روشی تکاملی برای بهینه‌سازی طبقه‌بندی مبتنی بر جمع‌سپاری

مریم حبیبی پور[۱]، الهام آفرنده[۲]، اکرم حبیبی پور[۳]

۱-آموزشکده فنی و حرفه‌ای سما، دانشگاه آزاد اسلامی، واحد مشهد، مشهد، ایران

۲-عضو هیات علمی گروه کامپیوتر ،موسسه آموزش عالی توس

۳-دانشگاه آزاد اسلامی، واحد مشهد، مشهد، ایران

چکیده

یادگیری مبتنی بر جمع‌سپاری یکی از انواع مهم یادگیری است که سعی در کاهش هزینه‌های برچسب‌زنی دارد.در بحث یادگیری مبتنی بر جمع‌سپاری، این امکان را فراهم می کنند که از خرد جمعی برای حل مسائل استفاده کنند.اکثر روش‌های مبتنی بر جمع‌سپاری به دلیل استفاده از الگوریتم EM در محاسبه پارامترها، دارای مشکل گیرافتادن در بهینه‌های محلی ناشی از نقطه شروع نامناسب هستند. در این مقاله روشی مبتنی بر الگوریتم های تکاملی به منظور رفع مشکلات الگوریتم EM در محاسبه پارامترهای مدل جمع سپاری ارائه شده است. در این روش، مقادیر اولیه پارامترهای مدل طبقه‌بندی مبتنی بر جمع‌سپاری، از جمله دقت نظردهندگان و ضرایب خط جداساز از طریق الگوریتم ازدحام ذرات محاسبه می‌شوند. سپس مقادیر بدست آمده به عنوان نقطه شروع الگوریتم EM مورد استفاده قرار گرفته و بهینه‌سازی می‌شوند.آزمایشات بر روی مجموعه داده های واقعی با نظردهندگان شبیه‌سازی شده، و مجموعه داده‌های واقعی با نظردهندگان واقعی، انجام شده است. نتیجه این آزمایشات برای روش ارائه شده در مقایسه با سایر روش‌ها، نشان دهنده ی برتری روش پیشنهادی از نظر صحت طبقه‌بندی بر روی داده‌های آزمون است.

واژگان کلیدی: یادگیری مبتنی بر جمع‌سپاری،خرد جمعی،بهینه‌های محلی، الگوریتم تکاملی، الگوریتم ازدحام ذرات

مقدمه

یکی از نظریه‌هایی که امروزه موردتوجه قرارگرفته است واژه جمع‌سپاری یا انبوه‌سپاری است که توسط آقای «هاو» در سال ۲۰۰۶ مطرح‌شده است(F. mansori, 2013). در مسائل طبقه بندی مبتنی بر جمع سپاری، بر خلاف طبقه بندی سنتی رایج، داده ها دارای یک برچسب صحیح نیستند و بجای آن، هر داده شامل چندین برچسب از طرف چندین نظردهنده است که با توجه به عدم تخصص کافی نظردهنده ها، این برچسب ها کیفیت مطلوبی ندارند و در نظرات خود گاهاً با مقادیری از خطا همراه هستند. مثال مربوط به این قضیه را میتوان از نظرات رادیولوژیست در مورد تصویر MRI در مقایسه با پزشک متخصص نام برد. جمع سپاری به منظور کاهش هزینه های مربوط به برچسب زنی مورد استفاده قرار میگیرد. زیرا معمولا هزینه چندین نظر از افراد غیر خبره از نظر فرد متخصص کمتر خواهد بود.

درمرجع (R. Jin and Z. Ghahramani, 2003) روش تشخیص برچسب‌های مناسب از برچسب‌های تصادفی بیان شده است، درصورتی که تعداد کلاس‌ها و به دنبال آن تعداد نظردهندگان افزایش یابد، روش ارائه شده توسط Raykar دارای بار محاسبات بالایی خواهد بود زیرا در این روش به‌ازای هر نظردهنده، تعداد k2 پارامتر در نظر گرفته می‌شود که kمشخص کننده‌ی تعداد کلاس‌ها است. بنابراین با افزایش تعداد نظردهندگان یا کلاس‌ها تعداد پارامترها افزایش یافته و این مساله منجر به بیش‌برازش می‌شود. در این روش با حذف پارامترهای اعتماد به نظردهنده‌ها در کلاس‌های مختلف و معرفی تنها یک پارامتر اعتماد، مدلی ارائه می‌شود که ضمن برخورداری از نتیجه مشابه با روش Raykar با افزایش پارامترها دارای بار محاسبات کمتری است.

