یازدهمین کنفرانس سیستم های هوشمند ایران
۹ و ۱۰ اسفند ۱۳۹۱
(ICIS2013)
11th Iranian Conference on Intelligent Systems
February 27th & 28th, 2013

دسته بندی داده‌های چند برچسبی با استفاده از وابستگی بین کلاس‌ها و تخمین تعداد برچسب‌ها

سجاد کمالی[۱]، حمید بیگی[۲]
[۱]دانشجوی ارشد دانشگاه شریف، skamali@ce.sharif.edu
[۲]استاددانشگاه شریف ، beigy@sharif.edu

چکیده - با افزایش حجم داده‌ها نیاز به دسته‌بندی و تحلیل داده‌ها به صورت خودکار، از جایگاه ویژه‌ای بر خوردار شده است. در این مقاله به دسته‌بندی داده‌های چندبرچسبی خواهیم پرداخت. داده‌های چند برچسبی داده‌هایی هستند که در آن‌ها نمونه‌ها می‌توانند بیش از یک برچسب داشته باشند. به عبارت دیگر هر نمونه توسط یک بردار از برچسب‌ها نمایش داده می‌شود.

در این مقاله یک روش مبتنی بر نزدیک‌ترین همسایگی برای دسته‌بندی داده‌های چندبرچسبی پیشنهاد شده است. در روش ارائه شده نخست تعداد برچسب‌ها برای هر نمونه جدید تخمین زده می‌شود و سپس برای هر نمونه جدید، با استفاده از ارتباط و وابستگی بین برچسب‌ها، برچسب‌های آن را پیش‌بینی می‌گردد. در این مقاله از دسته‌بند نزدیک‌ترین همسایگی به عنوان دسته‌بند پایه استفاده شده است اما با این تفاوت که برای محاسبه فاصله بین نمونه‌ها از معیار فاصله اقلیدسی استفاده نمی‌شود. با ارزیابی روش پشنهادی بر روی مجموعه‌داده‌های موجود در حوزه داده‌های چندبرچسبی و مقایسه آن با دیگر روش‌های مطرح در این زمینه، نشان داده می‌شود که روش پیشنهادی در مقایسه با روش‌های دیگر در پیش‌بینی برچسب‌های نمونه‌های جدید، از لحاظ پیچیدگی زمانی و دقت از کارایی بهتری برخوردار است.

کلید واژه-داده‌های چندبرچسبی ، دسته‌بندی داده‌ها، الگوریتم‌های دسته‌بندی

۱-مقدمه

در دسته بندی تک برچسبی هر نمونه با یک برچسب در ارتباط می‌باشد که این برچسب ویژگی‌های آن نمونه را مشخص می‌کند. در دسته بندی چند برچسبی هر نمونه ممکن است با چندین برچسب در ارتباط باشد و اجتماع این برچسب‌ها ویژگی‌های نمونه‌ها را مشخص می‌کنند به عبارت دیگر در دسته‌بندی چندبرچسبی هر نمونه توسط برداری از برچسب‌ها مشخص می‌شود. در بساری از کاربردهای واقعی نیز داده‌ها چند برچسبی می‌باشند. به عنوان مثال یک فیلم سینمایی ممکن است مربوط به چندین ژانر مختلف باشد. و یا یک خبر مربوط به چندین حوزه مختلف باشد. معمولا برای حل این مسائل از راهکارهای تک برچسبی استفاده می شود که اغلب، این راهکارها،باعث افزایش هزنیه‌های نگه داری و بازیابی می‌شود و دارای پیچیدگی زمانی زیادی نیز هستند. با افزایش تعداد برچسب ها یادگیری داده‌های چند برچسبی دچار چالش می شود زیرا با این افزایش، تعداد حالت‌های ممکن برای مجموعه برچسب‌های متفاوت به صورت نمایی افزایش می یابد و نمی‌توان از الگوریتم‌های موجود در داده‌های تک برچسبی برای حل این مسائل استفاده کرد. زیرا استفاده از دسته‌بندهای داده‌های تک برچسبی باعث پیچیدگی زمانی زیادی می شود.روش‌های یادگیری چند برچسبی سعی در کاهش این هزینه‌ها دارند.از طرفی برخی از کاربردها همانند توابع تاثیر ژنها، ذاتا چند برچسبی هستند و حتی با هزنیه‌های بالا هم نمی‌توان آنها را به مسئله تک برچسبی تبدیلn کردونیاز به یادگیری چند برچسبی برای این مسائل ضروری است. در یادگیری چند برچسبی هدف یافتن تابع نگاشتی است که فضای ویژگی‌های را به فضای مجموعه برچسب‌ها نگاشت کند[1].

یک رویکرد مناسب برای حل این مشکل استفاده از ارتباط و وابستگی بین برچسب می‌باشد.

