ششمین کنفرانس ملّی
محاسبات توزیعی و پردازش داده های بزرگ
۱۳ اسفند ۱۳۹۹
6th National Conference on Distributed Computing and Big Data Processing
ISC
دانشگاه شهید مدنی آذربایجان

انتخاب ویژگی در متون فارسی با استفاده از ترکیب الگوریتم‌های فراشناختی

لیلا هاوشکی[۱]، محبوبه شمسی[۲]، عبدالرضا رسولی کناری[۳]

[۱] دانشجوی کارشناسی ارشد نرم افزار، دانشکده برق و کامپیوتر، دانشگاه صنعتی قم،
haveshki.l@qut.ac.ir

[۲] استادیار، دانشکده برق و کامپیوتر، دانشگاه صنعتی قم،
shamsi@qut.ac.ir

[۳] استادیار، دانشکده برق و کامپیوتر، دانشگاه صنعتی قم،
rasouli@qut.ac.ir

چکیده

انتخاب ویژگی[۱]به طور گسترده در زمینه‌ی متن کاوی[۲]برای ایجاد یک مدل با تعداد ویژگی‌های کمتر استفاده می‌شود. انتخاب ویژگی متن یک گام مهم در طبقه‌بندی متن است و به طور مستقیم بر عملکرد آن تأثیر می‌گذارد. در این مقاله یک روش بهبود انتخاب ویژگی برای طبقه‌بندی داده‌های بزرگ[۳]با استفاده از ترکیب الگوریتم‌های فراشناختی[۴]پیشنهاد می‌شود. در این روش از ترکیب الگوریتم بهینه‌سازی ازدحام ذره‌ها براساس لینک[۵]و الگوریتم جست و جوی گرانشی[۶]استفاده می‌شود. الگوریتم بهینه‌سازی ازدحام ذره‌ها بر اساس لینک، بهبودی از الگوریتم بهینه‌سازی ازدحام ذره‌ها[۷]است. در این پژوهش برای افزایش توان جست و جوی محلی این الگوریتم از الگوریتم جست و جوی گرانشی استفاده می‌شود. براساس ویژگی‌های انتخاب شده، مدل طبقه‌بندیِ کلاس‌بند نزدیک‌ترین همسایه[۸] ساخته می‌شود. در آخر نتایج بر اساس معیار ارزیابی مورد بررسی قرار می‌گیرد.

عملکرد الگوریتم پیشنهادی بر روی مجموعه داده همشهری مورد بررسی قرار گرفته است. این مجموعه داده با استفاده از کتابخانه هضم پیش پردازش شده است و دو مجموعه داده تصادفی ایجاد شده است. نتایج ما نشان می‌دهد که روش پیشنهادی با انتخاب تعداد ویژگی کمتر در دو مجموعه داده تولید شده به ترتیب به افزایش دقت ۳/۹۷٪ و ۱/۵۸٪ رسیده است.

کلمات کلیدی

انتخاب ویژگی، داده‌های بزرگ، الگوریتم جست و جوی گرانشی، الگوریتم بهینه سازی ازدحام ذره‌ها، متن کاوی، پیش پردازش متن.

۱- مقدمه

در طی سال‌های اخیر حجم داده‌هایی با ابعاد بالا افزایش قابل توجهی پیدا کرده است. بنابراین روش‌های یادگیری ماشین در مواجهه با این حجم داده‌ها با تعداد ویژگی‌های زیاد دچار مشکل شده است. به همین جهت انتخاب ویژگی به یکی از زمینه‌های تحقیقی مهم و چالش برانگیز در یادگیری ما شین تبدیل شده ا ست. برای اینکه روش‌های یادگیری ماشین به طور موثر عمل کنند، اعمال پیش‌پردازش روی داده ها ضروری به نظر می‌رسد. یکی از تکنیک‌های پرکاربرد در پیش‌پردازش داده ها در یادگیری ماشین، انتخاب ویژگی است. انتخاب ویژگی فرآیندی جهت تشخیص ویژگی‌های مناسب و حذف ویژگی‌های نامناسب، تکراری یا داده‌های دارای اختلال است [۳]. چارچوب یک فرآیند انتخاب ویژگی معمولی شامل سه مرحله‌ی ا سا سیِ تابع تولید کننده، تابع ارزیابی و شرط خاتمه است. در مرحله‌ی اول زیر مجموعه‌ی

کاندید انتخاب می شود. در مرحله‌ی دوم این زیر مجموعه مورد ارزیابی قرار می‌گیرد. در مرحله‌ی سوم بهترین معیار قبلی مطابق با یک معیار ارزیابی خاص، مقایسه می‌شود. اگر زیر مجموعه‌ی تازه تاسیس شده از زیر مجموعه‌ی قبلی بهتر باشد، آنگاه آن آخرین زیر مجموعه خواهد بود. دو مرحله‌ی اول انتخاب ویژگی مبتنی بر جست و جو تا زمانی تکرار می‌شوند که یک معیار توقف معین ارضا شود. با توجه به معیار ارزیابی، الگوریتم های انتخاب ویژگی در مدل های فیلتر، لفافه و ترکیبی یا تعبیه شده[۱] طبقه‌بندی می‌شوند [۴]. در روش فیلتر، ویژگی‌های مربوطه بدون استفاده از الگوریتم یادگیری تعریف می شود. از سوی دیگر، روش لفافه از روش یادگیری برای انتخاب ویژگی اطلاعاتی استفاده می‌کند. به طور کلی روش لفافه از روش فیلتر از نظر دقت طبقه بندی بهتر است اما روش فیلتر سریع تر و ساده تر است. روش ترکیبی از ادغام استفاده می‌کند. یعنی انتخاب ویژگی شامل ساختاری از مدل است [۵]. تکنیک‌های سنتی بسیاری برای انتخاب ویژگی وجود دارند. بهبود این تکنیک‌ها مشکل است زیرا آن‌ها فاقد مدل‌های ریاضی هستند [۶].

