

## ارائه یک مدل طبقه‌بندی چند بر چسبی با قابلیت تعمیم بالا

سلیم حسینی دادوکلائی<sup>۱</sup>، علی امیری<sup>۲</sup>، حسین صفری<sup>۳</sup> و داریوش نجفی<sup>۴</sup>

<sup>۱</sup> دانشجوی کارشناسی ارشد مهندسی نرم افزار، دانشکده مهندسی کامپیوتر دانشگاه زنجان، salim.hoseini@znu.ac.ir

<sup>۲</sup> عضو هیئت علمی و استادیار دانشکده مهندسی کامپیوتر دانشگاه زنجان. a\_amiri@znu.ac.ir

<sup>۳</sup> عضو هیئت علمی و دانشیار دانشکده فیزیک دانشگاه زنجان. safari@znu.ac.ir

<sup>۴</sup> عضو هیئت علمی و مربی دانشکده مهندسی کامپیوتر دانشگاه زنجان. d\_najafi@znu.ac.ir

چکیده - امروزه با افزایش حجم داده‌ها، امکان جمع‌آوری و طبقه‌بندی سریع داده‌ها توسط انسان غیرممکن شده است و نیاز به طبقه‌بندی و تحلیل خودکار از جایگاه ویژه‌ای برخوردار است. در طبقه‌بندی داده‌ها، ابتدا طی یک فرآیند نمونه‌های آموزشی به همراه برچسب‌های آن‌ها به یک الگوریتم یادگیری داده می‌شود تا از تباط بین نمونه‌ها و برچسب‌ها را یاد بگیرد و سپس برچسب داده‌های آزمایشی را پیش‌بینی کند. داده‌های چند برچسبی داده‌هایی هستند که در آن نمونه‌ها می‌توانند بیش از یک برچسب کلاس داشته باشند، به عبارت دیگر هر نمونه توسط یک مجموعه از برچسب‌ها نمایش داده‌ها می‌شود.

در این مقاله، روش ماشین بردار پشتیبان ساختاری را با ماشین بردار پشتیبان چند برچسبی برای ایجاد یک طبقه‌بند داده‌های چند برچسبی ترکیب می‌کنیم. این روش از دو مرحله تشکیل شده است: مرحله اول شامل خوشه‌بندی داده‌ها است که کمک می‌کند مرز تصمیم بهتر نسبت به ماشین بردار پشتیبان داشته باشیم و مرحله دوم فرآیند یادگیری است که همانند SVM است. با ارزیابی روش پیشنهادی بر روی مجموعه داده‌های موجود در مقایسه با روش‌های موجود، به این نتیجه می‌رسیم که از لحاظ دقت و صحت پیش‌بینی از کارایی بهتری برخوردار است.

کلید واژه - ماشین بردار پشتیبان، طبقه‌بندی چند برچسبی، خوشه‌بندی.

### ۱- مقدمه

استراتژی توسعه الگوریتم، یک الگوریتم طبقه‌بندی چند کلاسه خاص را برای همه نمونه‌های آموزشی و همه کلاس‌های یک مجموعه داده آموزشی چند برچسبی، عمومیت می‌بخشد. چنین استراتژی می‌تواند در مسائل بهینه‌سازی پیچیده مورد استفاده قرار گیرد. برای مثال برنامه‌نویسی درجه دو بزرگ مقیاس در ماشین بردار پشتیبان چند برچسبی و بهینه‌سازی بدون قید در شبکه‌های عصبی چند برچسبی [۸-۱۰].

استراتژی ترکیبی از برخی الگوریتم‌های طبقه‌بندی تک برچسب با یک یا دو ترفند تجزیه داده که به صورت ضمنی یا صریح بکار رفته‌اند را گسترش می‌دهد. به عبارت دیگر برخی از محاسن دو گروه فوق از روش‌های چند برچسبی را در یک جا جمع می‌کند. این استراتژی برای طراحی و پیاده‌سازی چندین طبقه‌بند چند برچسبی مورد استفاده قرار می‌گیرد [۱۱ و ۱۲].

در این مقاله، ما روی نوع خاصی از طبقه‌بندی داده‌های چند برچسبی مبتنی بر ماشین بردار پشتیبان تمرکز داریم. ماشین بردار پشتیبان ساختاری به عنوان گونه متفاوت از یک ماشین بردار پشتیبان شناخته می‌شود. این ماشین بردار پشتیبان از دو مرحله تشکیل شده است مرحله اول شامل خوشه‌بندی<sup>۱</sup> داده‌ها است؛ برای خوشه‌بندی داده‌ها از تکنیک پیوند Ward استفاده می‌کند که کمک می‌کند مرز تصمیم بهتر نسبت به ماشین بردار پشتیبان

طبقه‌بندی داده‌های چند برچسبی<sup>۱</sup> یک مسئله یادگیری با نظارت می‌باشد که در آن هر نمونه می‌تواند همزمان با چندین کلاس یا برچسب همراه شود و این کلاس‌ها مختص یک نمونه خاص نیز نمی‌باشند [۱]. اخیراً طبقه‌بندی چند برچسبی توجه زیادی نسبت به کاربردهای دنیای واقعی پرداخته است به عنوان مثال، تقسیم‌بندی متن [۵]، بیوانفورماتیک [۶]، تفسیر صحنه و فیلم و دسته‌بندی موسیقی [۷] نام برد. سه نوع دسته برای طبقه‌بندی چند برچسبی وجود دارد: تجزیه داده<sup>۵</sup>، توسعه الگوریتم<sup>۶</sup> و استراتژی ترکیبی<sup>۷</sup>.

