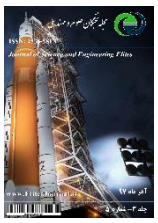
مجله نخبگان علوم و مهندسی

Journal of Science and Engineering Elites

جلد ۳- شماره ۵- سال ۱۳۹۷

www.ElitesJournal.ir



تکنیکی با هدف بهبود سرعت و دقت الگوریتم دسته بندی کنندهKNN

شیما شفیعی[۱*]، ناهید نخعی[۲]

۱- کارشناس ارشد کامپیوتر و مدرس آموزش و پرورش

۲- کارشناس ارشد کامپیوتر و مدرس دانشگاه آزاد اسلامی نهبندان

*Sh.shafiee2018@gmail.com

ارسال: آذر ماه ۹۷ پذیرش: آذر ماه ۹۷

چکیده

در سال های اخیر فرآیند داده‌ها، اطلاعات و دسته بندی آنان، به علت سر و کار داشتن بشر عصر ارتباطات و فناوری با حجم عظیمی از داده‌ها و دستیابی به اهدافی از قبیل : قابلیت مدیریت، اعمال نفوذ برحجم زیاد اطلاعات ، مقایسه متون مختلف و رتبه بندی مهم‌ترین متون ارتباطی، مسائل سیاسی و امنیتی یک کشور، از اهمیت خاصی برخودار می‌باشد. از اینرو، دسته بندی اسناد عبارت است از: انتساب سندهای دسته بندی نشده به یک یا چند دسته از پیش تعریف شده، به گونه ای که هر کلاسه بند، نقاط قوت و ضعف خود را دارد و به کار گیری یک تکنیک کارا هم چون الگورتیم KNN می‌تواند سرعت دست یابی به هدف اصلی را افزایش دهد. از اینرو در چند دهه اخیرتکنیک های زیادی برای مسأله دسته بندی اسناد انجام شده که در این مطالعه، تآکید بر بهینگی الگورتیم KNN است و سعی گردیده که با بکارگیری نرم افزار متلب، تکنیکی به منظور بهبود سرعت و دقت الگوریتم دسته بندی کننده KNNو مبتنی بر تابع ارزیاب پیشنهادی ارائه شود.

کلمات کلیدی: دسته بندی ، الگوریتم KNN، بهبود کیفیت ، بهبود سرعت ، استخراج ویژگی.

۱- مقدمه

باتوجه به گستردگی حجم اطلاعات متنی الکترونیکی که به طور قابل توجهی از طریق اینترنت و سایر منابع قابل دسترسی می‌باشند، در صورت نبودن ایندکس گذاری و دسته بندی مناسب، کار بازیابی و پردازش اطلاعات متنی دسته بندی نشده با محدودی‌تهای بسیاری، مواجه می‌گردد [۱]: به بیانی دیگر مسئله مورد مطالعه، همان ارائه الگورتیم KNN بهبود یافته برای دسته بندی اسناد، مادامی‌که از دقت و سرعت بالاتری در فضای ویژگی بزرگتر برخوردار باشد. از اینرو مسئله KNN عبارت است از: مجموعه S شامل تعدادی نقطه در یک فضای متریک مانند M و نیز یک نقطه ی پرس و جوی $q\in M$ داده شده که هدف پیدا کردن نزدیک ترین نقطه در S به q است [۲]. در بسیاری از موارد، فضای M بصورت یک فضای اقلیدسی d- بعدی و فاصله بین نقاط با معیار فاصله اقلیدسی، فاصله منهتن یا دیگر فاصله‌های متریک سنجیده می‌شود.

مبنای الگوریتم KNN همان پیدا کردن تعداد معینی از نزدیکترین عناصر موجود در جامعه آماری به عنصر جدید واردشده در آن جامعه است که بر اساس آن بتوان نزدیکترین داده (ها) موجود به عنصر جدید را از لحاظ ویژگی های مختلف پیدا نموده و بتوان عنصر جدید را در همان طبقه ای قرار داده که عناصر نزدیک به آن قرار دارند [۳].