در دو دهه‌ی گذشته الگوریتم‌های فراشناختی به هنگام حل مسائل بهینه‌سازی متنوع مانند طراحی مهندسی، یادگیری ماشین، برنامه‌ریزی استخراج داده‌ها، مشکلات تولید و ... بسیار قابل‌اطمینان بوده‌اند. انتخاب ویژگی یکی از حوزه هایی است که در آن الگوریتم های فراشناختی مورد استفاده قرار گرفته‌اند. الگوریتم‌های فراشناختی عملکرد برتری را در مقایسه با جست و جوی دقیق و کامل نشان می‌دهند زیرا جست و جوی کامل منجر به تولید همه راه‌حل‌های ممکن برای این مشکل می شود. به عنوان مثال، اگر مجموعه‌ی داده‌ها شامل N ویژگی باشد، آنگاه 2^N راه‌حل باید تولید و ارزیابی شود که به هزینه‌ی محاسبه بالایی نیاز دارد. جست و جوی تصادفی یک استراتژی دیگر برای انتخاب ویژگی است که گام بعدی را به طور تصادفی جست و جو می‌کند. الگوریتم‌های فراشناختی از جست و جوی تصادفی نیز بهتر در نظر گرفته می‌شوند، زیرا ممکن است به عنوان یک جست و جوی کامل در بدترین حالت عمل کنند. با وجود آنکه الگوریتم‌های فراشناختی، استراتژی پیدا کردن بهترین راه‌حل را در هر اجرا تضمین نمی‌کنند اما ممکن است راه‌حل قابل قبولی را در زمان معقول تعیین کنند [۷]. در سال‌های اخیر بسیاری از پژوهشگران در زمینه‌ی بهینه‌سازی انتخاب ویژگی کار کرده‌اند. با توجه به تحقیق‌های فراوان، ترکیب الگوریتم‌های فراشناختی کارایی بالاتری را در حل بسیاری از مسائل عملی نشان داده‌اند. تاکنون اکثر پژوهش‌های انجام شده بر روی سندهایی به زبان انگلیسی بوده است. در ادامه به کارهای مرتبط پیشین برای انتخاب ویژگی بر روی زبان‌های فارسی و انگلیسی پرداخته‌ایم.

در پژوهشی یک سیستم برای طبقه‌بندی متون فارسی ارائه شده است. در این سیستم پس از پیش پردازش متن‌ها و استخراج ویژگی، برای کاهش ابعاد بردار ویژگی یک روش بهبود یافته انتخاب ویژگی مبتنی بر الگوریتم ازدحام ذره‌ها استفاده شده است. در نهایت روش‌های طبقه‌بندی بر بردار ویژگیِ کاهش داده شده، اعمال شده است. برای ارزیابی روش انتخاب ویژگی، طبقه بندی کننده های ماشین بردار پشتیبان[۲] و بیزین ساده[۳] به کار گرفته شده است [۱]. پژوهشی یک روش ترکیبی دو مرحله‌ای در قسمت انتخاب ویژگی با الگوریتم‌های بهره اطلاعات[۴] و همبستگی براساس زیرمجموعه‌ی انتخاب ویژگی[۵] پیشنهاد داده است. در روش بهره اطلاعات با داشتن یک مجموعه‌ی آموزشی می‌توان بهره اطلاعات را برای هر واژه محاسبه نمود. حملاتی که بهره اطلاعاتی آن‌ها از یک حد آستانه کمتر است از فضای ویژگی حذف می‌شوند. سپس در روش همبستگی براساس زیرمجموعه‌ی انتخاب ویژگی، زیرمجموعه‌ی ویژگی‌ها براساس قابلیت پیش بینی به همراه درجه افزونگی بین آن‌ها ارزیابی می‌شود. در نتیجه ویژگی‌هایی که همبستگی بالایی با متغیر هدف داشته باشند و در عین حال همبستگی بین خود ویژگی‌ها پایین باشد ترجیح داده می‌شوند. در آخر از الگوریتم‌های ترکیبی ماشین برای بررسی نتایج استفاده شده است [۲]. در تحقیقی الگوریتمی پیشنهاد شده است که از یک استراتژی جست و جوی محلی استفاده می‌کند. این روش به گونه‌ای عمل می‌کند که ویژگی‌های غیر متمایز با احتمال بالا از ویژگی‌های مرتبط‌تر انتخاب می‌شوند. در این روش تعداد قابل توجهی از ویژگی‌های برجسته را با استفاده از یک طرح تعیین اندازه زیر مجموعه انتخاب می‌کند. این طرح بر روی یک منطقه محدود کار می‌کند و تلاش می‌کند تا یک زیرمجموعه با تعداد کم از ویژگی‌ها را ارائه دهد [۸]. دو مدل ترکیبی برای طراحی تکنیک های انتخاب ویژگی مختلف براساس الگوریتم بهینه‌سازی نهنگ[۶] پیشنهاد شده است. در مدل اول، الگوریتم تبرید شبیه سازی شده[۷] در الگوریتم بهینه سازی نهنگ تعبیه شده است. برای بهبود بهترین راه‌حل بعد از هر تکرار، الگوریتم بهینه سازی نهنگ در مدل دوم استفاده می‌شود. در واقع با استفاده از الگوریتم تبرید شبیه سازی شده، افزایش بهره‌برداری از طریق جست و جوی مناطق امیدبخش در الگوریتم بهینه‌سازی نهنگ اتفاق می‌افتد [۷]. پژوهشی با استفاده از انتخاب ویژگی به تجزیه و تحلیل احساسات پرداخته است. در این پژوهش یک روش انتخاب ویژگی بر اساس فیلتر برای بخش حریصانه‌ی الگوریتم IG ساخته شده است. نتایج نشان می‌دهد که بهبود در انتخاب ویژگی باعث شده است که دقت طبقه‌بندی درصد قابل قبولی افزایش داشته باشد[۹]. پژوهشی الگوریتم انتخاب ویژگی به نام MA-HS را براساس دو الگوریتم Mayfly و Harmon Search ارائه کرده است. این دو الگوریتم‌های فراشناختی هستند که برای مسئله‌های بهینه سازی استفاده می‌شوند. و نتایج نشان می‌دهد ترکیب

