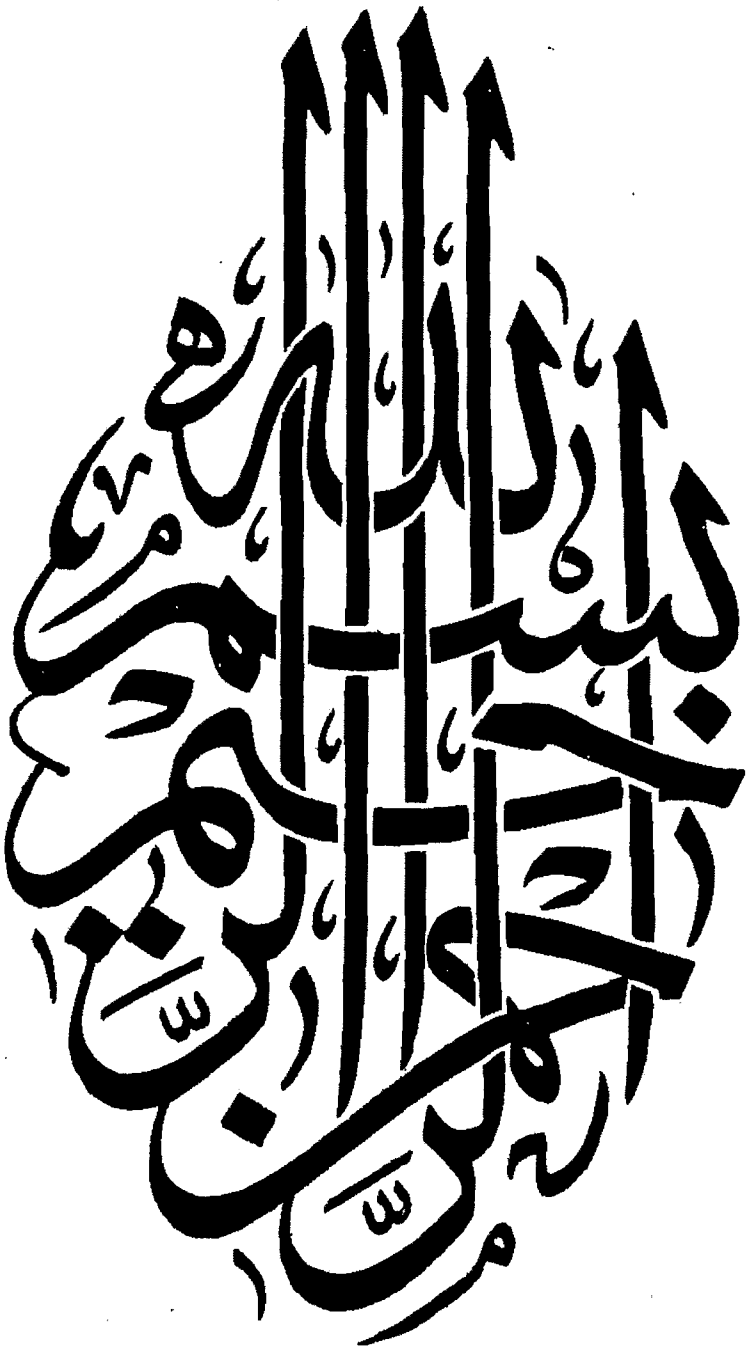

دانشگاه تربیت مدرس

دانشکده فنی و مهندسی

پایان‌نامه‌ی دوره کارشناسی ارشد مهندسی کامپیوتر - نرم‌افزار

ارتقای کیفیت انتخاب خصیصه متون با روش‌های الگوریتم ژنتیک و بهینه‌سازی گروه ذرات

شیما گرانی

استاد راهنما

دکتر سعید جلیلی

زمستان ۱۳۸۶

بسمه تعالی

تاییدیه اعضای هیات داوران حاضر در جلسه دفاع از پایان

خانم شیما گرانی پایان نامه واحدی خود را با عنوان ارتقای کیفیت انتخاب خصیصه متون با روشهای الگوریتم ژنتیک وبهینه سازی گروه ذرات در تاریخ ۱۳۸٦/۱۱/۳ ارائه کردند.