KNN یکی از روش های غیر پارامتریک برای بدست آوردن تابع توزیع از روی داده های توزیع شده می‌باشد. همچنین این تکنیک، یکی از متداول ترین روش ها برای دسته بندی داده‌ها می باشدکه [۴] : در آن، یک سند و یا داده ی آموزشی برای دسته بندی وجود دارد و الگوریتم K همسایه نزدیک در میان سندهای آموزشی پیش دسته بندی شده، بر اساس یک معیار شباهت پیدا نموده و دسته های این K همسایه نزدیک برای پیش بینی دسته سند آزمایشی به وسیله امتیاز دهی سندهای هر دسته منتخب، استفاده خواهد شد.ودر باب انتخاب الگورتیم KNN هم بایستی مطرح نمود که:

کاهش بعد یکی از زیر مجموعه های استخراج ویژگی و به معنا استخراج ویژگی از مجموعه داده مفید، متناهی و معین است. از اینرو الگوریتم KNN بهبود یافته برای دسته بندی داده های مفید پیشنهاد شده که از دقت و سرعت بهینه تر، که در فضای ویژگی بزرگتری اعمال شده نیز برخوردار باشد[۵]. این الگورتیم که یکی از پر کاربردترین الگوریتم‌های طبقه بندی، می‌باشد و در اکثر مسائل طبقه بندی دقت خوبی از خود نشان می‌دهد. این الگوریتم دارای پیچیدگی محاسباتی زیادی می باشد؛ زیرا برای طبقه بندی کردن هر الگوی تست باید فاصله آن را تا تمام داده های آموزشی پیدا کند. در ضمن برای نگهداری تمام داده های آموزشی نیاز به فضای ذخیره سازی زیادی دارد. سیستم‌های فعلی کاهش بعد با مشکلاتی از جمله دقت پایین دسته‌بندی، غیر قابل تفسیر بودن داده‌ها و کم شدن ارزش آن‌ها، روبرو هستند. به منظور رفع محدودیت‌های مذکور، تکنیکی مبتنی بر الگوریتم بهینه‌سازی گروه ذرات ارائه شده که خطای دسته‌بندی داده‌های اولیه به روش دسته‌بندی نزدیکترین همسایه را کاهش دهد. روش پیشنهادی با بازدهی و سرعت همگرایی بالا در استخراج پاسخ بهینه نسبت با سایر روش‌های موجود موفق‌تر عمل می‌کند.به طور کلی مزایای الگوریتم KNN عبارتند از[۶,۷]:

I. این الگوریتم محدود به مدل خاصی نیست و برای دسته بندی اسناد کاربرد زیادی دارد
II. سادگی و موثر بودن در طبقه بندی متون، جلوگیری از دخالت ویژگی مانند نویزو ویژگی های نامرتبط با متن اصلی ویا داده اصلی ،جلوگیری از کاهش ابعاد فضای ویژگی، دسته بندی کارا و هدفمند اخبار، مدیریت اسناد به گونه ای مناسب تر اشاره کرد
III. می تواند توابع پیچیده را مدل کند
IV. اطلاعات موجود در مثال‌های آموزشی از بین نمی رود
V. می‌تواند از نمایش سمبلیک نمونه‌ها استفاده نموده

در مجموع، هدف از انجام این مقاله، تحقق موارد ذیل می‌باشد که عبارتند از:

الف- بررسی مفهوم کاهش بعد و روش‌های کاهش بعد ارائه شده‌ی پیشین (جهت الهام گرفتن برای ارائه تکنیک نوین)

ب- ارائه روش جدید به منظور بهبود کارایی و دقت الگوریتم دسته بند KNN با استفاده از PSO

ج- ارزیابی مبتنی بر نتایج حاصل از تابع شایستگی مسئله دسته‌بندی با روش KNN

پرسش اصلی این مقاله این است که آیا می توان به روشی با استفاده از روش های پیشین به گونه ای دست یافت که بر سرعت و دقت الگورتیم افزوده گردد؟

در ادامه و در ارائه روش جدید کاهش ابعاد سعی شده تا از معایب موجود در کارهای مرتبط پیشین پرهیز شود و یک روش جدید با تابع شایستگی کارا پیشنهاد گردد. هدف تابع شایستگی پیشنهادی همان افزایش دقت دسته‌بندی می باشد. در ادامه، بخش های‌مقاله عبارتند از: در بخش دوم پیشینه ی پژوهش، بخش سوم شامل ارائه تکنیک پیشنهادی ، بخش چهارم به ارزیابی نتایج تجربی و سرانجام در بخش پنجم به نتیجه گیری و پیشنهادات آتی کار پرداخته می‌شود.

۲- پیشینه پژوهش

کارآمدترین تکنیک‌ها در کاهش ابعاد، روش‌های مبتنی بر استخراج ویژگی و روش‌های مبتنی بر انتخاب ویژگی هستند. بکارگیری الگوریتم بهینه سازی ازدحام ذرات، در مسائل کاهش ابعاد، باعث همگرایی سریع‌تر و پیدا کردن راه حل در زمان کمتر می‌شود. الگوریتم بهینه سازی ازدحام ذرات، در اکثر مسائل بهینه‌سازی سراسری کارا عمل می‌کند. در واقع، الگورتیم بهینه سازی ازدحام ذرات، برای بدست آوردن نتیجه‌ی بهینه تر یا مشابه، نسبت به الگوریتم‌های موجود، به زمان محاسباتی کمتری نیاز دارد تا جایی که نتایج آزمایشات شبیه‌سازی شده در ادبیات، نشان می‌دهد که الگورتیم بهینه سازی ازدحام ذرات، نتایج بهتری نسبت به الگوریتم ژنتیک دارد. از اینرو در این قسمت به بررسی کارهای مشابه انجام شده در زمینه روش‌های کاهش ابعاد که در حوزه استخراج ویژگی ایجاد شده اند، پرداخته.