در ادامه این مقاله در بخش ۲ مروری بر پژوهش‌های پیشین در حوزه یادگیری چند برچسبی خواهیم داشت در بخش ۳ روش پیشنهادی ارائه می‌شود در بخش ۴ به ارزیابی و مقایسه نتایج حاصل از پیاده سازی روش پیشنهادی می‌پردازیم و در بخش ۵ نتایج و جمع بندی ارائه می شود.

۲-مروری بر پژوهش‌های پیشین

انواع پژوهش‌های انجام شده در حوزه یادگیری داده‌های چندبرچسبی به سه دسته کلی روش‌های تطبیق الگوریتم ، روش‌های تغییر مسئلهو روش‌های جمعی تقسیم بندی شده است[2].

۲-۱-روش‌های تغییر مسئله

این روشها تلاش می‌کنند که داده‌های چندبرچسبی را به گونه‌ای تغییر دهند که این داده‌ها تبدیل به داده‌های تک‌برچسبی شوند. سپس با استفاده از روش‌ها و الگوریتم‌های موجود در حوزه یادگیری تک‌برچسبی کار یادگیری و پیش‌بینی برچسب‌ها را برای این داده‌ها انجام می‌شود. که برخی از مهم‌ترین روش ها در این حوزه عبارتند از:

روش ارتباط دودویی(BR):در این روش به ازای تعداد برچسب‌ها، دسته‌بند ساخته می‌شود. و یادگیری و پیش بینی برای هر کدام از این دسته‌بندها به صورت جداگانه انجام می‌شود.یعنی به ازای هر برچسب یک دسته‌بند جداگانه خواهیم داشت.

روش مجموعه توانی(LP):در این روش به ازای هر زیر مجموعه‌ای از مجموعه توانی برچسب‌های موجود، یک برچسب جداگانه در نظر گرفته می‌شود. برای نمونه اگر L برچسب داشته باشیم 2^L مجموعه برچسب خواهیم داشت که هر کدام به عنوان یک برچسب مجزا در نظر گرفته می‌شود و به این ترتیب داده‌های چندبرچسبی به داده‌های تک برچسبی تبدیل می‌شوند [۳].

روش فاصله جفتی:در این روش به ازای هر جفت برچسب یک دسته‌بند ساخته می‌شود. یعنی اگر L برچسب داشته باشیم L*(L-1)/2 دسته‌بند ساخته می‌شود. هر دسته‌بند برای مقایسه دو برچسب به کار می‌رود. برای داده‌های آزمایشی هر کدام از این دسته‌بندها یکی از دو برچسب را به عنوان برچسب آن‌ها پیش‌بینی می‌کند در پایان برای ترکیب این دسته‌بندها از روش رای‌گیری و رتبه‌بندی استفاده می‌شود. یک روش جدید موثر برای رای‌گیری بر اساس روش وزن دهی Qweighted به نام QWML برای روش فاصله جفتی ارائه شده است[۴].

۲-۲-روش‌های تطبیق الگوریتم

در این دسته از روش‌ها تلاش می‌شود که الگوریتم‌های موجود در دسته‌بندی داده‌های تک برچسبی به گونه‌ای تغییر داده شوند که توانایی دسته‌بندی داده‌های چندبرچسبی را داشته باشند.

روش درخت تصمیم: در این روش سعی شده الگوریتم C4.5 به گونه تغییر داده شود تا این الگوریتم برای دسته‌بندی داده‌های چندبرچسبی هم مناسب باشد. در الگوریتم C4.5ML- رابطه موجود در الگوریتم C4.5 برای محاسبه اینتروپی تغییر یافته و به صورت رابطه (۱) محاسبه می‌شود[۵].

$$entropy(E) = -\sum_{i=1}^{N}\left(P((c_i)log(c_i) + q(c_i)logq(c_i)\right) \qquad (1)$$

منظور از E مجموعه نمونه‌های آموزشی و منظور از نسبت تکرارهای برچسب i ام به مجموع تکرارهای کل برچسب‌ها می باشد و است.

روش نزدیک‌ترین همسایگی:در این روش ابتدا با استفاده از فاصله اقلیدسی نزدیکترین همسایه‌های نمونه جدید پیدا می‌شود سپس در بین این همسایه‌ها احتمال این که نمونه جدید هر کدام از برچسب‌ها را بگیرد محاسبه می شود. سپس این احتمالات را به عنوان احتمال اولیه اینکه یک نمونه جدید با چه کلاس‌هایی در ارتباط است به یک دسته بند بیز داده می شود و برچسبهای خروجی دسته بند بیزین که احتمال بیش ۰.۵ دارند به عنوان برچسب های نمونه جدید درنظر گرفته می شوند[۶].

