
بسم الله الرحمن الرحیم

دانشگاه کردستان

دانشکده مهندسی

گروه مهندسی کامپیوتر و فناوری اطلاعات

عنوان:

بهبود طبقه‌بندی متن با استفاده از روش‌های ترکیب

پژوهشگر:

علی دانش

اساتید راهنما:

دکتر فردین اخلاقیان

دکتر بهروز مینایی

پایان‌نامه کارشناسی ارشد رشته مهندسی کامپیوتر گرایش هوش مصنوعی

اسفند ماه ۱۳۹۱

تعهد نامه

اینجانب علی دانش دانشجوی کارشناسی ارشد رشته مهندسی کامپیوتر گرایش هوش مصنوعی دانشگاه کردستان، دانشکده مهندسی گروه کامپیوتر و فناوری اطلاعات تعهد می‌نمایم که محتوای این پایان‌نامه نتیجه تلاش و تحقیقات خود بوده و از جایی کپی‌برداری نشده و به پایان رسانیدن آن نتیجه تلاش و مطالعات مستمر اینجانب و راهنمایی و مشاوره اساتید بوده است.

با تقدیم احترام

علی دانش

۱۳۹۱/۱۲/۲۳

دانشگاه کردستان

دانشکده مهندسی

گروه مهندسی کامپیوتر و فناوری اطلاعات

پایان‌نامه کارشناسی ارشد رشته مهندسی کامپیوتر گرایش هوش مصنوعی

عنوان:

بهبود طبقه‌بندی متن با استفاده از روش‌های ترکیب

پژوهشگر:

علی دانش

در تاریخ ۱۳۹۱/۱۲/۲۳ توسط کمیته تخصصی و هیأت داوران زیر مورد بررسی قرار گرفت و با نمره ..... و درجه ..... به تصویب رسید.

هیأت داوران	نام و نام خانوادگی	مرتبه علمی	امضاء
۱- استاد راهنما	دکتر فردین اخلاقیان	استادیار	
۲- استاد داور خارجی	دکتر محمد رزاقی	استادیار	
۳- استاد داور داخلی	دکتر پرهام مرادی	استادیار	

مهر و امضاء گروه

مهر و امضاء معاون پژوهشی و تحصیلات تکمیلی دانشکده

سپاسگذاری و قدردانی

شکر خداوند متعال را به جای آورده که توفیق نصیب من کرد تا این پایان‌نامه را به سرانجام برسانم.

در آغاز لازم می‌دانم از پدر، مادر و همسرم به دلیل حمایت‌ها و تشویق‌هایشان که باعث رشد و پیشرفت اینجانب گردید کمال تشکر را داشته باشم.

همچنین از تمامی اساتید و دوستانی که مرا در انجام مراحل مختلف این پایان‌نامه یاری نمودند، بخصوص اساتید راهنمای گرامی جناب آقایان دکتر اخلاقیان و دکتر مینایی که در تمام مراحل انجام این پایان‌نامه پشتیبان بنده بودند، تشکر ویژه‌ای دارم.

در پایان از جناب آقای دکتر مرادی و جناب آقای دکتر رزاقی که برای داوری این پایان‌نامه قبول زحمت نمودند و وقت گران‌بهای خود را در اختیار اینجانب قرار دادند، تشکر می‌کنم.

با تقدیم احترام

علی دانش

چکیده

با توجه به گسترش اینترنت و افزایش چشمگیر حجم مطالب متنی، ابزارها و روش‌های خودکار برای مدیریت اسناد و مطالب متنی، مورد توجه بسیار زیادی قرار گرفته‌اند. از جمله‌ی این ابزارها می‌توان به طبقه‌بند متن اشاره کرد که می‌تواند در این زمینه بسیار مؤثر باشد. این ابزار کاربردهای فراوانی در سیستم‌های بررسی متون مانند موتورهای جستجو، سیستم‌های فیلترینگ، تشخیص هرزنامه‌ها، کتابخانه‌های دیجیتال و سایر سیستم‌های مشابه دارد.