در مقاله [۸] از یادگیری ماشینی برای ارائه روشی درطبقه بندی متون فارسی استفاده نموده و روش ارائه شده، تحت سیستم نرم افزاری طبقه‌بند متون فارسی، طراحی و پیاده‌سازی شده است. سیستم طبقه‌بند متون فارسی در فاز یادگیری،مجموعه ای از متون آموزشی را برای استخراج ویژگی های دسته ها بررسی می‌کند تا خصوصیات اصلی هر دسته را بدست آورد. به طوری‌که در فاز تست سیستم طبقه بند متون فارسی، این ویژگی‌های مختص دسته، برای طبقه بندی متون دسته بندی نشده به کار می‌روند. از ریشه یابی برای کاهش بعد بردارهای ویژگی استفاده نموده و دقت روش پیشنهادی روی مجموعه جمع آوری شده از اخبار فارسی در هفت دسته مورد آزمایش قرار گرفته است.

پژوهش [۹] در قالب مقاله‌ای است که سعی دربرطرف نمودن معایب الگوریتم KNN دارد. تمرکز اصلی در این پژوهش بر مشکل تأثیر یکسان همه ی خصیصه‌ها در محاسبه ی فاصله ی رکورد جدید با همسایه های آن رکورد می‌باشد، درصورتی‌که برخی از این خصیصه ها برای عمل دسته بندی کم اهمیت ترند.به منظور رفع این مشکل به هر یک از خصیصه‌ها یک وزن اختصاص داده و برای محاسبه ی فاصله ی بین دو نقطه به جای استفاده از فاصله ی اقلیدسی ازفاصله ی مانهتن استفاده می‌کنند.

در پژوهشی [۱۰] یکی دیگر از معایب الگوریتم KNN مورد بررسی قرار گرفته‌که همان پیچیدگی محاسباتی الگورتیم است؛ زیرا برای طبقه بندی کردن هر الگوی تست باید فاصله آن را تا تمام داده های آموزشی پیدا نموده مادامی که برای نگهداری تمام داده‌های آموزشی نیاز به فضای ذخیره سازی زیادی دارد. این محدودیت ها باعث شده که با استفاده از روش های تولید الگوهای نماینده، سعی در کاهش حجم داده ها و افزایش سرعت الگوریتم گردد. در این مقاله با ترکیب یکی از روش‌های Instance Filtering با الگوریتم AIRS الگوهای نماینده از روی داده های اصلی تولید کرده است که در واقع الگوریتم AIRS از سیستم ایمنی بدن الهام گرفته و با استفاده از مکانیزم این سیستم الگوهای نماینده را تولید می‌کند.

در مقاله بصیری استفاده از الگوریتم یادگیری ماشین K نزدیکترین همسایه برای متن فارسی، ارزیابی شده و نیز از اسناد تهیه شده توسط افراد استفاده شده و اسناد شامل ۶۰۰ سند است، که متعلق به ۶ دسته می باشد. همچنین از مجموعه ویژگی با ۶۱۵ کلمه کلیدی استفاده شده و سندها در مجموعه داده بوسیله حذف کلمات انتهایی، پیش پردازش می‌شوند و سپس با استفاده از مدل فضای برداری بر اساس مجموعه کلمات کلیدی استخراجی از هر سند نیز نرمالسازی می شوند. برای اندازه گیری شباهت بین دو سند، معیار اندازه گیری شباهت کسینوسی استفاده شده است[۱۱].

در مقاله عارفیان اکثر کارهای انجام گرفته با رویکرد فازی- ناهموار ، در زمینه یادگیری بانظارت بوده مادامی که تعداد کمی از آن‌ها برای یادگیری نیمه نظارتی و یادگیری بدون نظارت انجام گرفته اند. در اغلب پایگاه داده هایی که وجود دارد، برچسب کالس برای تعداد اندکی از داده ها مشخص است و این هم قابل ذکر است که برچسب زنی تمام داده‌ها، کاری دشوار، پرهزینه و زمان بر است [۱۲]. در مقاله [۱۳] روش کاهش ابعاد پیشنهادی برای مسائل کلاس‌بندی شده مناسب است. این روش براساس کمینه کردن احتمال خطای دسته‌بندی نزدیکترین همسایه بنا شده و یک تصویر خطی و مجموعه کوچکی از نمونه‌های اصلی مورد نظر را می‌آزماید. علت استفاده از روش دسته‌بندی نزدیکترین همسایه در مقاله خاصیت بازدهی بسیار بالای این روش

دسته‌بندی از لحاظ محاسباتی می‌باشد. به طوریکه عمل طبقه‌بندی را خیلی سریعتر از روش‌های دسته‌بندی رایج دیگر از قبیل SVM انجام می‌دهد و در عین حال از قدرت تشخیص خوبی نیز برخوردار است.

