تاریخ دریافت مقاله: ۹۶/۱۰/۰۵
تاریخ پذیرش مقاله: ۹۷/۰۶/۰۷

بهبود الگوریتم فازی C-Means با الگوریتم ژنتیک برای انتخاب ویژگی‌ها در دسته‌بندی اسناد متنی

ندا محمودی جاریحان
مربی، گروه مهندسی کامپیوتر، واحد ارومیه، دانشگاه آزاد اسلامی، ارومیه، ایران
پست الکترونیکی:nmahmoodi510@gmail.com

فرهاد سلیمانیان قره چپق*
استادیار، گروه مهندسی کامپیوتر، واحد ارومیه، دانشگاه آزاد اسلامی، ارومیه، ایران
پست الکترونیکی: farhad@iaurmia.ac.ir

چکیده

افزایش روزافزون مستندات الکترونیکی در وب، لزوم دسته‌بندی آنان در دسته‌های مختلف را نشان می‌دهد. با توجه به حجم و دامنه وسیع اسناد متنی که به‌طور قابل توجهی از طریق محیط‌های برخط و سایر منابع قابل دسترسی می‌باشند، در صورت عدم دسته‌بندی مناسب، عمل بازیابی و پردازش اسناد متنی دسته‌بندی نشده با مشکلات زیادی مواجه می‌گردد. این نیاز منجر به ایجاد روش‌های نوین برای دسته‌بندی اطلاعات شده است. دسته‌بندی، تخصیص اسناد متنی یا ویژگی‌ها به یک یا چندین دسته است، به‌طوری که اسناد متنی با توجه به موضوعات یا میزان مشابهت ویژگی‌ها می‌توانند دسته‌بندی گردند. در ارائه روش‌های دسته‌بندی، استخراج و انتخاب ویژگی‌های کلیدی اسناد متنی از اهمیت بالایی برخودار می‌باشد. در این مقاله روشی براساس بهبود الگوریتم فازی C-Means با الگوریتم ژنتیک برای انتخاب ویژگی‌ها در دسته‌بندی اسناد متنی ارائه شده است که در روش پیشنهادی انتخاب ویژگی‌های کلیدی متون از طریق الگوریتم فازی C-Means انجام می‌شود و این ویژگی‌ها به الگوریتم ژنتیک جهت بهبود در دسته‌بندی ارسال می‌گردند. روش پیشنهادی بر روی سه مجموعه‌داده مختلف Reuters21578, WEBKB, CADE 12 و بر اساس معیارهای ارزیابی مختلفی مورد آزمایش و ارزیابی قرار گرفته است. مقایسه نتایج روش پیشنهادی با سایر روش‌های مطرح در دسته‌بندی متون نشان می‌دهد که روش پیشنهادی عملکرد بهینه‌ای را در دسته‌بندی اسناد متنی دارد.

واژه‌های کلیدی: دسته‌بندی اسناد متنی، بهینه سازی، انتخاب ویژگی، الگوریتم فازی C-Means، الگوریتم ژنتیک.

۱- مقدمه

رشد چشمگیر منابع اطلاعاتی، فناوری‌های جدید و افزایش حجم پایگاه‌داده‌های الکترونیکی نیاز به دسته‌بندی متون را بیش از پیش آشکار می‌سازد. از آنجایی که حجم اطلاعات الکترونیکی و برخط روز به روز بیشتر می‌شود، دسترسی سریع و صحیح به منابع مهم، یکی از دغدغه‌های استفاده از این منبع اطلاعاتی بزرگ است. بسیاری از داده‌ها

* نویسندهٔ مسئول

در کاربردهای مختلف از جمله در اسناد متنی الکترونیکی، صفحـات وب، ایمیل‌هـا، اخبـار و مقالات، نیـاز به ابزار متن‌کاوی جهت دسته‌بندی خودکار متون را بیش از گذشته مبرم کرده اسـت. بدیهی است دسـته‌بندی اسناد موجب تسـهیل در روند جستجو و بازیابی اطلاعات می‌شوند. از جمله زمینه‌های کاربردی دسته‌بندی می‌توان به متن کاوی، پیدا کردن مشـابهت اسناد، ابهام زدایی، بازیابی اطلاعات، روش‌سازی و دسته‌بندی متن اشاره نمود. در تعریف کلی دسته‌بندی متون تخصیص اسناد متنی یا ویژگی‌ها (کلمات) به یک یا چندین دسـته اسـت، به‌طوری که اسناد متنی با توجه به موضوعات یا میزان مشابهت ویژگی‌ها می‌توانند دسته‌بندی گردند. دسته‌بندی متون تمام فعالیت‌هایی را که به نوعی به دنبال کسب دانش از متن هستند شامل می‌گردد و در سـطوح بالا زیرمجموعه‌ای از متن‌کاوی و داده‌کاوی اسـت. امروزه، روش‌های زیادی برای دسـته‌بندی اسناد متنی خودکار خصوصا روش‌های یادگیری ماشین وجود دارند.

دسته‌بندی متن به روش‌های یادگیری نظارت شده دلالت دارد، به‌طوری که این روش‌ها از نمونه‌های برچسب‌گذاری شده و از پیش تعریف شده در قسمت آموزش بهره برده و از اطلاعات به‌دست آمده در نمونه‌های آزمایشی در قسمت آزمایش سیستم برای دسته‌بندی استفاده می‌کنند. یادگیری نظارت شده یکی از انواع روش‌های یادگیری ماشین است که در آن ورودی و خروجی مشـخص بوده و به عبارت بهتر ناظری در این روش یادگیری وجود دارد که اطلاعات را برای یادگیرنده فراهم می‌نماید و سیسـتم با استفاده از این اطلاعات سـعی در پیداکردن سـاختاری مشخص در نمونه‌های داده دارد.

روش‌های دسته‌بندی متون مختلفی از جمله روش‌های آماری برای دسته‌بندی خودکار اسناد متنی وجود دارند. از جمله آن‌ها روش‌های k نزدیک‌ترین همسایه [۱]، الگوریتم ژنتیک[۲] و ماشـین بردار پشتیبان[۵] را می‌توان نام برد. در اکثر روش‌هایی که برای دسته‌بندی متون ارائه شده‌اند اسـتخراج ویژگی‌های کلیدی از اهمیـت بالایی برخوردار اسـت و براسـاس این ویژگی‌ها دسـته‌بندی متون انجام می‌گیرد. ثابت شده که تنها ۳۳درصد کلمات در یک متن مفید هستند و می‌توان از آن‌ها برای استخراج اطلاعات استفاده نمود[٦]. اکثر کلمات موجود در متون برای رساندن هدف یک متن اسـتفاده می‌شوند و گاهی اوقات تکراری هستند. به‌طـور کلی هدف از انتخاب ویژگی‌هـای متون کم کردن حجم داده‌ها، تسـریع در عملیات دسته‌بندی متون، کاهش زمان لازم برای آموزش، کاهش زمان محاسباتی و افزایش سـرعت عملکرد روش پیشـنهادی در دسـته‌بندی متون می‌باشـد. از این‌رو در این مقاله روشـی براساس بهبود الگوریتـم فازی C-Means با الگوریتم ژنتیک برای انتخاب ویژگی‌ها در دسـته‌بندی اسـناد متنی ارائه شده است. در روش پیشـنهادی ویژگی‌های کلیدی و مرتبط با موضوع اصلی متـون از طریق الگوریتم فـازی C-Means انتخاب خواهد شـد. سـپس این ویژگی‌ها به‌عنـوان ورودی‌های الگوریتم ژنتیک در نظر گرفته می‌شود. با توجه به عملکرد الگوریتـم ژنتیک که یک الگوریتم بهینه‌سـازی و مبتنی بر تکرار اسـت و برای جسـتجو و یادگیری ماشین استفاده می‌شـود[۷]، بهبود در عملکرد الگوریتـم فازی C-Means برای دسته‌بندی متون ایجاد خواهد شد.

ساختار این مقاله به‌شـرح زیر ساماندهی شده است: در بخـش ۲ کارهای قبلی انجام شـده در این حوزه بیان شده اسـت. در بخش ۳ روش پیشنهادی به صورت کامل ارائه شده است. در بخش ٤ روش پیشنهادی به‌طور کامل بررسـی و ارزیابی می‌شـود و در نهایـت در بخش ٥ به نتیجه‌گیری و کارهای آینده اشاره شده است.