استراتژی تجزیه داده مجموعه داده چند برچسبی را به یک یا چند زیرمجموعه تک برچسب (دودویی یا چند کلاسه) تقسیم می‌کند و سپس برای هر زیرمجموعه با استفاده از طبقه‌بندهای موجود، یک طبقه‌بند برای آنان می‌سازد و سپس همه این طبقه‌بندها در یک طبقه‌بند چند برچسبی کلی جمع می‌کند. عمدتاً ۳ نوع ترفند تجزیه داده وجود دارد: یک در مقابل یک، یک در مقابل بقیه، برچسب ابرمجموعه. روش چند برچسبی تجزیه داده برای پیاده‌سازی مناسب است چون تعدادی از طبقه‌بندها در این حوزه و نرم افزارهای رایگان آنان، موجود می‌باشند [۲-۴].

در این موارد یک طبقه‌بند برای اینکه برچسب‌های یک نمونه را مشخص کند از رویکرد مقدار آستانه<sup>۱۰</sup> استفاده می‌کند که این مقدار آستانه می‌تواند یک مقدار ثابت یا یک تابع از مقادیر برچسب‌ها باشد. در ضمن رتبه‌ای که یک برچسب مرتبط می‌گیرد باید بیشتر از رتبه‌ای که یک برچسب نامرتب همان نمونه می‌گیرد، باشد [۴].

$$f(x) = \{k | r_k(x) > t(x), k = 1, \dots, q\} \quad (1)$$

که در آن  $t(x)$  مقدار آستانه و  $r_k(x)$  رتبه هر برچسب است. در ادامه مقاله در بخش ۲ روش پیشنهادی ارائه می‌شود و در بخش ۳ به ارزیابی و مقایسه نتایج حاصل از پیاده‌سازی می‌پردازیم و در نهایت بخش ۴ نتایج و جمع‌بندی ارائه می‌شود.

## ۲- روش پیشنهادی

هدف نهایی در طبقه‌بندی داده‌های چند برچسبی، ارائه الگوریتم‌هایی است که توانایی پیش‌بینی برچسب‌های نمونه‌های جدید را داشته باشند. بنابراین لازم است الگوریتم‌های ارائه شده در این راستا دارای دقت پیش‌بینی خوبی باشند و همزمان بتوانند بر روی مجموعه داده‌های خیلی بزرگ نیز اجرا شوند. به این دلیل از گونه خاصی از ماشین بردار پشتیبان برای طبقه‌بندی استفاده می‌کنیم.

ماشین بردار پشتیبان<sup>۱۱</sup> [۱۳] طبقه‌بندی است که با استفاده از یک مرز تصمیم<sup>۱۲</sup> می‌خواهد داده‌های با کلاس‌های مختلف را از هم جدا کند. مرز تصمیم می‌تواند هم به صورت خط، صفحه یا ابر صفحه<sup>۱۳</sup> و هم به صورت یک منحنی باشد.

فرض می‌کنیم  $w x + b = 0$  (همان‌طور که در شکل (۱) مشخص است)، مرز تصمیم باشد؛ در این شکل داده‌ها در دو کلاس توزیع شده‌اند. اگر فاصله بین خطی موازی با مرز تصمیم با نزدیک‌ترین نقطه در کلاس اول و از آن طرف با خط موازی که از نزدیک‌ترین نقطه به مرز تصمیم در کلاس دوم عبور می‌کند را حاشیه<sup>۱۴</sup> تصمیم بنامیم، هدف ماشین بردار پشتیبان این است مرز تصمیم را طوری پیدا کند که این مرز تصمیم بیشترین حاشیه را داشته باشد. طبیعی است هر چه این مرز تصمیم بزرگ‌تر باشد خطای طبقه‌بند هم کمتر خواهد بود.

فرم اصلی ماشین بردار پشتیبان برای حالت داده‌های جدا ناپذیر غیرخطی در رابطه زیر نشان داده است که یک مسئله بهینه‌سازی مقید<sup>۱۵</sup> با قیدهای نامساوی است که با تشکیل تابع لاگرانژ و معادلات KKT دوگان رابطه زیر به دست می‌آید که می‌توان با استفاده از روش برنامه‌نویسی درجه دو<sup>۱۶</sup> حل نمود.

استاندارد داشته باشد و مرحله دوم فرآیند یادگیری است که همانند ماشین بردار پشتیبان استاندارد می‌باشد و در نهایت تابع هدف بهینه‌سازی تعریف شده از این مساله، با برنامه نویسی درجه دو حل خواهد شد.<sup>۹</sup> در ادامه نتایج آزمایشی روی چند مجموعه داده با چند روش موجود نظیر Rank-SVM، ADTree، ML-KNN مقایسه خواهیم کرد.

