Content list available at tspixplore.ir

پژوهشکده پردازشگر علم و فن

Journal homepage: http://aeis.ir

ارائه روشی جدید در دسته‌بندی خودکار متون فارسی مبتنی بر ترکیب ژنتیک و بهینه‌سازی جمعی ذرات

هادی رمضانی*[۱]، احمد فراهی[۲]

[۱]دانشجوی کارشناسی ارشد، گروه مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه پیام‌نور، واحد تهران شمال، تهران، ایران
h.ramazani47@gmail.com
*مسئول مکاتبات: هادی رمضانی
[۲] استادیار، گروه مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه پیام‌نور، ایران
afarahi@pnu.ac.ir

چکیده

با وجود حجم بالای اطلاعات متنی، ملزم به داشتن روش کارا جهت دسته‌بندی خودکار داریم. بنابراین باید دسته‌بندی را طوری انجام داد که ضمن افزایش دقت، سبب کاهش زمان و هزینه شود. یکی از گام‌های تعیین کننده در دسته‌بندی، استخراج ویژگی است. این موضوع در زبان فارسی به دلیل وجود ویژگی‌های زیاد و تکراری، فرایندی پیچیده است. هدف از این مقاله، بهبود عملکرد از طریق استخراج بهترین ویژگی‌ها از متون فارسی با استفاده از تجزیه و تحلیل مولفه‌های اصلی و روش ترکیبی ژنتیک و بهینه‌سازی جمعی ذرات می‌باشد، تا از مزیت جستجوی سراسری در ژنتیک و قابلیت جستجوی موضعی در بهینه‌سازی جمعی ذرات برای افزایش دقت استخراج ویژگی استفاده کنیم. نتیجه آزمایش بهبود عملکرد با درصد ۹۱,۲۴۲۱ و کاهش مدت زمان به مقدار ۰,۷۰۴۷۵ تست را برای متون فارسی نشان می‌دهد.

کلمات کلیدی:

شبکه‌های اجتماعی، نفوذ اجتماعی، بیشینه‌سازی نفوذ، کشف رهبر، بازاریابی ویروسی، کشف انجمن.

Proposing New Method For Automatic Classification of Persian Contexts Based on Combination of Genetic and Particle Swarm Optimization Algorithms

Hadi Ramazani*, Master of Science student [1], Ahmad Farahi, Assistant Professor [2]

1- Department of Computer and Information Technology, Payam Noor University, Tehran North Branch, Tehran, Iran

2- Department of Computer and Information Technology, Pajou Noor University, Tehran, Iran

1- h.ramazani47@gmail.com

2- afarahi@pnu.ac.ir

Abstract: Due to the high volume of text information, using an effective automatic classification method is necessary. This classification should improve accuracy and decrease time and cost. An important step in classification is feature extraction. In Persian language, this is complicated due to high dimensionality and redundancy of some features. Aim of this paper is to improve functionality with extraction of important features from Persian contexts with analyzing major components and combination of genetic and particle swarm optimization algorithms. Result shows improvements in accuracy with 91.2421% and test time with 0.70475.

تاریخ ارسال مقاله: ۹۲/۰۵/۲۹
تاریخ پذیرش مقاله: ۹۵/۱۱/۰۵
نام نویسنده مسئول: هادی رمضانی
نشانی نویسنده مسئول: گروه مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه پیام‌نور، واحد تهران شمال

ارائه روشی جدید در دسته بندی خودکار متون فارسی مبتنی بر ترکیب ژنتیک و بهینه‌سازی جمعی ذرات

هادی رمضانی[۱*]، احمد فراهی[۲]

[۱]دانشجوی کارشناسی ارشد، گروه مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه پیام نور واحد تهران شمال، تهران، ایران
h.ramezani۴۷@gmail.com
[۲]استادیار، گروه مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه پیام نور، ایران
afaraahi@pnu.ac.ir

چکیده

با وجود حجم بالای اطلاعات متنی، ملزم به داشتن سیستمی کارا جهت دسته بندی خودکار داریم. بنابراین، باید دسته بندی را طوری انجام داد که ضمن افزایش دقت، سبب کاهش زمان و هزینه شود. یکی از گام‌های تعیین کننده در دسته بندی، استخراج ویژگی است. این موضوع در زبان فارسی به دلیل وجود ویژگی‌های زیاد و تکراری، فرایندی پیچیده است. هدف از این مقاله، بهبود عملکرد از طریق استخراج بهترین ویژگی‌ها از متون فارسی با استفاده از تجزیه و تحلیل مؤلفه‌های اصلی و روش ترکیبی ژنتیک و بهینه‌سازی جمعی ذرات می‌باشد، تا از مزیت جستجوی سراسری در ژنتیک و قابلیت جستجوی موضعی در بهینه سازی جمعی ذرات برای افزایش دقت استخراج ویژگی استفاده کنیم. نتیجه آزمایش، بهبود عملکرد با درصد دقت ۹۱.۲۴۲۱ و کاهش مدت زمان به مقدار ۰.۷۰۴۷۵ تست را برای متون فارسی نشان می‌دهد.

کلمات کلیدی:

دسته بندی متون، ژنتیک، بهینه‌سازی جمعی ذرات، نزدیک‌ترین همسایگی.