۲- مروری برکارهای گذشته

در سال‌های اخیر، حوزه دسته‌بندی متون به‌دلیل رشد زیاد پایگاه‌داده‌ها، اطلاعات در صفحات وب و حجم بالای داده‌های متنی در فضای اینترنت اهمیت ویژه‌ای پیدا کرده است و در این حوزه روش‌های مختلفی توسط پژوهشگران

ارائه گردیده است که در ذیل به تعدادی از آن‌ها اشاره شده است.

بینا و رهگذر در [۳] دسته‌بندی خودکار متون فارسی با استفاده از الگوریتم یادگیری ماشینی، k نزدیک‌ترین همسایه را پیاده‌سازی کردند که نتایج دسته‌بندی خودکار متون فارسی، با استفاده از معیارهای شاخص‌گذاری همایند n تایی و حذف کلمه نهایی‌ها و برای دسته‌بندی متون از الگوریتم یادگیری ماشینی و k نزدیک‌ترین همسایه استفاده شده است. در نهایت به منظور ارزیابی و مقایسه نتایج، دو معیار دقت و یادآوری برای هر روش شاخص‌گذاری نیز محاسبه کرده‌اند. نتایج به‌دست آمده نشان داد که بهترین روش شاخص‌گذاری متون فارسی همایند n تایی می‌باشد و حذف کلمه نهایی‌ها نتایج را اندکی بهبود می‌بخشد.

موضوع برچسب زنی موضوعی متون فارسی، با استفاده از الگوریتم بردار فاصله اطلاعات را دسته‌بندی کرده و پیاده سازی شده است. نعمتی و بصیری در [٤] به ارائه روشی برای استخراج کلمات کلیدی با استفاده از معیار tf-idf پرداخته‌اند و می‌توان با در نظر گرفتن احتمال وقوع کلمات و نیز فاکتور، تمرکز کلمات مناسب برای دسته‌بندی متون را شناسایی نمود. همچنین الگوریتمی بر پایهٔ فاصلهٔ بین بردارهای بسامد کلمات کلیدی ارائه دادند و مشاهده شد که هرس دستی کلمات غیر کلیدی که به اشتباه در زمره کلمات کلیدی قرار گرفته‌اند، سبب ایجاد تاثیر بسزایی در افزایش معیار دقت و بازخوانی در ارزیابی می‌گردد که کار هرس بر پایه دانش زبان شناختی و معنی شناختی انجام شد.

در[۸] از یادگیری با روش سنتی بیز ساده با برچسب‌گذاری چندگانه برای دسته‌بندی متون استفاده شده است. در ابتدا، روش‌های استخراج ویژگی‌ها برمبنای روش تحلیل مولفه‌های اصلی به حذف ویژگی‌های بی‌ربط و تکراری و زاید می‌پردازد. سپس در روش‌های انتخاب خصیصه برای افزایش کارایی این روش از الگوریتم ژنتیک استفاده شده تا زیرمجموعه‌ای از ویژگی‌هایی که بیشترین مشابهت را دارند انتخاب شود. روش پیشنهادی از کارایی قابل قبولی بر روی داده‌های واقعی برخوردار است.

Feng و همکاران در [۹] روشی بر مبنای روش بیز ساده ارائه کرده‌اند که از ویژگی‌های انتخاب شده برای دسته‌بندی بهینه‌تر استفاده می‌نماید. قابلیت‌های منحصربه‌فرد این روش با استفاده از روش شاخص‌گذاری ویژگی‌ها با روش شاخص‌گذاری انتخاب عمومی است. نتایج این روش حاکی از آن است که با روش‌های شناختهٔ شدهٔ دیگر دسته‌بندی مثل ماشین بردار پشتیبان قابل رقابت است.

Chen و همکاران در [۱۰] با استفاده از الگوریتم‌های ژنتیک و تکاملی تفاضلی به دسته‌بندی اسناد متنی پرداخته‌اند. آن‌ها از عملگرهای ادغام و جهش به منظور بهینه‌سازی تعداد دسته‌ها استفاده می‌کنند. نتایج آزمایش‌های آن‌ها نشان می‌دهد که الگوریتم ژنتیک کاراتر از الگوریتم تکاملی تفاضلی می‌باشد و همچنین الگوریتم ترکیبی نسبت به دو الگوریتم ژنتیک و تکاملی تفاضلی کاراتر است.

استفاده از روش k نزدیک‌ترین همسایه به منظور مشابه بودن داده‌ها در دسته‌ها در [۱۱] استفاده شده است. نتایج آزمایش‌ها نشان می‌دهد که k نزدیک‌ترین همسایه کارایی خوبی در مقایسه با K-Means دارد و اسناد متنی را بهتر دسته‌بندی می‌کند. در دسته‌بندی اسناد متنی هدف بهینه کردن دسته‌ها در حداقل یا حداکثر تعداد خوشه‌ها می‌باشد. برای دسته‌بندی اسناد متنی از الگوریتم K-Means استفاده کرده‌اند. همچنین Luo و همکاران برای افزایش کارایی K-Means از الگوریتم ژنتیک به‌منظور انتخاب کلمات مرتبط به هم برای دسته‌بندی مشابه استفاده کردند. الگوریتم ژنتیک این توانایی را دارد که تکرار و ایجاد نسل‌های مختلف دقت دسته‌بندی را بهبود دهد و هزینه محاسباتی را کاهش دهد[۱۲]. نتایج آزمایش‌ها نشان می‌دهد که الگوریتم ژنتیک دقت الگوریتم K-Means را تا حد زیادی افزایش داده است. از جمله زمینه‌های کاربردی دسته‌بندی می‌توان به

متن کاوی، پیدا کردن مشابهت اسناد، ابهام زدایی، بازیابی اطلاعات، روش‌سازی و دسته‌بندی متن اشاره نمود. در [۱۳] Revathi و همکاران از شبکه عصبی مصنوعی و الگوریتم ژنتیک برای دسته‌بندی پویای متون بر مبنای محتوای آنان در رده‌های تعریف شده استفاده کرده‌اند. الگوریتم ژنتیک به منظور بهینه‌سازی دسته‌بندی و انتخاب ویژگی‌های مناسب برای دسته‌بندی به‌کار رفته است. مجموعه داده رویترز برای یادگیری و آزمایش استفاده شده است. نتایج این روش پویا که حاصل از ترکیب شبکه عصبی مصنوعی و الگوریتم ژنتیک می‌باشد نشان دهنده آن است که این روش کارایی خوبی در دسته‌بندی متون دارد.

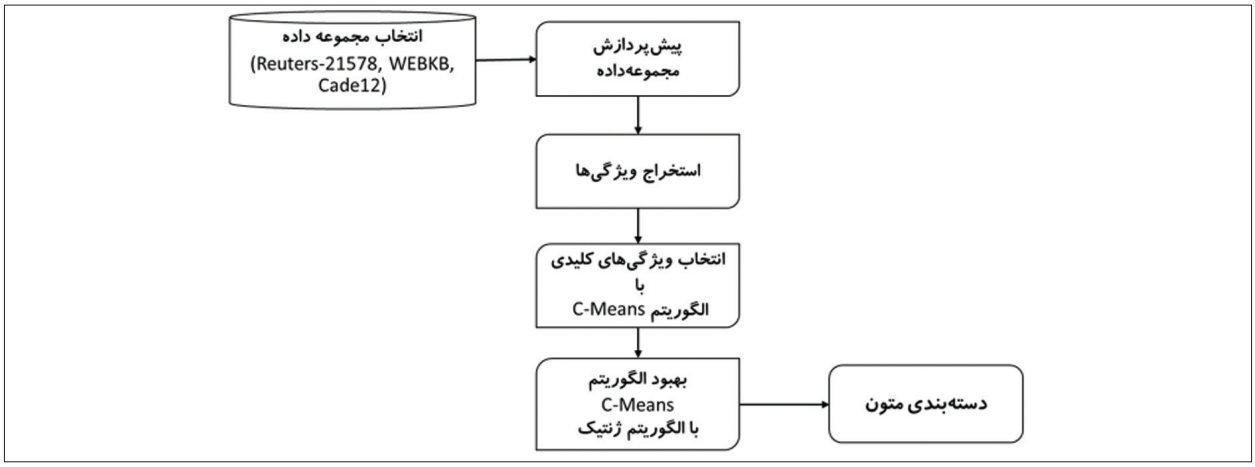
روش‌های یادگیری ماشین کاربرد گسترده‌ای در دسته‌بندی متون دارند، از جمله روش ماشین بردار پشتیبان در[۱۴] برای دسته‌بندی متون ارائه شده است. روش ارائه شده از روش انتخاب ویژگی چندرده‌ای ماشین بردار پشتیبان برای دسته‌بندی متون بهره می‌گیرد. نتایج حاکی از آن است که روش ارائه شده دقت خوبی در دسته‌بندی متون دارد. همچنین، Patra و Sing از شبکه‌های عصبی مصنوعی با بازخورد برای دسته‌بندی و یادگیری با نظارت برای دسته‌بندی خودکار اسناد متنی استفاده کرده‌اند. روش یادگیری در این روش از شبکه عصبی مصنوعی، الگوریتم پس‌انتشار می‌باشد و پنج مرحله برای پیش‌پردازش داده‌ها و متون ارائه کرده است. نتایج روش با الگوریتم پس‌انتشار به‌دست آمده وفاکتور مربوطه به‌عنوان روش وزن‌دار برای دسته‌بندی استفاده شده است.

