علوم و مهندسی کامپیوتر
نشریه علمی پژوهشی انجمن کامپیوتر ایران
مجلد ۵، شماره ۳، ۱۳۸۶
صفحات ۲۷-۳۸
مقاله عادی

یک روش دو مرحله‌ای برای حل مسأله یادگیری چند نمونه‌ای

محمدرضا کیوان‌پور　　نصراله مقدم چرکری

دانشکده فنی و مهندسی، دانشگاه تربیت مدرس، تهران، ایران

چکیده

یادگیری چند نمونه‌ای مدل کلی‌تر یادگیری با نظارت است که در سال‌های اخیر در حوزه یادگیری ماشین مطرح گردیده است. این مدل قادر به یادگیری براساس مثال‌های آموزشی است که هر یک از آنها با چندین بردار ویژگی نمایش داده می‌شوند. در این مقاله ضمن معرفی و بررسی روش‌های مختلف ارائه شده برای حل مساله یادگیری چند نمونه‌ای، روشی دو مرحله‌ای برای حل این مساله ارائه می‌شود. روش پیشنهادی در دو مرحله مقدماتی و تکمیلی عمل می‌کند و از یک ابر مکعب در فضای ویژگی n بعدی برای نمایش و توصیف مفهوم مورد نظر بهره می‌برد. این روش دارای کاربرد عمومی است و نیازمند تنظیمات خاص وابسته به مثال‌های آموزشی نمی‌باشد. ساختار درونی الگوریتم‌های مورد استفاده در هر یک از مراحل روش پیشنهادی مستقل از یکدیگر می‌باشند. بنابراین روش پیشنهادی از انعطاف قابل توجهی جهت توسعه توانمندی‌هایش برخوردار است. براساس نتایج حاصل از آزمون‌های انجام شده بر روی مجموعه داده‌های دارویی Musk1 و Musk2، که محک‌هایی همه‌پذیر و شناخته شده در حوزه یادگیری چند نمونه‌ای تلقی می‌شوند، دقت روش پیشنهادی حدود ۹۰٪ می‌باشد. در این تحقیق روش پیشنهادی در حوزه طبقه‌بندی تصاویر نیز مورد استفاده و آزمون قرار گرفته است، براساس این آزمون روش پیشنهادی در قیاس با سایر روش‌های مطرح در این حوزه از دقت مناسبی برخوردار است.

کلمات کلیدی: یادگیری ماشین، یادگیری چند نمونه‌ای، مثال‌های آموزشی، فضای ویژگی.

۱- مقدمه

یادگیری چند نمونه‌ای[۱] مدل کلی‌تر یادگیری با نظارت[۲] است که در سال‌های اخیر توجه بسیاری از محققین را به خود جلب نموده است [۴] ، [۲۸] ، [۳۸]. در یادگیری چند نمونه‌ای مجموعه آموزشی[۳] مشتمل بر تعداد زیادی بسته[۴] است. هر بسته می‌تواند شامل تعداد متفاوتی از نمونه‌ها باشد و هر نمونه توسط یک بردار ویژگی نمایش داده می‌شود. بنابراین در این نوع یادگیری برخلاف یادگیری با نظارت، هر نمونه برچسب مستقل ندارد بلکه برچسب در سطح بسته موجود است. مطابق رویکرد رایج در روش‌های ارائه شده برای یادگیری چند نمونه‌ای یک بسته دارای برچسب منفی است اگر هیچیک از نمونه‌های موجود در آن به مفهوم هدف[۵] (مفهوم موردنظر جهت یادگیری) مرتبط نباشد. همچنین یک بسته دارای برچسب مثبت است اگر حداقل یکی از نمونه‌های موجود در آن با مفهوم هدف مرتبط باشد. به این ترتیب امکان وجود نمونه‌های نامرتبط با مفهوم هدف حتی در بسته‌های مثبت وجود دارد. بر این اساس چالش اساسی در مسأله یادگیری چند نمونه‌ای شناسایی نمونه‌های مطلوب در هریک از بسته‌های مثبت و تشخیص مفهومی است که توسط آنها توصیف می‌شود [۸]. هدف یادگیری چند نمونه‌ای کسب دانش در مورد مفهوم معرفی شده توسط بسته‌های برچسب زده شده می‌باشد به گونه‌ای که با بهره‌گیری از این دانش بتوان بسته‌های جدید را براساس نمونه‌های موجود در آنها برچسب زد [۳].

یادگیری چند نمونه‌ای می‌تواند در حوزه‌های کاربردی متعددی مورد استفاده قرار گیرد. یکی از این حوزه‌ها بازیابی تصویر مبتنی بر محتوا[۶] است [۲]. در این حوزه بازیابی تصاویر مورد نظر کاربر براساس مجموعه ای از تصاویر برچسب زده شده انجام می‌شود. کاربر در صورت وجود ناحیه (شیء) موردنظر خود در هر تصویر به آن برچسب مثبت منتسب می‌کند و در غیر این صورت به تصویر برچسب منفی می‌دهد. بنابراین با وجود آنکه کاربر تنها به ناحیه خاصی از تصویر نظر دارد برچسب مثبت یا منفی را به کل تصویر منتسب می‌نماید. بر این اساس چالش مهم فراروی سیستم‌های بازیابی تصویر شناسایی ناحیه مورد نظر کاربر (وجه مشترک تصاویر مثبت) بر مبنای بازخوردهای دریافتی از او می‌باشد. با تعریف هر ناحیه از

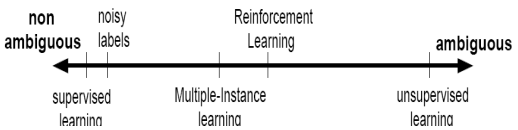
تصویر به عنوان یک نمونه و کل تصویر به عنوان یک بسته می‌توان از یادگیری چند نمونه‌ای به عنوان راه‌حلی مناسب برای رفع چالش مذکور استفاده کرد. دسته‌بندی متون [۱] و پیش‌بینی تأثیر داروها در حوزه داروسازی [۳] را می‌توان به عنوان دیگر حوزه‌های کاربردی یادگیری چند نمونه‌ای نام برد.

در این تحقیق روشی دو مرحله‌ای برای حل مساله یادگیری چند نمونه‌ای ارائه می‌شود. در مرحله مقدماتی عملیات شناسایی نمونه‌های مطلوب در بسته‌های مثبت و نمایش مفهوم موردنظر در قالب یک ابر مکعب[۷] در فضای ویژگی n بعدی انجام می‌شود. مرحله تکمیلی مشتمل بر توسعه روشمند ابرمکعب پیش گفته می‌باشد. روش پیشنهادی دارای کاربرد عمومی است و برخلاف برخی روش‌های مطرح گذشته نیازمند تنظیمات خاص وابسته به مثال‌های آموزشی نمی‌باشد. براساس آزمون‌های انجام شده بر روی بر روی مجموعه داده‌های دارویی Musk1 و Musk2، که محک‌های[۸] همه‌پذیر و شناخته شده در حوزه یادگیری چند نمونه‌ای تلقی می‌شوند، روش پیشنهادی در قیاس با سایر روش‌های ارائه شده در این حوزه از دقت مناسبی برخوردار است. همچنین در این تحقیق روش پیشنهادی در حوزه طبقه‌بندی تصاویر نیز مورد استفاده و آزمون قرار گرفته است، براساس این آزمون روش پیشنهادی در حوزه طبقه‌بندی تصاویر در قیاس با سایر روش‌های مطرح در این حوزه از دقت مناسبی برخوردار است. این روش با افزایش تعداد بسته‌های آموزشی و تعداد و تنوع نمونه‌های موجود در آنها با کاهش دقت مواجه نمی‌شود و قادر به حفظ دقت خود در حد قابل قبول می‌باشد. استقلال نسبی مراحل این روش و ظرفیت قابل توجه آن در بهره‌گیری از الگوریتم‌های جدید آتی در هر یک از این مراحل به عنوان دیگر مزیت مهم روش پیشنهادی به شمار می‌رود.

این مقاله در شش بخش تدوین شده است. در بخش دوم پیشینه تحقیق در این زمینه معرفی می‌شود. بخش سوم دربرگیرنده تعریف دقیق مساله یادگیری چند نمونه‌ای می‌باشد. در بخش چهارم روش پیشنهادی برای حل مساله یادگیری چند نمونه‌ای معرفی می‌شود. نتایج حاصل از پیاده‌سازی و آزمون این روش در بخش پنجم ارائه می‌گردد. نتیجه‌گیری و توسعه‌های آتی این تحقیق در بخش ششم معرفی می‌شود.

۲- پیشینه تحقیق

جایگاه یادگیری چند نمونه‌ای نسبت به سایر روش‌های یادگیری از منظر میزان ابهام در مثال‌های آموزشی مورد استفاده در فرایند یادگیری به‌صورت شکل ۱ قابل نمایش است. چنانکه در شکل ملاحظه می‌شود می‌توان یادگیری چند نمونه‌ای را نوعی یادگیری شبه نظارت شده[۹] تلقی کرد [۴]. زیرا در این نوع یادگیری برچسب فقط در سطح بسته وجود دارد و در سطح نمونه برچسبی موجود نمی‌باشد. این امر موجب افزایش ابهام در فرایند یادگیری می‌شود [۳۸].



شکل ۱- جایگاه یادگیری چند نمونه‌ای نسبت به سایر روش‌های یادگیری از منظر میزان ابهام در مثال‌های آموزشی [۲۹]

مساله یادگیری چند نمونه‌ای نخستین بار توسط Dietterich و همکارانش در حوزه داروسازی معرفی و استفاده شد [۵]. این گروه در روشی به نام Iterated-discrim از ابرمکعب‌ها با ابعاد موازی محورهای مختصات[۱۰] برای یادگیری چند نمونه‌ای بهره برده‌اند. هزینه پردازشی زیاد و لزوم انجام تنظیمات خاص[۱۱] وابسته به مثال‌های آموزشی به عنوان مهمترین نارسائی‌های این روش قابل ذکر می‌باشند [۶]. در [۷] روشی تحت عنوان MULTINST برای یادگیری چند نمونه‌ای ارائه شده است. این روش نیز مفهوم موردنظر را در قالب یک ابرمکعب نمایش می‌دهد. اما در این روش بدون نیاز به تنظیمات خاص وابسته به مثال‌های آموزشی و تنها با استفاده از برخی محاسبات آماری، عملیات یادگیری با پیچیدگی کمتر نسبت به روش گذشته انجام می‌گیرد. روش DD[۱۲] [۸] یکی دیگر از روش‌های مطرح در این زمینه است. ایده اصلی این روش شناسایی نقطه‌ای در فضای ویژگی است که هر یک از بسته‌های مثبت حداقل یک نمونه نزدیک آن داشته باشند و هیچ بسته منفی نمونه‌ای نزدیک آن نداشته باشد. روش EM-DD[۱۳] به عنوان یکی دیگر از روش‌های مهم در این حوزه مطرح می‌باشد، این روش شکل توسعه یافته روش DD به شمار می‌رود [۶].