روش HOMER:این‌روش یک سلسله مراتب از دسته‌بندها می‌سازد؛که هر کدام از این دسته‌بندها با یک مجموعه کوچک از برچسب‌ها (نسبت به کل برچسب‌ها) در ارتباط هستند. ایده اصلی این است که الگوریتم دسته‌بندی داده‌های چند برچسبی، در مجموعه‌داده‌های با برچسب‌های زیاد، به شکل یک درخت سلسله مراتبی تبدیل شوند. در این درخت هر گرهشامل

مجموعه ای از برچسب‌ها می‌باشد. این درخت |L| برگ دارد که هر کدام از این برگ‌ها شامل یک مجموعه مولفه با برچسب جدا می‌باشد{λj}و هر گره داخلی شامل مجموعه برچسب‌هایی است که از اجتماع برچسب‌های فرزندانش بدست می‌آید و ریشه شامل همه برچسب‌ها می‌باشد. هر گره داخلی از سلسله مراتب نیز شامل یک دسته‌بند چند برچسبی(h_n) می‌باشد. کار این دسته‌بند این است که یک یا چند ابر برچسب برای فرزندانش پیش‌بینی کند[۷].

۲-۳- روش‌های جمعی

در دسته‌بندی داده‌ها به کمک روش‌های جمعی، ابتدا نمونه‌ها را توسط چندین دسته‌بند جداگانه، دسته‌بندی می‌کنند. سپس برای مشخص کردن برچسب نهایی نمونه‌ها، برای هر نمونه، بین دسته‌بندها رای‌گیری انجام می‌دهند.

روشK مجموعه برچسب تصادفی(RAKEL):در این روش ابتدا به صورت تصادفی مجموعه برچسب‌های اولیه به K مجموعه برچسب مساوی افراز می‌شود. سپس برای هر کدام از این مجموعه برچسب‌های ایجاد شده، به صورت جداگانه دسته‌بندی انجام می‌شود. یعنی برای هر کدام از این مجموعه برچسب‌ها، تمام مجموعه توانی آن‌ها را ساخته و کار دسته‌بندی انجام می‌شود. در پایان برای مشخص کردن خروجی نهایی یک نمونه جدید، بین خروجی‌های این K دسته‌بند رای گیری انجام می‌شود[۸].

مدل دسته‌بندی زنجیره‌ای(CC):در مدل دسته‌بندی زنجیره‌ای L دسته‌بند دودویی وجود دارد. که هر دسته‌بند دودویی برای یک برچسب است و این دسته‌بندها با هم یک زنجیر را می‌سازنند. در این زنجیر هر دسته‌بند مربوط به یکی از برچسب‌های موجود در مجموعه برچسب‌ها است[۹].

۳-روش پیشنهادی

هدف نهایی در دسته بندی داده‌های چندبرچسبی ارائه الگوریتم‌هایی است که توانایی پیش‌بینی برچسب‌های نمونه‌های جدید را داشته باشند. بنابراین لازم است الگوریتم‌های ارائه شده در این راستا سه مسئله مهم زیر را در نظر بگیرند:

۱. مشخص کردن احتمال حضور هر برچسب برای هر نمونه

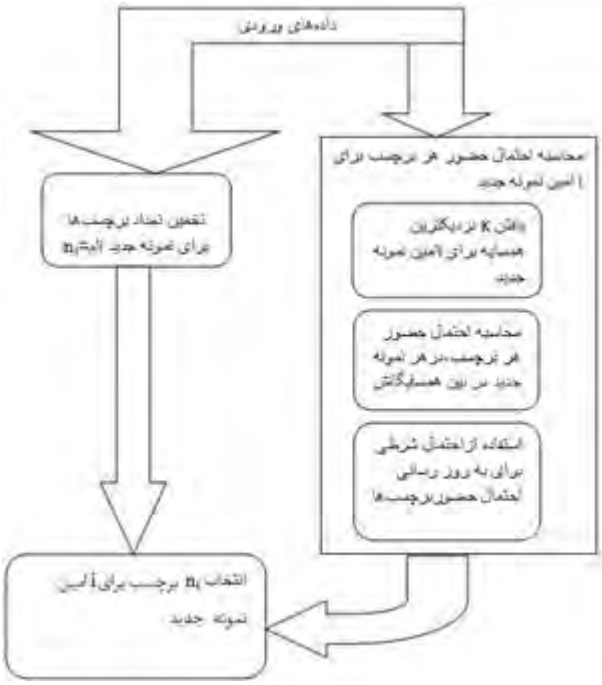
۲. تعیین کردن تعداد برچسب‌های هر نمونه

۳. هزینه محاسباتی پایین

روش‌های تغییر مسئله (همانند روش ارتباط دودویی، روش مجموعه توانی و روش فاصله جفتی) نسبت به روش‌های دیگر در محاسبه احتمال حضور هر برچسب برای نمونه‌های جدید موفق‌تر هستند اما کاستی بزرگ این روش‌ها هزینه زمانی بسیار بالای آنهاست و زمانی که حجم مجموعه داده‌ها یا تعداد برچسب‌ها زیاد شود، استفاده از این روش‌ها در عمل غیر ممکن می‌شود. از طرفی این روش‌ها برای مشخص کردن تعداد برچسب‌های نمونه‌های جدید هیچ راه‌حلی ارائه نمی‌دهند و معمولا تعداد برچسب‌های نمونه‌های جدید را به اشتباه تشخیص می‌دهند.