اعضای هیات داوران نسخه نهایی این پایان نامه را از نظر فرم و محتوا تایید کرده و پذیرش آنرا برای تکمیل درجه کارشناسی ارشد مهندسی برق - مهندسی کامپیوترـنرم افزار پیشنهاد می کنند.

عضو هیات داوران	نام و نام خانوادگی	رتبه علمی	امضا
استاد راهنما	دکتر سعید جلیلی	استادیار	
استاد ناظر	دکتر احسان اله کبیر	استاد	
استاد ناظر	دکتر نصراله مقدم چرکری	استادیار	
استاد ناظر	دکتر جعفر حبیبی	استادیار	
مدیر گروه (یا نماینده گروه تخصصی)	دکتر نصراله مقدم چرکری	استادیار	

این نسخه به عنوان نسخه نهایی پایان نامه/رساله مورد تأیید است.

امضای استاد راهنما:

آیین نامه چاپ پایان‌نامه (رساله)های دانشجویان دانشگاه تربیت مدرس

نظر به بنکه چاپ و نشار پایان نامه (رساله)های تحصیلی دانشجویان دانشگاه تربیت مدرس، مبین بخشی از فعالیتهای علمی - پژوهشی دانشگاه است بنابراین به منظور آگاهی و رعایت حقوق دانشگاه،دانش آموختگان این دانشگاه نسبت به رعایت موارد ذیل متعهد می‌شوند:

ماده ۱: در صورت اقدام به چاپ پایان نامه (رساله)ی خود، مراتب را قبلاً به طور کتبی به «دفتر نشر آثارعلمی» دانشگاه اطلاع دهد.

ماده ۲: در صفحه سوم کتاب (پس از برگ شناسنامه) عبارت ذیل را چاپ کند:

«کتاب حاضر، حاصل پایان نامه کارشناسی ارشد/ رساله دکتری نگارنده در رشته مهندسی کامپیوتر است که در سال ۱۳۸۶ در دانشکده فنی دانشگاه تربیت مدرس به راهنمایی سرکار خانم/جناب آقای دکتر سعید جلالی و مشاوره سرکار خانم/جناب آقای دکتر از آن دفاع شده است.»

ماده ۳: به منظور جبران بخشی از هزینه‌های انتشارات دانشگاه، تعداد یک درصد شمارگان کتاب (در هر نوبت‌چاپ) را به «دفتر نشر آثارعلمی» دانشگاه اهدا کند. دانشگاه می‌تواند مازاد نیاز خود را به نفع مرکز نشر درمعرض فروش قرار دهد.

ماده ۴: در صورت عدم رعایت ماده ۳، ۵۰٪ بهای شمارگان چاپ شده رابه عنوان خسارت به دانشگاه تربیت‌مدرس، تأدیه کند.

ماده ۵: دانشجو تعهد و قبول می کند در صورت خودداری از پرداخت بهای خسارت، دانشگاه می‌تواندخسارت مذکور را از طریق مراجع قضایی مطالبه و وصول کند؛ به علاوه به دانشگاه حق می‌دهد به منظور استیفای حقوق خود، از طریق دادگاه، معادل وجه مذکور در ماده ۴ را از محل توقیف کتابهای عرضه شده‌نگارنده برای فروش، تامین نماید.

ماده ۶: اینجانب شهاب گرایی دانشجوی رشته مهندسی کامپیوتر-نرم افزار مقطع کارشناسی ارشد

تعهد فوق وضمانت اجرایی آن را قبول کرده، به آن ملتزم می شوم.