## ۱-۱- تعریف مساله

اگر  $X$  یک رکورد از مجموعه داده باشد که این رکورد دارای  $d$  ویژگی یا مشخصه باشد و  $Y$  یک مجموعه از برچسب‌های ممکن به اندازه  $q$  باشد؛ این مجموعه را به صورت زیر نمایش می‌دهیم:  $\{\lambda_1, \lambda_2, \dots, \lambda_q\}$ . آنگاه مجموعه داده  $S$  یک مجموعه داده چند برچسبی است که به صورت  $\{(x_i, Y_i) | 1 \leq i \leq l\}$  می‌باشد که  $x_i$  یک نمونه آموزشی و  $Y_i$  زیرمجموعه‌ای از برچسب‌های ممکن می‌باشد. در نهایت یک طبقه‌بند چند برچسبی برای هر یک از نمونه‌های آزمایشی یکی از  $2^y$  حالت از زیرمجموعه برچسب‌های ممکن را پیش‌بینی می‌کند.

فرض کنیم مجموعه آموزشی دارای ۵ برچسب  $M = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5\}$  می‌باشد و یک نمونه آموزشی مانند  $x$  را در آن در نظر می‌گیریم، چنانچه نمونه  $x$  در مجموعه داده‌ها، دارای برچسب‌های  $\{\lambda_1, \lambda_4\}$  باشد آنگاه مجموعه برچسب‌ها برای این نمونه به دو حالت تقسیم می‌شود؛ مجموعه برچسب‌های مرتبط که همان  $\{\lambda_1, \lambda_4\}$  می‌باشند و مجموعه برچسب‌های غیر مرتبط که شامل بقیه برچسب‌ها  $\{\lambda_2, \lambda_3, \lambda_5\}$  می‌باشند تقسیم می‌شوند.

## ۱-۲- مسئله رتبه‌بندی برچسب‌ها

منظور از رتبه‌بندی برچسب‌ها این است مدلی ساخته شود که توانایی رتبه‌بندی مجموعه برچسب‌های از پیش تعریف شده را داشته باشد. برای هر نمونه، این مدل یک خروجی از تمام برچسب‌ها را در مجموعه برچسب‌ها می‌دهد که بر اساس میزان ارتباطشان با نمونه داده شده، مرتب شده‌اند. اگرچه بسیاری از طبقه‌بندهای چند برچسبی قادر به این کار هستند اما رتبه‌بندی یک کار متفاوت است و نیاز به شیوه ارزیابی دارد [۴].

فرض کنیم مجموعه آموزشی دارای ۵ برچسب  $M = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5\}$  می‌باشد و یک نمونه آموزشی مانند  $x$  را در آن در نظر بگیریم که این مدل به هر یک از این برچسب‌ها یک رتبه، مانند موارد زیر در نظر می‌گیرد.

$$r(\lambda_1) < r(\lambda_2) < r(\lambda_3) < r(\lambda_4) < r(\lambda_5)$$

که  $r(\lambda)$  رتبه هر یک از برچسب‌ها برای هر نمونه را مشخص می‌کند.

linkage است. تفاوت روش‌های ذکر شده در نحوه بروزسانی ماتریس شباهت است که کاملاً وابسته به نحوه سنجش شباهت است. به عبارت دیگر تفاوت، در پیدا کردن دو خوشه‌ای است که نزدیکترین شباهت را با هم دارند. ما در این تحقیق از خوشه‌بندی Ward استفاده می‌کنیم چون با استفاده از این روش خوشه‌ها به صورت متراکم و کروی بدست می‌آیند که به طور معنادار این مبنای محاسبه ماتریس کواریانس بهبود می‌بخشد و سپس برای محاسبه فاصله منتهی استفاده می‌کند [۱۶].

در [۱۷] الگوریتمی را فراهم کرده‌اند که به صورت خودکار می‌توان تعداد خوشه‌ها را بدست آورد؛ تعداد خوشه‌ها را متناظر با نقطه شکست انتخاب می‌کند. نقطه شکست<sup>۲۳</sup> نقطه‌ای از نمودار است که در آنجا بیشترین شکست وجود دارد.

بعد از عمل خوشه‌بندی، مجموعه خوشه‌های  $c_1$  و  $c_2$  را به ترتیب برای دو کلاس به دست می‌آوریم. خوشه‌های هر کلاس را به ترتیب با  $P_1, \dots, P_{c_1}$  و  $N_1, \dots, N_{c_2}$  مشخص می‌کنیم. ماشین بردار پشتیبان یک حد پایینی برای جداسازی بین کلاس‌ها توسط قیدها (شرط‌ها) را ارائه می‌دهد. در اینجا توجه بیشتری به تراکم نمونه‌های هر کلاس می‌کنیم. برای مثال خوشه‌ها اطلاعات ساختاری مختلفی در کلاس‌های مختلف را پوشش می‌دهد. هدف ما بیشینه‌سازی حاشیه و همزمان کمینه‌سازی تراکم‌ها است. بنابراین مدل ماشین بردار پشتیبان ساختاری به صورت رابطه (۴) فرموله‌سازی می‌شود:

$$\min_{w, b} \|w\|^2 / 2 + \lambda w^T \sum w / 2 + C \sum_{i=1}^L \varepsilon_i \quad (4)$$

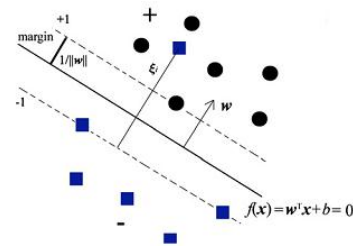
که در آن  $\sum = \sum_{P_1} + \dots + \sum_{P_{c_1}} + \sum_{N_1} + \dots + \sum_{N_{c_2}}$  و  $\sum_{N_i}$  و  $\sum_{P_i}$  به ترتیب ماتریس‌های کواریانس خوشه‌ها در هر کلاس می‌باشند. پارامتر  $\lambda$  نشان دهنده اهمیت اطلاعات ساختاری درون هر کلاس است. با تشکیل معادل لاگرانژ رابطه (۴) و همچنین تشکیل معادلات KKT معادل دوگان<sup>۲۴</sup> رابطه به دست می‌آید که به صورت رابطه (۵) است.

