Journal of Soft Computing and Information Technology (JSCIT)
Babol Noshirvani University of Technology, Babol, Iran
Journal Homepage: jscit.nit.ac.ir
Volume 9, Number 2, Summer 2020, pp. 61-73
Received: 10/15/2017, Revised: 04/29/2019, Accepted: 10/26/2019

Search Space Reduction for Farsi Printed Subwords Recognition by Simple Features, Feature Quantization and Fusion of Classifiers

Esmail Miri[1], Seyyed Mohmad Razavi[2*] and Nasser Mehrshad[3]
1- Faculty of Electrical and Computer Engineering, University of Birjand, Birjand, Iran.
2- Faculty of Electrical and Computer Engineering, University of Birjand, Birjand, Iran.
3- Faculty of Electrical and Computer Engineering, University of Birjand, Birjand, Iran.
[1]Miri.esmail@birjand.ac.ir, [2*]Smrazavi@birjand.ac.ir and [3]Nmehrshad@birjand.ac.ir

Corresponding author's address: Seyyed Mohmad Razavi, Faculty of Electrical and Computer Engineering, University of Birjand, Birjand, Iran.

Abstract- In this paper, a method is presented for search space reduction in Farsi Printed Subwords recognition. First 10 simple features from subword are extracted. These features are quantized according to the interval changes of each feature in training data, and are converted to integers. A score is given to every class, using each feature and its distance to corresponding feature of each training sample. By applying all features, each class has a score per feature. A final score is obtained, by fusion of these scores using algebra operations, for each class. Search space is reduced using sorting of final scores and selection of some subwords with more scores. For fusion of scores, sum, prod, max, min and weighted sum operations are used. The weighted sum method, which Optimized weights are obtained by particle swarm optimization (PSO), has given the best response.

Keywords- Farsi Subwords Recognition, Particle Swarm Optimization (PSO), Fusion of Classifiers, Search Space Reduction, Feature Quantization.

مجله علمی پژوهشی رایانش نرم و فناوری اطلاعات
دانشگاه صنعتی نوشیروانی بابل
صفحه مجله: jscit.nit.ac.ir
جلد ۹، شماره ۲، تابستان ۱۳۹۹، صفحه ۶۱-۷۳
دریافت: ۱۳۹۶/۰۷/۲۳، بازنگری: ۱۳۹۸/۰۲/۰۹، پذیرش: ۱۳۹۸/۰۸/۰۴

کاهش فضای جستجو برای بازشناسی زیرکلمات تایپی فارسی با استفاده از ویژگی‌های ساده، کوانتیزاسیون ویژگی و ترکیب طبقه‌بندها

اسماعیل میری[۱]، سید محمد رضوی*[۲]، ناصر مهرشاد [۳]
۱- دانشکده مهندسی برق و کامپیوتر، دانشگاه بیرجند، بیرجند، ایران.
*۲- دانشکده مهندسی برق و کامپیوتر، دانشگاه بیرجند، بیرجند، ایران.
۳- دانشکده مهندسی برق و کامپیوتر، دانشگاه بیرجند، بیرجند، ایران.
[1]Miri.esmail@birjand.ac.ir, [2*]Smrazavi@birjand.ac.ir and [3]Nmehrshad@birjand.ac.ir

* نشانی نویسنده مسئول: سید محمد رضوی، بیرجند، شوکت آباد، دانشگاه بیرجند، دانشکده مهندسی برق و کامپیوتر

چکیده- در این مقاله روشی برای کاهش فضای جستجو در بازشناسی زیرکلمات چاپی فارسی ارائه می‌شود. ابتدا ۱۰ ویژگی ساده از زیرکلمه استخراج می‌شود. با استفاده از مفهوم کوانتیزاسیون و با توجه به بازه تغییرات هر ویژگی روی همه داده‌های آموزشی ویژگی‌ها کوانتیزه شده و به اعداد صحیحی تبدیل می‌شوند. با استفاده از هر ویژگی و فاصله آن تا ویژگی متناظر هر کدام از نمونه‌های آموزشی، به هر کلاس امتیازی داده می‌شود. با اعمال همه ویژگی‌ها، هر کلاس به ازای هر ویژگی یک امتیاز دارد که با ترکیب این امتیازات با اعمال جبری یک امتیاز نهایی برای هر زیرکلمه بدست می‌آید که با مرتب کردن آنها و انتخاب تعدادی از آنها که امتیاز بیشتری دارند، فضای جستجو محدود می‌شود. از اعمال جبری جمع، ضرب، بیشینه، کمینه و جمع وزن‌دار برای ترکیب امتیازات استفاده شده است. روش جمع وزن دار، که وزن‌های بهینه با الگوریتم بهینه‌سازی جمعیت ذرات تعیین شده‌اند، بهترین پاسخ را داده است.

واژه‌های کلیدی: بازشناسی زیرکلمات فارسی، بهینه‌سازی جمعیت ذرات، ترکیب طبقه‌بندها، کاهش فضای جستجو، کوانتیزاسیون ویژگی

۱- مقدمه

یکی از رویکردها برای بازشناسی زیرکلمات تایپی فارسی استفاده از شکل کلی زیرکلمه است. در این رویکرد زیرکلمه به حروف یا زیرحروف شکسته نمی‌شود و بصورت کلی بازشناسی می‌شود و هر زیرکلمه یک کلاس منحصر بفرد دارد. فضای جستجوی بزرگ در این رویکرد (بیش از ۱۲۰۰۰ زیرکلمه پرتکرار به ازای هر قلم)، باعث کاهش دقت و سرعت در بازشناسی می‌شود. فضای جستجوی بزرگ و مشکلات آن مسئله مشترکی در حوزه بازشناسی کلمات پیوسته اعم از دستنویس انگلیسی و فارسی (عربی) و تایپی فارسی (عربی) است. که تحقیقات متنوعی در این زمینه انجام شده است.

برای اجتناب از این مشکل در تحقیقات مختلف سه رویکرد، روش‌های ابتکاری، یافتن ویژگی‌های دقیق، سلسله مراتبی و یا ترکیبی از این‌ها ارائه شده است. بطور نمونه تحقیقات [۱] تا [۳] بیشتر به روش‌های ابتکاری، تحقیق [۴] بر ارائه ویژگی‌های دقیق و تحقیقات [۵] تا [۱۴] بیشتر بر روش‌های سلسله مراتبی برای حل مشکل تکیه دارند.

در تحقیق [۵] از ویژگی‌های ناحیه‌ای شکل زیرکلمه برای بازشناسی کلمات دستنویس لاتین استفاده شده است که در آن از اطلاعات وجود یا عدم وجود بالارونده‌ها، پایین‌رونده‌ها، مکان تقریبی و ترتیب قرارگیری آنها برای توصیف شکل کلی کلمات دستنویس لاتین استفاده شده است. در [۵]، بالارونده‌ها و پایین‌رونده‌ها با استفاده از اطلاعات ناحیه‌ای کلمه که با دو روش

هیستوگرام افقی نقاط سیاه تصویر کلمه و بیشینه محلی تعداد نقاط سیاه تصویر کلمه استخراج می‌شوند بدست می‌آیند.

تحقیق [۶] یک استراتژی دو مرحله‌ای برای حذف کاندیدهای غیر شبیه قبل از بازشناسی کلمات دستنویس عربی برای افزایش سرعت بازشناسی ارائه داده است. اصول این تکنیک شامل استخراج نقاط و زیرکلمه‌ها از تصویر کلمه پیوسته عربی برای توصیف شکل آن است. در اولین قدم از کاهش فضای جستجو، تعداد زیرکلمات کلمه ورودی تخمین زده می‌شود. سپس در دومین مرحله از اطلاعات نقاط در کاندیدهای مرحله قبل استفاده می‌شود. نتایج روی دیتابیس IFN/ENIT که شامل ۲۶۴۵۹ تصویر زیرکلمه است کاهش ۹۲/۵ درصدی فضای جستجو با دقت ۷۴٪ را نشان می‌دهد.

تحقیق [۷] یک روش کاهش فضای جستجو برای اسناد تاریخی عربی را ارائه می‌دهد که تصویر زیرکلمه ورودی را با نمونه‌های دیکشنری مقایسه کرده و موارد با بیشترین تشابه را انتخاب می‌کند. برای مقایسه تصاویر زیرکلمه ورودی و زیرکلمات دیکشنری، اهمیت بیشتری به مناطق شاخص شکل داده می‌شود، مناطق شاخص مناطق محلی از زیرکلمه تعریف می‌شود که آن را از سایر زیرکلمات دیکشنری متفاوت می‌کند. در این روش ابتدا در یک معیار مبتنی بر میزان تاثیر ناحیه شاخص بر صحت بازیابی، برای منطقه شاخص محلی امتیازی محاسبه می‌گردد که نشان می‌دهد که آن ناحیه از شکل چقدر شاخص است. از این امتیازات برای تعدیل وزن‌های ویژگی‌های شکل مربوطه استفاده می‌شود، بطوریکه که مناطق با تمایز بیشتر، وزن بیشتری را به خود اختصاص می‌دهند. یک مرحله کاهش فضای ویژگی مبتنی بر ویژگی‌های کلی‌نگر مکان مشخصه به منظور تکمیل این توصیفگر محلی استفاده شده است. با روش پیشنهادی در پایگاه داده ابن سینا حاوی بیش از ۱۲۰۰۰زیرکلمه استخراج شده از یک سند تاریخی عربی، میزان کاهش ۹۸/۱۵٪ با دقت ۹۰/۱۵٪ به دست آمده است.

تحقیق [۸] یک روش بازشناسی کلمات دستنویس فارسی مبتنی بر کاهش فضای جستجو ارائه داده است. در این روش پس از استخراج ویژگی، کلمات در فرهنگ لغت خوشه‌بندی می‌شوند. میانگین هر خوشه در فضای ویژگی به عنوان نماینده خوشه و مدخل مشترک اعضای آن خوشه در فرهنگ لغت در نظر گرفته می‌شود. در آزمایش این روش روی مجموعه داده ایرانشهر (مجموعه‌ای متشکل از نام ۵۰۳ شهر ایران در قالب بیش از ۱۷۰۰۰ نمونه) در مرحله بازشناسی با انتخاب ۵ خوشه نزدیکتر به کلمه مورد آزمون با دقت ۹۳/۳۷٪ حدود ۷۶/۶۵٪ کاهش فضای جستجو را شاهد هستیم.