از آن جایی که صحت طبقه‌بندی متن یکی از مهمترین شاخص‌های موفقیت این سیستم‌ها به شمار می‌آید، از اهداف اساسی این پایان نامه افزایش میزان صحت طبقه‌بندی متون می‌باشد. با توجه به میزان کارهای انجام شده قبلی، افزایش کارایی طبقه‌بندهای منفرد مشکل می‌باشد، لذا رهیافت ما برای رسیدن به این هدف استفاده و بهبود روش‌های ترکیب طبقه‌بندها است.

در این پایان‌نامه برای بهبود صحت طبقه‌بندی متن، و بر مبنای روش ترکیب رأی‌گیری وزن‌دار، دو رهیافت جدید برای وزن‌دهی طبقه‌ها و طبقه‌بندها پیشنهاد شده است. رهیافت اول مبتنی بر در نظر گرفتن وزن مستقل برای هر طبقه و هر طبقه‌بند است. رهیافت دوم تعمیم رهیافت اول است. بدین شکل که برای جواب مثبت یا منفی هر طبقه‌بند در مورد هر طبقه وزن مستقلی در نظر گرفته می‌شود. برای محاسبه‌ی اوزان در هر دو رهیافت، علاوه بر الگوریتم ژنتیک، معادله تجربی خاصی هم پیشنهاد گردیده است که در زمان بسیار کمتری نسبت به الگوریتم ژنتیک اجرا می‌گردد. نتایج طبقه‌بندی بر مبنای محاسبه اوزان با استفاده از معادله‌ی پیشنهادی، کاملاً با نتایج استفاده از الگوریتم ژنتیک قابل مقایسه و حتی گاهی بهتر هم می‌باشد.

آزمایش‌ها با استفاده از طبقه‌بندهای رُکیو، نزدیک‌ترین همسایه و بیز، و سه روش انتخاب ویژگی شامل اطلاعات متقابل، خی۲ و MCFS انجام شده است. نتایج تجربی حاصل از اعمال الگوریتم‌های ترکیب پیشنهادی بر روی مجموعه داده‌های آموزشی رایج و مقایسه با نتایج حاصل از سایر روش‌های ترکیب طبقه‌بندها، مانند رأی‌گیری وزن‌دار، عملگر میانگین وزن‌دار رتبه یافته و روش قالب تصمیم، نشان می‌دهد که رهیافت‌های پیشنهادی دقت طبقه‌بندی را بنحو چشمگیری افزایش داده‌اند. این نتایج از آزمایش بر روی چهار مجموعه داده‌های آموزشی متفاوت و رایج بدست آمده است.

کلمات کلیدی: طبقه‌بند، طبقه‌بند متن، ترکیب طبقه‌بندها، رأی‌گیری وزن‌دار

فهرست مطالب

عنوان ..... صفحه

فهرست شکل‌ها

عنوان ..... صفحه

فهرست جدول‌ها

عنوان ..... صفحه

فصل ۱. مقدمه

۱-۱ انگیزه‌ی تحقیق

با توجه به گسترش اینترنت و افزایش حجم مطالب متنی، نیاز به روش‌های مدیریت متون و مطالب بطور چشمگیری افزایش یافته است [۱]. طبقه‌بند متن یکی از ابزارهایی مطرح در هوش مصنوعی است که می‌تواند در این زمینه مؤثر باشد. طبقه‌بندی متن کاربردهای فراوانی دارد و از جمله کاربردهای آن می‌توان به موتورهای جستجو [۲-۴]، سیستمهای فیلترینگ [۳]، سیستم‌های دسته بندی متون [۵-۸]، کتابخانه‌های دیجیتال [۶]، دسته‌بندی خبرهای دریافت شده در گروه‌های خبری [۴, ۹]، تشخیص هرزنامه‌ها در سیستم‌های پست الکترونیک[۱۰, ۱۱] و بسیاری کاربردهای دیگر اشاره کرد.