$$\begin{aligned} \max_{\alpha} \sum_{i=1}^L \alpha_i - \frac{1}{2} \sum_{i=1}^L \sum_{j=1}^L \alpha_i \alpha_j y_i y_j [x_i^T (I + \lambda \sum)^{-1} x_j] \\ \text{s.t. } 0 \leq \alpha_i \leq C, \quad i = 1, \dots, L \\ \sum_{i=1}^n \alpha_i y_i = 0 \end{aligned} \quad (5)$$

این رابطه یک مسئله بهینه‌سازی محدب<sup>۲۵</sup> می‌باشد که با تکنیک‌های برنامه‌نویسی درجه دو می‌توان جواب مسئله یعنی  $\alpha$  را به دست آورد. سپس تابع هدف طبقه‌بند استنتاج شده را می‌توان به صورت زیر فرموله‌سازی کرد که برای پیش‌بینی برچسب-های کلاس داده‌های دیده نشده، استفاده می‌شود.

$$\begin{aligned} \min \frac{1}{2} w^T w + C \sum_{i=1}^L \varepsilon_i \\ \text{s.t. } y_i (w^T x_i + b) > 1 - \varepsilon_i, \varepsilon_i > 0 \end{aligned} \quad (2)$$

در حالتی که داده‌ها جداپذیر غیرخطی باشند داده‌ها را از فضای اصلی به فضایی با ابعاد بالاتر نگاشت می‌کنیم به طوری که در فضای جدید داده‌ها جداپذیر خطی باشند، در این صورت می‌توان از ماشین بردار پشتیبان بخش قبلی برای طبقه‌بندی داده‌ها در فضای جدید استفاده کرد و پس از تعیین مرز تصمیم بردارهای پشتیبان مجدداً به فضای اصلی بازگشت. که برای حل این‌گونه مسائل از تکنیک kernel trick [۱۳] استفاده می‌کنیم.



شکل ۱. شمای تصویری از ماشین بردار پشتیبان

## ۱-۲- ماشین بردار پشتیبان ساختاری

اخیراً روش‌هایی برای اینکه بتوانند از اطلاعات ساختاری داده‌ها استفاده کنند طراحی شده‌اند. آنها دید جدیدی در طراحی یک طبقه‌بند ایجاد کرده‌اند که یک طبقه‌بند باید به ساختار توزیع داده‌ها حساس باشند. اخیراً با استفاده از اطلاعات ساختاری داده‌های هر کلاس می‌توان توانایی طبقه‌بند در پیش‌بینی داده‌ها بهبود بخشید که در ماشین بردار پشتیبان شامل بدست آوردن یک مرز تصمیم بهتر است.

ماشین بردار پشتیبان ساختاری<sup>۱۷</sup> [۱۴ و ۱۵] که از دانه‌بندی خوشه‌ای<sup>۱۸</sup> برای مدل خود استفاده می‌کند که این الگوریتم شامل دو گام می‌باشند: خوشه‌بندی و یادگیری. الگوریتم ماشین بردار پشتیبان ساختاری از برخی تکنیک‌های خوشه‌بندی برای ضبط توزیع داده در هر کلاس استفاده می‌کند. در ادامه هر یک از گام‌ها را توضیح می‌دهیم.

گام اول خوشه‌بندی: به منظور بررسی ساختار یک مجموعه، خوشه‌بندی سلسله‌مراتبی<sup>۱۹</sup> برای تشخیص خوشه‌ها در هر کلاس مناسب است. برای مسائلی که داده‌ها جداپذیر خطی باشند به صورت سلسله‌مراتب خوشه‌ها در فضای ورودی خوشه‌بندی می‌شوند و در حالت غیرخطی خوشه‌بندی سلسله‌مراتبی در فضای کرنلی انجام می‌شود.

خوشه‌بندی سلسله‌مراتبی دارای روش‌های نظیر خوشه‌بندی پیوند مجرد<sup>۲۰</sup>، خوشه‌بندی پیوند کامل<sup>۲۱</sup>، پیوند مرکزی<sup>۲۲</sup> و خوشه‌بندی Ward

$$\min \frac{1}{2} \sum_{k=1}^q (w_k^T w_k) + C \sum_{i=1}^l \frac{1}{|L_i| * |L_i|} \sum_{(m,n) \in (L_i * \bar{L}_i)} \varepsilon_{imn} \quad (9)$$

$$s.t. (w_m - w_n)^T x_i + (b_m - b_n) = \sum_{k=1}^q c_{imn}^k (w_k^T x_i + b_k) \geq 1 - \varepsilon_{imn},$$

$$\varepsilon_{imn} \geq 0, (m,n) \in (L_i * \bar{L}_i), i = 1, \dots, l$$

که با تشکیل تابع لاگرانژ و شرایط KKT معادل دوگان رابطه بالا به دست می‌آید.