نام و نام خانوادگی: شهاب گرایی

تاریخ و امضا:

۸۶/۱۱/۲۰

دستورالعمل حق مالکیت مادی و معنوی در مورد نتایج پژوهشهای علمی دانشگاه تربیت مدرس

۴۹۱/ط

مقدمه: با عنایت به سیاست‌های پژوهشی دانشگاه در راستای تحقق عدالت و کرامت انسانها که لازمه شکوفایی علمی و فنی است و رعایت حقوق مادی و معنوی دانشگاه و پژوهشگران، لازم است اعضای هیات علمی، دانشجویان، دانش آموختگان و دیگر همکاران طرح، در مورد نتایج پژوهشهای علمی که تحت عناوین پایان‌نامه، رساله و طرحهای تحقیقاتی که با هماهنگی دانشگاه انجام شده است، موارد ذیل را رعایت نمایند:

ماده ۱- حقوق مادی و معنوی پایان نامه‌ها / رساله‌های مصوب دانشگاه متعلق به دانشگاه است و هرگونه بهره‌برداری از آن باید با ذکر نام دانشگاه و رعایت آیین‌نامه‌ها و دستورالعمل‌های مصوب دانشگاه باشد.

ماده ۲- انتشار مقاله یا مقالات مستخرج از پایان نامه/ رساله به صورت چاپ در نشریات علمی و یا ارائه در مجامع علمی باید به نام دانشگاه بوده و استاد راهنما مسئول مکاتبات مقاله باشند.

تبصره: در مقالاتی که پس از دانش آموختگی بصورت ترکیبی از اطلاعات جدید و نتایج حاصل از پایان‌نامه/ رساله نیز منتشر می‌شود نیز باید نام دانشگاه درج شود.

ماده ۳- انتشار کتاب حاصل از نتایج پایان نامه / رساله و تمامی طرحهای تحقیقاتی دانشگاه باید با مجوز کتبی صادره از طریق حوزه پژوهشی دانشگاه و بر اساس آئین نامه های مصوب انجام می‌شود.

ماده ۴- ثبت اختراع و تدوین دانش فنی و یا ارائه در جشنواره‌های ملی، منطقه‌ای و بین‌المللی که حاصل نتایج مستخرج از پایان نامه/ رساله و تمامی طرح‌های تحقیقاتی دانشگاه باید با هماهنگی استاد راهنما یا مجری طرح از طریق حوزه پژوهشی دانشگاه انجام گیرد.

ماده ۵- این دستورالعمل در ۵ ماده و یک تبصره در تاریخ ۱۳۸۴/۴/۲۵ در شورای پژوهشی دانشگاه به تصویب رسیده و از تاریخ تصویب لازم الاجرا است و هرگونه تخلف از مفاد این دستورالعمل، از طریق مراجع قانونی قابل پیگیری خواهد بود.

تقدیم به پدر و مادر عزیزم

برادر مهربانم

و همسر عزیزم

تشکر و قدردانی

سپاس خداوند منان را که بار دیگر دری از درهای رحمت خود را بر من گشود و به من توفیق کسب علم، در راستای افزایش درک خویش از عظمت باری تعالی و شکوه جهان هستی مرحمت نمود. زیرا هر آنچه که بر انسان اتفاق می‌افتد بهانه‌ای است برای شناخت بهتر از جهان هستی و نزدیکی بیشتر با باری تعالی؛ این پژوهش نیز از این قاعده مستثنی نیست هر چند که در نگاه اول موضوع پژوهش چنین حسی را تداعی نمی‌کند.

در فرایند انجام این پژوهش از استاد گرانقدر جناب آقای دکتر سعید جلیلی درس‌های علمی و اخلاقی بی‌شماری را آموختم. از ایشان به دلیل این‌که طی مدت زمان انجام این پژوهش تجربیات ارزنده و چندین ساله‌ی خود را در اختیار اینجانب قرار دادند، نهایت تشکر را دارم.

از جناب آقای دکتر نصراله مقدم، به دلیل حمایت‌های بی‌دریغ‌شان و اعضای محترم هیات داوران به دلیل صرف وقت برای ارزیابی این پژوهش، نهایت تشکر را دارم.

آخرین و نه کمترین سپاس من از آقایان محسن ملانوری، مهدی آبادی و محرم منصوری‌زاده و خانمها پریسا جلیلی، پرواز مهدابی و نرگس خاکپور به دلیل صرف وقت و راهنمایی‌هایشان است.

