera-News هستند. نتایج نشان می‌دهد که روش انتخاب ویژگی مبتنی بر SVM با بهینه‌ساز GWO باینری پیشنهادی نتایج بهتری را حاصل کرده است.

رشدی و همکارش در مقاله‌ای با عنوان روش ترکیبی انتخاب ویژگی برای متن کاوی فارسی مبتنی بر الگوریتم-های تکاملی بیان کردند، امروزه با افزایش روز افزون حجم اطلاعات متنی، وجود روش‌های طبقه بندی متون ضروری به نظر می‌رسد. همچنین با رشد فزاینده‌ی منابع متنی فارسی این مهم بیشتر احساس می‌شود هرچند که هنوز کارهای صورت گرفته مخصوصاً در زمینه‌ی طبقه‌بندی متون فارسی به گستردگی لاتینی، چینی و غیره نیست. در این مقاله یک سیستم برای طبقه بندی متون فارسی ارائه شده است که توانسته معیارهای دقت، فراخوانی و کارایی کل را بهبود ببخشد. برای رسیدن به این هدف در این سیستم پس از پیش پردازش متون و استخراج ویژگی، برای کاهش ابعاد بردار ویژگی، یک روش بهبودیافته جدید انتخاب ویژگی مبتنی بر الگوریتم بهینه‌سازی ازدحام ذرات نوآوری شده است. نهایتاً روشهای طبقه‌بندی بر روی بردار ویژگی کاهش داده شده

اعمال شده است. برای ارزیابی روش انتخاب ویژگی در سیستم طبقه‌بندی ارائه شده، طبقه‌بندی کننده‌های ماشین بردار پشتیبان بیزین ساده به کارگرفته شده است. به منظور اعمال روشهای طبقه‌بندی متون و نیز اعمال روش‌های استخراج ویژگی‌های متون، بایستی ساختاری مناسب جهت نمایش سندها در نظر گرفته شود. ساده‌ترین و عمومی‌ترین روش نمایش متون، ایجاد یک فضای ویژگی از تمام کلمات بوده که در سندها وجود دارد. در این فضای ویژگی‌ها، پس از حذف کلمات خاص و گاهاً ریشه‌یابی، لیستی از تمامی کلمات در متون ایجاد شده و هر سند با توجه به اینکه در بردارنده چه کلماتی از لیست بوده و با چه وزنی این کلمات در سند اتفاق افتاده‌اند، به روش‌های مختلف نمایش داده می‌شود. در این مقاله از نسخه ۲ مجموعه داده استاندارد همشهری استفاده شده است. این پیکره شامل بیش از ۳۱۸ هزارخبربین سال های ۱۳۷۵ و ۱۳۷٦ است. الگوریتم پیشنهادی یک مدل بهبودیافته مبتنی برالگوریتم بهینه‌سازی انبوه ذرات برای انتخاب ویژگی در متن است. در واقع در این مقاله برای بهبود سرعت و دقت همگرایی الگوریتم بهینه‌سازی انبوه ذرات، از عملگر جهش(مطابق الگوریتم ژنتیک) استفاده روش ارائه شده به طور همزمان هم دارای جستجوی سراسری مناسب (حرکت ذرات با استفاده از روابط الگوریتم اجتماع ذرات) و هم دارای جستجوی محلی (به دلیل اعمال جهش بر روی ذرات) است. با افزودن عملگرجهش به الگوریتم بهینه‌سازی انبوه ذرات، احتمال فرار از مینیمم محلی به شدت افزایش پیدا کرده، و احتمال دستیابی به راهحل بهینه افزایش می‌یابد. برای افزایش دقت، فراخوانی و همچنین کارایی کلی الگوریتم-های طبقه‌بندی از روشهای متعدد پیش پردازش از قبیل شاخص گذاری اسناد و حذف کلمات اضافه در ساخت مجموعه آموزش استفاده کرده و از چهار روش و درخت تصمیم‌گیری برای طبقه‌بندی اسناد بهره برده‌ایم. همچنین برای انجام مراحلNaiveBayes, KNN ، SVM طبقه‌بندی استفاده شده است. نتایج آزمایش‌های به

دست آمده از اجرای سیستم ارائه شده بر روی مجموعه متون همشهری، حاکی از بهبود دقت، فراخوانی و کارایی کل آن است. هر چند که طبقه‌بندی کننده‌ی SVM در این تحقیق از عملکرد بهتری برخوردار است[٦].

حیجازی و همکارانش در سال ۲۰۲۱ در مقاله‌ی خود بیان داشتند، طبقه‌بندی متن یک روش محبوب در داده‌کاوی است. برای بدست آوردن اطلاعات ارزشمند از حجم وسیع داده‌ها استفاده می شود. انتخاب ویژگی یک مرحله مهم در طبقه بندی متن است. این یک روش پیش پردازش حیاتی برای تجزیه و تحلیل قدرتمند داده است، به طوریکه که فقط زیر مجموعه‌ای از ویژگی‌های اصلی داده با حذف ویژگی‌های اضافی یا زائد انتخاب می‌شود. در این مقاله، یک روش انتخاب ویژگی با استفاده از ترکیب روش Chi-Square و زنبور عسل مصنوعی (ABC) پیشنهاد شده است. از Chi-Square، یک روش فیلتر که از نظر محاسباتی سریع، ساده و توانایی مواجه شده با ویژگی‌های ابعادی بزرگ را دارد، به عنوان سطح اول فرایند انتخاب ویژگی استفاده می شود. و این روش به عنوان روشی برای کاهش تعداد ویژگی مورد استفاده قرار می‌گیرد. پس از آن، روش طبقه‌بندی، الگوریتم زنبورعسل، به عنوان سطح دوم که در آن از Naive Base به عنوان تابع ارزیابی در نظر گرفته می‌شود، استفاده می‌شود. در این بخش فرآیند انتخاب ویژگی برای ارائه زیرمجموعه‌ای از ویژگی‌ها (کلمات) که دارای دقت طبقه‌بندی بالاتری هستند، می‌باشند. در هر مرحله تعداد ویژگی‌های انتخاب شده ۷ تا ۵۰ انتخاب می‌شوند. ویژگی‌ها با استفاده از الگوریتم زنبور عسل به عنوان منابع غذایی در نظر گرفته می‌شوند که در هر مرحله اگر منبع غذایی جدید دارای ویژگی‌های بهتری باشد به عنوان ویژگی در مرحله‌ی بعد انتخاب می‌گردد. اطلاعات در هر مرحله از زنبورهای موجود جمع‌آوری شده و در نهایت بهترین انتخاب در نظر گرفته می‌شود. مجموعه داده‌ی مورد استفاده در این پژوهش داده‌های BBC که یک دیتاست زبان عربی است استفاده شده است. نتایج نشان داد که انجام این روش با تعداد کمتری از ویژگی‌ها از دقت طبقه‌بندی با استفاده از مجموعه ویژگی‌های اصلی بهتر

عمل می‌کنند. علاوه بر این، روش پیشنهادی در مقایسه با روش مجذور کای و الگوریتم ABC به عنوان یک روش انتخاب ویژگی، عملکرد بهتری داشت[٤].

شانگ و همکارانش در سال ۲۰۰۷ روشی نوین به منظور طبقه‌بندی متن ارائه کردند در این مقاله بیان شده، با توسعه وب، تعداد زیادی اسناد در اینترنت بوجود آمده‌اند. کتابخانه‌های دیجیتال، منابع خبری و داده‌های داخلی شرکت‌ها رو به افزایش هستند. از این رو طبقه‌بندی خودکار متن برای سازماندهی داده‌های انبوه اهمیت بیشتری پیدا کرده است. با این حال، مشکل اصلی طبقه بندی متن، ابعاد بالای فضای ویژگی است. در حال حاضر روش‌های زیادی برای استفاده از روش انتخاب ویژگی متن و سازماندهی این روش وجود دارد. برای بهبود عملکرد طبقه بندی متن، روش دیگری برای مواجه شدن با انتخاب ویژگی متن ارائه شده است. مطالعه ما در این مقاله بر اساس تئوری شاخص Gini است این الگوریتم در سال ١٩٨٤ ارائه شده و و به طور گسترده‌ای در الگوریتم‌های درخت تصمیم گیری CART ، SLIQ، SPRINT و روشهای هوشمند مورد استفاده قرار گرفت. در حالت کلی روش شاخص Gini به منظور طبقه‌بندی خصوصیات نامتناسب به کار می‌رود. در این روش تاکید بر این است که در انتخاب ویژگی‌ها باید به تکرار کلمات توجه گردد.

در این مقاله ما یک الگوریتم جدید مبتنی بر شاخصGini را برای کاهش ابعاد بردار ویژگی‌ها طراحی می‌کنیم. یک تابع اندازه گیری جدید از شاخص شاخصGini ساخته شده و برای طبقه بندی متن ساخته شده است. به منظور ارزیابی الگوریتم انتخاب ویژگی جدید، از سه طبقه بندی استفاده می‌کنیم که شامل ماشین بردار پشتیبانی SVM، kNN و fkNN می‌باشند این بررسی به منظور این است که تا نشان دهیم که الگوریتم جدید شاخص Gini در طبقه بندی های مختلف به چه میزان موثر است. نتایج بدست آمده از این ارزیابی‌ها در جداول زیر ارائه شده اند.

جدول ۲- ۱) نتایج بدست آمده

Measure function	SVM		kNN		fkNN	
	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1
Gini index	**69.940**	**88.591**	66.584	85.620	**67.999**	**86.537**
Inf Gain	69.436	88.445	**66.860**	85.326	67.032	86.134
Cross Entroy	69.436	88.445	66.579	85.326	67.518	86.207
CHI	67.739	88.225	66.404	**85.761**	66.846	86.060
Weigh of Evid	68.731	88.481	66.766	85.180	67.509	86.280

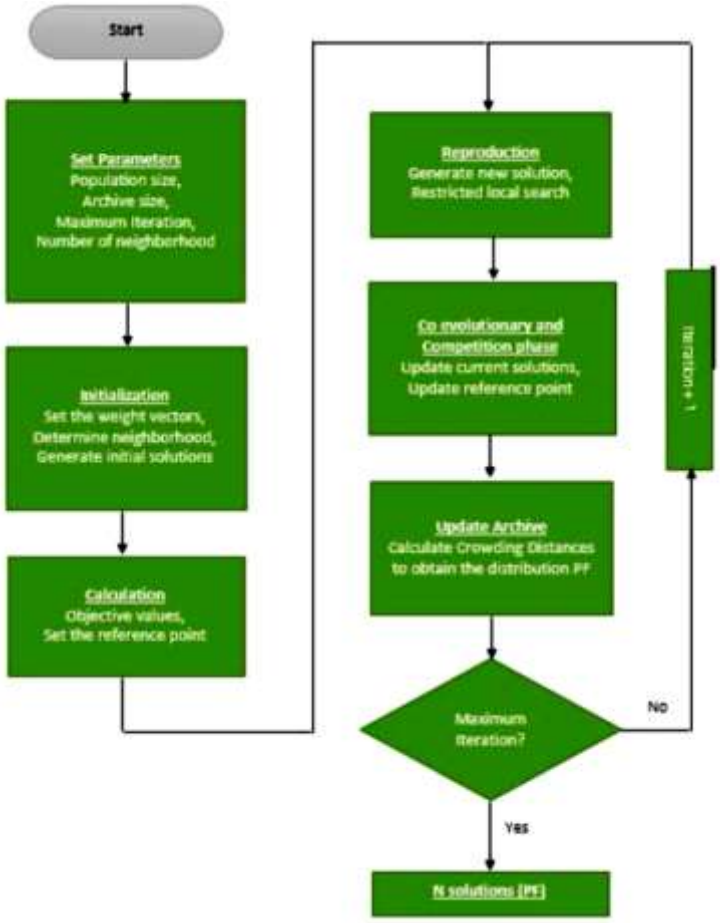
Measure function	SVM		kNN		fkNN	
	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1
Gini index	91.577	90.941	84.176	**83.043**	84.763	83.856
Inf Gain	91.531	90.708	83.318	81.301	84.346	82.811
Cross Entroy	91.481	90.708	83.318	81.301	84.216	82.578
CHI	**91.640**	**91.057**	**84.491**	82.811	85.256	84.008
Weigh of Evid	91.407	90.825	84.073	82.927	**85.867**	**85.017**

Measure function	SVM		kNN		fkNN	
	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1
Gini index	91.421	91.222	86.272	**85.222**	87.006	**86.556**
Inf Gain	**91.799**	**91.556**	**86.326**	**85.222**	**87.305**	86.556
Cross Entroy	91.419	91.222	85.764	85.111	86.999	86.444
CHI	91.238	91.000	85.770	85.000	86.898	86.444
Weigh of Evid	**91.799**	**91.556**	85.914	85.111	87.138	86.444

نتایج آزمایشات نشان می‌دهد که بهبودهای ما در شاخص جینی رفتار بهتری نسبت به سایر روش های انتخاب ویژگی دارد. آزمایش ها نشان می‌دهد که شاخص Gini بهبود یافته ما عملکرد بهتری دارد و محاسبه ساده‌تری نسبت به سایر روش‌های انتخاب ویژگی دارد .این یک روش امیدوار کننده برای انتخاب ویژگی متن است[۱۶]. طاهری و همکارانش در سال ۲۰۱۷ در مقاله‌ای بیان داشتند، مساله متن کاوی، به دلیل حجم بالای داده‌های متنی ایجاد شده در شبکه های اجتماعی مختلف، وب و دیگر اپلیکیشن های اطلاع، در سال های اخیر توجهات زیادی را به خود اختصاص داده است [۱]. به طور معمول، وظایف متن‌کاوی شامل طبقه‌بندی متن، خوشه‌بندی متن، استخراج مفهوم، تولید رده‌بندی دانه دانه، تجزیه و تحلیل احساسات، خلاصه‌سازی سند و مدلسازی رابطه مولفه است. تعدادی از خصوصیات کلیدی، داده های متن را از دیگر داده‌ها مانند داده‌های رابطه‌ای یا کمی متمایز

می‌سازد. این مساله، به طور طبیعی بر روی تکنیک‌های داده‌کاوی مورد استفاده روی این نوع داده تاثیر می‌گذارد. مهمترین خصوصیت داده‌های متن، پراکندگی و ابعاد بالای آن است. در این مورد، باید روش‌هایی برای فائق آمدن بر این خصوصیت داده‌های متن، به طور خاص طراحی شود.

در انتخاب ویژگی یکی از مهمترین بخش‌های پیش پردازش در حوزه‌ی متن‌کاوی و طبقه‌بندی متن می‌باشد. معیارهای فراوانی برای انتخاب ویژگی وجود دارد که این معیارها می‌تواند در قالب یک مساله بهینه‌سازی ارائه شود. در این تحقیق، مدلی برای انتخاب ویژگی در قالب بهینه‌سازی همزمان چند هدفه به صورت تکاملی پیشنهاد شده است. اهداف مورد بررسی در این مدل شامل دو هدف با رابطه‌ای متقابل است که به صورت همزمان مینیمم‌سازی می‌شود. اولین هدف انتخاب زیرمجموعه‌ای از کلمات با کمترین طول و دومین هدف انتخاب زیرمجموعه‌ای از کلمات با بیشترین حجم اطلاعاتی است. در بخش انتخاب ویژگی برای متن‌کاوی از معیارهای خوب بودن هر کلمه استفاده می شود. این معیارها عبارت است از: تکرار سندها (DF)، ناشناسی کلمه (TS)، رتبه‌بندی بر اساس انتروپی (ER)، توزیع کلمه (TC)، شاخص جینی (GI)، اطلاعات بدست آمده (IG)، اطلاعات متقابل (MI) و Statistic (CHI). که از بین این معیارها DF , ER , TS و TC مربوط به خوشه‌بندی متن می‌باشد و بقیه برای طبقه‌بندی متن مورد استفاده قرار می‌گیرند. در مدل پیشنهادی، از این دو معیار برای دو مساله بهینه‌سازی دو هدفه استفاده می‌کنیم. کل مجموعه داده معمولاً به دو بخش داده‌های آموزشی و داده‌های آزمایشی U تقسیم می شوند.