روش‌های تطبیق الگوریتم‌ها (همانند روش درخت تصمیم، روش نزدیکترین همسایگی، روش HOMERو..) نسبت به دیگر روش‌ها اغلب دارای پیچیدگی زمانی پایین‌تری هستند و با بزرگ شدن حجم مجموعه داده‌ها یا تعداد برچسب‌ها دچار چالش جدی نمی‌شوند. اما این روش‌ها نسبت به دیگر روش‌ها دارای دقت پایین‌تری در پیش‌بینی برچسب‌ها هستند این روش‌ها نیز همانند روش‌های تغییر مسئله هیچ راهکاری برای مشخص کردن تعداد برچسب‌های نمونه‌های جدید ارائه نمی‌دهند. این روش‌ها معمولادر تشخیص تعداد برچسب‌های نمونه‌های جدید دچار مشکل می‌شوند.

روش‌های جمعی رفتارهای متفاوتی دارند، برخی روش‌های جمعی (همانند روش RAKEL) با وجود اینکه دارای پیچیدگی زمانی پایینی هستند اما در تشخیص احتمال برچسب‌های هر نمونه، دقت کافی ندارند و بسیار حساس به پارامترهایشان هستند. برخی دیگر از روش‌های جمعی (همانند روش دسته‌بندی زنجیره‌ای) مانند روش‌های تغییر مسئله دارای پیچیدگی زمانی زیادی هستند و زمانی که حجم مجموعه داده‌ها یا تعداد برچسب‌ها زیاد شود استفاده از این روش‌ها در عمل غیر ممکن می‌شود.در راستای حل مسئله دسته‌بندی داده‌های چند برچسبی، ساختار شکل(۱) برای حل مساله بیان می‌شود که در ادامه هر کدام از زیر بخش‌ها با جزئیات بیش‌تر بیان می‌شود.



شکل ۱: ساختار روش پیشنهادی

۳-۱- تخمین تعداد برچسب‌های هر نمونه

بسیاری از الگوریتم‌های ارائه شده برای دسته بندی داده‌های چند برچسبی، اغلب پس از مشخص کردن احتمال حضور هر برچسب در هر نمونه جدید i، از یک تابع آستانگی برای تعیین برچسب‌های نهایی استفاده می‌نمایند. این تابع آستانگی معمولا به صورت رابطه (۲) تعریف می‌شود.

که منظور از PL_i^j احتمال حضور برچسب iام برای نمونه جدید jام و منظور از L_i^j برچسب iام برای نمونه jام است.

میزان دقت و کارایی الگوریتم‌هایی که از رابطه (۳-۱) برای تعیین تعداد برچسب‌های نهایی نمونه‌های جدید استفاده می‌کنند، به میزان زیادی به نحوه انتخاب متغیر آستانگی a وابسته است. به عنوان نمونه در الگوریتم نزدیکترین همسایگی a=۰.۵ در نظر گرفته می شود.

به منظور مدیریت چالش مذکور برای تخمین تعداد برچسب‌های نمونه‌های جدید، تخمین تعداد برچسب‌های یک نمونه جدید قبل از دسته‌بندی به عنوان یک فاز اولیه در نظر گرفته می‌شود و در آن تخمین تعداد برچسب‌ها انجام می‌پذیرد.

۳-۲- استفاده از احتمال شرطی برچسب‌ها

برای اینکه درپیش‌بینی برچسب‌های نمونه‌های جدید کارایی بهتری حاصل شود می‌توان از ارتباط و وابستگی بین برچسب‌ها استفاده کرد. اگر بدانیم که یک نمونه حتما یک برچسب خاص را می‌گیرد می‌توان از وابستگی بین این برچسب و دیگر برچسب‌ها برای دسته بندی داده‌ها استفاده کرد. به عنوان نمونه، فرض کنید برچسب L_1 جزء برچسب‌های واقعی یک نمونه است. اکنون اگر بدانیم که برچسب L_2 در هیچ یک از مثال‌های آموزشی به همراه برچسب L_1 نیامده است و برچسب L_3 همواره با برچسب L_1 رخ داده است می‌توان گفت، اگر نمونه‌ای برچسب L_1 را داشته باشد به احتمال بالایی برچسب L_3 را دارد و برچسب L_2 را ندارد. رابطه (۳) احتمال شرطی مربوط به برچسب‌های L_1 و L_2 را نشان می‌دهد.