۱) مقدمه

برای هر سازمانی با حجم بالای مستندات، دغدغه بزرگی وجود دارد که با صرف هزینه و زمان زیاد، نیروی انسانی استفاده کند که آگاه به مفهوم کل مطالب بوده و با حداکثر دقت دسته بندی‌ها را در کوتاه‌ترین زمان انجام دهد. رسیدن به این مقوله برای استفاده کنندگان منابع اطلاعاتی با توجه به سرعت بالای تولید مستندات الکترونیکی، خطای انسانی زیاد و رشد پردازشگرهای سخت افزاری سریع، ضروری است. آنچه امروزه مهم است، کمبود یا نبود اطلاعات نیست، بلکه کمبود روش‌هایی در جهت استخراج از اطلاعات، به صورت بهینه می‌باشد؛ لذا نیازمند استفاده از سیستم‌های دسته بندی خودکار خواهیم بود. دسته بندی متون، فرآیندی است که در آن متن‌ها به یک یا چند دسته از قبل تعریف شده بر اساس محتوا یا زبان نگارش متن نسبت داده می‌شود (Khreisat, ۲۰۰۴). دسته بندی ایمیل‌ها، تشخیص موضوع، فیلتر نمودن متون از جمله موارد کاربرد سیستم دسته بندی خودکار متون می‌باشند.

بیشتر الگوریتم‌هایی که در زمینه دسته بندی متون بکار گرفته شده، بر روی متون انگلیسی بوده است که از جمله پرکاربردترین آن‌ها در سال‌های اخیر، عبارتند از: نزدیک‌ترین همسایگی، درخت تصمیم‌گیری، یادگیری آماری، شبکه‌های عصبی، ماشین‌های برداری پشتیبان (Zahedi & Sarkardei, ۲۰۱۱). در این مقاله دسته بندی نزدیک‌ترین همسایه KNN به خاطر سادگی، کار آیی و همچنین کاربردش با تعداد کمی الگوهای آموزش استفاده شده (Khreisat, ۲۰۰٤) و به عنوان یکی از بهترین روش‌های دسته بندی برای متون انگلیسی و متون فارسی بیان شده است (نعمتی و بصیری، ۱۳۸۶) (Tan, ۲۰۰۶).

وزن دهی ویژگی، نقش بسیار مهمی در شاخص بندی با کیفیت متون ایفا می‌کند. از روش‌های زیادی برای وزن دهی ویژگی‌ها استفاده شده است، که از جمله می‌توان به روش‌های زیر اشاره کرد: وزن دهی مبتنی بر تعداد تکرار کلمات (TF)[۱]، روش وزن دهی دودویی، روش مبتنی بر تعداد تکرار کلمه در مستندات مختلف (IDF)، روش‌های مبتنی بر الگوریتم ژنتیک، شبکه‌های عصبی و غیره. در این مقاله از روش TFCRF برای وزن دهی به ویژگی استفاده شده است (Maleki & Abdollahzadeh, ۲۰۰۷)، که در آن برای وزن دهی ویژگی‌ها علاوه بر توجه به چگونگی توزیع آن‌ها در مستندات مختلف و

[۱] Term Frequency

مستندات کل مجموعه به چگونگی توزیع آن‌ها در طبقات مختلف نیز توجه شده است.

حال نوبت به انتخاب ویژگی‌های مناسب می‌رسد که در روش‌های انتخاب ویژگی، باید بهترین خصیصه‌ها انتخاب شده و خصیصه‌هایی که حاوی اطلاعات نیستند حذف گردند. برخی از این روش‌ها عبارتند از روش آستانه یابی تکرار سند (DF)، بهره اطلاعاتی (IG)، اطلاعات متقابل (MI) و ضریب همبستگی(CC). در این مقاله به منظور کاهش زمان محاسبات در مرحله انتخاب ویژگی، از روش کاهش ویژگی تجزیه و تحلیل مؤلفه‌های اصلی[۱] (PCA) استفاده می‌نماییم.

روش مرسوم محاسبات تکاملی که جهت بهینه سازی استفاده می‌شود سه گام دارد: الف: با جمعیتی تصادفی از جواب‌های ممکن آغاز می‌شود ب: با به روز رسانی نسل‌ها، جستجو برای جواب بهینه را انجام می‌دهد ج: ارزشیابی جمعیت بر اساس نسل‌های گذشته صورت می‌گیرد.

الگوریتم ژنتیک یکی از انواع الگوریتم‌های تصادفی است که از عملگرهای انتخاب، تقاطع و جهش استفاده می‌نماید. این الگوریتم از مشهورترین الگوریتم‌های تکاملی است و به صورت گسترده در حل مسائل بهینه سازی مورد استفاده قرار گرفته است.

الگوریتم PSO[۲] یا الگوریتم بهینه‌سازی جمعی ذرات، یک الگوریتم بهینه‌سازی مبتنی بر رفتار اجتماعی حرکت پرندگان است. یکی از مهم‌ترین مسائل در PSO تعداد پرندگان یا همان تعداد ذرات است. در حالت کلی تعداد ذرات بستگی به نوع و حل مسئله دارد. با افزایش ذرات فضای بیشتری مورد بررسی قرار می‌گیرد و زمان انجام محاسبات زیاد می‌شود ولی تعداد کم ذرات ممکن است باعث شود PSO در بهینه محلی قرار گرفته و به جواب مسئله نزدیک نشود. PSO در مقایسه با الگوریتم‌های مشابه دارای توابع کمتر، سرعت همگرایی و نتایج بهتری است (Hu & Eberhart, ۲۰۰۳).

