شناسایی نویسندگان پیام‌های الکترونیکی از طریق واکاوی نوع و سبک نگارش آنها مبتنی بر روش‌های یادگیری ماشین (WKF based on SVM-PHGS)

سمیرا زنگویی* | کارشناسی ارشد مهندسی فناوری اطلاعات

حسنعلی نعمتی شمس‌آباد[1] | دکتری مدیریت فناوری اطلاعات
دانشکده مدیریت فناوری اطلاعات، دانشگاه تهران

دریافت: ۱۳۹۲/۰۲/۱۶ | پذیرش: ۱۳۹۲/۰۴/۱۲

فصلنامهٔ علمی پژوهشی
پژوهشگاه علوم و فناوری اطلاعات ایران
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISA، SCOPUS و ISC
http://jimp.irandoc.ac.ir
دورهٔ ۲۹ | شمارهٔ ۲ | ص‌ص ۴۵۳-۴۷۶
زمستان ۱۳۹۲

نوع مقاله: پژوهشی

چکیده: شناسایی نویسنده یکی از مسائل مهم در دسته‌بندی متن و پردازش زبان‌های طبیعی به‌شمار می‌رود. این نوشتار دستاورد پژوهشی با هدف تعیین هوشمند نوشته‌های ۵۰ نویسنده سایبری (۵۰ نفر از مشتریان بالقوه وب‌سایت آمازون با توجه به پیام‌ها و مراجعاتی که به این وب‌سایت داشته‌اند انتخاب شده‌اند)، به کمک روش‌های یادگیری ماشین است. برای سنجش کارآیی روش پیشنهادی، دقت تصمیم‌گیری آزموده و نتایج آنها با بازدهی روش‌های یادگیری ماشین مقایسه شده است. همچنین در هنگام استخراج ویژگی‌های گوناگون نوشته‌های نویسندگان برای ارزیابی توسط ماشین، کوشش شده تا حداکثر ویژگی‌های مورد نیاز برای تشخیص نویسنده شبیه‌سازی شود. بدین‌منظور، نزدیک به ۱۰۰۰۰ ویژگی گوناگون از نوشته‌های مختلف استخراج شده و در چهار دسته ویژگی‌های لغوی، ویژگی‌های نحوی، ویژگی‌های خاص و ویژگی‌های ساختاری قرار گرفته‌اند. در این پژوهش به‌طور میانگین دقت تعیین نویسنده به کمک روش پیشنهادی تا ۹۸/۷۸ درستی نیز رسیده است.

کلیدواژه‌ها: تعیین نویسنده؛ روش‌های یادگیری ماشین؛ ویژگی‌های سبک نوشتاری؛ ماشین بُردار پشتیبان

* پدیدآور رابط:
samirazangoei@yahoo.com
1. nemati@ut.ac.ir

۱. مقدمه

شناسایی نویسنده از روی نثر، سبک و شیوه نوشتاری، یا به‌عبارت دیگر ویژگی‌های نهفته در متون نوشته‌شده توسط وی، یکی از مباحث جدید در زمینه هوش مصنوعی و پردازش زبان طبیعی[1] به‌شمار می‌رود. در این مبحث کوشش می‌شود تا با استخراج ویژگی‌هایی از درون متن و پردازش و تحلیل آن به کمک انواع روش‌های هوش مصنوعی، نویسنده متن شناسایی شود، که به این سلسله امور متن‌کاوی[2] گفته می‌شود.

ایده شناسایی نویسنده از مبحث طبقه‌بندی متن[3] که خود سرفصلی از دانش فهم زبان‌های طبیعی است گرفته شده (D. D. Lewis 2004) و در آن کوشش می‌شود تا با تجزیه و تحلیل واژگان، دستور زبان و مفهوم یک جمله، و نیز با کمک‌گرفتن از دانش مربوط به واژگان، معنای آن جمله برای ماشین قابل درک شود (Allen 1987).

مهم‌ترین عواملی که موجب پیدایش و گسترش دانش طبقه‌بندی متن (که گاهی دسته‌بندی متن[4] نیز خوانده می‌شود) در چند سال گذشته شده است، تولید فزاینده دانش و اطلاعات در جهان، گسترش چشمگیر اینترنت و ارتباطات و تبادل اطلاعات، و نیز پیشرفت‌های چشمگیر در فناوری حافظه‌های جانبی بوده است. این موارد موجب‌شده که کاربر انسانی در طبقه‌بندی اطلاعات نوشتاری مانند اسناد، مقالات، نامه‌های الکترونیکی و غیره با مشکلاتی روبه‌رو شود. از این‌رو نیاز به گسترش طبقه‌بندی خودکار متن روزافزون شده است (A. N. Pavlov 2001). مدیریت پایگاه داده، پردازش زبان‌های طبیعی، یادگیری ماشین[5] و شناسایی الگو[6]، برخی از مهم‌ترین دانش‌هایی هستند که با مسائل متن‌کاوی به‌طور عام و مسئله تعیین نویسنده به‌طور خاص ارتباط دارند (de Vel 2001).

این نوشتار گزارش پژوهشی در زمینه تعیین نویسنده پیام‌های الکترونیکی به‌صورت هوشمند است که در آن نوشته‌ای به مدل طرح‌شده داده می‌شود و این مدل بعد از تجزیه و تحلیل، نویسنده آن را تعیین می‌کند. برای این کار نیاز به جمع‌آوری پیام‌ها، نویسندگان و

1. Natural Language Understanding/ Processing (NLP)
2. Text Mining
3. Text Categorization
4. Text Classification
5. Machine Learning
6. Pattern Recognition

ویژگی‌های سبکی است و بعد از طرح چارچوبی برای شناسایی، می‌توان به شناسایی نویسنده پرداخت. به‌خاطر پیچیدگی بسیار این مسئله که ناشی از گمنامی و ناشناس‌بودن فرد نگارنده است، به‌دست آوردن جواب قابل قبول در شناسایی و تشخیص هویت با کمک روش‌های یادگیری ماشین، بالابردن سرعت در تشخیص و اینکه چه ویژگی‌هایی از سبک نگارش در این شناسایی تأثیر می‌گذارند، هدف این تحقیق است.

۲. ادبیات پژوهش

با آنکه دانش طبقه‌بندی متن و شناسایی نویسنده گرایشی نوظهور از دانش دسته‌بندی و فهم زبان طبیعی است و عمری نزدیک به یک دهه دارد، اما پژوهش‌های انجام‌شده در مورد آن بسیار و خوشبختانه همراه با نتایج عملی و کاربردیِ ثمربخش بوده است.

«گوتنر» و همکارانش در سال ۱۹۹۳ به کمک شبکه عصبی مصنوعی سیستمی را برای دسته‌بندی نامه‌های الکترونیکی پدید آوردند که از یک فرهنگ واژگان بهره می‌برد و قابلیت یادگیری داشت. دسته‌بندی با این روش با دقت ۷۹/۱ درصد، درستی نزدیک به دقت انسان با ۷۹/۴ درصد درستی بوده است (Petra Geutner 1993). «ریلوف» در سال ۱۹۹۴ برای اثبات اینکه استخراج ویژگی‌ها از متن، مهم‌ترین بخش از فرایند طبقه‌بندی متن است، سه روش را به نام‌های نشان‌های ارتباطی، نشان‌های ارتباطی تکامل‌یافته[1] و الگوریتم دسته‌بندی متن موردگرا، با هم آزمود و نشان داد که روش‌های دوم و سوم دقت بسیار بالایی دارند (Ellen Riloff 1994). «جویه» در سال ۱۹۹۹ پیشنهاد کرد که یک متن به‌صورت مجموعه‌ای فازی از کلیدواژگان آگاهی‌بخش درنظر گرفته شود (Taeho C. Joe 1999). «استاتاماتوس» و همکارانش توانستند بر اساس محتوا و شیوه نگارش، شناسایی نویسنده و گروه متون به زبان یونانی جدید موجود بر روی اینترنت را به انجام رسانند (Stamatatos 2001). «پاولوف» و همکارانش ادعا کردند که می‌توانند با الگوریتم‌هایی مانند DFA و WTMM – که تعداد، افزایش و کاهش حروف را تحلیل می‌کند – رشته‌های طولانی از واژگان را از نظر کمّی مقیاس‌گذاری کنند (A. N. Pavlov 2001). «سوشی» و «مینیاو» در سال ۲۰۰۰ الگوریتم k همسایه نزدیک ساده را برای کاهش ویژگی‌های