در (Q. Liu et al, 2013) موضوع عدم توازن برچسب‌های مثبت و منفی را در طبقه‌بندی دوکلاسه مبتنی بر جمع‌سپاری مورد تحلیل قرار می‌دهد. ادعای وی بدین گونه است که نظردهندگان نه‌چندان خبره در اکثر مواردی که برچسب صحیح نمونه مورد را نتوانند حدس بزنند آن نمونه را منفی تلقی می‌نمایند. این مسئله منجر به عدم توازن بین برچسب‌های کلاس‌ها می‌گردد. وی با تحلیلی ریاضی، به تنظیم توازن بین دو کلاس این‌گونه موارد پرداخته است.اکثر کارهای انجام شده در حوزه‌ی یادگیری مبتنی بر جمع‌سپاری بر پایه‌ی الگوریتم EM است. الگوریتم EM به مقداردهی اولیه حساس است ضمن اینکه نتایج خیلی قابل‌اعتماد نیستند.

در (V.C. Raykar and Shipeng Yu, 2012) مدل یادگیری از چندیدن متخصص بیان شده که این روش، مدلی را مبتنی‌بر شبکه بیز ارائه می‌دهد که با معرفی پارامترها و روشی متفاوت از روش Raykar، علاوه بر صفحه‌ی جداساز نهایی که از نظرات چند متخصص به‌دست می‌آید صفحه‌ی جداساز مربوط به هر یک از متخصصین را تخمین بزند. مزیت این روش استفاده از تابع زیان لولا است که باعث می‌شود هنگامی که ابعاد مساله زیاد شود ماتریس کرنل اسپارس‌تر و مقدار زیان حاصل از تخمین اشتباه برچسب‌ها حداقل شود.

در (Kartik Audhkhasi and Shrikanth (Shri) Narayanan, 2013) مدل کارتیک استفاده شده است که این روش سعی در مدل کردن رفتار متفاوت نظردهندگان روی مجموعه داده‌های مربوط به یک کلاس را دارد. به‌صورتی

که فرض می‌شود متخصصین بر روی داده‌هایی که نسبتاً شبیه به هم هستند (به صوت محلی) یکسان عمل می‌کنند و دقت آن‌ها روی این مجموعه ثابت است؛ اما دقت آن‌ها که از آن به‌عنوان ماتریس قابلیت یاد می‌شود، روی کل نمونه‌ها متفاوت خواهد بود. این مدل پارامترهای طبقه‌بندی و ماتریس قابلیت را بدون فرض دانش در مورد برچسب‌های نمونه‌ها، با الگوریتم EM تقریب می‌زند.

در مدل‌های مختلف جمع‌سپاری متداول مانند روش ریکار (R. Jin and Z. Ghahramani, 2003) و کارتیک (Kartik Audhkhasi and Shrikanth (Shri) Narayanan, 2013)، روشی به‌منظور پیش‌بینی برچسب‌های اصلی پنهان داده‌ها از روی نظرات پیشنهاد گردیده است. در این روش‌ها، پارامترهای پنهانی برای تعیین میزان خبرگی نظردهندگان روی داده‌ها، عنوان شده است که یکی از اهداف اصلی، تعیین میزان هرچه دقیق‌تر این پارامترها است.
اکثر روش‌های مبتنی بر جمع‌سپاری به دلیل استفاده از الگوریتم EM در محاسبه پارامترها، دارای مشکل گیرافتادن در بهینه‌های محلی ناشی از نقطه شروع نامناسب هستند. این مشکلات باعث شد تا روشی مبتنی بر الگوریتم های تکاملی به منظور رفع مشکلات الگوریتم EM در محاسبه پارامترهای مدل جمع سپاری پیشنهاد گردد.
در این مقاله در بخش یک روش جمع سپاری در بخش دوم روش ریکار و الگوریتم EM توضیح داده خواهد شد. در بخش سوم روش پیشنهادی و در بخش جهارم نتایج شبیه سازی شده آورده می شود و درانتها جمع بندی مقاله خواهد بود