در مقاله [۱۴] الگوریتم جدیدی ارائه شده که هدف آن کاهش ابعاد است و در دسته‌بندی صفحات وب کاربرد دارد. از اینرو از یک مجموعه ی ناهمگن از صفحات وب به عنوان مجموعه داده ورودی استفاده می‌شود. صفات انتخابی برای دسته‌بندی، محتوای متنی صفحات است. با استفاده از الگوریتم پیشنهادی می توان ابعاد بالای صفحات و کلمات استخراجی از صفحات را به یک ابرصفحه‌ی جدید تصویر نمود که ابعاد آن برابر با تعداد کلاس‌هاست و نتایج نشان می‌دهد که زمان پردازش الگوریتم دسته‌بندی با استفاده از الگوریتم کاهش بعد پیشنهادی به طرز قابل توجهی، کاهش می‌یابد. میزان کاهش آن وابسته به تعداد صفحات ارائه شده به دسته بندهاست. همچنین در قیاس با تست‌های اجرا شده بدون استفاده از الگوریتم کاهش ابعاد پیشنهادی، دقت الگوریتم‌های دسته‌بندی پیشنهادی افزایش یافته است.

در حین انجام این مقاله، منابع زیادی مورد مطالعه قرار گرفتند و مهمترین آن‌ها در این بخش ذکر گردید. از مطالعه این منابع می‌توان به این نتیجه دست یافته که: الگوریتم‌ها و تکنیک‌های متفاوتی برای برای دسته بندی بکار می‌روند و از آنجایی که هدف این مقاله هم ارائه الگوریتم KNN بهبود یافته برای د سته بندی ا ست که از قیاس و ا ستناج با تکنیک‌های پیشین حا صل شده. به بیانی دیگر یعنی ارائه تکنیکی نوین برای الگوریتم KNN با دقت و سرعت بالاتری و با فضای ویژگی بزرگتری نیز می‌باشد. الگورتیم پیشنهادی که یکی از بهترین و پرکاربردترین الگوریتم‌های دسته بندی است که ازجمله کاربردهای مختلف آن: پزشکی تجارت، بیوانفورماتیک، بازیابی اطلاعات، فیلتر کردن اطلاعات، تفکیک مهمترین اخبار و امور سیاسی می‌گردد. و هدف نویسندگان، از استناج و آوردن تکنیک‌های پیشین هم ارائه اهدافی می‌باشد که عبارتند از:

۱) بدست آوردن روشی برای محاسبه فاصله داده در فضای جستجو با هدف بازگردانی نزدیک ترین کلاس آموزشی
۲) قابلیت کاراتر در مدیریت و اعمال نفوذ برحجم زیاد اطلاعات
۳) مقایسه داده‌های مختلف و رتبه بندی مهم‌ترین آن‌ها