1. Augmented Relevancy Signatures Algorithm

استخراج‌شده در طبقه‌بندی متن به‌کار گرفتند (Pascal Soucy 2000). بر اساس برخی از تعاریف از گری و همکارانش، کارهای انجام‌شده در آنالیز نویسنده را به سه فیلد اصلی طبقه‌بندی می‌کنیم:

شناسایی نوع نگارش (تألیف) نویسنده[1] که به تعیین احتمال تعلق قطعه‌ای از نوشتار به یک نویسنده خاص از طریق بررسی سایر نوشته‌ها توسط آن نویسنده می‌پردازد. همچنین در برخی از ادبیات، به‌ویژه توسط محققان زبان‌شناسی با نام «خصوصیات نویسنده»[2] هم کاربرد دارد. کامل‌ترین و قانع‌کننده‌ترین تحقیقی که در این زمینه انجام شد، توسط «موستلر» و «والیس» بود (Mosteller and Wallace 1964). در تحقیقی که توسط این محققان انجام شد، انتخاب کلمات تابع در شناسایی نوع نگارش تأثیر به‌سزایی داشت. متعاقباً نویسندگان دیگری با مشخصه ممتاز کلمات تابع و قدرت آن در شناسایی نوع نگارش نویسنده موافق بودند. نتایج آنها به‌طور کلی توسط محققان تاریخی پذیرفته شد و نقطه عطفی در این زمینه تحقیقاتی شد. از آنجایی که کاربردکلمات تابع از روی قواعد نحو در جمله تعیین می‌شود، در این تحقیق نیز کلمات تابع جزء ویژگی‌های نحوی درنظر گرفته می‌شود.

توصیف سبک نگارش[3] که به جمع‌آوری خصوصیات یک نویسنده و تولید مشخصات او بر اساس نوشته‌های او می‌پردازد. برخی از این ویژگی‌ها عبارتند از: جنس، پس‌زمینه آموزشی و فرهنگی، و آشنایی با زبان. این تحقیقاتِ نسبتاً جدید در جهتی خارج از تحقیقات شناسایی نویسنده رشد پیدا کردند. برای اولین بار «کریج» ارتباط میان هویت و خصوصیات نویسندگی با تجزیه و تحلیل نمایشنامه‌های نوشته‌شده توسط میدلتون توماس و دیگران را ارائه کرد (Craig 1999). او از کلمات رایج برجسته که به بهترین وجه می‌توانست تبعیض را نشان دهد برای توصیف عادات نوشتاری استفاده کرد. کورنی و همکارانش در سال ۲۰۰۲ به کاوش اختلاف سبک نوشتن با پس‌زمینه‌های مختلف تحصیلاتی از نویسندگان پرداختند (Corney, M 2002). «کوپل» و همکارانش شواهد قطعی ارائه دادند که نشان می‌دهد سبک نوشتن مردان در استفاده از ضمایر و انواع خاصی از اصلاحات و اسامی، با زنان متفاوت است (M. Koppel 2002).

---

1. Authorship Identification
2. authorship attribution
3. Authorship characterization

تشخیص وجوه مشترک سبک نگارش[1]، بدون شناخت نویسنده به مقایسه قطعات متعدد نوشتار و تعیین اینکه آیا آنها توسط یک نویسنده واحد تولید شده، می‌پردازد. اکثر مطالعات انجام‌شده که در این رده قرار دارند مربوط به کشف سرقت ادبی است. دزدی ادبی شامل تکرار کامل و یا بخش و قطعه‌ای از کار بدون کسب اجازه از نویسنده اصلی است. دزدی ادبی، تلاش‌های تشخیصِ فعالیتِ سرقت ادبی از طریق بررسی شباهت بین دو قطعه نوشتار است. از آنجایی که تشخیص شباهت بسیار متفاوت از شناسایی نویسنده در جنبه‌های مختلف است، این موضوع فراتر از محدوده این مقاله است.

تحلیل نویسنده در سال‌های اخیر بر روی پیام‌های الکترونیکی اعمال شده است. با بررسی تحقیقات انجام‌شده مشاهده می‌شود که بیشترین تحقیقات در فیلد شناسایی نوع نگارش نویسنده است. با توجه به مهم‌بودن این فیلد و بررسی نتایج گرفته‌شده از تحقیقات انجام‌شده در این زمینه، مشاهده می‌شود که برای دستیابی به دقت بالا در این فیلد، جای کار بسیاری دارد. این تحقیق با درنظر گرفتن پارامترهای مهم‌تر و تکمیل ویژگی‌های مؤثر، در شناسایی نویسنده و استفاده از روشی مؤثرتر، به دقت بالاتر و سرعت بالاتری در شناسایی نویسنده دست یافته است.

۳. روش پژوهش

محققان برای جلوگیری از برخی سوء استفاده‌ها در محیط‌های مجازی، به‌دنبال ایجاد ابزارهای امنیت بیشتر کاربران در جوامع مجازی هستند. ما نیز در این تحقیق به ارائه طرحی پرداختیم که می‌تواند به شناسایی افراد بر اساس سبک نگارش آنها کمک کند. شناسایی نویسنده پیام‌های الکترونیکی تاکنون به روش‌های مختلف توسط محققان مورد بررسی قرار گرفته است، ولی برای دستیابی به دقت بالاتر باید تحقیقات گسترده‌تری انجام شود تا بتوان احتمال خطا را کاهش و امنیت را افزایش داد.

با توجه به اینکه زبان انگلیسی زبانی بین‌المللی در استفاده از اینترنت است و به‌دلیل بنیادی‌بودن این تحقیق، جامعه آماری زبان انگلیسی است. مجموعه داده این تحقیق برگرفته از پیام‌های مشتریان وب‌سایت تجاری آمازون برای شناسایی نویسنده است. بیشتر تحقیقات انجام‌شده شناسایی نویسنده برای دو تا ده نویسنده صورت گرفته است، اما در

1. Similarity Detection

فضاهای مجازی پیام‌هایی که شناسایی می‌شوند اغلب متعلق به نویسندگان بالقوه بسیاری هستند و به‌طور معمول الگوریتم‌های دسته‌بندی با تعداد زیاد کلاس‌های هدف سازگار نیستند. برای آزمایش قدرت الگوریتم‌های دسته‌بندی، تعداد ۵۰ نفر از فعال‌ترین کاربران که اخیراً نیز در این گروه‌های خبری پیام گذاشته‌اند، انتخاب شده است. تعداد پیام‌هایی که برای هر نویسنده جمع شده ۳۰ تاست. به‌طورکلی طرح پیشنهادی ما به این‌صورت است که چارچوبی برای شناسایی نویسنده پیام‌های الکترونیکی ارائه می‌دهیم که شامل چهار گام است؛ با انجام به‌ترتیب این گام‌ها و استفاده از روش پیشنهادی می‌توان به شناسایی نویسنده با دقت بالایی دست یافت. در ادامه به بررسی مدل و روش موردنظر خواهیم پرداخت. در این تحقیق مهم‌ترین داده، خصوصیات نویسندگان برای سبک نوشتاری آنها بود. بنابراین در مرحله اول به جمع‌آوری پیام‌های الکترونیکی پرداخته شد. همچنین برای تجزیه و تحلیل پیام‌ها و به‌کار بردن آنها در روش پیشنهادی، نیاز به جمع‌آوری ویژگی‌های سبک نوشتاری بود. سپس بعد از جمع‌آوری ویژگی‌ها و اضافه‌کردن آنها به پایگاه داده، از این پایگاه داده برای آزمایش و به‌دست آوردن نتایج روش پیشنهادی استفاده شد. بر اساس نتایج به‌دست آمده از این تحقیق، ما معتقدیم روش‌های شناسایی نویسنده پیام‌های الکترونیکی می‌تواند به اجرای قانون در شناسایی مجرمان اینترنتی، و همچنین به تأیید هویت کاربران (در میان اعضای آنلاین) به منظور جلوگیری از فریب سایبری کمک کند.