روش‌های Citation-KNN و Bayesian-KNN از دیگر روش‌های مهم مطرح شده در این زمینه هستند [۹]. در این روش‌ها با استفاده از Hausdroff Distance برای محاسبه فواصل میان بسته‌ها در فضای ویژگی و همچنین بهره‌گیری از انواع خاصی از همسایگی‌ها راه حل مناسبی با دقت بالا برای مساله یادگیری چند نمونه‌ای ارائه می‌شود. همچنین mi-SVM و MI-SVM [۱۰] و روش ارائه شده در [۱۱] روش‌هایی مبتنی بر ماشین بردار پشتیبان[۱۴] برای حل مساله یادگیری چند نمونه‌ای می‌باشند.

در روش MI-NN [۱۲] و برخی پژوهش‌های دیگر [۲۷] ، [۲۸] برای انجام یادگیری چند نمونه‌ای از شبکه عصبی استفاده شده است. در یکی از جدیدترین پژوهش‌های انجام شده در این زمینه تلاش شده است تا با گزینش یکی از نمونه‌های هر بسته به عنوان نماینده آن بسته بستر مناسبی برای استفاده از روش‌های یادگیری با نظارت جهت یادگیری چند نمونه‌ای فراهم شود[۴]. همچنین در [۱۳] ، [۱۴] ، [۱۵] ، [۱۶] ، [۲۱] ، [۳۸] و [۴۰] نیز روش‌های دیگری برای حل این مساله ارائه شده است. توجه به این نکته ضروری است که بخش عمده‌ای از روش‌های ارائه شده در این زمینه متکی بر فرضیه وجود تک نقطه ایده‌آل در فضای ویژگی به عنوان مفهوم هدف هستند. بنابراین چنانچه مفهوم هدف مشتمل بر چندین نقطه در فضای ویژگی باشد روش‌های مذکور قادر به یادگیری آن نمی‌باشند [۱۷].

یادگیری چند نمونه‌ای عمومیت یافته[۱۵] نامی است که محققین به این مساله داده‌اند. این مساله به عنوان یکی از محورهای پژوهشی مطرح در این زمینه مورد توجه پژوهشگران قرار گرفته است [۱۱] ، [۱۸] ، [۱۹] ، [۲۰]. محور پژوهشی مهم دیگری که در این زمینه مطرح است چگونگی بکارگیری یادگیری چند نمونه‌ای در حوزه‌های کاربردی متنوع می‌باشد، چنانکه پیش از این ذکر شد بازیابی تصویر [۲] ، [۲۲] ، [۲۴] ، [۲۵] ، [۳۹] و دسته‌بندی و بازیابی متون [۱] ، [۶] ، [۱۰] به عنوان مهمترین حوزه‌های کاربردی استفاده کننده از یادگیری چند نمونه‌ای تلقی می‌شوند.

۳- تعریف مساله

در یادگیری چند نمونه‌ای هر مثال آموزشی بسته‌ای برچسب زده شده مشتمل بر چندین نمونه است. در این روش برچسب در سطح بسته‌ها وجود دارد و در سطح نمونه‌ها برچسبی ارائه نمی‌شود اما برچسب هر بسته به گونه‌ای وابسته به نمونه‌های موجود در آن است. نحوه وابستگی برچسب هر بسته به نمونه‌های آن فرض پایه‌ای چند نمونه‌ای[۱۶] نامیده می‌شود. در مدل رایج یادگیری چند نمونه‌ای فرض پایه‌ای مذکور به صورت زیر مطرح می‌شود [۸]:

- یک بسته به عنوان بسته مثبت تلقی می‌شود اگر و فقط اگر حداقل یکی از نمونه‌های موجود در آن مثبت باشد.
- یک بسته به عنوان بسته منفی تلقی می‌شود اگر و فقط اگر کلیه نمونه‌های موجود در آن منفی باشند.

بر این اساس مساله یادگیری چند نمونه‌ای عبارت است از ایجاد فرضیه‌ای مانند h_b که وظیفه نگاشت یک بسته جدید به یک برچسب (مثبت یا منفی) را براساس نمونه‌های آن دارد. در تعریف ۱ مساله یادگیری چند نمونه‌ای به‌صورت رسمی معرفی گردیده است. همچنین فرض پایه‌ای مدل رایج یادگیری چند نمونه‌ای در تعریف ۲ ارائه شده است.

تعریف ۱. اگر I فضای نمونه‌ها، $B=2^I$ فضای بسته‌ها[۱۷]، $\Phi = \{T, F\}$ مجموعه برچسب‌ها و $E=<Bag,Label>$ مجموعه مثال‌های آموزشی باشد به صورتی که $Bag = \{Bag_i | Bag_i \in B, i = 1,...,U\}$ مجموعه‌ای از U بسته و $Label = \{Label_i | Label_i \in \Phi, i = 1,...,U\}$ مجموعه برچسب‌های منتسب به بسته‌ها باشد، آنگاه مساله یادگیری چند نمونه‌ای عبارت است از ایجاد یک فرضیه به‌صورت $h_b : B \rightarrow \Phi$ که می‌تواند برچسب *بسته‌های ناشناخته را تعیین نماید.*

تعریف ۲. براساس فرض پایه‌ای مدل رایج یادگیری چند نمونه‌ای، مفهوم هدف معادل تک نقطه ایده‌آل در فضای ویژگی است به گونه‌ای که هر بسته مثبت حداقل یک نمونه نزدیک آن نقطه دارد و هیچ بسته منفی نمونه‌ای نزدیک آن ندارد.

نقطه ایده‌آل در فضای ویژگی، نقطه‌ای است که معرف مفهوم مورد نظر جهت یادگیری است و هر بسته مثبت دارای حداقل یک نمونه نزدیک آن است و هیچ بسته منفی نمونه‌ای نزدیک آن ندارد. اگر مفهوم مورد نظر از پیچیدگی بیشتری برخوردار باشد و چندین نقطه را در فضای ویژگی در برگیرد آنگاه فرض پایه‌ای پیش گفته و قسمت‌هایی از تعاریف ۱ و ۲ باید اصلاح شود. یادگیری چند نمونه‌ای عمومیت یافته سعی در پوشش چنین مواردی با توسعه فرض پایه‌ای دارد [۱۷].

روش پیشنهادی در این مقاله مبتنی بر فرض پایه‌ای مدل رایج یادگیری چند نمونه‌ای یعنی وجود تک نقطه ایده آل در فضای ویژگی که در تعریف ۲ ارائه گردید می‌باشد. اما ساختار درونی این روش به گونه‌ای است که از قابلیت و انعطاف لازم جهت توسعه یافتن و پوشش برخی انواع یادگیری چند نمونه‌ای عمومیت یافته برخوردار است.

۴- روش پیشنهادی

روش پیشنهادی در این پژوهش برای حل مساله یادگیری چند نمونه‌ای مشتمل بر دو مرحله مقدماتی و تکمیلی است. مرحله مقدماتی شامل شناسایی نمونه‌های مطلوب از بسته‌های مثبت و نمایش مفهوم موردنظر در قالب یک ابرمکعب در فضای ویژگی n بعدی می‌باشد.
در مرحله تکمیلی عملیات توسعه روشمند ابرمکعب پیش گفته انجام می‌شود. الگوریتم‌های مورد استفاده در هر یک از این مراحل مستقل از یکدیگر می‌باشد. در این بخش ضمن معرفی هر یک از این دو مرحله، چگونگی بهره‌گیری از آنها در قالب یک مدل مناسب برای حل مساله یادگیری چند نمونه‌ای بیان می‌شود.

۴-۱- مرحله مقدماتی

نخستین مرحله روش پیشنهادی، که از اهمیت بیشتری نیز برخوردار است، نوعی یادگیری اولیه را به انجام می‌رساند و مشتمل بر دو زیر مرحله است: ۱) الگوریتم شناسایی نمونه‌های مطلوب ۲) ایجاد یک ابرمکعب اولیه برای نمایش مفهوم مورد نظر. در ادامه نحوه کار هر یک از این دو بخش بیان می‌شود.

۴-۱-۱- الگوریتم شناسایی نمونه‌های مطلوب

براساس فرض پایه‌ای مدل رایج یادگیری چند نمونه‌ای مفهوم مورد نظر برای یادگیری معادل تک نقطه ایده‌آل در فضای ویژگی است و یک بسته مثبت است اگر و فقط اگر حداقل یکی از نمونه‌های موجود در آن نزدیک نقطه مذکور باشد [۵]. در این مقاله نمونه‌ای از یک بسته مثبت که موجب مثبت شدن آن بسته شده است نمونه مطلوب نامیده می‌شود. به جز نمونه مطلوب بسیاری از نمونه‌های موجود در هر بسته مثبت نمونه‌های زائد و نوعی نویز محسوب می‌شوند. به عنوان مثال شکل ۲ سه بسته مثبت و دو بسته منفی را به همراه نمونه‌های موجود در آنها در فضای ویژگی دو بعدی نشان می‌دهد، در این مثال مفهوم هدف معادل نقطه Cp=(2,5) در فضای ویژگی می‌باشد و نمونه‌های مطلوب در هر بسته مثبت به‌صورت پر رنگ نشان داده شده است.

$B_1^+ = \{(17,6),(3.8,19),(5,14),\mathbf{(1.8,5.2)},(27,8)\}$

$B_2^+ = \{(5,10),\mathbf{(2,4.8)},(12,7)\}$

$B_3^+ = \{\mathbf{(2.2,4.9)},(7,9),(5.8,11.5),(16,6),(25,5),(14,5)\}$

$B_4^- = \{(20,9),(3,12),(18,13),(18,5),(22,10)\}$

$B_5^- = \{(15,10),(8,7),(23,4),(11,11)\}$

شکل ۲- چند بسته مثبت و نمونه‌های مطلوب موجود در آنها

با شناسایی نمونه‌های مطلوب موجود در هر یک از بسته‌های مثبت و استفاده مناسب از آنها می‌توان مفهوم مورد نظر را یادگیری نمود. بر این اساس شناسایی نمونه‌های مطلوب از بسته‌های مثبت به عنوان یک ضرورت برای حل مساله یادگیری چند نمونه‌ای مطرح است [۶].