لازم به ذکر است که این پروژه با استفاده از حمایت مالی مرکز تحقیقات مخابرات ایران انجام شده است.

شیما گرانی

۸۶/۱۱/۲

چکیده

در دسته‌بندی متون، بطور معمول از کلمات متن به‌عنوان خصیصه‌های متن استفاده می‌شود. با توجه به آنکه هر مجموعه متون دارای تعداد زیادی کلمه است، روش‌های دسته‌بندی متون با تعداد زیادی خصیصه مواجه هستند. این تحقیق، دو فرایند و چهار روش جهت بهبود انتخاب خصیصه در حوزه متن پیشنهاد می‌دهد و به ارزیابی و مقایسه روش‌های پیشنهادی با یکدیگر و با سایر روش‌های موجود می‌پردازد.

در فرایند اول(SFSP)، ابتدا با استفاده از کل خصیصه ها دسته‌بند یادگیری شده و کارایی آن محاسبه می‌شود. سپس خصیصه‌هایی که باعث کاهش کارایی شده‌اند و در ادامه خصیصه‌هایی که تأثیری در کارایی دسته‌بند ندارد، حذف می‌گردند. در نهایت اگر حذف بیشتر خصیصه‌ها مطلوب باشد، حذف خصیصه‌ها با کارایی کمتر ولی قابل قبول ادامه می‌یابد. برای پیاده‌سازی فرایند SFSP دو روش پیشنهاد شده است که در روش اول(SGA) از الگوریتم ژنتیک و در روش دوم(SBPSO) از BPSO برای جستجوی فضای خصیصه‌ها استفاده شده است. برای ارزیابی زیرمجموعه خصیصه‌های انتخاب شده، از روش دسته‌بندی SVM در هر دو روش استفاده شده است. روش SGA، با کارایی ۰/۹٦۷٦ حدود ۹۸/۸۳ درصد خصیصه‌ها و روش SBPSO با کارایی ۰/۹٦۸٤ حدود ۸۵/۳٤ درصد خصیصه‌ها را از متون رویترز-۲۱۵۷۸ گونه ModApte حذف می‌نماید.

در فرایند پیشنهادی دوم،MFSP، برای فرار از بهینه محلی، در شرایطی که ممکن است GA و یا PSO در بهینه محلی گیر کنند، بعد از تعدادی تکرار که جمعیت تا حدی به سمت یک راه حل همگرا شد، خصیصه‌هایی که تا این لحظه به عنوان نویز شناخته شده‌اند دور ریخته می‌شوند و در فضای خصیصه‌های باقی مانده به دنبال حذف نویزهای بیشتر می‌گردیم. به‌این ترتیب، کرومزوم‌ها در GA و یا ذرات در PSO با توجه به بردار خصیصه‌های جدید، دوباره به صورت تصادفی مقداردهی می‌شوند. شرط پایان برای الگوریتم کاهش خصیصه‌ها در فرایند MFSP، آن است که در مرحله جدید قادر به افزایش کارایی نسبت به مرحله قبل نباشیم. پیاده‌سازی MFSP با استفاده از GA به عنوان روش جستجو در روش پیشنهادی سوم،MGA و با استفاده از BPSO به عنوان روش جستجو در روش پیشنهادی چهارم،MBPSO، انجام شده است. در این دو روش نیز، برای ارزیابی زیرمجموعه خصیصه‌های انتخاب شده، از روش دسته‌بندی SVM استفاده شده است. ارزیابی روش‌های MGA و MBPSO نشان داد که روش MGA، ۹۹/۰۵ درصد خصیصه‌ها را با کارایی ۰/۹۷۲ و روش MBPSO نیز ۹۸/۵ درصد خصیصه‌ها را با کارایی ۰/۹۸ حذف نمود. مقایسه روش‌های پیشنهادی با IG و CHI به عنوان دو نمونه از بهترین روشهای فیلتری، برتری روشهای پیشنهادی را نشان می‌دهد.