۳-۱. چارچوب شناسایی نویسنده پیام‌های الکترونیکی

در این بخش چارچوبی را برای شناسایی نگارنده پیام‌های الکترونیکی پیشنهاد می‌کنیم (شکل ۱). همانطور که در شکل ۱ نشان داده شده است، روند شناسایی نویسنده را می‌توان به چهار مرحله تقسیم کرد:

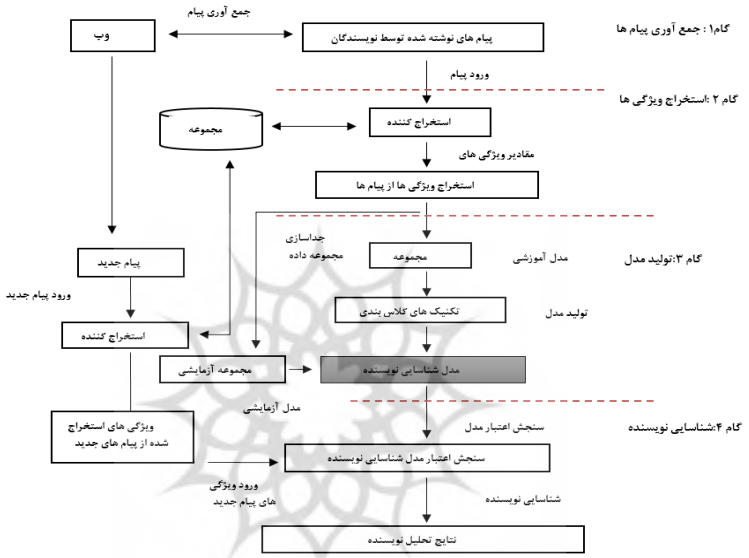
گام ۱: جمع‌آوری پیام[1]

اولین گام در شناسایی نویسنده، جمع‌آوری مجموعه‌ای از پیام‌های الکترونیکی نوشته‌شده توسط نویسندگان بالقوه برای نمایش سبک نوشتن هر نویسنده است.

1. Message Collection

گام ۲: استخراج ویژگی[1]

پیام‌های الکترونیکی در اینترنت در قالب متن بدون ساختار هستند. بر اساس ویژگی‌های سبک نوشتن از پیش تعریف‌شده، استخراج‌کننده ویژگی می‌تواند پیام‌ها را آنالیز و ویژگی‌های پیام‌های متنی الکترونیکی را استخراج کند. از استخراج ویژگی، هر متن بدون ساختار به‌عنوان یک بُردار از ویژگی‌های سبک نوشتاری نشان داده می‌شود.



شکل ۱. چارچوب شناسایی نویسنده

گام ۳: ایجاد مدل[2]

جمع‌آوری پیام الکترونیکی به دو زیرمجموعه تقسیم شده است: یکی از زیرمجموعه‌ها، مجموعه آموزشی[3] نامیده می‌شود که برای آموزش مدل طبقه‌بندی استفاده شده است. تکنیک‌های طبقه‌بندی اعمال‌شده در این فرایند ممکن است به مدل با قدرت‌های مختلف پیش‌بینی منجر شود. زیرمجموعه دیگر، مجموعه آزمایش[4] است که

1. Feature Extraction
2. Model Generation
3. training set
4. testing set

مدل تولیدشده شناساییِ نویسنده را بررسی می‌کند و قدرت پیش‌بینیِ آن را می‌سنجد. اگر مدل طبقه‌بندی توسط مجموعه تست تأیید شود، می توان آن را برای شناسایی نگارش پیام‌های الکترونیکی تازه یافت‌شده مورد استفاده قرار داد. توسعه یک مدل خوبِ پیش‌بینیِ نویسنده نیازمند دو فرایند آموزش و آزمایش است.

گام ٤: شناسایی نویسنده[1]

پس از توسعه مدل شناسایی نویسنده، می‌توان آن را برای پیش‌بینی نگارش پیام‌های ناشناخته الکترونیکی استفاده کرد. نتایج شناسایی نویسنده به محقق کمک خواهد کرد که تلاش خود را بر روی مجموعه کوچکی از پیام‌ها و نویسندگان متمرکز کند.

٤. الگوریتم ماشین بُردار پشتیبان و روش پیشنهادی (روش تجمیع کرنل‌های وزن‌دار)[2]

٤-١. الگوریتم ماشین بُردار پشتیبان

الگوریتم ماشین بُردار پشتیبان یکی از الگوریتم‌های معروف درزمینه یادگیری با ناظر است که برای دسته‌بندی و رگرسیون استفاده می‌شود و یکی از خصوصیات مهم ماشین بُردارِ پشتیبان این است که به‌طور همزمان خطای تجربی دسته‌بندی را کمینه کرده و حاشیه‌های هندسی را بیشینه می‌کند، بنابراین دسته‌بند بیشینه‌کننده حاشیه نیز نامیده می‌شود (Lin 2007).

برای یک مسئله دوبُعدی با دو دسته نتیجه، خطوط بی‌شماری ممکن است وجود داشته باشد که توسط آنها دسته‌بندی انجام شود، ولی فقط یکی از این خطوط دارای بیشینه تفکیک و جداسازی است. نقاط داده‌ای ممکن است ضرورتاً نقاط داده‌ای در فضای R^2 نباشند و به فضای چندبُعدی R^n متعلق باشند. نکته جالب توجه این است که چگونه می‌توان داده‌ها را به دامنه‌های ابرصفحه‌ای تفکیک کنیم. دسته‌بندی‌های خطیِ بسیاری ممکن است این خصوصیت را ارضاء کنند، اما ماشین بُردارِ پشتیبان به‌دنبال جداکننده‌ای است که حداکثر جداسازی را برای دسته‌ها انجام دهد. نقاط داده‌ای به‌صورت $(X_1,C_1), (X_2,C_2), \ldots, (X_n,C_n)$ نشان داده می‌شوند. C_1 می‌تواند مقدار ۱ یا ۱- را دریافت کند که توسط این ثابت‌ها دسته‌های نقاط Xi مشخص می‌شود. هر Xi یک بُردار n

1. Author Identification
2. Weighted Kernel Fusion based on SVM-Parallel Hierarichal Grid Search

بعدی به تعداد خصیصه‌هاست. هنگامی که داده‌های آموزشی‌ای را در اختیار داریم که در دسته‌های صحیح دسته‌بندی شده‌اند، ماشین بُردارِ پشتیبان توسط تقسیم‌بندی فراصفحه‌ای آنها را از هم جدا می‌کند و در دسته‌های جداگانه‌ای قرار می‌دهد به‌طوری که W.X-b=0.

افزودن پارامتر حاشیه b، حاشیه‌ها را افزایش می‌دهد. برای داشتن حاشیه، حداکثر بُردارهای پشتیبان و فراصفحه‌های موازی نزدیک‌تر به این بُردارهای پشتیبان لحاظ می‌شوند. فراصفحات موازی می‌توانند توسط معادله‌ی (۱) بیان شوند.

$$W.X - b = 1$$
$$W.X - b = -1 \qquad \text{(معادله ۱)}$$

اگر داده‌های آموزشی توانایی جداسازی به‌صورت خطی را داشته باشند، می‌توان فراصفحه‌ها را طوری انتخاب کرد که هیچ نقطه‌ای بین آنها نباشد؛ و سپس تلاش کرد تا فاصله آنها را از هم به حداکثر رساند. با استفاده از علم هندسه، می‌توان فاصله بین فراصفحه را با رابطه $\frac{2}{|W|}$ یافت، به‌طوری که |W| حداقل شود. برای تردید نداشتن در مورد داده‌ها نیاز داریم که برای هر i رابطه زیر را داشته باشیم:

$$W\ .X_i - b \geq 1 \text{ or } W\ .X_i - b \leq -1 \qquad \text{(معادله ۲)}$$

که معادله (۳) را نتیجه خواهد داد:

$$C_i.(W\ .X_i - b) \geq 1, \quad 1 \leq i \leq n \qquad \text{(معادله ۳)}$$

حال مسئله حداقل‌کردن |W| است که یک مسئله بهینه‌سازی است. به‌طور شهودی اینگونه می‌توان تصور کرد که ماشین بُردارِ پشتیبان با استفاده از رویکرد آماری، امکانی را در اختیار ما می‌گذارد که از بین دو دسته داده صفحه‌ای عبور و داده‌ها را در دو طرف این صفحه تفکیک کند. موقعیت قرارگرفتن این فراصفحه به‌گونه‌ای است که ابتدا توسط دو بُردار که از همدیگر دور می‌شوند، در بین داده‌ها به‌گونه‌ای حرکت کنند که هر یک به اولین داده نزدیک به خود برسند. سپس صفحه‌ای که از حد واسط این دو بُردار رسم می‌شود، حداکثر فاصله را از داده‌ها خواهد داشت و تقسیم‌کننده بهینه است. با افزایش ابعاد، صفحه‌های متقاطعی خواهیم داشت که داده‌ها را در داخل خود محبوس کرده‌اند.