در [۱۵] Zhang و همکاران روشی نیمه نظارت شده برای دسته‌بندی متون ارائه کرده‌اند. این روش برمبنای دسته‌بندی دسته‌های متون به‌صورت مجموعه‌ای از مولفه‌ها می‌باشد. در قسمت دسته‌بندی، از متون برچسب‌گذاری شده برای تعیین دسته‌های متنی و از متونی که برچسب‌گذاری نشده‌اند برای تعیین مراکز اقلیدسی دسته‌ها بهره می‌گیرد. زمانی که متن بدون برچسب‌گذاری برای دسته‌بندی در مرحله آزمایش وارد سیستم پیشنهادی می‌گردد، میزان مشابهت آنان با دسته‌های متنی سنجیده شده و بر مبنای این‌که نزدیک‌ترین دسته کدام است، برچسب آن دسته به آن متن جدید تعلق می‌گیرد. آزمایش سیستم پیشنهادی برروی مجموعه داده‌های Reuters-21578 و TanCrop V1.0 انجام گرفته است. نتایج حاکی از آن است که رویکرد نیمه نظارت شده پیشنهادی از روش ماشین بردار پشتیبان و روش شبکه عصبی مصنوعی با الگوریتم پس‌انتشار بهتر عمل می‌کند. همچنین این روش کارایی قابل مقایسه‌ای با روش بیز ساده با حداکثر بازده داشته و پیچیدگی محاسباتی کمتری دارد. بروز رسانی مراکز دسته‌ها، در دسته‌بندی اسناد متنی کاری مهم می‌باشد زیرا با بهینه‌سازی مراکز دسته‌ها، روند دسته‌بندی ویژگی‌ها در اسناد متنی موجب کیفیت بهتر آن خواهد شد.

Karimov و همکاران در [۱۶]، به ایجاد روشی بهینه برای بهبود روند دسته‌بندی در الگوریتم K-Means پرداخته‌اند. آن‌ها روش تکاملی ترکیبی با استفاده از الگوریتم K-Means را برای دسته‌بندی اسناد متنی مطرح کردند. از روش فراابتکاری در این روش برای شناسایی نامزدهای مناسب برای تعیین مراکز اقلیدسی انتخابی در الگوریتم K-Means استفاده شده است. بهبود نتایج دسته‌بندی ۳۰٪ بیشتر از الگوریتم K-Means استاندارد است. از جمله روش‌های فراابتکاری دیگر استفاده از شبیه‌سازی تبریدی در دسته‌بندی است. کار انجام یافته در دسته‌بندی اسناد متنی با روش شبیه‌سازی تبریدی [۱۷] به دسته‌بندی اسناد متنی در زبان چینی با روش همایند n تایی می پردازد.

۳- روش پیشنهادی

عامل اصلی در دسته‌بندی اسناد متنی، استخراج و انتخاب ویژگی‌های کلیدی متون می‌باشد و انتخاب ویژگی قبل از دسته‌بندی متن امری اساسی و ضروی است. در واقع انتخاب ویژگی‌های کلیدی متون از میان بیشمار ویژگی‌های استخراج شده، دشوار بوده و از طرفی انتخاب



شکل۱: نحوه عملکرد روش پیشنهادی

ایـــن ویژگی‌ها باعث افزایش کارایی در دســـته‌بندی متون می‌گردد. لذا در این مقاله از الگوریتم فازی C-Means برای انتخاب ویژگی‌های کلیدی، به صورت جســـتجو در فضای کل ویژگی‌هـــا با حذف کلمات تکـــراری، زائد و غیر مرتبط استفاده شده است. در شکل(۱) نمای کلی از عملکرد روش پیشنهادی نشان داده شده است.

به‌طـــور کلی برای انتخاب ویژگی‌هـــای کلیدی متون با الگوریتـــم C-Means فازی، تمام ویژگی‌هـــای متون ابتدا باید استخراج شوند که از میان ویژگی‌های استخراج شده ویژگی‌های کلیدی و مرتبط با موضوع اصلی متن انتخاب گردد. بـــرای اســـتخراج ویژگی‌های متـــون از فرمول (۱) استفاده کردیم.

$$Ltc - Weighting: a_{ik} = \frac{\log(f_{ik} + 1.0) * \log(N/n_i)}{\sqrt{\sum_{j=1}^{M}[\log(f_{ik} + 1.0) * log(N/n_i)]^2}} \quad (۱)$$

در فرمـــول (۱)، f_{ik} فرکانس کلمه i در متن N، k تعداد متون در مجموعه، M تعداد کلمات مجموعه پس از انجام عملیات کاهش و حذف کلمات اضافی و n_i: مجموع تعداد دفعاتی که کلمه i در هر مجموعه اتفاق افتاده است می‌باشد.

پـــس از اســـتخراج ویژگی‌هـــا و لغـــات یک متـــن، این ویژگی‌ها تبدیل به برداری از کلمات می‌شوند و تمام لغات کلیدی، اسامی، افعال و ... مشـــخص می‌گردند. بعد از این مرحلـــه باید ویژگی‌های بی‌فایده و نامرتبط حذف گردند و ویژگی‌های کلیدی متون اســـتخراج شود به‌طوری که این ویژگی‌ها، ویژگی‌های موثری در دســـته‌بندی باشـــند. در روش پیشـــنهادی از الگوریتم فازی C-Means برای یافتن اطلاعات و ویژگی‌های مهم متون استفاده شده است.

۳-۱ الگوریتم فازی C-Means

الگوریتم فـــازی C-Means به‌طور کلی نمونه‌ها رابه C دسته تقسیم می‌شوند و تعداد C از قبل مشخص شده است. الگوریتم C میانگین دقیقا شبیه به الگوریتم K میانگین است با فرق این‌که در آنجا K دسته داشتیم و در C میانگین تعداد دســـته‌ها برابر با C است. در این بخش الگوریتم پایه فازی C-Means به‌طور خلاصه معرفی خواهد شد. تابع هدف این الگوریتم به صورت زیر تعریف می‌شود:

$$J = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m}\, d^2(x_i - v_j) \quad (۲)$$

در فرمول (۲) u_{ij} میزان تعلق نمونه iام به مرکز دســـته jام و m درجـــه فـــازی بـــودن را تعیین می‌کنـــد. در اینجا $d^2(x_i - v_i)$ همـــان فاصلـــه اقلیدســـی و برابـــر بـــا $(x_i - v_i)$ اســـت، به‌طـــوری کـــه x_i نمونـــه iام و v_j مرکـــز دســـته jام اســـت. بر اســـاس تابع هدف معرفی شـــده در فرمول، معادلات بروزرســـانی مراکز و توابـــع تعلق بدین صورت خواهد بود:

$$u_{ij=} \frac{1}{\sum_{i=1}^{C}\left(\frac{d^2(x_i - v_j)}{d^2(x_i - v_i)}\right)^{\frac{1}{m-1}}} \quad (۳)$$

$$v_j = \frac{\sum_{i=1}^{N} {u_{ij}}^m \; x_i}{\sum_{i=1}^{N} {u_{ij}}^m} \qquad (۴)$$