شکل ۲- ۴) مدل پیشنهادی مقاله

ابتدا الگوریتم پیشنهادی انتخاب ویژگی، بر روی دادههای آموزشی، جهت بدست آوردن مجموعه‌ای از ویژگی-های مرتبط F ، اعمال می‌گردد. سپس داده‌های آزمایشی با ویژگی‌های انتخاب شده به عنوان ورودی مدل طبقه‌بند، جهت فاز آزمایش و ارزیابی ایفای نقش می‌کنند. در نتایج تجربی، دقت در طبقه‌بندی را معیار ارزیابی قرار داده شده است و به همین جهت از سه طبقه‌بند ماشین بردار پشتیبان، درخت تصمیم و شبکه بیزین استفاده شده است. برای این مجموعه داده، داده‌های آزمایشی و آموزش رسما از یکدیگر جدا شده‌اند. در نتایج تجربی ارائه شده در این تحقیق، برای ارزیابی مدل پیشنهادی برای طبقه‌بندی متن، از داده واقعی ۲۰ -NEWSGROUPS استفاده شده

است. در این تحقیق، مدلی برای انتخاب ویژگی در قالب بهینه‌سازی همزمان چند هدف به صورت محاسبات تکاملی پیشنهاد شده است. انتخاب ویژگی شامل دو هدف اصلی، افزایش دقت طبقه‌بندی و کاهش تعداد ویژگیها است. این اهداف معمولاً در تناقض با یکدیگر هستند. یکی از ابزارهای برخورد با این موضوع، استفاده از الگوریتم‌های بهینه‌سازی چندهدفه است. از آنجاییکه تکنیک‌های محاسبات تکاملی روش‌های جمعیت محور می‌باشند، منحصراً مناسب برای بهینه‌سازی چندهدفه نیز هستند. از دیگر مزایای استفاده محاسبات تکاملی در انتخاب ویژگی می‌توان به مکانیزم جستجو اشاره کرد. مسائل انتخاب ویژگی دارای فضای جستجو بزرگی هستند که اغلب بسیار پیچیده است. در مقایسه با روش‌های جستجو سنتی، محاسبات تکاملی، نیازی به داشتن و دانستن اطلاعات یا فرضیاتی در مورد دامنه فضای جستجو ندارند. از دیگر مسائل مورد بحث در حوزه انتخاب ویژگی مساله تعداد ویژگی های انتخاب شده بهینه است که اغلب ناشناخته است. با انتخاب تعداد زیادی از ویژگی‌ها ممکن است ریسک انتخاب ویژگی‌های بی‌ربط و زاید بیشتر شود و در نتیجه باعث به خطر افتادن عملکرد یادگیری شود. از طرف دیگر، کوچک بودن زیر مجموعه انتخاب ویژگی‌ها ممکن است باعث شود یک سری ویژگی‌های مرتبط و خوب، حذف شوند. در نهایت این موضوع تبدیل به مساله چالش برانگیزی در حوزه انتخاب ویژگی شده است. با استفاده از محاسبات تکاملی به این موضوع فائق آمده‌ایم، بطوریکه محاسبات تکاملی با اتکا بر جمعیت در حال تکاملش، به مجموعه‌ای از جواب‌های بهینه در یک بار اجرا می‌رسد. در نهایت، فرد تصمیم گیرنده می‌تواند با توجه به تعداد ویژگی‌ها و دقت طبقه‌بندی، به جواب دلخواه خود برسد. محاسبات تکاملی، پتانسیل خوبی از خود در مسائل بهینه‌سازی در مقیاس بزرگ نشان داده‌اند که این موضوع فرصت مناسبی برای استفاده از محاسبات تکاملی در مساله انتخاب ویژگی بخصوص در حوزه متن‌کاوی است[۱۰].

قاریب و همکارانش در سال ۲۰۱۵ مقاله‌ای ارائه کردند در این مقاله روشهای انتخاب ویژگی ترکیبی بر اساس الگوریتم ژنتیک (GA) ارائه شده است. این روش از یک روش جستجوی ترکیبی استفاده می‌کند که مزایای روش انتخاب ویژگی فیلتر را با یک GA (EGA) پیشرفته در یک رویکرد دسته بندی ترکیب می‌کند تا از پس ابعاد بالای فضای ویژگی برآید و عملکرد دسته‌بندی را به طور همزمان بهبود بخشد.

عدم حساسیت GA به نویز و نیاز آن به دانش کمتری در مورد دامنه مشکل می تواند آن را به روشی قدرتمند برای کنترل کاهش ابعاد و FS برای طبقه‌بندی متن تبدیل کند. با این حال، زمان مصرف، تنظیم پارامترها و انتخاب تصادفی راه‌حل اولیه چالش‌های اصلی مرتبط با GA هستند. علاوه بر این، ابعاد بالای داده‌های متنی به دلیل وجود نویز، یک مشکل اساسی در دسته‌بندی متن است. ویژگی‌ها، که بر دقت طبقه‌بندی تأثیر منفی می‌گذارد. بنابراین، برای رفع این مشکلات، ما یک EGA برای FS پیشنهاد می‌کنیم و سپس برخی از روشهای ترکیبی FS را که ترکیبی از یک روش FS فیلتر واحد با EGA در یک چارچوب است، معرفی می‌کنیم.

در ابتدا، ما EGA را با بهبود عملگرهای crossover و جهش پیشنهاد می‌دهیم. عملیات crossover بر اساس پارتیشن بندی کروموزوم (زیر مجموعه ویژگی) با فرکانس اصطلاح و متن ورودی کروموزوم (ویژگی ها) انجام می‌پذیرد، در حالی که جهش بر اساس عملکرد طبقه بندی شده از والدین اصلی و اهمیت آن انجام می‌شود. بنابراین، عملیات cr و جهش به جای استفاده از احتمال و انتخاب تصادفی، بر اساس اطلاعات مفید انجام می‌شود. در مرحله ی بعد، شش روش شناخته شده انتخاب ویژگی فیلتر را با EGA ترکیب می‌کنیم تا رویکردهای انتخاب ویژگی ترکیبی را ایجاد کنیم. در روش ترکیبی، EGA به چندین زیر مجموعه ویژگی در اندازه‌های مختلف اعمال می‌شود که براساس اهمیت آنها در ترتیب کمتری رتبه‌بندی می‌شوند و کاهش ابعاد انجام می‌شود. عملیات EGA برای مهمترین ویژگی‌هایی که دارای رتبه بالاتری هستند اعمال می‌شود.

تکنیک‌های طبقه‌بندی برای اندازه‌گیری قدرت روش‌های پیشنهادی و نشان دادن چگونگی تأثیر آنها بر عملکرد دسته‌بندی استفاده می‌شود. ما برای این منظور دو روش طبقه‌بندی را بررسی کردیم. اولین روش محبوب طبقه‌بندی متن، NB است که براساس اطلاعات احتمالی در مورد ویژگی های متن است. مورد دوم AC است که مبتنی بر ساخت قوانین طبقه بندی است که به الگوریتم های استخراج قاعده ارتباط بستگی دارد.

عملکرد تابع برازندگی عملکرد زیرمجموعه ویژگی را ارزیابی می‌کند و در انتخاب زیرمجموعه‌هایی که در نسل بعدی (جمعیت جدید) گنجانده می‌شود، نقش اصلی را بازی می‌کند. زیرمجموعه‌هایی که عملکرد بهتری دارند شانس بیشتری برای انتخاب و تولید مثل برای تشکیل جمعیت جدید دارند. در این کار ما عملکرد طبقه‌بندی یک طبقه‌بندی کننده NB (از نظر اندازه‌گیری F کل متوسط) را با اندازه زیر مجموعه ویژگی ترکیب می‌کنیم تا عملکرد تابع برازندگی را شناسایی کنیم و زیر مجموعه ویژگی را بر این اساس ارزیابی کنیم.

اثربخشی روش پیشنهادی با استفاده از بیز ساده و طبقه بندی انجمنی در سه مجموعه مختلف از مجموعه داده‌های متن عربی ارزیابی می شود. نتایج تجربی برتری EGA بر GA را نشان می دهد ، مقایسه GA با EGA نشان داد که مورد دوم از نظر کاهش ابعاد، زمان و عملکرد طبقه‌بندی نتایج بهتری را به دست می‌آورد. علاوه بر این، شش روش پیشنهادی ترکیبی FS متشکل از یک روش فیلتر و EGA در زیر مجموعه های مختلف ویژگی اعمال می‌شود. نتایج نشان داد که این روشهای ترکیبی نسبت به روش‌های فیلتر یکپارچه برای کاهش ابعاد موثرتر هستند زیرا در بیشتر شرایط بدون کاهش دقت طبقه‌بندی قادر به تولید نرخ کاهش بالاتر بودند.

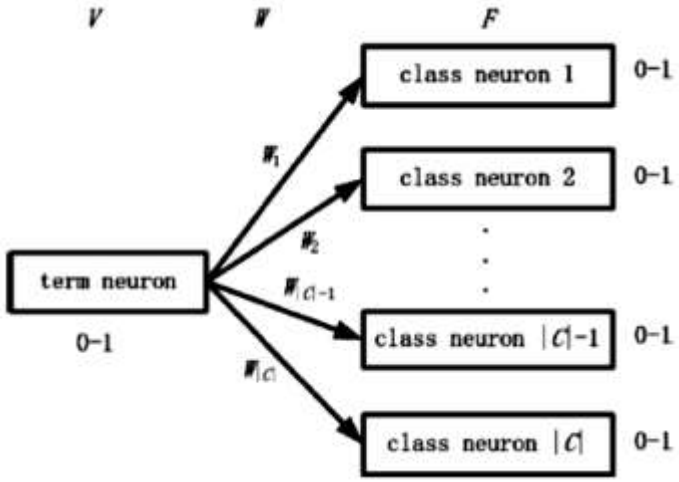
نتایج رویکردهای ترکیبی FS نشان داد که رویکردهای ترکیبی FS در کاهش ابعاد موثرتر بوده و می‌توانند در اکثر شرایط نسبت به روش تک فیلتر و GA به صورت جداگانه میزان کاهش بالاتر و دقت طبقه بندی بالاتری ایجاد کنند. همانطور که نتایج تجربی نشان می‌دهد، پتانسیل EGA پیشنهادی به صورت جداگانه و زمانی که در مرحله

دوم رویکردهای ترکیبی FS مورد استفاده قرار گرفت، اثبات شد. با این حال، استفاده از EGA در مرحله ساخت از نظر تأثیر آن در بهبود FS و ساده‌سازی فرایند دسته بندی که توسط الگوریتم دسته‌بندی دیگری انجام می‌شود، محدود است. بنابراین، می توان از EGA به عنوان الگوریتم طبقه‌بندی برای ایجاد یک طبقه‌بندی متن مبتنی بر قاعده استفاده کرد که چندین مزیت را به جای استفاده از آن به عنوان ابزار پیش پردازش برای الگوریتم دیگر، ترکیب می‌کند. علاوه بر این، افزایش GA به دو اپراتور GA (کراس اوور و جهش) اعمال شد. با این وجود می‌توان GA را با کار روی سایر عملیات مانند عملکرد تناسب اندام و طرح‌های انتخابی افزایش داد[۱۲].

چکیک و همکارش در سال۲۰۲۰ مقاله‌ای به منظور پردازش متون کوتاه مبتنی بر روش انتخاب ویژگی ارائه کردند در این مقاله بیان شده است، مشکل ابعاد بالا به دلیل تأثیر آن بر هزینه محاسباتی و دقت طبقه‌بندی کننده‌ها، یک مسئله مهم در طبقه‌بندی متن کوتاه است. همچنین، داده های متن کوتاه علاوه بر اینکه دارای ابعادی بالا هستند، دارای ساختاری ناقص، ناسازگار و پراکنده هستند. انتخاب ویژگی های مهمی که نمایش بهتری ارائه می دهند، یک راه حل برای مسئله ابعاد بالا است. در این مطالعه، یک روش انتخاب ویژگی مبتنی بر فیلتر، متناسب با انتخاب ویژگی ناهنجاری (PRFS) ، که از مجموعه ناهنجاری‌های موجود برای تشخیص یک تمایز منطقه‌ای مطابق با مقدار مجموعه استفاده می کند تا اسنادی را مشخص کند که دقیقاً متعلق به یک کلاس هستند یا احتمالاً متعلق به یک کلاس هستند. اسناد احتمالی متعلق به یک کلاس با ضریبی بنام a مقداردهی می‌شوند. روش پیشنهاد شده با پیشرفته‌ترین روش های انتخاب ویژگی فیلتر مانند شاخص جینی، کسب اطلاعات، انتخاب ویژگی‌های متمایز، نسبت حداکثر–حداقل ارائه شده‌ی اخیر و روش‌های اندازه‌گیری اختلاف نرمال مقایسه می‌شود. پراکندگی/ کم بودن یکی از مهمترین مشکلات در طبقه‌بندی متن کوتاه است. مسئله پراکندگی نوعی مشکل است که در متون کوتاه حاوی کلمات بسیار کمی وجود دارد. با استفاده از RST می‌توان یک راه‌حل موثر و موفق برای این

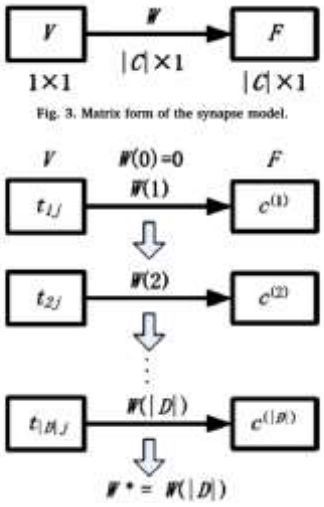
مشکل ارائه داد. RST الگوی پنهان را در داده‌ها کشف می‌کند و در معرض نمایش قرار دادن اطلاعات تکراری و بی‌معنی که باعث ناسازگاری در سیستم اطلاعات می‌شوند بسیار موفق است. در این مطالعه، یک روش انتخاب ویژگی جدید بر اساس فیلتر، یعنی انتخاب کننده ویژگی‌های ناهنجار(PRFS) ، با استفاده از RST پیشنهاد شده است. در حالت ایده آل، یک روش انتخاب ویژگی فیلتر انتظار می‌رود که امتیازات بالا را به ویژگی‌های با تبعیض بالا و امتیازات کم را به ویژگی‌های کمتر تمایز اختصاص دهد. مقایسه با استفاده از اندازه‌های مختلف ویژگی در چهار مجموعه داده متن کوتاه کوتاه با اندازه گیری موفقیت Macro-F1 انجام شده است. نتایج تجربی نشان داد که PRFS با توجه به سایر روش‌های انتخاب ویژگی از نظر Macro-F1 عملکرد بهتر یا رقابتی را ارائه می‌دهد. این مطالعه ممکن است یک مطالعه پیشگام در این زمینه تحقیقاتی باشد زیرا یک روش انتخاب ویژگی جدید برای طبقه‌بندی متن کوتاه با استفاده از یک نظریه مجموعه ناهنجار پیشنهاد می‌کند[۵].

هیونگ و همکارانش در سال ۲۰۱۹ به منظور دسته‌بندی و پردازش صحیح متون روشی مبتنی بر انتخاب ویژگی پیشنهاد کردند که در این روش بیان شده است، اسناد متنی معمولاً شامل اصطلاحات متفاوت با ابعاد بالا (بی ربط و پرنویز) هستند که منجر به هزینه‌های محاسباتی زیاد و عملکرد ضعیف یادگیری طبقه‌بندی متن می‌شوند. یکی از راه‌حل‌های موثر برای این مسئله انتخاب ویژگی است که هدف آن شناسایی اصطلاحات افتراقی از داده‌های متنی است. در این مقاله روشی با عنوان انتخاب ویژگی مبتنی بر قانون Hebb (HRFS) پیشنهاد شده است. در این مقاله یک روش انتخاب ویژگی جدید تحت نظارت HRFS پیشنهاد می‌شود. این روش بر اساس مدل سیناپس عصبی است و اصطلاحات افتراقی را با قانون Hebb مشخص می‌کند. روش انتخاب ویژگی پیشنهادی HRFS را بر اساس مدل سیناپس نشان می‌دهد. مدل را می‌توان با نمایش ماتریس ساده کرد.



شکل ۲- ۵) مدل روش پیشنهادی

روش پیشنهادی HRFS از مزایای اطلاعات کلاس با اندازه‌گیری همبستگی بین اصطلاحات و کلاسها برای تضمین عملکرد خوب طبقه‌بندی‌ها بهره می‌برد. قانون Hebb شبکه‌های عصبی را به صورت عملکرد ماتریس آموزش می‌دهد. HRFS را می‌توان از نظر عملکرد ماتریس طبق قانون Hebb به راحتی ساخت. HRFS امتیاز اصطلاح tj را با توجه به فرآیندهای نشان داده شده در شکل زیر نمایش می‌دهد[۷].



شکل ۲- ۶) فرایند روش پیشنهادی

روش پیشنهادی ما با هفت روش انتخاب ویژگی دیگر در مورد عملکرد انتخاب ویژگی در طبقه‌بندی متن مجموعه داده های معیار مقایسه شده است. نتایج تجربی نشان می‌دهد که HRFS برای دستیابی به عملکرد بهتر از روشهای مقایسه شده موثر است. HRFS می‌تواند اصطلاحات افتراقی را از نظر سیناپس بین سلولهای عصبی شناسایی کند. علاوه بر این، HRFS نیز کارآمد است زیرا می‌تواند از نظر عملکرد ماتریس توصیف شود تا از پیچیدگی انتخاب ویژگی کاسته شود. همچنین نتایج نشان می‌دهد روش ما می‌تواند با استفاده از ماتریس اجرا شود تا پیچیدگی برنامه‌نویسی و پیچیدگی زمان را کاهش دهد. به طور خلاصه، روش ما برای انتخاب ویژگی برای طبقه‌بندی متن موثر و کارآمد است[۷].

۲- ٤) خلاصه فصل

فصل حاضر به بررسی و مطالعه‌ی مفاهیم نظری و ادبیات پیشین و متونی که در راستای موضوع مورد بحث وجود داشت پرداخته است در فصل آتی رویکرد پیشنهادی ارائه خواهد شد. که با توجه به معایب روشهای قبلی سعی شده روشی بهبود یافته ارائه شود که سعی در رفع معایب روشهای پیشین و پیشنهاد روشی بهینه دارد.