برای اینکه بتوانیم مسئله ابعاد خیلی بالا را با استفاده از این روش‌حل کنیم، ازتوابعی به‌نام کرنل استفاده می‌کنیم. توابع مختلفی از جمله: هسته‌های نمایی، چندجمله‌ای و

سیگموید را می‌توان استفاده کرد. هدف اصلی استفاده از این ابرصفحه‌ها، ایجاد فاصله زیادتر بین ابرصفحات است. هر کدام از این توابع، خود شامل پارامترهای مختلفی هستند که باید توسط کاربر تعیین شود. از جمله این پارامترها، پارامتر γ است. دو پارامتر C و γ از جمله پارامترهایی هستند که بهینه‌کردن آنها می‌تواند در دقت روش‌های مبتنی بر بُردار پشتیبان تأثیر فراوانی داشته باشد. روش پیشنهادی ما نیز تلاش می‌کند این مقادیر را بهینه کند. برای بهینه‌کردن این پارامترها ما از روش الگوریتم ژنتیک چندهدفی و شبیه‌سازی التهابی[1] استفاده کرده‌ایم.

٤-٢. روش تجمیع کرنل‌های وزن‌دار جست‌وجوی سلسله‌مراتبی موازی در مشبکه

این روش شامل مراحل: الف) نرمال‌سازی داده‌ها[2]، ب) جست‌وجوی سلسله‌مراتبی موازی در مشبکه[3]، ج) ساختن مدل[4]، و د) انتخاب بهترین کرنل ماشین بُردارِ پشتیبان[5] است که در ادامه، هر مرحله را توضیح می‌دهیم.

٤-٢-١. نرمال‌سازی داده‌ها

برای نرمال‌سازی داده‌ها می‌توان از فرمول زیر استفاده کرد:

$$X_{Normalized}(\square_{\square}Xmin) = \frac{\square}{X\square ax(\square_{\square}Xmin)}$$

٤-٢-٢. جست‌وجوی سلسله‌مراتبی موازی در مشبکه

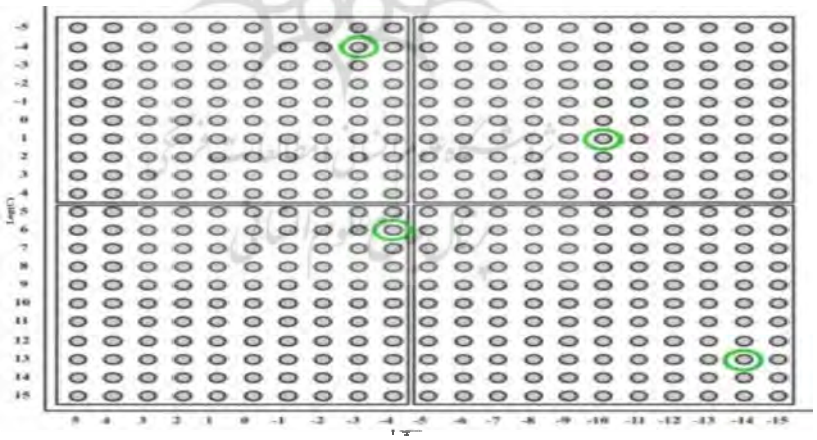
در روش‌های یادگیری ماشین بُردارِ پشتیبان، انتخاب کرنل و همچنین مقداردهی مناسب پارامترهای کرنل می‌تواند تأثیر بسیار زیادی در دقت آن داشته باشد. کرنل‌های روش ماشین بُردار پشتیبان دارای دو پارامتر C و γ هستند. به‌دست‌آوردن مقدار بهینه برای این دو پارامتر ماشین بُردار پشتیبان، هنوز یکی از مسائل باز[6] است. در این قسمت می‌خواهیم مقدار این دو پارامتر را نزدیک به بهینه کنیم.

روش پیشنهادی این است که از یک مشبکه دوبُعدی استفاده کنیم که هر بُعد آن شامل مقادیر C و γ (که توسط کاربر مشخص می‌شود) است. تجربه نشان داده است

1. Simulated Annealing
2. Data Normalization
3. Parallel Hierarchical Grid Search
4. Constructing Model
5. Optimal SVM Kernel Selection
6. Open Problem

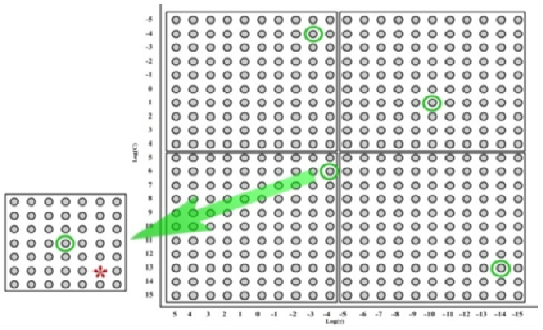
هنگامی که مقادیر این پارامترها توانی از دو (و بین 2^{15} تا 2^{-15}) باشد، معمولاً دقت روش ماشین بُردار پشتیبان بالا و قابل قبول می‌شود؛ بنابراین مقادیر این دو پارامتر، توانی از دو (و بین 2^{15} تا 2^{-15}) درنظر گرفته شده است. حال باید به ازای همه مقادیر جفت (C, γ) در بازه $[2^{15}, 2^{-15}]$ دقت به‌دست‌آمده را قرار داد (دقت به‌دست‌آمده از مجموعه آزمون). به‌عبارت دیگر به ازای هر مقدار جفت (C, γ)، با استفاده از روش یادگیری بر روی مجموعه، آزمون عمل یادگیری را انجام می‌دهیم و دقت به‌دست‌آمده را در مشبکه می‌گذاریم. برای اینکه بتوانیم دقیق‌تر و سریع‌تر مقادیر نزدیک به بهینه را برای پارامترهای ماشین بُردار پشتیبان به‌دست بیاوریم، مشبکه را به ۴ قسمت شکسته، و جست‌وجو در هر قسمت را به یک پردازه[1] واگذار کنیم تا به‌صورت موازی، هر پردازه در بازه مربوط به خود جست‌وجو کند (شکل ۲). هنگامی که همه پردازه‌ها بهترین جفت (C, γ) خود را به‌دست آورند، از بین آنها مجدداً بهترین (C, γ) را انتخاب می‌کنیم. حال یک پنجره دور بهترین (C, γ) تشکیل می‌هیم، به‌طوری که بهترین (C, γ) در مرکز آن باشد (شکل 3).

پنجره به‌دست‌آمده را به‌عنوان مشبکه جدید درنظر می‌گیریم و مجدداً عملیات گفته‌شده را بر روی آن تکرار می‌کنیم. این کار را آنقدر انجام می‌دهیم تا به‌دقت و یا تعداد تکرار موردنظر برسیم.



شکل۲. جست‌وجوی موازی در مشبکه

1. Process



شکل ۳. انتخاب بهترین (C, γ) و جست‌وجوی سلسله‌مراتبی در آن

٤-۲-۳. ساخت مدل

در مرحله قبل بهترین (C, γ) را به‌دست آوردیم. با داشتن بهترین (C, γ) می‌توانیم مدل خود را بسازیم. بدین منظور از دسته‌بندهای باینری[1] استفاده می‌کنیم. اگر مسئله ما شامل n کلاس باشد آنگاه $\binom{n}{2} = \frac{n!}{2!*(n-2)!}$ دسته‌بند باینری ایجاد می‌شود.

روش‌های مبتنی بر ماشین بُردار پشتیبان دارای چندین کرنل هستند که چهار کرنل معروف روش ماشین بُردار پشتیبان عبارتند از: کرنل‌های خطی[2]، چندجمله‌ای[3]، هلالی[4]، و RBF[5]. بیشتر روش‌های مبتنی بر ماشین بُردار پشتیبان، از کرنل RBF برای ساخت مدل خود استفاده کرده‌اند، ولی نتایج آزمایش نشان داد که هیچ کرنلی نمی‌تواند دقت قابل قبولی را به‌دست بیاورد. در ادامه به بررسی چگونگی استفاده از مهارت هر سه کرنل برای پیش‌بینی نویسنده خواهیم پرداخت.