مرجع [۹] به دنبال روش‌هایی برای کاهش دامنه جستجوی زیرکلمه ناشناخته در واژه‌نامه تصویری، ایده استخراج حروف شاخص در ابتدا و انتهای زیرکلمات را مطرح کرده است. از آنجا که این حروف از لحاظ سبک نگارش فارسی، شکل منحصر به فردی دارند، بدون جداسازی حروف زیرکلمات، قابل بازشناسی هستند. در این تحقیق دو روش برای استخراج حروف شاخص در متن چاپی فارسی، ارائه شده است. الگوریتم اول، استخراج حروف شاخص بر مبنای برچسب‌زنی به کانتور بالایی زیرکلمه است. به این ترتیب که با استفاده از تشکیل کد زنجیره‌ای برای حروف شاخص بر اساس برچسب‌های کانتور بالایی، به جستجوی حروف شاخص در ابتدا و انتهای زیرکلمه می‌پردازد. ویژگی‌های استخراج شده، بالارونده‌ها و پایین‌رونده‌ها نسبت به خط زمینه، مشتق اول منحنی پیرامونی و زاویه انحنای کانتور زیرکلمه می‌باشند.

در الگوریتم پیشنهادی دوم، با استفاده از عملیات مورفولوژی، حروف شاخص اول و آخر را به ترتیب، با بخش ابتدایی و انتهای زیرکلمه ورودی تطبیق داده، به بازشناسی این حروف در زیرکلمات می‌پردازد. به دلیل حساسیت بالای عملگرهای مورفولوژی به نویز، الگوریتم شامل یک مرحله پیش‌پردازش نیز می‌باشد. الگوریتم پیشنهادی روی مجموعه پایگاه داده شامل ۷۳۱۷ زیرکلمه چاپی فارسی در چهار قلم و سه اندازه آزمایش شده است و فضای جستجو به طور میانگین تا میزان ۹۱/۱۵٪ با استخراج حروف شاخص اول و ۹۳/۱۵٪ پس از استخراج حروف شاخص آخر، کاهش می‌یابد.

مرجع [۱۰] از تخمین توصیفگر سازگار با شکل زیرکلمه جهت کاهش فضای جستجو استفاده کرده است. روش پیشنهادی به این ترتیب است که ویژگی‌های مکان مشخصه، ناحیه‌بندی و تبدیل فوریه بعنوان سه ویژگی مناسب جهت بازشناسی زیرکلمات فارسی انتخاب شده‌اند. در مرحله آموزش، یک شبکه عصبی جهت تشخیص توصیف کننده مناسب از بین سه توصیف کننده انتخاب شده آموزش می‌بینند. در مرحله آزمایش ابتدا شبکه عصبی آموزش دیده، ویژگی توصیفگر مناسب جهت زیرکلمه ورودی را تخمین می‌زند. سپس بر اساس این ویژگی تعدادی از نزدیکترین زیرکلمات به زیرکلمه ورودی بعنوان فضای جستجوی کاهش یافته انتخاب می‌شوند.

در تحقیقی دیگر ابراهیمی برای کاهش دامنه جستجو ابتدا زیرکلمات پایگاه داده را که از قلم‌ها و اندازه‌های متفاوت بودند به۳۰۰ خوشه تقسیم کرد. در مرحله طبقه‌بندی زیرکلمه ورودی

تنها با مرکز خوشه‌های یاد شده مقایسه شده و در مرحله بعد تنها نزدیکترین خوشه‌ها با استفاده از ویژگی‌های توصیفگرهای فوریه مورد جستجو قرار می‌گیرند [۱۱].

در تحقیق دیگری داودی کار ابراهیمی را به اینصورت بهبود داد که پس از خوشه‌بندی با ویژگی‌های سراسری از افزایش میزان اطمینان به خوشه انتخابی بر اساس ویژگی‌های محلی شکل زیرکلمه جهت کاهش فضای جستجو استفاده کرده است و به این ترتیب فضای جستجوی ۱۰ خوشه‌ای ارائه شده در تحقیق ابراهیمی بطور متوسط به ۴/۸ خوشه کاهش پیدا کرد [۱۲].

در تحقیق دیگری میری از خوشه‌بندی، کد موقعیت علائم و نسبت پهنا به ارتفاع زیرکلمات برای محدودسازی فضای جستجو استفاده کرده است [۱۳]. در این تحقیق در اولین مرحله با استخراج ویژگی‌های ساده‌ای از نمایه‌های افقی و عمودی، فضای جستجو به تعدادی از خوشه‌های منتخب محدود شده است. در دومین مرحله با تعیین نسبت پهنا به ارتفاع زیرکلمه، دامنه جستجو به زیرکلماتی با محدوده‌ای از این نسبت محدود شده است. در مرحله سوم با توجه به موقعیت علائم تنها زیرکلمه‌هایی مورد جستجو قرار می‌گیرند که موقعیت علائم آنها با زیرکلمه ورودی یکسان باشند. با اعمال روش پیشنهادی فضای جستجو تا حد قابل قبولی کاهش یافته است.

در مرجع [۱۴] جهت کاهش دامنه جستجو از ویژگی‌های ساده و در عین حال کارا استفاده شده است. کاهش فضای جستجو در چند مرحله صورت می‌گیرد. در اولین مرحله فضای جستجو تنها به زیرکلماتی که نسبت پهنا به ارتفاع (با علائم و بدون علائم) آنها در محدوده مشخصی باشند محدود می‌گردد در سطوح بعدی همین مراحل با نسبت تعداد نقاط سیاه تصویر به سطح تصویر زیرکلمه، نسبت تعداد نقاط سیاه نیمه بالایی تصویر به نقاط سیاه نیمه پایینی تصویر زیرکلمه و نسبت تعداد نقاط سیاه نیمه راست تصویر به نقاط سیاه نیمه چپ تصویر زیرکلمه در فضای محدود شده قبلی تکرار می‌شود و در نهایت برای کاهش فضای جستجو از موقعیت علائم استفاده شده است. در مراحل طراحی شده جهت کاهش فضای جستجو، این فضا با کاهش ۹۹/۳۶٪ (برای مجموعه داده‌های آزمون) از ۱۲۷۰۰ زیرکلمه به متوسط ۸۳/۷ زیرکلمه کاهش یافته است.

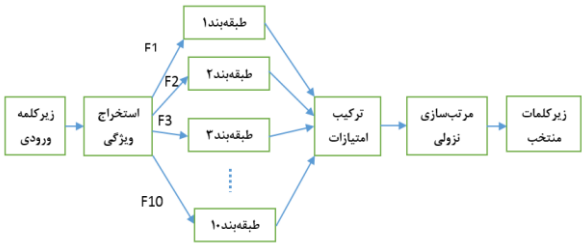
مصالحه بین دقت و اندازه فضای محدود شده می‌تواند بعنوان معیار درست یا غلط بودن نتیجه محدودسازی مد نظر قرار گیرد که منظور از دقت محدودسازی، این است که از تعداد کل نمونه‌های آزمون که محدودسازی در مورد آنها انجام می‌شود، برای چه درصدی از آنها نمونه هم‌کلاس با نمونه آزمون در فضای محدود شده وجود دارد. دو ضعف عمده روش‌های قبلی عدم دقت کافی و پیچیدگی برخی از این روش‌ها است. در این تحقیق یک روش جدید که علاوه برسادگی از دقت بالایی نیز برخوردار است جهت کاهش فضای جستجوی زیرکلمات فارسی ارائه می‌شود. در این راستا ابتدا ۱۰ ویژگی ساده از زیرکلمه استخراج می‌شود. با استفاده از مفهوم کوانتیزاسیون و با توجه به بازه تغییرات هر ویژگی روی همه داده‌های آموزشی، ویژگی‌ها کوانتیزه شده و به اعداد صحیحی تبدیل می‌شوند. با استفاده از هر ویژگی و فاصله آن تا ویژگی متناظر هرکدام از نمونه‌های آموزشی، به هر کلاس امتیازی داده می‌شود. با اعمال همه ویژگی‌ها، هر کلاس به ازای هر ویژگی یک امتیاز دارد که با ترکیب این امتیازات با اعمال جبری یک امتیاز نهایی برای هر زیرکلمه بدست می‌آید که با مرتب کردن آنها و انتخاب تعدادی از آنها که امتیاز بیشتری دارند، فضای جستجو محدود می‌شود.

نوآوری ارائه شده در این روش در درجه اول سادگی و دقت روش و در درجه بعد استفاده از الگوریتم‌های هوشمند جهت تخصیص وزن‌های بهینه به ویژگی‌ها و همچنین انتخاب هوشمند تعداد سطوح کوانتیزه برای هریک از ویژگی‌ها است.

پایگاه داده استفاده شده در این تحقیق که به منظور امکان مقایسه صحیح نتایج، با پایگاه داده استفاده شده در مراجع [۱۲] و [۱۳] یکسان است، عبارت است از مجموعه ۱۲۷۰۰ زیرکلمه رایج زبان فارسی که با قلم‌های لوتوس، میترا، نازنین، زر و یاقوت با اندازه قلم ۱۴ نگارش و چاپ شده و با درجه تفکیک ۴۰۰ نقطه در اینچ روبش شده‌اند (قسمتی از پایگاه ارائه شده در [۱۱])، که به عنوان داده‌های آموزشی مورد استفاده قرار گرفته‌اند و داده‌های آزمون عبارتند از ۵۰۰۰ تصویر زیرکلمه که از ۱۰۰۰ زیرکلمه با اندازه قلم و درجه تفکیک متفاوت تولید شده‌اند.

۲- روش پیشنهادی

در این تحقیق ۱۰ ویژگی ساده از هر زیرکلمه استخراج می‌شود. هر ویژگی، ورودی یک طبقه‌بندی‌کننده با خروجی امتیازدار است. یعنی هر طبقه‌بندی‌کننده به هر کلاس یک امتیاز با توجه به زیرکلمه ورودی می‌دهد. خروجی ۱۰ طبقه‌بندی‌کننده با هم ترکیب شده و برای هر کلاس یک امتیاز نهایی حاصل می‌شود. کلاس‌ها بر اساس امتیازشان به صورت نزولی مرتب شده و تعدادی از کلاس‌ها با بیشترین امتیاز انتخاب می‌شوند. بلوک دیاگرام روش پیشنهادی در شکل ۱ دیده می‌شود. در ادامه به شرح بخش‌های مختلف روش پیشنهادی پرداخته می‌شود.

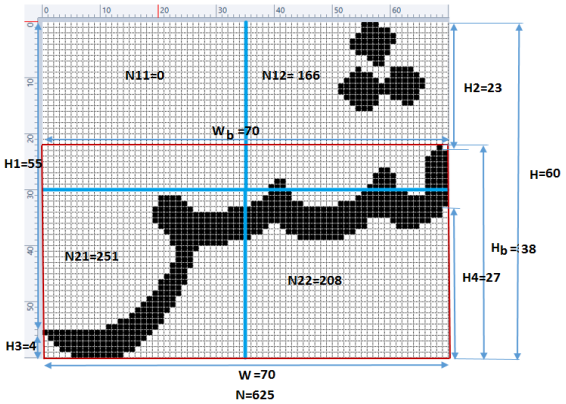


شکل ۱: بلوک دیاگرام روش پیشنهادی

۲-۱- استخراج ویژگی

در این مرحله از زیرکلمه ورودی ۱۰ ویژگی استخراج می‌شود. برای بیان روش استخراج ویژگی از شکل ۲ که تصویر یک زیرکلمه نمونه است استفاده شده است.