الگوریتم‌های فراشـــناختی باعث بهبود انتخاب ویژگی می‌شـــوند [۱۰]. پژوهشی یک روش انتخاب جدید ویژگی برای خوشه بندی متن، به نام بهینه سازی ازدحام ذره‌ها برا ساس لینک را پی شنهاد داده ا ست. این روش یک اســـتراتژی انتخاب همســـایه جدید در الگوریتم بهینه ســـازی ازد حام ذره ها را برای انتخاب ویژگی های برجســـته معرفی می‌کند. عملکرد بهینه ســـازی ذره‌ها در الگوریتم ازدحام ذره‌ها با اســـتفاده از بهترین مکانی سم به روز ر سانی سرا سری برای انتخاب ویژگی محدود شده ا ست. در روش پی شنهادی به جای ا ستفاده از بهترین سرا سری، ذره‌ها بر اســـاس بهترین موقعیت همســـایگی برای افزایش قابلیت بهره برداری و اکت شاف به روز می شوند. این ویژگی‌های برج سته سپس با اســـتفاده از الگوریتم خوشـــه بندی کی - میانگین برای بهبود عملکرد و کاهش هزینه زمان محا سبه الگوریتم پی شنهاد شده، مورد آزمایش قرار گرفته است [۵].

پژوهشگران در اکثر کارهای پی شین به این مو ضوع توجه دا شته‌اند که الگوریتم‌های فراشـــناختی را به صـــورتی ترکیب کنند تا دقت در انتخاب ویژگی را افزایش دهند. گرچه این الگوریتم‌ها در جست و جوی سرا سری موفق عمل می‌کنند اما اغلب در ج ست و جوی محلی دچار مشـــکل می‌شـــوند. ما در این پژوهش با اســـتفاده از الگوریتم گرانشـــی، جست و جوی محلی را در الگوریتم بهینه سازی ازدحام ذره‌ها براساس لینک بهبود بخشیده‌ایم.

۲- مطالب اصلی

یکی از مسائل مهم در متن کاوی، طبقه بندی متن‌ها است. طبقه بندی متن، انتساب سندهای متنی براساس محتوا به یک یا چند طبقه از قبل تعیین شده است. در این پژوهش نیز روشی بر اساس داده‌کاوی و الگوریتم‌های انتخاب ویژگی برای طبقه‌بندی متن‌ها ارائه شده است.

۲-۱- مفاهیم اولیه مورد نیاز

روش پیشنهادی در این پژوهش در سه مرحله انجام می‌شود:

- در مرحله‌ی اول پردازش متن صورت می‌گیرد. در این مرحله با استفاده از ابزارهای مناسب و بر اساس نیاز، متن جهت سایر پردازش‌ها و تعیین دسته آماده می‌شود.
- در مرحله‌ی دوم پس از پردازش متن و استخراج ویژگی از متن، با توجه به اینکه ویژگی‌های استخراج شده همگی جهت ساخت مدل مناسب نیستند، انتخاب ویژگی به کمک ترکیب دو الگوریتم بهینه‌سازی ازدحام ذره‌ها براساس لینک و جست و جوی گرانش صورت خواهد گرفت.
- در مرحله‌ی سوم بر اساس ویژگی‌های انتخاب شده مدل طبقه‌بندی کلاس‌بند نزدیک‌ترین همسایه ساخته خواهد شد و مورد ارزیابی قرار خواهد گرفت.

۲-۱-۱- پردازش متن

برای استخراج ویژگی از متن که داده‌ای غیرساخت یافته است ابتدا باید عملیات پیش پردازش متن انجام شـــود. عملیات شـــامل مراحل نرمال‌ســـازی[۱۹] جداســـازی[۲۰] حذف کلمه های توقف[۲۱]اســـت. در نرمال‌ســـازی همه‌ی نویســـه‌های متن با معادل اســـتاندارد آن جایگزین می‌شـــوند. در این مرحله مطابق با یک ســـری قاعده دقیق و مشـــخص، فاصله‌ها و نیم فاصله‌های موجود در متن برای علامت‌هایی نظیر "ها" و "ی" غیرچسبان در انتهای کلمه‌ها و همچنین پیشوندها و پسوندهای فعل‌ساز نظیر "می"، "ام"، "ایم"، "اید" و موارد مشابه اصلاح می‌شوند. در جداســـازی، تمام متن به صـــورت کلمه‌های جداگانه در می‌آیند. در مرحله‌ی حذف کلمه‌های توقف، کلمه‌هایی مانند: از، با، در، به، برای، اگر و ... که با وجود تکرار بسیار و حضور در اغلب سندها فاقد اطلاعات معنایی هستند حذف می‌شوند [۱۳].