٤-۲-٤. انتخاب بهترین کرنل

همانطور که گفته شد، انتخاب نوع کرنل و مقادیر پارامترهای کرنل در دقت روش‌های مبتنی بر روش ماشین بُردار پشتیبان مؤثر است. تا اینجا با استفاده از روش جست‌وجوی سلسله‌مراتبی موازی در مشبکه تلاش کردیم تا مقادیر پارامترهای کرنل را به بهینه نزدیک کنیم. در این بخش می‌خواهیم سه کرنل خطی، چندجمله‌ای و RBF را با

1. Binary Classifier
2. Linear
3. Polynomial
4. Sigmoid
5. Radial Basis Function

یکدیگر تجمیع کنیم. همانطور که می‌دانیم انتخاب روش مناسب برای تجمیع نظرات کرنل‌ها، می‌تواند بر روی دقت کلی تأثیر بسیار زیادی بگذارد.

روش‌های همجوشی زیادی[1] برای تجمیع خروجی روش‌های دسته‌بند پیشنهاد شده است (Brunak 2001). بر اساس ویژگی آنها، می‌توان این روش‌ها را به گروه‌های مختلفی از جمله: روش‌های همجوشی خطی، غیرخطی، مبتنی بر آمار و روش‌های محاسبات هوشمند دسته‌بندی کرد.

روش‌های همجوشی خطی ساده‌ترین روش همجوشی است که از میانگین وزنی و جمع برای همجوشی استفاده می‌کند. روش‌های غیرخطی معمولاً از رأی‌گیری اکثریت، روش‌های مبتنی بر آمار از تکنیک‌های دمپستر- شافر[2] و ترکیب بیزین[3]، و روش‌های هوشمند از یک روش یادگیری ماشین برای همجوشی استفاده می‌کنند.

همجوشی کرنل‌ها هنگامی می‌تواند مفید باشد که رفتار کرنل‌ها شبیه هم نباشد و کرنل‌ها مکمل یکدیگر باشند. اگر کرنل‌ها مکمل یکدیگر باشند هنگامی که یک کرنل در پیش‌بینی دچار خطا شود، بقیه کرنل‌ها قادر خواهند بود خطای او را جبران کنند. در روش پیشنهادی بر اساس مهارت هر کرنل در تشخیص هر کلاس شناسایی نویسنده، یک وزن به آنها داده شود و سپس تلاش می‌کنیم نظر نهایی را از تجمیع نظر سه کرنل (با توجه به وزن داده شده به هرکرنل) به‌دست آوریم.

برای وزن‌دهی به کرنل‌ها از سه تکنیک ایستا، نیمه‌پویا و پویا استفاده می‌شود. در ادامه تکنیک وزن‌دهی پویا را شرح می‌دهیم. نتایج آزمایش‌ها نشان می‌دهد که سه کرنل خطی، چندجمله‌ای و RBF دقت بالاتری نسبت به بقیه کرنل‌های پیاده‌سازی‌شده ماشین بُردار پشتیبان (Gussian، Bessel، Sigmoid و Tangent) دارند و به همین دلیل از این سه کرنل فقط برای همجوشی کرنل‌ها استفاده می‌کنیم.

(معادله ۴)

$$StaticWeight_{i,j} = \frac{Metric_{i,j}}{\sum_{j} Metric_{i,j}},$$

$$WhereMetric = \{FAR, Q_i, Q_i^{pre}, F-Measure, SOV\}, j = \{Linear, Poly, RBF\}, i = classes$$

1. Fusion
2. Dempster-Shafer
3. Bayesian Combination

٤-۲-۵. تکنیک وزن‌دهی پویا

وزن‌دهی پویا بر اساس احتمال اینکه یک کرنل نظر درستی را بدهد، تخصیص داده می‌شود. در این تکنیک وزن هر کرنل را در واقع احتمال اینکه نظر آن کرنل چقدر درست است، تعیین می‌کند، سپس در زمان اجرا یک عدد تصادفی در بازه [0...1] تولید می‌شود. این بازه به نسبت سه احتمال (نظر کرنل‌ها) به سه زیربازه شکسته می‌شود که طول هر زیربازه متناسب با مقدار سه احتمال است. عدد تصادفی در زیربازه هر کرنل که باشد، نظر همان کرنل را به‌عنوان نظر نهایی درنظر می‌گیریم. معادله (۴) نحوه به‌دست‌آوردن وزن‌دهی پویا را نشان می‌دهد.

(معادله ۵)

$$w_{i,x} = \frac{\sum_{j \in Predict^{i}} j}{\sum_{l \in Total^{i}} i}, where$$

$$Predict = \{Samples \mid Sample\ in\ area\ X\ and\ Linear, Poly\ and\ RBF\ kernels\ predict\ their\ label\ is\ equal\ to\ L, P\ and\ R\ respectively\ and\ its\ label\ is\ i\}$$

$$, Total = \{Samples \mid Sample\ in\ area\ X\ and\ Linear, Poly\ and\ RBF\ kernels\ Predict\ their\ labels\ equal\ to\ L, P\ and\ R\ respectively\}$$

$$, L, P, R\ and\ i = \{H, E, C\}, x = \begin{cases} A, if\ L = R = P \\ B, if\ L = R \neq P \\ C, if\ L = P \neq R \\ D, if\ P = R \neq L \\ E, if\ L \neq P \neq R \end{cases}$$

۵. نتایج و مشاهدات

۵-۱. جمع‌آوری داده‌ها

مجموعه ویژگی‌ها[1]: مجموعه ویژگی‌ها شامل ویژگی‌های سبک نوشتن از پیش تعریف‌شده توسط محققان است. به‌عنوان یک جزء مهم چارچوب ما، مجموعه ویژگی‌ها ممکن است تأثیر قابل توجهی در عملکرد شناسایی نویسنده داشته باشد. بر اساس بررسی و تجزیه و تحلیل مطالعات قبلی، چهار نوع از ویژگی‌ها را برای مجموعه ویژگی‌ها درنظر گرفته‌ایم که عبارتند از: لغوی[2]، نحوی[3]، ویژگی‌های خاص[4]، و ویژگی‌های ساختاری[5].

1. Feature set
2. lexical
3. syntactic
4. content- specific
5. structural features

مبتنی‌بر کلمه تقسیم شود. تحقیق ما شامل ویژگی‌های لغوی مبتنی بر کاراکتر پیشنهادشده توسط دی‌ول[1]، فورسیث و هولمز[2]، و لجر و مریام[3]، ویژگی‌های غنایی واژگان پیشنهادشده توسط توییدی و باین[4] و ویژگی‌های با طول فرکانس کلمه پیشنهادشده توسط مندنهال[5] است. در مجموع، ما ۸۷ ویژگی لغوی برای پیام‌های انگلیسی به تصویب رساندیم (جدول ۱).

جدول ۱. ویژگی‌های لغوی

ویژگی	توضیح ویژگی
ویژگی‌های مبتنی بر کاراکتر	
تعداد کل کاراکترها ©	
تعداد کل حروف/ تعداد کل کاراکترها	
تعداد کل کاراکترهای حرف بزرگ/ تعداد کل کاراکترها	
تعداد کل کاراکترهای رقمی/ تعداد کل کاراکترها	
تعداد کل کاراکترهای فضای خالی/ تعداد کل کاراکترها	
تعداد کل فضاهای Tab/ تعداد کل کاراکترها	
کل حروف	A–Z
کل کاراکترهای خاص	~ , @, #, $, %, ^, &, *, -, _, = ,+, >, <, [,], {, }, /, \, |
ویژگی‌های مبتنی بر کلمه	
تعداد کل کلمه‌ها (M)	
تعداد کل کلمات کوتاه/ تعداد کل کلمات	e.g., and, or
تعداد کل کاراکترهای موجود در کلمه/ تعداد کل کاراکترها	
میانگین طول کلمات	
متوسط طول جملات در قالب کاراکتر	
متوسط طول جملات در قالب کلمه	
کل کلمات متفاوت/ تعداد کل کلمات	

1. De Vel
2. Forsyth and Holmes
3. Ledger and Merriam
4. Tweedie and Baayen
5. Mendenhall

◇ **ویژگی‌های نحوی:** ویژگی‌های نحوی، از جمله: کلمات تابع[1]، علائم نقطه‌گذاری[2]، و تکیه‌کلام‌ها[3] می‌تواند سبک نوشتن نویسنده را در سطح جمله مشخص کند. قدرت تشخیص تبعیض ویژگی‌های نحوی از عادات مختلف مردم مشتق می‌شود. در این تحقیق، ما مجموعه بزرگی از ۱۵۰ عبارت تابع را که بر اساس پژوهش‌های قبلی انتخاب شده‌اند، بیان کرده‌ایم. ما نیز ویژگی‌های نقطه‌گذاری را بیان کردیم. در این چارچوب ما ۱۵۸ ویژگی نحوی که شامل کلمات تابع و ویژگی‌های نقطه‌گذاری است برای پیام‌های انگلیسی الکترونیکی درنظر گرفته‌ایم (جدول ۲).