هر زیرکلمه فارسی دارای یک بدنه اصلی است و ممکن است تعدادی علامت مثل نقطه، سرکش، همزه و مد داشته باشد. در شکل ۲ زیرکلمه "شر" نمایش داده شده که بدنه اصلی آن "سر" و شامل سه نقطه است.



شکل ۲: تصویر یک زیرکلمه فارسی، نحوه استخراج ویژگی زیرکلمات

اولین ویژگی نسبت پهنا به ارتفاع چارچوب زیرکلمه است. در شکل ۲ پهنای کل زیرکلمه ۷۰ نقطه و ارتفاع آن ۶۰ است. ویژگی اول که با f_1 آن را نشان می‌دهیم برابر است با:

$$f_1 = \frac{W}{H} = \frac{70}{60} = 1.167 \qquad (۱)$$

این ویژگی در عین سادگی قابلیت ایجاد تمایز خیلی خوبی بین بعضی از زیرکلمات ایجاد می‌کند. به عنوان مثال زیرکلمه "شر" به راحتی از زیرکلمه "سر" می‌تواند جدا شود، چون این ویژگی برای زیرکلمه "شر" خیلی کمتر از مقدار آن برای زیرکلمه "سر" است. یا زیرکلمه "شر" از زیرکلمه "شمشیر" براحتی قابل تفکیک است.

با استفاده از این ویژگی موقعی که یک زیرکلمه با تعداد حروف کم وارد سیستم شود به راحتی از زیرکلماتی با تعداد حروف زیاد تفکیک خواهد شد و برعکس. در مرجع [۱۳] نیز از این ویژگی استفاده شده است ولی نحوه استفاده از این ویژگی در این تحقیق با مرجع مذکور متفاوت است که در ادامه تشریح شده است. در بعضی از تحقیقات گذشته نقاط و علائم زیرکلمه را کنار گذاشته‌اند و فقط از بدنه اصلی زیرکلمه برای محدودسازی فضای جستجو استفاده نموده‌اند. ولی در این تحقیق با توجه به این ویژگی و سایر ویژگی‌هایی که در ادامه می‌آیند. وجود نقاط و علائم به کاهش فضای جستجو کمک می‌کنند.

دومین ویژگی نسبت پهنا به ارتفاع چارچوب بدنه اصلی زیرکلمه است. در مثال فوق پهنای بدنه اصلی زیرکلمه ۷۰ نقطه و ارتفاع آن ۳۸ است. ویژگی دوم که با f_2 آن را نشان می‌دهیم برابر است با:

$$f_2 = \frac{W_b}{H_b} = \frac{70}{38} = 1.842 \qquad (۲)$$

با کمی دقت به این ویژگی و سایر ویژگی‌هایی که در ادامه معرفی می‌شوند، می‌توان دریافت که ویژگی‌های ساده و در عین حال موثری هستند و در بخش نتایج اثر آنها مشخص خواهد شد.

برای استخراج ویژگی‌های سوم تا ششم، چارچوب کل زیرکلمه به چهار ناحیه تقسیم شده، تعداد نقاط سیاه هر ناحیه شمارش شده و بر کل نقاط سیاه تصویر تقسیم می‌شود . تعداد نقاط نواحی به ترتیب با N11، N12، N21 و N22 و تعداد کل نقاط سیاه با N نشان داده شده‌اند و ویژگی‌های مرتبط با آنها به صورت زیر بدست می‌آیند.

$$f_3 = \frac{N_{11}}{N} = \frac{0}{625} = 0 \qquad (۳)$$

$$f_4 = \frac{N_{12}}{N} = \frac{166}{625} = 0.266 \qquad (۴)$$

$$f_5 = \frac{N_{21}}{N} = \frac{251}{625} = 0.402 \qquad (۵)$$

$$f_6 = \frac{N_{22}}{N} = \frac{208}{625} = 0.333 \qquad (۶)$$

برای استخراج ویژگی‌های هفتم تا دهم که با نگاه به اول و آخر زیرکلمه سعی در کاهش فضای جستجو با استفاده از پروفایل‌های ابتدا و انتهای زیرکلمات دارند (با الهام از مرجع [۹] که به حروف اول و آخر زیرکلمه توجه دارد) از پروفایل عمودی از بالا و پایین برای ستون‌های ابتدا و انتهای بدنه اصلی نسبت به چارچوب کل زیرکلمه استفاده می‌شود. این مقادیر که برای یک زیرکلمه نمونه در شکل ۲ به ترتیب با H1، H2، H3 و H4 نشان داده شده‌اند و

ویژگی‌های مرتبط با آنها از مقدار نسبی آنها نسبت به ارتفاع زیرکلمه بدست می‌آیند.

$$f_7 = \frac{H_1}{H} = \frac{55}{60} = 0.9167 \quad (۷)$$

$$f_8 = \frac{H_2}{H} = \frac{23}{60} = 0.3833 \quad (۸)$$

$$f_9 = \frac{H_3}{H} = \frac{4}{60} = 0.0667 \quad (۹)$$

$$f_{10} = \frac{H_4}{H} = \frac{38}{60} = 0.6333 \quad (۱۰)$$

لازم به ذکر است که برای محاسبه ۴ ویژگی اخیر برای بدست آوردن پروفایل‌ها ابتدا برای سه ستون اول و سه ستون آخر محاسبه شده و میانگین آنها استفاده شده است. اینکار اثر وجود نویز در ابتدا و انتهای بدنه زیرکلمه را از بین می‌برد.

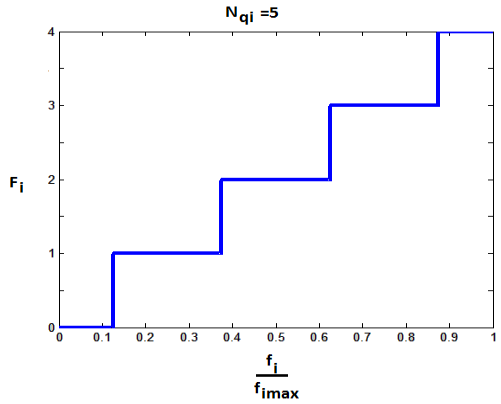
طی بررسی‌ها و آزمایشات صورت گرفته با استفاده از ۱۰ ویژگی مذکور، بعلت ایجاد خطا در حالت تطبیق کامل ولی نادرست برخی گزینه‌ها از نظر برخی از ویژگی‌ها که با اختصاص امتیاز بالا ولی نادرست به گزینه‌های نامناسب باعث ایجاد خطا در نتیجه نهایی می‌شود نتایج مطلوبی حاصل نشد. برای حل این مشکل با استفاده از کوانتیزاسیون، با ایجاد بازه تطبیق به جای تطبیق نقطه‌ای، امتیاز تطبیق‌های کامل ولی نادرست که موجب نتایج نامطلوب می‌شود تصحیح می‌گردد. در واقع با ایجاد سطوح محدود کوانتیزه شده که در ادامه توصیف شده‌اند به نوعی خوشه‌بندی روی داده‌ها را شاهد خواهیم بود و زیرکلمه‌های با ویژگی نزدیک به هم (در هر ویژگی بصورت مجزا) به یک سطح کوانتیزه شده تعلق خواهند گرفت و در مرحله بازشناسی نیز پس از استخراج ویژگی و کوانتیزه کردن آن، فاصله مقدار کوانتیزه زیرکلمه ورودی از سطوح کوانتیزه داده‌های آموزشی معیار امتیاز زیرکلمه ورودی خواهد بود..

۲-۲- کوانتیزاسیون ویژگی

روش استفاده شده برای کوانتیزاسیون به این ترتیب است که بعد از بدست آمدن همه ویژگی‌ها برای نمونه‌های آموزشی، بیشینه هر ویژگی بدست می‌آید و با توجه به آن ویژگی‌ها کوانتیزه می‌شوند. نحوه تبدیل ویژگی‌های اولیه به ویژگی‌های نهایی (کوانتیزه) در شکل ۳ نشان داده شده است.

در این شکل N_{qi} تعداد سطوح کوانتیزاسیون ویژگی i ام، f_i ویژگی اولیه i ام، f_{imax} بیشینه ویژگی اولیه i ام روی همه نمونه‌های آموزشی و F_i ویژگی نهایی i ام می‌باشند.

تعداد سطوح کوانتیزاسیون یکی از پارامترهایی است که تاثیر زیادی در دقت سیستم دارد که اثر آن در ادامه بررسی شده است. ضمنا تعیین تعداد سطوح کوانتیزه مناسب برای حصول بالاترین



شکل ۳: تبدیل ویژگی‌های اولیه به ویژگی‌های نهایی (کوانتیزه)

دقت ممکن برای هر ویژگی در این مقاله به الگوریتم‌های بهینه سازی جمعیت ذرات (PSO) [۱۶] سپرده شده است.

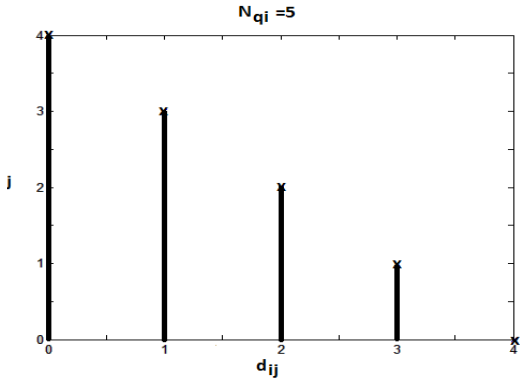
۲-۳- طبقه‌بندی‌کننده

همان‌طور که در شکل ۱ دیده می‌شود در روش پیشنهادی ۱۰ طبقه‌بندی‌کننده استفاده شده است. این طبقه‌بندی‌کننده‌ها فقط در ویژگی استفاده شده در آنها با هم تفاوت دارند. در هر طبقه‌بندی‌کننده فاصله ورودی تا هر کلاس محاسبه می‌شود. اگر ورودی طبقه‌بندی‌کننده i ام F_i و ویژگی i ام نماینده کلاس j ام برابر F_{ij} باشد، فاصله ورودی تا کلاس j ام از رابطه۱۱ بدست می‌آید.

$$d_{ij} = \left|F_{ij} - F_i\right| \quad (۱۱)$$

همانطور که گفته شد F_{ij} ویژگی i ام نماینده کلاس j ام است. شکل ۴ نحوه امتیازدهی با توجه به فاصله را نشان می‌دهد. اگر هر کلاس فقط یک نمونه آموزشی داشته باشد آن نمونه نماینده آن کلاس خواهد بود. اگر تعداد نمونه‌های آموزشی بیشتر از یکی باشد می‌توان از روش‌های مختلفی برای محاسبه فاصله (d_{ij}) استفاده کرد. مثلا می‌توان از میانگین آنها استفاده نمود یا اینکه هر نمونه را به عنوان یک زیرکلاس در نظر گرفت و کمترین فاصله بین نمونه آزمون و نمونه‌های آموزشی یک کلاس را بعنوان فاصله نمونه آزمون تا آن کلاس درنظر گرفت (روش استفاده شده در این تحقیق).