الگوریتم پیشنهادی در این بخش برای شناسایی نمونه‌های مطلوب مبتنی بر خوشه‌بندی[۱۸] نمونه‌های موجود در کلیه بسته‌های مثبت است. از آنجا که تمام نمونه‌های مطلوب در اطراف نقطه ایده‌آل در فضای ویژگی قرار گرفته‌اند بنابراین خوشه‌بندی مذکور، اگر به درستی انجام شود، موجب تجمیع نمونه‌های مطلوب در یکی از خوشه‌ها (خوشه مطلوب) خواهد شد. به این ترتیب با تشخیص خوشه مطلوب می‌توان نمونه‌های مطلوب را شناسایی نمود. در الگوریتم پیشنهادی در این بخش برای خوشه‌بندی نمونه‌های موجود در بسته‌های مثبت از روش K-Median [۲۶] استفاده می‌شود. این روش شکل تغییر یافته روش K-Means [۳۲] ، [۳۳] می‌باشد.

در روش K-Means هدف گروه‌بندی کلیه نمونه‌ها در k خوشه است به گونه‌ای که نمونه‌های موجود در هر خوشه حداقل پراکندگی و حداکثر تجانس[۱۹] را دارا باشند. در این روش برای حصول هدف پیش گفته باید مقدار H در رابطه (۱) مینیمم شود [۳۲].

$$H = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \qquad (۱)$$

در رابطه (۱) k نشان دهنده تعداد خوشه‌ها، C_i نشانگر خوشه iام ، p معرف هر یک از نمونه‌های موجود در خوشه C_i و m_i بیانگر مرکز خوشه C_i است. در روش K-Means مرکز هر خوشه یک بردار n بعدی است که مقدار آن در هر بعد برابر میانگین کلیه نمونه‌های خوشه در آن بعد می‌باشد. مراحل کار در روش K-Means را می‌توان به‌صورت زیر بیان کرد [۳۳]:

۱- انتخاب k نمونه اولیه به عنوان مراکز خوشه‌های اولیه
۲- انتساب هر یک از نمونه‌ها به نزدیک ترین خوشه
۳- محاسبه مجدد مرکز هر یک از خوشه‌ها (محاسبه میانگین کلیه نمونه‌های منتسب شده به هر یک از خوشه‌ها)
۴- تکرار مراحل ۲ و ۳ تا هنگامی که مراکز خوشه‌ها تثبیت شده و دیگر تغییر نکند

برای محاسبه فواصل در این روش از فاصله اقلیدسی استفاده می‌شود. روش K-Means نسبت به نقاط مرزی[۲۰] حساس است، برای کاهش تأثیرات منفی این نقاط می‌توان از روش K-Median استفاده کرد. روش K-Median شکل تغییر یافته روش K-Means می‌باشد. در این روش برای محاسبه مجدد مرکز خوشه‌ها در مرحله سوم الگوریتم از میانه[۲۱] به جای میانگین استفاده می‌شود. همچنین در این روش برای محاسبه فواصل به جای فاصله اقلیدسی از رابطه (۲) استفاده می‌شود [۲۶].

$$Dis(x,y)=\sum_{i=1}^{n}\left|y_i-x_i\right| \qquad (۲)$$

نتایج بدست آمده از روش K-Median وابستگی زیادی به مقدار k و همچنین k نمونه اولیه انتخاب شده به عنوان مراکز خوشه‌های اولیه دارد [۳۴]. بر این اساس در روش پیشنهادی با انتخاب بسته پایه[۲۲] بستر مناسبی برای تعیین مقدار k و انتخاب k نمونه اولیه به عنوان مراکز خوشه‌های اولیه فراهم می‌شود. براساس تعریف ۲ حداقل یکی از نمونه‌های هر بسته مثبت نزدیک نقطه ایده‌آل در فضای ویژگی قرار دارد. بدیهی است هر چقدر تعداد نمونه‌های موجود در یک بسته مثبت کمتر باشد تعداد گزینه‌های مطرح در آن بسته جهت انتخاب نمونه مطلوب محدودتر خواهد شد. بر همین اساس در روش پیشنهادی ابتدا بسته پایه انتخاب می‌شود و سپس بسته پایه مبنای خوشه‌بندی به روش پیش گفته قرار می‌گیرد. برای انتخاب بسته پایه با هدف مذکور ابتدا لازم است کلیه بسته‌های مثبت کمینه[۲۳] شناسایی شوند.

تعریف ۳. B_m^+ یک بسته مثبت کمینه است اگر و فقط اگر یک بسته مثبت بوده و کمترین تعداد نمونه‌ها را دارا باشد.

اگر در بین بسته‌های مثبت تنها یک بسته مثبت کمینه وجود داشته باشد همان بسته به عنوان بسته پایه (B_b^+) انتخاب می‌شود و چنانچه بیش از یک بسته مثبت کمینه وجود داشته باشد یکی از آنها به‌صورت تصادفی به عنوان بسته پایه انتخاب می‌شود. به عنوان مثال در شکل ۲ بسته دوم (B_2^+) به عنوان بسته پایه انتخاب می‌شود.

چون بسته پایه یک بسته مثبت است بنابراین طبق تعریف ۲ حداقل یکی از نمونه‌های آن نزدیک نقطه ایده‌آل در فضای ویژگی قرار دارد. به همین دلیل می‌توان نمونه‌های موجود در بسته پایه را به عنوان مراکز k خوشه اولیه در نظر گرفت و خوشه‌بندی به روش پیش گفته را انجام داد. به عنوان مثال اگر در شکل ۲ بسته دوم (B_2^+) به‌عنوان بسته پایه مبنای خوشه‌بندی نمونه‌ها قرار گیرد نتایج حاصل از خوشه‌بندی نمونه‌های موجود در بسته‌های مثبت به‌صورت شکل ۳ قابل نمایش است. چنانکه در این شکل ملاحظه می‌شود کلیه نمونه‌هایی که در شکل ۲ بصورت پر رنگ نمایش داده شده بودند، در خوشه C_2 و سایر نمونه‌ها در دو خوشه دیگر خوشه‌بندی شده‌اند.

C_1 ={(5,10),(3.8,19),(5,14),(7,9),(5.8,11.5)}

C_2 ={**(2,4.8),(1.8,5.2), (2.2,4.9)**}

C_3 = {(12,7),(17,6),(16,6),(25,5),(14,5),(27,8)}

شکل ۳- خوشه‌های بدست آمده از خوشه‌بندی نمونه‌های موجود در بسته‌های مثبت

پس از خوشه‌بندی نمونه‌های موجود در بسته‌های مثبت لازم است خوشه مطلوب شناسایی شود. مطابق فرض پایه‌ای مدل رایج یادگیری چند نمونه‌ای، هر بسته مثبت حداقل یک نمونه مطلوب دارد. بر طبق همین فرض نمونه‌های مطلوب موجود در بسته‌های مثبت در اطراف نقطه ایده‌آل در فضای ویژگی قرار دارند ولی سایر نمونه‌ها از توزیع مشخصی برخوردار نیستند و در قسمت‌های مختلف فضای ویژگی پراکنده‌اند [۸]. همچنین بر طبق تعریف ۲ هیچ نمونه منفی[۲۴] در نزدیک نقطه ایده‌آل در فضای ویژگی قرار ندارد. بر این اساس می‌توان دریافت که عوامل مؤثر بر میزان مطلوبیت یک خوشه عبارتند از: ۱) میزان تجانس نمونه‌های موجود در آن خوشه ۲) میزان فاصله نمونه‌های منفی با مرکز خوشه ۳) میزان تنوع محتوایی خوشه.

تعریف ۴. تنوع محتوایی یک خوشه برابر است با تعداد بسته‌های مثبت که نمونه‌ای از آنها در آن خوشه وجود دارد.

از آنجا که هر بسته مثبت حداقل یک نمونه مطلوب دارد بنابراین بدیهی است که اکثریت نسبی بسته‌های مثبت باید نمونه‌ای در خوشه مطلوب داشته باشند. بر این اساس تنها بخشی از خوشه‌های ایجاد شده به عنوان کاندید برای انتخاب خوشه مطلوب مطرح می‌باشند.

تعریف ۵. اگر C یک خوشه، V_C تنوع محتوایی خوشه C و B^+ تعداد کل بسته‌های مثبت باشد آنگاه C یک خوشه کاندید است اگر $V_C > \frac{B^+}{2}$، بر طبق تعریق ۵ باید بیش از نیمی از بسته‌های مثبت نمونه‌ای در یک خوشه داشته باشند تا آن خوشه به عنوان یک کاندید برای انتخاب خوشه مطلوب مطرح شود. براساس آنچه پیش از این ذکر شد تعریف زیر برای خوشه مطلوب ارائه می‌شود.

تعریف ۶. خوشه مطلوب خوشه‌ای است که حداکثر تجانس نمونه‌های عضو، بیشترین میزان فاصله با نمونه‌های منفی، و بالاترین تنوع محتوایی را در میان خوشه‌های کاندید دارا باشد.

در روش پیشنهادی ابتدا خوشه‌های کاندید مطابق تعریف ۵ شناسایی می‌شوند، سپس ارزش هر خوشه کاندید مانند C بر طبق رابطه (۳) محاسبه می‌شود. پس از آن خوشه کاندیدی که بیشترین ارزش را دارا است به عنوان خوشه مطلوب شناسایی می‌شود.

$$R_C=\frac{1}{\sum_{j=1}^{w} Dis(p_j,m_C)}+\sum_{h=1}^{f} Dis(I_h^-,m_C)+V_C \qquad (۳)$$

در رابطه (۳) مؤلفه اول بیانگر میزان تجانس نمونه‌های موجود در خوشه C، مؤلفه دوم نشان دهنده مجموع فواصل نمونه‌های منفی با مرکز خوشه C و مؤلفه سوم معرف تنوع محتوایی خوشه C می‌باشد. در این رابطه m_C مرکز خوشه C، w تعداد نمونه‌های موجود در خوشه C، p_j نشانگر هر یک از نمونه‌های موجود در خوشه C، f تعداد کل نمونه‌های منفی، I_h^- معرف هر یک از نمونه‌های منفی و Dis بیانگر فاصله میان نقاط در فضای ویژگی و V_C نشانگر میزان تنوع محتوایی خوشه C می‌باشد. پس از محاسبه ارزش هریک از خوشه‌های کاندید براساس رابطه (۳)، این خوشه‌ها برطبق ارزش محاسبه شده برای آنها رتبه‌بندی می‌شوند و خوشه کاندیدی که از بالاترین ارزش برخوردار است به عنوان خوشه مطلوب تشخیص داده می‌شود و نمونه‌های موجود در آن به عنوان نمونه‌های مطلوب تلقی می‌شوند. بنابراین در مراحل بعد این نمونه‌های گزینش شده به نمایندگی از بسته‌های مثبت در فرایند یادگیری مورد استفاده قرار می‌گیرند. نمونه‌های مطلوب گزینش شده در این مرحله مجموعه I^+ را تشکیل می‌دهند. همچنین در پایان این مرحله کلیه نمونه‌های موجود در سایر خوشه‌ها به همراه تمامی نمونه‌های موجود در بسته‌های منفی مجموعه I^- را به عنوان مجموعه نمونه‌های منفی تشکیل می‌دهند.