۲-۱-۲- الگوریتم ازدحام ذره‌ها بر اساس لینک

الگوریتم ازدحام ذره‌ها در ابتدا برای مسائل بهینه سازی پیوسته طراحی شـــد اما پس از گذشـــت زمان با اســـتفاده از تابع ســـیگموید[۲۲]آن را به خوبی برای حل کردن مسائل بهینه سازی گسسته به الگوریتم دودویی ازدحام ذره‌ها[۲۳]تبدیل کردند. همانند دیگر الگوریتم‌های فراشـــناختی، الگوریتم ازدحام ذره‌ها از بعضی مشکل‌های ذاتی مانند همگرایی قبل از موقع و گیر انداختن خود در حداقل‌های محلی رنج می‌برد. در چنین شرایطی جست و جوی همسایه‌های شخصی برای پیدا کردن بهترین راه حل مفید اســـت. در این الگوریتم رتبه هر ذره از طریق ماتریس ارتباط و همسایه هر ذره بر ا ساس رتبه‌ها محا سبه می شود. درواقع به جای یادگیری از بهترین مکان ســـراســـری یافت شـــده از بهترین مکان همسایه یافت شده ا ستفاده می شود. سرعت و موقعیت الگوریتم بهینه سازی ازدحام ذرات براساس لینک به ترتیب با توجه به معادله‌های (۱) و (۲) به روز می‌شود [۵].

$$v_i^j(t+1) = w * v_i^j(t) + rand_1 * c_1 * (pbest_i^j(t) - x_i^j(t)) + rand_2 * c_2 * (nbest_i^j(t) - x_i^j(t)) \quad (۱)$$

$$x_i^j(t+1) = x_i^j(t) + v_i^j(t+1) \quad (۲)$$

v(t): سرعت فعلی ذره i

x(t): مکان فعلی ذره i

rand: عدد تصادفی بین ۰ و ۱

c: ضریب شتاب

pbest: بهترین مکان شخصی یافت شده توسط ذره i

nbest: بهترین مکان همسایه یافت شده توسط ذره i

۲-۱-۳- الگوریتم جست و جوی گرانش

هدف این الگوریتم پیدا کردن بهترین راه حل در فضای جست و جوی مسأله با استفاده از نظریه نیوتن است. این نظریه بیان می‌کند: «هر ذره در جهان، ذره دیگر را با یک نیرویی که به طور مستقیم با حاصل ضرب جرم آن دو ذره و به طور معکوس با مجذور فاصله‌ی بین آن‌ها رابطه دارد، جذب می‌کند.» الگوریتم جست و جوی گرانش را می‌توان به عنوان مجموعه‌ای از راه حل‌های نامزد در نظر گرفت که دارای جرم متناسب با مقدار خود در تابع شایستگی هستند. در طول اجرای الگوریتم تمام توده‌ها توسط نیروی جاذبه بین خودشان یکدیگر را جذب می‌کنند. هر چه جرم سنگین‌تر باشد، نیروی جاذبه بیشتر است. بنابراین، جرم‌های سنگین‌تر که به احتمال زیاد نزدیک‌ترین جرم‌ها به کمینه کلی هستند، جرم‌های دیگر را متناسب با فاصله آن‌ها جذب می‌کنند.

در الگوریتم جست و جوی گرانش هر عامل یک راه حل نامزد تلقی می‌شود. در این الگوریتم تمام عوامل با مقادیر تصادفی مقداردهی اولیه می‌شوند. پس از مقداردهی اولیه، سرعت و موقعیت همه عوامل به ترتیب با استفاده از معادله‌های (۳) و (۴) محاسبه می‌شود.

$$v_i^d(t+1) = rand_i \times v_i^d(t) + a_i^d(t) \quad (۳)$$

$$x_i^d(t+1) = x_i^d(t) + v_i^d(t+1) \quad (۴)$$

که در معادله‌ها d نشان از بعد مسأله، t نشان از یک زمان خاص، $rand$ عدد تصادفی بین ۰ و ۱ است. a نیز نشان‌دهنده‌ی شتاب است. با توجه به قانون حرکت، شتابِ یک عامل به طور مستقیم با نیروی برآیند و به طور معکوس با جرم آن رابطه دارد. شتاب با توجه به معادله (۵) محاسبه می‌شود.

$$a_i^d(t) = \frac{(F_i^d(t))}{M_{ii}(t)} \quad (۵)$$

که در آن F_i^d نیروی گرانش و M_i جرم جسم i-ام است. پارامترهای دیگر مانند ثابت گرانش و نیروی گرانش از روابط (۶) و (۷) محاسبه می‌شوند.

$$G(t) = G_0 \times \exp(-a \times iter / mat iter) \quad (۶)$$

که در آن a ضریب نزولی، G_0 ثابت گرانش اولیه، $iter$ تعداد تکرار فعلی و $\max iter$ نشان دهنده‌ی حداکثر تعداد تکرار است.

$$F_{ij}^d(t) = G(t)\frac{(M_{pi}(t) \times M_{aj}(t))}{(R_{ij}(t)+\varepsilon)} \times (x_j^d(t) - x_i^d(t)) \quad (۷)$$