$$P(L_1|L_2) = 0 \quad (۳)$$

رابطه (۴) احتمال شرطی مربوط به برچسب‌های L_1 و L_3 را نشان می‌دهد.

$$P(L_1|L_3) = 1 \quad (۴)$$

نکته اصلی برای استفاده از وابستگی بین برچسب‌ها در دسته‌بندی داده‌های چندبرچسبی این است که ابتدا برای یک نمونه جدید برچسبی یافته شود که با احتمال زیادی مربوط به آن نمونه باشد. سپس برای بقیه برچسب‌ها که احتمال کمتری دارند از احتمال شرطی استفاده شود و در پایان کار دسته‌بندی انجام شود. برای یافتن احتمال هر برچسب برای هر نمونه جدید از روش نزدیک‌ترین همسایگی استفاده خواهد شد.

برای هر نمونه جدید ابتدا با استفاده از روش نزدیکترین همسایگی، K نزدیکترین همسایه آن پیدا شده سپس در بین برچسب‌های این K همسایه، احتمال وقوع هر برچسب محاسبه می‌شود. یعنی برای هر برچسب تعداد رخداد آن برچسب در بین K همسایه، بر مجموع رخدادهای کل برچسب ها در K همسایه تقسیم می‌شود. سپس برای هر نمونه، برچسبی که بیش‌ترین تکرار را داشته باشد به عنوان برچسب اصلی این نمونه در نظر گرفته شده و احتمالات بدست آمده برای برچسب‌های دیگر از روش نزدیکترین همسایگی با استفاده از احتمال شرطی این برچسب و دیگر برچسب‌ها به روز رسانی می‌شود.

۳-۳- تغییر معیار فاصله اقلیدسی

در روش دسته‌بندی داده‌های چندبرچسبی به روش K نزدیکترین همسایگی، ابتدا نزدیکترین همسایه‌ها در بین مثال‌های آموزشی برای نمونه جدید پیدا شده و سپس احتمال وقوع هر برچسب بین k همسایه برای نمونه جدید به عنوان احتمال اولیه بدست می‌آید. منظور از فاصله بین دو نمونه، این است که ویژگی‌های این دو نمونه به چه میزان به یکدیگر شبیه هستند. فاصله اقلیدسی از رابطه (۵) برای محاسبه فاصله میان دو نمونه X,Y که هریک سه ویژگی با مقادیر x_1,x_2,x_3 ,y_1,y_2,y_3 استفاده می‌کند.

$$\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+(x_3-y_3)^2} \quad (۵)$$

اگر از فاصله اقلیدسی برای یافتن نزدیکترین همسایه برای نمونه جدید موجود در جدول (۱) استفاده شود مشاهده می‌شود که معیار فاصله اقلیدسی فاصله هر سه نمونه را با نمونه آزمایشی برابر بدست می‌آورد.

جدول ۱: نمونه‌های چند برچسبی

نمونه	ویژگی				برچسب			
	F1	F2	F3	F4	L1	L2	L3	L4
نمونه اول	۱	۰	۱	۱	۱	۰	۱	۱
نمونه دوم	۰	۱	۱	۱	۱	۱	۰	۱
نمونه سوم	۰	۰	۱	۰	۱	۰	۰	۰
نمونه آزمایشی	۰	۰	۱	۱	?	?	?	?

دلیل این موضوع این است که فاصله اقلیدسی دو نمونه‌آموزشی اول، که نسبت به نمونه جدید (نمونه آزمایشی) یک ویژگی اضافی دارند و نمونه آموزشی سوم که نسبت به نمونه جدید یک ویژگی کم دارد برابر است.

نمونه‌های اول و دوم تمام ویژگی‌هایی که نمونه آزمایشی دارا می‌باشد را دارند و نیز یک ویژگی بیشتر دارند ولی نمونهسومیکی از ویژگی‌هایی که نمونه آزمایشی دارد، را ندارد. در دسته‌بندی داده‌های تک برچسبی این سه نمونه فاصله برابر با نمونه آزمایشی دارند، چون قرار است در نهایت یک برچسب برای نمونه جدید پیش‌بینی شود، مشکلی بوجود نمی‌آید. اما در داده‌های چندبرچسبی تا حدودی شرایط متفاوت است. دلیل آن هم این است که در یک داده چندبرچسبی ویژگی‌های موجود در بردار ویژگی‌های هر نمونه مربوط به چندین برچسب هستند بنابراین یک نمونه با چند برچسب دارای اجتماعی از ویژگی‌هایی است که هر برچسب دارا است. به عبارت دیگر وجود هر برچسب در مجموعه برچسب‌های یک نمونه موجب می‌شود آن نمونه دارای تعدادی از ویژگی‌ها شود. بنابراین در جدول (۳-۱) می‌توان گفت (به صورت نادقیق) نمونه‌های اول و دوم تمام برچسب‌های نمونهآزمایشی را دارند به اضافه چند برچسب اضافی (که باعث شده است ویژگی‌های بیشتری را داشته باشند) و نمونه سوم چند مورد از برچسب‌های موجود در مجموعه برچسب‌های نمونه آزمایشی را ندارد. پس می‌توان گفت نمونه‌های اول و دوم نسبت به نمونه سوم به نمونه آزمایشی نزدیک‌تر هستند و فاصله‌ها یکسان نیست.