۱-۲ بیان مسأله‌ی تحقیق

با توجه به افزایش چشمگیر حجم مطالب متنی در سال‌های اخیر، و با توجه به کاربردهای مهم طبقه‌بندی متن، نیاز به طبقه‌بندهای متنی که دارای صحت طبقه‌بندی بالایی باشند اهمیت زیادی پیدا کرده است. خطا در طبقه‌بندی متن در بعضی از کاربردها هزینه‌های زیادی را ممکن است داشته باشد. برای نمونه اگر در یک سیستم پست الکترونیک به اشتباه یک نامه‌ی مهم به عنوان هرزنامه تشخیص

داده شود و کاربر متوجه نامه‌ی مذکور نشود ممکن است هزینه‌ی گزافی برای کاربر یا شخص ارسال‌کننده داشته باشد. با توجه به پیچیدگی حوزه‌ی پردازش متن، هنوز بهبود میزان صحت طبقه‌بندی متن از مسائل مطرح در این حوزه می‌باشد.

۱-۳ اهداف و رویکردها

با توجه به مطالب ذکر شده درباره‌ی اهمیت صحت طبقه‌بندی متن، مهمترین هدف در این پایان‌نامه رسیدن به میزان صحت بیشتر در طبقه‌بندی متن می‌باشد. برای این منظور از روش‌های ترکیب در طبقه‌بندی متن استفاده شده است. روش‌های رأی‌گیری، عملگر میانگین وزن‌دار رتبه یافته و روش قالب تصمیم از جمله روش‌های استفاده شده برای رسیدن به هدف اصلی در این پایان‌نامه است. همچنین تاثیر نحوه انتخاب ویژگی بر ترکیب طبقه‌بندها بررسی شده است.

۱-۴ دستاوردهای پایان‌نامه

در راستای رسیدن به اهداف پایان‌نامه، دستاوردهای مهمی حاصل شده است که در ادامه بصورت مختصر هر یک از این دستاوردها شرح داده می‌شوند و در فصل‌های بعد به تشریح جزئیات آنها پرداخته خواهد شد.

همانطور که در بخش‌های قبل نیز اشاره شد، اصلی‌ترین هدف این پایان‌نامه رسیدن به میزان صحت بیشتر در طبقه‌بندی متن است. در این پایان‌نامه، ما به بررسی روش‌های مختلف طبقه‌بندی متن پرداخته‌ایم. همچنین روش‌های مختلف ترکیب طبقه‌بندها نیز بررسی شده است. مبنای روش‌های پیشنهادی ما در این پایان‌نامه، روش رأی‌گیری وزن‌دار است. برای وزن‌دهی دو رهیافت جدید ارائه داده‌ایم. رهیافت اول مبتنی بر در نظر گرفتن وزن مستقل برای هر طبقه از هر طبقه‌بند است. رهیافت دوم تعمیم رهیافت اول است. بدین شکل که برای جواب مثبت یا منفی هر طبقه‌بند و برای هر طبقه

وزن مستقلی در نظر گرفته شود. پس از پیاده‌سازی چندین طبقه‌بند، با در نظر گرفتن اوزان تطبیقی وابسته به طبقه، طبقه‌بند و همچنین با توجه به خروجی آن در روش‌های ترکیب طبقه‌بندها، به نتایج بهتری در صحت طبقه‌بندی متن رسیدیم.

دستاورد دیگری که ترکیب طبقه‌بندها به همراه دارد، افزایش پایداری طبقه‌بندی می‌باشد. زیرا چندین طبقه‌بند به صورت موازی با هم عمل طبقه‌بندی را انجام می‌دهند و در صورتی که بعضی از آنها دچار مشکل شوند، سایر طبقه‌بندها می‌توانند یک جواب مناسب را تولید کنند و سیستم متوقف نخواهد شد. از طرف دیگر چون طبقه‌بندها بصورت موازی می‌توانند اجرا شود، با استفاده از چندین پردازشگر و با استفاده از تکنیک‌ها پردازش موازی می‌توان در زمان طبقه‌بندی کل سیستم صرفه‌جویی قابل ملاحظه‌ای کرد [۱۲].