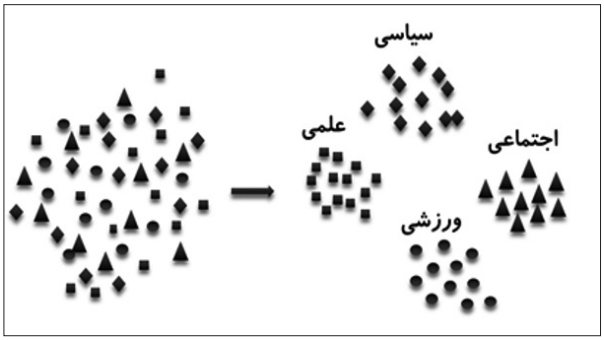
الگوریتم‌هـــای سلســـله مراتبی از یک معیار شـــباهت استفاده کرده و در هر مرحله داده‌ها را به دودسته تقسیم و درنهایت یک ساختار درختی از این تقسیم بندی تحت عنوان دندروگرام ایجاد می‌کنند[۱۸]. اما الگوریتم‌های افرازکننده، داده‌ها را به‌طور مســـتقیم درون چند دسته قرار می‌دهند. این الگوریتم‌ها خود به دو دسته سخت (انحصاری) و نرم (فازی) تقسیم می‌شوند.

در روش پیشـــنهادی الگوریتم فـــازی C-Means برای انتخـــاب ویژگی‌هـــای موثر، ابتـــدا سردســـته‌های (علمی، ورزشـــی، سیاســـی و...) از دســـته‌ها را ایجـــاد می‌کند و براســـاس فاصلهٔ اقلیدســـی تمام ویژگی‌های کلیدی متون رده‌بندی می‌شوند. با این فرایند ویژگی‌های با فاصلهٔ خیلی دور از سردســـته حذف شده و ویژگی‌های کلیدی انتخاب می‌گردند. در شکل (۲) نمایی از دسته‌بندی نشان داده شده است.

در الگوریتـــم فازی C-Means برای مشـــخص شـــدن تعداد مناســـب ویژگی‌ها این الگوریتم در هر تکرار از میان بیشـــمار ویژگی‌های استخراج شـــده از متون تعداد ۱/۲یا ۱/۳ یـــا ۱/٤ و .... ویژگی‌ها را انتخـــاب می‌کند و در نهایت فقط ۱/۳ این ویژگی‌ها را که جزء مهم‌ترین و باارزش‌ترین ویژگی‌ها در تشـــخیص متنی از متن دیگر می‌باشد انتخاب می‌نمایـــد. پس از این‌که ویژگی‌های مهم و باارزش انتخاب شـــدند تمام ایـــن ویژگی‌ها بـــه الگوریتم ژنتیک ارســـال می‌گردند.

۳-۲ الگوریتم ژنتیک

الگوریتم ژنتیک روش بهینه‌سازی الهام گرفته از طبیعت جاندار(موجودات زنده) اســـت که بر اســـاس قانون تکامل داروی‌ـــن (بقا بهتریـــن) که می‌گوید: موجـــودات ضعیف‌تر از بیـــن می‌روند و موجودات قوی‌تر باقـــی می‌مانند ارائه گردیده است. به‌طور کلی این الگوریتم با تولید راه‌حل‌های



شکل ۲: نمایی از الگوریتم فازی C-Means

تصادفی که جمعیت اولیه نامیده می‌شـــود شروع می‌شود و هر کروموزوم به‌عنوان فـــردی از جمعیت اولیه در نظر گرفته می‌شـــود. برای تولید نسل‌ها تمامی کروموزوم‌ها با یک تابع برازش متناســـب با هر مسئله مورد ارزیابی قرار می‌گیرند و بهترین کروموزوم‌ها برای تولید نســـل بعدی انتخاب می‌شـــوند. این الگوریتم، الگوریتم مبتنی بر تکرار است و برای بهینه‌سازی، جستجو و یادگیری ماشین مورد استفاده قرار می‌گیرد[۷].

باتوجه به این‌کـــه الگوریتم ژنتیک در هر تکرار چندین کروموزوم از فضای جستجو را در نظر می‌گیرد بنابراین شـــانس این‌که به یک ماکزیمم محلی همگرا شـــود کاهش می‌یابد. این الگوریتـــم جمعیت‌های کاملی از کروموزوم‌ها را تولید می‌کند. سپس هر کروموزوم را به‌صورت انفرادی امتحـــان می‌کند و بـــا ترکیب محتویـــات آن‌ها یک جمعیت جدید را که شامل نســـل بهبود یافته است تشکیل می‌دهد. صـــرف نظر از انجام یک جســـتجو و ملاحظـــهٔ هم‌زمان، الگوریتم ژنتیک تعدادی از کروموزوم‌ها را با ماشـــین‌های مـــوازی تطبیق می‌دهد زیرا در اینجا تکامل هر کروموزوم یک فرآیند مســـتقل اســـت. لذا الگوریتم ژنتیک فقط نیاز به اطلاعاتی در مورد کیفیت رده حل‌های ایجاد شده به وسیلهٔ هـــر مجموعه از متغیرهـــا دارد، در صورتی که بعضی از روش‌هـــای بهینه‌ســـازی نیاز به اطلاعات یـــا حتی نیاز به شناخت کامل از ســـاختمان مسئله و متغیرها دارند. چون الگوریتم ژنتیک نیاز به چنین اطلاعاتی مشخصی از مسئله ندارد بنابراین قابل انعطاف‌تر از بیشتر روش‌های جستجو است.

شکل ۳: هم‌ارزی عناصر ژنتیک در روش پیشنهادی

در الگوریتم ژنتیک تمام ویژگی‌های انتخاب شده یک متن توسط الگوریتم C-Means فازی به‌عنوان یک کروموزوم و ورودی‌های الگوریتم ژنتیک در نظر گرفته می‌شود. مجموع کروموزوم‌های الگوریتم ژنتیک، مجموع متون یک مجموعه‌داده را تشکیل می‌دهد. در شکل(٤)، شبه کد و در شکل ٥ روندنمای روش پیشنهادی نشان داده شده است. در شبه کد روش پیشنهادی مراحل مختلف روش پیشنهادی تشریح شده است.

در اجرای الگوریتم ژنتیک در هر مرحله، از ادغام دو طرفه و انتخاب تورنومت تعدادی از متون انتخاب شده و به نسل بعدی انتقال داده می‌شود. بعد از اعمال ادغام به منظور اجتناب از همگرایی به بهینه محلی و ایجاد تنوع و گوناگونی در جمعیت با استفاده از عملگر جهش با نرخ ۰/۱ یک تعداد از کروموزوم‌های به‌دست آمده را تغییر می‌دهیم. بدین شکل هر متن براساس این‌که از چه ویژگی‌ها و کلمات کلیدی برخوردار است طبقهٔ مربوط به آن‌ها، برایش پیش‌بینی می‌شود. در نهایت پس از تخصیص هر متن به یک دسته و پیش‌بینی دستهٔ هر یک از متون، ستون هدف با رده‌بندی اولیه مقایسه می‌شود تا تعداد پیش‌بینی درست یا اشتباه روش پیشنهادی مشخص شود. در صورت تعداد زیاد پیش‌بینی روش پیشنهادی الگوریتم ژنتیک از

جدول۱: لیست دسته‌ها و موضوعات مجموعه‌داده Reuters 21578

Reuters 21578				
Topics #	train # docs	test # docs	other #	Total # docs
0	1828	280	8103	10211
1	6552	2581	361	9494
2	890	309	135	1334
3	191	64	55	310
4	62	32	10	104
5	39	14	8	61
6	21	6	3	30
7	7	4	0	11
8	4	2	0	6
9	4	2	0	6
10	3	1	0	4
11	0	1	1	2
12	1	1	0	2
13	0	0	0	0
14	0	2	0	2
15	0	0	0	0
16	1	0	0	1

ابتدا اجرا می‌شود تا تعداد دسته‌بندی نادرست را به حداقل برساند.