فصل سوم

روش پیشنهادی

۳- ۱) مقدمه

به طور تقریبی بیش از ۹۰ درصد از دانش امروزی به صورت متن، مستندات و سایر صورت‌های رسانه‌ای نظیر صوت، تصویر و ویدیو نگهداری می‌شود. اگر از منظر علوم کامپیوتری به این مستندات نگاه شود اکثر آنها به نحوی غیرساخت یافته ذخیره شده‌اند. با این حال با رشد سریع اینترنت، طبیعی است که از متون نه به صورت کاغذی بلکه به صورت اطلاعات الکترونیکی و برخط استفاده شود. امروزه می‌توان کتاب‌ها و اخبار را به صورت الکترونیکی جستجو کرد. تقریباً همه شرکت‌ها، ادارات و سازمان‌ها دارای صفحات تار هستند و اطلاعات خود را در این صفحات ارائه می‌کنند. در نتیجه از طریق اینترنت بسیاری از اطلاعات در دسترس عموم قرار می‌گیرند.

یکی از مهم‌ترین قسمت‌های علم داده‌کاوی، متن‌کاوی است. متن کاوی هنر و علم استخراج اطلاعات و دانش از متن است. رده‌بندی و دسته‌بندی متون از مهمترین مسائل در متن‌کاوی هستند که هم به تنهایی دارای کاربرد می‌باشند و هم به عنوان بخشی از مسائل کاوش متن به کار می‌روند. به طور کلی همواره اطلاعات زیاد، نیازمند رده‌بندی می‌باشند. مهمترین موضوع در بررسی یک سری داده متنی، امر رده‌بندی و دسته‌بندی آنها محسوب می‌شود. رده‌بندی به این مفهوم است که یک سری داده فراهم شده را به چند رده از قبل تعریف شده اختصاص داد. زمانی که با مجموعه‌ای از اسناد متنی کار می‌شود بحث رده‌بندی به صورت طبقه‌بندی متون

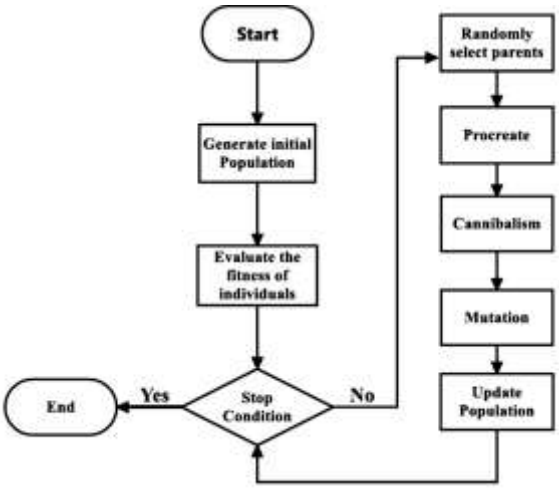
و یا رده‌بندی متون تعریف می‌شود. تعریف رده‌بندی متون به این صورت است که مجموعه‌ای از رده‌ها (موضوعات متن‌ها) به همراه یک سری سند متنی تهیه می‌شود و هدف یافتن رده و یا به عبارتی موضوع صحیح برای هر متن است. در حال حاضر اطلاعات متنی بسیار زیادی در اینترنت موجود است که این موجب شده تا دیگر قادر به رده‌بندی به صورت دستی نبود. بنابراین رده‌بندی اطلاعات متنی به صورت خودکار به رده‌های مختلف به امری ضروری تبدیل شده است.

۳- ۲) الگوریتم بیوه سیاه

الگوریتم‌های بهینه‌سازی الهام گرفته از طبیعت به دلیل سهولت و انعطاف‌پذیری می‌توانند مشکلات مختلف مهندسی و علمی را حل کنند. برای اعمال الگوریتم‌های فراابتکاری بر روی آنها نیازی به اصلاحات ساختاری در مشکلات بهینه‌سازی نیست. اخیراً، الگوریتم‌های فراابتکار در حال تبدیل شدن به روش‌های قدرتمندی برای حل مسائل NP هستند. الگوریتم بهینه‌سازی بیوه سیاه (BWO)، از رفتار منحصر به فرد جفت‌گیری عنکبوت‌های بیوه سیاه الهام گرفته شده است [۱۰].

شکل ۳- ۱) بیوه سیاه ماده در حین تخم‌گذاری

الگوریتم بیوه سیاه در ابتدا با جمعیت اولیه عنکبوت آغاز می‌شود، به طوری که هر عنکبوت یک راه‌حل بالقوه را نشان می‌دهد. عنکبوت‌های اولیه موجود، به صورت جفت، سعی در تولید مثل نسل جدید دارند. بیوه سیاه ماده عنکبوت‌های نر را در حین جفت‌گیری یا بعد از آن می‌خورند. سپس او اسپرم‌های ذخیره شده را در حفره‌های اسپرم خود حمل کرده و آنها را در کیسه‌های تخمک آزاد می‌کند. پس از گذشت ۱۱ روز عنکبوت‌ها از تخم‌ها خارج می‌شوند. آنها چندین روز تا یک هفته در شبکه مادر زندگی می‌کنند و در این مدت هم نوع خواری خواهر و برادر در جمعیت عنکبوت‌های تازه متولد شده، مشاهده می‌شود، سپس آنها آن محیط را ترک می‌کنند. روندنمای این الگوریتم به نحوی که در شکل ۳- ۲ نمایش داده شده است، می‌باشد.



شکل ۳- ۲) روندنمای الگوریتم بیوه سیاه

۳- ۲- ۱) جمعیت اولیه

برای حل یک مسئله بهینه‌سازی، مقادیر متغیرهای مسئله باید به عنوان یک ساختار مناسب برای حل مسئله فعلی تشکیل شوند. در اصطلاحات GA و PSO، این ساختار به ترتیب کروموزوم و موقعیت ذره نامیده می‌شود، اما درالگوریتم بهینه‌سازی بیوه سیاه (BWO) به آن بیوه گفته می‌شود. در الگوریتم بهینه‌سازی بیوه سیاه(BWO) ،

راه‌حل بالقوه هر مسئله به عنوان عنکبوت بیوه سیاه در نظر گرفته شده است. هر عنکبوت بیوه سیاه مقادیر متغیرهای مسئله را نشان می‌دهد. برای حل توابع معیار، ساختار باید به عنوان یک آرایه در نظر گرفته شود. مسئله بهینه‌سازی N بعدی، یک بیوه یک آرایه‌ی $1*N_{var}$ است که نشان‌دهنده‌ی حل مسئله است، که به صورت زیر تعریف می‌شود.

$$Widow=[x_1,x_2,\ldots.x_{Nvar}] \quad (۱-۳)$$

هر یک از مقادیر متغیر $(x_1,x_2,\ldots.x_{Nvar})$ عدد شناور است، تابع برازندگی الگوریتم بیوه سیاه به صورت زیر می‌باشد.

$$Fitness = f\,(widow) = f\,(x_1, x_2,\ldots., x_{Nvar}) \quad (۲-۳)$$

برای شروع الگوریتم بهینه‌سازی، یک ماتریس بیوه با اندازه‌ی $N_{pop} \times N_{var}$ با جمعیت اولیه عنکبوت‌ها تولید می‌شود. سپس جفت والدین به طور تصادفی انتخاب می‌شوند تا مرحله تولید مثل را با عمل جفت‌گیری انجام دهند، که در آن بیوه سیاه نر در طی جفت‌گیری یا بعد از آن توسط عنکبوت ماده از بین می‌رود.

۳- ۲- ۲) تولید مثل

از آنجایی که این جفت‌ها از یکدیگر مستقل هستند، به منظور تولید مثل نسل جدید، به طور موازی و همچنین در طبیعت، هر یک از جفت‌ها جدا از بقیه جفت می‌شوند. در دنیای واقعی، در هر جفت‌گیری تقریباً ۱۰۰۰ نوزاد عنکبوت تولید می‌شوند، اما سرانجام، برخی از نوزادان عنکبوتی زنده می‌مانند که قوی‌تر هستند. این روش شامل یک مرحله انحصاری، یعنی هم نوع خواری است. با توجه به این مرحله، گونه‌هایی با تناسب نامناسب از محدوده حذف می‌شوند، بنابراین منجر به همگرایی اولیه می‌شوند. برای تولید مثل، باید آرایه‌ای به نام آلفا ایجاد شود تا

زمانی که آرایه بیوه با اعداد تصادفی، سپس فرزندان با استفاده از α با معادله زیر تولید می‌شوند که در آن x_1 و x_2 والدین هستند، y_1 و y_2 فرزندان هستند.

$$\begin{Bmatrix} y_1 = \alpha \times x_1 + (1-\alpha) \times x_2 \\ y_2 = \alpha \times x_2 + (1-\alpha) \times x_1 \end{Bmatrix} \quad (۳-۳)$$

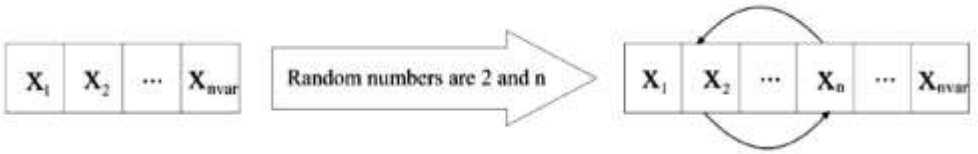
سرانجام، فرزندان و مادران به یک آرایه اضافه می‌شوند و براساس ارزش تناسب آنها مرتب می‌شوند، برخی از بهترین افراد به جمعیت تازه تولید شده اضافه می‌شوند و این مراحل برای همه جفت‌ها اعمال می‌شود[۱۰].

۳- ۲- ۳) همنوع خواری

در این قسمت سه نوع گونه خوری وجود دارد. اولین مورد همنوع خواری جنسی است که در آن بیوه سیاه ماده، در حین جفت‌گیری یا بعد از آن جفت نر خود را می‌خورد. در این روش، می‌توان عنکبوت نر و ماده را توسط تابع برازندگی تشخیص داد. نوع دیگر همنوع خواری خواهر و برادر است که در آن عنکبوت‌های قوی خواهر و برادر ضعیف تر خود را می‌خورند. در این روش، یک درجه بندی همنوع خواری (CR) تنظیم می‌شود که بر اساس آن تعداد بازماندگان را می‌توان تعیین کرد. در بعضی موارد، نوع سوم گونه خوری اغلب مشاهده می‌شود که در آن بچه عنکبوت مادر خود را می‌خورد. با توجه به مقدار برازندگی برای تعیین عنکبوت قوی یا ضعیف استفاده می‌شود[۱۰].

۳- ۲- ٤) جهش

در این مرحله، به طور تصادفی Mutepop تعداد افراد از جمعیت انتخاب می‌شود. همانطور که شکل ۳- ۳ نشان می‌دهد، هر یک از راه‌حل‌های انتخاب شده به طور تصادفی دو عنصر را در آرایه رد و بدل می‌کند Mutepop با نرخ جهش محاسبه می‌شود.



شکل ۳- ۳) جهش

۳- ۲- ۵) همگرایی

رسیدن به حالت همگرا و شرایط ثابت و بدون تغییر به منزله رسیدن به حالتی است که الگوریتم به همگرایی رسیده و می‌تواند متوقف شود. در این الگوریتم نیز مانند سایر الگوریتم‌های تکاملی، سه شرط توقف را می‌توان در نظر گرفت که شرحی که در ادامه ذکر شده است می‌باشند.

۱- تعداد تکرار از پیش تعریف شده

۲- رعایت عدم تغییر در مقدار تابع برازندگی بهترین بیوه برای چندین تکرار

۳- رسیدن به سطح مشخصی از دقت[۱۰].

الگوریتم BWO در ۵۱ عملکرد مختلف ارزیابی می‌شود تا کارایی آن در دستیابی به راه حل‌های بهینه برای مشکلات تأیید شود. نتایج به دست آمده نشان می دهد که الگوریتم پیشنهادی از جنبه‌های مختلف از جمله همگرایی اولیه و دستیابی به مقدار تابع شایستگی (برازندگی) در مقایسه با الگوریتم‌های دیگر مزایای بی شماری دارد. همچنین، این توانایی ارائه نتایج رقابتی و امیدوار کننده را دارد.

۳- ۳) الگوریتم نروفازی انفیس

یک واژه اختصاری است که از حروف اول ساخته شده است. در واقع انفیس یک سیستم استنتاج عصبی- فازی سازگار نوعی شبکه عصبی مصنوعی است که براساس سیستم فازی تاکاگی-سوگنو[1]می باشد. این شیوه در اوایل ۱۹۹۰ ایجاد شده است. از آنجایی که این سیستم، شبکه‌های عصبی و مفاهیم منطق فازی را یکی می‌کند، می-تواند از امکانات هر دو آنها در یک قاب بهره برد. سیستم سازگار[2]آن مطابق با مجموعه قوانین فازی اگر-آنگاه است که قابلیت یادگیری برای تقریب زدن توابع غیرخطی را دارد. از این رو، انفیس به عنوان یک برآورد جهانی[3]مطرح شده است.

۳- ۳- ۱) یادگیری مدل و استنتاج از طریق انفیس

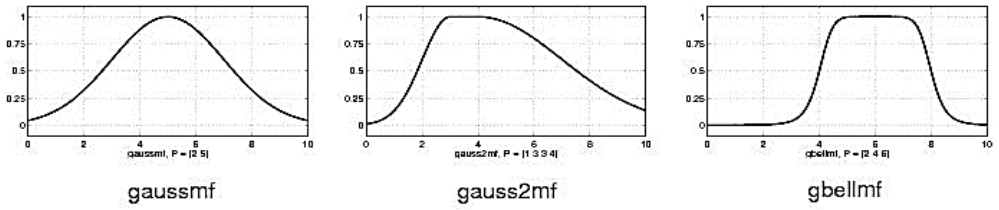
یادگیری عصبی- انطباقی دارای عملکردی مشابه با شبکه‌های عصبی می‌باشد. تکنیک‌های یادگیری عصبی-انطباقی روشی را برای ایجاد یک رویه مدل‌سازی فازی در راستای یادگیری اطلاعات از یک مجموعه داده فراهم می‌آورند. جعبه‌ی ابزار منطق فازی پارامترهای تابع عضویت را طوری محاسبه می‌کند که سیستم استنتاج فازی بر مجموعه داده‌های ورودی/ خروجی منطبق گردد[۲۳].

۳- ۳- ۲) شبکه‌های یادگیرنده تطابقی عصبی فازی انفیس

شبکه‌های یادگیرنده عصبی فازی در واقع نوعی از سیستم‌های استنتاج فازی هستند، با این تفاوت که از الگوریتم‌های توضیح داده شده در قسمت قبل می‌توان برای آموزش آن‌ها استفاده نمود تا با یادگیری داده‌های آموزشی، به مرور عملکرد آن‌ها بهبود پیدا کند. برای تشریح معماری داخلی این شبکه‌ها مثال ساده ارائه شده در

[1] Takagi–Sugeno
[2] inference
[3] universal estimator

شکل (۳-٤) را در نظر بگیرید در این شکل یک سیستم استنتاج فازی از نوع سوم که داری دو ورودی و یک خروجی می‌باشد نشان داده شده است.



شکل۳-۴) الف: gbellmf تابع عضویت ناقوس تعمیم یافته ب: gauss2mf تابع عضویت ترکیب دو منحنی گاوسی ج: gaussmf تابع عضویت منحنی ساده گاوسی

فرض کنید که این سیستم دارای دو قانون فازی اگر- آنگاه از نوع TSK به صورت زیر باشد.

قانون اول : اگر $x \in A_1$ باشد و $y \in B_1$، باشد آنگاه $f_1 = p_1x + q_1y + r_1$

قانون دوم : اگر $x \in A_2$ باشد و $y \in B_2$، باشد آنگاه $f_2 = p_2x + q_2y + r_2$

در شکل (۳-٤) شبکه انفیس معادل این سیستم نشان داده شده است. همان‌طور که از شکل مشخص است، این شبکه دارای ۵ لایه می‌باشد.

لایه ۱: هر گره i در این لایه که با علامت مستطیل نشان داده شده است دارای تابع درونی زیر می‌باشد:

$$O_i^1 = \mu A_i(x) \qquad (٤-٣)$$

که در آن X ورودی گره i و A مجموعه فازی مشخص کننده یک بر چسب زمانی (Linguistic Labels) مانند کوچک، متوسط، بلند و یا غیره می‌باشد. به عبارت دیگر O_i^1 تابع عضویت مجموعه A_i می‌باشد و به عنوان

خروجی درجه عضویت عضو X در این مجموعه را بر می‌گرداند معمولاً این تابع عضویت از نوع زنگی شکل و یا از نوع تابع گوس به صورت‌های زیر تعریف می‌گردد:

$$\mu_{Ai}(x) = \frac{1}{1 + \left[\left(\frac{x - c_i}{a_i}\right)^2\right]^{bi}} \qquad (۵-۳)$$

که دارای حداکثر یک و حداقل صفر می‌باشد و با تغییر پارامترهای a_i، b_i و c_i شکل و محل قرارگیری این توابع تغییر می‌کند. در واقع در این گره‌ها از هر نوع توابعی که در بازه مورد نظر مشتق پذیر باشد مانند توابع ذوزنقه‌ای و یا مثلثی می‌توان استفاده نمود. پارامترهای این لایه به پارامترهای مقدماتی معروف می‌باشد و چون خروجی شکل نسبت به این پارامترها خطی نیست باید از روشی مانند BP برای کشف مقادیر بهینه این پارامترها استفاده کرد.