برای حل این مشکل، برای محاسبه فاصله نمونه آزمایشی با هریک از نمونه‌های آموزشی، به جای استفاده از فاصله اقلیدسی برای محاسبه فاصله، بردار ویژگی‌های نمونه‌ها با هم مقایسه می‌شوند. به عنوان نمونه فرض کنید یک نمونه جدید، ویژگی f1 را دارد و ویژگی f2 را ندارد. برای پیدا کردن نزدیکترین همسایه‌های این نمونه جدید در بین داده‌های آموزشی، به این صورت عمل می‌شود.

برای هریک از نمونه‌های آموزشی یک متغیر امتیاز تعریف می‌شود. که مقدار این امتیاز برابر است با مجموع امتیازهایی که هر یک از ویژگی‌های نمونه آموزشی به دست آورده‌اند.

اگر یک نمونه از داده‌های آموزشی ویژگی f1 را دارا باشد به متغیر امتیاز آن مقدار α اضافه می‌شود. اگر یک نمونه از داده‌های آموزشی یک ویژگی‌ای که نمونه جدید دارد، را دارا باشد به این معنی است که در این ویژگی، این نمونه آموزشی با نمونه جدید همسو است. بنابراین یک امتیاز مثبت می‌گیرد که نشان دهنده این است که در این ویژگی این نمونه آزمایشی به نمونه جدید نزدیک است.

اگر یک نمونه از داده‌های آموزشی ویژگی f2 را نداشته باشد به متغیر امتیاز آن مقدار β($\beta < \alpha$) اضافه می‌شود. اگر یک نمونه از داده‌های آموزشی یک ویژگی‌ای که نمونه جدید ندارد را نداشته باشد به این معنی است که این دو نمونه به یکدیگر نزدیک هستند ولی نه به اندازه حالت قبل بنابراین امتیازی پایین‌تر از

حالت قبل می‌گیرد.

اگر یک نمونه از داده‌های آموزشی ویژگی f2 را داشته باشد به متغیر امتیاز آن مقدار γ ($\gamma<\beta$) اضافه می‌شود. اگر نمونه جدید یک ویژگی را نداشته باشد ولی نمونه‌های آموزشی داشته باشند این بدان معنی است که این نمونه آموزشی با نمونه جدید همسو نیست. بنابراین امتیاز کمتری خواهد گرفت (امتیاز منفی) تا مشخص کننده این عدم تطابق باشد. در واقع می‌توان فرض کرد که نمونه آموزشی، برچسبی را دارا است که جز برچسب‌های اصلی نمونه جدید نیست و به خاطر این برچسب، نمونه آموزشی ویژگی که این نمونه دارا نیست را دارد.

اگر یک نمونه از داده‌های آموزشی ویژگی f1 را نداشته باشد به متغیر امتیاز آن مقدار θ ($\theta<\gamma$) اضافه می‌شود. بیشترین فاصله بین یک نمونه جدید و نمونه‌های آموزشی در یک ویژگی، زمانی حاصل می‌شود که یک نمونه آموزشی ویژگی‌ای که نمونه جدید دارد را نداشته باشد. این حالت عکس حالت قبل است، که می‌توان فرض کرد در این حالت، نمونه جدید یک برچسبی را دارد که نمونه آموزشی ندارد و به دلیل وجود این برچسب، نمونه جدید این ویژگی را دارد و نمونه آموزشی این ویژگی را ندارد.

در پایان K تا از نمونه‌های آموزشی که بیش‌ترین امتیاز را دارند به عنوان همسایه‌های نمونه جدید مشخص می‌شوند.

۴- ارزیابی و مقایسه نتایج

برای ارزیابی روش پیشنهادی از مجموعه‌داده‌های موجود در جدول (۲) که شاملسه مجموعه‌داده enron,bibtex,medical می‌باشند،استفاده می‌کنیم.