### ۲-۳- ماشین بردار پشتیبان ساختاری چند برچسبی

در این بخش، ما Rank-SVM را با ماشین بردار پشتیبان ساختاری ترکیب کرده و یک طبقه بند چند برچسبی جدیدی را برای ماشین بردار پشتیبان ایجاد می‌کنیم. الگوریتم جدید دارای همان شرایط و قیود روش Rank-SVM می‌باشد با این تفاوت که بخش جدیدی به تابع هدف اضافه می‌شود که این بخش مجموع ماتریس کواریانس‌های خوشه‌ها می‌باشد. تابع هدف ما از سه بخش تشکیل شده است. ابتدا قیود تابع هدف را مشخص می‌کنیم؛ سپس مقدار ضرایب خطای آموزشی را برای تابع هدف طبقه‌بند خود مشخص می‌کنیم و در نهایت با ترکیب موارد ذکر شده با ماتریس کواریانس خوشه‌های به دست آمده، تابع هدف کلی طبقه‌بند را تشکیل می‌دهیم.

برای قیود تابع هدف مطلوب است که هر برچسب مرتبط باید رتبه‌ای بیشتر از یک برچسب نامرتب بگیرد. بر همین اساس شرط به صورت زیر تعریف می‌شود و مقدار خطای آموزشی برابر با متوسط مجموع ضرایب خطای آموزشی برای هر نمونه است. اکنون رابطه اصلی مسئله ما به صورت زیر خواهد بود:

$$\min \frac{1}{2} \sum_{k=1}^q (w_k^T w_k + \frac{\lambda}{2} w_k^T \Sigma w_k) + C \sum_{i=1}^l \frac{1}{|L_i| * |L_i|} \sum_{(m,n) \in (L_i * \bar{L}_i)} \varepsilon_{imn}$$

$$s.t. (w_m - w_n)^T x_i + (b_m - b_n) = \sum_{k=1}^q c_{imn}^k (w_k^T x_i + b_k) \geq 1 - \varepsilon_{imn}, \quad (10)$$

$$\varepsilon_{imn} \geq 0, (m,n) \in (L_i * \bar{L}_i), i = 1, \dots, l$$

پس از تشکیل معادلات لاگرانژ بالا و شرایط KKT با جایگذاری در رابطه بالا معادل دوگان به صورت زیر خواهد بود:

$$\min \frac{1}{2} \sum_{k=1}^q \sum_{i,j=1}^n \beta_{ki} \beta_{kj} x_i^T (I + \lambda \Sigma)^{-1} x_j - \sum_{i=1}^l \sum_{(m,n) \in (L_i * \bar{L}_i)} \alpha_{imn}$$

$$s.t. 0 < \alpha_{imn} < C_i, \sum_{i=1}^l \sum_{(m,n) \in (L_i * \bar{L}_i)} \alpha_{imn} c_{imn}^k = 0 \quad (11)$$

پس از به دست آوردن معادل دوگان، معادله بالا را می‌توان با استفاده از روش برنامه‌نویسی درجه دو حل نمود. به طور مشابه، معادله بالا را می‌توان در فضای کرنل نیز حل نمود. ما در اجرای تابع هدف طبقه‌بند پیشنهادی خود از حالت حاشیه نرم غیرخطی یعنی استفاده از تکنیک کرنل برای توسعه کارایی طبقه‌بندی خود استفاده می‌کنیم.

$$f(x) = \text{sgn} \left[ \sum_{i=1}^l \alpha_i y_i x_i^T (I + \lambda \Sigma)^{-1} x + b \right] \quad (6)$$

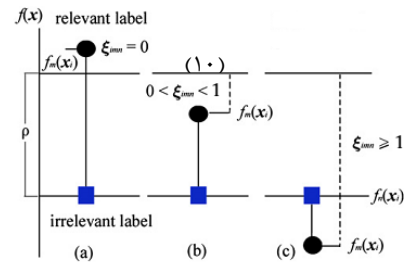
### ۲-۲- ماشین بردار پشتیبان چند برچسبی

اکنون روش Rank-SVM را به اختصار معرفی می‌کنیم. Rank-SVM یک حالت گسترش یافته‌ای از ماشین بردار پشتیبان است که برای طبقه‌بندی داده‌های چند برچسبی طراحی شده است [۱۰ و ۹]. از آن جایی که در مدل‌های رتبه‌بندی، رتبه برچسب‌های مرتبط هر نمونه باید بیشتر از رتبه برچسب‌های غیر مرتبط آن نمونه باشد، Rank-SVM از این ایده برای طبقه‌بندی داده‌های چند برچسبی استفاده می‌کند. پس در این رویکرد رابطه بین هر برچسب مرتبط و برچسب غیر مرتبط برای هر نمونه به صورت رابطه (۷) خواهد بود.

مطلوب است که هر برچسب مرتبط باید رتبه‌ای بیشتر از یک برچسب نامرتب بگیرد. بر همین اساس شرط به صورت زیر تعریف می‌شود.

$$f_m(x_i) - f_n(x_i) = (w_m - w_n)^T x_i + (b_m - b_n) \geq 1 - \varepsilon_{imn}, (m,n) \in (L_i * \bar{L}_i) \quad (7)$$

همان طور که در شکل (۲) مشخص است سه حالت برای مقادیر برچسب مرتبط و برچسب غیر مرتبط برای هر نمونه وجود دارد که در حالت اول که به حالت بدون خطا یا کامل معروف است تفاضل دو مقدار بزرگ‌تر از یک است و مقدار ضریب آموزش در این مورد برابر صفر خواهد بود. در دو مورد دیگر ضریب خطای آموزش نمونه بزرگ‌تر از صفر خواهد بود.