در این مقاله روشی جدید مبتنی بر ترکیب Genetic (Yi-Shian & Hou-Chiang, ۲۰۱۲) و PSO ارائه می‌شود که توسط نرم افزار متلب، کلیه مراحل آن پیاده سازی می‌شود. کائو و زاهارا (Kao & Zahara, ۲۰۰۷) یک ترکیب Genetic و PSO برای تابع چند نمایی بکار بردند. کیواو و هان (Kuo & Han, ۲۰۱۱) ترکیب Genetic و PSO را برای حل مسائل برنامه‌ریزی خطی به کار بردند. همچنین فن و زاهارا (Fan & Zahara, ۲۰۰۷) Genetic را برای فرایند محاسبه PSO معرفی کردند و مقدار توافقی را با مقدار Genetic و PSO به صورت تکی مقایسه کردند. نتایج نشان داد که روش ترکیبی اجرای بهینه سازی بهتری دارد.

۲) فرایند دسته بندی متون

فرایند دسته بندی متون در این مقاله، بر مبنای روش یادگیری ماشین است که دو فاز آموزش و تست را شامل می‌شود. در فاز آموزش، جهت یادگیری از دسته‌های از قبل مشخص شده‌ای استفاده می‌کنیم و در فاز تست، مستندات شناخته نشده‌ای به سیستم داده می‌شود تا آن را به دسته‌ای که بیشترین شباهت را دارد، نسبت دهد. برای دسته بندی، فرایندی شامل چهار فاز را طی می‌کنیم. این چارچوب را به صورت شکل ۱ انجام خواهیم داد.

شکل ۱) مدل مفهومی استخراج شده از فرایند دسته بندی متون

در فاز پیش پردازش، با حذف ویژگی‌های بی ارزش و کم ارزش سند را برای پردازش آماده می‌کند سپس با وزن دهی، اهمیت هر یک را مشخص، در مرحله بعد به استخراج ویژگی با دقت بالا پرداخته و بعد برای بهینه سازی محاسبات، آن‌ها را کاهش می‌دهیم. در نهایت با انتخاب یک الگوریتم، دسته بندی را انجام می‌دهیم.

• فاز پیش پردازش

این فاز جهت تطابق مستندات متنی با نمایش آن‌ها در یک قالب مناسب می‌باشد تا از کلمات و علامت‌هایی که ارزش ساختاری ندارند استفاده نکنیم. این فرآیند شامل دو گام است (مقصودی و همایون پور، ۱۳۸۹). ابتدا حذف tagهای html یا xml و دوم حذف کلمات بی‌فایده (حروف ربط، اضافه، نشانه، علایم نقطه گذاری، ضمایر، افعال ربطی).

مجموعه داده‌های ورودی جهت پیاده سازی، پیکره روزنامه همشهری است که یکی از معتبرترین پیکره‌ها در زمینه زبان فارسی بوده و توسط آزمایشگاه پایگاه داده دانشگاه تهران تهیه و در اختیار پژوهشگران قرار گرفته است. متون آن ساخت یافته به صورت نیمه خودکار دسته بندی شده‌اند. این ویژگی باعث شده است که این پیکربندی علاوه بر کاربردهای متوالی مثل بازیابی اطلاعات، برای تحقیقات و کاربردهایی چون فرایند دسته بندی مناسب باشد.

در مقاله حاضر از ۸۶، Stop Word استفاده شده که علاوه بر دو مورد بالا چند کلمه که از نظر معنایی (مقصودی و همایون پور، ۱۳۸۹) در رده دسته بندی‌ها نمی‌گنجد مانند

---

[۱] Principal Component Analysis

[۲] Particle Swarm Optimization

(من، رفتن و ...) در دایره لغات کلمات بی اهمیت در دسته بندی استفاده شده است.

- **وزن دهی به ویژگی**

روش‌های وزن دهی متفاوتی برای اندیس گذاری وجود دارد، ولی همه آن‌ها در دو مورد مشترک هستند اول اینکه هرچه یک عبارت در متن بیشتر تکرار شود، بیشتر با موضوع متن مرتبط است و دوم اینکه هر چه یک عبارت بیشتر در همه متون تکرار شود، اهمیت و وزن کمتری می‌گیرد.

در وزن دهی انجام شده، ویژگی‌هایی را که در اسناد موجود نبود، وزن صفر در نظر گرفتیم. در مرحله آزمایش کلیه اسناد دارای مجموعه ویژگی‌های یکسان فرض شدند و سنجش فاصله بین ویژگی‌ها را برای ویژگی‌های موجود در مجموعه مذکور انجام دادیم. در این مقاله از روش وزن دهی TFCRF برای انتخاب ویژگی ها استفاده شده است که توسط (Maleki & Abdollahzadeh, ۲۰۰۷) ارائه شده است.

از آنجا که اکثر روش‌های وزن دهی ویژگی در ابتدا برای کاربردهای بازیابی اطلاعات مطرح شده‌اند و سپس در حوزه دسته بندی مستندات به کار گرفته شده‌اند، در این روش‌ها چگونگی توزیع ویژگی t_k در دسته C_j نادیده گرفته شده است. به عنوان مثال در روش IDF وزن دهی ویژگی t_k رابطه معکوسی با تعداد مستنداتی که دارای این ویژگی هستند، دارد. یعنی هرچه تعداد مستنداتی که دارای ویژگی t_k هستند بیشتر باشد قدرت آن ویژگی در متمایز کردن مستندات از یکدیگر پایین‌تر و در نتیجه وزن کمتری به آن ویژگی اختصاص داده می شود. اگرچه این فرض در حوزه بازیابی مستندات صحیح می‌باشد اما در حوزه دسته بندی مستندات نیازمند اعمال اصلاحاتی است تا وزن ویژگی تابعی از دسته مستنداتی که دارای آن ویژگی هستند نیز باشد.