جدول ۲:مجموعه داده‌های مورد ارزیابی

مجموعه داده‌ها	ویژگی‌ها	برچسب‌ها	نمونه‌ها
Medical	1449b	45	978
Enron	1001b	53	1702
Bibtex	1836b	159	7393

از آنجا که روش پیشنهادی مبتنی بر رتبه‌بندی برچسب‌ها می‌باشد. بنابراین از بین معیارهای موجود در حوزه دسته بندی داده‌های چند برچسبی معیارهای پوشش، خطای همینگ و میانگین دقتکه در روش‌های مبتنی بر رتبه‌بندی استفاده می‌شوند[۲۹] برای ارزیابی استفاده شده‌اند.

معیار پوشش: این معیار مشخص می‌کند که به طور میانگین در بردار احتمال برچسب‌ها (که به صورت نزولی مرتب هستند) چه تعداد برچسب (چه تعداد گام) باید بررسی شود تا برچسب‌های صحیح نمونه جدید مشخص گردد. هرچه مقدار این معیار کوچک‌تر باشد به معنای کارایی بهتر در الگوریتم پیش‌بینی برچسب‌ها است. رابطه ریاضی این معیار در رابطه (۶) آورده شده است.

$$coverage(f)=\frac{1}{N}\sum_{i=1}^{N} max_{l_i\in Y}\, rank_f(x_i,l)-1 \qquad (۶)$$

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

جدول ۳: مقایسه نتایج اجرای روش پیشنهادی و با دیگر روش‌ها بر اساس معیارcoverage.

Data Set	BR	CC	HOMER	MLC4.5	MLKNN	RAKEL	**Proposed method**
Medical	1.61	1.47	5.32	3.05	2.84	8.52	**2.80**
Enron	12.5	12.4	24.2	17.0	13.1	.5	**13.9**
Bibtex	2.9	21.1	65.6	58.0	56.2	DNF	**43.6**

معیار میانگین دقت: این معیار مشخص می‌کند که به طور میانگین چه نسبتی از برچسب‌هایی که رتبه بالاتر از یک برچسب واقعی $l\in y_i$ دارند عضو y_i هستند. بنابراین هرچه مقدار این معیار بزرگتر باشد به معنای کارایی بهتر در الگوریتم پیش‌بینی برچسب‌ها است. رابطه این معیار در رابطه (۷) آورده شده است.

$$average\ precision(f)=\frac{1}{N}\sum_{l\in y_i}\frac{L_i}{rank_f(x_i,l)} \qquad (۷)$$

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

جدول ۴ :مقایسه نتایج اجرای روش پیشنهادی و با دیگر روش ها بر اساس معیار average precision

Data Set	BR	CC	HOMER	MLC4.5	MLKNN	RAKEL	Proposed method
Medical	.896	.901	.786	.823	.784	.676	**.801**
Enron	.693	.695	.604	.629	.635	.522	**.661**
Bibtex	.597	.599	.407	.392	.349	DNF	**.507**

معیار دیگر برای ارزیابی صحت در یادگیری چندبرچسبی، معیار خطای همینگ است. این معیار جز روش مبتنی بر برچسب است که به صورت رابطه (۴-۳) تعریف می‌شود.هرچه مقدار این معیار کمتر باشد به معنای کارایی بهتری می‌باشد.

$$Hamming - Loss(D) = \frac{1}{NL}\sum_{i=1}^{N}|\hat{y}_i \Delta y_i| \qquad (۸)$$

در جدول (۵) نتایج مربوط به اجرای روش پیشنهادی و دیگر روش‌ها برای معیار فقدان همینگ آورده شده است. منظور از DNF در این جدول‌ها این است که الگوریتم مربوطه نیاز به زمان زیادی داشته است.

جدول ۵ : مقایسه نتایج اجرای روش پیشنهادی و با دیگر روش ها بر اساس معیار hamming loss.

Data Set	BR	CC	HOMER	MLC4.5	MLKNN	RAKEL	Proposed method
Medical	.077	.077	.012	.015	.018	.012	**.014**
Enron	.045	.064	.051	.053	.051	.045	**.048**
Bibtex	.012	.012	.014	.016	.014	DNF	**.014**

اما نکته دیگری که باید در ارزیابی روش پیشنهادی با روش‌های پیشن بررسی شود، میزان پیچیدگی زمانی الگوریتم‌ها است. برخی از الگوریتم‌ها مانند الگوریتم ارتباط دودویی، اغلب از لحاظ دقت پیش‌بینی برچسب‌های نمونه‌های جدید از دیگر الگوریتم‌ها کاراتر است اما از لحاظ پیچیدگی زمانی، نیاز به زمان اجرای بالایی دارد. به همین دلیل در ارزیابی روش پیشنهادی با دیگر روش‌های موجود، علاوه بر مقایسه معیارهای ذکر شده، روش پیشنهادی از نظر میزان زمان مورد نیاز با دیگر الگوریتم‌های موجود مقایسه شده است. در جدول (۶) زمان اجرای الگوریتم های مختلف برحسب دقیقه آورده شده است.