۴- بررسی وارزیابی

در این بخش به منظور مشاهده نتایج، الگوریتم ژنتیک و روش پیشنهادی بر روی مجموعه‌داده WebKB[۲۰], Reuters 21578[۱۹] و Cade12[۲۱] اجرا شده است. در جدوال(۱) تا (۳) انواع دسته‌ها و موضوعات در مجموعه‌داده در دو بخش آزمایش و آموزش نشان داده شده است. برای محاسبه و ارزیابی عملکرد روش پیشنهادی از پنج معیار مطرح در طبقه‌بندی متون که عبارتند از: تعداد داده‌های درست طبقه‌بندی شده، تعداد داده‌های نادرست طبقه‌بندی شده دقت، بازخوانی و F Measure استفاده شده است.

همچنین بعضی از پارامترهای الگوریتمهای فازی C-Means و ژنتیک که بر روی روند روش پیشنهادی تاثیر مهمی دارند در جدول(٤) نشان داده

بهبود الگوریتم فازی C-Means با الگوریتم ژنتیک برای انتخاب ویژگی‌ها در دسته‌بندی اسناد متنی	
مجموعه‌داده انتخابی کاربر	ورودی‌ها
متون دسته‌بندی شده درست، نادرست و مقدار معیارهای ارزیابی دقت، دوباره‌خوانی، میانگین دقت و بازخوانی	خروجی‌ها
گام‌ها	گام
۱- بارگذاری داده‌های مجموعه‌داده‌ی انتخابی توسط کاربر ۲- تبدیل داده‌های موجود در مجموعه‌داده به شکل استاندارد و حذف داده‌های مبهم و تهی	پیش‌پردازش مجموعه‌داده
۳- استخراج ویژگی‌های متون با استفاده از فرمول (۱)	استخراج ویژگی‌های متون
۴- انتخاب ویژگی‌های کلیدی متون از میان ویژگی‌های استخراج شده متون	الگوریتم فازی C-Means
۵-کروموزوم‌ها براساس ویژگی‌های متون موجود در مجموعه‌داده جدید، برای ایجاد مجموعه جمعیت اولیه تولید می‌شوند ۶-تعدادی از کروموزوم‌ها به‌طور تصادفی انتخاب می‌شوند (از قانون انتخاب تورنومت استفاده می‌شود). ۷-عمل ترکیب بر روی دو کروموزوم دارای بالاترین تابع ارزیابی برای تولید نسل جدید اعمال می‌شود (تعداد پیش‌بینی درست و غلط به‌عنوان تابع ارزیاب الگوریتم ژنتیک در نظر گرفته شده است). ۸-عمل جهش بر روی کروموزوم اعمال شود. ۹-کروموزوم جدید به لیست کروموزوم‌های موجود اضافه شده و از گام ۲ ادامه می یابد ۱۰-شرط توقف دسته‌بندی تمام متون موجود در مجموعه‌داده است.	الگوریتم ژنتیک
۱۱-ذخیره‌سازی نتایج در روش پیشنهادی ۱۲-دسته‌بندی داده‌ها با روش آموزش دیده ۱۳-محاسبه معیارهای ارزیابی	دسته‌بندی متون موجود در مجموعه‌داده

شکل ۴: شبه کد روش پیشنهادی

شکل۵: روندنمای مدل پیشنهادی

جدول ۳ : لیست دسته‌ها و موضوعات مجموعه‌داده Cade 12

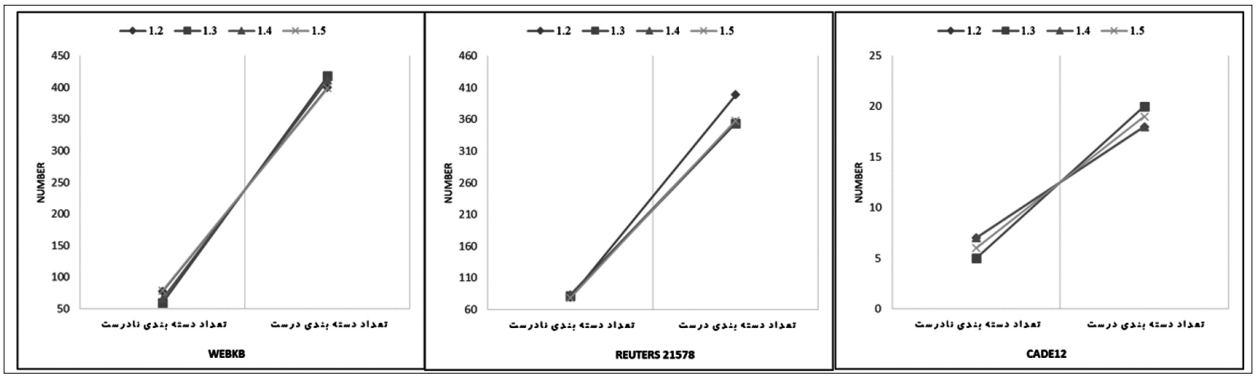
Cade12			
Class	# train docs	# test docs	Total # docs
01--servicos	5627	2846	8473
02--sociedade	4935	2428	7363
03--lazer	3698	1892	5590
04--informatica	2983	1536	4519
05--saude	2118	1053	3171
06--educacao	1912	944	2856
07--internet	1585	796	2381
08--cultura	1494	643	2137
09--esportes	1277	630	1907
10--noticias	701	381	1082
11--ciencias	569	310	879
12--compras-online	423	202	625
Total	27322	13661	40983

جدول ۲: لیست دسته‌ها و موضوعات مجموعه‌داده WebKB

WebKB			
Class	train docs #	test docs #	Total # docs
project	336	168	504
course	620	310	930
faculty	750	374	1124
student	1097	544	1641
Total	2803	1396	4199

جدول ۴: پارامترهای مورد استفاده در روش پیشنهادی

پارامتر	مقدار
تعداد جمعیت اولیه	۱۰۰
تعداد تکرار	۵۰
نرخ ادغام	۰,۲۵
نرخ جهش	۰,۱
تابع ارزیابی	تعداد پیش‌بینی درست و غلط



شکل۶: ارزیابی معیارهای تعداد دسته‌بندی درست و تعداد دسته‌بندی نادرست روش‌ها بر روی مجموعه داده‌های WEBKB ,Reuters 21578 و CADE12 با ۱/۲، ۱/۳، ۱/۴و ۱/۵ ویژگی‌های انتخابی از کل ویژگی‌ها

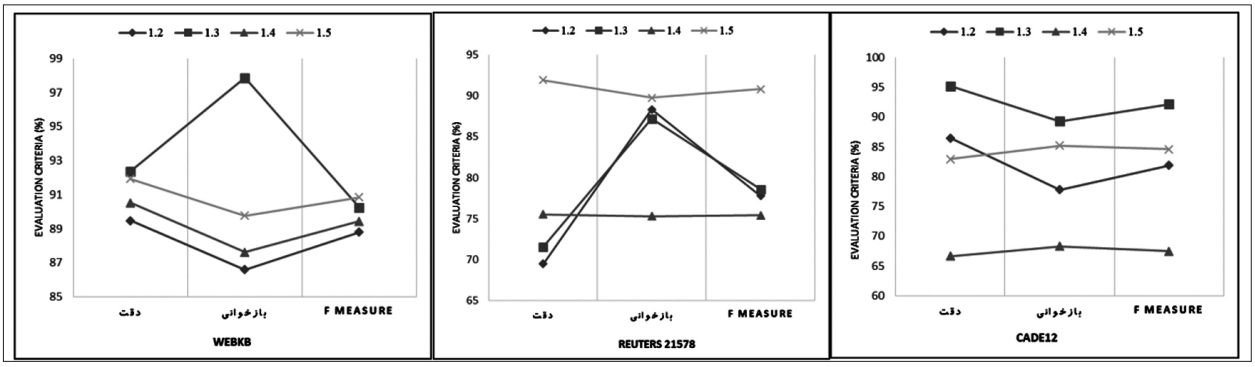
اســت. مقدار تمام این پارامترها بــر مبنای آزمایش تعیین شده است.