واضح است هرچه تعداد مستنداتی که دارای ویژگی t_k هستند زیاد باشد ولی اکثر آن مستندات متعلق به دسته C_j باشند، ویژگی t_k نه تنها ویژگی نامناسبی نبوده بلکه باید به عنوان یک ویژگی مهم جهت تمایز دسته C_j از سایر طبقات در نظر گرفته شود و وزن بالایی در آن دسته به خود اختصاص دهد. از طرفی هرچه مستنداتی که ویژگی t_k در آن‌ها وجود دارد متعلق به طبقاتی غیر از دسته C_j باشند باید وزن آن ویژگی در دسته C_j پایین باشد. در معیار r_f تعریف شده در (Maleki & Abdollahzadeh, ۲۰۰۷) راه حل اولیه ای برای مسئله فوق ارائه شده است. زیرا در آن وزن ویژگی t_k در مستند d_j رابطه مستقیمی با تعداد مستنداتی دارد که از آن دسته بوده و رابطه معکوسی با تعداد مستنداتی دارد که از دسته‌ای غیر از هستند. اما فاکتور r_f در روش فوق مستقل از تعداد مستندات موجود در هر دسته محاسبه می‌شود. در صورتی که توجه به همین عامل می‌تواند کارایی دسته بندی کننده مستندات را تا حد قابل توجهی افزایش دهد.

لذا در روش استفاده شده برای وزن دهی دقیق‌تر به ویژگی‌ها به جای r_f دو فاکتور PositiveRF (فاکتور ارتباط مثبت) و negativeRF (فاکتور ارتباط منفی) تعریف می‌شود. PositiveRF نسبت تعداد مستنداتی از دسته C_j را که ویژگی t_k را دارند به کل مستندات آن دسته نشان می‌دهد و negativeRF نسبت مجموع تعداد مستنداتی از دسته غیر C_j را که ویژگی t_k را دارند به کل مجموع مستندات طبقات غیر C_j را نشان می‌دهد که به صورت رابطه (۱) و (۲) تعریف می‌شوند.

$$PositiveRF(tk,Cj) = \frac{|D(C_j,t_k)|}{|D(C_j)|} \qquad (۱)$$

$$negativeRF(t_k, C_i) = \frac{\sum_{m=1,m\neq j}^{|c|} |D(t_k,C_m)|}{\sum_{m=1,m\neq j}^{|c|} |D(C_m)|} \qquad (۲)$$

که در روابط فوق داریم $|C_j(D)|$ تعداد مستندات دسته C_j و $|D(C_j, t_k)|$ تعداد مستنداتی از مجموعه d و دسته C_j که دارای ویژگی t_k می‌باشند است. از دو رابطه بالا مقدار ارزش فاکتور ارتباط هر دسته (crfValue) به صورت رابطه (۳) تعریف می‌شود.

$$\text{crfValue}(t_k, C_i) = \frac{positiveRF(t_k,C_j)}{negativeRF(t_k,C_j)} \qquad (۳)$$

مشخص است ارزش فاکتور ارتباط هر دسته، رابطه مستقیم با فاکتور ارتباط مثبت و رابطه معکوس با فاکتور ارتباط منفی دارد. رابطه استفاده شده برای وزن دهی ویژگی t_k در مستند d_i به صورت رابطه (۴) است.

$$w_{ki} = \log\left(tf(t_k, d_i) * crfValue(t_k, c_{d_i})\right) \qquad (۴)$$

- **دسته بندی متون با الگوریتم KNN**

KNN یک روش کارا و ساده برای دسته بندی متن می‌باشد. در مقاله‌هایی که الگوریتم‌های مختلف دسته بندی متون انگلیسی را با هم مقایسه نموده‌اند، KNN نتایج بسیار خوبی داشته است. به همین دلیل در اینجا نیز از این الگوریتم برای دسته بندی متون فارسی استفاده شده است.

مبنای کار این الگوریتم، مقایسه متن تست شده با متون آموزشی داده شده و بدست آوردن میزان شباهت بین آن‌ها می‌باشد (Tan, ۲۰۰۶). ایده KNN به این است که، یک

مجموعه آموزشی برای دسته بندی وجود دارد، الگـوریتم K همسایه نزدیک در میان مجموعه آموزشی پیش دسته بندی شده، بر اساس یک معیار شباهت پیدا کرده و دسته‌های این K همسایه نزدیک برای پیش بینی دسته سند آزمایشـی بـه وسیله امتیاز دهی سـندهای هـر دسـته منتخـب، اسـتفاده می‌شود. اگر بیشتر از یک همسایه به دسته‌های مشابه تعلـق داشته باشد، مجموع امتیاز آن‌هـا بـه عنـوان وزن آن دسـته استفاده می‌شود و دسته با بـالاترین امتیـاز بـه سـند مـورد آزمایش انتساب می‌یابد، که اگر از یک مقدار آسـتانه تجـاوز کند، بیشتر از یک دسته می‌تواند به سند آزمایشـی انتسـاب یابد (نعمتی و بصیری، ۱۳۸۶). یک مشـکل در روش KNN تعیین مقدار K می باشد و برای تعیین آن باید یک سـری از آزمایشات با مقادیر مختلف K انجام شود تـا بهتـرین مقـدار برای K را تعیین کند. عیـب دیگـر KNN پیچیـدگی زمـانی محاسباتی مورد نیاز برای پیمایش همـه سـندهای آموزشـی مـی‌باشـد (He & Tan, ۲۰۰۰). افـزایش K یعنـی اینکـه سندهای بیشتری در تصمیم گیـری دسـته بنـدی دخالـت دارند و در نتیجه می‌تواند کارایی دسته بندی را کاهش دهد زیرا بیشتر سندها با امتیاز شباهت پایینی دخالت دارند کـه امتیاز پایین نشان دهنده این است که مجموعـه آزمایشـی و آموزشی احتمالاً از دسته‌های متفاوتی هستند.