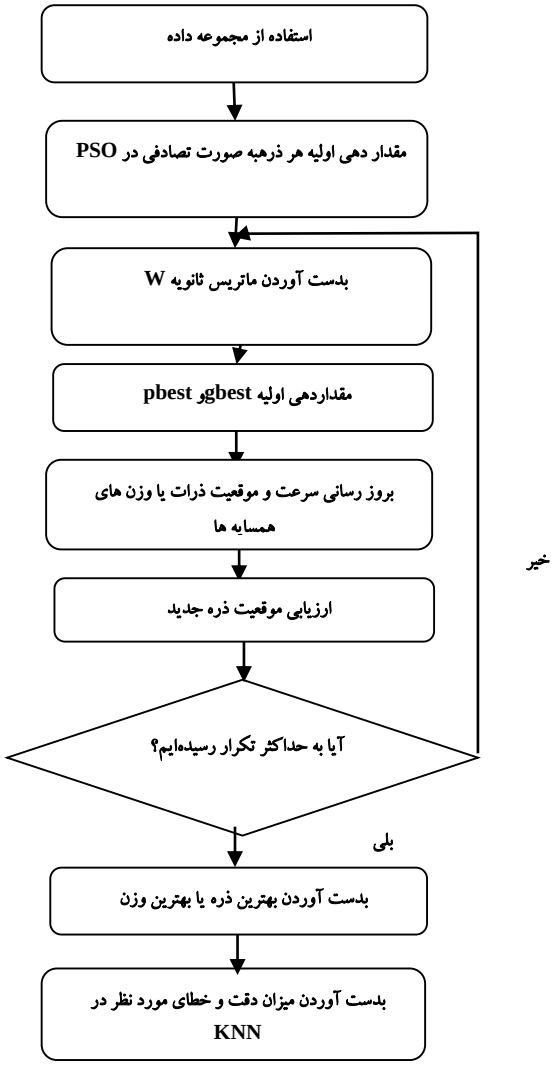
۳- ارائه تکنیک پیشنهادی

الگوریتم بهینه سازی ازدحام ذرات، طبیعتاً یک روش سریع و جدید مطلوب است که شامل تبدیل ماتریس اولیه به ماتریس با ابعاد ثانویه می‌باشد که هدف اصلی آن، همان کاهش بعد ماتریس اولیه است. از اینرو، هدف نهایی در این مقاله دستیابی به پاسخ این پرسش است که: آیا می توان با استفاده از الگوریتم بهینه سازی ازدحام ذرات با وزن دهی همسایه های الگوریتم نزدیک ترین همسایه، خطای دسته‌بندی داده‌های اولیه به روش دسته‌بندی نزدیکترین همسایه (KNN) را کمینه کرده؟

الگوریتم بهینه سازی ازدحام ذرات، از چندین نقطه جستجو استفاده می‌کند که Pbest و Gbest این نقاط نزدیک به نقطه‌ی بهینه است و از آن‌جایی که الگوریتم بهینه سازی ازدحام ذرات ، می‌تواند برای مسائل پیوسته و گسسته بکار رود و توانایی کارایی در جستجوی سراسری در فضای مسئله دارد. اما قدرت آن در جستجوی محلی ضعیف است و ممکن است در یک بهینه‌ی محلی گیر کند. در روش پیشنهادی، با ترکیب و ادغام دو الگوریتم PSO و KNN سعی شده تا الگوریتمی جدید ارائه شود به گونه‌ای که این الگورتیم، با دستیابی به حداکثر تعداد تکرار به اتمام می‌رسد و راه‌حل اکتسابی، بوسیله‌ی آن، به جواب بهینه نزدیک است. از اینرو در پیاده‌سازی روش پیشنهادی که آن را PSO-WeightedKNN نامیده و هدف: وزن دهی به فاصله داده ورودی با همسایه‌های آن است بطوری‌که خطای دسته‌بندی مجموعه داده کمینه و دقت این دسته‌بندی بیشینه باشد. پس در مرحله ی نخست اقدام به کاهش ابعاد ماتریس داده اولیه نموده و سپس اقدام به وزن دهی ماتریس ثانویه با استفاده از الگوریتم بهینه سازی ازدحام ذرات نموده و در تابع ارزیابی برای الگورتیم بهینه سازی ذرات، هم دقت الگوریتم دسته بند نزدیک ترین همسایه در نظر گرفته می شود. همانطور که در شکل (۱) آمده، تابع شایستگی برای الگوریتم بهینه سازی ازدحام ذرات با استفاده از کارایی الگوریتم KNN تعریف می شود.بدین صورت که در ابتدا وزن های تصادفی به ذره ها یا همان همسایه ها داده می

شود وسپس این وزن های تصادفی بر روی k همسایه داده تست، اعمال شده و عمل دسته بندی بر روی داده مورد نظر انجام می گیرد.میزان خطای دسته بند به عنوان تاب fitness بر روی الگوریتم بهینه سازی ازدحام ذرات، اعمال می‌گردد.بعد از آن با توجه به تابع شایستگی هر کدام از پارامتر های الگوریتم بهینه سازی ازدحام ذرات از قبیل مکان ذره و سرعت ذره و همچنین Gbest و Pbest به روز رسانی می شوند و این روند تا یافتن، بهترین وزن برای همسایه ها که منجر به کمینه شدن میزان خطا و بیشینه شدن دقت دسته بندی ادامه می یابد [۱۵]. تعداد تکرار ، تعداد ذرات و نیز تعداد همسایه ها در این الگوریتم قابل تغییر می باشد و به همین دلیل می توان بهترین مقدار مناسب برای هر پارامتر را کشف نمود، همچنین الگوریتم قابلیت تغییر سایز ماتریس ثانویه را هم داراست و بدین ترتیب می توان تصمیمی در این باره اتخاذ نموده که بهترین ماتریس ثانویه در نظر گرفته شود.