جمع سپاری

یکی از نظریاتی که امروزه موردتوجه قرارگرفته است واژه جمع‌سپاری یا انبوه‌سپاری است که توسط آقای «هاو» در سال ۲۰۰۶ مطرح‌شده است(F. mansori, 2013) و می‌تواند در دستور کار کسب‌وکارها قرار گیرد. جمع‌سپاری به‌عنوان مدل کسب‌وکار در حال ظهور می‌باشد که تمرکز آن بر مشارکت دادن جمعیت در فعالیت‌هایی چون حل مسئله، تولید و توسعه مفاهیمی چون مشارکت در ایده‌سازی، نوآوری، تولید و فرآیندهای ارائه خدمات می‌باشد که بر کیفیت محصول، وفاداری و خشنودی مشتری اثری مستقیم دارد (Amazon Mechanical Turk). جمع‌سپاری به کاربردی و اجرایی کردن خرد جمعی دلالت دارد و سازوکاری برای به‌کارگیری اهرم دانش جمعی کاربران آنلاین به سمت نتایج پرسود می‌باشد (Q. Liu et al, 2013).
روش‌های مبتنی بر جمع‌سپاری و یادگیری از نظرات شامل روش های مدل رأی اکثریت ، مدل ریکار ،مدل کارتیک و.... می باشد که هر کدام مزایا و معایب خود را دارند. در این مقاله به دلیل مقایسه نتایج با روش ریکار این روش توضیح مختصری داده می‌شود.

مدل ریکار و الگوریتم EM

مدل ریکار که در راستای رفع نواقص مربوط به روش رأی اکثریت تلاش می‌کند، از بستر EM به منظور بهینه‌سازی پارامترهای جمع‌سپاری استفاده می‌نماید. در حوزه‌ی یادگیری مبتنی بر جمع‌سپاری، توزیع نمونه‌ها و برچسب‌ها به‌صورت زیر نمایش داده می‌شود:

$$D=\{(x_i,y_i^1,...,y_i^R)\}_{i=1}^N \quad (۱)$$

متغیر y_i^j برابر برچسبی است که نظردهنده jام به داده‌ی iام نسبت می‌دهد. هدف، یافتن برچسب‌های مخفی y_i برای همه نمونه‌ها و یافتن میزان اعتماد به نظردهندگان و نیز یافتن وزن‌های طبقه‌بند است به‌نحوی‌که برای داده‌های جدید و بدون برچسب بتوانیم برچسب آن را حدس بزنیم.در ریکار پارامترهایی به نام α و β برای هر یک از نظردهنده ها در نظر می گیرد که به ترتیب معرف میزان دقت آن‌ها روی کلاس یک و صفر است. درصورتی‌که μ_i، تخمین برچسب واقعی مربوط به داده iام در نظر گرفته شود، روابط مربوط به محاسبه پارامتر α^j و β^j (دقت نظردهنده ی jام) و μ_i به‌صورت زیر خواهد بود:

$$\beta^j=\frac{\sum_{i=1}^{N}(1-\mu_i)(1-y_i^j)}{\sum_{i=1}^{N}(1-\mu_i)} \qquad \alpha^j=\frac{\sum_{i=1}^{N}\mu_i y_i^j}{\sum_{i=1}^{N}\mu_i} \quad (۲)$$

$$\omega^{t+1}=\omega^t-\eta H^{-1}g \qquad \mu_i=\frac{a_i p_i}{a_i p_i+b_i(1-p_i)}$$

محاسبه α و β نیاز به دانستن مقادیر μ دارد. همچنین محاسبه مقادیر μ نیاز به دانستن مقدار ω دارد که خود نیاز به نقطه شروع مشخصی دارد. درمجموع می‌توان گفت که نقطه شروع مناسب، امری است که در فرایند محاسبه پارامترهای جمع‌سپاری نقش مهمی ایفا می‌نماید.

الگوریتم[1] EM روش تکراری کارا برای محاسبه بیشینه درست‌نمایی در هنگامی است که پارامترها پنهان یا مخفی هستند(A.P. Dempster, 1977) . به دلیل آن‌که الگوریتم EM به کمینه‌های محلی همگرا می‌شود بهتر است با شروع‌های متفاوت یعنی مقادیر اولیه برای ω آزمایش شود تا بتوان احتمال رسیدن به نتیجه بهتر را انتظار داشت.