با توجه به فاصله بدست آمده امتیازی برای هرکلاس به ازای هر ویژگی از رابطه ۱۲ بدست می‌آید. به زبان ساده اگر فاصله صفر باشد، امتیاز یک واحد کمتر از تعداد سطوح کوانتیزاسیون است. به ازای هر واحد فاصله یک واحد امتیاز کم می‌شود. اگر فاصله حداکثر ممکن باشد، امتیاز صفر می‌شود.



شکل ۴: نحوه امتیازدهی با توجه به فاصله

۲-۴- ترکیب امتیازات

بعد از محاسبه امتیازات هر کلاس توسط طبقه‌بندی‌کننده‌های ۱۰گانه باید امتیاز نهایی هر کلاس با ترکیب امتیازات همه طبقه‌بندی‌کننده‌ها محاسبه شود. روش‌های مختلفی برای ترکیب امتیازات و بدست آوردن امتیاز نهایی می‌توان بکار برد که از جمله آنها می‌توان به جمع، ضرب، بیشینه، کمینه و جمع وزن‌دار امتیازها اشاره کرد که در این تحقیق برای بدست آوردن امتیاز نهایی از روش جمع وزن‌دار امتیازها استفاده شده و برای بدست آوردن ضرایب بهینه با معیار حصول بالاترین دقت ممکن، از الگوریتم بهینه‌سازی جمعیت ذرات (PSO) استفاده شده است. در این روش برای محاسبه امتیاز نهایی هر کلاس، امتیازاتی که هر کدام از طبقه‌بندی‌کننده‌ها به آن کلاس داده‌اند در ضرایب بهینه پیشنهادی الگوریتم PSO ضرب و سپس با هم جمع می‌شوند.

۳- آزمایشات و نتایج

در ابتدای این فصل برای بیان روشن‌تر روش پیشنهادی، مثالی با فضا و ابعاد کوچکتر آورده می‌شود. فرض کنید می‌خواهیم فضای جستجو را برای بازشناسی حروف مجزایی که در پایگاه داده استفاده شده وجود دارند از ۳۶ کلاس به ۴ کلاس کاهش دهیم. (علاوه بر ۳۲ حرف فارسی حروف {آ، ء، أ، ؤ} نیز در پایگاه داده وجود دارند.) از سه ویژگی اول، سوم و هفتم که در بخش استخراج ویژگی معرفی شدند می‌خواهیم استفاده کنیم، بنابراین ۳ طبقه‌بندی‌کننده خواهیم داشت. برای ترکیب امتیازات هم از روش جمع کمک می‌گیریم. تعداد سطوح کوانتیزاسیون نیز ۵ در نظرگرفته می‌شود.

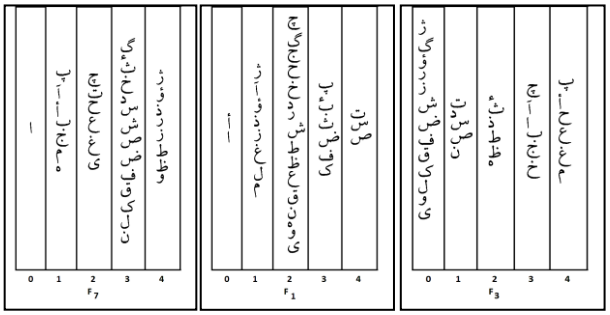
ابتدا ویژگی‌های مورد نظر از نمونه‌های آموزشی استخراج شده است. این ویژگی‌ها درستون‌های دوم تا چهارم جدول ۱ آمده است. سپس مقادیر بیشینه هرکدام از ویژگی‌ها بدست آمده است. این مقادیر برای ۳ ویژگی استفاده شده به ترتیب ۱/۷۷۲۷، ۰/۴۳۴۷ و ۰/۹۱۳۸ می‌باشند.

جدول ۱: حروف موجود در داده‌های آموزشی و ویژگی‌های مستخرج از آنها

حرف	f_1	f_3	f_7	F_1	F_3	F_7
پ	۱.۱۴۸۹	۰.۴۰۲۰	۰.۱۲۷۷	۳	٤	۱
چ	۰.۹٤۱۲	۰.۳۳۸۷	۰.٤۹٦۷	۲	۳	۲
ژ	۰.٤٤۸۳	۰.۰۳۳۸	۰.۹۱۳۸	۱	۰	٤
گ	۰.۹۸۱۸	۰.۰۲۰۲	۰.٦٤۲٤	۲	۰	۳
ء	۱.۱۱۱۱	۰.۲۲۱٦	۰.۷۵۹۳	۳	۲	۳
آ	۰.۳۰۹۱	۰.۳۱۱۲	۰.۳۰۳۰	۱	۳	۱
أ	۰.۱٦٦۷	۰.۳۸۳۹	۰.۳۰۵٦	۰	٤	۱
ؤ	۰.۵۰۰۰	۰.۰۰۰۰	۰.۸۸٦۹	۱	۰	٤
ا	۰.۱۳٦٤	۰.۳۵۹۲	۰.۰٤۵۵	۰	۳	۰
ب	۱.۲۵۵۸	۰.۳۵۳۱	۰.۱٦۲۸	۳	۳	۱
ت	۱.۵۸۸۲	۰.۱۲٤۰	۰.٤۳۱٤	٤	۱	۲
ث	۱.۳۰۹۵	۰.۱٦۵۷	۰.۵۸۷۳	۳	۲	۳
ج	۰.۹٤۱۲	۰.۳٦۸۲	۰.۲۵٤۹	۲	۳	۱
ح	۰.۹۲۱٦	۰.۳۸۳٦	۰.۳۵۹۵	۲	٤	۲
خ	۰.٦۹۵۷	۰.۳۷٤٦	۰.۵۷٤۹	۲	۳	۳
د	۰.۷۱٤۳	۰.۱٤٤۵	۰.٦۹۰۵	۲	۱	۳
ذ	۰.٤٦۵۱	۰.۲۵٤۲	۰.۸۱٤۰	۱	۲	٤
ر	۰.۷۸۱۳	۰.۰۰۰۰	۰.۸۳۳۳	۲	۰	٤
ز	۰.۵۵۳۲	۰.۰۰۰۰	۰.۸۹۳٦	۱	۰	٤
س	۱.۷۵٦۸	۰.۰۵۸۷	۰.٦۲۱٦	٤	۱	۳
ش	۱.۰٦۵٦	۰.۰۰٤۷	۰.۷۱۰٤	۲	۰	۳
ص	۱.۷۷۲۷	۰.۰۹۵۲	۰.۵۹۸۵	٤	۱	۳
ض	۱.۳۰۰۰	۰.۰۰۷۷	۰.٦۹٤٤	۳	۰	۳
ط	۰.۸۹۵۸	۰.۱۸۸۳	۰.۸۵٤۲	۲	۲	٤
ظ	۰.۹۱٦۷	۰.۱٦۵۳	۰.۸۵٤۲	۲	۲	٤
ع	۰.۷۱٤۳	۰.٤۳۱۳	۰.۵۳٤٤	۲	٤	۲
غ	۰.۵۹۷٤	۰.٤۲۹۰	۰.۵۵۸٤	۱	٤	۲
ف	۱.۲۷۹۱	۰.۰۰۳٤	۰.۵۹٦۹	۳	۰	۳
ق	۰.٦۹٦٤	۰.۰۰۱۷	۰.۵۷۷٤	۲	۰	۳
ک	۱.۱٤۸۹	۰.۰۱۵۹	۰.۵۹۵۷	۳	۰	۳
ل	۰.۵۹۲٦	۰.۰۰۰۰	۰.٦۱۱۱	۱	۰	۳
م	۰.۳۸۱۸	۰.٤۳٤۷	۰.۲۳۰۳	۱	٤	۱
ن	۰.٦۷۳۹	۰.۱۲۸٦	۰.٦۰۱٤	۲	۱	۳
ه	۰.۷٦۱۹	۰.۲٦۲۷	۰.۲۳۸۱	۲	۲	۱
و	۰.۸۰۰۰	۰.۰۰۰۰	۰.۸۲۸٦	۲	۰	٤
ی	۰.۹۷۷۸	۰.۰۳۹٤	۰.۵۲۵۹	۲	۰	۲

با توجه به این مقادیر و روش کوانتیزاسیون ویژگی توضیح داده شده در بخش ۲ ویژگی‌ها کوانتیزه شده و در ستون‌های پنجم تا هفتم جدول ۱ ذخیره شده‌اند. مقادیر این ۳ ستون از جدول و ۳ مقدار ماکزیمم بعنوان داده آموزشی ذخیره شده و در مرحله آزمون از آنها استفاده خواهد شد.

برای نشان دادن قابلیت ویژگی‌های استخراجی در جداسازی حروف قسمت‌های مختلف اجزاء شکل ۵ به ترتیب نشان می‌دهند که تصاویر چه حروفی از داده‌های آموزشی از نظر ویژگی‌های F1، F3 و F7 یکسانند و نزدیکی یا دوری حروف از همدیگر مشخص است. حروفی که در یک ستون قرار دارند ویژگی یکسان دارند و هر چه ستون‌های بین دو حرف بیشتر باشد شباهتشان نسبت به هم با توجه به آن ویژگی کمتر است.



شکل ۵: تصاویر داده های آموزشی از نظر ویژگی های F1، F3 و F7 و نزدیکی یا دوری حروف از همدیگر

نمونه آزمون ورودی در شکل ۶ دیده می‌شود. این نمونه از همان قلم نمونه‌های آموزشی ولی با اندازه کوچکتر و دقت روبش کمتر است. از این تصویر ویژگی‌های اولیه استخراج و کوانتیزه می‌شوند. ویژگی‌های اولیه [۰/۶۶۶۷ ۰ ۰/۹۷۰۶] و ویژگی‌های کوانتیزه شده [۲ ۰ ۳] می‌باشند. با استفاده از جدول ۱ فاصله این ویژگی‌ها از ویژگی‌های نمونه‌های آموزشی محاسبه شده و در ستون‌های ۲ تا ۴ جدول ۲ آمده است. امتیازات مربوطه نیز محاسبه شده و در ستون‌های ۵ تا ۷ جدول ۲ آورده شده‌است. ستون آخر امتیاز نهایی هر حرف با استفاده از جمع ۳ امتیاز ۳ طبقه‌بند است.