- **معیارهای ارزیابی**

سه معیار دقت (Precision) و یادآوری (Recall) معمولاً در دسته بندی متون استفاده می‌شود و به صورت رابطه (۵) و (۶) نشان داده می‌شوند.

$$Precision = \frac{TP}{TP+FP} \qquad (۵)$$

$$Recall = \frac{TP}{TP+FN} \qquad (۶)$$

TP: تعداد متونی که درست به یک دسته منسوب شده‌اند.

FP: تعداد متونی که نادرست به یک دسته منسوب شده‌اند.

FN: تعداد متونی که نادرست از یک دسته رد شده‌اند.

۳) روش پیشنهادی

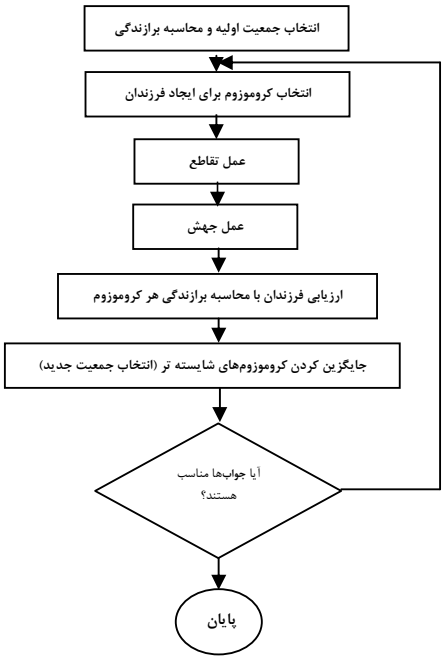
روش پیشنهادی در این مقاله برای دسته بندی متون فارسی بر اساس روش یادگیری ماشین است که دو فاز آموزش و تست دارد. برای بالا بردن معیارهای ارزیابی و کاهش زمان تست از روش تجزیه و تحلیل مؤلفه‌های اصلی همراه با انتخاب بهترین ویژگی‌ها از طریق روش پیشنهادی ترکیب ژنتیک و PSO، استفاده شده است.

- **تجزیه و تحلیل مؤلفه‌های اصلی** (PCA)

تجزیه تحلیل مؤلفه اصلی ابزاری برای آنالیز داده‌ها می‌باشد که می‌توان برای کاربردهایی مانند کاهش ابعاد، فشرده سازی یا از دست دادن اطلاعات، استخراج ویژگی و نمایش داده استفاده کرد. این روش به عنوان تصویر متعامد داده‌ها درون یک فضای خطی با ابعاد کمتر است به قسمی که واریانس بین داده‌های تصویر شده، حداکثر شود.

- **الگوریتم ژنتیک**

الگوریتم ژنتیک، یک تکنیک جستجوی سراسری تصادفی می‌باشد که مسائل را به وسیله تقلید از فرایندهای مشاهده شده در طی تکامل طبیعی حل می‌کند. ویژگی‌های خاص این الگوریتم باعث می‌شود که نتوانیم آن را یک جستجوگر تصادفی ساده قلمداد کنیم. در ابتدا، متغیرهایی که باید تعیین شوند، مشخص می‌شوند. سپس این متغیرها به نحو مناسبی کد گذاری شده و به شکل کروموزوم نمایش داده می‌شوند. بر اساس تابع هدف، یک تابع برازندگی برای کروموزوم‌ها تعریف می‌گردد و یک جمعیت اولیه دلخواه نیز تصادفی انتخاب می‌شوند. به دنبال آن، میزان تابع برازندگی برای هر کروموزوم جمعیت اولیه حساب می‌شود. سپس مراحلی که در شکل ۲ نمایش داده شده است به ترتیب انجام می‌گیرد (Parsi, Salehi & Doosmohammadi, ۲۰۱۲). خروجی این الگوریتم مجموع ژن‌ها است که در رسیدن به ویژگی‌های مناسب، اهمیت دارند.

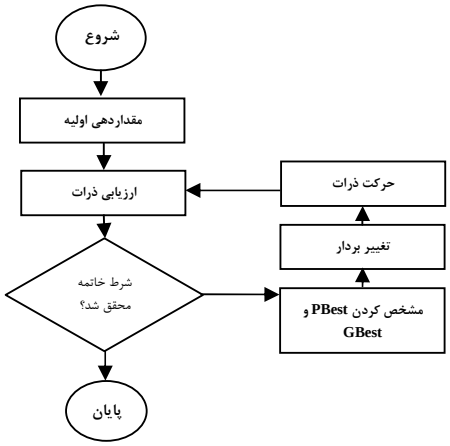


شکل ۲) فرایند الگوریتم ژنتیک (Kim, Kim, Tek & Kyungki, ***2000***)

- **الگوریتم PSO**

الگوریتم PSO یک الگوریتم بهینه‌سازی مبتنی بر رفتار اجتماعی حرکت پرندگان است. این الگوریتم اولین بار توسط Kennedy و Eberhart در سال ۱۹۹۵ ارائه شد (Kennedy & Eberhar, ۱۹۹۵).