از آنجائی که انتخاب ویژگی‌های متون و انتخاب تعداد مختلف آن‌ها، اهمیت بالایی در دســته‌بندی درست اسناد متنــی دارد. از این‌رو بــرای اطمینان از درســتی انتخاب ۱/۳ تعــداد ویژگی‌هــا از میان بیشــمار ویژگی‌های متون توســط الگوریتم فازی C-Means در روش پیشــنهادی، ابتدا به بررســی تاثیر تعداد مختلــف ویژگی‌های انتخاب شده از طریق الگوریتم فازی C-Means می‌پردازیم. نتایج حاصل از ارزیابی روش پیشــنهادی با تعداد انتخاب ۱/۲، ۱/۳، ۱/٤ و ۱/۵ ویژگی‌های متون توســط الگوریتم فازی C-Means در جدول (۵) و شــکل‌های (٦) و (۷) نشان داده شده است.

بــا توجه بــه ارزیابی نتایــج حاصل از انتخــاب ۱/۲، ۱/۳، ۱/٤ و ۱/۵ ویژگی‌ها از میان بیشــمار ویژگی توسط الگوریتــم فــازی C-Means انتخاب تعــداد ۱/۳ ویژگی‌ها نتایــج بهتری در مقایســه بــا تعداد ویژگی‌هــای مختلف انتخــاب شــده دارد. همچنیــن بــرای نشــان دادن دقت به‌کارگیری همزمــان دو الگوریتم ژنتیک و الگوریتم فازی C-Means نتایج پیاده‌سازی الگوریتم ژنتیک بدون دخالت الگوریتــم فازی C–Means نیز در جدول (٦) نشــان داده شده است.

جدول۵: ارزیابی روش پیشنهادی با ویژگی‌های متفاوت انتخاب شده توسط الگوریتم فازی C-Means

	مجموعه‌داده WebKB				مجموعه‌داده Reuters 21578				مجموعه‌داده Cade 12			
	۱/۲	۱/۳	۱/۴	۱/۵	۱/۲	۱/۳	۱/۴	۱/۵	۱/۲	۱/۳	۱/۴	۱/۵
دسته‌بندی نادرست	۷۸	۶۰	۶۶	۷۹	۸۰	۸۲	۸۴	۷۹	۷	۵	۷	۶
دسته‌بندی درست	۴۰۰	۴۱۸	۴۱۲	۳۹۹	۳۵۸	۳۵۶	۳۵۴	۳۹۹	۱۸	۲۰	۱۸	۱۹
دقت	۸۹/۵	۹۲/۴	۹۰,۵	۹۲	۶۹,۵	۷۱/۵۵	۷۵/ ۵	۹۲	۸۶/۴۶	۹۵	۶۷	۸۳
بازخوانی	۸۶/۶	۹۷/۸۷	۸۷/۶	۸۹/۷۷	۸۸/۳۴	۸۷/۱۸	۷۵/۲۹	۸۹/۷۷	۷۷/۷۸	۸۹/۲۹	۶۸/۳	۸۵/۱۹
F Measure	۸۸/۸	۹۰/۲۳۸۰	۸۹/۴	۹۰/۸	۷۷/۷۹	۷۸/۵۹	۷۵/۳۹	۹۰/۸۳	۸۱/۸۸	۹۲/۲	۶۷/۵	۸۴/۵۹



شکل(۷): ارزیابی معیارهای دقت، صحت و F Measure روش‌ها بر روی مجموعه داده‌های WEBKB, Reuters 21578 و CADE12 با ۱/۲، ۱/۳، ۱/۴و ۱/۵ ویژگی‌های انتخابی از کل ویژگی‌ها

جدول۶: ارزیابی روش پیشنهادی بر روی مجموعه‌داده‌ها

	مجموعه‌داده WebKB		مجموعه‌داده Reuters 21578		مجموعه‌داده Cade 12	
	الگوریتم ژنتیک	روش پیشنهادی	الگوریتم ژنتیک	روش پیشنهادی	الگوریتم ژنتیک	روش پیشنهادی
تعداد دسته‌بندی نادرست	۱۲۲	۶۰	۱۰۷	۸۲	۷	۵
تعداد دسته‌بندی درست	۳۵۶	۴۱۸	۳۳۱	۳۵۶	۱۸	۲۰
دقت	۸۷/۷۲	۹۲/۳۸	۶۸/۶۳	۷۱/۵۵	۷۵/۹۳	۹۵/۲۴
بازخوانی	۸۵/۲۸	۹۷/۸۷	۷۵/۷۳	۸۷/۱۸	۷۹/۷۶	۸۹/۲۹
F Measure	۸۶/۵	۹۰/۲۴	۷۲/۲	۸۷/۵۹۴	۷۷/۸	۹۲/۲

براســاس مقادیر جدول(٦) می‌تــوان نتیجه گرفت که روش پیشنهادی در تمام معیارها بر روی سه مجموعه‌داده WebKB, Cade 12 و Reuters 21578، عملکــرد بهتری از الگوریتم ژنتیک داشــته اســت و به‌کارگیری همزمان دو الگوریتم C-Means فازی و ژنتیک در دســت‌یابی به نتایج بهتر در دســته‌بندی اســناد متنی ضروری بوده است. در آخــر نتایــج معیارهای دقــت، بازخوانــی و F-Measure روش پیشــنهادی با تعدادی از روش‌های ارائه شده مورد مقایســه قرار داده شــده اســت. نتایج حاصل از مقایسه در جداول(۷) تا (۹) و شــکل(۸) جمع‌آوری و نشــان داده شده است.

۵. نتیجه‌گیری و کارهای آینده

نکتهٔ حائز اهیمت در ارائهٔ یک روش دســته‌بندی دقیق، انتخــاب ویژگی‌های مفیــد و حذف ویژگی‌هــای غیرمفید متون از میان تمام ویژگی‌های اســتخراج شده می‌باشد که براساس انتخاب ویژگی‌های مفید، دسته‌بندی متون با دقت و سرعت بالاتری امکان‌پذیر شود. برای تحقق این هدف در

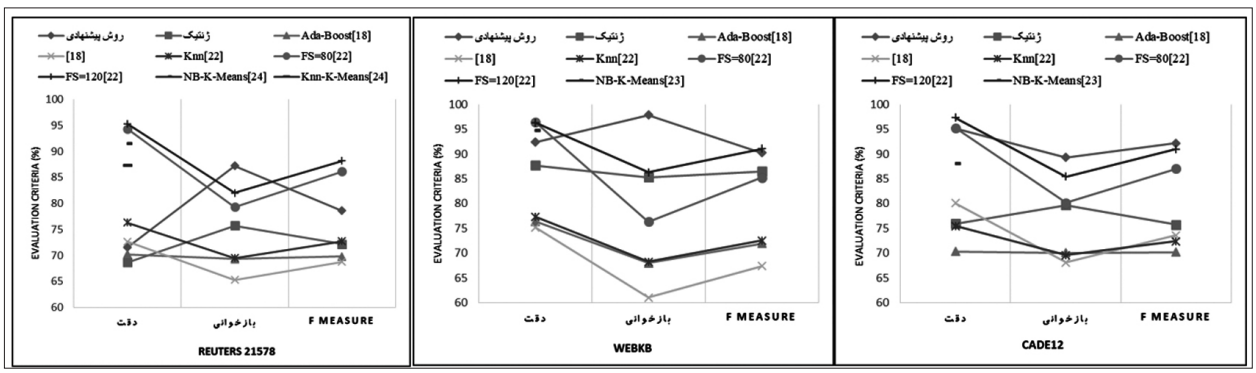
جدول۷: مقایسه روش پیشنهادی با سایر روش‌های ارائه‌شده در طبقه‌بندی اسناد متنی در مجموعه‌داده Reuters 21578