جدول ۶ : زمان مورد نیاز برای اجرای الگوریتم های مختلف برحسب دقیقه

Data Set	BR	CC	HOMER	MLC4.5	MLKNN	RAKEL	Proposed method
Medical	7.3	11	5.8	1.03	0.4	35.3	**2.5**
Enron	122	164	60	5.06	3	215	**11**
Bibtex	3889	4365	1017	190	62	DNF	**150**

۵- نتیجه‌گیری

در این مقاله یک روش با استفاده از احتمال بین برچسب‌ها مبتنی بر روش نزدیکترین همسایگی برای دسته‌بندی داده‌های چندبرچسبی ارائه شد. ایده اصلی ارائه شده در این مقاله استفاده از احتمال شرطی بین برچسب‌ها برای دسته‌بندی داده‌های چند برچسبی می‌باشد برای اینکه بتوان از احتمال شرطی بین برچسب‌ها استفاده کرد باید در گام اول برای هر نمونه یک برچسب را به درست پیش‌بینی کرد (تا جای ممکن معیار تک خطا را کاهش داد). در این مقاله برای تشخیص این برچسب از روش نزدیکترین همسایگی بر اساس فاصله غیر اقلیدسی استفاده شده است. نتایج آزمایش‌ها بر روی مجموعه‌داده‌ها بیانگر این نکته است که روش پیشنهادی با توجه به پیچیدگی زمانی آن، نسبت به روش‌های دیگر، در پیش‌بینی برچسب‌های نمونه‌های جدید بهبود داشته است.

همچنین نتایج آزمایشات بیان‌گر این موضوع می‌باشد که با انتخاب مقادیر $\alpha=1$، $\beta=0$، $\gamma=-1$ و $\theta=-4$ بهترین نتایج برای الگوریتم پیشنهادی حاصل می‌شود به همین دلیل برای محاسبه معیارهای ارزیابی از این مقادیر استفاده شده است.ویژگی‌های روش پیشنهادی را می‌توان به صورت زیر خلاصه کرد.

زمان اجرای روش پیشنهادی برخلاف روش‌های BR,RAKEL,CC,ECC به تعداد برچسب‌ها مجموعه‌داده وابستگی شدیدی ندارد و با افزایش تعداد برچسب‌ها هزینه زمانی بالایی را تحمیل نمی‌کند.

روش پیشنهادی نسبت به روش‌هایی مانند MLKNN,MLC4.5,HOMER که هزینه زمانی قابل قبولی

دارند تقریبا دارای دقت بالاتری در پیش‌بینی برچسب‌های نمونه‌های جدید است .

در روش پیشنهادی سعی شده است از ارتباط بین برچسب‌ها برای تشخیص برچسب‌های نمونه‌های جدید استفاده شود. که در روش‌های قبلی از این ویژگی استفاده نشده است.

روش پیشنهادی دارای دو فاز مجزا است. در فاز نخست تعداد برچسب‌ها برای هر نمونه جدید تخمین زده می‌شود و در فاز دوم احتمال اینکه هر نمونه چه برچسب‌هایی را خواهد گرفت محاسبه می‌شود.

مراجع

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

[2] G. Madjarov, D. Kocev and D. Gjorgjevikj, "*an extensive experimental comparison of methods for multi-label learnin*, p*attern Recognition*," Vol. 45, No. 9, pp. 3084-3104, 2012.

[3] G. Tsoumakas and I. Katakis, "*Multi label classification An overview*," International Journal of Data Mining and Warehousing, Vol. 3, No. 3, pp. 1–13, 2007.

[4] E. L. Mencía, S. H. Park and J. Furnkranz,"*Efficient voting prediction for pairwise multilabel classificationNeurocomputing*,"in Proceedings of the 18th European Conference on Machine Learning, pp. 658-668, 2010.

[5] A. Clare and R. D. King, "*Knowledge discovery in multi-label phenotype data*,"in Proceedings of the 5th European Conference on PKDD, pp. 42–53,2001.

[6] M. L. Zhang and Z. H. Zhou, "*ML-kNN:a lazy learning approach to multi-label learning*," Pattern Recognition,Vol. 40, No. 7, pp. 2038–2048, 2007.

[7] G. Tsoumakas, I. Katakis and P. Vlahavas,"*Effective and efficient multilabel classification in domains with large number of labels*," In ECML/PKDD Workshop on Mining Multidimensional Data, 2008.

[8] G. Tsoumakas and I. P. Vlahavas, "*Random k-labelsets: An ensemble method for multilabel classification*," In Proceedings of the 18th European Conference on Machine Learning, PP. 406-417, Springer, 2007.

[9] J. Read, B. Pfahringer and G. Holmes, "*Multi-label classification using ensembles of pruned sets,"* in Proceedings of the 8th IEEE International Conference on Data Mining, pp. 995-1000, 2008.