جدول ۲. ویژگی‌های نحوی

ویژگی‌های نحوی	توضیح ویژگی
علایم	“ ” ’ ’ ; ; : ! ? . . ” ” , , “
کلمات تابع	

◇ **ویژگی‌های ساختاری:** به‌طور کلی، ویژگی‌های ساختاری نشان‌دهنده راهی است که یک نویسنده قطعه‌ای از نوشتارش را سازماندهی می‌کند. دی‌ول[4] در سال ۲۰۰۱ چندین ویژگی ساختاری را به‌طور خاص برای ای‌میل معرفی کرد. از آنجایی که ایمیل شامل ویژگی‌های ساختاری عمومی زیادی از پیام‌های الکترونیکی است، ما نیز این خصوصیات را برای پیام‌های الکترونیکی اتخاذ کردیم. به‌طور کلی در این تحقیق ۱۴ ویژگی ساختاری که شامل ۱۰ ویژگی پیشنهادشده از طرف دی‌ول و چهار ویژگی جدید دیگر است، مورد استفاده قرار می‌گیرد (جدول ۳).

جدول ۳. ویژگی‌های ساختاری

ویژگی‌های ساختاری	توضیح ویژگی
تعداد کل خطوط	
تعداد کل جملات	

1. function words
2. punctuation
3. part of speech
4. De Vel

ویژگی‌های ساختاری	توضیح ویژگی
تعداد کل پاراگراف‌ها	
تعداد جملات در پاراگراف	
تعداد کاراکترها در پاراگراف	
جداکننده بین پاراگراف‌ها	
متن نقل قول	
مکان متن نقل‌قول	
فرورفتگی پاراگراف	
استفاده از ای‌میل برای امضا	
استفاده از تلفن برای امضا	

◇ **ویژگی‌های خاص متن:** علاوه بر ویژگی‌های «مستقل از متن»[1]، ویژگی‌های خاص محتوا هم‌ویژگی‌های ممتاز مهمی برای پیام‌های الکترونیکی هستند. انتخاب چنین ویژگی‌هایی وابسته به حوزه‌های کاربرد خاص هستند. در این تحقیق با مشاهده و تجزیه و تحلیل یک‌سری از پیام‌ها، تعداد ۱۱ کلمه کلیدی به‌عنوان ویژگی‌های خاص منحصربه‌فرد برای پیام‌های فروش آنلاین تعریف کرده‌ایم (جدول ۴).

جدول ٤. ویژگی‌های خاص متن

ویژگی‌های خاص متن	توضیح ویژگی
کلمات کلیدی خاص متن	"deal", "obo", "sale", "wtb", "thx", "paypal", "check", "windows", "software", "offer"

استخراج‌کننده ویژگی‌ها[2]: خصوصیات سبک نوشتاری نویسنده از متن بدون ساختار برای مقاصد تحلیلی بیشتر استخراج شده است. با توجه به مجموعه‌ای از ویژگی‌های از پیش تعریف‌شده، انسان قادر به استخراج ویژگی‌های با دقت بسیار بالاست؛ با این حال، با توجه به مقدار زیادی از پیام‌های الکترونیکی و تعداد زیادی از ویژگی‌های سبک نوشتن، استخراج ویژگی‌های کاربر کار فشرده و وقت‌گیری است. در این مطالعه، ما به‌طور تصادفی ۳۰ پیام برای اعتبارسنجی دقت استخراج ویژگی‌ها انتخاب کرده‌ایم. تمام

1. content-free
2. Feature extractor

ویژگی‌های استخراج‌شده به‌صورت دستی در اصل سند مورد آزمایش قرار گرفتند. ما صحت را (به‌عنوان مثال، درصد ویژگی‌های استخراج‌شده به‌درستی از تمام ویژگی‌های استخراج‌شده)، به‌عنوان اندازه‌گیری اثربخشی برای فرایند استخراج ویژگی استفاده می‌کنیم. نتایج نشان داد که استخراج ویژگی برای زبان انگلیسی، به دقت بیش از ۹۵ درصد رسید. با این حال برخی از ویژگی‌ها، به‌خصوص ویژگی‌های ساختاری به‌سختی به استخراج دقیق این مورد دست می‌یابند. به‌عنوان مثال، اگر نویسنده از کلمه‌ای غیرمعمول برای تبریک استفاده کند، این برنامه ممکن است قادر به استخراج این ویژگی نباشد. مثلاً اگر کاراکتر «@» در چند خط آخرین پیام یافت شود، این برنامه ممکن است به اشتباه آن را آدرس پست الکترونیکی شناسایی کند. چک‌کردن دستی این ویژگی‌ها پس از استخراج ویژگی‌ها توصیه می‌شود.

۵-۲. نتایج به‌دست آمده از روش همجوشی کرنل‌ها

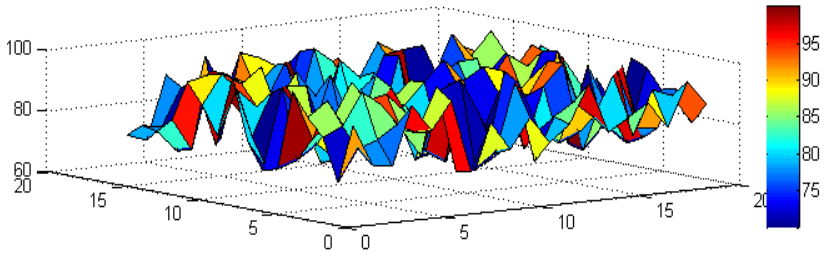
◇ روش پایه‌ای

همانطور که گفته شد می‌توان در فاز یادگیری، رفتار کرنل‌های مختلف را بررسی کرد و بهترین کرنل (که دارای بیشترین دقت است) را انتخاب کرده و در فاز آزمون از آن استفاده کنیم. هر کرنل نیز با استفاده از جست‌وجوی سلسله‌مراتبی موازی در مشبکه، پارامترهای خود را (C, γ) نزدیک به بهینه می‌کند. لگاریتم بازه پارامتر C برابر با [۱۵، ۵-] و لگاریتم بازه پارامتر γ برابر با [۵، ۱۵-] درنظر گرفته شده است. با استفاده از تکنیک اعتبارسنجی متقاطع kگانه بر روی مجموعه‌های آموزش، به‌ازای هر جفت (C, γ) امتیاز[1] (که همان دقت است) آن جفت (C, γ) را به‌دست می‌آوریم و در مشبکه قرار می‌دهیم. در نهایت مقادیر بهترین جفت (C, γ) که بیشترین امتیاز (دقت) را داشته‌اند، به‌عنوان (C, γ) نزدیک به بهینه درنظر می‌گیریم. شکل ۴ دقت به‌دست‌آمده را بر اساس مقادیر مختلف (C, γ) نشان می‌دهد. شکل ۵ نیز مقایسه دقت به‌دست‌آمده برای هر نویسنده در کرنل خطی را نشان می‌دهد. دقت کل به‌دست‌آمده برای کرنل خطی به‌صورت زیر است:

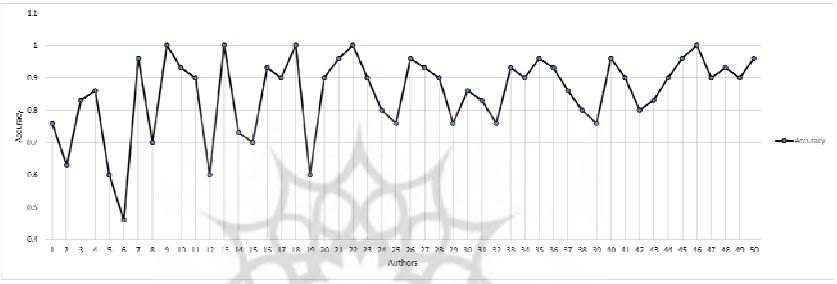
تعداد نویسندگان/ جمع دقت‌ها = دقت کل

که دقت کل به‌دست آمده ۸۵ درصد است.