لایه ۲: هر گره در این لایه که در شکل (۳-۹) با دایره و علامت Π نشان داده شده است، ورودی‌های خود را از کمان‌های نشان داده شده به آن وارد می‌شود در هم ضرب می‌شود و حاصل را به عنوان خروجی ارایه می‌نماید. به عبارت دیگر:

$$w_1 = \mu_{Ai}(x) \times \mu_{Bi}(y) \quad i = 1.2 \qquad (۶-۳)$$

خروجی این گره‌ها در واقع نشان دهنده میزان قوت هر کدام از گزاره‌های فازی می‌باشد و عمل ضرب انجام شده در آن‌ها نوعی عملگر «و یا And» در مجموعه‌های فازی می‌باشد، که به جای آن می‌توان از هر نوع عملگر دیگر مانند «حداقل یا Min» نیز استفاده نمود.

لایه ۳: در این لایه هر گره با علامت دایره و یک حرف بزرگ N نشان داده شده است. در واقع هر گره در این لایه نسبت قوت هر قانون فازی را نسبت به جمع قوای تمام قوانین محاسبه و به عنوان خروجی برمی گرداند. برای i امین گره این لایه می‌توان نوشت:

$$\overline{w_i} = \frac{w_i}{w_1 + w_2} \quad i = 1.2 \tag{۳-۷}$$

خروجی این لایه قوت‌های نرمال شده نامیده می‌شود.

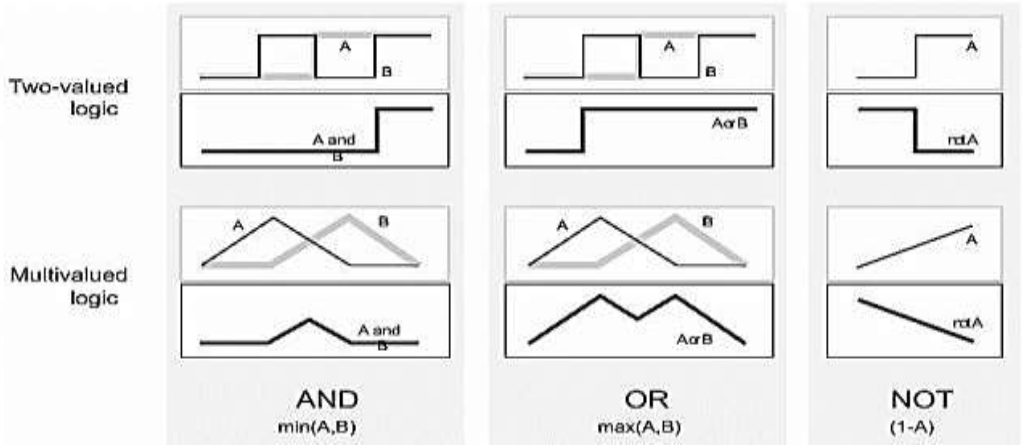
لایه ٤: هر گره در این لایه که با علامت مستطیل نشان داده شده است دارای تابع درونی زیر می‌باشد:

$$O_i^4 = \overline{w_i} f_i = \overline{w_i}(p_i x + q_i y + r_i) \tag{۳-۸}$$

که در آن $\overline{w_i}$ خروجی لایه قبلی و r_i و q_i و p_i پارامترهای قابل تعیین در این لایه می‌باشند. پارامترهای این لایه معمولاً پارامترهای نتیجه گیری نامیده می‌شوند چون خروجی شبکه نسبت به این پارامترها خطی می‌باشد، می‌توان از روش LSM برای تعیین پارامترهای این لایه استفاده نمود.

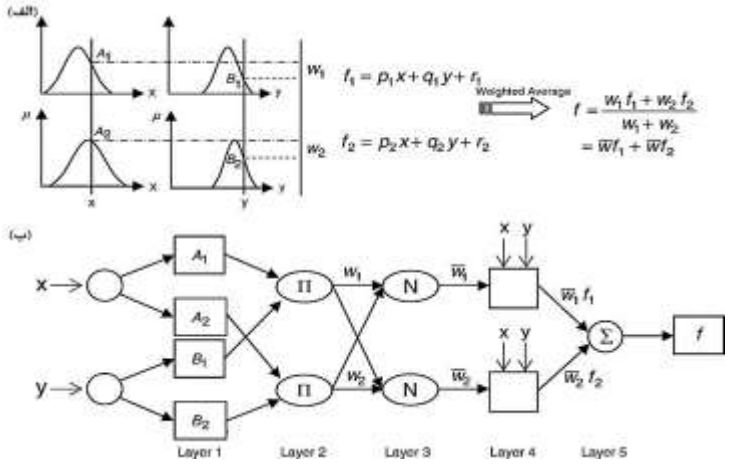
لایه ۵: تنها گره‌ای که در این لایه وجود دارد یک گره دایره‌ای است که با علامت Σ مشخص شده است و خروجی آن تمام ورودی‌هایش می‌باشد:

$$O_i^5 = overalloutput = \Sigma_i \overline{w_i} f_i = \frac{\Sigma_i w_i f_i}{\Sigma_i w_i} \tag{۳-۹}$$



شکل۳-۵) جداول درستی استانداردAND, OR, NOT دو مقداری و چند مقداری

بنابراین با توضیحات فوق شبکه نشان داده شده در شکل (۳-۶) یک شبکه یادگیرنده می‌باشد که خروجی آن معادل با یک شبکه استنتاج فازی از نوع سوم است.



شکل۳-۶) الف: سیستم استنتاج فازی از قوانین اگر-آنگاه به صورت TSK ب: شبکه انفیس با دو متغیر ورودی z معادل با سیستم ارائه شده در الف

در این نوع سیستم استنتاج از قوانین اگر-آنگاه فازی به صورت تاکاگی-سوگنو-کانت (TSK) استفاده می‌شود. خروجی هر کدام از گزاره‌ها یک ترکیب خطی از مقادیر ورودی به اضافه یک مقدار ثابت می‌باشد و خروجی نهایی میانگین وزنی مقادیر خروجی هر گزاره با توجه به قوت آن گزاره می‌باشد.

۳-۳-۳) معتبرسازی مدل با استفاده از مجموعه داده‌های آزمایشی[۱] و داده‌های وارسی[۲]

معتبرسازی مدل فرایندی است که طی آن بردارهای ورودی/ خروجی که در فرایند آموزش مدل شرکت نداشته‌اند، به مدل اعمال می‌شوند تا به این ترتیب میزان صحت عمل مدل مشخص شود. یکی از مشکلات فرایند معتبرسازی مدل، نحوه‌ی انتخاب داده‌های معتبرسازی است. زیرا این داده‌ها علاوه بر اینکه باید با داده‌های آموزشی دارای همخوانی باشند، در عین حال باید به حد کافی از آن‌ها متمایز باشند تا به این ترتیب فرایند معتبرسازی از اعتبار ساقط نشود. مجموعه داده‌های آزمایشی به شما قابلیت بررسی عمومیت سیستم استنتاج فازی را می‌دهد.

۳- ۳- ٤) محدودیت‌های انفیس

تابع انفیس فقط در مورد سیستم‌های سوگنو دارای خصوصیات زیر باشند، دارای کاربرد است:

- سیستم‌های سوگنویی که از درجه صفر و یا یک باشند.
- تنها دارای یک خروجی باشند که این خروجی باید با استفاده از روش غیر فازی سازی میانگین وزن دار شده[۳]فراهم شده باشد.
- نباید دارای اشتراک قوانین باشند. به این معنی که تعداد توابع عضویت خروجی باید با تعداد قوانین برابر باشد و قواعد متفاوت نمی‌توانند در مورد خروجی توابع عضویت با هم مشترک باشند.

۱Testing Dataset
۲Checking Dataset
3 Weighted Average Defuzzification

- تمام قواعد باید دارای وزن واحد (۱) باشند.

۳- ۳- ۵) ساختار و نحوه‌ی‌ایجاد مدل نروفازی

در مدل انفیس در لایه اول، مقادیر هر متغیر ورودی باید به چند کلاس برای ساختن قوانین مربوط، دسته‌بندی شود که قوانین فازی از ترکیب ۲ یا تعداد بیشتری توابع عضویت در لایه دوم ساخته می‌شوند برای طبقه‌بندی داده‌های ورودی و ساخت قوانین، روش‌های متعددی پیشنهاد شده که رایج‌ترین آن‌ها عبارتنداز: افراز شبکه‌ای[1] خوشه‌بندی فازی کاهشی[2] روش کلاسترینگ فازی (FCM)[3] در مدل فازی عصبی افراز شبکه‌ای توسط genfis1، خوشه‌بندی فازی کاهشی توسط genfis2 و روش کلاسترینگ فازی (FCM) با استفاده از genfis3 تولید می‌شود.

- افراز شبکه‌ای

زمانی که تعداد متغیرهای مورد استفاده کم است، افراز شبکه‌ای، یک روش مناسب برای طبقه‌بندی داده‌ها می‌باشد به عنوان مثال اگر n متغیر ورودی و برای هر متغیر m تابع عضویت داشته باشیم، تعداد قوانین برابر n^m خواهد بود که برآورد فراسنجه‌های این مدل با توجه به تعداد داده‌های موجود و محاسبه‌ی طولانی مدت میسر نمی‌باشد. لذا در این تحقیق از روش خوشه‌بندی فازی کاهشی و روشی کلاسترینگ FCM برای ساخت سیستم استنتاج فازی FIS استفاده می‌شود.

Genefis1: این نوع سیستم فقط قادر به تولید یک خروجی می‌باشد زیرا از سیستم نوع سوگنو استفاده می‌کند؛ و نوع تابع عضویت خروجی باید به صورت خطی یا ثابت و تعداد توابع عضویت خروجی برابر همان تعداد قوانین‌ایجاد شده توسط genfis1 می‌باشد. درgenfis1 پیش فرض تعداد توابع عضویت برابر num_MFS = ۲،

[1] Grid partition
[2] Subtractive Fuzzy clustering
[3] Fuzzy c-means clustering

پیش فرض نوع توابع عضویت ورودی ناقوس تعمیم یافته (gbellmf) و نوع توابع عضویت خروجی نیز خطی در نظر گرفته شده است و با استفاده از روش میانگین وزنی عمل غیر فازی انجام می‌پذیرد[۲٤].

- کلاسترینگ تفاضلی

روش خوشه‌بندی فازی کاهشی بر مبنای اندازه‌گیری تراکم نقاط موجود در فضای متغیرهای مورد استفاده انجام می‌شود. به این منظور، فضای متغیرهای موجود استاندارد سازی می‌شود به نحوی که ابعاد تمامی متغیر به بازه [۱،۰] منتقل می‌شود. در ابتدا هر یک از نقاط به عنوان نقطه‌ای که پتانسیل مرکزیت یک خوشه را دارد در نظر گرفته می‌شود، سپس شاخص تراکم (D_i) نقاط موجود در اطراف نقطه x_i محاسبه می‌شود. اگر یک نقطه دارای تعداد زیادی نقاط همبستگی باشد، شاخص تراکم بالایی خواهد داشت. پس از محاسبه‌ی تراکم برای هر نقطه، نقطه‌ای که دارای بالاترین شاخص تراکم می‌باشد به عنوان مرکز اولین خوشه انتخاب می‌شود. برای تعیین خوشه‌های بعدی، اگر نقطه‌ی x_{c1} با شاخص تراکم D_{c1} به عنوان مرکز انتخاب شده باشد، شاخص تراکم هر یکی از نقاط باقی مانده برای انتخاب مرکز خوشه بعدی تصحیح می‌شود.

Genfis2: در این روش خوشه‌بندی داده‌ها در سیستم تطابقی فازی توسعه یافته، توسط کلاسترینگ کاهشی صورت می‌پذیرد. زمانی که یک خروجی داریم می‌توان از Genfis2 برای‌ایجاد یک FIS اولیه برای آموزش انفیس استفاده کرد. در Genfis2 از تابع[1]کلاسترینگ تفاضلی برای تعیین تعداد قوانین فازی و توابع عضویت مرجع استفاده می‌نماید و سپس از تخمین خطی حداقل مربعات[2]برای تعیین معادلات متعاقب هر قانون استفاده می‌نماید؛

[1] Subclust
[2] Linear least square

پارامتر مهم و قابل تغییر در اینجا (radii) می‌باشد که یک بردار شعاعی است که محدوده نفوذ یک مرکز خوشه را در هر یک از ابعاد مشخص می‌نماید.

- C – Means فازی

c-means فازی یک تکنیک کلاسترینگ است که در آن هر نقطه با درجه خاصی (که با توجه به امتیاز عضویت تعیین می‌شود) به یک کلاستر متعلق می‌باشد. این تکنیک اولین بار توسط جیم بزدک[1] در سال ۱۹۸۱ در راستای بهبود کارایی روش‌های پیشین کلاسترینگ مطرح گشت. در این روش نحوه‌ی گروه‌بندی داده‌ها در فضای چند بعدی به تعداد معینی از کلاسترهای مختلف تشریح شده است[۲۶].

Genfis 3: در این روش ابتدا با استفاده از تابع FCM برای استخراج توابع عضویت و تعیین تعداد قوانین استفاده می‌شود. در این روش‌ایجاد یک ساختار FIS از نوع ممدانی یا سوگنو داده شده است؛ و به شما اجازه مشخص کردن تعداد خوشه‌ها[2]توسط FCM داده می‌شود. تابع عضویت ورودی پیش فرض، از نوع گوسی gaussmf و تابع عضویت خروجی پیش فرض از نوع خطی است.

۳-٤) طراحی روش پیشنهادی

متن‌کاوی می‌تواند همچنین مشابه داده‌کاوی به عنوان کاربرد الگوریتم‌ها و روش‌ها در زمینه یادگیری ماشین و آمار با هدف یافتن الگوهای مفید تعریف شود. به همین منظور لازم است متن پیش پردازش شود. بسیاری از نویسنده‌ها روش‌های استخراج اطلاعات و پردازش زبان طبیعی را به منظور استخراج داده از متن استفاده می‌کنند. متن کاوی، کشف به وسیله اطلاعات ناشناخته قبلی و استخراج خودکار اطلاعات از منابع نوشته شده مختلف

[1] Jim Bezdek
[2] n–cluster

است. برای طبقه‌بندی متن‌ها قبل از استفاده از هرگونه روش، تبدیل متن‌ها به فرم نمایش مناسب ضروری است. پس از اینکه متن‌ها به فرم نمایش مناسبی تبدیل شدند، الگوریتم‌های انتخاب ویژگی روی آنها اعمال می‌شوند. پس از اعمال انتخاب ویژگی، متون با استفاده از ویژگی‌های انتخاب شده طبقه‌بندی می‌شوند. به منظور اعمال روش‌های طبقه‌بندی متون و نیز اعمال روش‌های استخراج ویژگی‌های متون، بایستی ساختاری مناسب جهت نمایش سندها در نظر گرفته شود. ساده‌ترین و عمومی‌ترین روش نمایش متون، ایجاد یک فضای ویژگی از تمام کلمات بوده که در متن‌ها وجود دارد. در این فضای ویژگی‌ها، پس از حذف کلمات خاص و گاهاً ریشه‌یابی، لیستی از تمامی کلمات در متون ایجاد شده و هر سند با توجه به اینکه در بردارنده چه کلماتی از لیست بوده و با چه وزنی این کلمات در سند اتفاق افتاده‌اند، به روش‌های مختلف نمایش داده می‌شود.

دسته‌بندی متن اغلب به صورت انتساب یک یا بیشتر از یک دسته به متن بر اساس محتوای آن تعریف می‌شود. طبقه‌بندی متن از دو نظر قابل بررسی است: اول از دیدگاه بازیابی اطلاعات؛ چون با رشد سریع منابع متنی نیاز به میزان زیادی از پردازش بیشتر می‌شود. دسته‌بندی متن می‌تواند به عنوان یک طبقه‌بندی کننده داده‌های متنی و یا به عنوان یک مرحله در پیش‌پردازش برای بازیابی اطلاعات، با انجام اعمالی مانند فیلتر کردن اسناد یا استخراج اطلاعات استفاده می‌شود. دوم از دیدگاه یادگیری ماشین است، دسته‌بندی متن در زمینه یادگیری ماشین کاربرد زیادی دارد. در یادگیری ماشین، دسته‌بند خودکاری به نام طبقه‌بندی کننده تولید می‌شود، این طبقه‌بندی کننده به وسیله استنتاج از نمونه‌هایی که قبلاً دسته‌بندی شده‌اند تولید می‌شود.

انتخاب ویژگی فرآیندی است که زیرمجموعه‌ای از ویژگی‌های مجموعه متون را بر اساس یک معیار انتخاب می‌کند. انتخاب ویژگی باعث می‌شود فضای با ابعاد زیاد داده‌های ورودی به فضایی با ابعاد کوچکتر تبدیل

شود. البته در صورتی که این انتخاب به درستی صورت نگیرد می‌تواند موجب کاهش دقت رده‌بندی متون شود.