PSO در حوزه وسیعی از حل مسائل از جمله در بهینه سازی توابع پیچیده توانایی دارد. این روش مانند اکثر روش‌های جستجوی تکاملی، با یک جمعیت جستجو به شکل موازی شروع می‌کند. سپس شایستگی هر یک از افراد جمعیت را بر اساس یک تابع هزینه تعیین کرده و با استفاده از مقادیر شایستگی، اطلاعات جمعیت را بروز می‌کند. این روند تا همگرایی الگوریتم ادامه می‌یابد. فرایند الگوریتم PSO در شکل ۳ بیان شده است. با وجود نقاط مشترکی که الگوریتم PSO با دیگر الگوریتم‌های تکاملی دارد، وجود چند خاصیت این الگوریتم را ممتاز می‌کند (Xue, ۲۰۱۲). اول اینکه پیاده‌سازی این الگوریتم بسیار آسان است و برنامه نویسی آن در زمان کوتاهی انجام می‌شود. دوم اینکه این سادگی موجب بالا رفتن سرعت محاسبات و رسیدن به جواب دلخواه با حجم کم حافظه خواهد شد. به عنوان مثال در مقایسه PSO و الگوریتم وراثتی، مکانیزم تسهیم اطلاعات در PSO بسیار آسان‌تر از الگوریتم وراثتی است. در الگوریتم وراثتی تمام کروموزوم‌ها و افراد جمعیت اطلاعات را با یکدیگر به مشارکت می‌گذارند تا نسل بعد ساخته شود، اما در PSO تنها از اطلاعات دو فرد از جمعیت استفاده می‌شود که به نوعی بهترین جواب‌ها هستند. بنابراین تمام جمعیت به سرعت به سوی جواب‌های بهینه حرکت کرده و الگوریتم سریع‌تر همگرا می‌شود. اولین کار در زمینه انتخاب ویژگی با استفاده از الگوریتم PSO باینری در مرجع (Firip & Goodman, ۲۰۰۴) صورت گرفت.



شکل ۳) مدل مفهومی استخراج شده از الگوریتم PSO

- **پیاده سازی روش پیشنهادی**

مجموعه داده‌ها آزمایشی روزنامه همشهری، شامل ۶۰۳ سند متنی به فرمت XML می‌باشد. هر سند در دسته شامل ادب و هنر، ورزشی، حوادث، اقتصاد، اجتماعی و سیاسی تقسیم شده است. حال گام‌های زیر را طی می‌کنیم:

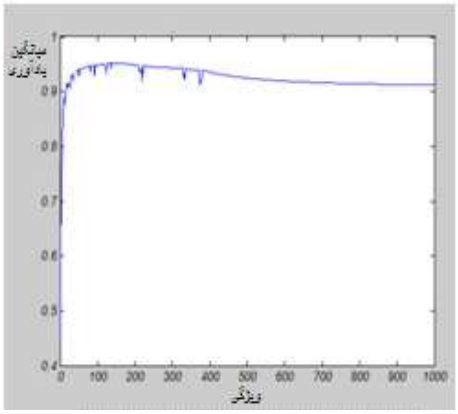
۱. بعد از حذف کلمات زائد، تعداد دفعات تکرار هر مستند را بدست می‌آوریم. تعداد ویژگی استخراج شده یا کلمات برابر با ۱۲۸۲۰ می‌باشد.

۲. یک حد آستانه تعریف شده و ویژگی‌هایی که کمتر از حد آستانه دیده شدند، جهت بالا بردن دقت و سرعت حذف شدند، و تعداد ویژگی‌ها به ۵۸۱۴ کاهش پیدا نمود که در تعداد ۵۶۴ دسته بندی قرار می‌گیرد.

۳. با الگوریتم TFCRF به هر ویژگی وزنی اختصاص داده شده، سپس مستندات به دو دسته آموزش و تست تقسیم شدند. ۲/۳ کل مستندات (۳۷۶) مستند برای آموزش و ۱/۳ کل مستندات (۱۸۸) برای تست می‌باشد.

در ابتدا دقت و زمان دسته بندی متون با استفاده از کل ویژگی‌های استخراج شده تا این مرحله با الگوریتم KNN محاسبه می‌شود که در جدول (۱) نتیجه آورده شده است. میانگین دقت نشان دهنده نتایج خوبی برای پیش بینی متون فارسی می‌باشد. ولی از آنجا که از تمام ویژگی‌ها برای تست استفاده شده، مدت زمان تست زیاد است.

۴. بر روی مجموعه داده PCA را اعمال کرده و مجموعه داده آموزش را به دو قسمت آموزش Develop و تست Develop تقسیم می‌کنیم. هدف انتخاب چند ویژگی اول است. برای استفاده از معیارهای یادآوری، ابتدا ویژگی‌ها را یکی یکی افزایش می‌دهیم و روش دسته بندی KNN استفاده می‌نماییم و مقدار یادآوری را در هر مرحله بدست می‌آوریم. نمودار میانگین یادآوری ماکرو بدست آمده از گام اول با روش دسته بندی KNN به صورت شکل (۴) است.