1. Cross Validation Score (CVS)



شکل ٤. دقت به‌دست آمده بر اساس (C, γ)های مختلف

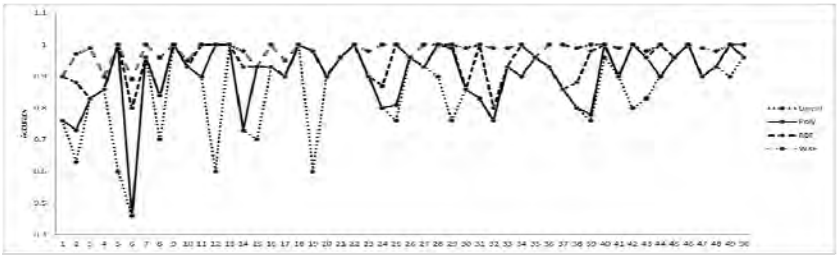


شکل ٥. مقایسه دقت به‌دست آمده برای هر نویسنده در کرنل خطی

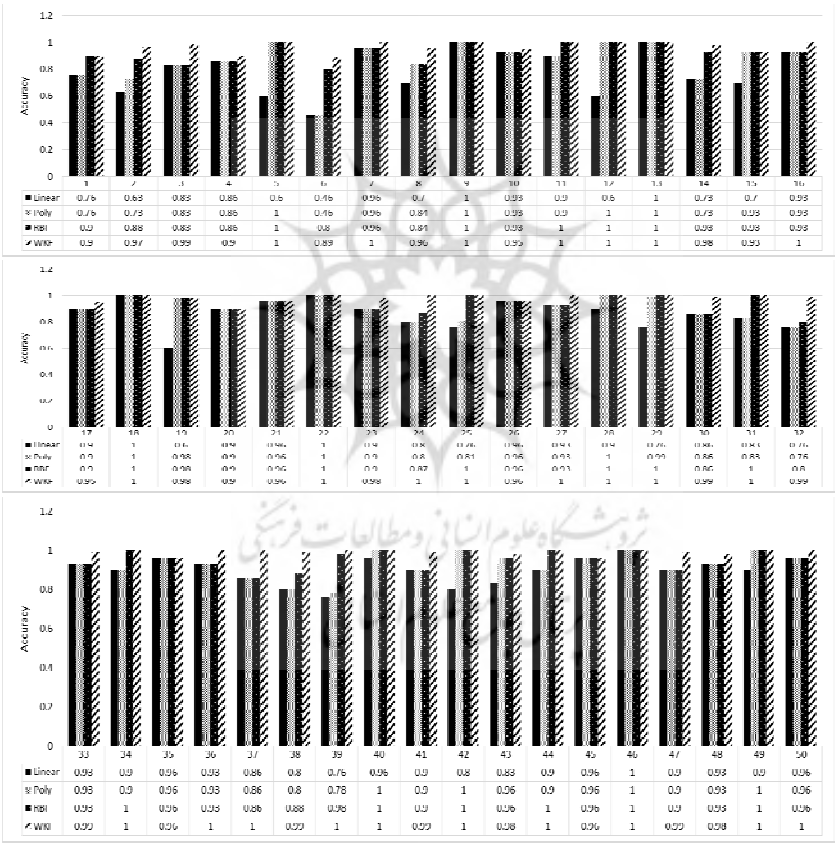
به همین ترتیب دقت‌های به‌دست آمده برای هر نویسنده را برای کرنل poly، کرنل RBF و کرنل WKF به‌دست می‌آوریم. بعد از مقایسه این دقت‌ها نتیجه می‌گیریم بالاترین دقت متعلق به کرنل WKF است که برابر ۹۸ درصد می‌شود. همانطور که در دقت‌های کل به‌دست‌آمده برای کرنل‌های مختلف مشاهده شد، کرنلی وجود ندارد که بتواند همه کلاس‌ها را با دقت قابل قبولی تشخیص دهد. به همین دلیل از مهارت سه کرنل (همجوشی کرنل‌ها) برای تشخیص نویسنده استفاده می‌کنیم.

◇ روش توسعه‌یافته

همانطور که گفته شد ابتدا با جست‌وجوی سلسله‌مراتبی موازی در مشبکه؛ پارامترهای کرنل‌ها را نزدیک به بهینه می‌کنیم تا دقت هر کرنل بالا برود. از طرف دیگر به این نتیجه رسیدیم که هیچ کرنلی وجود ندارد که بتواند همه کلاس‌ها را با دقت قابل قبولی پیش‌بینی کند؛ به همین دلیل از روش همجوشی کرنل‌ها استفاده کردیم. شکل‌های ۶ و ۷ مقایسه سه کرنل بیان‌شده با کرنل پیشنهادی را نشان می‌دهند.



شکل ٦. مقایسه دقت چهار کرنل



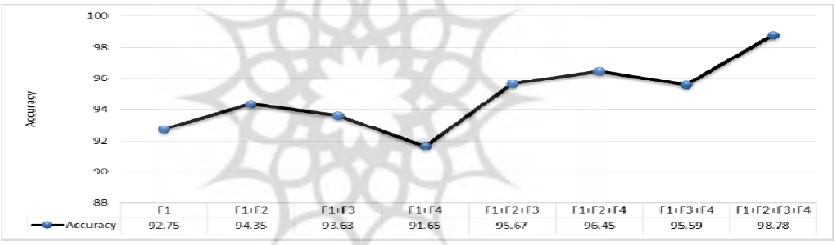
شکل ۷. مقایسه دقت چهار کرنل برای نویسندگان (تعداد نویسندگان به سه دسته تقسیم شده‌اند)

۵-۳. نتایج به‌دست آمده از فرایند تکاملی ویژگی‌ها

برای انجام مقایسه ویژگی‌ها، چهار دسته از ویژگی‌ها را درنظر گرفته‌ایم که شامل ویژگی‌های لغوی، نحوی، ویژگی‌های خاص، و ویژگی‌های ساختاری است. در آزمایشات انجام‌شده ابتدا فقط ویژگی‌های لغوی درنظر گرفته می‌شوند، سپس ویژگی‌های نحوی به مجموعه ویژگی‌های لغوی اضافه می‌شود و به همین ترتیب روند تکاملی ویژگی‌ها صورت می‌گیرد و آزمایشات بر روی آنها انجام می‌شود. جدول ۵ دقت به‌دست آمده از ترکیب ویژگی‌ها و شکل ۸ نیز نتایج این مقایسات را نشان می‌دهند.

جدول ۵. مقایسه تکامل ویژگی‌ها

ویژگی	F1	F1+F2	F1+F3	F1+F4	F1+F2 +F3	F1+F2 +F4	F1+F3 +F4	F1+F2+ F3+F4
دقت	92.75	94.35	93.63	91.65	95.67	96.45	95.59	98.78

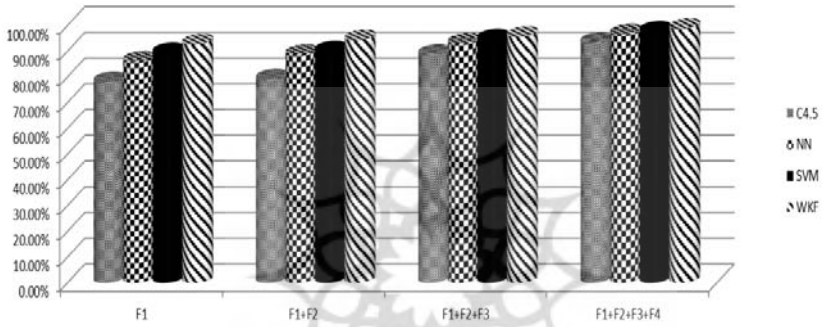


شکل ۸. مقایسه تکامل ویژگی‌ها

همانطور که در شکل ۸ مشاهده می‌شود با تکمیل ویژگی‌ها، دقت به‌دست آمده روند صعودی داشته که به دقت ۹۸/۷۸ درصد رسیده است. در مرحله بعد روش‌های C۴/۵، NN و SVM درنظر گرفته می‌شوند و با همان روند تکاملی ویژگی‌ها با روش پیشنهادی مقایسه شده و همانطور که در جدول ۶ مشاهده می‌شود در تمامی روش‌های بیان‌شده، دقت به‌دست آمده با تکامل ویژگی‌ها افزایش یافته است. همچنین دقت روش WKF در تمامی مراحل تکاملی ویژگی‌ها بالاتر از سایر روش‌هاست. شکل ۹ این مقایسات را نشان می‌دهد.