شکل ۶: تصویر یک نمونه آزمون

جدول ۲: فاصله‌های ویژگی‌های نمونه آزمون از داده‌های آموزشی و امتیازات مربوطه

حرف	d۱	d۳	d۷	s۱	s۳	s۷	s
پ	۱	٤	۲	۳	۰	۲	٥
چ	۰	۳	۱	٤	۱	۳	۸
ژ	۱	۰	۱	۳	٤	۳	۱۰
گ	۰	۰	۰	٤	٤	٤	۱۲
ء	۱	۲	۰	۳	۲	٤	۹
آ	۱	۳	۲	۳	۱	۲	٦
أ	۲	٤	۲	۲	۰	۲	٤
ؤ	۱	۰	۱	۳	٤	۳	۱۰
ا	۲	۳	۳	۲	۱	۱	٤
ب	۱	۳	۲	۳	۱	۲	٦
ت	۲	۱	۱	۲	۳	۳	۸
ث	۱	۲	۰	۳	۲	٤	۹
ج	۰	۳	۲	٤	۱	۲	۷
ح	۰	٤	۱	٤	۰	۳	۷
خ	۰	۳	۰	٤	۱	٤	۹
د	۰	۱	۰	٤	۳	٤	۱۱
ذ	۱	۲	۱	۳	۲	۳	۸
ر	۰	۰	۱	٤	٤	۳	۱۱
ز	۱	۰	۱	۳	٤	۳	۱۰
س	۲	۱	۰	۲	۳	٤	۹
ش	۰	۰	۰	٤	٤	٤	۱۲
ص	۲	۱	۰	۲	۳	٤	۹
ض	۱	۰	۰	۳	٤	٤	۱۱
ط	۰	۲	۱	٤	۲	۳	۹
ظ	۰	۲	۱	٤	۲	۳	۹
ع	۰	٤	۱	٤	۰	۳	۷
غ	۱	٤	۱	۳	۰	۳	٦
ف	۱	۰	۰	۳	٤	٤	۱۱
ق	۰	۰	۰	٤	٤	٤	۱۲
ک	۱	۰	۰	۳	٤	٤	۱۱
ل	۱	۰	۰	۳	٤	٤	۱۱
م	۱	٤	۲	۳	۰	۲	٥
ن	۰	۱	۰	٤	۳	٤	۱۱
ه	۰	۲	۲	٤	۲	۲	۸
و	۰	۰	۱	٤	٤	۳	۱۱
ی	۰	۰	۱	٤	٤	۳	۱۱

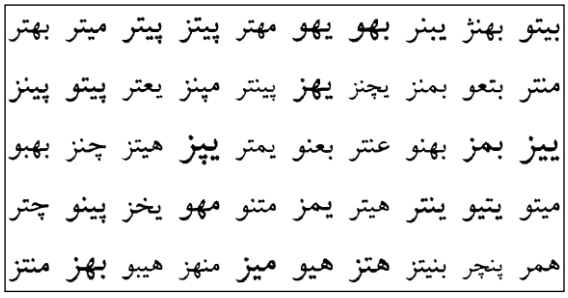
جدول ۳ حروف و امتیازات هر کدام را که بصورت نزولی و ترتیب حروف الفبا مرتب شده‌اند نشان می‌دهد. با توجه به جدول ۳ و با توجه به هدفی که قرار بود فضای جستجو به ۴ کلاس محدود

شود، ۴ حرف اول لیست مرتب شده { گ، ش، ق، د} انتخاب می‌شوند. در مرحله بازشناسی نهایی نمونه آزمون باید از طبقه‌بندی استفاده شود که خروجی آن فقط این ۴ حرف هستند.

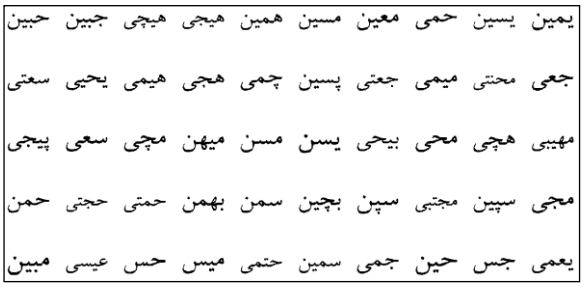
جدول ۳: امتیازات مرتب شده براساس امتیاز کسب شده (ترتیب الفبا)

حرف	s
گ	۱۲
ش	۱۲
ق	۱۲
د	۱۱
ر	۱۱
ض	۱۱
ف	۱۱
ک	۱۱
ل	۱۱
ن	۱۱
و	۱۱
ی	۱۱
ژ	۱۰
ؤ	۱۰
ز	۱۰
ء	۹
ث	۹
خ	۹
س	۹
ص	۹
ط	۹
ظ	۹
چ	۸
ت	۸
ذ	۸
ه	۸
ج	۷
ح	۷
ع	۷
آ	٦
ب	٦
غ	٦
پ	٥
م	٥
أ	٤
ا	٤

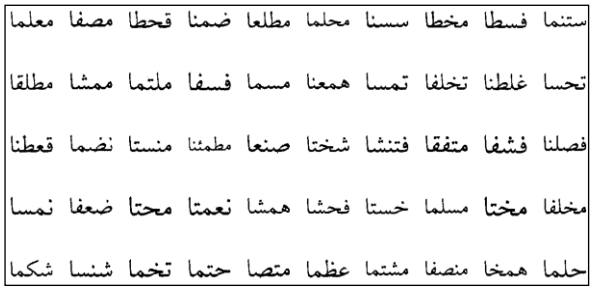
شکل ۷ چند نمونه از فضای محدود شده به پنجاه کلاس پیشنهادی نخست ناشی از اجرای کامل الگوریتم با ترتیب نزدیکی به گزینه وارده را برای چند زیرکلمه نمونه نشان می‌دهد. تشابه و تفاوت‌های زیرکلمات موجود در مجموعه محدود شده مبنای طراحی مراحل بعدی بازشناسی زیرکلمه هدف از این مجموعه خواهد بود. بطور مثال اگر کد موقعیت علائم معرفی شده در مرجع [۱۳] مورد استفاده قرار گیرد می‌توانند بسیار راهگشا باشند. مثلا در نمونه الف که مجموعه محدود شده جهت بازشناسی زیرکلمه "بهتر" را نشان می‌دهد، استفاده از موقعیت علائم فضا را به مجموعه ۱۵ زیرکلمه‌ای {میتر، بهتر، بتعو، یهز، یعتر، بمز، بهنو، بعنو، یمتر، میتو، هیتر، یمز، چتر، میز، بهز} کاهش می‌دهد در حالی که همین موضوع در قسمت ج شکل ۷، فضای جستجو برای زیرکلمه "معلما" را به مجموعه {محلما، مطلعا، معلما، مسلما، مسما، حلما } کاهش داده و باعث می‌شود تنها فضای جستجوی شامل این ۶ زیرکلمه برای مرحله بعد باقی بماند.



الف: زیرکلمه ورودی"بهتر"



ب: زیرکلمه ورودی"هیچی"



ج: زیرکلمه ورودی"معلما"

شکل ۷: چند نمونه از فضای محدود شده به نخستین پنجاه کلاس پیشنهادی برای چند زیرکلمه ورودی

در مثال ارائه شده، روش پیشنهادی برای بازشناسی حروف و تعداد

ویژگی‌های کمتر برای درک بهتر و بیشتر خوانندگان این مقاله بیان شد. ولی روش پیشنهادی همانطور که در بخش ۲ بیان شده برای زیرکلمات تایپی فارسی که ۱۲۷۰۰ کلاس دارند و با استفاده از ۱۰ ویژگی ارائه، پیاده‌سازی و آزمایش شده است. هدف‌گذاری مولفین مقاله این بوده است که فضای جستجو را از ۱۲۷۰۰ به نزدیک به ۵۰ محدود نمایند و دقت محدودسازی نیز در حد کارهای قبلی و حتی بهتر از آنها باشد.

در انجام آزمایشات این قسمت، داده‌های آموزشی تصاویر ۱۲۷۰۰ زیرکلمه با قلم لوتوس اندازه ۱۴ اند که با دقت ۴۰۰ نقطه بر اینچ روبش شده‌اند. داده‌های آزمون نیز عبارتند از تصاویر ۵۰۰۰ نمونه زیرکلمه که با قلم لوتوس و اندازه قلم و دقت روبش متفاوت جمع آوری شده و به دو گروه ۲۵۰۰ تایی تقسیم شده‌اند. با استفاده از۲۵۰۰ نمونه اول پارامترهای روش پیشنهادی بدست آمده و علاوه بر این ۲۵۰۰ نمونه، روی ۲۵۰۰ نمونه دیگر نیز که نقشی در تنظیم پارامترها نداشته‌اند آزمایشات صورت گرفته است (پایگاه داده استفاده شده در مراجع [۱۲] و [۱۳]).

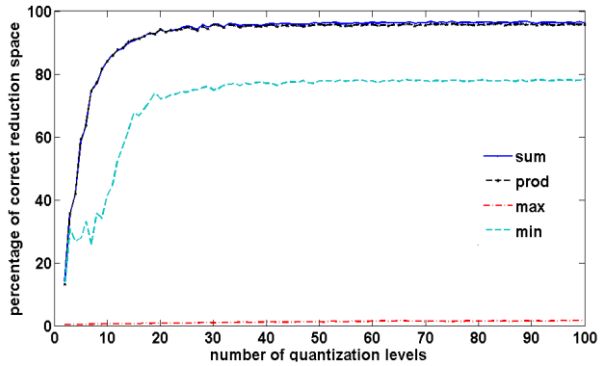
برای روشن شدن نقش ترکیبی ویژگی‌های استخراجی، در اولین قدم هرکدام از ویژگی‌ها به تنهایی و ترکیبات مختلف ویژگی‌ها (جمع ساده امتیاز ویژگی‌های ترکیب شده) بر روی ۲۵۰۰ نمونه اول مورد آزمون قرار گرفت که تعدادی از نتایج بدست آمده در جدول ۴ نشان داده شده است. همانطور که در نتایج حاصل مستتر است استفاده از هریک از ویژگی‌ها به تنهایی کارایی چندانی ندارد اما ترکیب این ویژگی‌ها با روش پیشنهادی منجر به نتایج قابل توجهی می‌گردد.