[1] Expectation-Maximization, EM

روش پیشنهادی

یکی از مشکلات اصلی الگوریتم EM بهینه های محلی است در این مقاله، روشی مبتنی بر روش های تکاملی به منظور کاهش رخداد بهینه های محلی، در فرایند یافتن پارامترهای بهینه مدلهای طبقه بندی مبتنی بر جمع سپاری، پیشنهاد شده است. لذا در صورت استفاده از این روش، به دلیل بهبود فرایند بهینه سازی، انتظار افزایش میزان صحت طبقه بند نیز وجود دارد.

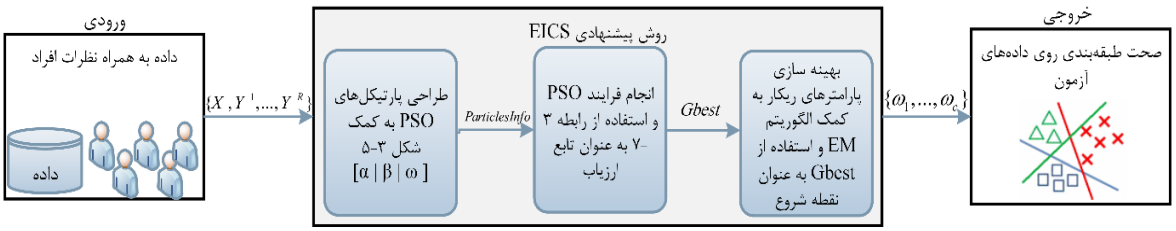
برای پیدا کردن بهترین برچسب برای داده از بین برچسب های ارائه شده و تخمین میزان دقت و همچنین تعیین مدل طبقه بند از روش ریکار استفاده شده است.در فرایند حل مساله بعضی داده ها مخفی یا گم هستند که روش ریکار از الگوریتم EM برای بدست اوردن پارامتر ها استفاده می کند.

در این مقاله رویکرد زیر ارائه شده است:

- یافتن نقطه شروع مناسب برای الگوریتم EM در جمع سپاری (نقطه شروع ω و α و β)

روش EICS۲

در این روش ، نقطه اولیه مناسب پارامترها برای انجام جستجو به روش EM در جمع سپاری را میتوان به عنوان مسئله ای جستجو تعریف نمود. این روش در شبکه های عصبی عمیق مرسوم است. زیرا در آن ها مشکل کمینه‌های محلی متداول است و از روش های مختلف (مانند روش های تکاملی) در جهت بهبود نقطه شروع استفاده میشود. برای این منظور روش تکاملی جستجوی PSO (Kennedy and James, 2011)پیشنهاد شده است. نمای کلی این روش مانند شکل ۱ است.



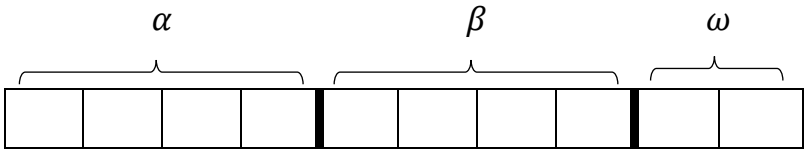
شکل ۱: فلوچارت کلی روش پیشنهادی (EICS)

با توجه به شکل ۱، داده های جمع سپاری که شامل ویژگی های داده به همراه چندین برچسب از نظردهندگان است به عنوان ورودی سیستم داده میشود. در باکس اول، برداری شامل ωو α و β تشکیل میگردد که پارتیکل های مربوط به [هستند. این پارتیکل ها، جواب‌های کاندید نقطه شروع را بیان می‌کنند. الگوریتم PSO با در نظر گرفتن تابع ارزیاب

[2] Evolutionary Initialize Crowdsourcing

مناسب، نقطه ی بهینه‌ای برای این پارامترها جستجو می کند. پس از آن، از نقاط یافته شده‌ی بهینه‌ی نسبی، به عنوان نقطه شروع استفاده می گردد و بهینه سازی متداول جمع سپاری برای تنظیم دقیق تر این پارامترها انجام می‌شود.