برای سنجش کارایی روش‌های پیشنهادی در سایر دامنه‌ها، روش SGA در انتخاب خصیصه هرزنامه‌های متنی و همچنین روشهای SGA و MBPSO در زمینه انتخاب ژن برای دسته‌بندی نسوج تومور به‌کار گرفته شد. مقایسه روش SGA با سایر روش‌های جداسازی هرزنامه‌ها، نشان داد که روش SGA به خوبی قابل مقایسه با بهترین روش‌های آماری موجود است. اعمال روش SGA و MBPSO بر روی مجموعه داده سرطان روده بزرگ برای جداسازی نسوج تومور نشان داد که این روش‌ها می‌توانند با انتخاب تعداد بسیار کم (۳ یا ٤) ژن کارایی دسته‌بند تومور را افزایش دهند. روش‌های پیشنهادی می‌توانند با شناساندن ژن‌های مهم در دسته‌بندی نسوج تومور، در سایر جنبه‌های پزشکی مفید واقع شوند.

در زمینه دسته‌بندی متون با روش‌های تکاملی، روش Liu و همکاران، بهبود داده شده به طوری که روش دسته‌بندی پیشنهادی، از نظر دقت، فراخوانی، F1 و صحت به ترتیب ۲، ۱۵، ۹ و ۲ درصد بهتر از روش Liu و همکارانش عمل کرده است.

کلمات کلیدی: دسته‌بندی متون، انتخاب خصیصه، روش‌های تکاملی، روش بهینه‌سازی گروه ذرات

فهرست مطالب

فهرست جداول

عنوان.....صفحه

فهرست اشکال

عنوان.....صفحه

فصل اول

کلیات

۱-۱- مقدمه

امروزه با توجه به گسترش روز افزون اسناد متنی الکترونیکی، توسعه روش‌های پردازش خودکار اسناد اهمیت زیادی پیدا کرده است.

یکی از کارهای بنیادی در مدیریت سندهای متنی، اختصاص اسناد به دسته‌های از پیش تعیین شده می‌باشد که دسته‌بندی(طبقه بندی) متون نامیده می‌شود[۱]. این عمل مستلزم فهم محتوای اسناد و دانش قبلی از حوزه‌های مربوطه است. به همین دلیل، دسته‌بندی اسناد متنی در گذشته توسط افراد متخصص انجام می‌شد. با توجه به‌این که رویکرد سنتی دسته‌بندی برای حجم زیاد اسناد غیر عملی است، در دهه اخیر تلاش‌های فراوانی در زمینه توسعه الگوریتم‌هایی برای دسته‌بندی خودکار متون صورت گرفته است. در این میان، رویکرد یادگیری ماشین[1] بسیار مورد توجه بوده است که در آن، با استفاده از مجموعه کوچکی از اسناد برچسب خورده (اسنادی که دسته آنها مشخص است) می‌توان دسته‌بندهایی را یادگرفت که به صورت خودکار به دسته‌بندی اسناد جدید بپردازد[۱]. مهمترین روش‌های استفاده شده در زمینه دسته‌بندی متون عبارتند از: ۱) روش بیزین ساده[2]، ۲) روش Roccio، ۳) روش K همسایه نزدیکتر[3]، ٤) روش‌های رگرسیون، ٥) درخت‌های تصمیم گیری، ٦) شبکه‌های عصبی، ۷) روش SVM[4] ، ۸)روش‌های بر اساس قاعده[5] و ۹) روش‌های تکاملی.

به طور کلی، روش‌های دسته‌بندی بر اساس خصیصه‌ها[6] شکل می گیرند و در حوزه متون، خصیصه‌ها همان کلمات هستند. از آنجا که تعداد کلمات در متن بسیار زیاد است، تعداد خصیصه‌ها و در نتیجه تعداد ابعاد فضای خصیصه‌ها بسیار زیاد می‌باشد و این مسئله می‌تواند برای بسیاری از

---

[1] Machine Learning
[2] Naïve Bayesian
[3] K Nearest Neighbor
[4] Support Vector Machines
[5] Decision Rule Classifier
[6] Features

الگوریتم‌های یادگیری ماشین مشکل‌ساز باشد. در نتیجه، کاهش فضای خصیصه‌ها به صورت خودکار و بدون از دست دادن دقت دسته‌بندی، یکی از اهداف مهم دسته‌بندی متون به شمار می‌آید.