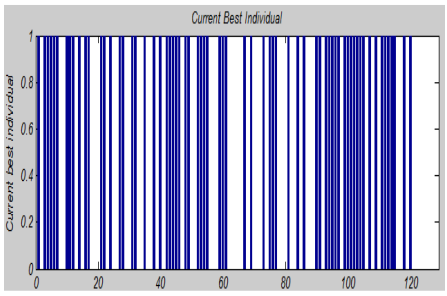


شکل ۴) ویژگی ۱۴۳ به عنوان بالاترین حد میانگین یاد آوری

همان طور که در شکل (۴) نشان داده شده است، با افزایش ویژگی از ۴۵۰ به بعد میانگین یادآوری ماکرو زیاد تغییر نمی‌کند. با توجه به نمودار، اگر ویژگی‌های ۱ تا ۱۴۳ در نظر گرفته شود، بیشترین مقدار میانگین یادآوری ماکرو برابر با ۰.۹۵۱۸۴ ارائه می‌شود. از این نمودار می‌توان نتیجه گرفت که ویژگی ۱۴۳ ویژگی مهمی بوده که سبب بالا بردن معیار ارزیابی شده است.

پس از اجرای الگوریتم PCA، ویژگی‌های بدست آمده از این مرحله به الگوریتم ژنتیک داده می‌شود. الگوریتم ژنتیک در هر دوره به تعداد تعیین شده ژن تولید می‌کند و هر ژن دو مقدار صفر و یک دارد، این عدد تصمیم الگوریتم در انتخاب شدن یا نشدن هر ویژگی را مشخص می‌کند. الگوریتم ژنتیک با توجه به نوع مسئله دارای توابع مختلفی برای انتخاب و همچنین انجام اپراتورهای ژنتیکی همانند ترکیب و جهش در کروموزوم‌ها می‌باشد. در هنگام اجرای الگوریتم ژنتیک در هر مرحله می‌توان برای همگرایی زودتر به جواب میانگین برازندگی بهترین کروموزوم و میانگین برازندگی تمام کروموزوم‌ها را در هر دوره نمایش داده و در هنگام اجرای الگوریتم ژنتیک این مقادیر را بررسی نمود. سپس با ثابت شدن نرخ تغییرات میانگین برازندگی‌های کروموزوم‌ها در هر دوره نسبت به دوره قبلی، می‌توان دریافت که مجموعه جواب‌ها در یک محدوده بسته در حال تغییر هستند و ممکن است روند جستجو در یک بهینه محلی گیر افتاده باشد، در نتیجه در این شرایط با افزایش نرخ جهش و کاهش نرخ ترکیب می‌توان با بیشتر کردن نرخ تولید تصادفی کروموزوم‌ها از این بهینه‌های محلی فرار کرد. باید توجه داشت که با نگهداری بهترین کروموزوم‌ها در هر دوره به تعداد مشخص که در این تحقیق ۱۰ است، امکان از دست دادن کروموزوم بهینه از بین می‌رود.

در مرحله دوم، ۱۴۳ ویژگی اول مرحله قبل به عنوان ورودی (ژن‌های هر کروموزوم) الگوریتم ژنتیک در نظر گرفته می‌شود که شکل (۵) ویژگی‌های انتخاب شده از روش ژنتیک را نشان می‌دهد که ۹۲ ویژگی می‌باشد. بقیه ویژگی‌ها لنتخاب نمی‌شوند.



شکل ۵) انتخاب ۹۲ ویژگی بعد از اعمال الگوریتم ژنتیک

دسته بندی KNN را بر روی ویژگی‌های استخراج شده از الگوریتم ژنتیک اعمال می‌کنیم. نتیجه در جدول (۱) ردیف دو قابل مشاهده است.

الگوریتم PSO به تعداد ۱۰۰ تکرار اجرا می‌گردد. خروجی الگوریتم، بهینه عمومی می‌باشد که دنباله‌ای دودویی از صفرها و یک‌ها تصور شده است، این دنباله، تصمیم الگوریتم در انتخاب شدن یا نشدن هر ویژگی با شماره اندیس هر ستون را مشخص می‌کند. هنگام اجرای الگوریتم PSO در هر مرحله می‌توان برای همگرایی زودتر به جواب میزان برازندگی بهترین ذره را در هر دوره نمایش داد و در هنگام اجرای الگوریتم PSO مقادیر را بررسی نمود.

حال تعداد ویژگی‌های انتخاب شده از الگوریتم ژنتیک را به الگوریتم PSO اعمال می‌کنیم. با استفاده از روش پیشنهادی ۶۱۵۶ ویژگی به ۶۳ ویژگی کاهش پیدا کرد. دسته بندی را با مجموعه داده‌های آموزش و تست با استفاده از ویژگی‌های بدست آمده از روش پیشنهادی با روش KNN انجام می‌دهیم که نتایج در جدول ۱ نشان داده شده است.