۴-۱-۲- ایجاد یک ابرمکعب اولیه جهت نمایش مفهوم موردنظر

یکی از روش‌های نمایش مفهوم یادگیری شده توسط مدل‌های یادگیرنده نمایش آن در قالب یک ابرمکعب با ابعاد موازی محورهای مختصات در فضای ویژگی می‌باشد. با استفاده از یک ابرمکعب برای نمایش مفهوم یادگیری شده امکان پوشش مناسب نمونه‌های مطلوب متناسب با نحوه توزیع آنها در فضای ویژگی فراهم می‌شود[۵] ، [۷]. بنابراین در این بخش از مرحله اول روش پیشنهادی بر طبق نمونه‌های گزینش شده در بخش قبل (I^+) یک ابرمکعب اولیه برای نمایش مفهوم یادگیری شده ایجاد می‌شود. ابعاد این ابرمکعب موازی با محورهای مختصات در فضای ویژگی n بعدی است. اندازه این ابرمکعب در هر یک از ابعاد فضای ویژگی براساس کمترین و بیشترین مقدار نمونه‌های مطلوب در آن بعد محاسبه می‌شود به گونه‌ای که کلیه نمونه‌های مطلوب شناسایی شده در آن واقع شوند. رابطه (۴) نحوه محاسبه ابتدا (BHC_{1h}) و انتهای (BHC_{2h}) ابرمکعب در بعد h را نشان می‌دهد. در این رابطه I_{jh} مقدار نمونه jام در بعد h را بیان می‌کند ($I_j \in I^+$) همچنین در این رابطه n معرف ابعاد فضای ویژگی و r نشانگر تعداد کل نمونه‌های مطلوب شناسایی شده است.

$$BHC_{2h} = \mathrm{Max}\,(I_{jh})$$
$$BHC_{1h} = \mathrm{Min}\,(I_{jh}) \qquad (۴)$$
$$1 \le h \le n \quad 1 \le j \le r$$

با توجه به رابطه (۴) می‌توان دریافت که ابرمکعب فوق کلیه نمونه‌های مطلوب شناسایی شده را در بر می‌گیرد. در روش پیشنهادی این ابرمکعب به عنوان *ابرمکعب پایه*[۲۵] (BHC) نامیده می‌شود.

تعریف۷. ابرمکعب پایه کوچکترین ابرمکعب با ابعاد موازی با محورهای مختصات در فضای ویژگی می‌باشد که همه بسته‌های مثبت حداقل یک نمونه در آن دارند و هیچ بسته منفی نمونه‌ای در آن ندارد.

بنابراین طبق تعریف یک بسته مثبت است اگر و فقط اگر حداقل یک نمونه در داخل ابرمکعب پایه داشته باشد.

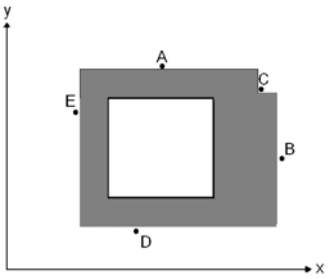
۴-۲- مرحله تکمیلی

مطابق رابطه (۴) ابرمکعب پایه کوچکترین ابرمکعبی است که کلیه نمونه‌های مطلوب شناسایی شده در مرحله مقدماتی را شامل می‌شود و انتظار می‌رود که توصیفگر مناسبی برای مفهوم مورد نظر باشد، اما با تعریف مفهوم *ناحیه بحرانی* اطراف ابرمکعب پایه و توجه به این مفهوم مشخص می‌شود که ابرمکعب مذکور از جامعیت لازم برای توصیف مفهوم هدف برخوردار نیست. بر این اساس توسعه روشمند ابرمکعب پایه به عنوان یک الزام برای روش پیشنهادی مطرح می‌شود.

تعریف۸. اگر I^- معرف مجموعه نمونه‌های منفی و BHC معرف ابرمکعب پایه باشد و $\exists \beta, BHC \subset \beta, \forall I_j^- \in I^-, I_j^- \not\subset \beta$ آنگاه ناحیه بحرانی ابر مکعب BHC بصورت زیر تعریف می‌شود

$$\lambda = (\beta - BHC)$$

مطابق تعریف ۸ اگر β زیر فضایی از فضای ویژگی باشد که اولا ابرمکعب پایه را در بر گیرد و ثانیا هیچ نمونه منفی در آن وجود نداشته باشد آنگاه ناحیه بحرانی به‌صورت $\beta - BHC$ تعریف می‌شود، در واقع ناحیه بحرانی زیرفضای میان ابرمکعب پایه و نزدیکترین نمونه‌های منفی اطراف آن را در بر می‌گیرد و هیچ نمونه مثبت یا منفی از مثال‌های آموزشی در این ناحیه قرار نگرفته است. به عنوان مثال شکل ۴ یک مستطیل (تصویر ابرمکعب پایه در فضای دو بعدی) و پنج نمونه منفی اطراف آن را نشان می‌دهد. مطابق تعریف ۸ قسمت خاکستری رنگ در این شکل ناحیه بحرانی اطراف مستطیل را نشان می‌دهد. چنانکه در شکل ملاحظه می‌شود ناحیه بحرانی فضای میان ابرمکعب پایه و نزدیکترین نمونه‌های منفی اطراف آن را در بر می‌گیرد.



شکل ۴- یک مستطیل در فضای دو بعدی و ناحیه بحرانی اطراف آن

براساس تعریف ۸ می‌توان دریافت که طبق بسته‌های آموزشی[۲۶] در ناحیه بحرانی اطراف ابرمکعب پایه هیچ نمونه مطلوب یا نمونه منفی وجود ندارد. بنابراین تعلق این ناحیه به قلمرو نمونه‌های مطلوب یا نمونه‌های منفی ابهامی جدی را فراروی روش پیشنهادی قرار می‌دهد، به همین دلیل ناحیه مذکور در این تحقیق ناحیه بحرانی نامیده شده است. اگر این ناحیه قلمرو نمونه‌های منفی تلقی شود، یعنی مفهوم مورد نظر در قالب ابرمکعب پایه، بدون در نظر گرفتن ناحیه بحرانی، نمایش داده شود آنگاه ممکن است در مورد برخی بسته‌های آزمایشی[۲۷] که باید مثبت تشخیص داده شوند، نمونه مطلوب مربوطه در ناحیه بحرانی واقع شود و در نتیجه این بسته‌ها به اشتباه منفی تشخیص داده شوند، در این صورت نرخ FN[۲۸] افزایش می‌یابد. اما اگر ناحیه بحرانی، قلمرو نمونه‌های مثبت در نظر گرفته شود و به ابرمکعب پایه افزوده گردد آنگاه ممکن است برخی بسته‌های آزمایشی که باید منفی تشخیص داده شوند، دارای نمونه‌ای در ابرمکعب پایه شوند و به اشتباه مثبت

قلمداد شوند، در این حالت نرخ FP[۲۹] افزایش خواهد یافت. بر این اساس تشخیص آن بخش از ناحیه بحرانی که به قلمرو نمونه‌های مطلوب تعلق دارد و افزودن آن به ابرمکعب پایه موجب افزایش دقت روش پیشنهادی خواهد شد، به همین دلیل توسعه روشمند ابرمکعب پایه در ناحیه بحرانی اطراف آن به‌صورت یک الزام برای روش پیشنهادی مطرح است. روش پیشنهادی توسعه مذکور را در مرحله تکمیلی انجام می‌دهد. مرحله تکمیلی روش پیشنهادی مشتمل بر دو بخش مجزا است: ۱) رتبه‌بندی ابعاد فضای ویژگی براساس میزان اهمیت آنها ۲) توسعه روشمند ابرمکعب پایه براساس رتبه‌بندی ابعاد فضای ویژگی.

۴-۲-۱- رتبه‌بندی ابعاد فضای ویژگی

توسعه روشمند ابرمکعب پایه مستلزم توجه دقیق به اهمیت هر یک از ابعاد فضای ویژگی است. بنابراین لازم است پیش از انجام توسعه مذکور کلیه ابعاد فضای ویژگی براساس میزان اهمیت آنها رتبه‌بندی شوند. ذکر این نکته ضروری است که هدف این بخش از روش پیشنهادی انجام نوعی انتخاب ویژگی[۳۰] یا کاهش بعد[۳۱] با استفاده از روش‌های پیچیده‌ای نظیر آنچه در [۳۰] و [۳۱] برای انجام طبقه‌بندی[۳۲] نمونه‌ها ارائه می‌شود نیست. بلکه تنها هدف رتبه‌بندی پیشنهادی آن است که با روشی ساده مبنایی مناسب برای توسعه روشمند ابرمکعب پایه در هر یک از ابعاد فضای ویژگی فراهم آورد. در این بخش کلیه نمونه‌های مثبت (I^+) و منفی (I^-) نقش ورودی را ایفا می‌نمایند. از سوی دیگر فهرست رتبه‌های ابعاد فضای ویژگی براساس میزان اهمیت آنها خروجی این بخش را تشکیل می‌دهد. در روش پیشنهادی آنچه میزان اهمیت هر یک از ابعاد فضای ویژگی را تعیین می‌نماید و مبنای رتبه‌بندی آنها را تشکیل می‌دهد نحوه توزیع نمونه‌های مثبت و نمونه‌های منفی در فضای ویژگی است. در واقع ارزش هر بعد از فضای ویژگی معادل میزان توانمندی آن بعد در تفکیک نمونه‌های مثبت و منفی می‌باشد.

روش پیشنهادی برای رتبه‌بندی ابعاد فضای ویژگی از ایده‌ای مشابه معیار Fisher[۳۳] [۳۵] بهره می‌برد[۳۴]. در ساده‌ترین حالت می‌توان ارزش (وزن) هر یک از ابعاد فضای ویژگی را با استفاده از معیار Fisher پس از طبقه‌بندی نمونه‌های آموزشی در دو کلاس مختلف براساس رابطه (۵) محاسبه نمود [۳۷].

$$S(i)=\frac{\left|\mu_{ia}-\mu_{ib}\right|}{cv_{ia}+cv_{ib}} \qquad (۵)$$

در رابطه (۵) μ_{ia} و μ_{ib} به ترتیب عبارتند از میانگین نمونه‌های کلاس a و کلاس b در بعد i و cv_{ia} و cv_{ib} معادل کواریانس نمونه‌های دو کلاس a و b در بعد i می‌باشند. بنابراین برطبق معیار Fisher ارزش هر بعد از فضای ویژگی عبارت است از نسبت واریانس بین کلاسی[۳۵] به واریانس درون کلاسی[۳۶] در آن بعد [۳۰]. در صورت طبقه‌بندی نمونه‌های آموزشی در بیش از دو کلاس استفاده از معیار Fisher برای تعیین ارزش ابعاد فضای ویژگی با پیچیدگی بیشتری همراه خواهد بود [۳۶].