۳- ۵) پیاده‌سازی روش پیشنهادی

انتخاب ویژگی فرآیندی است که زیرمجموعه‌ای از ویژگی‌ها را بر اساس یک معیار به عنوان ویژگی‌های اصلی انتخاب می‌کند. ویژگی‌های انتخاب شده دارای معنا و مفهوم مشخص می‌باشند و فرآیند یادگیری را ساده‌تر و سریع‌تر می‌کنند. این مرحله بعد از مرحله پیش پردازش انجام می‌گیرد و هم در رده‌بندی متون و هم در خوشه‌بندی متون نیاز می‌باشد و می‌تواند به صورت نظارتی و یا غیرنظارتی باشد. در سیستم‌های رده‌بندی پیاده‌سازی شده هر کلمه در متن به عنوان یک ویژگی در نظر گرفته می‌شود. یک سیستم رده‌بندی متن می‌توان از سه قسمت استخراج ویژگی، انتخاب ویژگی و آموزش رده‌بند تشکیل شود.

روش‌های مختلفی برای نمایش یک متن وجود دارد. در الگوریتم‌های پیشنهادی برای نمایش متن از روش بسامد کلمه استفاده شده است. در این روش تنها از فاکتور محلی وزن‌دهی کلمه برای نمایش سند استفاده می‌شود. در روش بسامد کلمه از رابطه‌ی زیر به منظور محاسبه وزن کلمه iام استفاده می‌شود.

$$w_i = tf_i \qquad (۱۰-۳)$$

در رابطه‌ی فوق W_i وزن، tf_i تعداد تکرار کلمه iام، در سند متنی مورد نظر است. بردار ویژگی با استفاده از روش بسامد کلمه نشان داده شده است. پس از مشخص شدن بردار ویژگی کلمات، مراحل پیش پردازش بر روی آن اجرا خواهد شد. در مرحله انتخاب ویژگی برای رده بندی متون از الگوریتم بیوه سیاه استفاده شده است.

الگوریتم بیوه سیاه مورد استفاده در این بخش یکی از اولین روش‌های فرامکاشفه‌ای است که استراتژی بسیار صریحی برای فرار از بهینه‌های محلی دارد. ایده اصلی آن اجازه حرکت به سمت جواب‌های بد برای نجات از بهینه محلی است. این الگوریتم با تکرار مراحل، بردار اولیه را بهبود می‌بخشد. بردار اولیه برداری است که شامل تمام کلمات در مجموعه متون می‌باشد. این بردار با استفاده از استخراج ویژگی از تمام سندهای متنی به دست آمده است. ایده‌ای که در این الگوریتم پیشنهادی به کار رفته است بکارگیری الگوریتم بیوه سیاه پیاده‌سازی شده برای کاربرد رده‌بندی متون و تلفیق این الگوریتم با روش بسامد سند است. الگوریتم بیوه سیاه جستجویی فرامکاشفه‌ای و تصادفی است که در مسائل بهینه‌سازی کاربرد فراوان دارد و بسامد سند روشی معمول و نسبتاً کارا در امر انتخاب ویژگی برای رده‌بندی متون است.

در الگوریتم پیشنهادی، بردار ویژگی اولیه شامل تمام کلمات در مجموعه متون می‌باشد. این بردار ویژگی همان راه‌حل مسئله است که الگوریتم در جهت بهبود آن تلاش می‌کند. الگوریتم پیشنهادی به این صورت شروع به کار می‌کند که بردار ویژگی اولیه را به بردار ویژگی نهایی که شامل کلمات انتخاب شده خواهد بود نسبت می‌دهد.

الگوریتم بیوه سیاه نیاز به تابعی برای ارزیابی و تشخیص میزان کارایی هر راه‌حل دارد. در راستای کارایی محاسباتی کل استفاده از توابع ارزیابی کارا و مناسب بسیار مهم می‌باشد، مخصوصاً زمانی که نیاز به توابعی داریم که پیچیدگی محاسباتی کمتری داشته باشند. در الگوریتم پیشنهادی تابع ارزیابی به صورت زیر فراخوانی می‌شود.

Evaluate(offspring vector) (۱۱ -۳)

تابع ارزیابی که در اینجا به کار گرفته شده است مبتنی بر روش بسامد سند می‌باشد. در امر رده‌بندی متون پس از اینکه الگوریتم بیوه سیاه در هر مرحله راه‌حلی جدید ارائه می‌دهد نیاز به بررسی آن راه‌حل داریم. به عبارتی

دقیق‌تر خروجی هر مرحله الگوریتم بیوه سیاه برداری از ویژگی‌ها می‌باشد، بنابراین باید الگوریتم یادگیری را با این بردار جدید آموزش داده و نتیجه آموزش را بر روی داده‌های آزمون بررسی کنیم. اگر بخواهیم برای هر مرحله از اجرای الگوریتم بیوه سیاه این فرآیند را تکرار کنیم، الگوریتم بسیار طولانی، هزینه بر و زمان گیر خواهد شد. برای محاسبه برازندگی کلی یک بردار ویژگی میانگین مقادیر مربوط به بسامد تمام ویژگی‌ها را محاسبه می‌کند.

در الگوریتم پیشنهادی از روش بسامد سند برای ارزیابی ویژگیهای جدید به دست آمده استفاده خواهیم کرد. روش بسامد سند سادهترین تکنیک برای ارزیابی یک مجموعه واژگان است. این معیار برای مجموعه داده های بزرگتر با پیچیدگی محاسباتی تقریباً خطی نسبت به داده های آموزشی، به راحتی قابل تعمیم است. این معیار برای هر کلمه برابر با نسبت تعداد سندهای متنی شامل کلمه مورد نظر به کل تعداد سندها می‌باشد.

برای محاسبه برازندگی کلی یک بردار ویژگی میانگین مقادیر مربوط به بسامد تمام ویژگی‌ها را محاسبه می‌کند. در هر تکرار، الگوریتم یک راه‌حل همسایه را تولید می‌کند و سپس تفاوت ارزش بین راه‌حل فعلی و راه‌حل جدید را بررسی خواهد کرد. اگر راه حل کاندید بهتر باشد آن را می‌پذیرد. در غیر این صورت راه‌حل با احتمالی که به تفاوت ارزش و دما بستگی دارد مورد قبول قرار می‌گیرد. سپس دما برای اجرای تکرار بعدی کاهش می‌یابد. در الگوریتم پیشنهادی، راه‌حل همسایه با حذف ویژگی‌ها به دست می‌آید. در کاربرد رده‌بندی متون تعداد ویژگی‌ها بسیار زیاد می‌باشد، بنابراین در این مرحله از الگوریتم به صورت تصادفی و با استفاده از بررسی معیار بسامد سند یک ویژگی حذف و یا نگهداری خواهد شد. برای ایجاد همسایه جدید بایستی احتمال حذف هر کدام از ویژگی‌ها بررسی شوند. در هر مرحله راه‌حل همسایه ممکن است شامل همان بردار ویژگی قبلی با تغییرات اندک باشد و یا اینکه بسیاری از ویژگی‌ها در بردار قبلی حذف شده باشند.

۳- ۵- ۱) پیش پردازش

اولین مرحله در دسته‌بندی متن تبدیل اسناد به صورت رشته‌ای از کاراکترها با فرمت‌های مختلف می‌باشد که برای روش‌های یادگیری و طبقه‌بندی نمایش داده می‌شود. همواره بهتر است در بازیابی اطلاعات ریشه کلمه را پیدا کرده تا بتوان آن کلمه را به صورت واحد در اسناد به کار برد و این کلمه‌ی واحد، منجر به نمایش مقدار ویژگی در متن می‌شود. در مرحله‌ی بعد توکن بندی صورت می‌پذیرد که این فرآیند به این صورت است که جریان متن به کلمه‌ها، عبارات، نشانه‌ها یا عناصر معنی‌دار شکسته می‌شود که به هر کدام از آن‌ها توکن گفته شده و به این فرآیند توکن‌بندی Tokenization می‌گویند.

در مرحله ریشه‌یابی، ریشه کلمه‌ها به فرم اصلی در می‌آید و هرگونه پیشوند و پسوندی از ابتدا و انتهای آن حذف می‌شود. حذف کلمات ایست یا توقف کلمات ایست به کلمه‌هایی گفته می‌شود که حاوی هیچ‌گونه معنی مفیدی نیستند مانند حروف ربط و حروف اضافه. لیست کلمات توقف برای اکثر زبان ها باید استخراج شود که با استفاده از این لیست می‌توان دید که اگر این کلمه‌ها در داخل اسناد متنی وجود داشته باشد از اسناد حذف می‌شود و اگر کلمه در لیست نبود حذف نمی‌شود و با این کار از تعداد ویژگی‌ها کم می‌شود. نمونه‌ای از کلمات ایست می‌توان به کلمه‌های و، در، به، که، از، این، را، است، با، برای و غیره اشاره کرد. عبارات تاکید یا نشانه‌گذاری هم شامل ” “ ? : ,] [: / ... . ! } { *) (# _ – می‌باشد. برای نمایش متون می‌توان از روش فضای برداری استفاده نمود. با این نمایش می‌توان ویژگی‌ها را از داخل اسناد استخراج کرد. در مدل فضای برداری، سندها به وسیله برداری از کلمه‌ها نمایش داده می‌شوند و مجموعه سندها به وسیله ماتریس کلمه در سندA ، نمایش داده می‌شوند.

در مرحله استخراج کلمات و ویژگی‌ها مجموعه‌ای از جداکننده‌ها که مهمترین آنها فاصله می‌باشد، برای یافتن کلمات استفاده شده است. در بخش حذف کلمات زائد نیاز به وجود اطلاعات زبان شناسی زبان فارسی می‌باشد.

در بین تمام کلمات در مجموعه متون، بعضی کلمات به دلیل تکرار بسیار زیاد، ویژگی مفید محسوب نمی‌شوند که زائد هستند. کلمات زائد در قالب یک پرونده متن توسط کارشناس زبان‌شناسی جمع‌آوری و فراهم شده‌اند. پس از استخراج کلمات از متن، کلمات زائد از داخل متن حذف می‌شوند. این کلمات مانند "از"، "در"، "را"، "به" هستند و در رده‌بندی متون ارزشی ندارند. باید به این نکته توجه داشت که بسیاری از این کلمات زائد حروف اضافه و قیدها هستند و به طور خودکار توسط روش‌های انتخاب ویژگی معیار ارزیابی بالایی را کسب نمی‌کنند و کنار گذاشته می‌شوند اما حذف آنها در مرحله پیش پردازش سرعت سیستم رده‌بندی را بالا می‌برد، علاوه بر اینکه در روش‌های انتخاب ویژگی هیچ تضمینی وجود ندارد که تمامی آنها از بردار ویژگی کنار گذاشته شوند. همچنین در این مرحله پس از حذف کلمات زائد، کلماتی را که تعداد تکرارشان در کل اسناد کمتر از چهار بود، حذف گردید. کلمات "خوشخو"، "خیابانی"، "هفدهم"، "خلاقیت"، "بیژن" و "تعطیلی" نمونه‌هایی از کلمات با تعداد تکرار کمتر از چهار هستند. در این مرحله کلماتی را که بیش از پانزده حرف داشتند، نیز حذف شدند. این امر بدین دلیل است که معمولاً این کلمات ارزش اطلاعاتی کمی داشته و داده نویز محسوب می‌شوند. مثال‌هایی از کلمات با بیش از پانزده حرف "غیرانگلیسی زبان"، ""،"berkshirefilmfestival"حادثه گروگان-گیری" و "نانوکامپوزیت‌های" هستند

۳- ۵- ۲) استخراج ویژگی

با پیشرفت علم حجم اسناد متنی موجود بر روی رسانه‌های دیجیتال و اینترنت، افزایش یافته است و این موضوع ضرورت استفاده از سیستم‌های خودکار تشخیص و دسته‌بندی متن را بیشتر پررنگ می‌کند. روش‌های دسته‌بندی متن جزو روش‌های یادگیری ماشین هستند و استخراج و انتخاب ویژگی مرحله‌ی بسیار مهم در رویه‌ی دسته‌بندی متون به شمار می‌رود، زیرا در این مرحله واژه‌های کلیدی انتخاب می‌شوند تا به‌عنوان بهترین نمایش‌دهنده برای

سند متنی مورد استفاده قرار بگیرند. هدف روش‌های انتخاب ویژگی به دست آوردن یک مجموعه‌ی کوچک‌تر از ویژگی‌های موجود در سند می‌باشند که به طرز مؤثری محتوای سند را بیان می‌کند. الگوریتم‌های مختلفی برای دسته‌بندی متون وجود دارد. مشکلی که در دسته‌بندی متن وجود دارد، حجم زیاد ویژگی‌ها است که باعث کاهش دقت نتایج دسته‌بندی می‌شود. برای انتخاب و برای حل این مشکل و کاهش ابعاد ویژگی‌ها از متدهای انتخاب ویژگی استفاده می‌کنند.

در واقع استخراج ویژگی فرایندی است که در آن با انجام عملیاتی بر روی داده‌ها، ویژگی‌های بارز و تعیین‌کننده آن مشخص می‌شود. هدف استخراج ویژگی این است که داده‌های خام به شکل قابل استفاده‌تری برای پردازش‌های آماری بعدی درآیند. روش‌های مختلف استخراج ویژگی بنا به فلسفه پشت سرشان ممکن است یک یا چند کار زیر را انجام دهند.

- حذف نویز داده‌ها
- جداسازی اجزای مستقل داده‌ها
- کاهش ابعاد برای تولید بازنمایی مختصرتر
- افزایش بعد برای تولید بازنمایی جدایی‌پذیرتر

استخراج ویژگی، نقش بسیار مهمی در میزان دقت طبقه‌بندی دارد. اگر ویژگی‌ها با روش مناسبی استخراج نشده باشند، ممکن است طبقه‌بند گمراه شده و به پاسخ مناسبی دست پیدا نکند. که در این روش از ویژگی‌های متعددی استفاده می‌شود، هدف استفاده از این ویژگی‌های متعدد، یافتن مناسب‌ترین ویژگی‌ها برای استفاده در طبقه‌بندی

تومورهای مغزی می‌باشد. در این پژوهش از فیلتر گابور به منظور استخراج ویژگی استفاده شده است. گابور یک تابع گوسی هسته‌ای است که با یک موج هموار سینوسی فرمول شده است.

$$G(x.y) = \frac{f^2}{\pi\gamma\eta}\exp(\frac{-\acute{x}^2 + \gamma^2\acute{y}^2}{2\sigma^2})\exp(j^2\pi x \acute{+} \varphi) \quad (۱۲-۳)$$

$$\acute{x} = x\cos\theta + y\sin\theta \quad , \quad \acute{y} = -x\sin\theta + y\cos\theta \quad (۱۳-۳)$$

در روابط فوق پارامتر f فرکانس عامل سینوسی، θ نمایش دهنده‌ی جهت نرمال تابع گابور، φ تأثیر فاز، σ انحراف استاندارد و γ نرخ فضایی است.

فیلتر گابور ماهیتی مختلط دارد، به این معنی که دارای یک قسمت حقیقی و یک قسمت موهومی (اندازه و زاویه) می‌باشد. این فیلتر دارای دو پارامتر اصلی فرکانس و زاویه یا جهت فیلتر است. با تغییر این دو پارامتر، شکل فیلتر تغییر می‌کند. به طور معمول فیلتر گابور در ۱ اندازه و ۸ جهت مختلف محاسبه می‌شود. بنابراین مجموعاً ٤٠ فیلتر به صورت نشان داده شده است.

- **تابع گوسی**

تابع گاوسی، تابعی است به شکل نمایی که به صورت زیر تعریف می‌شود.

$$f(x) = ae^{-\frac{(x-b)^2}{2c^2}} \quad (۱٤-۳)$$

که در آن a ، b و c ضرایب ثابت حقیقی و e عدد اویلر است. شکل این تابع زنگوله‌ای متقارن است که به سرعت به صفر نزول می‌کند. ثابت a تعیین کننده‌ی ارتفاع قله‌ی منحنی، b تعیین کننده‌ی محل مرکز قله و c انحراف معیار، تعیین کننده‌ی میزان کشیدگی یا پهن شدگی زنگوله است. تابع گوسی در علوم احتمال، آمار و هوش مصنوعی و به ویژه در توزیع نرمال، استفاده ی فراوان دارد.

- **کاهش ابعاد بردار ویژگی**

استخراج ویژگی یکی از مولفه‌های بسیار مهم در مبحث شناسایی الگو می‌باشد. در دنیای واقعی اغلب با هزاران و یا حتی صدها هزار ویژگی سروکار داریم که به دلایل زیادی مثل زمان بر بودن و هزینه‌بر بودن کار در فضا با ابعاد بالا می‌تواند نامطلوب باشد. کاهش ابعاد در زمینه‌هایی که با تعداد زیادی متغیر سروکار دارند مانند پردازش سیگنال، تشخیص گفتار، نورفورماتیک و بیوانفورماتیک معمول است. در کل همه ویژگی‌ها با هم برابر نیستند و هدف از استخراج ویژگی برای کاهش ابعاد این است که ویژگی‌ها جوری تغییر کنند که در انتها به مجموعه ی جدید است برسیم در حالی که بسیاری از اطلاعات اساسی را حفظ می‌کنند.