در شکل (۱) مراحل اجرای الگوریتم پیشنهادی آمده است:



شکل ۱- مراحل انجام تکنیک PSO-WeightedKNN

در الگوریتم اصلی و پایه‌ای PSO، جمعیت اولیه‌ی ذرات بصورت تصادفی تولید می‌شوندو هر ذره به صورت یک بعدی می‌باشد، که این اصل در الگوریتم پیشنهادی نویسندگان نیز رعایت شده است، ولی در انتهای هر ذره مقدار K نیز به عنوان ستون آخر قرار می‌گیرد. زیرا هر کدام از این جمعیت های تولید شده در حقیقت وزن ها مورد استفاده در الگوریتم KNN وزن دار محسوب می شوند و ماتریس ثانویه که W نامیده را نیز در همان ابتدای تعریف جمعیت اولیه ایجاد کرده و برای استفاده توسط

الگوریتم دسته بند آماده نموده. در نهایت الگوریتم دسته بند هر بار با استفاده از وزن های تولید شده عمل دسته بندی را بر روی ماتریس کاهش بعد داده شده اعمال می کند و از سویی دیگر هم، تعداد همسایه ها در این الگوریتم پیشنهادی قابل تغییر می باشد.

۴- ارزیابی نتایج تجربی

الگوریتم‌های استفاده شده در این مقاله، به منظور مقایسه و ارزیابی کارآیی الگوریتم پیشنهاد شده WeightedKNN-PSO ، عبارتند از الگوریتم‌های PSO، KNN، که به ترتیب اولین الگورتیم، جزء الگوریتم‌های بهینه‌سازی و دومین الگورتیم، جزء الگوریتم‌های طبقه‌بندی می‌باشد. از اینرو در این مقاله، مجموعه‌ای از آزمایش‌ها برای ارزیابی الگوریتم پیشنهادی بر روی سه مجموعه داده استاندارد یادگیری ماشین دانشگاه UCI و به نام های Sonar,Iris,Wine انجام شده است. کارآیی الگوریتم پیشنهادی در این مقاله، با الگوریتم‌های شناخته شده در حوزه دسته بندی و ا لگوریتم های دیگر KNN که با رویکردی خاص دسته بندی را انجام می دهد، مقایسه شده که تنظیمات پارامترهای خاص الگوریتم PSO در جدول ۱ آورده شده است.

جدول۱- تنظیمات پارامترهای الگوریتم PSO

مقدار پارامتر	توضیح پارامتر
۵۰+۱	اندازه جمعیت
۲	ضریب c1
۲	ضریب c2
دقت دسته‌بندی KNN	تابع ارزیابی

۴-۱- ارزیابی عملکرد الگوریتم PSO-WeightedKNN در مقایسه با کارهای مرتبط

به منظور سنجش روش پیشنهادی از سه تا مجموعه داده که شامل Sonarو Wineو Liver می‌باشد، استفاده نموده که جدول ۲ خلاصه‌ای از ویژگی‌های این سه مجموعه را نشان می‌دهد:

جدول ۲- خصوصیات مجموعه داده‌های کاربردی

Data set	No .of classes	No .of attribute	No .of data
Liver	۲	۳۴	۳۴۵
Wine	۳	۱۳	۱۷۸
Sonar	۲	۶۰	۲۰۸

در این مقاله، کارایی مدل پیشنهادی را با روش‌هایی همچون الگورتیم KNN و برای دسته‌بندی داده‌ها را انجام نموده که با در نظر گرفتن معیار خطای دسته‌بندی و دقت دسته‌بندی برای هر یک از الگوریتم‌ها، در جدول ۳ و بعد از آن ارزیابی و مقایسه شده.

جدول ۳- خطای دسته بندی دیتاست wine

Dataset = Wine	Model (k=invariant),itmax=200-500					
	KNN		1NN		PSO-WeightedKNN	
	Error rate (mse)	Classification rate	Error rate (mse)	Classification rate	Error rate (mse)	Classification rate
Kbest=10	0.169	0.84	0.159	0.85	0.130	0.871
Kbest=9	0.178	0.83	0.159	0.85	0.13	0.89
Kbest=20	0.159	0.85	0.159	0.85	0.074	0.929

همانطور که در جدول۳، نتایج اولین آزمایش بر روی دیتاست wine نشان داده شدهو مطرح میکند که با درنظر گرفتن K ثابت برای هر ذره بهترین مقدار برای K، ۲۰ خواهد بود که بالاترین دقت دسته بندی یعنی ۹۲٪ را به همراه دارد.