روش پیشنهادی در مرحله مقدماتی کلیه نمونه‌های موجود در تمام بسته‌ها را در دو مجموعه نمونه‌های مثبت و نمونه‌های منفی تقسیم‌بندی می‌نماید. مطابق فرض پایه مدل رایج یادگیری چند نمونه‌ای، نمونه‌های مطلوب در اطراف نقطه ایده‌آل در فضای ویژگی قرار دارند و از توزیع مشخصی برخوردارند، بنابراین می‌توان مجموعه نمونه‌های مطلوب را به عنوان یک کلاس در نظر گرفت. اما نمونه‌های منفی از توزیع مشخصی در فضای ویژگی برخوردار نیستند و در نواحی مختلف این فضا پراکنده‌اند به همین دلیل نمی‌توان آنها را به عنوان یک کلاس تلقی کرد [۱۶] ، [۳۹]. به این ترتیب چون فقط یک کلاس موجود می‌باشد این نوع یادگیری نوعی طبقه‌بندی تک کلاسی[۳۷] [۴۲] محسوب می‌شود. در طبقه‌بندی تک کلاسی تلاش می‌شود تا کلاس هدف از سایر نمونه‌ها (Outliers) تشخیص داده شود [۴۳]. در چنین حالتی استفاده از معیار Fisher مناسب نخواهد بود، زیرا طبق تعریف معیار Fisher مبتنی بر وجود دو کلاس متفاوت از نمونه‌ها با توزیع مشخص می‌باشد. بنابراین لازم است تغییراتی در رابطه (۵) اعمال شود تا بتوان از آن به عنوان معیاری مناسب برای رتبه‌بندی ابعاد فضای ویژگی در روش پیشنهادی استفاده کرد. رابطه (۶) معیار پیشنهادی را برای ارزش‌دهی و رتبه‌بندی ابعاد فضای ویژگی بیان می‌کند.

$$S(i)=\frac{\sum_{j=1}^{q}\left|\mu_i^+-I_{ji}^-\right|}{cv_i^+} \qquad (۶)$$

در رابطه (۶) μ_i^+ و cv_i^+ به ترتیب عبارتند از میانگین و کواریانس نمونه‌های مطلوب در بعد i ، q بیانگر تعداد نمونه‌های منفی و I_{ji}^- نشانگر مقدار نمونه منفی j ام در بعد i می‌باشد. از آنجا که نمونه‌های منفی در نواحی مختلف فضای ویژگی پراکنده‌اند و از توزیع مشخصی برخوردار نیستند بنابراین کلاس خاصی را تشکیل نمی‌دهند، به همین دلیل در رابطه (۶) اشاره‌ای به کواریانس نمونه‌های منفی نشده است. همچنین به دلیل مذکور به جای استفاده از میانگین نمونه‌های منفی در رابطه (۶) از مجموع فواصل نمونه‌های منفی با میانگین نمونه‌های مثبت در هر بعد از فضای ویژگی استفاده شده است. براساس معیار پیشنهادی برای محاسبه ارزش هر بعد بیشتر به نمونه‌های مثبت که دارای توزیع مشخصی هستند و تشکیل یک کلاس را می‌دهند توجه شده است، بر این اساس هر قدر پراکندگی نمونه‌های مطلوب در یک بعد کمتر باشد و میانگین نمونه‌های مطلوب در آن بعد فاصله بیشتری با نمونه‌های منفی داشته باشد، آن بعد از قدرت بیشتری برای تفکیک نمونه‌های مطلوب و نمونه‌های منفی برخوردار است و بنابراین ارزش آن بعد افزایش خواهد یافت. پس از محاسبه ارزش هر یک از ابعاد فضای ویژگی براساس رابطه (۶) این ابعاد رتبه‌بندی می‌شوند و به هر یک از آنها یک شماره ردیف از ۱ تا n (n معادل تعداد ابعاد فضای ویژگی می‌باشد) اختصاص داده می‌شود، به این صورت که بعدی در آن مقدار S(i) ماکزیمم است مهم‌تر (Rank=1) و بعدی که کمترین مقدار S(i) را دارد کم اهمیت‌تر (Rank=n) است. رتبه‌بندی انجام شده در این بخش مبنای اصلی توسعه روشمند ابرمکعب پایه در بخش بعدی مرحله تکمیلی را تشکیل می‌دهد.

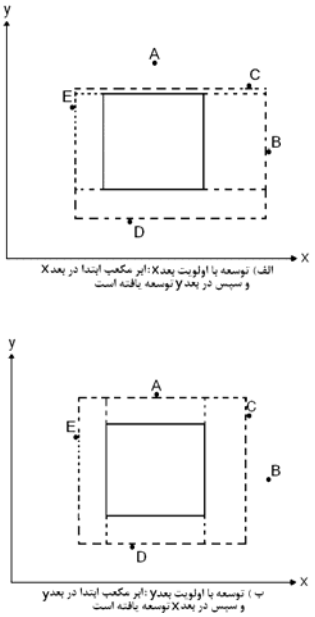
۴-۲-۲- توسعه روشمند ابرمکعب پایه

براساس آنچه در ابتدای مرحله تکمیلی در بخش ۴-۲ ذکر شد توسعه روشمند ابرمکعب پایه در ناحیه بحرانی اطراف آن به‌صورت یک الزام برای روش پیشنهادی مطرح است. در این بخش دو روش برای توسعه ابرمکعب پایه پیشنهاد می‌شود.

۴-۲-۳- روش خوشبینانه

نخستین روش پیشنهادی برای توسعه ابرمکعب پایه روشی با رویکرد خوشبینانه است. مطابق این روش ابرمکعب پایه در هر یک از ابعاد فضای ویژگی، در دو جهت مثبت و منفی، تا رسیدن به اولین نمونه منفی توسعه می‌یابد. پس از توسعه در یک بعد ابرمکعب توسعه یافته مبنای توسعه در ابعاد دیگر می‌باشد. بنابراین ترتیب ابعاد در روند توسعه ابرمکعب پایه می‌تواند نتیجه نهایی را تحت تاثیر قرار دهد. به

عنوان مثال شکل (۵) چگونگی توسعه مستطیل نمایش داده شده در شکل (۴) را نشان می‌دهد. در این شکل در قسمت الف ابتدا مستطیل در بعد X توسعه داده شده و سپس در بعد Y در حالیکه در قسمت ب ابتدا توسعه در بعد Y صورت پذیرفته و سپس در بعد X. چنانکه ملاحظه می‌شود نتیجه نهایی در این دو حالت یکسان نیست.



شکل ۵- توسعه مستطیل در فضای دو بعدی

برای تعیین ترتیب مناسب ابعاد در این قسمت از رتبه‌بندی پیش گفته استفاده می‌شود. در این راستا توجه به دو نکته ضروریست. نکته اول آنکه مطابق رابطه (۶) پراکندگی نمونه‌های مثبت در ابعاد مهم (Rank=1) کمتر است و بنابراین توسعه در این ابعاد باید با احتیاط بیشتری صورت پذیرد، برعکس در ابعادکم اهمیت (Rank=n) پراکندگی نمونه‌های مثبت بیشتر است و توسعه ابرمکعب پایه در این ابعاد می‌تواند گستره بیشتری را شامل شود. نکته دوم این است که با انجام هر توسعه‌ای در یک بعد تعداد نمونه‌های منفی نزدیک ابرمکعب افزایش می‌یابد، این امر فرایند توسعه را مرحله به مرحله محدودتر می‌نماید. با دقت در نکات پیش گفته می‌توان دریافت که توسعه بایداز ابعاد کم اهمیت با کمترین محدودیت‌ها آغاز و به ابعاد مهم با محدودیت‌های بیشتر ختم شود.

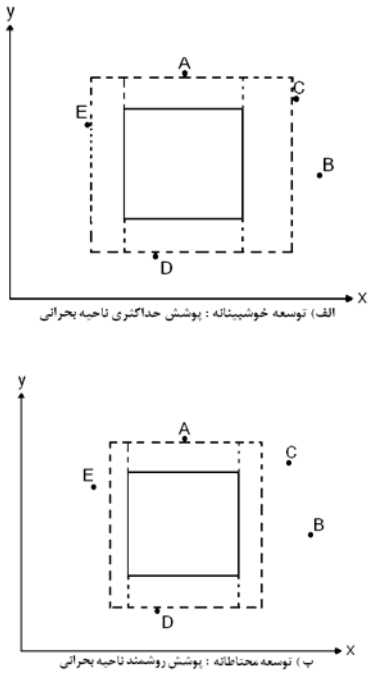
روش خوشبینانه با استفاده از رویکردی ایستا در محاسبه میزان توسعه در هر بعد سعی در پوشش حداکثری ناحیه بحرانی در فرایند توسعه ابرمکعب پایه دارد. اما آزمون‌های عملی نتایج مناسبی را برای این رویکرد نشان نمی‌دهد زیرا این روش متکی بر فرضیه تعلق بخش عمده ناحیه بحرانی به قلمرو نمونه‌های مثبت است در حالیکه طبق تعریف ارائه شده برای ناحیه بحرانی، بسته‌های آموزشی ارائه شده به مدل یادگیرنده این فرضیه را تایید نمی‌نمایند و تنها بر عدم وجود نمونه‌های منفی و مثبت در این ناحیه دلالت می‌کنند.