ابعاد ویژگی‌های استخراج شده با استفاده از روش ایجاد همسایگی و حرکت به همسایگی جدید کاهش می‌یابد. این روش، روشی شناخته شده برای استخراج ویژگی و کاهش بعد به شمار می‌رود که در زمینه‌های مختلف کاربرد دارد و یک تکنیک رایج برای کشف الگوهای موجود در داده‌های با ابعاد بالا است.

۳- ۵- ۳) نرمال‌سازی داده‌ها

بعد از استخراج ویژگی‌ها از داده‌های متنی، به دلیل اینکه هریک از ویژگی‌ها در محدوده‌ی متفاوت از یکدیگر قرار دارند نرمال‌سازی داده‌ها صورت می‌گیرد. بعد از استخراج ویژگی‌ها از متون مورد بررسی، به دلیل اینکه هریک از ویژگی‌ها در محدوده‌ی متفاوت از یکدیگر قرار دارند نرمال‌سازی داده‌ها صورت می‌گیرد. نرمال‌سازی در آمار معانی متفاوتی دارد که ساده‌ترین کاربرد آن، نرمال‌سازی داده‌ها یا نرمال‌سازی متغیرها است و عبارت است از روشی که داده‌ها را در زمانی که در یک دامنه نیستند، در دامنه مشابه قرار می‌دهد. به بیان دیگر ممکن است یک داده کاو با موقعیت‌هایی مواجه گردد که ویژگی‌های داده، شامل مقادیری باشند که در محدوده یا دامنه متفاوتی قرار داشته باشند. این ویژگی‌های با مقادیر بزرگ ممکن است اثر بسیار زیادتری در تابع هزینه نسبت به ویژگی‌های با مقادیر کم داشته باشند. این مشکل با نرمال‌سازی ویژگی‌ها طوری که مقادیرشان در دامنه‌های مشابه

قرار گیرند برطرف خواهد شد. در ساخت مدل از روی داده‌ها پیش از شروع آموزش مدل‌ها، داده‌ها را به بزرگترین مقدار متناظرشان تقسیم می‌کنند تا به مقدارهای بین صفر و یک مقیاس شوند. این کار باعث می‌شود که اثر مقیاس واقعی کمینه شود و همه ورودی‌ها تقریباً در یک دامنه باشند. بنابراین بدلیل اینکه ویژگی‌های به دست آمده در محدوده‌های متفاوتی قرار دارند، جهت هم مقیاس کردن ویژگی‌ها و سهولت در پردازش‌های بعدی، باید آنها را نرمال‌سازی نمود. نرمال‌سازی ویژگی‌ها در محدوده‌ی [0,1] به صورت زیر نشان داده شده است.

$$\text{Normalizedx(i)} = \frac{\text{x(i)} - \min(\text{x})}{\max(\text{x}) - \min(\text{x})} \qquad (۳-۱۵)$$

داده‌های استخراج شده و کاهش داده شده به عنوان ورودی در مرحله‌ی بعدی به سیستم نروفازی داده می‌شود

داده‌های موجود در دو دسته train و test کلاس‌بندی می‌شوند. در پژوهش حاضر تابع genfis تعداد ۲۰ عدد متد نروفازی‌ایجاد می‌کند همینطور تابع عضویت در این روش ۵ می‌باشد. تابع عضویت (MF) منحنی‌ای است که نحوه‌ی نگاشت هر نقطه از فضای ورودی را به یک مقدار عضویت (درجه‌ی عضویت) بین ۰ و ۱ تعریف می‌کند. تنها شرطی که تابع عضویت باید ارضا کند این است که خروجی آن باید بین ۰ و ۱ باشد. براساس تعداد و نوع تابع محدودیت ایجاد شده و بر داده‌های موجود اعمال می‌شود داده‌ها بر اساس نوع فازی آموزش داده شده و مورد ارزیابی قرار می‌گیرند کلاس‌بندی ایجاد شده برای افزایش دقت و اطمینان از تشخیص درست متن مورد تست قرار گرفته است. بر اساس ویژگی‌های استخراج شده میزان صحت داده‌های بدست آمده مورد ارزیابی قرار می‌گیرد.

۳-٦) پارامترهای مورد ارزیابی

معیارهای ارزیابی عملکرد الگوریتم‌های یادگیری معمولاً شامل درستی[1]، دقت[2] صحت[3]می‌باشند. با توجه به نتایج این معیارها می‌توان در مورد عملکرد الگوریتم‌های ارائه شده بحث نمود. برای توضیح این چهار معیار بایستی ابتدا اجزای مورد نیاز برای محاسبه‌ی آنها بررسی شوند. این اجزا شامل چهار مورد مثبت درست[4]، مثبت نادرست[5]، منفی درست[6]و منفی نادرست[7]می‌باشند.

مثبت درست: نمونه‌های بیماری که درست بیمار تشخیص داده شده‌اند.

مثبت نادرست: نمونه‌های سالمی که نادرست بیمار تشخیص داده شده‌اند.

منفی درست: نمونه‌های سالمی که درست سالم تشخیص داده شده‌اند.

نرخ درستی رده بندی متون ورودی یکی از مسائل مهمی می‌باشد که در روش‌های مختلف به عنوان یک پارامتر اصلی برای تعیین دقت تشخیص و درستی کار مورد نظر قرار می‌گیرد. در حالت کلی برای روش‌های یادگیری با نظارت به طور معمول ۸۰٪ از داده‌ها در یادگیری ماشین به داده آموزش Train و ۲۰٪ برای داده تست تعلق می‌گیرد.

$$precision=TP / (TP + FP) \quad (۳-۱٦)$$

$$Accuracy=Tp+TN/P+N \quad (۳-۱۷)$$

[1] accuracy
[2] precision
[3] recall
[4] True Positive (TP)
[5] False Positive (FP)
[6] True Negative (TN)
[7] False Negative (FN)

$$recall = TP / (TP + FN) \quad (۱۸-۳)$$

$$specificity = TN / (FP + TN) \quad (۱۹-۳)$$

در روابط فوق N + P نشان‌دهنده‌ی تمام نمونه‌ها می‌باشد.

۳- ۷) خلاصه فصل

در این فصل در ابتدا به بررسی الگوریتم بیوه سیاه و سپس روش نروفازی که در استخراج و رده‌بندی متون استفاده شده اند پرداخته شد، در ادامه‌ی فصل به بیان روش پیشنهادی و کارکرد سیستم طراحی شده پرداخته شده است.

فصل چهارم

نتایج

٤- ١) مقدمه

امروزه به طور تقریبی بیش از ۹۰ درصد از دانش امروزی به صورت متن، مستندات و سایر صورت‌های رسانه‌ای نظیر صوت، تصویر و ویدیو نگهداری می شود. اگر از منظر علوم کامپیوتری به این مستندات نگاه کنیم اکثر آنها به نحوی غیر ساخت یافته ذخیره شده‌اند. با این حال با رشد سریع اینترنت، طبیعی است که از متون نه به صورت کاغذی بلکه به صورت اطلاعات الکترونیکی و برخط استفاده شود. امروزه می‌توان کتابها و اخبار را به صورت الکترونیکی جستجو کرد. تقریباً همه شرکت‌ها، ادارات و سازمان‌ها دارای صفحات تار هستند و اطلاعات خود را در این صفحات ارائه می‌کنند. در نتیجه از طریق اینترنت بسیاری از اطلاعات در دسترس عموم قرار می‌گیرند. شرح میزان تاثیر اینترنت بر گسترش اطلاعات کار بسیار دشواری است. اشخاص، شرکت‌ها و سازمان‌ها دارای صفحات تار هستند که می‌توانند اطلاعات خود را به راحتی به اشتراک بگذارند. یک تخمین ساده اندازه صفحات تار را بالغ بر بیلیون صفحه نشان می‌دهد. با افزایش بیش از حد متون و در اختیار قرار گرفتن اطلاعات زیاد، استفاده از مفهوم هوش تجاری در متن، تبدیل به جزئی ضروری شده است. هوش تجاری اطلاعات لازم برای درک، مدیریت و هدایت اشخاص، شرکت‌ها و سازمان‌ها را در اختیار تصمیم‌گیرندگان قرار می‌دهد و در این راستا استفاده از منبعی همانند متن امری اجتناب‌ناپذیر است. بایستی به این حقیقت توجه داشت که اینترنت در قالب یک پروژه تحقیقاتی برای بررسی شبکه بین مراکز کامپیوتری بنا نهاده شده است. علاوه بر پژوهشگران علم کامپیوتر، پس از آن سایر دانشمندان و مهندسین نیز شروع به استفاده از اینترنت برای به اشتراک گذاشتن اطلاعات و داده‌های خود کردند. پیدایش و معرفی شبکه جهانی تار به عنوان وسیله‌ای

برای ایجاد ارتباط بین دانشمندان و رد و بدل کردن اطلاعات و مقالات علمی، دروازه‌ها را برای جهان بیرون گشود تا بتوانند از طریق اینترنت و بهره‌گیری از مزایای آن داده‌ها و اطلاعات خود را منتشر کنند. بنابراین امروزه ما به صورت مجازی به تمام اطلاعات موجود و عمومی دسترسی داریم و این تنها به دلیل وجود شبکه گسترده جهانی تار است.

٤- ٢) معرفی نرم افزار

برخلاف بسیاری از زبان‌های کامپیوتری دیگر، متلب دستورات بسیاری را برای رسم و تصویربرداری متلب یک زبان برنامه‌نویسی سطح بالای نسل چهارم و یک محیط تعاملی برای محاسبات عددی، تجسم و برنامه‌نویسی می‌باشد که از ترکیب دو واژه ماتریس[1]و آزمایشگاه[2]ایجادشده است این نام حاکی از رویکرد ماتریس محور برنامه است که در آن حتی اعداد منفرد نیز به‌صورت یک ماتریس با ابعاد ۱×۱ در نظر گرفته می‌شود. نرم‌افزار متلب توسط شرکت MathWorks تولیدشده است. این شرکت در سال ۱۹۸٤ در ایالت ماساچوست امریکا تأسیس شد. در سال ۱۹۷۰ Cleve Moler رییس دانشکده نیومکزیکو نرم‌افزار متلب را بر پایه زبان فرترن نوشت. در سال ۱۹۸۳ این نرم‌افزار را بر پایه زبان برنامه‌نویسی C شکل دادند و پس از تأسیس شرکت گسترش آن سرعت گرفت. متلب توانایی کار با ماتریس‌ها، رسم انواع توابع و داده‌ها، پیاده‌سازی انواع الگوریتم‌ها، ایجاد رابط کاربری، ارتباط با برنامه‌های نوشته‌شده به زبان‌های دیگر ازجملهC ، C++، JAVA و فرترن و ایجاد مدل‌ها و برنامه‌های کاربردی را فراهم می‌کند. کار کردن با ماتریس‌ها در متلب بسیار ساده است. در حقیقت تمام داده‌ها در متلب به شکل یک ماتریس ذخیره می‌شوند. برای مثال یک عدد (اسکالر) به شکل یک ماتریس ۱×۱

[1] MATrix
[2] LABoratory

ذخیره می‌شود. یک رشته مانند «Whale is the biggest animal» به شکل ماتریسی با یک سطر و چندین ستون (که تعداد ستون‌ها به تعداد کاراکترهاست) ذخیره می‌شود. حتی یک تصویر به شکل یک ماتریس سه بعدی ذخیره می‌گردد که بُعد اول و دوم آن برای تعیین مختصات نقاط و بُعد سوم آن برای تعیین رنگ نقاط استفاده می‌شود. فایل‌های صوتی نیز در متلب به شکل ماتریس‌های تک ستون (بردارهای ستونی) ذخیره می‌شوند؛ بنابراین جای تعجب نیست که متلب مخفف عبارت آزمایشگاه ماتریس باشد. علاوه بر توابع فراوانی که خود متلب دارد، برنامه‌نویس نیز می‌تواند توابع جدید تعریف کند. ساخت رابط گرافیکی کاربر مانند دیالوگ‌هایی که در محیط‌های ویژوال مانند بیسیک و C وجود دارند، در متلب امکان‌پذیر است. این قابلیت، ارتباط بهتری را میان برنامه‌های کاربردی نوشته‌شده با متلب و کاربران برقرار می‌کند. هسته متلب برای سرعت و کارایی بالا به زبان سی[1] نوشته شده‌است ولی رابط گرافیکی آن به زبان جاوا پیاده‌سازی گشته‌است. برنامه‌های متلب اکثراً متن‌باز هستند و در واقع متلب (مانند بیسیک) مفسر (رایانه) است نه کامپایلر. قدرت متلب از انعطاف‌پذیری آن و راحت بودن کار با آن ناشی می‌شود، همچنین شرکت سازنده و گروه‌های مختلف، از جمله دانشگاه‌های سرتاسر جهان و برخی شرکت‌های مهندسی هر ساله جعبه‌ابزارهای خاص-کاربردی به آن می‌افزایند که باعث افزایش کارایی و محبوبیت آن شده‌است.

➢ کاربردهای متلب

متلب به‌طور گسترده به‌عنوان یک ابزار محاسباتی در علم و مهندسی مانند رشته‌های فیزیک، شیمی، ریاضی و تمام رشته‌های مهندسی استفاده می‌شود. در زیر بعضی از موارد استفاده از متلب مطرح‌شده است:

- پردازش سیگنال و ارتباطات
- پردازش تصویر و ویدئو
- سیستم‌های کنترل
- تست و اندازه‌گیری
- مهندسی مالی
- محاسبات زیستی

➢ مزایای متلب

۱- راحتی در استفاده: متلب یک زبان مفسری است که برنامه در محیط توسعه یکپارچه متلب به‌راحتی نوشته، اصلاح و ایجاد می‌گردد. ازآنجایی‌که زبان برنامه‌نویسی برای استفاده راحت است توسعه برنامه‌های جدید به‌راحتی امکان‌پذیر است.

۲- استقلال بستر نرم‌افزاری: متلب توسط بسیاری از سیستم‌های کامپیوتری مختلف پشتیبانی می‌شود. زبان متلب توسط سیستم‌عامل‌های لینوکس، ویندوز و مکینتاش پشتیبانی می‌شود.

۳- توابع از پیش تعریف‌شده: متلب هم را با کتابخانه گسترده‌ای از توابع از پیش تعریف‌شده است که برای بسیاری از کاربردها استفاده می‌شود.

٤- رسم مستقل از دسدارد. این تصاویر و رسم‌ها می‌تواند روی هر وسیله خروجی گرافیکی که توسط کامپیوتر پشتیبانی می‌شود قابل نمایش است.

٥- واسط گرافیکی کاربر: متلب شامل ابزاری است که به برنامه‌نویس اجازه می‌دهد که به‌صورت تعاملی یک واسط گرافیکی کاربر را ایجاد نماید. با این قابلیت برنامه‌نویس می‌تواند برنامه‌های پیچیده تجزیه‌وتحلیل داده‌ها را طوری طراحی کند که کاربران بی‌تجربه نیز بتوانند به‌راحتی با برنامه تعامل داشته باشند.

٤- ۳) داده‌های مورد بررسی

مجموعه داده‌های متنی زیادی وجود دارند که می‌توان به عنوان مجموعه داده‌های تست برای دسته‌بندی به کار برد. WebKB مجموعه داده‌ای است که شامل صفحات وب از گروه‌های علوم کامپیوتر دانشگاه‌های مختلف است. ٤٥١٨ صفحه وب در ٦ دسته نامتعادل (دانشجو، اساتید، کارکنان، گروه، دوره، پروژه) طبقه‌بندی می‌شوند.

مجموعه داده 20 Newsgroups مجموعه‌ای متشکل از ۲۰۰۰۰ سند گروه خبری است که (تقریباً) به طور مساوی در بین ۲۰ گروه خبری مختلف تقسیم شده است. این مجموعه داده به مجموعه داده‌ای محبوب برای آزمایش‌ها در کاربردهای متنی تکنیک‌های یادگیری ماشین، مانند طبقه‌بندی متن و خوشه‌بندی متن تبدیل شده است.

اسناد موجود در مجموعه Reuters-21578 در سال ۱۹۸۷ در شبکه خبری رویترز ظاهر شد. اسناد توسط پرسنل رویترز Ltd. (سم دابینز، مایک توپلیس، استیو واینستاین) و گروه کارنگی، شرکت (پگی اندرسن،) جمع‌آوری و با دسته‌بندی‌ها فهرست شدند.