۱-۲- اهداف و نتایج پژوهش

با توجه به بعد بالای فضای خصیصه‌ها در حوزه متون، استفاده از روش‌هایی به منظور حذف خصیصه‌های نا مربوط و خصیصه‌هایی که تأثیر چندانی در کارایی دسته‌بندی ندارند، لازم به نظر می‌رسد. همچنین، برخی خصیصه‌ها باعث ایجاد اغتشاش و کاهش کارایی دسته‌بندی می‌شوند. حذف چنین خصیصه‌هایی ضمن کاهش ابعاد فضای خصیصه‌ها، کارایی دسته‌بندی را افزایش می‌دهد.

دراین تحقیق با در نظر گرفتن سطح انتظار کاربر از انتخاب خصیصه، دو فرایند انتخاب خصیصه ارائه شده است. فرایند اول، SFSP[1]، شامل ٤ مرحله می‌باشد. در مرحله اول، کارایی دسته‌بند با استفاده از کل خصیصه‌ها و بدون انتخاب خصیصه محاسبه می‌شود. در مرحله دوم، خصیصه‌های نویز یعنی خصیصه‌هایی که حضورشان تأثیری در کارایی دسته‌بندی ندارد و یا آنهایی که باعث کاهش کارایی دسته‌بندی می‌شوند، حذف می‌گردند. به‌این ترتیب، ضمن کاهش خصیصه‌ها، کارایی دسته‌بند تا حد ممکن بالا می‌رود. در مرحله سوم، سیستم با حفظ کارایی اولیه(نتایج حاصل از مرحله اول) به حذف بیشتر خصیصه‌ها می‌پردازد. در صورتیکه کاربر مایل به کاهش بیشتر تعداد خصیصه‌ها باشد، سیستم وارد مرحله چهارم می‌شود و سعی می‌کند که با دقت کمتر ولی قابل قبولی نسبت به مرحله اول، حذف خصیصه‌ها را ادامه دهد.

به منظور پیاده‌سازی فرایند SFSP ابزاری برای جستجوی فضای خصیصه‌ها و انتخاب زیر مجموعه خصیصه بهینه و نیز ابزاری برای ارزیابی زیر مجموعه خصیصه‌ها نیاز است. از آنجا که پیدا کردن بهترین زیر مجموعه خصیصه یک مسأله NP می باشد[۲, ۳]، استفاده از روش‌های حل تقریبی برای یافتن زیر مجموعه خصیصه بهینه مناسب می‌باشد. در نتیجه در این پژوهش، دو روش برای پیاده‌سازی SFSP پیشنهاد شده است که در روش اول،SGA، الگوریتم ژنتیک[2](GA)[٤] و در روش دوم، SBPSO، نوع دودویی الگوریتم بهینه سازی گروه ذرات[3](PSO)[٥, ٦] برای جستجوی

---

[1] Sequentioal Feature Selection Process
[2] Genetic Algorithm
[3] Particle Swarm Optimization

زیرمجموعه خصیصه مناسب به‌کار رفته است. همچنین، از روش دسته‌بندی SVM برای دسته‌بندی متون با توجه به خصیصه‌های منتخب، و ارزیابی زیر مجموعه خصایص استفاده می‌شود.

حضور خصیصه‌های نویز در بردار خصیصه‌ها ممکن است موجب گمراه کردن ابزارهای جستجوی GA یا PSO شود و آنها را از ادامه جستجو برای حذف نویزهای بیشتر باز دارد. برای حل این مشکل، دراین تحقیق، فرایند دومی پیشنهاد شده است، MFSP[1] ، که در آن، در فواصل زمانی تعیین شده و یا در صورت ارضاء شرط خاصی، نویزهای تعیین شده تا کنون، دور ریخته شده و فضای خصیصه‌ها، به خصیصه‌های انتخاب شده کاهش می‌یابد. سپس حذف نویز در فضای خصیصه‌های جدید، ادامه می‌یابد. این عمل، یعنی کاهش بردار خصیصه‌ها بعد از گذشت چندین تکرار، تا زمانی ادامه می‌یابد که در مرحله جدید قادر به حذف خصیصه‌های بیشتر نباشیم.