۴-۲-۴- روش محتاطانه

این روش برای تعیین ترتیب مناسب ابعاد فضای ویژگی در عملیات توسعه شبیه روش خوشبینانه، و بر طبق رتبه‌بندی انجام شده، عمل می‌کند. اما در اندازه توسعه در هر بعد عملکردی متفاوت و پویا دارد. به دیگر بیان توسعه ابرمکعب پایه در این روش با احتیاط و دقت بیشتری صورت می‌گیرد. در این روش بر خلاف روش خوشبینانه توسعه در هر بعد، در دو جهت منفی و مثبت، تنها به محل قرار گرفتن نزدیکترین نمونه‌های منفی وابسته نیست و کل فاصله میان ابرمکعب پایه تا نزدیکترین نمونه‌های منفی را شامل نمی‌شود بلکه تنها بخشی از این فاصله را در بر می‌گیرد. در این روش دامنه توسعه در هر بعد علاوه بر محل قرار گرفتن نزدیکترین نمونه منفی به رتبه آن بعد نیز وابسته است. اگر n تعداد ابعاد فضای ویژگی، $Rank_j$ رتبه بعد j ام و I_{ij}^- مقدار نزدیکترین نمونه منفی به ابرمکعب پایه در جهت مثبت (منفی) بعد j باشد آنگاه دامنه توسعه در جهت مثبت (منفی) بعد j به‌صورت رابطه (۷) قابل تعریف است.

$$ED_j = \frac{Rank_j}{n} \times \left| BHC_{ij} - I_{ij}^- \right| \qquad (۷)$$
$$i \in \{1,2\}$$

بر طبق رابطه (۷) توسعه در کم اهمیت‌ترین بعد ($Rank_j=n$) کل فاصله میان ابرمکعب پایه تا نزدیکترین نمونه منفی را شامل می‌شود اما دامنه این توسعه به تدریج کاهش می‌یابد تا جائیکه در مهم‌ترین بعد ($Rank_j=1$) دامنه این توسعه تا 1/n فاصله مذکور کاهش می‌یابد. بر این اساس روش محتاطانه ضمن استفاده از رویکردی پویا در تعیین دامنه توسعه در هر بعد بر خلاف روش خوشبینانه سعی در پوشش حداکثری ناحیه بحرانی ندارد بلکه سعی در پوشش هدفمند مناسب‌ترین بخش‌های ناحیه بحرانی بر طبق چگونگی پراکندگی نمونه‌های مثبت و منفی دارد. شکل (۶) نحوه توسعه یک مستطیل (تصویر ابرمکعب پایه در فضای دو بعدی) را به دو روش خوشبینانه (الف) و محتاطانه (ب) نشان می‌دهد. در این شکل فرض بر آن است که توسعه ابتدا در بعد Y و سپس در بعدی X انجام می‌گیرد. چنانکه در شکل ملاحظه می‌شود در روش محتاطانه توسعه به‌صورت محدودتر نسبت به روش خوشبینانه انجام گرفته است.



شکل ۶- توسعه یک مستطیل به دو روش خوشبینانه و محتاطانه

۵- پیاده‌سازی و آزمون

در این بخش ابتدا مجموعه داده‌های آموزشی مورد استفاده در آزمون معرفی می‌شود. سپس نتایج بدست آمده از پیاده‌سازی و آزمون روش پیشنهادی ارائه و تحلیل می‌گردد. همچنین در این بخش روش پیشنهادی با سایر روش‌های مطرح در این زمینه براساس نتایج بدست آمده از آنها مقایسه می‌شود. روش پیشنهادی در دو حوزه متفاوت آزمون شده است، حوزه داروسازی و حوزه طبقه‌بندی تصاویر.

۵-۱- مجموعه داده Musk

برای انجام آزمون در حوزه داروسازی، از دو مجموعه داده Musk1 و Musk2 به عنوان مجموعه داده‌های آموزشی استفاده شده است. این داده‌ها به عنوان تنها داده‌های واقعی و محک‌های شناخته شده و همه‌پذیر در یادگیری چند نمونه‌ای تلقی می‌شوند [۱۱] ، [۱۶]. نتایج گزارش شده برای بسیاری از روش‌های مطرح در این زمینه مبتنی بر این مجموعه داده‌ها می‌باشد. در این داده ها هر بسته نشانگر یک مولکول و هر نمونه موجود در یک بسته یکی از شکل‌های متصور برای مولکول مذکور به شمار می‌رود. در این مجموعه داده‌ها یک بسته مثبت معادل است با مولکولی که حداقل یکی از اشکال مختلف آن برای مشارکت در یک ترکیب دارویی خاص مناسب است. همچنین یک بسته منفی بیانگر مولکولی است که هیچ یک از اشکال مختلف آن برای شرکت در ترکیب دارویی موردنظر مناسب نمی‌باشد [۵]. Musk1 و Musk2 به عنوان مجموعه داده‌های شناخته شده در دسترس عموم قرار دارد [۲۳] و رایج‌ترین محک مورد استفاده در حوزه یادگیری چند نمونه‌ای محسوب می‌شود. اطلاعات کاملی در مورد دو مجموعه داده پیش گفته در جدول ۱ ارائه شده است. چنانکه ملاحظه می‌شود مجموعه داده Musk1 هم از نظر تعداد بسته‌ها و هم از نظر متوسط تعداد و تنوع نمونه‌های موجود در بسته‌ها از مجموعه داده Musk2 محدودتر می‌باشد و مجموعه داده Musk2 در مجموع از پیچیدگی بیشتری برخوردار است.

جدول ۱- مشخصات مجموعه داده‌های Musk1 و Musk2

Data set	Musk1	Musk2
Dimensionality	۱۶۶	۱۶۶
Number of bags	۹۲	۱۰۲
Number of positive bags	۴۷	۳۹
Number of negative bags	۴۵	۶۳
Number of instances	۴۷۶	۶۵۹۸
Average number of instances per bag	۵/۱۷	۶۴/۶۹
Maximal number of instances in a bag	۴۰	۱۰۴۴
Minimal number of instances in a bag	۲	۱

روش رایج برای آزمون الگوریتم‌های یادگیری چند نمونه‌ای روش 10-Fold Cross Validation [۸] ، [۴۰] می‌باشد. بر این اساس برای انجام آزمون بر روی دو مجموعه داده Musk1 و Musk2 به روش مذکور عمل شده است. در روش 10-Fold Cross Validation ابتدا کلیه داده‌ها به ده قسمت مساوی تقسیم می‌شود. سپس آزمون در ده مرحله انجام می‌شود. در هر مرحله از آزمون نه گروه از داده‌ها به عنوان داده‌های آموزشی و یک گروه به عنوان داده‌های آزمایشی تلقی می‌شود. در نهایت متوسط میزان دقت در مراحل مختلف آزمون به عنوان دقت کلی الگوریتم یادگیری ارائه می‌شود [۲۹]. همچنین از آنجا که رایج‌ترین معیار مورد استفاده برای سنجش توانمندی روش‌های یادگیری چند نمونه‌ای معیار دقت[۳۸] است [۶] ، [۱۰] در این بخش برای نشان دادن میزان توانمندی روش پیشنهادی و فراهم آوردن امکان مقایسه آن با سایر روش‌های مطرح در این زمینه از معیار دقت استفاده شده است.

یک الگوریتم یادگیری چند نمونه‌ای باید قادر باشد با دریافت مجموعه‌ای از بسته‌های برچسب زده شده، برچسب بسته (های) جدید و ناشناخته را پیش‌بینی نماید. طبق تعریف میزان دقت یک الگوریتم یادگیری عبارت است از نسبت پیش‌بینی‌های صحیح به کل پیش‌بینی‌های انجام شده توسط آن الگوریتم [۵] ، [۴۱].

جدول ۲ نتایج اجرای روش پیشنهادی را بر روی MUSK1 و MUSK2 در سه حالت مختلف نشان می‌دهد. در حالت اول (بدون توسعه) از ابرمکعب پایه بدون توسعه برای نمایش مفهوم موردنظر استفاده شده است. حالت دوم (توسعه خوشبینانه) مبتنی بر شکل توسعه یافته ابرمکعب پایه براساس رویکرد خوشبینانه است و در حالت سوم (توسعه محتاطانه) شکل توسعه یافته ابرمکعب پایه براساس رویکرد محتاطانه برای نمایش مفهوم موردنظر استفاده شده است.

چنانکه ملاحظه می‌شود در هر سه حالت میزان دقت روش پیشنهادی در یادگیری از مجموعه داده Musk2 بیشتر از زمانی است که از Musk1 استفاده شده است، با توجه به مشخصات ذکر شده در جدول ۱ برای این دو مجموعه داده، می‌توان دریافت که روش پیشنهادی با افزایش تعداد بسته‌های آموزشی و تعداد و تنوع نمونه‌های موجود در آنها با کاهش دقت مواجه نمی‌شود و قادر به حفظ دقت در حد قابل قبول می‌باشد. بنابراین ویژگی مذکور یکی از مزایای روش پیشنهادی به شمار می‌رود. علت این ثبات دقت در حالت‌های دوم و سوم، توسعه مناسب ابرمکعب پایه می‌باشد. اما در حالت اول این موضوع اهمیت کمتری دارد زیرا علت آن صرفاً کمتر بودن تعداد نمونه‌های مثبت در مجموعه داده Musk2 نسبت به Musk1 می‌باشد. همچنین براساس جدول ۲ می‌توان دریافت در حالت سوم، که شکل توسعه یافته ابرمکعب پایه براساس رویکرد محتاطانه برای نمایش مفهوم یادگیری شده استفاده شده است، بیشترین میزان دقت بدست آمده است و در این حالت روش پیشنهادی در یادگیری از این دو مجموعه داده از ثبات رفتاری مناسبی برخوردار است.

جدول ۲- میزان دقت الگوریتم پیشنهادی در سه حالت مختلف

روش یادگیری	Musk1 دقت (%)	Musk2 دقت (%)
حالت اول (بدون توسعه)	۷۱/۸	۷۴/۵
حالت دوم (توسعه خوشبینانه)	۸۲/۶	۸۵/۳
حالت سوم (توسعه محتاطانه)	**۸۹/۲**	**۹۰/۲**

جدول ۳ نتایج اجرای روش پیشنهادی و سایر روش‌های مهم ارائه شده توسط پژوهشگران برای یادگیری چند نمونه‌ای را بر روی دو مجموعه داده Musk1 و Musk2 نشان می‌دهد. بر طبق جدول ۳ می‌توان دریافت که روش پیشنهادی در مقایسه با سایر روش‌های مهم ارائه شده در این زمینه از دقت مناسبی برخوردار است. علاوه بر این اطلاعات جدول مذکور نشان می‌دهد که بسیاری از روش‌های ارائه شده جهت یادگیری چند نمونه‌ای در یادگیری از مجموعه داده Musk2، که درآن تعداد و تنوع نمونه‌ها و بسته‌ها بیشتر است، دچار افت دقت می‌شوند و فقط روش‌های MI-SVM، MULTINST و روش پیشنهادی در یادگیری از مجموعه داده Musk2 از دقت بیشتری نسبت به Musk1 برخوردارند. روش پیشنهادی علاوه بر آنکه یادگیری از Musk2 را با دقت بالاتری نسبت به Musk1 انجام می‌دهد در مقایسه با دو روش پیش گفته نیز دقت مناسبی دارد. علت این افزایش دقت در روش پیشنهادی توسعه مناسب ابرمکعب پایه می‌باشد.