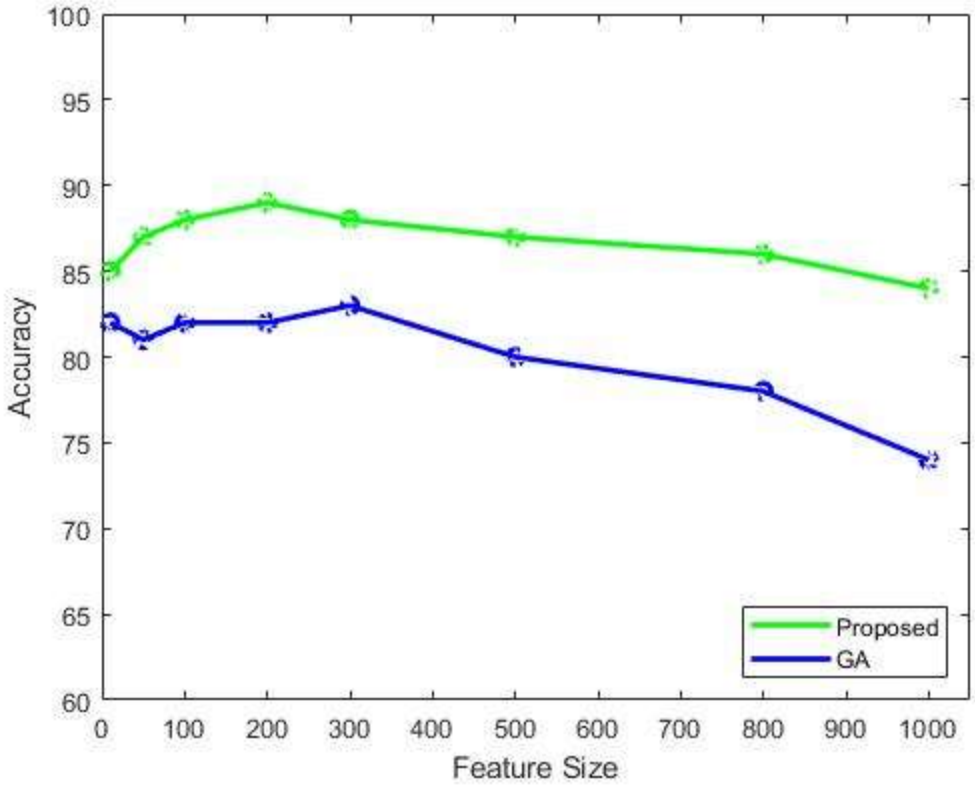
در هر بار ۷۵ درصد از داده‌هایی که از قبل وجود دارد را به عنوان آموزش و ۲۵درصد مابقی را به عنوان تست به الگوریتم داده می‌شود. از قبل مشخص است که این ۲۵درصد باید در کدام گروه‌ها طبقه‌بندی شوند. پس از دادن این ۲۵درصد به الگوریتم می‌توان با گروه‌بندی واقعی مقایسه کرد و میزان طبقه‌بندی درست داده‌ها را به دست آورد. وقتی کار با این ۲۵ درصد تمام شد، ۲۵درصد دیگر از کل داده‌ها را برای تست انتخاب و ۷۵درصد مابقی به آموزش اختصاص داده می‌شود.

٤- ٤) نتایج

امروزه با توجه به رشد روزافزون دسترسی به اسناد الکترونیکی، دسته‌بندی خودکار اهمیت ویژه‌ای یافته است. دسته‌بندی متون به عمل برچسب گذاری موضوعی متون زبان طبیعی برمبنای یک مجموعه از پیش تعیین شده می‌باشد. در طول سالیان اخیر، طبقه‌بندی‌های مختلف و با نگرش‌های مختلفی برای این کار مطرح شده است. نکته قابل توجه، امکان خودکارسازی این طبقه‌بندی‌ها به ازای متون جدید است. مسئله انتخاب ویژگی ناشی از زیادی نویز و ویژگی‌های نامربوط و اضافی در مجموعه داده‌ها است، به وسیله حذف این ویژگی‌ها از مجموعه داده‌ها کارائی مدل‌های یادگیری به طور چشمگیری افزایش پیدا می‌کند. هدف از انتخاب ویژگی پیدا کردن کوچکترین زیرمجموعه از ویژگی‌های ورودی با بیشترین خاصیت پیش‌گویانه است. انتخاب ویژگی در زمینه‌های زیادی همچون شناسایی الگو، داده‌کاوی، یادگیری ماشین، پردازش سیگنال، بازیابی اطلاعات چندرسانه‌ای و دیگر زمینه‌ها کاربرد دارد.

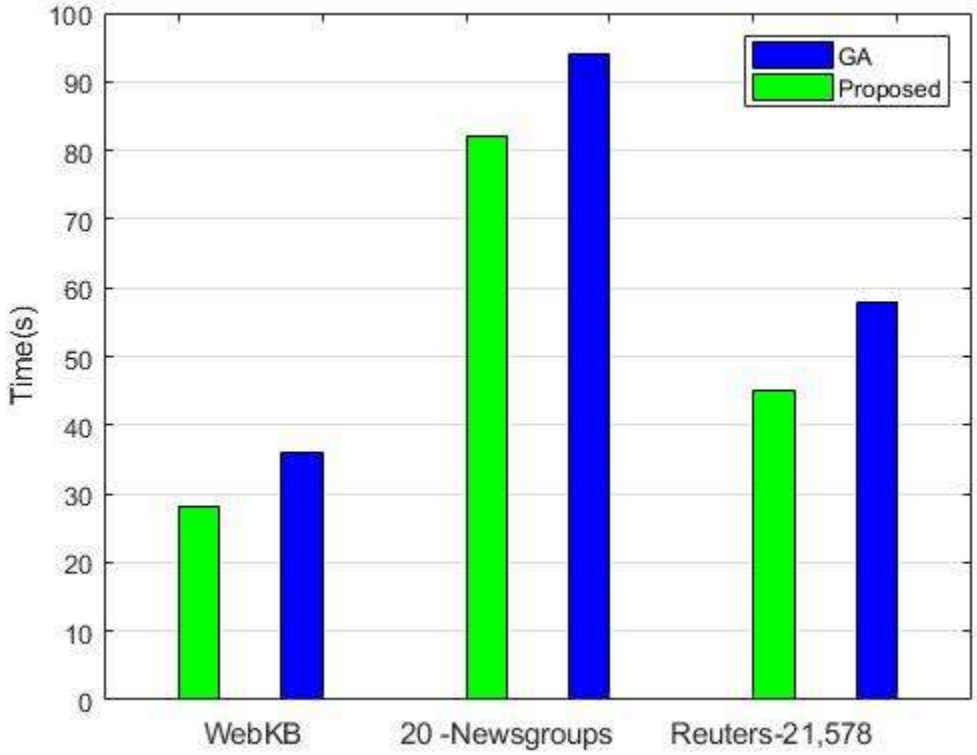
امروزه پیشرفت امکانات نرم‌افزاری و سخت‌افزاری، موجب آسانی ذخیره شدن مقادیر زیادی داده شده است. تعداد مستندات متنی روز به روز در حال افزایش است، نامه‌های الکترونیکی، صفحات وب، متون خبری و مقالات تنها بخشی از این گستره رو به افزایش هستند. بنابراین نیاز به تکنیک‌های متن‌کاوی همانند روش‌های خودکار برای رده‌بندی متون احساس می‌شود. در امر رده‌بندی خودکار متون، انتخاب ویژگی از درون متن جزء مهمترین مراحل می‌باشد. انتخاب ویژگی برای کاهش ابعاد فضای ویژگی استفاده می‌شود، چرا که فضای ویژگی برای متون شامل ده‌ها هزار کلمه خواهد بود که پردازش‌های بعدی سیستم را امکان ناپذیر می‌کند. تاکنون روشهای مختلفی برای انتخاب ویژگی برای داده‌های متنی طراحی شده‌اند که هر یک دارای معایب و مزایایی هستند، ولی روشی کلی که اکثر سیستم‌های رده‌بندی متون از آن استفاده کنند و میزان کارایی بالایی نیز داشته باشد معرفی نشده است.

انتخاب ویژگی در داده‌های با ابعاد بالا همانند پردازش متون و طبقه‌بندی متون کاربرد فراوانی دارد. ازآنجاییکه انتخاب ویژگی دارای پیچیدگی زمانی نمایی است، استفاده از روش‌های کلاسیک موجب افزایش زمان اجرا می‌شود. در بیشتر مواقع این روش‌ها قادر به یافتن راه‌حل بهینه نمی‌باشند. یکی از اهداف انتخاب ویژگی، یافتن راه‌حل‌های بهینه و یا نیمه بهینه است. در این پایان‌نامه روشی جدید بر مبنای الگوریتم بهینه‌سازی بیوه سیاه برای حل مسئله انتخاب ویژگی ارائه شده است.



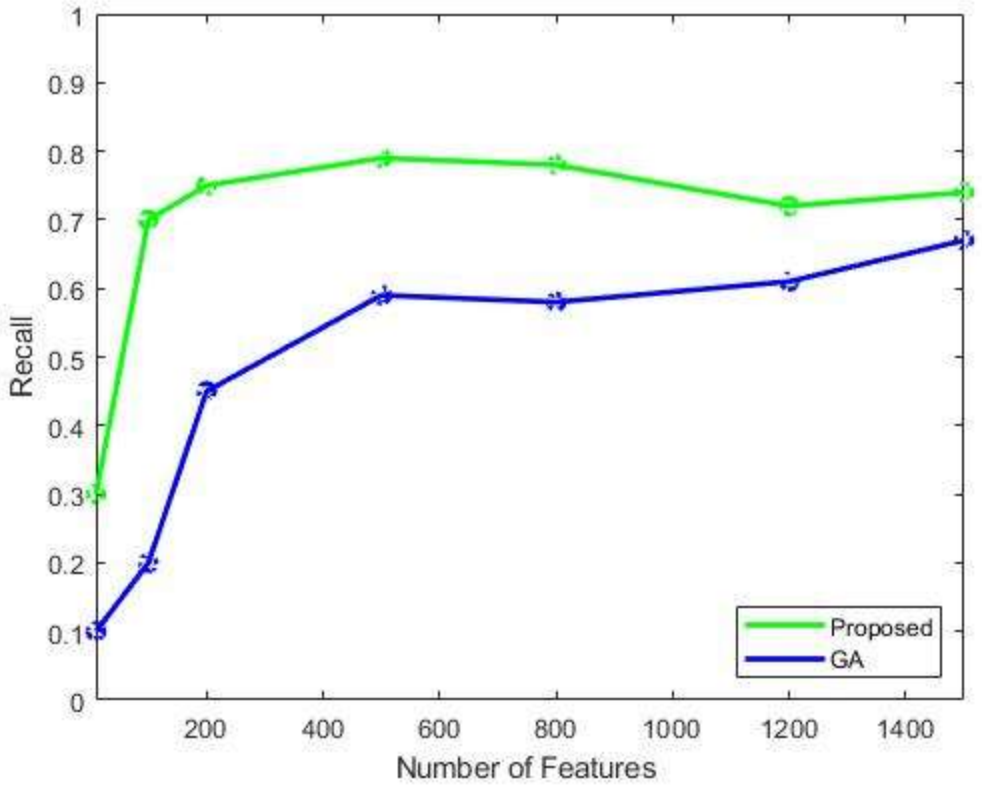
شکل ۴- ۱) میزان درستی تشخیص

در شکل ٤- ۱ میزان درستی و دقت تشخیص و اجرای الگوریتم در بین مقادیر متعدد از ویژگی‌ها برای دو الگوریتم بیوه سیاه و ژنتیک مورد بررسی قرار گرفته است، هر دو روش پیشنهاد شده در محیط متلب پیاده‌سازی شده و نتایج نشان داده شده در نمودارها حاصل شده است. طبق نتایج بدست آمده میزان دقت روش پیشنهاد شده مبتنی بر الگوریتم بیوه سیاه به مقدار ٪۹۰ نیز در تعداد ویژگی ۳۰۰ رسیده است. روند اجرای الگوریتم نشان می‌دهد با افزایش تعداد ویژگی‌ها میزان دقت کاهش یافته در تعداد ویژگی‌های ۱۰۰۰ برابر با مقدار ٪۸۵ می‌باشد، میزان دقت بدست آمده از الگوریتم مورد بررسی در این پژوهش در مقایسه با روش‌های مشابه مانند ژنتیک به میزان قابل توجهی افزایش یافته است. میزان دقت برای الگوریتم ژنتیک از میزان ٪۸۳ شروع شده و در مقدار ٪۷۵ در تعداد ویژگی‌های ۱۰۰۰ کاهش می‌یابد.



شکل ۴- ۲) زمان انتخاب ویژگی با استفاده از الگوریتم‌های مختلف

در شکل ٤- ۲ زمان صرف شده برای انتخاب ویژگی در مجموعه داده‌های مختلف توسط دو الگوریتم ژنتیک و بیوه سیاه بررسی شده است. طبق نتایج بدست آمده می‌توان گفت، زمان مورد نیاز برای انتخاب ویژگی برای هر سه مجموعه داده توسط الگوریتم بیوه سیاه کمتر از الگوریتم ژنتیک می‌باشد، همچنین زمان صرف شده برای مجموعه داده‌ی 20-newsgroup در مقایسه با دو مجموعه داده‌ی دیگر مقدار بیشتری دارد. کمترین زمان صرف شده برای بررسی مجموعه داده مربوط به مجموعه داده‌ی WebKB می‌باشد.



شکل ۴- ۳) بررسی معیار پوشش داده‌ها

در این شکل به بررسی و محاسبه‌ی معیار پوشش داده‌ها پرداخته شده است این معیار به منظور بررسی میزان پوشش‌دهی درست داده‌ها مورد توجه قرار می‌گیرد مفهوم این پارامتر میزان داده‌هایی است که به صورت درست و صحیح تشخیص داده شده است و داده‌های بدون خطا را نشان می‌دهد. همانطور که در شکل ٤- ۳ نمایش داده شده است میزان پارامتر پوشش در الگوریتم پیشنهادی مبتنی بر الگوریتم بیوه سیاه نسبت به الگوریتم مشابه در وضعیت بهتری داشته و نشان‌دهنده‌ی این است که روش پیشنهاد شده قدرت پوشش دهی بهتری نسبت به سایر روشهای مشابه داشته است و تقریبا ۸۰٪ داده‌ها را پوشش داده و به درستی تشخیص می‌دهد.

جدول ۴- ۱) بررسی میزان معیار پوشش در الگوریتم‌های مختلف در مقایسه با الگوریتم پیشنهادی

دیتاست	CSO Greedy	GA	BWO
PCMAC	۷۵	۷۵ /۵۳	–
SPAM	۸۲/۸۵	۸۰	–
BOOK	۸۳/۴۵	۸۱/۹۵	–
Web KB	–	۷۰	۸۳/۸
20 Newsgroups	–	۷۳/۷۵	۸۲
Reuters-21,578	–	۷۲/۷۱	۸۰/۷۶

در جدول ٤- ۱ مقادیر بدست آمده برای معیار پوشش در الگوریتم‌های مختلف مورد بررسی قرار گرفته است. مقادیر بدست آمده با مقادیر بدست آمده در روش پیشنهادی تحت الگوریتم‌های مختلف مورد بررسی و مطالعه قرار گرفته است که در جدول ٤- ۱ ارائه شده‌اند. در مقایسه صورت گرفته با توجه به نتایج حاصل می‌توان نتیجه گرفت مقادیر بدست آمده برای روش پیشنهادی نسبت به روش‌های پیشین بر روی دیتاست‌های متفاوت مقادیر بهتری ارائه کرده و یک پیشرفت نسبی در مقایسه با روش‌های مشابه داشته است.

جدول ۴- ۲) بررسی میزان معیار صحت و درستی تشخیص در الگوریتم‌های مختلف در مقایسه با الگوریتم پیشنهادی

دیتاست	CSO Greedy	GA	BWO
PCMAC	۷۴/۸۰	۷۷/۲۷	–
BOOK	۸۵/۲۰	۸۲/۱۰	–
Web KB	–	۸۴/۳	۹۲/۵
20 Newsgroups	–	۸۰/۵۴	۹۱/۵۴
Reuters-21,578	–	۸۵/۸۳	۹۲/۳۰

در جدول ٤- ٢ میزان درستی و صحت مقادیر تشخص داده شده نشان داده شده است. طبق نتایج بدست آمده در این جدول می‌توان نتیجه گرفت مقادیر بدست آمده برای درستی برای روش پیشنهادی در دیتاست‌های مختلف در مقایسه با روش‌های مشابه یک پیشرفت نسبی داشته و نسبت به روش‌های قبلی بهتر نتیجه داده است.

٤-٥) خلاصه فصل

در فصل جاری در ابتدا به معرفی نرم افزار متلب پرداخته شد و در ادامه به ارائه‌ی نتایج بدست آمده از پیاده‌سازی روش پیشنهادی مبتنی بر الگوریتم بیوه سیاه ارائه شده است، همچنین مقایسه‌ای از مقادیر بدست آمده برای روش پیشنهادی در مقایسه با کارهای پیشین صورت گرفته است.