به عنوان مثال در صورتی که تعداد نظردهندگان $R = 4$ در نظر گرفته شده و داده ها در فضای $d = 2$ بعدی قرار داشته باشند، بنابراین $\omega \in \mathbb{R}^d$ و α و β هر کدام بردارهایی 4 تایی هستند. بنابراین هر پارتیکل شکلی مانند شکل۲ را خواهد داشت:



شکل ۲: نحوه قرارگیری محل پارامترها در پارتیکل نمونه در روش EICS

همانطور که مشخص است، سایز هر پارتیکل در الگوریتم PSO یا تعداد پارامترها در این حالت برابر $2R + d$ خواهد بود. بدیهی است که برای مسائل چند کلاسه، به ازای هر کلاس، خط (ابرصفحه) جداکننده ای در نظر گرفته می‌شود. همچنین پارامتر β حذف شده و به ازای هر فرد، تعداد K پارامتر مشخص کننده دقت وی روی هر کدام از کلاس ها وجود دارد. لذا در صورتی که مسئله‌ای K کلاسه مفروض باشد، تعداد پارامترهای هر پارتیکل برای $KR + Kd$ خواهد بود.

پس از تشکیل پارتیکل های اولیه، همانطور که در باکس دوم شکل ۱ مشخص است، مراحل الگوریتم PSO که شامل به روزرسانی مقادیر پارتیکل ها است انجام میشود. اما نکته قابل ذکر در این روش این است مقادیر دو قسمت اول پارتیکل‌ها که برای α و β است میبایستی اعدادی بین ۰ و ۱ باشند؛ زیرا این پارامترها بیان کننده‌ی دقت افراد بوده و در بازه ۰ و ۱ است. در طی فرایند PSO، این محدودیت را میتوان به صورت خودکار با اعمال جریمه روی پارتیکل هایی که مقادیر α و β آنها از محدوده مجاز تخطی کرده است کنترل کرد. اما محدودیتی برای مقدار اولیه بخش سوم پارتیکل که متخص پارامتر ω است وجود ندارد. این اعداد را از نمونه گیری تصادفی از توزیع نرمال میتوان مقداردهی نمود.

سپس با توجه به باکس سوم شکل ۱، از بهترین پارتیکل شناسایی شده (Gbest) در فرایند PSO به عنوان مقادیر اولیه در روش ریکار استفاده می‌شود. لذا پس از آن محاسبات تکراری EM برای بهینه کردن مجدد این پارامترها و محاسبه تخمین برچسب اصلی داده های آموزش یعنی بردار μ انجام می‌شود. نهایتاً همانطور که در خروجی شکل۱مشخص است، مدل طبقه بندی حاصل از الگوریتم، به منظور طبقه بندی داده های آزمون (که دارای هیچ برچسبی حتی از نظرات نیستند) استفاده شده و صحت طبقه بندی گزارش میشود.

به صورت کلی، میتوان روش EICS را بدین گونه خلاصه نمود. مقادیر اولیه پارامترها توسط الگوریتم PSO محاسبه شده و این مقادیر برای بهینه سازی بیشتر توسط الگوریتم EM استفاده میشود. پس از آن، از مدل طبقه بندی محاسبه شده برای برچسب زنی داده های آزمون و محاسبه صحت طبقه بندی استفاده می گردد.

تابع ارزیاب

با توجه به روش گفته شده نیاز به استفاده از الگوریتم تکاملی PSO می‌باشد. همه‌ی الگوریتم های جستجو تکاملی نیاز به تابعی دارند که با کمک آن بتوانند میزان بهینگی یا کیفیت هر پارتیکل را ارزیابی نمایند. این تابع به تابع ارزیاب یا fitness معروف است. در مسائل طبقه بندی رایج که از الگوریتم PSO برای بهینه سازی پارامترهای مختلف استفاده میشود، به راحتی میتوان مقدار تابع ارزیاب را برابر عکس میزان خطای سیستم بر روی بخشی از داده ها (داده های ارزیابی) در نظر گرفت. زیرا برچسب واقعی آن داده ها موجود بوده و عبارت خطا قابل محاسبه است.