که در آن M_{aj} جرم گرانشی مربوط به عامل j-ام، M_{pi} جرم گرانشی غیرفعال مربوط به عامل i-ام، ε یک ثابت کوچک و R_{ij} فاصله‌ی اقلیدسی بین دو عامل i و j است. جرم تمام عامل‌ها با استفاده از رابطه (۸) به روز رسانی می‌شود.

$$m_i(t) = \frac{fit_i(t) - worst(t)}{best(t) - worst(t)} \quad (۸)$$

که در آن، $fit_i(t)$ ارزش شایستگی عامل i-ام در زمان t، $best(t)$ قوی‌ترین عامل در زمان t و $worst(t)$ ضعیف‌ترین عامل در زمان t است. در نهایت الگوریتم جست و جوی گرانش با رسیدن به معیار نهایی متوقف می‌شود [۱۱].

۲-۱-۴- الگوریتم کلاس‌بند نزدیک‌ترین همسایه

الگوریتم کلاس‌بند نزدیک‌ترین همسایه یک الگوریتم دسته‌بندی مبتنی بر نمونه، با ناظر است. اساس کار این الگوریتم مقایسه‌ی میزان شباهت نمونه‌ی جدید با نمونه‌های موجود در داده آموزشی است. جهت اینکه الگوریتم، کلاسِ نمونه‌ی جدید را تشخیص دهد، K نزدیک‌ترین عضو مجموعه‌ی آموزشی نسبت به نمونه جدید را انتخاب می‌کند. کلاسی که دارای بیش‌ترین عضو در این K عضو باشد به نمونه‌ی جدید تعلق می‌گیرد [۱۲].

۲-۲- طرح مسأله

یکی از مسأله‌های موجود در انتخاب ویژگی به وسیله‌ی الگوریتم‌های فراشناختی، همگرایی زودرس و گیر افتادن در نقاط محلی است. این مسأله باعث می‌شود که با انتخاب تعداد ویژگی‌های زیاد به خصوص در داده‌های بزرگ، زمان پردازش افزایش یابد. با بهبود عملکرد در نقاط محلیِ یک الگوریتم فراشناختی که در پژوهش‌های پیشین نتایج قابل قبولی داشته است می‌توان به این هدف رسید که با انتخاب تعداد ویژگی کمتر به دقت بالاتری دست یافت. همچنین از الگوریتم‌های ترکیبی فراشناختی برای انتخاب ویژگی در متون فارسی کم استفاده شده است. با پیش پردازش سندها در زبان فارسی و با استفاده از الگوریتم کلاس‌بند نزدیک‌ترین همسایه می‌توان انتخاب ویژگی در زبان فارسی را بدون به خطر انداختن دقت بهبود بخشید.

۲-۲- راه حل پیشنهادی

ایده اصلی ترکیبِ الگوریتم‌های ازدحام ذره‌ها و جست و جوی گرانشی در واقع توان جست و جوی سراسری ازدحام ذره‌ها و جست و جوی محلی الگوریتم گرانشی است. در همه‌ی الگوریتم‌های مبتنی بر جمعیت دو ویژگیِ توانایی الگوریتم برای جست و جوی تمام بخش‌ها و توانایی بهره‌برداری از بهترین راه حل در نظر گرفته می‌شود. یک الگوریتم مبتنی بر جمعیت باید این دو ویژگی را برای تضمین پیدا کردن بهترین راه حل داشته باشد. الگوریتم ازدحام ذره‌ها توانایی بالایی را در جست و جوی سراسری در تمام بخش‌ها دارد. در الگوریتم ازدحام ذره‌ها بر اساس لینک که پیش‌تر توضیح داده شد، با انتخاب بهترین همسایه، بهبودی در توانایی بهره‌برداری از بهترین راه حل ایجاد شد. اما الگوریتم همچنان برای توانایی بهره‌برداری از بهترین راه حل، دچار گیر افتادن در نقاط محلی می‌شود. توانایی بالای الگوریتم گرانش در جست و جوی محلی و بهره‌برداری از بهترین راه حل، باعث بهبود الگوریتم ازدحام ذره‌ها بر اساس لینک شده است. در نتیجه، این الگوریتمِ ترکیبی به دقت بالایی در انتخاب ویژگی رسیده است.

۲-۲-۱- ترکیب الگوریتم بهینه‌سازی ازدحام ذره‌ها براساس لینک و الگوریتم جست و جوی گرانشی

برای ترکیب این دو الگوریتم رابطه‌ی (۹) پیشنهاد شده است.

$$v_i(t+1) = w \times v_i(t) + c_1^{'} \times rand \times ac_i(t) + c_2^{'} \times rand \times (nbest - x_i(t)) \quad (۹)$$