برای پیاده سازی MFSP دو روش پیشنهاد شده است که در روش اول،MGA، از الگوریتم ژنتیک و در روش دوم، MBPSO، از نوع دودویی الگوریتم بهینه سازی گروه ذرات برای جستجوی زیرمجموعه خصیصه مناسب به کار رفته است. در این دو روش نیز، از روش دسته‌بندی SVM برای دسته‌بندی متون با توجه به خصیصه‌های منتخب، و ارزیابی زیر مجموعه خصایص استفاده می‌شود.

نتایج ارزیابی‌ها نشان می‌دهد، در کل می‌توان نتیجه گرفت، در صورتی‌که کم بودن تعداد خصیصه‌های انتخابی و بالا بودن معیار F1 از اهمیت یکسانی برخوردار باشند و همچنین در شرایطی که بالا بودن معیار F1 اهمیت بیشتری داشته باشد، روش MBPSO بهترین روش بوده و روش‌های MGA، SGA و SBPSO به ترتیب در رده‌های بعدی قرار دارند. در صورتی‌که کمتر بودن تعداد خصیصه‌های انتخاب شده از اهمیت بیشتری نسبت به بالا بودن معیار F1 برخوردار باشد، روش MGA بهترین بوده و سپس روش‌های MBPSO، SGA و SBPSO به ترتیب در رده‌های بعدی قرار دارند.

همچنین مقایسه روش‌های SGA و SBPSO و همچنین روش‌های MGA و MBPSO، نشان می‌دهد که GA وBPSO ، هر دو در قالب MFSP بهتر از SFSP عمل می‌کنند و میزان بهبود عملکرد BPSO در MFSP ، بیشتر از بهبود عملکرد GA در SFSP مشهود است.

---

[1] Multi start Feature Selection Process

مقایسه روش‌های پیشنهادی با روش‌های انتخاب خصیصه IG و CHI که دو تا از بهترین روش‌های انتخاب خصیصه فیلتری در حوزه متن می‌باشند، نشان داد که روش‌های پیشنهادی، نتایج بهتری را نسبت به روش‌های انتخاب خصیصه IG و CHI از خود نشان می‌دهند. در ضمن، روش IG بعد از حذف حدود ۸۵ درصد خصیصه‌ها و روش CHI بعد از حذف حدود ۹۰ درصد خصیصه‌ها، دیگر قادر به حفظ کارایی اولیه نمی‌باشند.

به منظور سنجش کارایی روش‌های پیشنهادی در سایر دامنه‌ها، روش SGA را در راستای انتخاب خصیصه در جداسازی هرزنامه‌های متنی بکار برده‌ایم. نتایج اعمال این روش روی مجموعه استاندارد LingSpam[7] نشان داده است که روش پیشنهادی علاوه بر جداسازی هرزنامه‌ها با دقت ۱ و فراخوانی ۰/۹۹۸، قدرت کاهش خصایص تا حدود ۱۲/۷٤٪ تعداد اولیه را دارد. همچنین، مقایسه روش پیشنهادی با سایر روش‌های جداسازی هرزنامه‌ها، نشان داده است که روش پیشنهادی به خوبی قابل مقایسه با بهترین روش‌های آماری موجود است. همچنین روش‌های SGA و MBPSO را به منظور انتخاب حداقل ژن‌های لازم برای دسته‌بندی درست نسوج تومور به کار برده‌ایم. روش‌های پیشنهادی بر روی مجموعه داده سرطان روده بزرگ آزمایش شده اند. نتایج نشان داده‌اند که روش‌های SGA و MBPSO برای انتخاب ژن و دسته‌بندی نسوج تومور مناسب بوده و می‌توانند با انتخاب تعداد بسیار کم (به طور متوسط ۳) ژن، کارایی دسته‌بندی را افزایش دهند. همچنین روش‌های پیشنهادی می‌توانند با شناساندن ژن‌های مهم در دسته‌بندی نسوج تومور در سایر جنبه‌های پزشکی و درمان مفید واقع شوند.