					معیارهای ارزیابی		
					دقت	بازخوانی	F-Measure
Reuters 21578	الگوریتم ادابوست[۱۸]	TF			۷۰,۲۳	۶۹,۳۲	۶۹,۷۷
		NORMTF			۷۲,۳۲	۶۵,۱۰	۶۸,۵۲
		LOGTF			۷۶,۰۹	۶۶,۸۷	۷۱,۱۸
		ITF			۷۷,۹۴	۶۸,۲۴	۷۲,۷۷
		SPARCK			۷۴,۲۹	۶۶,۰۷	۶۹,۹۴
	[۱۸]	TF			۷۲,۶۵	۶۵,۳۲	۶۸,۷۹
		NORMTF			۶۹,۳۵	۶۴,۷۷	۶۶,۹۸
		LOGTF			۷۳,۰۰	۶۲,۰۰	۶۷,۰۵
		ITF			۷۶,۰۰	۶۹,۳۵	۷۲,۵۲
		SPARCK			۷۲,۱۴	۶۸,۱۱	۷۰,۰۷
	K نزدیک‌ترین همسایه[۲۲]	تعداد k مختلف	K=3		۷۶,۳۴	۶۹,۴۷	۷۲,۷۴
			K=4		۷۳,۲۴	۶۷,۲۸	۷۰,۱۳
			K=5		۷۰,۰۸	۶۵,۱۲	۶۷,۵۱
	FS=80[22]	تعداد k مختلف	K=3		۹۴,۳۴	۷۹,۳۲	۸۶,۱۶
			K=4		۹۳,۰۰	۷۶,۲۴	۸۳,۷۹
			K=5		۹۱,۲۷	۷۴,۲۰	۸۱,۸۵
	FS=120[22]	تعداد k مختلف	K=3		۹۵,۳۰	۸۲,۰۲	۸۸,۱۶
			K=4		۹۶,۰۰	۸۱,۱۵	۸۷,۹۵
			K=5		۹۵,۱۸	۸۰,۲۰	۸۷,۰۵
	NB-K-Means[23]	تعداد k مختلف	K=3		۹۱,۶۰	–	–
			K=4		۹۰,۱۰	–	–
			K=5		۸۳,۴۰	–	–
	KNN-K-Means[24]	تعداد k مختلف	K=3		۸۷,۳۲	–	–
			K=4		۷۲,۳۷	–	–
			K=5		۶۷,۱۵	–	–
	الگوریتم ژنتیک	تعداد ویژگی‌های متفاوت انتخاب شده	۱/۲		۵۶/۱۶	۷۵/۹۶	۶۴/۶
			۱/۳		۶۸/۶۳	۷۵/۷۳	۷۲/۲
			۱/۴		۷۵/۷۶	۸۲/۹۶	۷۹/۲
			۱/۵		۸۷/۸۹	۸۶/۱۷	۸۷/۲۴
	روش پیشنهادی	تعداد ویژگی‌های متفاوت انتخاب شده	۱/۲		۶۹,۴۹	۸۸,۳۴	۷۷,۷۹۰
			۱/۳		۷۱,۵۵	۸۷,۱۸	۷۸,۵۹
			۱/۴		۷۵,۵۱	۷۵,۲۹	۷۵,۳۹۷
			۱/۵		۹۱,۹۲	۸۹,۷۷	۹۰,۸۳

این مقاله روشی براساس بهبود الگوریتم فازی C-Means با الگوریتم ژنتیک برای انتخاب ویژگی‌ها در دسته‌بندی اسناد متنی ارائه شده است که از طریق الگوریتم فازی C-Means از بین تمامی ویژگی‌های استخراج شده ۱/۳ ویژگی‌ها که مفیدترین ویژگی‌ها در جداسازی متنی از متن دیگر است، انتخاب شد و به الگوریتم ژنتیک برای بهبود در دسته‌بندی متون ارسال گردید. در روش پیشنهادی دسته‌بندی متون فقط براساس ۱/۳ویژگی‌های انتخاب شده از طریق الگوریتم C-Means فازی انجام گرفت که این عمل باعث افزایش سرعت و دقت روش پیشنهادی در

جدول۸: مقایسه روش پیشنهادی با سایر روش‌های ارائه‌شده در طبقه‌بندی اسناد متنی در مجموعه‌داده WEBKB

					معیارهای ارزیابی		
					دقت	بازخوانی	F-Measure
WEBKB	الگوریتم ادابوست[۱۸]	TF			۷۶,۳۸	۶۸,۰۰	۷۱,۹۵
		NORMTF			۷۲,۲۴	۶۹,۹۴	۷۱,۰۷
		LOGTF			۷۶,۵۴	۶۵,۳۷	۷۰,۵۲
		ITF			۷۵,۹۸	۶۴,۸۷	۶۹,۹۹
		SPARCK			۷۶,۰۰	۶۷,۴۷	۷۱,۴۸
	[۱۸]	TF			۷۵,۲۳	۶۱,۰۲	۶۷,۳۸
		NORMTF			۷۱,۷۷	۶۶,۵۵	۶۹,۰۶
		LOGTF			۷۳,۶۸	۶۸,۰۸	۷۰,۷۷
		ITF			۷۶,۰۴	۶۹,۷۷	۷۲,۷۷
		SPARCK			۷۹,۵۸	۷۰,۳۶	۷۴,۶۹
	K نزدیک‌ترین همسایه[۲۲]	تعداد k مختلف	K=3		۷۷,۳۵	۶۸,۲۴	۷۲,۵۱
			K=4		۷۳,۸۷	۶۵,۰۷	۶۹,۱۹
			K=5		۶۹,۳۴	۶۲,۰۱	۶۵,۴۷
	FS=80[22]	تعداد k مختلف	K=3		۹۶,۴۷	۷۶,۳۴	۸۵,۲۳
			K=4		۹۵,۲۴	۷۴,۲۱	۸۳,۴۲
			K=5		۹۳,۰۳	۷۳,۴۱	۸۲,۰۶
	FS=120[22]	تعداد k مختلف	K=3		۹۶,۳۴	۸۶,۳۰	۹۱,۰۴
			K=4		۹۴,۱۵	۸۲,۲۴	۸۷,۷۹
			K=5		۹۲,۲۷	۸۰,۲۰	۸۵,۸۱
	NB-K-Means[23]	تعداد k مختلف	K=3		۹۴,۸۰	–	–
			K=4		۹۳,۲۰	–	–
			K=5		۷۸,۲۰	–	–
	الگوریتم ژنتیک	تعداد ویژگی‌های متفاوت انتخاب شده	۱/۲		۸۹/۳۷	۸۵/۷۷	۸۷/۵
			۱/۳		۸۷/۷۲	۸۵/۲۸	۸۶/۵
			۱/۴		۸۷/۱۳	۸۴/۳۵	۸۵/۷۱۸
			۱/۵		۸۷/۸۹	۸۶/۱۷	۸۷/۲۴
	روش پیشنهادی	تعداد ویژگی‌های متفاوت انتخاب شده	۱/۲		۸۹/۴۷	۸۶/۵۹	۸۸/۸
			۱/۳		۹۲/۳۸	۹۷/۸۷	۹۰/۲۳۸۰
			۱/۴		۹۰/۵۲	۸۷/۶۲	۸۹/۴۳
			۱/۵		۹۱/۹۲	۸۹/۷۷	۹۰/۸۳۶



شکل۸: ارزیابی روش‌ها بر روی مجموعه داده‌های WEBKB, Reuters 21578 و CADE12 با ۱/۳ ویژگی‌های انتخابی از کل ویژگی‌ها

جدول(۹): مقایسه روش پیشنهادی با سایر روش‌های ارائه‌شده در طبقه‌بندی اسناد متنی در مجموعه‌داده CADE۱۲