اما همانطور که گفته شد، در جمع‌سپاری، برچسب اصلی داده ها موجود نیست. بنابراین نمیتوان از میزان صحت (یا خطا) بدست آمده برای ارزیابی کیفیت پارتیکل های الگوریتم تکاملی استفاده کرد. در این مقاله روش دیگری به منظور ارزیابی پارتیکل ها پیشنهاد میشود. با توجه به اینکه حل اصلی روش پایه در پی بیشینه سازی $E\{\ln \Pr(D,y\mid\theta)\}$ بود، لذا میتوان از خروجی همین عبارت به عنوان معیار مناسبی برای fitness استفاده نمود. به بیانی دیگر، پارتیکل هایی که مقدار $E\{\ln \Pr(D,y\mid\theta)\}$ را بیشتر می کنند، بهتر هستند. بنابراین تابع ارزیاب به صورت زیر خواهد بود:

$$Fitness = E\{\ln \Pr(D,y\mid\theta)\} = \sum_{i=1}^{N} \mu_i \ln p_i a_i + (1-\mu_i)\ln(1-p_i)b_i \quad (۳)$$

که در آن:

$$\begin{aligned} \mu_i &= \frac{a_i p_i}{a_i p_i + b_i(1-p_i)} \\ p_i &:= \sigma(\omega^T x_i) \\ a_i &:= \prod_{j=1}^{R} [\alpha^j]^{y_i^j} [1-\alpha^j]^{1-y_i^j} \\ b_i &:= \prod_{j=1}^{R} [\beta^j]^{1-y_i^j} [1-\beta^j]^{y_i^j} \end{aligned} \quad (۴)$$

اما با توجه به اینکه در بخش قبل ذکر شد که محدودیت بازه ی $(0,1)$ برای پارامترهای α و β وجود دارد و این تخطی از این مقادیر مجاز نیاز به اعمال جریمه در تابع ارزیاب است، تابع ارزیاب فوق به صورت زیر اصلاح می گردد:

$$Fitness = E\{\ln \Pr(D,y\mid\theta)\} - \lambda Pendty(\alpha,\beta) \quad (۵)$$

که در آن، $E\{\ln Pr(D,y|\theta)\}$ همان عبارت قبل بوده و تابع $Pendty(\alpha,\beta)$ در صورت خروج α و β از محدوده مجاز، عدد ۱ و در غیراین صورت، عدد صفر را برمی‌گرداند. همچنین ثابت λ را میتوان به منظور تنظیم اهمیت عبارت جریمه استفاده کرد. بدیهی است که در صورتی که λ را مقدار بزرگی قرار دهیم، الگوریتم PSO به پارتیکل هایی که قسمت مربوط به α و β آنها از محدوده مجاز تجاوز کرده باشند نمره یا fitness کمی اختصاص خواهد داد.