جدول ٦. مقایسه تکامل ویژگی‌ها و روش‌ها

Feature	C4.5	NN	SVM	WKF
F1	٪۷۸/۰۶	٪۸۶/۰۹	٪۸۹/۳۶	٪۹۲/۷۵
F1+F2	٪۷۹/۰۹	٪۸۸/۷۲	٪۹۰/۰۳	٪۹۴/۳۵
F1+F2+F3	٪۸۸/۷۵	٪۹۳/۰۶	٪۹۴/۶۶	٪۹۵/۶۷
F1+F2+F3+F4	٪۹۳/۳۶	٪۹۶/۶۶	٪۹۷/۶۹	٪۹۸/۷۸



شکل ۹. دقت به‌دست آمده بر اساس تکامل ویژگی‌ها

جدول ۷ نیز به مقایسه روش پیشنهادی با نتایج به‌دست آمده از تحقیقی که در سال ۲۰۱۱ بر اساس روش شبکه‌های عصبی در شناسایی نویسنده انجام شد، می‌پردازد (Sanya Liu 2011). همانطور که در جدول مشاهده می‌شود این تحقیق نسبت به تحقیق انجام‌شده از دقت بالاتری برخوردار است.

جدول ۷. مقایسه روش پیشنهادی با روش Sanya Liu, Zhi Liu, Jianwen Sun, Lin Liu

روش	Balanced attention parameter	Self-adaptive attention parameter	SVM	WKF
دقت	68.31	80.49	78.61	98.78%

٦. نتیجه‌گیری و پیشنهادات

هدف این مقاله شناسایی نویسندگان پیام‌های الکترونیکی است. تعداد زیاد خصیصه‌ها و کلاس‌های این مسئله باعث پیچیدگی حل این مسئله شده است. در گذشته روش‌های مبتنی بر ماشین بُردار پشتیبان برای حل این مسئله به‌کار گرفته شده‌اند، ولی در این روش‌ها یا تلاشی برای بهینه‌سازی پارامترها نشده و یا از روش‌های ساده‌ای برای بهینه‌کردن آنها استفاده شده است. از طرف دیگر، کرنل روش‌های پیشنهادی مبتنی بر ماشین بُردار پشتیبان نیز به‌صورت پیش‌فرض انتخاب شده است. به‌عبارت دیگر، تلاشی برای استفاده از مهارت هر سه کرنل به‌عمل نیامده است. همچنین در این مقاله روشی برای شناسایی نویسنده پیشنهاد شد که مبتنی بر ماشین بُردار پشتیبان است. در این روش با استفاده از جست‌وجوی سلسله‌مراتبی موازی، مقادیر پارامترهای آن (C, γ) را نزدیک بهینه می‌کنیم. همچنین مشاهده شد که هیچ کرنلی وجود ندارد که به‌تنهایی بتواند تمامی کلاس‌ها را با دقت قابل قبولی پیش‌بینی کند؛ بنابراین همجوشی پویای کرنل‌ها را به‌کار بردیم. این روش از بقیه روش‌ها بهتر است. تعداد کرنل استفاده‌شده برای این همجوشی سه کرنل است. با افزایش تعداد کرنل‌ها، دقت روش‌های پیشنهادی در مرحله یادگیری کاهش می‌یابد. به‌دست آوردن تکنیکی که بتواند از تعداد کرنل بیشتری استفاده کند و یا به‌عبارت دیگر از کرنل‌هایی که دقت پایین‌تری نیز دارند استفاده کند، از جمله پژوهش‌هایی است که در آینده قصد بررسی آن را داریم. بدین منظور باید روش هوشمندانه‌تری برای همجوشی کرنل‌ها ارائه شود. ایرادی مانند محدودبودن فضای جست‌وجوی پارامترهای ماشین بُردار پشتیبان که به تمام روش‌های مبتنی بر ماشین بُردار پشتیبان گرفته شده، برای این تحقیق نیز محقق است. روشی که بتواند فضای بیشتری برای جست‌وجوی این پارامترها درنظر بگیرد، بدون اینکه سرعت پیش‌بینی را خیلی پایین بیاورد، می‌تواند دقت بالاتری داشته باشد و از پژوهش‌های آتی ما به‌شمار می‌رود.

٧. منابع

A. N. Pavlov, Werner Ebeling, Lutz Molgedey, Amir R. Ziganshin and Vadim S. 2001. Scaling features of texts, images and time series. *Statistical Mechanics and its Applications* , vol. 300, issue 1, pages 310-324.

Allen, J. 1987. *Natural language understanding*. Benjamin/ Cummings Publishing Company.

Brunak, P. B. 2001. *Bioinformatics: the machine learning approach*. MIT press. Cambridge, Massachusetts.

Craig, H. 1999. Authorial attribution and computational stylistics: If you can tell authors apart, have you learned anything about them?. *Literary and Linguistic* ,14 (1), 103–113.

Corney, M., de Vel, O., Anderson, A., & Mohay, G. (2002, December). Gender-preferential text mining of E-mail discourse. *Paper presented at the 18th annual Computer Security Applications Conference* (ACSAC 2002), Las Vegas, NV.

D. D. Lewis, Y. Y. 2004. Rcv1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, 5: 361-397

de Vel, O, Anderson, A., Corney, M., & Mohay. 2001. Mining e-mail content for author identification forensics. *SIGMOD Record*. 30 (4), 55–64.

Ellen Riloff and Wendy Lehnert, Information extraction as a basis for high-precision text classification, *ACM Transactions on Information Systems*, No.3, Vol. 12, pp. 296-333, 1994.

Lin, C. C. 2007. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*.

M. Koppel, S. Argamon. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*. 17(4), 401–412.

Mosteller, F. & Wallace, D.L.. 1964. *Inference and disputed authorship:The Federalist. Reading, MA: Addison-Wesley*.

Pascal Soucy, G. W. Mineau. 2000. *A simple KNN algorithm for text categorization*. Laval Univ., Que

Petra Geutner and Uli Bodenhausen and Alex Waibel, Flexibility Through Incremental Learning: Neural Networks for Text Categorization, Proceedings of WCNN-93, *World Congress on Neural Networks*, pp. 24-27, 1993.

R. S. Michalski, I. B. (1998). Machine Learning and Data Mining; Methods and Applications, pp.71-122

Ripley, B. D. (2008). *Pattern recognition and neural networks*. Cambridge university press.

S. Russell, P. N. (1995). *Artificial Intelligence, "A modern approach," Artificial Intelligence*. Prentice-Hall, Egnlewood Cliffs.

Sanya Liu, Z. L. 2011. Application of Synergetic Neural Network in Online Writeprint Identification. *International Journal of Digital Content Technology and its Applications*. Vol. 5, No. 3, pp. 126 ~ 135.

Stamatatos, E. Fakotakis, N., & Kokkinakis. 2001. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*. 35 (2), 193–214.

Taeho C. Joe, Text categorization with the concept of fuzzy set of informative keywords, *IEEE International Fuzzy Systems Conference Proceeding* 0-7803-5406-0, 1999.

Yule, G. (1938). On sentence length as a statistical characteristic of style in prose. *Biometrika*. 30, 363–390

Zheng, R. Qin, Y., Huang, Z., & Chen, H. (2003). Authorship analysis in cybercrime Investigation. *Proceedings of the 1st NSF/ NIJ Symposium, ISI2003*. (pp. 59–73). Springer-Verlag.

Identify the Authors of Electronic Messages Through the Analysis of the Type and Style Based on Machine Learning Technique

Samira Zangoei*
Masters Of Information Technology Engineering; Iran

Hassanali Nemati Shamsabad[1]
PhD of Information Technology Management
Department of Information Technology Management
Tehran University; Tehran, Iran

Iranian Journal of Information Processing & Management

Iranian Research Institute Iranian for Science and Technology
ISSN 2251-8223
eISSN 2251-8231
Indexed in LISA, SCOPUS & ISC
Vol.29 | No.2 | pp: 453-476
Winter 2014

Abstract: Identifying the author of an electroni message is one of the main problems in text classification and natural language processing. The aim of this article is to determine the authors of 50 cyber messages (by 50 potential customers, according to Amazon 's website), by a machine learning methods. To evaluate the effectiveness of the proposed method, the decision was carefully tested and the results were compared with the performance of machine learning methods. Also, when extracting various features of authors' writing style for evaluation by machine, we tried to maximize the features required to identify a writer. Therefore, nearly 10,000 different features were extracted from different entries in four categories: lexical features, syntactic features, special features and structural features. In this study, the average accuracy of the proposed method reached to 98. 78.

Keywords: Identification of Authors; Machine Learning Methods; Characteristics of Writing Styles

* Corresponding Author:
samirazangoei@yahoo.com
1. nemati@ut.ac.ir