					معیارهای ارزیابی		
					دقت	بازخوانی	F-Measure
CADE۱۲	[۱۸]Ada-Boost	TF			۷۰،۳۲	۷۰،۰۵	۷۰،۱۸
		NORMTF			۷۶،۹۴	۶۸،۳۴	۷۲،۳۹
		LOGTF			۷۴،۰۰	۶۵،۰۸	۶۹،۲۵
		ITF			۷۶،۵۷	۷۱،۲۴	۷۳،۸۱
		SPARCK			۷۴،۰۱	۶۵،۴۶	۶۹،۴۷
	[۱۸]	TF			۸۰،۱۰	۶۸،۰۸	۷۳،۶۰
		NORMTF			۷۳،۹۸	۶۹،۱۱	۷۱،۴۶
		LOGTF			۷۶،۵۸	۶۵،۰۰	۷۰،۳۲
		ITF			۸۲،۳۶	۸۰،۳۵	۸۱،۳۴
		SPARCK			۷۵،۰۹	۷۱،۶۸	۷۳،۶۸
	Knn[22]	تعداد k مختلف	K=3		۷۵،۴۸	۶۹،۵۸	۷۲،۴۱
			K=4		۷۲،۰۰	۶۷،۳۲	۶۹،۵۸
			K=5		۶۸،۴۶	۶۶،۴۱	۶۷،۴۲
	FS=80[22]	تعداد k مختلف	K=3		۹۵،۳۰	۸۰،۱۴	۸۷،۰۷
			K=4		۹۳،۰۷	۷۹،۳۳	۸۵،۶۵
			K=5		۹۱،۵۶	۷۶،۸۴	۸۳،۵۶
	FS=120[22]	تعداد k مختلف	K=3		۹۷،۳۴	۸۵،۴۷	۹۱،۰۲
			K=4		۹۴،۹۰	۸۴،۱۹	۸۹،۲۲
			K=5		۹۳،۲۷	۸۱،۲۲	۸۶،۸۳
	NB-K-Means[23]	تعداد k مختلف	K=3		۸۸،۱۰	-	-
			K=4		۸۵،۸۰	-	-
			K=5		۷۰،۳۰	-	-
	الگوریتم ژنتیک	تعداد ویژگی‌های متفاوت انتخاب شده	۱/۲		۶۱/۶۷	۶۷/۷۸	۶۴/۵۷۷
			۱/۳		۷۵/۹۳	۷۹/۷۶	۷۷/۷۹۶
			۱/۴		۷۲/۲۲	۷۹/۶۹	۷۵/۷۷۱
			۱/۵		۸۴/۳۸	۷۰/۳۷	۷۶/۷۳۸
	روش پیشنهادی	تعداد ویژگی‌های متفاوت انتخاب شده	۱/۲		۸۶/۴۶	۷۷/۷۸	۸۱/۸۸۸
			۱/۳		۹۵/۲۴	۸۹/۲۹	۹۲/۱۶۵
			۱/۴		۶۶/۶۷	۶۸/۳	۶۷/۴۷۵
			۱/۵		۸۲/۹۶	۸۵/۱۹	۸۴/۵۹

دسته‌بندی اسناد متنی می‌شود.

در این مقاله از پنج معیار ارزیابی مطرح در طبقه‌بندی متون که عبارتند از : تعداد داده‌های درست طبقه‌بندی شده، تعداد داده‌های نادرست طبقه‌بندی شده، دقت، بازخوانی و F Measure برای محاسبه و ارزیابی عملکرد روش پیشنهادی استفاده شده است. براساس ارزیابی روش پیشنهادی در تمام معیارها بر روی سه مجموعه‌داده WebKB, Reuters 21578 و Cade12 نتایج حاصله بیانگر عملکرد بهتر روش پیشنهادی در مقایسه با الگوریتم ژنتیک بود و به‌کارگیری همزمان دو الگوریتم C-Means فازی و ژنتیک در دست‌یابی به نتایج بهتر در دسته‌بندی اسناد متنی ضروری بوده است.

منابع:

[۱] نعمتی شهلا، بصیری محمد احسان.، دسته بندی اسناد با استفاده از الگوریتم KNN، دهمین کنفرانس مهندسی برق ایران، ۱۳۸۶.

[۲] جلیلی سعید، گرانی شیما.، دسته‌بندی متون با رویکرد الگوریتم ژنتیک، پانزدهمین کنفرانس مهندسی برق ایران، ۶۷ – ۷۳، ۱۳۸۶.

[۳] بینا بهاره و رهگذر مسعود و ده موبد آذین.، دسته‌بندی خودکار متون فارسی، انجمن کامپیوتر ایران، ۴-۱۳، ۱۳۸۶.

[۴] نعمتی شهلا، بصیری محمد احسان، دسته‌بندی اسناد فارسی با استفاده از الگوریتم knn، ۱-۸، ۱۳۹۲.

[5] Shao, Z., Yang, S., Gao, F., Zhou, K., & Lin, P. A New Electricity Price Prediction Strategy Using Mutual Information-Based SVM-RFE Classification. Renewable and Sustainable Energy Reviews, 70, 330-341, (2017).

[6] Liu, L., Kang, J., Yu, J., & Wang, Z. (2005). "A Comparative Study on Unsupervised Feature Selection Methods for Text Clustering", Proc. of the Conference on Natural Language Processing and Knowledge Engineering, IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on (pp. 597-601). IEEE.

[7] Webber, W. E. 2010. Measurment in Information Retrieval Evaluation. Ph.D Thesis. The university of Melburne http://www.umiacs.umd.edu/~wew/wew-thesis-PhD.pdf [Last Access: 08.08.2017]

[8] M. L. Zhang, J. M. Pena, V. Robles, Feature Selection for Multi-LabelNaive Bayes Classification, Information Sciences, Vol. 179, Iss. 19, pp. 3218-3229, 2009

[9] G. Feng, J. Guo, B. Y. Jing, T. Sun, Feature Subset Selection Using Naive Bayes for Text Classification, Pattern Recognition Letters, Vol. 65, pp. 109-115, 2015.

[10] Chen Y., Qin B., Liu T., Liu Y., Li Sheng, The Comparison of SOM and K-Means for Text Clustering", Vol. 3, No. 2, pp. 268-274, 2010.

[11] C. Luo, Y. Li, S.M. Chung, Text Document Clustering Based on Neighbors, Data & Knowledge Engineering, Vol. 68 , pp.1271–1288, 2009.

[12] Verma H., Kandpal E., Pandey B., Dhar J., A Novel Document Clustering Algorithm Using Squared Distance Optimization Through Genetic Algorithms", International Journal on Computer Science and Engineering, Vol. 02, No. 05, pp. 1875-1879, 2010.

[13] N. Revathi, A. Peter, S.J. Kumar, Web Text Classification Using Genetic Algorithm and a dynamic neural NetworK Model, International Journal of Advanced Research in Computer Engineering and Technology (IJARCET), Vol. 2, Iss. 2, February 2013.

[14] B. Ramesh, J. G. R. Sathiaseelan, An advanced Multi Class instance selection based Support Vector Machine for Text Classification, Procedia Computer Science, Vol. 57, pp. 1124-1130, 2015.

[15] W. Zhang, X. Tang, T. Yoshida, TESC: An Approach to Text Classification Using Semi-Supervised Clustering, Knowledge-Based Systems, Vol. 75, pp. 152-160, 2015

[16] J. Karimov, M. Ozbayoglu, Clustering Quality Improvement of K-Means Using a Hybrid Evolutionary Model, Procedia Computer science, Vol. 61, pp. 38-45, 2015.

[17] C. H. Chang, Simulated Annealing Clustering of Chinese Words for Contextual Text Recognition, Pattern Recognition Letters, Vol. 17, No. 1, pp. 57–66, 1996.

[18] Majidpour, Hiwa, and Farhad Soleimanian Gharehchopogh." An Improved Flower Pollination Algorithm with AdaBoost Algorithm for Feature Selection in Text Documents Classification". Journal of Advances in Computer Research, pp: 1-11(2008).

[19]http://archive.ics.uci.edu/ml/datasets/Reuters 21778+Text+Categorization+Collection [Last Access: 08.08.2017]

[20] Craven, M., McCallum, A., PiPasquo, D., Mitchell, T., & Freitag, D. ((1998(. Learning to extract symbolic knowledge from the world wide web. In: DTIC Document.

[21] http://ana.cachopo.org/datasets-for-single-label-text-categorization [Last Access: 08.08.2017]

[22]Allahverdipour, Ali, and Farhad Soleimanian Gharehchopogh. "An Improved K-Nearest Neighbor with Crow Search Algorithm for Feature Selection in Text Documents Classification." Journal of Advances in Computer Research 9.2, pp: 1-11(2018).

[23]A. Allahvirdipour, F.S.Gharehchopogh, A New Hybrid Model of K-Means and Naïve Bayes Algorithms for Features Selection in Text Documents Classification, Journal of Advances in Computer Research,Vol.8, No. 4.pp:73-86, 2017.

[24] R. Habibpour, K. Khalilpour, A New Hybrid K-means and K-Nearest-Neighbor Algorithms for Text Document Clustering, International Journal of Academic Research, Vol. 6 Issue 3, pp. 7984, 2014.