جدول۱) مقایسه روش‌های دسته بندی

ردیف	روش دسته بندی	الگوریتم کاهش ویژگی	میانگین دقت	میانگین یادآوری	زمان اجرا مرحله تست (ثانیه)
۱	KNN	بدون کاهش ویژگی	۸۸.۱۲۰۱	۸۷.۱۹۵۳	۴۱.۱۶۲۹
۲		ژنتیک	۹۱.۲۱۶۲	۸۸.۱۹۵۳	۵.۹
۳		PSO	۹۰.۱۱	۸۷.۲۴۵۶	۵.۲
۴		روش پیشنهادی (PSO&GEN)	۹۱.۲۴۲۱	۹۰.۱۴۵۲	۰.۷۰۴۷۵

همان‌طور که در نتایج دیده می‌شود، دسته بندی با استفاده از ویژگی‌های استخراج شده توسط روش پیشنهادی درصد بالایی از معیارهای ارزیابی و مدت زمان قابل توجه پایینی در مرحله آزمایش بر روی متون فارسی را نشان می‌دهد.

۴) نتیجه گیری

در این مقاله روشی جدید مبتنی بر Genetic و PSO، با استفاده از قابلیت‌های هر دو جهت کاهش فضای ویژگی‌ها و تجزیه و تحلیل مؤلفه‌های اصلی پرداخته شد. از آنجا که افزایش تعداد ویژگی، هزینه محاسباتی یک سیستم را

افزایش داده و زمان‌گیر است، انتخاب ویژگی‌های بهینه، می‌تواند تأثیر قابل توجه‌ای در نرخ بازشناسی درست الگوریتم دسته بندی داشته باشد. نتایج بدست آمده توسط روش پیشنهادی نشان می‌دهد که این روش به خوبی می‌تواند بهترین ویژگی‌ها را استخراج نموده و سبب بهبود درصد معیار میانگین دقت به اندازه ۹۱.۲۴۲۱ و همچنین کاهش زمان تست به میزان ۰.۷۰۴۷۵ ثانیه بر روی متون فارسی شود.

مراجع

مقصودی، ن. و همایون پور م. (۱۳۸۹) . **ارائه روشی جدید در طبقه بندی متون فارسی با استفاده از دانش معنایی**، دانشگاه امیر کبیر، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، آزمایشگاه پردازش هوشمند سیگنال و گفتار.

نعمتی ش. و بصیری م. (۱۳۸۶). **دسته بندی اسناد فارسی با استفاده از الگوریتم KNN**، دانشگاه صنعتی اصفهان، دانشگاه اصفهان.

Fan, S. K. S. & Zahara, E. (۲۰۰۷). A hybrid simplex search and particle swarm optimization for unconstrained optimization, *Eur. J. Oper. Res.* ۱۸۱ *(۲), pp.* ۵۲۷- ۵۴۸.

Firip, H. A. & Goodman, E. (۲۰۰٤). Swarmed feature selection, in IEEE Proceedings of the *33*rd Applied Imagery Pattern Recognition Workshop, AIPR'*04*.

He, J., Tan, A. & Tan, C. (*2000*). *Comparative Study on Chinese Text Categorization Method*, On the PRICAI *2000* Workshop on Text and Web Mining, Melbourne, pp. *25-31*.

Hu X., Eberhart R. & Shi Y. (*2003*). *Engineering optimization with particle swar*, In: IEEE swarm intelligence symposium (SIS *2003*), Indianapolis, IN, pp. *11–13.*

Kao, Y. T. & Zahara, E. (*2007). A hybrid genetic algorithm and a particle swarm optimization for multimodal function*s, Appl. Soft Compute. *8* (*2*) pp. *849-857*.

Kennedy, J. & Eberhart, R. (*1995*). *Particle Swarm Optimization*, IEEE International Conference on Neural Networks, Perth, Australia, pp. *1942-1948.*

Khreisat, L. (*2004*). *Arabic Text Classification Using N-Gram Frequency Statistics*, Tech. report Fairleigh Dickinson University, pp. ۹-۱۰.

Kim G, Kim S, Tek T, Kyungki S. (*2000*). *Feature Selection Using Genetic Algorithms for Hatd Written Character Recognition*, pp. *48-49*.

Kuo, R. J. & Han, Y. S. (*2011*). *A hybrid of genetic algorithm and a particle swarm optimization for hi-level linear programming problem- A case study on supply chain model*, Applied Mathematical Modeling. *35* pp. *3905-3917*.

Maleki, M. & Abdollahzadeh, A. (*2007*) TFCRF: *A Novel Feature Weighting Method Base on class Information in Text Categorization*, Accepted in the XLX. International conference on computer, Information and systems science and Engineering, Bangkok, Thailand.

Parsi, A., Salehi, M., & Doosmohammadi, A., (*2012*). *Swap Training: A Genetic Algorithm Based Feature Selection Method Applied on Face Recognition System*, World of Computer Science and Information Technology Journal (WCSIT), Vol. *2*, No. *4*, pp. *125-130*.

Yi-Shian .L, & Hou-Chiang T., (*2012*). *Constructing a novel Chinese readability classification model using principal component analysis and genetic programming*, IEEE International Conference on Advanced Learning Technologies *12*th.

Tan, S. (*2006*). *An effective refinement strategy for KNN text classifier*, Expert Systems with Application, vol. *30*, no. *2*, pp. *290-298*.

Xue, B. (*2012*). *Multi-Objective Particle Swarm Optimization (PSO) for Feature Selection*, GECCO'*12*, Philadelphia, Pennsylvania, USA, pp. *7-11*.

Zahedi, M. & Sarkardei, A. (*2011*). *Using MI method for feature weighting to improve Text classification Performance*. World of Computer Science and Information Technology Journal (WCSIT), Vol.*1*, No. *3*, pp. *92-95*.