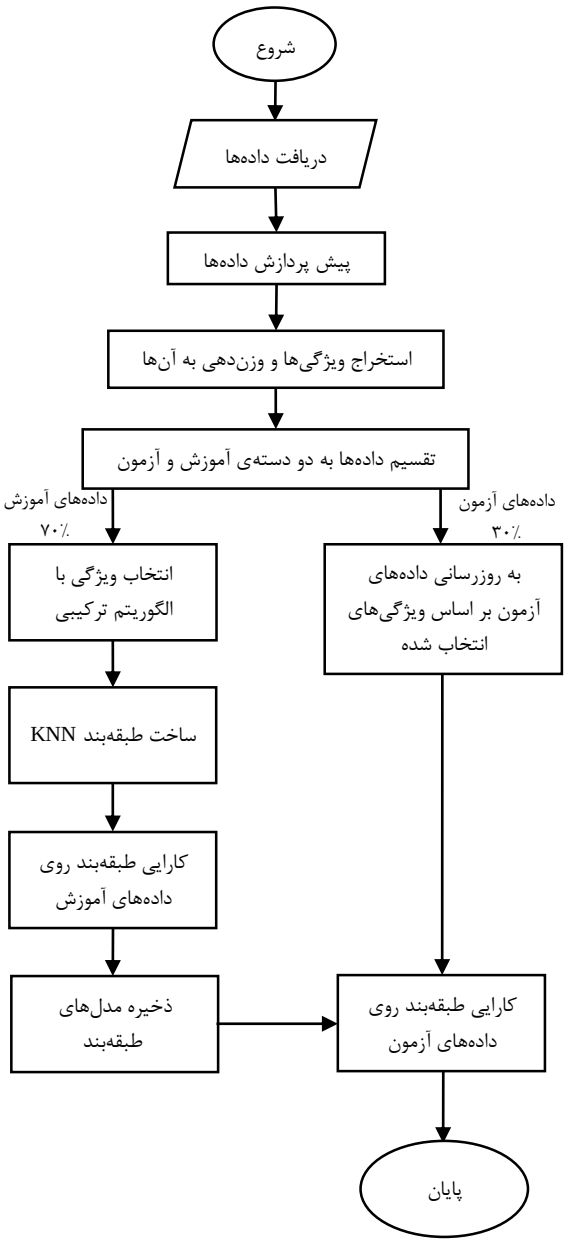
که در آن، c' ضریب شتاب، w تابع وزن و $ac_i(t)$ شتاب عامل i-ام در تکرار t-ام است. موقعیت عوامل نیز در هر تکرار با توجه به رابطه (۱۰) به روزرسانی می‌شود.

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad (۱۰)$$

در این روش هر عامل به عنوان یک راه حل نامزد در نظر گرفته می‌شود. در ابتدا همه عوامل به صورت تصادفی مقداردهی اولیه می‌شوند. پس از مقداردهی اولیه، ثابت گرانش و نیروی گرانش به ترتیب با استفاده از رابطه‌های (۶) و (۷) محاسبه می‌شوند. سپس شتاب ذره‌ها به وسیله رابطه (۵) تعریف می‌شود. بهترین راه حل پیدا شده در هر تکرار باید به روز شود. پس از محاسبه شتاب و به روزرسانی بهترین راه حل، سرعت همه عامل‌ها با استفاده از رابطه (۹) محاسبه می‌شود. در نهایت، موقعیت عامل‌ها توسط رابطه (۱۰) به روز می‌شود. پس از انتخاب ویژگی، الگوریتم کلاس‌بند نزدیک‌ترین همسایه، برای تخمین دسته‌های داده‌ها استفاده شده است. انتخاب ویژگی در روش پیشنهادی روشی مبتنی بر لفافه است. در این روش از کارایی مدل طبقه‌بند برای محاسبه کارایی هر ذره استفاده خواهد شد. بر همین اساس رابطه (۱۱) نحوه محاسبه کارایی ذره‌ها در روش پیشنهادی را نشان می‌دهد.

$$Error = 1 - \frac{Correct\ Detect\ Sample}{Total\ Sample} \quad (۱۱)$$

که Correct Detect Sample بیان‌گر تعداد نمونه‌هایی است که به درستی تشخیص داده شده‌اند. Total Sample تعداد کل نمونه‌ها است و Error میزان خطا را نشان می‌دهد. هدف از انتخاب ویژگی، کاهش مقدار خطا در رابطه (۱۱) است. روش پیشنهادی در شکل شماره (۱) نشان داده شده است.



شکل (۱) : مراحل روش پیشنهادی

برای شبیه سازی الگوریتم پیشنهادی از برنامه Matlab2020 استفاده شده است. برای ساخت مجموعه داده و استخراج اطلاعات از این سندها ابتدا عمل پیش پردازش بر روی سندها انجام شده است. عملیات پیش پردازش با استفاده از کتابخانه هضم [۱۴] در python و در محیط colab صورت گرفته است. پس از انجام پیش پردازش، خروجی در محیط متلب به صورت یک دیکشنری در آمده است. سپس عملیات TF-IDF روی آن انجام شده است. ما برای پیش پردازش متون فارسی از مجموعه داده همشهری استفاده کرده‌ایم. مجموعه داده همشهری مربوط به متن اخبار و گروه اخبارهای منتشر شده در روزنامه همشهری مربوط به سال های ۱۳۷۵-۱۳۸۷ است. مجموعه داده همشهری شامل ۱۰۰۰۰ سند در ده گروهی خبری به فرمت اکسل است. دو مجموعه داده از داده همشهری به صورت تصادفی با کلاس‌های متفاوت ایجاد کرده‌ایم. مجموعه داده اول با نام Data1، مجموعه ای با ۸۰ سند شامل کلاس‌های علم و فرهنگ، سیاسی و اتفاقات متفرقه است. مجموعه داده دوم با نام Data2، مجموعه ای با ۱۲۱ سند شامل کلاس‌های اقتصادی و سیاسی است. مشخصه‌های دو مجموعه داده در جدول شماره (۱) آورده شده است.