۱-۵ ساختار پایان‌نامه

این پایان‌نامه در شش فصل تنظیم گردیده است که رئوس مطالب و موضوعات اصلی هر فصل در ذیل بطور خلاصه بیان گردیده است:

- **فصل اول** به انگیزه، اهمیت و شرح موضوع تحقیق می‌پردازد. همچنین ساختار کلی پایان‌نامه و خلاصه مطالب ارائه شده در این تحقیق نیز بیان می‌گردد.
- **فصل دوم** مطالب پیش‌زمینه‌ای را که برای درک روش‌های ارائه شده در پایان‌نامه ضروری است، بیان می‌کند. از آنجاییکه هدف این پایان نامه بهبود طبقه‌بندی متن است، لازم است مقدمه‌ای درباره‌ی طبقه‌بندی و روش‌های مختلف آن ارائه شود. چون در طبقه‌بندی متن از

روش‌های نزدیک‌ترین همسایه، بیز ساده و رُکیو[1] بیشتر استفاده می‌شود، و در آزمایشات این پایان‌نامه از این سه روش استفاده شده است، مروری بر آنها انجام داده‌ایم.

برای استخراج ویژگی‌ها از متن و انتخاب ویژگی‌های مهم، الگوریتم‌های مختلفی در مقالات پیشنهاد شده است. ما از دو روش معروف اطلاعات متقابل[2][۱۳] و خی[3]۲[۱] و همچنین از یک روش جدید به نام MCFS [۱۴] استفاده کرده‌ایم. در ادامه‌ی این فصل روش‌های رُکیو و بیز ساده[۱۵] و نزدیک‌ترین همسایه[۱] برای طبقه‌بندی متن بررسی شده است.

- **فصل سوم** از آنجاییکه رهیافت ما در این پایان‌نامه استفاده از روش‌های ترکیب طبقه‌بندها است، این فصل را به مروری بر نحوه‌ی ترکیب طبقه‌بندها و مطالب پیش‌زمینه‌ای که برای درک روش‌های ترکیب طبقه‌بندها ضروری است، اختصاص داده‌ایم [۱۶, ۱۷]. روش‌هایی که مرور خواهند شد عبارتند از رأی اکثریت، رأی اکثریت وزن‌دار، ترکیب بیز ساده، روش‌های طبقه-هوشیار[4] و روش‌های طبقه-بی‌تفاوت[5].
- **فصل چهارم** به بررسی راهکارهای پیشنهادی برای بهبود صحت طبقه‌بندی متن می‌پردازد. رهیافت ما استفاده از روش‌های ترکیب طبقه‌بندها است. ما در این پایان‌نامه دو رهیافت جدید در روش ترکیب با استفاده از رأی‌گیری وزن‌دار ارائه داده‌ایم. رهیافت اول مبتنی بر در نظر گرفتن وزن مستقل برای هر طبقه از هر طبقه‌بند است. رهیافت دوم تعمیم رهیافت اول است. بدین شکل که برای جواب مثبت یا منفی هر طبقه‌بند و برای هر طبقه وزن مستقلی در نظر

[1] Rocchio
[2] Mutual Information
[3] Chi-2
[4] Class-Conscious
[5] Class-Indifferent

گرفته شود. برای محاسبه وزن در هر دو رهیافت، معادله‌ی تجربی نیز پیشنهاد شده است. البته از سایر روش‌های هوشمند مانند الگوریتم ژنتیک نیز برای محاسبه وزن‌ها بهره برده شده است.

- **فصل پنجم** نتایج حاصل از پیاده‌سازی روش‌های طبقه‌بندی متن که در فصل سوم و چهارم بیان شد را ارائه کرده‌ایم. با تقسیم داده‌های آموزشی به سه دسته و استفاده از سه روش مختلف انتخاب ویژگی (اطلاعات متقابل، خی۲ و MCFS)، و سه روش طبقه‌بندی متن (رُکیو، بیز ساده و نزدیک‌ترین همسایه)، مجموعه متنوعی از طبقه‌بندهای پایه را جهت انجام عملیات ترکیب آماده نمودیم. سپس با استفاده از روش‌های رأی‌گیری و میانگین وزن‌دار رتبه‌یافته و قالب تصمیم، نتایج حاصل از طبقه‌بندهای متن را با هم ترکیب کردیم و به نتایج بهتری رسیدیم. در ادامه روش‌های پیشنهادی که دو رهیافت جدید برای محاسبه‌ی وزن‌ها در روش رأی‌گیری بود را بکار بردیم. برای محاسبه وزن در هر دو رهیافت، هم از معادله‌ی تجربی پیشنهاد شده در فصل چهارم استفاده شده و هم از الگوریتم ژنتیک بهره برده شده است. نتایج تجربی نشان می‌دهد که رهیافت پیشنهاد شده، دقت طبقه‌بندی را نسبت به سایر روش‌های رأی‌گیری و همچنین سایر روش‌های ترکیب در مجموع افزایش داده است. این نتایج از آزمایش بر روی چهار داده آموزشی متفاوت بدست آمده است.
- **فصل ششم** در این فصل راهکارهای ارائه شده در این پایان‌نامه بصورت خلاصه بررسی شده و در ادامه پیشنهاداتی برای توسعه و بهینه‌سازی هرچه بیشتر این سیستم ارائه می‌شود.