جدول ۴- خطای دسته بندی دیتاست sonar

Dataset = Sonar	Model (k=invariant),itmax=200-500 1 <= K =<30					
	KNN		1NN		PSO-WeightedKNN	
	Error rate (mse)	Classification rate	Error rate (mse)	Classification rate	Error rate (mse)	Classification rate
Kbest=25	0.159	0.85	0.159	0.85	0.087	0.915
Kbest=24	0.154	0.848	0.159	0.85	0.13	0.88
Kbest=12	0.159	0.85	0.159	0.85	0.12	0.889

جدول ۵- خطای دسته بندی دیتاست iris

Dataset = Iris	Model (k=invariant),itmax=200-500 1<= K =<30					
	KNN		1NN		PSO-WeightedKNN	
	Error rate (mse)	Classification rate	Error rate (mse)	Classification rate	Error rate (mse)	Classification rate
Kbest=2	0.05	0.97	0.05	0.97	0.05	0.97
Kbest=13	0.05	0.97	0.05	0.97	0.028	0.983
Kbest=25	0.05	0.97	0.05	0.97	0.015	0.989

همانطور که در جدول شماره ۵ مشاهده می شود دیتا ست iris دقت بهتری نشان داده و به نسبت دیتاست های رقیب، زودتر همگرا می شود. به طوری که برای برخی مقادیر k در همان تکرار های اولیه به جواب بهینه دست پیدا نموده به گونه‌ای که در این آزمایش مشاهده شد که به طور مثال برای k=13 بعد از حداقل ۴۰ تکرار به همگرایی در جواب دست یافته.

جدول ۶- خطای دسته بندی دیتاست ها با استفاده ازروش پیشنهادی و روش های قبلی

Dataset	Model						
	KNN	PSO-KNN (invariance)	PSO-KNN	KNNDW	VWAKNN	KNNBA	PSO-WeightedKNN
	Classification rate	Classification rate	Classification rate	Classification rate	Classification rate	Classification rate	Classification rate
wine	79	91.5	89.2	--	--	98.6	99.66
sonar	85	--	--	81.4	77.9	80.5	95.71
iris	97	94.6	92.94	94.34	97.27	--	99.33…100

در جدول مذکور، مقادیر به درصد بیان شده اند و همانطور که ملاحظه می شود نتایج به دست آمده در این پژوهش تفاوت قابل توجهی با سایر روش های پیشین دارد. پس با توجه به آنچه که گفته شد:

در این مقاله، یک الگوریتم وزن دهی بر پایه‌ی الگورتیم PSO برای مسئله‌ی دسته‌بندی پیشنهاد شده که الگوریتم PSO می‌تواند برای وزن دهی مناسب به همسایه های داده ی تست استفاده شود و نتیجه مطلوبی را حاصل نماید. چرا که دارای سرعت همگرایی بالا بوده و حساسیت آن نسبت به جمعیت اولیه کمتر است و هر ذره، نشان‌دهنده‌ی یک راه‌حل ممکن و شدنی است. عملکرد نویسندگان، برای ایجاد یک دسته‌بندی بهینه است، به طوریکه خطای دسته‌بندی حداقل شده ودقت آن بیشینه شود. در این مقاله همواره سعی بر این بوده که آزمایشی صورت گیرد که بر اساس آن بتوان به بهترین دقت در الگوریتم دسته بندی KNN دست پیدا نموده و با توجه به جداول و نمودار ها مذکور ه بتوان، تا حد قابل قبولی به این مهم دست یافته که با استفاده از ترکیب وزن

دهی به همسایه ها و ارزیابی توسط الگوریتم PSO و کاهش ابعاد داده های اولیه دسته بندی بهینه صورت پذیرفته. الگوریتم PSO-WeightedKNN، عملکرد بهتری نسبت به KNN اساسی دارد. می‌توان در کارهای آتی از سایر تکنیک‌های ادغام برای بدست آوردن دقتی بالاتر از آنچه در این مقاله حاصل شده نیز استفاده نمود.

۵- نتیجه گیری

بسترهای داده‌ای که دارای ابعاد زیادی بوده که علیرغم فرصت‌هایی که به وجود می‌آورند، چالشهای محاسباتی زیادی را ایجاد می‌کنند و از آنجایی که یکی از محدودیت های، داده‌ها با ابعاد بالا این است که در بیشتر مواقع تمام ویژگی های داده‌ها برای یافتن دانشی که در داده‌ها نهفته نیز مهم و حیاتی نیستند روش‌های آماری سنتی به دو دلیل امروزه کارائی خود را از دست داده‌اند که عبارتند از:

افزایش تعداد مشاهدات[1] و افزایش تعداد متغیرهای مربوط به یک مشاهده. پس در بسیاری از زمینه‌ها کاهش ابعاد داده یکی از مباحث قابل توجه باقی مانده و در تشریح الگوریتم پیشنهادی هم باید ذکر کرد که کار الگوریتم به این صورت انجام می گیرد که در ابتدا با مقدار k دلخواه تعداد ذره ها تعیین و سپس کاهش بعد صورت می گیرد و بعد ازآن تعدا ذره‌های تعیین شده مقادیر تصادفی وزن ها به همسایه ها اختصاص می‌یابد و برنامه بعد از مقدار دهی اولیه در الگوریتم بهینه سازی وارد برنامه دسته بندی شده و وزن ها را اعمال می‌کند و بعد از انجام دسته بندی میزان خطا محاسبه گردیده و دوباره این مقدار تحت عنوان مقدار تابع ارزیاب به الگوریتم اولیه اعمال می شود.بعد از آخرین تکرار نتیجه حاصل شده و با الگوریتم خام دسته بندی مقایسه می گردد. از آنجا که سرعت پردازش اطلاعاتی چون تصویر با حجم بسیار بالا در بعضی موارد بحرانی است، استخراج و انتخاب ویژگی‌های مهم بسیاری از مشکلات را حل نموده است. از این رو در این مقاله سعی بر این شد تا قدمی برای حل این مشکل برداشته شود و با استفاده از ترکیب الگوریتم های داده کاوی با با الگوریتم هایی که در پی یافتن جواب بهینه بوده تا علاوه بر فهم آسان و کم کردن خطا بر دقت و کارایی دسته بندی بیافزاید.

۶- مراجع

1. Yufei Tao, Jun Zhang, Dimitris Papadias, and Nikos Mamoulis," An Efficient Cost Model for Optimization of Nearest Neighbor Search in Low and Medium Dimensional Spaces ", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 16, NO. 10, OCTOBER, 2004.

2. Songbo Tan," Neighbor-weighted K-nearest neighbor for unbalanced text corpus ", Expert Systems with Applications 28, 2005.

3. Yingquan Wu, Krassimir Ianakiev, Venu Govindaraju," Improved k-nearest neighbor classification ", Pattern Recognition 35, 2002.

4. Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer,"KNN Model-Based Approach in Classification ", Lecture Notes in Computer Science,Volume 2888, 2003.

5. N. Suguna, and Dr. K. Thanushkodi," An Improved k-Nearest Neighbor Classification Using Genetic Algorithm ", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 4, No 2, July 2010.

6. Bin Zhang, Sargur N. Srihari," Fast k-Nearest Neighbor Classification Using Cluster-Based Trees",IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 26, NO. 4, APRIL, 2004.

7. Pascal Soucy,Guy W. Mineau," A Simple KNN Algorithm for Text Categorization",ICDM '01 Proceedings of the 2001 IEEE International Conference on Data Mining, IEEE Computer Society, USA ,2001.

۸. محمدحسین سرابی ، آذر شاهقلیان ،کاوش متون فارسی بر مبنای روش طبقه بندی؛ مجلد ۸، شماره۱و۳ الف؛ نشریه علمی پژوهشی انجمن کامپیوتر ایران، ۱۳۸۹.

[1] observations

۹. مهدی مرادیان، محمدکاظم سپهری فر، بهبود دقت الگوریتم KNN در داده کاوی با استفاده از قوانین وابستگی، انجمن کامپیوتر ایران، پانزدهمین کنفرانس بین المللی سالانه انجمن کامپیوتر ایران،۱۳۸۸.

۱۰. امین زارع، منصور ذوالقدر جهرمی، رضا بوستانی، تولید الگوهای نماینده به کمک الگوریتم AIRS، نخستین کنگره سیستم های فازی، دانشگاه فردوس مشهد، ایران، ۱۳۸۶.

۱۱. محمد احسان بصیری؛ شهلا نعمتی؛ ناصر قاسم آقایی؛ مقایسه دسته بندی متون فارسی با استفاده از الگوریتمهایKNN و fkNN و انتخاب ویژگیها بر اساس بهره اطلاعات و فرکانس سند؛ سیزدهمین کنفرانس ملی انجمن کامپیوتر ایران، ۱۳۸۶.

۱۲. فاطمه عارفیان، مهدی افتخاری، روش جدید K نزدیکترین همسایه فازی و ناهموار برای طبقه بندی نیمه نظارتی، همایش ملی مهندسی کامپیوتر و توسعه پایدار با محوریت شبکه های کامپیوتری، مدل سازی و امنیت سیستمها،۱۳۹۲.

13. S. Hatipoglu,S.K Mitra, N.kigdbury, "Texture classification using Dual-Tree complex wavelet Transform", Image processing and Application, confrerence publication No. 456, IEEE 1999.

14. Göksel Biricik, Banu Diri, Yildiz Technical University, Computer Engineering Department, "Impact Of New Attribute Extarction Algorithm on Web Page Classification",2012.

15. Meizhen CHEN, Jin GOU, Cheng WANG, Fenlin WU ," PSO-based Adaptive Normalized Weighted KNN Classifier ", Journal of Computational Information Systems 11: 4 ,2015.