جدول ۴: دقت فضای محدود شده ناشی از هریک از ویژگی‌ها به تنهایی و ترکیب‌های مختلف از ویژگی‌ها

ویژگی استفاده شده	دقت حاصل شده (درصد)
ویژگی ۱ به تنهایی	۴/۸
ویژگی ۲ به تنهایی	۵/۱۲
ویژگی ۳ به تنهایی	۵/۴۴
ویژگی ۴ به تنهایی	۵/۲۴
ویژگی ۵ به تنهایی	۴/۲۸
ویژگی ۶ به تنهایی	۳/۶
ویژگی ۷ به تنهایی	۴/۴۴
ویژگی ۸ به تنهایی	۳/۶۴
ویژگی ۹ به تنهایی	۱/۵۲
ویژگی ۱۰ به تنهایی	۴/۶
ویژگی های ۱و۲	۱۷
ویژگی های ۱و۲و۳	۵۰/۶۴
ویژگی های ۱و۲و۳و۴	۸۲/۰۸
ویژگی های ۱و۲و۳و۴و۵	۹۰/۶۸

۳-۱- تاثیر تعداد سطوح کوانتیزاسیون و روش ترکیب امتیازات بر دقت محدودسازی

یکی از پارامترهای تاثیرگذار بر میزان دقت، تعداد سطوح کوانتیزاسیون است. در اولین آزمایش تعداد سطوح کوانتیزاسیون از ۲ تا ۱۰۰ تغییر کرده و دقت محدودسازی با اندازه فضای محدود شده ۵۰ تایی بدست آمده است. منظور از دقت محدودسازی، این است که از تعداد کل نمونه‌های آزمون که محدودسازی در مورد آنها انجام می‌شود، برای چه درصدی از آنها نمونه هم‌کلاس با نمونه آزمون در فضای محدودشده وجود دارد. روش‌های جمع، ضرب، بیشینه و کمینه نیز آزمایش شده‌اند. در شکل ۸ دقت محدودسازی به درصد برحسب تعداد سطوح کوانتیزاسیون با چهار روش ترکیب امتیازات رسم شده است که همانطور که دیده می‌شود روش‌های جمع و ضرب کارایی بیشتری دارند. بیشترین دقت بدست آمده روی ۲۵۰۰ نمونه اول ٪۹۶/۷۲ برای روش ترکیب جمع و تعداد سطوح کوانتیزاسیون ۷۵ است. با تعداد سطوح کوانتیزاسیون ۷۵ و روش ترکیب جمع، دقت بدست آمده روی ۲۵۰۰ نمونه دوم ٪۹۶/۵۲ بدست آمده است. لازم به ذکر است چون برای تبدیل ویژگی‌های اولیه به ویژگی‌های کوانتیزه فقط یک عمل ضرب، یک عمل تقسیم و یک عمل گردکردن لازم است، تعداد سطوح کوانتیزاسیون تاثیری در سرعت روش پیشنهادی ندارد. در ضمن همانطور که این شکل نشان می‌دهد در سطوح پایین افزایش تعداد سطوح باعث بهبود سریع در نتایج می‌گردد اما با اضافه شدن سطوح از حدی به بعد منحنی اشباع شده و افزایش تعداد سطوح تاثیر چندانی در بهبود نتیجه ندارد.

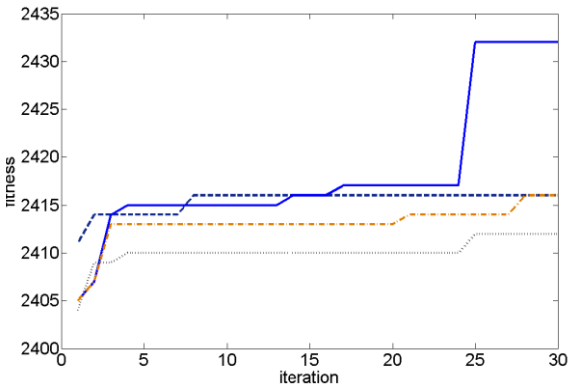


شکل ۸: تاثیر تعداد سطوح کوانتیزاسیون و روش ترکیب امتیازات بر دقت محدودسازی.

۳-۲- انتخاب تعداد سطوح متفاوت کوانتیزاسیون برای ویژگی‌ها

در بخش قبل تعداد سطوح کوانتیزاسیون برای همه ویژگی‌ها یکسان در نظر گرفته شده است. در این بخش تاثیر انتخاب سطح

کوانتیزاسیون مخصوص برای هر ویژگی مورد بررسی قرار می‌گیرد. برای نرمال کردن امتیازات، در این بخش امتیازی که هر طبقه‌بندی‌کننده به هر کلاس می‌دهد بر یک واحد کمتر از تعداد سطوح کوانتیزاسیون تقسیم می‌شود تا بیشترین امتیاز ممکن، یک و کمترین امتیاز ممکن صفر باشد. در غیر اینصورت هر ویژگی که تعداد سطوح کوانتیزاسیون بیشتری داشته باشد با وزن بیشتری در جمع اثر خواهد گذاشت. برای بدست آوردن تعداد سطوح بهینه هر ویژگی الگوریتم PSO با در نظرگرفتن c1=c2=2، W=0.8 و جمعیت اولیه ۲۰ و ۳۰ تکرار انجام شد. هدف در نظر گرفته شده در این مساله مینیمم کردن تعداد خطاهای محدودسازی در ۲۵۰۰ نمونه اول با یافتن سطوح بهینه کوانتیزاسیون برای همه ویژگی‌ها به صورت همزمان است که الگوریتم PSO این مقدار را در چندین مرحله تکرار آزمایش شده در بهترین حالت به ۶۸ خطا کمینه می‌کند. که با توجه به آن، دقت محدود سازی با ٪۰/۵۶ بهبود نسبت به حالت کوانتیزاسیون با سطوح یکسان، ٪۹۷/۲۸ گردیده است. روال چهار اجرای نمونه این الگوریتم در یافتن تعداد سطوح بهینه در شکل ۹ دیده می‌شود. مقادیر بهینه بدست آمده برای تعداد سطوح کوانتیزاسیون در جدول ۵ ارائه شده است. برای ۲۵۰۰ نمونه دوم با استفاده از این سطوح کوانتیزاسیون تعداد خطا ۶۴ شده است، لذا نرخ محدودسازی برای آنها با ٪۰/۹۲ بهبود ٪۹۷/۴۴ گردیده است.



شکل ۹: روندکمینه سازی تعداد خطاها با تغییر تعداد سطوح کوانتیزاسیون با استفاده از الگوریتم PSO. (محور عمودی تعداد نمونه‌هایی از ۲۵۰۰ زیرکلمه را که محدودسازی در مورد آنها به درستی انجام شده است را نشان می دهد.)

جدول ۵: تعداد سطوح کوانتیزاسیون بهینه برای هر ویژگی بدست آمده با الگوریتم PSO

Nq۱	Nq۲	Nq۳	Nq۴	Nq۵	Nq۶	Nq۷	Nq۸	Nq۹	Nq۱۰
۷۹	۵۳	۷۵	۵۵	۳۸	۵۲	۲۸	۵۱	۵۶	۴۶

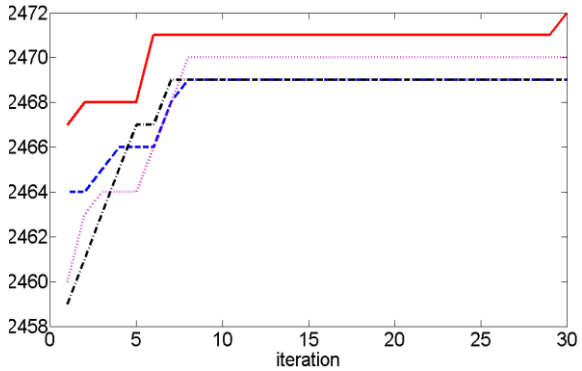
۳-۳- ترکیب امتیازات با جمع وزن دار

با توجه به اینکه همه ویژگی‌های استفاده شده از نظر قابلیت تمایز یکسان نیستند در این بخش وزن‌هایی برای امتیازات در نظرگرفته شده است. یعنی امتیاز هر کلاس از مجموع وزن‌دار امتیازات کل طبقه‌بندی‌کننده‌ها طبق رابطه زیر بدست می‌آید.

$$s_j = \sum_{i=1}^{10} w_i s_{ij} \qquad (۱۳)$$

برای بدست آوردن وزن‌های بهینه، الگوریتم PSO با در نظرگرفتن $c_1=c_2=2$، W=0.8 و جمعیت اولیه ۳۰ و ۳۰ تکرار انجام شد. هدف در نظرگرفته شده در این مساله کمینه کردن تعداد خطاهای محدودسازی در ۲۵۰۰ نمونه اول است که الگوریتم PSO این تعداد را کمینه می‌کند. شکل ۱۰ به عنوان نمونه چهار روند الگوریتم PSO را برای یافتن وزن‌های بهینه نشان می‌دهد. بهترین نتیجه بدست آمده در این اجراها ۲۴۷۲ (۲۸خطا) است که مفهوم آن این است که در ۲۵۰۰ نمونه فقط ۲۸ نمونه وجود دارند که اگر فضای جستجو را به ۵۰ زیرکلمه محدود کنیم زیرکلمه هم‌کلاسشان در فضای محدود شده وجود ندارد. یعنی دقت محدود سازی به ٪۹۸/۸۸ رسیده است. جدول ۶ وزن‌های بهینه پیشنهادی الگوریتم PSO را نشان می‌دهد.

با آزمایش روی ۲۵۰۰ نمونه دوم با استفاده از وزن‌های بدست آمده از مرحله قبل و با وجود اینکه این نمونه‌ها نقشی در یافتن وزن‌های بهینه نداشته‌اند با بهبود قابل توجه در دقت، فقط ۲۰ نمونه خطا وجود دارد. یعنی دقت محدودسازی روی این نمونه‌ها ٪۹۹/۲ و روی کل نمونه‌های آزمون ٪۹۹/۰۴ است.



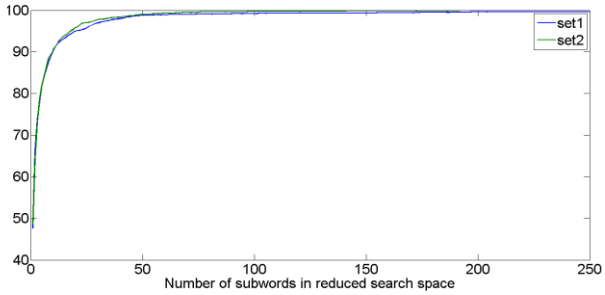
شکل ۱۰: روند کمینه سازی تعداد خطاها با تغییر وزن ویژگی‌ها با استفاده از الگوریتم PSO.