شکل ۲. ارتباط‌های ممکن بین یک جفت برچسب

بر اساس معیار خطای رتبه‌بندی که در بخش بالا بیان شده، مقدار خطای آموزشی برابر با متوسط مجموع ضرایب خطای آموزشی برای هر نمونه می‌باشد که به صورت رابطه (۸) بیان می‌شود.

$$RL = \frac{1}{l} \sum_{i=1}^l \frac{1}{|L_i| * |\bar{L}_i|} \sum_{(m,n) \in (L_i * \bar{L}_i)} \varepsilon_{imn} \quad (8)$$

بر اساس تعریف بالا مسئله بهینه‌سازی اصلی Rank-SVM به صورت رابطه (۹) تعریف می‌شود.

### ۳- ارزیابی و مقایسه نتایج

$$Coverage = \frac{1}{l} \sum_{i=1}^l (|C(x_i)| - 1) \in [0, q-1] \quad (12)$$

در جدول (۲) نتایج مربوط به اجرای روش پیشنهادی و دیگر روش‌ها برای معیار پوشش آورده شده است.

**جدول ۲.** مقایسه نتایج اجرای روش پیشنهادی و با دیگر روش‌ها بر اساس معیار پوشش

مجموعه داده	روش پیشنهادی	Rank-SVM	ADTree	BP-MLL	ML-kNN
Emotions	۱,۸۵۱	۱,۹۶	۲,۲۰۷	۱,۹۵۰	۱,۸۷۶
Image	۰,۹۴	۱,۸۵	۱,۱۰۰	۱,۶۲۷	۰,۹۶۶
Yeast	۶,۲۳۹	۶,۳۸	۶,۵۱۹	۶,۴۲۳	۶,۴۱۴
Plant	۱,۷۵	۱,۸۴	۲,۵۹۷	۵,۹۸۴	۲,۴۲۳

خطای همینگ<sup>۲۷</sup>: درصد برچسب‌هایی که برای یک نمونه درست پیش‌بینی نشده است. در جدول (۳) نتایج مربوط به اجرای روش پیشنهادی و دیگر روش‌ها برای معیار خطای همینگ آورده شده است.

میانگین دقت<sup>۲۸</sup>: این معیار مشخص می‌کند که به طور میانگین چه نسبتی از برچسب‌هایی که رتبه بالاتری از یک برچسب واقعی مانند  $k$  دارند، عضو مجموعه برچسب‌های نمونه هستند. هر چه مقدار این معیار بزرگ‌تر باشد به معنای کارایی بهتر الگوریتم است.

$$A - P = \frac{1}{l} \sum_{i=1}^l \left( \frac{1}{|L_i|} \sum_{k \in L_i} \frac{|\{k' \in L_i \mid f_k^P(x_i) \geq f_{k'}^P(x_i)\}|}{|\{k' \in Q \mid f_k^P(x_i) \geq f_{k'}^P(x_i)\}|} \right) \in [0,1] \quad (13)$$

**جدول ۳.** مقایسه نتایج اجرای روش پیشنهادی و با دیگر روش‌ها بر اساس معیار خطای همینگ

مجموعه داده	روش پیشنهادی	Rank-SVM	ADTree	BP-MLL	ML-kNN
Emotions	۰,۱۸۹	۰,۲۰۳	۰,۲۵۶	۰,۲۱۷	۰,۲۰۸
Image	۰,۱۴۱	۰,۱۶۵	۰,۱۹۱	۰,۲۲۹	۰,۱۷۲
Yeast	۰,۱۹۱	۰,۲۰۳	۰,۲۱۲	۰,۲۰۹	۰,۱۹۸
Plant	۰,۱۱۲	۰,۱۲۲	۰,۰۹۵	۰,۰۹۸	۰,۰۸۶

در جدول (۴) نتایج مربوط به اجرای روش پیشنهادی و دیگر روش‌ها برای معیار میانگین دقت آورده شده است.

**جدول ۴.** مقایسه نتایج اجرای روش پیشنهادی و با دیگر روش‌ها بر اساس معیار میانگین دقت

مجموعه داده	روش پیشنهادی	Rank-SVM	ADTree	BP-MLL	ML-kNN
Emotions	۰,۷۹۵	۰,۷۸۰	۰,۷۳۶	۰,۷۷۹	۰,۷۹۰
Image	۰,۸۰۰	۰,۷۹۱	۰,۷۵۲	۰,۶۳۷	۰,۷۹۱
Yeast	۰,۷۶۱	۰,۷۵۴	۰,۷۳۷	۰,۷۵۰	۰,۷۵۸
Plant	۰,۵۹۱	۰,۵۸۲	۰,۴۸۲	۰,۲۳۲	۰,۵۳۶

خطای رتبه‌بندی<sup>۲۹</sup>: این معیار مشخص می‌کند که به طور میانگین چه تعداد از جفت برچسب‌های مرتبط و نامرتبط از هر نمونه که برچسب‌های

برای ارزیابی روش پیشنهادی، ما چهار طبقه‌بند موجود در حوزه داده‌های چند برچسبی Rank-SVM [۱۰ و ۹]، ADTree [۱۸]، BP-MLL [۱۹] و ML-kNN [۱۲] را انتخاب می‌کنیم و با روش پیشنهادی خود مقایسه می‌کنیم. همچنین شایان ذکر است که سه روش اول و روش پیشنهادی ما متعلق به روش‌های توسعه یا وابسته به الگوریتم می‌باشد که همه نمونه‌های آموزشی و همه برچسب‌ها همزمان در نظر می‌گیرد در حالی که ML-kNN یک روش ترکیبی می‌باشد.