جدول شماره (۱): مشخصه‌های مجموعه داده

مجموعه داده	تعداد کلاس	تعداد سند	تعداد ویژگی‌ها
Data1	۳	۸۰	۴۷۷۹
Data2	۲	۱۲۱	۴۵۱۲

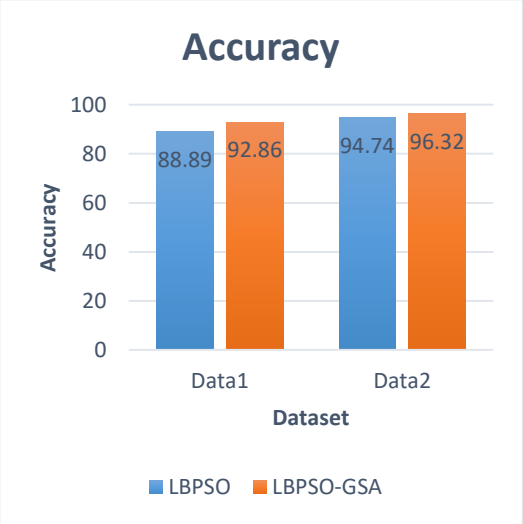
برای بررسی عملکرد الگوریتم پیشنهادی از معیار دقت برای طبقه-بندی و میزان تعداد ویژگی انتخاب شده برای رسیدن به دقت بالا استفاده کرده‌ایم. دقت درواقع محاسبه‌ی درصد برچسب‌ها[۵]برای تعیین این که مدل تا چه اندازه خروجی را درست پیش بینی می‌کند. هرچقدر دقت بالاتر باشد ویژگی‌های مهم‌تری انتخاب شده‌اند. نتایج به دست آمده در جدول شماره (۲) آورده شده است.

جدول شماره (۲): مقدار پارامترها و نتایج

مجموعه داده	تعداد ویژگی انتخاب شده	تعداد تکرار	دقت
Data1	۱۰۰۰	۱۰	۹۲/۸۶
Data2	۱۰۰	۱۰	۹۶/۳۲

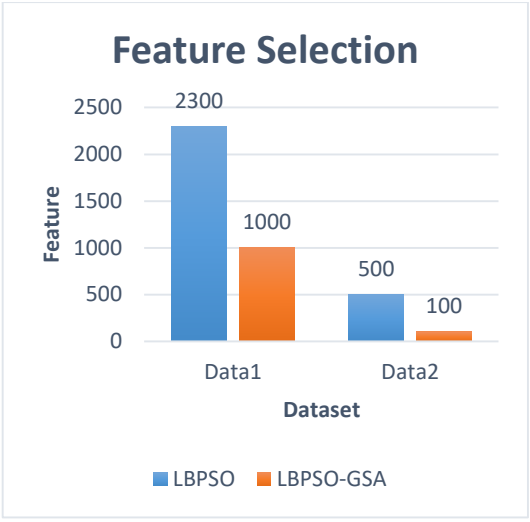
شکل شماره (۲) میزان دقت را در روش پیشنهادی و روش ازدحام ذره‌ها براساس لینک نشان می‌دهد. این نکته حائز اهمیت است که تعداد تکرار در روش پیشنهادی ۱۰ است و بسیار کمتر از تعداد تکرار ۱۰۰ در روش ازدحام ذره‌ها براساس لینک است. روش پیشنهادی در شکل‌های (۲) و (۳) با نام اختصاری LBPSO-GSA و روش ازدحام ذره‌ها بر اساس لینک با نام LBPSO نشان داده شده است.

همان طور که در جدول شماره (۱) آمده است تعداد ویژگی‌های هر مجموعه داده بسیار بالا است. Data1 دارای ۴۷۷۹ ویژگی است که با توجه به جدول شماره (۲) با انتخاب ۱۰۰۰ ویژگی توانسته‌ایم به دقت ۹۲/۸۶٪ برسیم اما همانطور که در شکل (۳) مشخص است روش ازدحام ذره‌ها بر اساس لینک (LBPSO)، تعداد ۲۳۰۰ ویژگی را انتخاب کرده است که با توجه به شکل (۲) به دقت ۸۸/۸۹٪ رسیده است. همین طور Data2 دارای ۴۵۱۲ ویژگی است که با توجه به جدول شماره (۲) ما در روش پیشنهادی (LBPSO-GSA) با انتخاب ۱۰۰ ویژگی توانسته‌ایم به دقت ۹۶/۳۲٪ برسیم. اما همانطور که در شکل (۳) مشخص است روش ازدحام ذره‌ها بر اساس لینک، تعداد ۵۰۰ ویژگی را انتخاب کرده است که با توجه به شکل (۲) به دقت ۹۴/۷۴٪ رسیده است. پس در این پژوهش ما با انتخاب ویژگی‌های کم‌تر و تعداد تکرار کم‌تر به دقت بالاتری رسیده‌ایم.



شکل (۲): مقایسه‌ی عملکرد دو الگوریتم LBPSO و LBPSO-GSA با استفاده از اندازه گیری دقت

شکل شماره (۳) تعداد ویژگی انتخاب شده برای هر داده را در دو روش نشان می‌دهد. تعداد ویژگی انتخاب شده در روش پیشنهادی به مراتب کمتر از روش LBPSO است. این موضوع نشان می‌دهد که روش پیشنهادی توانسته است ویژگی‌های برجسته‌تری را انتخاب کند.