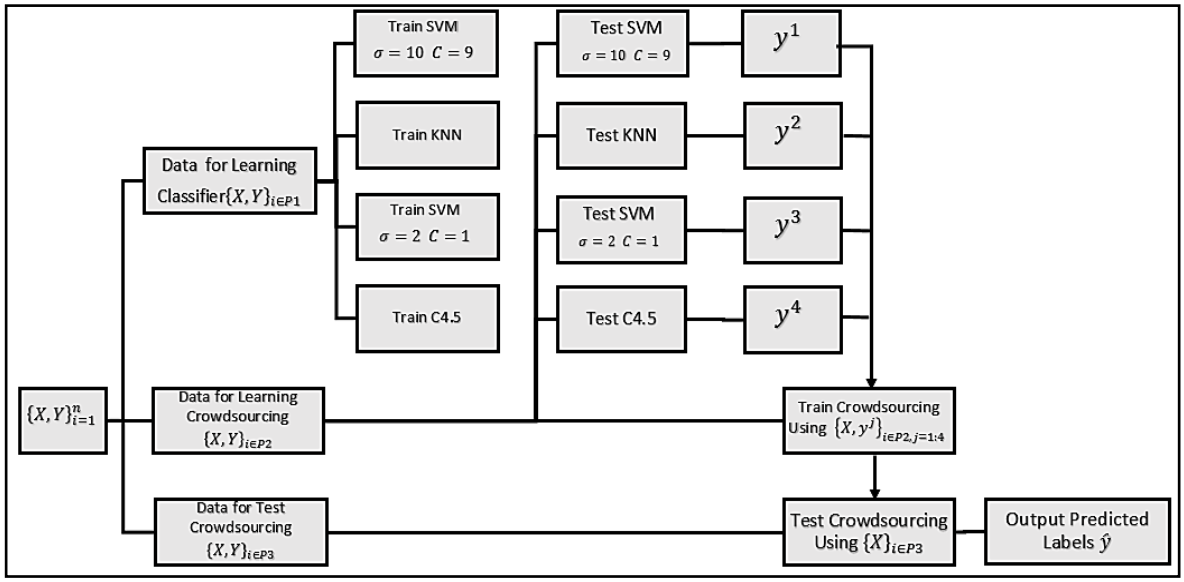
نتایج شبیه سازی شده

در این بخش، آزمایشات مربوطه برای مقایسه روش پیشنهادی با سایر روش های مرسوم انجام گرفته است. در ابتدا توضیحات در مورد نحوه ی تولید نظردهندگان شبیه سازی شده برای مجموعه داده استاندارد ارائه خواهد شد. لذا دو روش مختلف برای تولید نظرات ارائه میگردد. سپس نتایج مربوط به طبقه بندی داده های واقعی با نظرات شبیه سازی شده، ارائه خواهد شد. نهایتاً در بخش بعدی نتایج مربوط به مجموعه داده های واقعی با نظرات واقعی مورد بررسی قرار خواهد گرفت.

نحوه‌ی تولید نظردهندگان شبیه‌سازی‌شده

با توجه به این‌که در روش‌های مبتنی بر جمع‌سپاری، برچسب واقعی داده‌های آموزش موجود نیست، برای ارزیابی روش‌های یادگیری مبتنی بر جمع‌سپاری می‌توان از مجموعه داده‌های موجود (که دارای برچسب واقعی هستند) استفاده کرد. به‌عنوان‌مثال مجموعه داده‌های UCI3 که برچسب واقعی آن‌ها موجود است درنظر گرفته می‌شود. هدف این است که برچسب‌های متعددی به‌عنوان نظرات افراد خبره تهیه گردد. برای این منظور می‌توان از دو روش زیر استفاده کرد:

روش اول (Kartik Audhkhasi and Shrikanth (Shri) Narayanan, 2013) : انواع متعددی از طبقه‌بندها را به‌صورت ضعیف آموزش داد و از نظرات آن‌ها برای نمونه‌های دیگر به‌عنوان نظرات افراد استفاده نمود. با توجه به اینکه برچسب واقعی این داده‌ها موجود است می‌توان نتیجه نهایی طبقه‌بندی را با آن مقایسه کرده و نرخ صحت را اندازه‌گیری نمود. توجه شود که از برچسب‌های واقعی نمونه‌ها در امر یادگیری استفاده نمی‌گردد. شکل ۳ نحوه‌ی تولید برچسب‌ها توسط طبقه‌بندهای مختلف و همچنین فرایند آموزش جمع‌سپاری را نشان می‌دهد. در این شکل داده‌ها به سه قسمت تقسیم‌شده‌اند. بخش اول آن جهت آموزش چهار طبقه‌بند SVM, KNN (با دو مجموعه پارامتر) و C4.5 استفاده می‌شود. بخش دوم داده‌ها جهت آزمایش چهار طبقه‌بند و تولید برچسب‌ها به‌عنوان برچسب های نظردهندگان فرایند جمع‌سپاری استفاده می‌شود و بخش سوم داده‌ها (داده‌های آزمون) برای آزمایش فرایند یادگیری مبتنی بر جمع‌سپاری و استنتاج برچسب تخمینی مورداستفاده قرار می‌گیرند.



شکل ۳:نمودار نحوه‌ی تولید برچسب‌های فرایند آموزش در جمع‌سپاری

روش دوم: با استفاده از روش ذکرشده در (Rodrigues et al, 2017) می‌توان به ازای هر نظردهنده، ماتریس اغتشاش[3] دلخواهی درنظر گرفته شود و برچسب‌های آن نظردهنده با نمونه‌برداری از این ماتریس قابل تولید است. به عنوان مثال فرض شود میخواهیم در یک مساله طبقه بندی دو کلاسه، نظرات فردی را تولید کنیم که دقت وی روی کلاس ۰، برابر ۹۰ درصد و دقت آن روی کلاس ۱ برابر ۷۰ درصد باشد. ماتریس اغتشاش آن فرد به صورت شکل ۴ است.