برای مقایسه روش پیشنهادی با چهار طبقه‌بند بیان شده، از مجموعه داده‌های موجود در جدول (۱) که شامل مجموعه داده Emotions، Image، Yeast و Plant می‌باشند، استفاده می‌کنیم. اکثر این مجموعه داده‌های در سال‌های ۲۰۰۷ تا ۲۰۱۰ جمع‌آوری شده است. اما همچنان مجموعه داده‌های موجود در این حوزه بسیار اندک هستند و نیاز به جمع‌آوری مجموعه داده‌های جدید در این حوزه زیاد است.

**جدول ۱.** مجموعه داده‌های چند برچسبی مورد ارزیابی

مجموعه داده	تعداد نمونه‌ها	ویژگی	برچسب
	آموزشی	آزمایشی	
Emotions	۳۹۱	۲۰۲	۷۲
Image	۱۲۰۰	۸۰۰	۲۹۴
Yeast	۱۵۰۰	۹۱۷	۱۰۳
Plant	۵۸۸	۳۹۰	۴۴۰

در این کار، از کرنل RBF با معادله  $k(x,y) = \exp(-\sigma \|x-y\|_2^2)$  برای آزمایش روش خود و روش Rank-SVM استفاده می‌کنیم که  $\sigma$  فاکتور مقیاس کرنل می‌باشد.

در این مقاله برای تمامی مجموعه داده‌ها بهینه‌ترین مقدار پارامترها را با استفاده از روش اعتبار سنجی برای پارامترهای  $C$ ، سیگما و لامبدا به دست می‌آوریم. از آنجا که روش پیشنهادی مبتنی بر رتبه‌بندی برچسب‌ها می‌باشد. بنابراین از بین معیارهای موجود در طبقه‌بندی چند برچسبی معیارهای پوشش، خطای همینگ، میانگین دقت و خطای رتبه‌بندی برای ارزیابی استفاده می‌شوند.

معیار پوشش<sup>۲۶</sup>: این معیار مشخص می‌کند برای اینکه بتوانیم ساده‌تر همه برچسب‌های یک نمونه را پیش‌بینی کنیم چه تعداد برچسب از مجموعه اصلی برچسب‌ها باید بررسی شوند. هر چه مقدار این معیار کوچک‌تر باشد به معنای کارایی بهتر الگوریتم است.



- [2] G. Tsoumakas, I. Katakis, Multi-label classification: an overview, *International Journal of Data Warehousing and Mining* 3 (3) (2007) 1-13.
- [3] A.C.P.L.F. de Carvalho, A.A. Freitas, "A tutorial on multi-label classification techniques," in: A. Abraham, A.E. Hassanien, V. Snasel (Eds.), *Function Approximation and Classification*, Foundations of Computational Intelligence, vol. 5, Springer, Berlin/Heidelberg, 2009, pp. 177-195.
- [4] G. Tsoumakas, I. Katakis, I. Vlahavas, "Mining multi-label data," in: O. Maimon, L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook*, seconded., Springer, New York, 2010, pp. 667-685.
- [5] T. Joachims, "Text categorization with support vector machines: learning with many relevant features," in: *Proceedings of the 10th European Conference on Machine Learning (ECML1998)*, Lecture Notes in Computer Science, vol. 1398, Chemnitz, Germany, 1998, pp. 137-142.
- [6] K.C. Chou, H.B. Shen, Cell-PLoc 2.0: "an improved package of web-servers for predicting subcellular location of proteins in various organisms," *Natural Science* 2 (10) (2010) 1090-1103.
- [7] K. Trohidis, G. Tsoumakas, G. Kalliris, I. Vlahavas, "Multi-label classification of music into emotions," in: *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR2008)*, Philadelphia, PA, USA, 2008, pp. 325-330.
- [8] Xu, Jianhua. "Fast multi-label core vector machine." *Pattern Recognition* (2012)
- [9] Elisseeff, André, and Jason Weston. "A kernel method for multi-labelled classification." *Advances in neural information processing systems*. 2001.
- [10] Elisseeff, Andre, and Jason Weston. "Kernel methods for multi-labelled classification and categorical regression problems." *Advances in neural information processing systems* 14 (2002): 681-687.
- [11] Schapire, Robert E., and Yoram Singer. "BoosTexter: A boosting-based system for text categorization." *Machine learning* 39.2-3 (2000): 135-168.
- [12] Zhang, Min-Ling, and Zhi-Hua Zhou. "ML-KNN: A lazy learning approach to multi-label learning." *Pattern*
- [13] Cortes, Corinna, and Vladimir Vapnik. "Support vector machine." *Machine learning* 20.3 (1995): 273-297.
- [14] Xue, Hui, Songcan Chen, and Qiang Yang. "Structural regularized support vector machine: a framework for structural large margin classifier." *Neural Networks, IEEE Transactions on* 22.4 (2011): 573-587.
- [15] Xue, Hui, Songcan Chen, and Qiang Yang. "Structural support vector machine." *Advances in Neural Networks-ISNN 2008*. Springer Berlin Heidelberg, 2008. 501-511.
- [16] A.K. Jain and R. Dubes. "Algorithms for Clustering Data." New Jersey: Prentice Hall, 1988.
- [17] Salvador, Stan, and Philip Chan. "Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms." *Tools with Artificial Intelligence*, 2004. ICTAI 2004. 16th IEEE International Conference on. IEEE, 2004.
- [18] A. Clare, R.D. King, "Knowledge discovery in multi-label phenotype data," in: *Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD2001)*, Lecture Notes in Computer Science, vol. 2168, Freiburg, Baden-Württemberg, Germany,
- [19] M.L. Zhang, Z.H. Zhou, "Multilabel neural networks with application to function genomics and text categorization", *IEEE Transactions on Knowledge and Data Engineering* 18 (10) (2006) 1338-1351.2001, pp. 42-53.
- [20] M.L. Zhang, Matlab software of BP-MLL and ML-kNN, and Image data set, 2009 <http://cse.seu.edu.cn/people/zhangmls>