شکل (۳): تعداد ویژگی انتخاب شده برای دو مجموعه داده با روش‌های LBPSO و LBPSO-GSA

با توجه به نتایج و تحلیل داده‌ها می‌توان متوجه این موضوع شد که داده‌هایی با تعداد کلاس بالاتر، در اغلب موارد نیاز به انتخاب ویژگی بیشتری دارند تا بتوانند به دقت قابل قبولی برسند. زیرا به علت وجود موضوع‌های متنوع، ویژگی‌های برجسته بیشتری دارند. مجموعه داده اول با وجود آنکه تعداد سند کمتری دارد، دارای ویژگی بیشتری است. به همین علت برای رسیدن به دقت مطلوب تعداد ویژگی بالاتری برایش انتخاب شده است.

۳- نتیجه

این مقاله یک روش انتخاب ویژگی جدید با ترکیب الگوریتم‌های ازدحام ذره‌ها بر اساس لینک و جست و جوی گرانش ارائه می‌کند. انتخاب ویژگی نقشی مهم در کاهش هزینه محاسبه، ساده کردن یادگیری مدل و ارتقا توانایی طبقه‌بندی کننده در مسأله‌های طبقه‌بندی دارد. در این پژوهش روش انتخاب ویژگی روی داده‌های فارسی انجام شده است. برای ایجاد داده‌ها دو مجموعه داده تصادفی با کلاس‌های متفاوت از مجموعه داده همشهری ایجاد شده است. پیش از انجام انتخاب ویژگی، داده‌ها پیش پردازش شده‌اند و توسط روش TF-IDF وزن‌دهی شده‌اند. روش پیشنهادی جست و جوی محلی را توسط الگوریتم جست و جوی گرانش افزایش می‌دهد. این بهبود موجب شده است که با انتخاب تعداد ویژگی کمتر و انجام تعداد تکرار کمتر دقت بالاتری به دست آید.

مراجع

[۱] رشدی، اکرم، اکبرپور، شاهین، **"روش ترکیبی انتخاب ویژگی برای متن کاوی فارسی مبتنی بر الگوریتم‌های تکاملی"**، همایش ملی مهندسی رایانه و مدیریت فناوری اطلاعات، CSITM01، ۸، شرکت علم و طلوع فرزین، ۱۳۹۳.

[۲] هاشمی، سید محسن، **"بهبود دسته‌بندی متون فارسی با ترکیب روش دو مرحله‌ای انتخاب ویژگی و الگوریتم‌های یادگیری ماشین"**، کنفرانس بین‌المللی یافته‌های نوین پژوهشی در مهندسی برق و کامپیوتر، COMCONF01، ۹، موسسه آموزش عالی نیکان، ۱۳۹۴.

[3] Kumar, V.and S.Minz, *Feature selection: a literature review*. SmartCR, 2014. 4(3): p. 211-229.

[4] ۱. Li, Y., T. Li, and H. Liu, *Recent advances in feature selection and its applications*. Knowledge and Information Systems, 2017. 53(3): p. 551-577.

[5] Kushwaha, N. and M. Pant, *Link based BPSO for feature selection in big data text clustering.* Future Generation Computer Systems, 2018. 82: p. 190-۱۹۹.

[6] Abualigah, L.M. and A.T. Khader, *Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering.* The Journal of Supercomputing, 2017. ۷۳(۱۱): □. ۴۷۷۳-۴۷۹۵.

[7] Mafarja, M.M. and S. Mirjalili, *Hybrid whale optimization algorithm with simulated annealing for feature selection.* Neurocomputing, 2017. 260: p. 302-۳۱۲.

[8] Moradi, P. and M. Gholampour, *A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy.* Applied Soft Computing, ۲۰۱۶. ۴۳: □. ۱۱۷-۱۳۰.

[9] Lim, H. and D.-W. Kim, *Generalized Term Similarity for Feature Selection in Text Classification Using Quadratic Programming.* Entropy, 2020. 22(4): p. 395.

[10] Bhattacharyya, T., et al., Mayfly in harmony: *A new hybrid meta-heuristic feature selection algorithm.* IEEE Access, ۲۰۲۰. ۸: □. ۱۹۵۹۲۹-۱۹۵۹۴۵

[11] Mirjalili, S. and S.Z.M. Hashim. *A new hybrid PSOGSA algorithm for function optimization.* in 2010 international conference on computer and information application. ۲۰۱۰.

[12] Kumar, R.R., P. Viswanath, and C.S. Bindu, *Nearest neighbor classifiers: a review.* Int. J. Comput. Intell. Res, ۲۰۱۷. ۱۳(۲): □. ۳۰۳-۳۱۱.

[13] https://blog.text-mining.ir/text-preprocessing o2wzgzkagzyj/

[14] https://www.sobhe.ir/hazm/

زیرنویس‌ها

۱ Feature Selection

۲ Text Mining

[۳] Big Data
[٤] meta-heuristic
[٥] Link based BPSO (LBPSO)
[٦] Gravitational Search Algorithm (GSA)
[۷] Particle swarm optimization (PSO)
[۸] K-nearest neighbors (KNN)
[۹] Filter
[۱۰] Wrapper
[۱۱] Hybrid or embedded
[۱۲] Support vector machines (SVM)
[۱۳] Naive Bayes
[۱٤] Information gain
[۱٥] Correlation based on Feature Selection (CFS)
[۱٦] Whale Optimization Algorithm
[۱۷] Simulated Annealing Algorithm
[۱۸] iterated greedy
[۱۹] Normalize
[۲۰] Tokenize
[۲۱] Stop Words
[۲۲] Sigmoid function
[۲۳] BPSO
[۲٤] Term Frequency- Inverse Document Frequency
[۲٥] label