۵) فصل پنجم

نتیجه‌گیری و پیشنهادات

۵- ۱) مقدمه

محدوده موضوعات پوشش داده شده از طریق منابع در اینترنت تا حدی گسترده است که هر مطلب با عنوان دلخواه قابل یافته شدن است. با این حال صفحات تار تنها منبع برای رشد اطلاعات اینترنتی نیست. شرکت‌ها و سازمان‌ها دارای سیستم‌های مدیریت اسنادی هستند که به سرعت در حال افزودن متن به منابع متنی جهان هستند. امروزه معدود شرکت‌هایی وجود دارند که بتوانند و یا بخواهند بدون استفاده از پست الکترونیکی فعالیت کنند. اکثر دانشگاه‌ها و موسسات آموزشی نیاز به استفاده از پژوهش‌ها، مقالات و سایر منابع متنی دارند. مهندسین بایستی به طرح، مشخصات و اسناد پروژه‌ها دسترسی یابند. فروشنده‌ها و مدیران فروشگاه‌ها نیاز به بررسی میزان و نحوه فروش دارند که برای این بررسی و آزمون باید از اسناد متنی استفاده شود. بنابراین به طور کلی متون و استفاده از اسناد متنی در تمام امور و کاربردها از پزشکی تا تجارت نقشی اساسی را بر عهده دارند. اما در نقطه مقابل تاثیر منفی فزونی اطلاعات متنی در تار را می‌توان به دو شاخه تقسیم کرد. نکته اول این است که راه مناسبی برای یافتن تمام اطلاعات مورد نیاز که به جستجوی آنها پرداخته می‌شود، وجود ندارد. یا به عبارت دیگر رسیدن به تنها اطلاعاتی که مورد نیاز است، در بین انبوه اطلاعات موجود در تار، کار بسیار مشکل و هزینه‌بری است. مسئله دوم این است که حتی اگر بخشی از اطلاعاتی را که مورد نیاز است، یافته شود، مطالب زیادی برای خواندن خواهیم داشت. این دو مسئله به خاطر فراوانی اطلاعات موجود است. با تمام این مزایا و مشکلات نیاز به سیستم‌هایی برای بازیابی اطلاعات وجود دارد. داده‌کاوی یا کشف دانش در پایگاه‌های داده‌ها

علم نسبتاً تازه‌ای است که با توجه به پیشرفت جهانی در زمینه فناوری اطلاعات، فزونی بیش از اندازه اطلاعات و نفوذ استفاده از سیستم‌های کامپیوتری در صنعت و ایجاد بانک‌های اطلاعاتی بزرگ توسط ادارات دولتی، بانک‌ها و بخش خصوصی نیاز به استفاده از آن به طور عمیقی احساس می‌شود.

انتخاب ویژگی فرآیندی است که زیرمجموعه‌ای از ویژگی‌ها را بر اساس یک معیار به عنوان ویژگی‌های اصلی انتخاب می‌کند. ویژگی‌های انتخاب شده دارای معنا و مفهوم مشخص می‌باشند و فرآیند یادگیری را ساده‌تر و سریع‌تر می‌کنند. این مرحله بعد از مرحله پیش پردازش انجام می‌گیرد و هم در رده‌بندی متون و هم در خوشه‌بندی متون نیاز می‌باشد و می‌تواند به صورت نظارتی و یا غیرنظارتی باشد. در واقع استخراج ویژگی یا انتخاب ویژگی یک مرحله مهم در یادگیری ماشین است. همچنین با حذف ویژگی‌های نویزی در دسته‌بندی باعث می‌شود تا بهره‌وری دسته‌بندی افزایش پیدا کند. ایده اساسی این الگوریتم جستجو در تمام ترکیبات ممکن از ویژگی‌ها و انتخاب یک زیرمجموعه با بهترین کارکرد برای پیش‌بینی و دسته‌بندی می‌باشد. این کار با حفظ ویژگی‌هایی که معنادارتر هستند و از بین بردن ویژگی‌هایی که بی‌ربط هستند انجام می‌شود. الگوریتم‌های انتخاب ویژگی با معیارهای ارزیابی مختلف طراحی شده‌اند و به طور کلی به سه دسته تقسیم می‌شوند. که در این پژوهش به بررسی یک نمونه از این نوع الگوریتم‌ها پرداخته شده است.

۵- ۲) نتیجه‌گیری

انتخاب ویژگی فرآیندی است که زیرمجموعه‌ای از ویژگی‌های مجموعه متون را بر اساس یک معیار انتخاب می‌کند. انتخاب ویژگی باعث می‌شود فضای با ابعاد زیاد داده‌های ورودی به فضایی با ابعاد کوچکتر تبدیل شود. البته در صورتی که این انتخاب به درستی صورت نگیرد می‌تواند موجب کاهش دقت رده‌بندی متون

شود. کاهش ابعاد مسائلی که ابعاد زیادی دارند موضوع خیلی مهمی در یادگیری ماشین و شناسایی الگو است. فرایند کاهش ابعاد به دو صورت انتخاب ویژگی و استخراج ویژگی انجام می‌شود. روش‌های تبدیل داده یا روش‌های استخراج ویژگی هنگام کاهش ابعاد معنای اصلی داده‌ها را از بین می‌برند و به همین دلیل برای حفظ معنای داده‌ها از روش‌های انتخاب ویژگی استفاده می‌شود، این روش‌ها هنگام کاهش ابعاد معنای مجموعه داده-های اصلی را حفظ می‌کنند. روش‌ها اکثراً در پیدا کردن راه‌حل‌های بهینه ناموفق هستند، برای این که در پیدا کردن راه‌حل‌ها از اکتشاف استفاده نمی‌کنند. از طرف دیگر، جستجوی کامل برای پیدا کردن راه‌حل‌های بهینه حتی در مجموعه داده‌هایی که اندازه متوسط دارند، غیرممکن است. فضای جستجوی مسئله انتخاب ویژگی یک فضای نمایی است NP-hard است، به همین دلیل مسئله انتخاب ویژگی جزء مسائل در این پایان‌نامه یک روش جدید بر مبنای الگوریتم‌های هوش ازدحامی برای انتخاب ویژگی ارائه شده است. هوش ازدحامی بر مبنای رفتارهای جمعی در گروه‌های نامتمرکز و خودسازمانده بنیان شده است. این گروه‌ها معمولاً از جمعیتی از عامل‌های ساده تشکیل شده‌اند که به طور محلی با یکدیگر و با محیط اطراف خود در تعامل هستند، با وجود اینکه هیچ نوع کنترل متمرکزی روی رفتار عامل‌ها وجود ندارد ولی تعاملات محلی آنها باعث پیدایش رفتار عمومی می‌شود. یکی از روش‌های جدید که در این حیطه وجود دارد الگوریتم بیوه سیاه می‌باشد که در این پژوهش از این الگوریتم استفاده شده است.

الگوریتم پیشنهادی که بر اساس روشی مبتنی بر الگوریتم بهینه‌سازی فاخته می‌باشد برای انتخاب ویژگی در طبقه‌بندی متن که جز دیتاست‌های بزرگ می باشد استفاده شد. نتایج به دست آمده از این روش در انتخاب ویژگی نشان می‌دهد این الگوریتم از سرعت همگرایی بالایی برخوردار است و توانایی زیادی در جستجوی فضای مسئله دارد و می‌تواند کوچکترین زیرمجموعه قابل قبول را پیدا کند. برای نشان دادن کارآمدی روش

پیشنهادی این روش را با الگوریتم ژنتیک در محیط متلب پیاده سازی و مورد مقایسه قرار داده شد همچنین از نتایج بدست آمده برای سایر الگوریتم‌ها به منظور مقایسه‌ی دقیق‌تر این روش بهره برده شده است. نتایج به دست آمده نشان داد که روش ارائه شده از لحاظ دقت بهتر از الگوریتم‌های دیگر عمل می‌کند، طبق نتایج بدست آمده در این پژوهش میزان درستی و دقت تشخیص در روش پیشنهاد شده نسبت به الگوریتم ژنتیک و سایر الگوریتم‌ها و بر روی دیتاست‌های مختلف به طور نسبی بهتر عمل کرده و میزان پوشش دهی داده‌ها در روش ارائه شده در مقایسه با سایر روش‌ها پیشرفت خوبی داشته است. همچنین این روش برای تعداد ویژگی‌های بالا دارای کیفیت و دقت بالایی بوده که در مقایسه با روش‌های مشابه بهتر عمل کرده و بهبود مورد نظر تا حد خوبی ایجاد شده است.

۵- ۳) پیشنهادات کارهای آتی

در این بخش به بیان پیشنهاداتی در راستای بهبود و توسعه تلاش‌های انجام شده در این پژوهش‌ها و سایر پژوهش‌های انجام شده صورت گرفته شده است.

۱- با توجه به اینکه در زمینه زبان فارسی و متون آن کارهای زیادی صورت نگرفته است، بنابراین در زمینه رده‌بندی متون فارسی نیز پژوهش‌های اندکی موجود است، در نتیجه کاوش بر روی متون فارسی و طراحی سیستم‌های بازیابی اطلاعات برای این زبان به عنوان مهمترین پیشنهاد در این پایان نامه مطرح می‌شود.
۲- در فاز پیش پردازش، مرحله ریشه‌یابی کلمات پیاده‌سازی نشده و مورد استفاده قرار نگرفته است. همان طور که میدانیم ریشه‌یابی کلمات موجب افزایش کارایی سیستم رده‌بندی و کاهش فضای ویژگی خواهد شد، بنابراین در زمینه کاوش بر روی متون فارسی بهتر است از مرحله ریشه‌یابی کلمات نیز استفاده کنیم.
۳- روش پیشنهادی اول بر اساس الگوریتم سرد شدن شبیه‌سازی شده به عنوان ابزار جستجو در فضای ویژگی

با ابعاد زیاد استفاده شده است. بنابراین می‌توان از روش‌های دیگر جستجو و همچنین سایر روش‌های مکاشفه‌ای و تکاملی نیز برای حل این مسئله استفاده کرد.

٤- تابع برازندگی ارائه شده، در روش پیشنهادی اول بر اساس بسامد سند می‌باشد و به گونه‌ای طراحی شده است که سرعت و کارایی بالایی داشته باشد. در مورد طراحی تابع برازندگی بهینه بودن تابع برازندگی از نظر زمان بسیار مهم می‌باشد زیرا بیشترین تأثیر را بر روی زمان الگوریتم دارد. طراحی تابع برازندگی مناسب می‌تواند باعث افزایش کارایی سیستم شود.

۵- همچنین به عنوان پیشنهاد می‌توان گفت سسایر محققین از روش‌های زیر بهره ببرند.

- استفاده از سایر توصیف‌گرهای متن
- استفاده از یادگیری عمیق برای تشخیص
- همچنین می‌توان از سایر روش‌های ترکیبی با استفاده از سایر الگوریتم‌های فراابتکاری مانند ژنتیک، گرگ خاکستری و.... برای تشخیص استفاده کرد.

منابع

[1] Labani, M., Moradi, P., Ahmadizar, F., & Jalili, M. (2018). A novel multivariate filter method for feature selection in text classification problems. Engineering Applications of Artificial Intelligence, 70, 25–37

[2] Liu, C. L., Hsaio, W. H., Lee, C. H., Chang, T. H., & Kuo, T. H. (2016). Semi-supervised text classification with universum learning. IEEE Transactions on Cybernetics, 46(2), 462–473.

[3] Maruf, S., Javed, K., & Babri, H. A. (2016). Improving text classification performance with random forests-based feature selection. Arabian Journal for Science & Engineering, 41(3), 951–964.

[4] Hijazi, M., Zeki, A., Ismail, A. (2021). Arabic Text Classification Using Hybrid Feature Selection Method Using Chi-Square Binary Artificial Bee Colony Algorithm, I nternational Journal of Mathematics and Computer Science, no. 1, 213–228.

[5] Cekik, R., Kursat Uysal, A. (2020). A novel filter feature selection method using rough set for short text data, Expert Systems with Applications 160, 113691.

[6] Chantar, H., Mafarja, M., Alsawalqah, H., Asghar Heidari, A., Aljarah, I., Faris, H. Feature selection using binary grey wolf optimizer with elite-based crossover for Arabic text classification, Neural Computing and Applications.

[7] Heyong, W., Ming, H. (2019). Supervised Hebb rule based feature selection for text classification, Information Processing and Management 56 (2019) 167–191

[8] Bahassine, S., Madani, A., Al-Sarem, M., Kissi, M. (2018). Feature selection using an improved Chi-square for Arabic text classification, Journal of King Saud University – Computer and Information Science.

[9] Roshdi, A. New Method Of Feature Selection For Persian Text Mining Based On Evolutionary Algorithms, ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 6, No.18.

[10] Hayyolalam, V., Pourhaji Kazem, A.A. (2020). Black Widow Optimization Algorithm: A novel meta-heuristic approach for solving engineering optimization problems, Engineering Applications of Artificial Intelligence 87,103249.

[11] Hosseinzadeh Aghdam, M., Ghasem-Aghaee, N., Ehsan Basiri, M., Text feature selection using ant colony optimization, Expert Systems with Applications 36, 6843–6853.

[12] Saeed Ghareb, A., Abu Bakar, A., Razak Hamdan, A. (2015). Hybrid feature selection based on enhanced genetic algorithm for text categorization, Expert Systems With Applications.

[13] Bahassine, S., Madani, A., Kissi, M. (2016). An improved Chi-sqaure feature selection for Arabic text classification using decision tree, 11th International Conference on Intelligent Systems: theories and Applications (SITA). Mohammedia, 1–5. https://doi.org/10.1109/SITA.2016.7772289.

[14] Bahassine, S., Madani, A., Kissi, M. (2017). Arabic text classification using new stemmer for feature selection and decision trees. J. Eng. Sci. Technol. 12 (6), 1475–1487.

[15] Baraa, S., Nazlia, O., Zeyad, S. (2014). An automated Arabic text categorization based on the frequency ratio accumulation. Int. Arab J. Inform. Technol. 11 (2), 213–221.

[16] Singh, V., Saxena, P., Singh, S., Rajendran, S. (2017). Opinion mining and analysis of movie reviews. Indian J. Sci. Technol. 10 (19)

[17] Suhad, I.E., Yousif, A., Venus, W., Samawi Zantout, R. (2015). The effect of combining different semantic relations on arabic text classification. World Comput. Sci. Inform. Technol. J. (WSCIT) 5 (6), 112–118.

[18] Ye-wang, C., Qing, Z., Wei, L., Ji-Xiang, D. (2016). Classification of Chinese texts based on recognition of semantic topics. Cognit. Comput. 8 (1), 114–124. https://doi.org/10.1007/s12559-015-9346-8.

[19] Rasha, M., Mahmoud, A. (2016). Arabic text stemming: comparative analysis, Conference of Basic Sciences and Engineering Studies (SGCAC). Khartoum, 88–93. https://doi.org/10.1109/SGCAC.2016.7458011.

[20] Onan, A., & Korukoğlu, S. (2017). A feature selection model based on genetic rank aggregation for text sentiment classification. Journal of Information Science, 43(1), 25–38

[21] Rehman, A., Javed, K., & Babri, H. A. (2017). Feature selection based on a normalized difference measure for text classification. Information Processing & Management,53(2), 473–489.

[22] Tommasel, A., & Godoy, D. (2018a). A social-aware online short-text feature selection technique for social media. Information Fusion, 40, 1–17

[23] Wang, H., & Hong, M. (2015). Distance variance score: An efficient feature selection method in text classification. Mathematical Problems in Engineering, 2015, 1–102015-5-112015.

[24] Wei, D., Wang, B., Lin, G., Liu, D., Dong, Z., Liu, H., et al. (2017). Research on unstructured text data mining and fault classification based on rnn-lstm with malfunction inspection report. Energies, 10(3), 406.

[25] Esposito, M., DePietro, G. (2011). An ontology-based fuzzy decision support system for multiple sclerosis, international conference on Engineering Applications of Artificial Intelligence Volume 24, Year, pages: 1340–1354.

[26] Aydogan, E., Karaoglan, I., Pardalos, P.h. (2012). GA: Hybrid genetic algorithm in fuzzy rule-based classification systems for high-dimensional problems, Applied Soft Computing, volume 12, Issue: 2, Y, pp:800–806.

[27] Nguyen, T., Khosravi, A. Creighton, D., Nahavandi, S. (2014). Medical Diagnosis by Fuzzy Standard AdditiveModel with Wavelets, IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) July 6-11, Beijing, China, Year: 2014, pages: 1937 – 1944

Abstract:

Today, Progress through software and hardware facilities, causing easily be stored amounts of data. Day by day the number of text documents is increasing, e-mail, web pages, texts, news and articles are only part of this range of increasing. Thus the need for text mining techniques such as methods for automatic text classification is felt. In the automatic text classification, feature selection from within any text is one of the most important steps. Feature selection is using for space dimension reduction, because the feature space includes tens of thousands of words that will cause the next processes of system be impossible. Different methods to feature selection data texts have been designed each with advantages and disadvantages. Feature selection is widely used in high-dimensional data such as text processing and text classification. Since feature selection has an exponential time complexity, the use of classical methods increases the execution time. In most cases, these methods are not able to find the optimal solution. One of the goals of feature selection is to find optimal or semi-optimal solutions. In this thesis, a new method based on the black widow optimization algorithm is presented to solve the problem of feature selection. To show the utility of proposed method and to compare it's with other approache, we select egith datasets in text categorization. From the obtained results of implementations and by comparison of this methods to some other feature selection methods, we conclude that proposed BWO intelligence based methods have very better performance with selecting less number of features. Also, this method by eliminating irrelevant features, improve the accuracy and speed of classifier.

Keyword: feature selection**,** classification, accuracy, black widow optimization algorithm.

University of Tabriz

School of Engineering -Department of Mechanic

Thesis Approve for Master of Science

in Mechatronics Engineering

Title

Optimal control of microbots using particle swarm optimization algorithm

Supervisor

Dr. Habib Ezadkhah

By

Ali Jalali

Date

Full 2022