توجه به این نکته ضروری است که برخی روش‌ها نظیر EM-DD و Iterated-discrim با استفاده از تنظیمات خاص وابسته به داده‌های آموزشی در مراحل آزمون به دقت‌های یاد شده دست می‌یابند [۵] ، [۶] بنابراین در این روش‌ها با تغییر مجموعه داده‌های آموزشی برای حصول دقت مناسب باید تنظیمات خاصی انجام شود. در حالی که روش پیشنهادی بدون نیاز به انجام تنظیمات خاص وابسته به داده‌های آموزشی، یادگیری را با دقت مذکور انجام می‌دهد. از سوی دیگر بر طبق [۶] ، [۹] و [۱۰] دقت نسبتاً بالای روش Iterated-discrim به علت آن است که این روش به صورت خاص برای مجموعه داده‌های Musk1 و Musk2 طراحی و بهینه‌سازی شده است. اما روش پیشنهادی هیچ وابستگی خاصی به این مجموعه داده‌ها ندارد.

جدول ۳- مقایسه میزان دقت روش‌های مختلف بر روی Musk1 و Musk2

روش یادگیری	MUSK1 دقت (٪)	MUSK2 دقت (٪)
[39]EM-DD [6]	۹۶/۸	۹۶
Iterated-discrim [5]	۹۲/۴	۸۹/۲
Citation-kNN [9]	۹۲/۴	۸۶/۳
Bayesian-kNN [9]	۹۰/۲	۸۲/۴
Diverse Density [8]	۸۸/۹	۸۲/۵
MI-NN [12]	۸۸/۹	۸۲/۵
mi-SVM [10]	۸۷/۴	۸۳/۶
MI-SVM [10]	۷۷/۹	۸۴/۳
MULTINST [7]	۷۶/۷	۸۴
روش پیشنهادی	**۸۹/۲**	**۹۰/۳**

۵-۲- مجموعه داده تصویری

در این بخش نتایج حاصل از پیاده‌سازی و اجرای روش پیشنهادی در درک مفهوم و طبقه‌بندی تصاویر طبیعت در سه گروه کوهستان، آبشار و غروب آفتاب ارائه می‌شود. برای انجام آزمون در این حوزه از یک مجموعه داده مشتمل بر ۵۰۰ تصویر طبیعت استفاده گردیده است. تمامی این تصاویر در قالب JPEG و در ابعاد ۲۵۶×۳۸۴ یا ۳۸۴×۲۵۶ می‌باشند. این پایگاه داده شامل دو بخش اصلی و فرعی می‌باشد. بخش اصلی مشتمل بر یکصد تصویر طبیعت از هر یک از سه گروه پیش گفته است که در مجموع سیصد تصویر را در بر می‌گیرد. بخش فرعی شامل ۲۰۰ تصویر از کلاس‌های مختلف COREL می‌باشد که برای افزایش دقت آزمون مورد استفاده قرار گرفته‌اند. شکل ۷ مثال‌هایی از تصاویر موجود در هر یک از گروه‌های پیش گفته در پایگاه داده تصویری را نشان می دهد.

گروه ۱: کوهستان

گروه ۲: آبشار

گروه ۳: غروب آفتاب

شکل ۷- مثال‌هایی از تصاویر پایگاه داده تصویری در سه گروه متفاوت

تصاویر موجود در مجموعه فوق ابتدا با استفاده از روش SBN [۴۴] به بسته‌های قابل استفاده در یادگیری چند نمونه‌ای می‌شوند. در این روش، که مبتنی بر ترکیب دو ویژگی رنگ و روابط مکانی[40] نواحی داخل تصویر است، هر بخش تشکیل دهنده تصویر تبدیل به یک نمونه و کل تصویر تبدیل به یک بسته می‌شود. هر یک از بسته‌های تولید شده توسط این روش، دارای ۹ نمونه می‌باشد و هر نمونه معادل یک بردار ویژگی ۱۵ بعدی است. در مرحله بعد، از بسته‌های متناظر با تصاویر برای آزمون استفاده می‌شود. بنابراین دقت فرایند ناحیه‌بندی[41] تصویر و تولید بسته‌ها، تأثیر زیادی بر دقت نتایج طبقه‌بندی خواهد داشت. برای انجام آزمون در این بخش در هر مرحله از آزمون از پنج مثال مثبت و پنج مثال منفی (5p5n) به عنوان داده‌های آموزشی استفاده شده و سایر تصاویر به عنوان داده‌های آزمایشی مورد استفاده قرار گرفته‌اند.

جدول ۴- مقایسه میزان دقت روش‌های مختلف بر روی مجموعه داده تصویری

روش یادگیری	کوهستان دقت (٪)	آبشار دقت (٪)	غروب آفتاب دقت (٪)
EM-DD	۸۵/۶۲	۸۴/۱۴	۸۷/۲
DD	۸۲/۵۷	۷۹	۸۶/۷۱
CkNN	۷۳/۵۶	۷۵/۵۷	۸۰/۴۳
روش پیشنهادی	**۸۴/۲۸**	**۸۲/۷۱**	**۸۹/۷۱**

جدول ۴ نتایج اجرای روش پیشنهادی و سه روش مهم دیگر در یادگیری چند نمونه‌ای را بر روی مجموعه داده تصویری پیش گفته نشان می‌دهد. مقادیر نشان داده شده در این جدول در واقع متوسط میزان دقت در کلیه مراحل آزمون مربوط به هر یک از گروه‌های تصویری را بیان می‌کند. بر طبق جدول ۳ می‌توان دریافت که روش EM-DD در مقایسه با سایر روش‌ها، همچنان بالاترین میزان دقت را داراست. همچنین براساس این نتایج، روش پیشنهادی در یادگیری از مجموعه داده تصویری مذکور نیز از دقت خوبی برخوردار است.

۶- نتیجه‌گیری و توسعه‌های آتی

در این پژوهش ضمن بررسی مساله یادگیری چند نمونه‌ای، روشی دو مرحله‌ای برای حل این مساله ارائه شد. این روش قادر به یادگیری مفهوم موردنظر و نمایش آن در قالب یک ابرمکعب در فضای ویژگی n بعدی با دقت قابل قبول می‌باشد. نتایج حاصل از پیاده‌سازی و آزمون این روش بر روی محک‌های شناخته شده Musk1 و Musk2 بیانگر دقت مناسب این روش در مقایسه با سایر روش‌های مطرح در این زمینه می‌باشد. همچنین براساس نتایج آزمون‌های انجام شده در این تحقیق روش پیشنهادی در حوزه طبقه‌بندی تصاویر نیز از دقت قابل قبولی برخوردار است. علاوه بر این طراحی روش پیشنهادی به گونه‌ای است که الگوریتم‌های مورد استفاده در هر یک از مراحل آن مستقل از یکدیگر است و امکان بکارگیری الگوریتم‌های جدید بصورت مستقل در هر یک از مراحل آن وجود دارد، بر این اساس روش پیشنهادی از ظرفیت و انعطاف زیادی برای بهسازی و توسعه‌های آتی برخوردار است.

مهمترین توسعه‌های آتی متصور برای روش پیشنهادی در دو محور قابل ذکر می‌باشند. محور اول، امکان بهره‌گیری از الگوریتم‌های جدید در هر یک از مراحل این روش است. از آنجا که مراحل روش پیشنهادی از استقلال نسبی برخوردار

هستند بنابراین می‌توان الگوریتم‌های مورد استفاده در هر یک از مراحل را به صورت جداگانه باز طراحی نمود. روش پیشنهادی متکی بر فرضیه وجود تک نقطه ایده‌آل در فضای ویژگی به عنوان مفهوم هدف است. بنابراین چنانچه مفهوم هدف مشتمل بر چندین نقطه در فضای ویژگی باشد این روش قادر به یادگیری آن نمی‌باشد بر این اساس دومین محور متصور برای توسعه این روش توانمند کردن آن برای انجام انواع مختلف یادگیری چند نمونه‌ای عمومیت یافته می‌باشد. بر خلاف برخی از روش‌های ارائه شده در این زمینه که فقط برای انجام نوع خاصی از یادگیری چند نمونه‌ای طراحی شده‌اند ساختار درونی روش پیشنهادی به گونه‌ای است که توسعه مذکور برای انجام برخی انواع یادگیری چند نمونه‌ای عمومیت یافته به راحتی امکان‌پذیر است.

فعالیت بعدی مؤلفین این مقاله، که در حال انجام است، استفاده از روش پیشنهادی در حوزه بازیابی تصاویر می‌باشد. با توجه به انطباق نیازهای این حوزه با توانمندی‌های یادگیری چند نمونه‌ای بهره‌گیری از روش پیشنهادی در این حوزه می‌تواند مفید و مؤثر باشد.

قدردانی

این پژوهش با حمایت مالی مرکز تحقیقات مخابرات ایران طبق قرارداد شماره ۵۰۰/۸۲۲۹/ت انجام شده است. بدینوسیله از مسئولین محترم این مرکز تشکر و قدردانی می‌شود.

مراجع

[1] Z. H. Zhou, K. Jiang, and M. Li , "Multi-Instance Learning Based Web Mining," Applied Intelligence, Vol. 22, No. 2, pp. 135-147, 2005.

[2] C. Zhang, S. C. Chen, and M. L. Shyu, "Multiple Object Retrieval for Image Databases Using Multiple Instance Learning and Relevance Feedback," Proc. IEEE Intl Conf. Multimedia and Expo, pp. 775-778, 2004.

[3] X. Huang, S. C. Chen, and M. L. Shyu, "An Open Multiple Instance Learning Framework and Its Application in Drug Activity Prediction Problems," Proc. 3rd IEEE Symposium on BioInformatics and BioEngineering, pp. 53-59, 2003.

[4] S. Ray, and M. Craven, "Supervised versus Multiple Instance Learning: An Empirical Comparison," Proc. Of the 22nd Intl Conf. on Machine Learning, pp. 697-704, 2005.

[5] T. Dietterich, R. Lathrop, and T. Lozano-Perez, "Solving the multiple-instance problem with axis-parallel rectangles," Artificial Intelligence,Vol. 89, pp. 31-71, 1997.

[6] Q. Zhang, and S. A. Goldman, "EM-DD: An improved multiple-instance learning technique," Advances in Neural Information Processing Systems, Vol. 14, pp. 1073-1080, 2002.

[7] P. Auer, "On Learning From Multi-instance Examples: Empirical Evaluation of a Theoretical Approach," Proc. Of the 14th Intl Conf. on Machine Learning, pp. 21-29, 1997.