نامرتبط رتبه بالاتری از بر چسب‌های مرتبط گرفته‌اند. هر چه مقدار این معیار کوچک‌تر باشد به معنای کارایی بهتر الگوریتم است.

$$R - L = \frac{1}{l} \sum_{i=1}^l \left( \frac{1}{|L_i| * |L_i|} \left| \{ (k, k') \in (L_i * \bar{L}_i) \mid f_k^P(x_i) \geq f_{k'}^P(x_i) \} \right| \right)$$

در جدول (۵) نتایج مربوط به اجرای روش پیشنهادی و دیگر روش‌ها برای معیار خطای رتبه‌بندی آورده شده است.

**جدول ۵.** مقایسه نتایج اجرای روش پیشنهادی و با دیگر روش‌ها بر اساس معیار

خطای رتبه‌بندی

مجموعه داده	روش پیشنهادی	Rank-SVM	ADTree	BP-MLL	ML-kNN
Emotions	۰,۱۷۵	۰,۱۶۲	۰,۲۲۸	۰,۱۷۹	۰,۱۶۱
Image	۰,۱۶۰	۰,۱۶۵	۰,۲۱۲	۰,۳۳۹	۰,۱۷۵
Yeast	۰,۱۷۰	۰,۱۷۱	۰,۱۸۷	۰,۱۷۵	۰,۱۷۱
Plant	۰,۱۵۱	۰,۱۸	۰,۲۲۸	۰,۵۲۵	۰,۲۱۰

#### ۴- نتیجه‌گیری

برای طبقه‌بندی داده‌های چند برچسبی، تقریباً همه محققان هدفشان کاهش زمان اجرای الگوریتم و بهبود کارایی روش است. این دو با هم در تضاد هستند که روش‌های تجزیه داده و وابسته به الگوریتم دارای چنین شرایطی هستند. در این مقاله، ما یک روش وابسته به الگوریتم را برای طراحی و پیاده‌سازی یک ماشین بردار پشتیبان جدید را برای طبقه‌بندی داده‌های چند برچسبی بکار برده‌ایم.

در این ماشین بردار پشتیبان جدید قبل از یادگیری، ابتدا داده‌های هر کلاس را خوشه‌بندی می‌کنیم سپس یک بخش جدید به تابع هدف ماشین بردار پشتیبان اضافه می‌کنیم که این بخش شامل ماتریس کواریانس داده‌ها می‌باشد که این ماتریس کواریانس، مجموع ماتریس کواریانس‌های خوشه‌های هر کلاس است. نتیجه آزمایشات نشان داده است که این روش تاثیر زیادی در بهبود کارایی طبقه‌بندی داده‌های چند برچسبی دارد. البته بهترین نتایج در مواردی که مقدار  $C=1$  و  $\sigma=0.5$  باشد بدست می‌آید.

#### مراجع

- [1] J. Read, "Scalable Multi-label Classification," Ph.D. These, Department of Computer Science, University of Waikato, 2010

- <sup>۱</sup> Multi-label classification
- <sup>۲</sup> Text Categorization
- <sup>۳</sup> Bioinformatics
- <sup>۴</sup> Music Emotion Categorization
- <sup>۵</sup> Data decomposition
- <sup>۶</sup> Algorithm extension
- <sup>۷</sup> Hybrid
- <sup>۸</sup> Clustering
- <sup>۹</sup> Quadratic programming
- <sup>۱۰</sup> Thresholds
- <sup>۱۱</sup> Support Vector Machine
- <sup>۱۲</sup> Decision Boundary
- <sup>۱۳</sup> Hyperplane
- <sup>۱۴</sup> Margin
- <sup>۱۵</sup> Constrained optimization problem
- <sup>۱۶</sup> Quadratic programming
- <sup>۱۷</sup> Structural Regularized Support Vector Machine
- <sup>۱۸</sup> Cluster granularity
- <sup>۱۹</sup> Hierarchical clustering techniques
- <sup>۲۰</sup> Single linkage
- <sup>۲۱</sup> Complete linkage
- <sup>۲۲</sup> Centroid linkage
- <sup>۲۳</sup> Knee point
- <sup>۲۴</sup> Dual problem
- <sup>۲۵</sup> Convex problems
- <sup>۲۶</sup> Coverage
- <sup>۲۷</sup> Hamming loss
- <sup>۲۸</sup> Average precision
- <sup>۲۹</sup> Ranking loss