آخرین کار انجام شده دراین پژوهش، استفاده از روش‌های تکاملی برای یادگیری دسته‌بند متون است. به این منظور، روش Liu و همکارانش[۸] به عنوان یکی از روش‌های مستقل از ساختار متن بر روی داده‌های استاندارد رویترز آزمون شده و با روش‌های آماری مقایسه شده است. در این پژوهش، تابع ارزیابی جدیدی تعریف شده است که موجب بهبود کارایی روش Liu و همکارانش، از نظر دقت: ۲٪، از نظر فراخوانی ۱۵٪، از نظر F1 ۹٪ و از نظر صحت ۲٪ بهتر از روش Liu و همکارانش عمل می‌کند. البته روش بهبود یافته دسته‌بندی نسبت به روش‌های آماری موجود ضعیفتر می‌باشد.

۱-۳- مروری بر ساختار پایان نامه

این پایان نامه شامل ۹ فصل است. فصل دوم به بررسی تاریخچه پژوهش در زمینه انتخاب خصیصه و دسته‌بندی متون پرداخته و روش‌های مطرح آماری و همچنین روش‌های برگرفته از طبیعت در زمینه انتخاب خصیصه در حوزه متن و غیر متن را شرح می‌دهد. در نهایت نیز انواع روش‌های مطرح دسته‌بندی متون توضیح داده می‌شود.

فصل سوم درباره متون و مفاهیم پایه مربوط به آن از جمله خصیصه، روش‌های نمایش اسناد متنی، مجموعه‌های آموزش و آزمون، مجموعه داده‌های استاندارد ومعیارهای ارزیابی کارایی دسته‌بندهای متون، صحبت می‌کند. در نهایت نیز، روش دسته‌بندی SVM به دلیل استفاده از آن در این پژوهش، به صورت دقیقتر توضیح داده می‌شود.

فصل چهارم به توصیف روش‌های محاسباتی برگرفته از طبیعت از جمله الگوریتم‌های تکاملی و روش بهینه سازی گروه ذرات، انواع آنها و مفاهیم مربوطه می‌پردازد.

در فصل پنجم دو فرایند پیشنهادی انتخاب خصیصه توضیح داده شده و هر یک از آنها با استفاده از دو روش GA و PSO پیاده سازی می‌شوند.

نتایج حاصل از آزمون روش‌های پیشنهادی انتخاب خصیصه روی داده‌های رویترز، در فصل ششم آورده شده است. همچنین روش‌های پیشنهادی با بهترین روش‌های آماری انتخاب خصیصه متون مقایسه می‌شوند.

فصل هفتم به بررسی شایستگی روش‌های پیشنهادی در سایر دامنه‌ها اختصاص دارد. بخش اول فصل هفتم به بحث جداسازی هرزنامه‌های متنی اختصاص دارد. دراین بخش، ابتدا مقدمه‌ای بر هرزنامه‌ها و روش‌های جداسازی آنها گفته می‌شود. سپس پژوهش‌های مرتبط در زمینه روش‌های جداسازی هرزنامه مطرح می‌شود. در ادامه، تابع ارزیابی و سایر شرایط محیطی استفاده شده برای مسأله انتخاب خصیصه در حوزه هرزنامه‌های متنی، توضیح داده می‌شود. سپس، نتایج ارزیابی کارایی روش SGA در مقایسه با روش‌های موجود مطرح می‌گردد.

بخش دوم فصل هفتم نیز به بحث دسته‌بندی نسوج تومور اختصاص دارد. ابتدا مقدمه‌ای بر دسته‌بندی نسوج تومور و روش‌های نوین دراین زمینه گفته می‌شود. سپس پژوهش‌های مرتبط در زمینه دسته‌بندی نسوج تومور مطرح می‌گردد. در ادامه، تابع ارزیابی و سایر شرایط محیطی استفاده