فصل ۲. مروری بر طبقه‌بندهای متن

طبقه‌بندها از مباحث مطرح در بازشناخت الگو می‌باشند. روش‌های بسیار زیادی برای طبقه‌بندی ارائه شده است. هدف اصلی از ارائه‌ی روش‌های مختلف، افزایش صحت طبقه‌بندی می‌باشد. یک ایده‌ی مهم که در این زمینه مطرح می‌باشد ترکیب چند طبقه‌بند برای رسیدن به نتایج بهتر می‌باشد. اما در ابتدا باید با نحوه کار طبقه‌بندها آشنا شد، لذا در این فصل به مرور اجمالی طبقه‌بندهای پرکاربرد می‌پردازیم.

۲-۱ تفاوت طبقه‌بندی[1] و خوشه‌یابی[2]

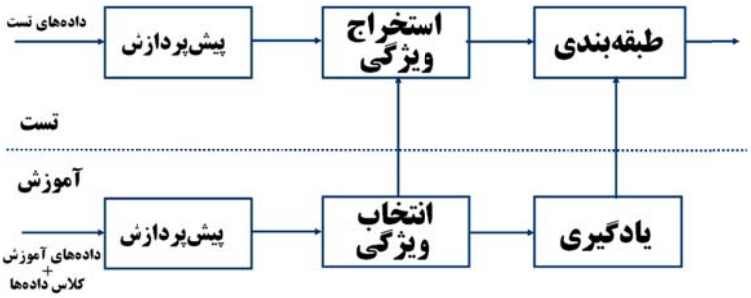
شکل ۲-۱: مقایسه‌ی طبقه‌بندی با خوشه‌یابی

[1] Classification
[2] Clustering

در طبقه‌بندی، طبقات از قبل شناخته‌شده هستند. هدف یک طبقه‌بند این است که تشخیص دهد ورودی به کدام طبقه تعلق دارد. طبقه‌بندها به صورت با سرپرست[1] آموزش داده می‌شوند. اما در خوشه‌یابی، طبقه‌ها لزوما از قبل شناخته‌شده نیستند و الگوریتم خوشه‌یابی با توجه به داده‌های ورودی، تعداد طبقه‌ها و نحوه‌ی طبقه‌بندی را تشخیص می‌دهد. در واقع می‌توان این الگوریتم‌ها را بدون سرپرست[2] نامید. تفاوت طبقه‌بندی و خوشه‌یابی در شکل ۲-۱ نمایش داده شده است.

۲-۲ بخش‌های مهم طبقه‌بند

در شکل ۲-۲ بخش‌های مهم یک طبقه‌بند را مشاهده می‌کنید.



شکل ۲-۲: بخش‌های مختلف یک طبقه‌بند

پیش پردازش[3]: این بخش داده‌های ورودی را به شکلی که برای طبقه‌بندی مناسب باشد بازنمایی می‌کند [۴].

استخراج ویژگی[4]: در این مرحله، بخشی از اطلاعات که برای طبقه‌بندی مهم نیست، حذف شده و اطلاعات مفید برای طبقه‌بندی تحت عنوان ویژگی استخراج می‌شود. این کار باعث می‌شود تا ضمن حفظ کارایی طبقه‌بند، پیچیدگی محاسباتی کاهش یابد [۴, ۱۸].

[1] Supervised
[2] Unsupervised
[3] Preprocessing
[4] Feature Extraction