[8] O. Maron, and T. Lozano-Perez, "A framework for multiple-instance learning," Advances in Neural Information Processing Systems, Vol. 10, pp. 570-576, 1998.

[9] J. Wang, and J. D. Zucker, "Solving the multiple instance problem: A lazy learning Approach," Proc. Of the 17th Intl Conf. on Machine Learning , pp. 1119-1125, 2000.

[10] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," Advances in Neural Information Processing Systems , Vol 15, pp. 561-568, 2003.

[11] Q. Tao, S. Scott, N. V. Vinodchandran, and T. T. Osugi, "SVM-Based Generalized Multiple-Instance Learning via Approximate Box Counting," Proc. Of the 21st Intl Conf. on Machine Learning, pp. 779-806, 2004.

[12] J. Ramon, and L. D. Raedt, "Multi instance neural networks," Proc. Of ICML-2000, Workshop on Attribute-Value and Relational Learning, 2000.

[13] Z. H. Zhou, and M. L. Zhang, "Ensembles of multi-instance learners," In N. Lavrac,D. Gamberger, H. Blockeel, and L. Todorovski, Eds. Lecture Notes in Artificial Intelligence 2837, Berlin: Springer, pp. 492–502, 2003.

[14] P. Auer, and R. Ortner, "A boosting approach to multiple instance learning," Proc. Of the 15th European Conf. on Machine Learning, pp. 63-74, 2004.

[15] A. Cannon, and D. Hush, "Multiple Instance Learning Using Simple Classifier," In Proc. Of International Conf. on Machine Learning and Applications, pp. 123-128, 2004.

[16] D. R. Dooly, Q. Zhang, S. A. Goldman, and R. A. Amar, "Multiple-instance learning of real-valued data," Journal of Machine Learning Research, Vol 3, pp. 651–678, 2002.

[17] N. Weidmann, E. Frank, and B. Pfahringer, "A two-level learning method for generalized multi-instance problems," Proc. Of the European Conf. on Machine Learning, LNCS 2837, pp. 468–479, 2003.

[18] Q. Tao, and S. Scott, "A Faster Algorithm for Generalized Multiple-Instance Learning," Proc. Of 17th Intl Florida Artificial Intelligence Research Society Conf. , pp. 550-555, 2004.

[19] Q. Tao, S. Scott, N. V. Vinodchandran, T. T. Osugi, and B. Mueller, "An Extended Kernel for Generalized Multiple-Instance Learning," Proc. Of the 16th IEEE Intl Conf. on Tools with Artificial Intelligence, pp. 272-277, 2004.

[20] M. R. Naphade, and J. R. Smith, "A Generalized Multiple Instance Learning Algorithm for Large Scale Modeling of Multimedia Semantics," Proc. Of IEEE Intl Conf. on Acoustics, Speech , and Signal Processing , pp. v/341-v/344, 2005.

[21] P. M. Cheung, and J. T. Kwok, "A regularization framework for multiple-instance learning," Proc. of the 23rd Intl Conf. on Machine Learning , pp.193-200, 2006.

[22] X. Huang, S.-C. Chen, and M.-L. Shyu, "Incorporating Real-Valued Multiple Instance Learning Into Relevance Feedback For Image Retrieval," Proc. Of the IEEE Intl Conf. on Multimedia & Expo, Vol. I, pp. 321-324, 2003.

[23] C. Blake, E. Keogh, and C. J. Merz, UCI repository of machine learning databases, www.ics.uci.edu/˜mlearn /MLRepository.html.

[24] Y. Chen, and J. Z. Wang, "Image categorization by learning and reasoning with regions," Journal of Machine Learning Research,Vol. 5, pp. 913–939, 2004.

[25] J. W. Hsieh, C. C. Chiang, Y. S. Huang, and W. E. L. Grimson, "Learning Visual Concepts from Image Instances," Journal of Information Science and Engineering, Vol. 20, pp. 1197-1212, 2004.

[26] P. S. Bradley, O. L. Mangasarian, and W. N. Street, "Clustering via concave minimization," In M. C. Mozer, M. I. Jordan, and T. Petsche,editors, Advances in Neural Information Processing Systems Vol. 9, pp. 368-374, 1997.

[27] Z. H. Zhou, and M. L. Zhang, "Neural networks for multi-instance learning," Proc. Of the Intl Conf. on Intelligent Information Technology, pp. 455-459, 2002.

[28] M. L. Zhang, and Z. H. Zhou, "Adapting RBF Neural Networks to Multi-Instance Learning," Neural Processing Letters, Vol. 23, No. 1, pp. 1-26 , 2006.

[29] O. Maron, Learning from ambiguity. PhD dissertation, Department of Electrical Engineering and Computer Science, MIT, 1998.

[30] I. Guyon, and A. Elisseeff, "An Introduction to Variable and Feature Selection," Journal of Machine Learning Research, Vol. 3, pp. 1157–1182, 2003.

[31] H. Liu, and L. Yu, "Toward Integrating Feature Selection Algorithms for Classification and Clustering," IEEE Trans. Knowledge and Data Engineering, Vol. 17, No. 4, pp. 491-502, 2005.

[32] T. Kanungo, D. M. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and Implementation," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 24, No. 7, pp. 881-892, 2002.

[33] C. Ding, and X. He, "K-means Clustering via Principal Component Analysis," Proc. of Intl Conf. Machine Learning, pp. 225-232, 2004.

[34] K. Chen, "On k-Median clustering in high dimensions," Proc. of the 17th annual ACM-SIAM symposium on Discrete algorithm, pp.1177-1185, 2006.

[35] J. Yang, A. F. Frangi, J. Yang, D. Zhang, and Z. Jin, "KPCA Plus LDA: A Complete Kernel Fisher Discriminant Framework for Feature Extraction and Recognition," IEEE Trans. on Pattern Analysis and Machine Intelligence Vol. 27, No. 2, pp. 230-244, 2005.

[36] R. O. Duda, P. E. Hart, and D. G. Stork. Pattern Classification. John Wiley & Sons, USA, 2nd edition, 2001.

[37] T. Cooke, "Two Variations on Fisher's Linear Discriminant for Pattern Recognition," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 24, No. 2, pp. 268-273, 2002.

[38] S. A. Goldman, and R. Rahmani, "MISSL: Multiple-Instance Semi-Supervised Learning" Proc. of the 23rd Intl Conference on Machine Learning , pp. 705-712, 2006.

[39] C. Zhang, X. Chen, M. Chen, S. C. Chen, and M. L. Shyu, "A Multiple Instance Learning Approach for Content Based Image Retrieval Using One-Class Support Vector Machine," Proc. IEEE Intl Conf. on Multimedia & Expo , pp. 1142-1145, 2005.

[40] Z. H. Zhou, "Multi-instance learning from supervised view," Journal of Computer Science & Technology , Vol. 21, No. 5, pp. 800-809, 2006.

[41] R. Kohavi, and F. Provost, "Glossary of Terms," Machine Learning, Vol. 30, pp. 271-274, 1998.

[42] D. M. J. Tax, One-class classification, PhD Thesis, Delft University of Technology, 2001.

[43] D. Wang, D. S. Yeung, and E. C. C. Tsang, "Structured one class classification," IEEE Trans. on System, Man and Cybernetics, Vol. 36, No. 6, pp. 1283-1295, 2006.

[44] Z. H. Zhou, X. B. Xue, and Y. Jiang, "Locating regions of interest in CBIR with multi-instance learning techniques," Lecture Notes in Artificial Intelligence 3809, Zhang S, Jarvis R (eds.), Berlin: Springer, pp. 92-101, 2005.

محمدرضا کیوان‌پور مدرک کارشناسی خود را در سال ۱۳۷۶ در رشته مهندسی کامپیوتر ـ نرم‌افزار از دانشگاه علم و صنعت ایران اخذ نمود و دوره کارشناسی ارشد مهندسی کامپیوتر گرایش نرم‌افزار را در سال ۱۳۷۹ در دانشگاه تربیت مدرس به پایان رساند. وی هم‌اکنون دانشجوی دکتری همان دانشگاه می‌باشد. زمینه‌های تحقیقاتی مورد علاقه ایشان، طبقه‌بندی و بازیابی تصویر، تصویرکاوی و یادگیری ماشین می‌باشد.

آدرس پست‌الکترونیکی ایشان عبارت است از:

keyvanm@modares.ac.ir

نصراله مقدم چرکری مدرک کارشناسی خود را در سال ۱۳۶۴ در رشته علوم کامپیوتر از دانشگاه شهید بهشتی تهران، و مدارک کارشناسی ارشد و دکتری خود را نیز به ترتیب در سال‌های ۱۳۷۱ و ۱۳۷۴ در رشته علوم کامپیوتر از دانشگاه یاماناشی ژاپن اخذ نمود. ایشان از سال ۱۳۷۴ با عنوان عضو هیات علمی به دانشگاه تربیت مدرس تهران پیوست. وی هم‌اکنون عضو انجمن‌های تخصصی

IEEE، INEC و CSI می‌باشد. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: پردازش تصویر، بازیابی تصویر، نهان‌نگاری و الگوریتم‌های موازی.
آدرس پست‌الکترونیکی ایشان عبارت است از:
moghadam@modares.ac.ir

---

[1] Multiple – Instance Learning (MIL
[2] Supervised
[3] Training Set
[4] Bag
[5] Target Concept
[6] Content-Based Image Retrieval
[7] Hyper Cube
[8] Benchmark
[9] Semisuprvised
[10] Axis-Parallel-Rectangles(APR)
[11] گزینش پارامترهای مناسب برای افزایش دقت در آزمون
[12] Diverse-Density
[13] Expectation-Maximization
[14] Support Vector Machine (SVM)
[15] Generalized Multiple Instance Learning
[16] Multiple-Instance Assumption
[17] Bag Space
[18] Clustering
[19] Homogeneity
[20] Outliers
[21] Median
[22] Base Bag
[23] Minimal
[24] نمونه‌های موجود در بسته‌های منفی
[25] Base Hyper Cube (BHC)
[26] Training Bags
[27] Test Bags
[28] False Negative
[29] False Positive
[30] Feature Selection
[31] Dimension Reduction
[32] Classification
[33] Fisher Criterion
[34] این معیار در تعیین میزان دقت روش‌های طبقه‌بندی نیز استفاده می‌شود [۳۰].
[35] Between Class Variance
[36] Within Class Variance
[37] One Class Classification
[38] Accuracy
[39] در [۱۰] برای روش EM-DD در یادگیری از مجموعه داده‌های Musk1 و Musk2 به ترتیب دقت‌های ۸۴/۸ و ۸۴/۹ درصد گزارش شده است.
[40] Spatial Relationship
[41] Segmentation