جدول ۶: وزن‌های بهینه بدست آمده با استفاده از الگوریتم PSO

w۱	w۲	w۳	w۴	w۵	w۶	w۷	w۸	w۹	w۱۰
۱/۴۸	۱/۷۱	۱/۴۳	۰/۷۳	۱/۱۱	۰/۴۸	۰/۸۷	۰/۹۶	۰/۱	۰/۷۵

۳-۴- دقت محدودسازی و تعداد کلاس‌های فضای محدود شده

در این بخش اثر تعداد کلاس‌های فضای محدود شده (میزان کاهش فضای جستجوی هدف تعیین شده) بر دقت محدودسازی (درصد وجود زیرکلمه مورد جستجو در فضای محدود شده) بررسی شده است. برای این منظور با استفاده از تعداد سطوح کوانتیزاسیون بهینه بدست آمده برای هر ویژگی و روش جمع وزن‌دار با وزن‌های بهینه بدست آمده در بخش قبل تعداد کلاس‌های فضای محدود شده از ۱ تا ۲۵۰ تغییر داده و دقت محدودسازی برای هر کدام از دو مجموعه ۲۵۰۰ عضوی آزمایش محاسبه شده است. نتایج حاصل در شکل ۱۱ نشان داده شده است. برای نشان دادن مقادیر عددی با وضوح بیشتر در جدول ۷ نتایج مربوط به آزمایش مربوط به ۲۵۰۰ نمونه دوم ارائه شده است. برای حالات مختلف قبلی می‌توان این اطلاعات را بدست آورد که برای جلوگیری از طولانی شدن گزارش و بعلت تشابه نتایج از ارائه آن اجتناب شده است.



شکل ۱۱: رابطه تعداد کلاس‌ها ودقت محدود سازی برای دو مجموعه آزمایشPSO .

جدول ۷: رابطه تعداد کلاس‌های انتخابی(تعداد اعضای فضای محدود شده) و دقت محدود سازی(مجموعه آزمایش دوم).

تعداد کلاس	۱	۲	۳	۴	۵	۷	۹	۱۲
میزان پوشش	۱۱۶۰	۱۵۷۳	۱۷۸۱	۱۹۱۷	۱۹۹۸	۲۱۲۷	۲۲۰۷	۲۲۹۳
تعداد کلاس	۱۵	۱۸	۲۱	۲۴	۲۸	۳۲	۳۶	۴۰
میزان پوشش	۲۳۳۹	۲۳۶۴	۲۳۹۳	۲۴۱۱	۲۴۳۳	۲۴۴۷	۲۴۵۷	۲۴۶۶
تعداد کلاس	۴۸	۵۶	۷۰	۹۰	۱۲۰	۱۷۰	۲۱۰	۲۵۰
میزان پوشش	۲۴۷۹	۲۴۸۴	۲۴۸۹	۲۴۹۱	۲۴۹۶	۲۴۹۷	۲۴۹۷	۲۴۹۸

همانطور که در جدول ۷ دیده می‌شود تنها با انتخاب اولین کلاس (انتخاب نزدیکترین زیرکلمه به زیرکلمه ورودی) با دقت ٪۴۶/۰۴ زیرکلمه انتخابی همان زیرکلمه ورودی است. با انتخاب ۴۴ کلاس اول با دقت ٪۹۹ زیرکلمات انتخابی حاوی زیرکلمه ورودی خواهد بود و با صرف نظر از این دقت و اکتفا به دقت ٪۹۸ تنها با انتخاب ۳۴ کلاس اول با این مقصود می‌رسیم که نشان دهنده کاهش فضای جستجو در حد ٪۹۹/۶۵ با دقت ٪۹۹ و ٪۹۹/۷۳ با دقت ٪۹۸ است. با توجه به حصول دقت ٪۹۹/۱۷ در [۱۳] و جهت مقایسه نتایج از جدول ۷ می‌توان استنتاج نمود که برای رسیدن به این دقت می‌بایست ۴۸ کلاس اول انتخاب گردد که منجر به کاهش فضای جستجو در حد ٪۹۹/۶۲ می‌گردد.

۳-۵- مقایسه نتایج با کارهای گذشته

در روش ارائه شده در این مقاله در مقایسه با روش‌های قبلی از جمله روش ارائه شده در [۱۱] و [۱۲] و [۱۳] علاوه بر سادگی و سرعت، نتایج نسبتا بهتری حاصل شده است که این نتایج که با پایگاه داده یکسان بدست آمده‌اند در جدول ۸ ارائه شده است. همانطور که قبلا نیز بیان شد در [۱۱] پس از خوشه‌بندی به ۳۰۰ خوشه ۱۰ خوشه اول بعنوان فضای کاهش یافته معرفی شده و در روش ارائه شده در [۱۲] داوودی موفق شده با یک مرحله بهبود فضای جستجو را با حفظ دقت ٪۹۹/۱۷ به ۴/۸ خوشه از ۳۰۰ خوشه کاهش دهد و در روش معرفی شده در [۱۳] فضای ویژگی در چند مرحله طراحی شده ٪۹۹/۲ کاهش یافته است.

در مقایسه با سه روش یاد شده روش ارائه شده در این مقاله در عین سادگی و سرعت با همان دقت فضای جستجو را تا ٪۹۹/۶۲ کاهش داده است. جدول ۹ زیرکلماتی که در مجموعه زیرکلمات آزمایش با انتخاب ۵۰ گزینه اول به درستی بازشناسی نمی‌شوند (گزینه صحیح در بین پنجاه گزینه اول نیست) را به همراه انتخاب صحیح که نشان دهنده تعداد کلاسی است که باید انتخاب شود تا منجر به بازشناسی درست زیرکلمه شود و اندازه قلم و دقت روبش شکل زیرکلمه را نشان می‌دهد. که در اتخاذ تصمیم جهت بهبود در کارهای بعدی کارساز خواهد بود.

جدول ۸: مقایسه نتایج روش‌های مختلف

روش	درصد کاهش دامنه جستجو
خوشه‌بندی با ویژگی‌های سراسری (ابراهیمی [۱۱])	٪۹۶/۶
خوشه‌بندی با ویژگی‌های سراسری و نواحی شاخص (داوودی [۱۲])	٪۹۸/۴
خوشه‌بندی با متوسط ۴ قلم و نسبت پهنا به ارتفاع و موقعیت علائم [۱۳]	٪۹۹/۲
روش ارائه شده در این مقاله	٪۹۹/۶

۳-۶- آزمایش با قلم‌های مختلف

به منظور بررسی عملکرد روش پیشنهادی در مواجه با قلم‌های مختلف دو آزمایش طراحی شد. در اولین آزمایش پایگاه داده معرفی شده قبلی (۱۲۷۰۰ زیرکلمه) با چهار قلم یاقوت، نازنین، میترا و زر با اندازه قلم ۱۴ که با دقت ۴۰۰ نقطه بر اینچ روبش شده‌اند بعنوان داده‌های آموزش و ۲۵۰۰ زیرکلمه بطور تصادفی از

جدول ۹: زیرکلماتی که در مجموعه زیرکلمات آزمایش با انتخاب ۵۰ گزینه اول به درستی بازشناسی نمی‌شوند. منظور از انتخاب صحیح رتبه کلاس زیرکلمه مورد جستجو است که در موارد زیر بزرگتر از ۵۰ است و در فضای محدود شده نمی باشد.

زیرکلمه	'فح'	'بعت'	'تمس'	'تمس'	'تمس'	'قسا'	'تمس'	'مکش'	'ینس'	'ینس'	'پختی'	'ئتخا'	'ببعد'
اندازه قلم	۱۲	۱۰	۱۰	۱۲	۱۲	۱۲	۱۲	۱۲	۱۰	۱۲	۱۰	۱۲	۱۰
دقت روبش	۲۰۰	۳۰۰	۳۰۰	۲۰۰	۲۰۰	۲۰۰	۲۰۰	۲۰۰	۳۰۰	۲۰۰	۳۰۰	۲۰۰	۳۰۰
انتخاب صحیح	۲۱۷	۱۱۸	۳۴۳	۶۷۶	۹۹	۱۰۹۲	۵۲	۱۰۶	۱۱۹	۱۰۳	۲۰۶	۷۳	
زیرکلمه	'ببعد'	'بعلم'	'تعمد'	'تنفس'	'سپید'	'سبتر'	'سنیا'	'متعد'	'متعد'	'مهنا'	'هیما'	'یضخا'	
اندازه قلم	۱۲	۱۰	۱۲	۱۲	۱۲	۱۲	۱۲	۱۲	۱۲	۱۲	۱۰	۱۰	
دقت روبش	۲۰۰	۳۰۰	۲۰۰	۲۰۰	۲۰۰	۲۰۰	۲۰۰	۲۰۰	۲۰۰	۲۰۰	۳۰۰	۳۰۰	
انتخاب صحیح	۹۵	۱۹۴	۸۷۴۳	۵۶	۸۸	۵۵	۹۲	۸۳	۶۱	۲۳۳	۶۱	۶۸	
زیرکلمه	'سینین'	'عملیه'	'عملیه'	'قلمشا'	'ینصو'	'پیچید'	'بعجله'	'تخیلی'	'سیکلت'	'سینین'	'سینین'	'سینین'	
اندازه قلم	۱۴	۱۰	۱۲	۱۰	۱۰	۱۲	۱۲	۱۲	۱۰	۱۰	۱۲	۱۴	
دقت روبش	۳۰۰	۳۰۰	۲۰۰	۳۰۰	۳۰۰	۲۰۰	۲۰۰	۲۰۰	۳۰۰	۳۰۰	۳۰۰	۲۰۰	
انتخاب صحیح	۱۸۱	۴۴۹	۶۴	۹۹	۷۲	۱۴۶	۶۱	۹۷	۶۵	۲۵۳	۱۲۳	۴۰۰	
زیرکلمه	'مچگیر'	'میثمی'	'نخجیر'	'نستیم'	'نستیم'	'نستیم'	'نستیم'	'نصیبش'	'نصیبش'	'تضمینی'	'شبختیم'	'مینستر'	
اندازه قلم	۱۰	۱۰	۱۲	۱۰	۱۲	۱۲	۱۴	۱۲	۱۲	۱۰	۱۰	۱۲	
دقت روبش	۳۰۰	۳۰۰	۳۰۰	۳۰۰	۲۰۰	۳۰۰	۲۰۰	۲۰۰	۳۰۰	۳۰۰	۳۰۰	۲۰۰	
انتخاب صحیح	۶۱	۳۲۹	۵۹۶	۶۱	۱۷۶	۱۶۲	۱۲۰	۷۶	۹۱	۱۲۰	۲۴۰	۵۴	

۱۲۷۰۰ زیرکلمه با قلم لوتوس اندازه ۱۴ که با دقت ۴۰۰ نقطه بر اینچ روش شده‌اند بعنوان داده‌های آزمایش مورد استفاده قرار گرفتند که نتایج حاصل در جدول ۱۰ ارائه شده است.

همانطور که اطلاعات جدول ۱۰ نشان می‌دهد علیرغم عدم حضور داده‌های با قلم لوتوس در داده‌های آموزشی نتایج نسبتا خوبی حاصل شده است. برای بدست آوردن نتایج مشابه حالت تک قلم از نظر دقت میبایست فضای محدود شده را به ۲۰۰ زیرکلمه (چهار برابر حالت تک قلم) محدود نمود. در این حالت برای رسیدن به فضای محدودتر باید مرحله جدیدی به روش پیشنهادی اضافه نمود. به عنوان مثال با ترکیب روش پیشنهادی با قسمتهایی از روش ارائه شده در [۱۳] (انتخاب زیرکلماتی با کد موقعیت نقاط مشابه) امکان کاهش فضای جستجو به میزان قابل توجهی را خواهیم داشت که نمونه تاثیر این محدود کننده در مثال تشریح شده ارائه شده است.

در آزمایشی دیگر مجموعه پنج قلم مرحله قبل بعنوان داده‌های آموزش و ۵۰۰۰ داده آزمون معرفی شده در قسمت تک قلم بعنوان داده‌های آزمایش مورد استفاده قرار گرفتند که نتیجه آزمایشات در جدول ۱۱ ارائه شده است. خطای اضافه شده در نتیجه گسترش فضای آموزشی در این آزمایش نیز با اضافه کردن محدوده فضای انتخابی و اضافه کردن محدودکننده ساده در مراحل بعدی بازشناسی قابل جبران خواهد بود. با توجه به نتایج حاصل اگرچه درحالت مواجهه با چند قلم نتایج ضعیف‌تری را شاهدیم اما کارایی روش تعریف شده در عین سادگی آن در شرایط مختلف اثبات می‌گردد. قطعا با ارائه راه حل‌های تکمیلی در کنار روش پیشنهادی می‌توان نسبت به افزایش کارایی آن در مواجهه با قلم‌های مختلف اقدام نمود. در یک روش پیشنهادی در این مقاله بدین منظور از الگوریتم مرجع [۱۱] با تغییراتی استفاده شده است. در روش پیشنهادی زیرکلمه ورودی بجای مقایسه با مراکز خوشه‌ها و انتخاب نزدیکترین خوشه‌ها، با زیرکلمه‌های فضای محدود شده مقایسه شده و تعدادی از زیرکلمه‌های نزدیکتر به زیرکلمه ورودی به عنوان فضای محدود شده جدید انتخاب می‌شوند و در نتیجه فضای جستجوی محدودتری را شاهد خواهیم بود.

جدول ۱۰: نتایج با آموزش با چهار قلم و آزمایش با قلم باقیمانده

تعداد زیرکلمه در فضای جستجوی محدود شده	۵۰	۱۰۰	۱۵۰	۲۰۰
تعدادخطا				
در ۲۵۰۰ زیرکلمه اول	۱۰۸	۵۱	۳۰	۲۰

جدول ۱۱: نتایج با آموزش با پنج قلم و آزمایش با داده‌های آزمون

تعداد زیرکلمه در فضای جستجوی محدود شده	۵۰	۱۰۰	۱۵۰	۲۰۰
تعدادخطا				
در ۲۵۰۰ زیرکلمه اول	۸۱	۳۴	۲۲	۱۶

خلاصه روش بدین ترتیب است که ابتدا با روش پیشنهادی در این مقاله، فضای جستجو به ۱۵۰زیرکلمه محدود می‌شود (فضای جستجویی که به دقت قابل قبولی منجر می‌گردد) و پس از آن ۲۵۶ ویژگی مکان مشخصه زیرکلمه ورودی استخراج می‌گردد. تعداد این ویژگی برای هر زیرکلمه با الگوریتم PCA از ۲۵۶ به ۲۷ کاهش می‌یابد. با مقایسه ویژگی‌های ۲۷ گانه زیرکلمه ورودی و ویژگی‌های مشابه استخراج شده ۱۵۰ زیرکلمه فضای محدود شده قبلی، تعدادی از زیرکلمه‌ها که فاصله کمتری نسبت به زیرکلمه ورودی دارند به‌عنوان فضای محدود شده نهایی انتخاب می‌شوند که بر اساس آزمایشات صورت گرفته، تعداد مناسب زیرکلمات انتخابی با این روش (فضای محدود شده جدید) با حفظ دقت قبلی به ۵۵ زیرکلمه کاهش می‌یابد.

۴- جمع‌بندی

در این تحقیق یک روش جدید جهت کاهش فضای جستجوی زیرکلمات فارسی ارائه شده است. در این روش ابتدا ۱۰ویژگی ساده از زیرکلمه استخراج می‌شود. با استفاده از مفهوم کوانتیزاسیون و با توجه به بازه تغییرات هر ویژگی روی همه داده‌های آموزشی، ویژگی‌ها کوانتیزه شده و به اعداد صحیحی تبدیل می‌شوند. با استفاده از هر ویژگی و فاصله آن تا ویژگی متناظر هرکدام از نمونه‌های آموزشی، به هر کلاس امتیازی داده می‌شود. با اعمال همه ویژگی‌ها، هر کلاس به ازای هر ویژگی یک امتیاز دارد که با ترکیب این امتیازات با اعمال جبری یک امتیاز نهایی برای هر زیرکلمه بدست می‌آید که با مرتب کردن آنها و انتخاب تعدادی از آنها که امتیاز بیشتری دارند، فضای جستجو محدود می‌شود. جهت حصول بهترین نتیجه ممکن سطوح کوانتیزاسیون برای ویژگی‌های مختلف متفاوت در نظر گرفته شده و یافتن مقدار بهینه برای هر سطح به الگوریتم *PSO* سپرده شده است که منجر به نتیجه به مراتب بهتری از انتخاب یک سطح یکنواخت برای همه ویژگی‌ها شده است. در مرحله ترکیب امتیازات نیز برای هر یک از ویژگی‌ها وزنی در نظر گرفته شده و وزن‌های بهینه نیز توسط الگوریتم بهینه‌یابی *PSO* انتخاب شده است. در مجموع در این تحقیق با روال طی شده یک روش ساده و سریع با نتایج بهبود یافته نسبت به روش‌های پیشنهادی قبلی بدست آمده که مقایسه نتایج در جدول ۸ ارائه شده است. در انتها نیز کارایی روش پیشنهادی در حالت مواجه با چند قلم متفاوت بررسی شده و علاوه بر تاکید بر کارایی روش پیشنهادی در حالت مواجه با چند قلم، چند پیشنهاد تکمیلی جهت بهبود عملکرد در این حالت نیز ارائه و بررسی گردیده است.

مراجع

[1] S.A.A. Abbaszadeh Arani and E. Kabir and R. Ebrahimpour. "Combining right-to-left and left-to-right HMMs to recognize handwritten Farsi words of small- and medium-sized vocabularies" IET Computer Vision, Vol. 12, Issue 6. 2018

[2] N. Aouadi Aouadi and A.K. Echi. "Word extraction and recognition in arabic. handwritten Text" International Journal of Computing and Information Sciences, Vol. 12, No. 01, 2016.

[3] M. Shafii. "Optical character recognition of printed persian/arabic documents", Ph.D. dissertation, Windsor Univ., Ontario, Canada, 2014.

[4] S. Nasrollah and A. Ebrahimi. "Printed persian subword recognition using wavelet packet descriptors", Journal of Engineering (Hindawi Publishing Corporation), 2013.

[5] P.K. Powalka and N. Sherkat and R.J. Whitrow. "The use of word shape information for cursive script recognition" In Fourth International Workshop on Frontiers of Handwriting Recognition, pp. 67-76. 1994.

[6] S. Mozaffari and K. Faez and V. Märgner and H. Elabed. "Two-stage lexicon reduction for offline arabic handwritten word recognition" International Journal of Pattern Recognition and Artificial Intelligence, Vol. 22, No. 07: pp. 1323-1341, November 2008.

[7] H. Davoudi and M. Cheriet and E. Kabir. "lexicon reduction of handwritten arabic subwords based on the prominent shape regions" International Journal on Document Analysis and Recognition (IJDAR), Vol. 19, Issue 2, pp. 139–153, 2016.

[۸] سمیه برومند، ایرانپور مبارکه، مجید، "بازشناسی کلمات دست‌نوشته با ویژگی‌های نوین و کاهش فرهنگ لغت"، مجله پردازش بینایی و تصویر، آماده چاپ، ۱۳۹۶.

[۹] فائقه فتحی، "استخراج حروف شاخص از زیرکلمات چاپی فارسی"، پایان‌نامه کارشناسی ارشد، دانشگاه صنعتی سهند، تبریز، ایران، ۱۳۸۸.

[10] H. Davoudi and E. Kabir. "Using compatible shape descriptor for lexicon reduction of printed farsi subwordsjournal" International Journal on Document Analysis and Recognition (IJDAR), Vol. 19, Issue 2. pp. 139-153, 2016.

[۱۱] افشین ابراهیمی، احسان الله کبیر "یک روش دو مرحله‌ای برای بازشناسی زیرکلمات چاپی"، نشریه مهندسی برق و مهندسی کامپیوتر ایران، سال ۲، شماره ۲، ۱۳۸۳.

[۱۲] هما داودی، احسان الله کبیر "استفاده از مناطق شاخص زیرکلمات چاپی فارسی برای کاهش فضای جستجو در بازشناسی آنها"، نشریه مهندسی برق و مهندسی کامپیوتر ایران، ب –مهندسی کامپیوتر، سال ۱۲، شماره۱، ۱۳۹۳.

[۱۳] اسماعیل میری، سیدمحمد رضوی، ناصر مهرشاد، "روشی ساده برای کاهش فضای جستجو در بازشناسی زیرکلمات تایپی فارسی "، نهمین کنفرانس ماشین بینایی و پردازش تصویر ایران، دانشگاه شهید بهشتی، آبان ماه ۱۳۹۴.

[14] E. Miri and S.M. Razavi and N. Mehrshad. "Recognition of the persian typed sub-words with a hierarchical manner" Journal of Engineering and Applied Sciences, 12 (8): 2009-2017, 2017.

[15] A. Ebrahimi and E. Kabir. "A pictorial dictionary for printed farsi subwords" Pattern Recognition Letters, Vol. 29, pp. 656-663, 2008.

[16] J. Kennedy and R Eberhart. "Particle swarm optimization". In Fourth IEEE International Conference on Neural Networks, pp. 1942–1948, 1995. doi:10.1109/ICNN.1995.488968