Annotation \ True Label	Class 0	Class 1
Class 0	90	30
Class 1	10	70

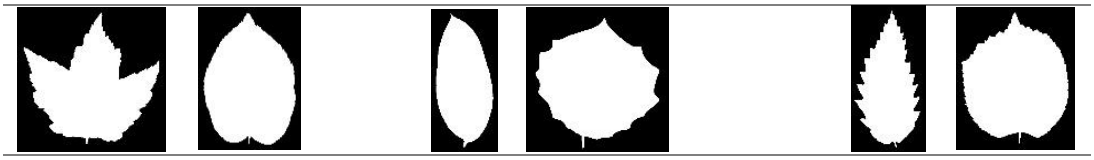
شکل ٤: نمونه ای از ماتریس اغتشاش برای مدل سازی نظرات فردی خاص

بدیهی است که در صورتی که برای یک داده، برچسب اصلی آن، برابر ۰ باشد، نظردهنده مربوط به شکل ۴، با احتمال ۹۰ درصد برچسب آن را ۰ تخمین زده و با احتمال ۱۰ درصد آن را اشتباهاً ۱ تلقی میکند. این نمونه برداری مبتنی بر ماتریس اغتشاش را میتوان توسط چرخ رولت پیاده سازی نمود.

[3] confusion matrix

طبقه‌بندی داده‌های واقعی با نظردهندگان شبیه‌سازی‌شده

در این بخش به‌منظور ارزیابی روش ارائه‌شده، از داده‌های واقعی با نظردهندگان شبیه‌سازی‌شده استفاده می‌کنیم. اطلاعات مربوط به داده‌ها در جدول۱ قابل‌مشاهده است. مجموعه کل داده‌ها به‌صورت تصادفی به ۷۰ درصد آموزش و ۳۰ درصد آزمون تقسیم‌بندی شده و میانگین صحت طبقه‌بندی در ۲۰ اجرا، گزارش شده است. به‌منظور شبیه‌سازی نظردهندگان از روش ذکرشده در (Rodrigues at al, 2017) یعنی ماتریس اغتشاش استفاده شده است. همانطور که گفته شد، در این روش به ازای هر نظردهنده، یک ماتریس اغتشاش دلخواه درنظر گرفته می‌شود و برچسب‌های آن نظردهنده با نمونه‌برداری از این ماتریس قابل‌تولید است. یکی از داده‌های مورداستفاده، مجموعه‌داده‌ها Leaves است که در شکل ۵ مشاهده می‌شود. این داده‌ها در مجموعه CEKA[۴] موجود است. هر نمونه در این دیتاست دارای یک برچسب درست است که توسط اوراکل مشخص شده است. هرچند در مورد این مجموعه داده، نظرات شبیه‌سازی‌شده است اما با توجه به شکل ۵، استفاده از روش جمع‌سپاری برای برچسب‌زنی این نمونه‌ها توسط افراد عادی مفید است زیرا هزینه و زمان برچسب‌زنی را کاهش می‌دهد. نظردهندگان ممکن است در برچسب‌زنی برگ‌های Alnus و Eucalyptus و برگ‌های Populus و Tilia دچار اشتباه گردند.



Acer (maple) | Alnus (Alder) | Eucalyptus | Populus (poplar) | Quercus4 (oak) | Tilia

شکل ۵: مجموعه داده Leaves6 که در ابزار Ceka موجود است و دارای ٦ کلاس مختلف می‌باشد.

میانگین نتایج مربوط به روش پیشنهادی در مقایسه با روش ریکار و رأی اکثریت پس از ۲۰ بار اجرا، در جدول ۲ نمایش داده شده است. توجه شود که نتایج الگوریتم رأی اکثریت ساده، فقط بر روی داده‌های آموزشی که دارای چندین برچسب هستند، قابل‌محاسبه است. بهترین نتیجه‌ی داده‌های آزمون در هر ردیف به‌صورت بولد، هایلایت شده است.

[۴] CEKA ابزاری در دسترس به زبان جاوا است که انواع روش های جمع سپاری از جمله (A.P. Dawid and A.M. Skene, 1979) ، (Filipe Rodrigues et al, 2013) ، Raykar (V.C. Raykar et al, 2010) ، (G. Demartini et al, 2012)، (Karger et al, 2011) و رای اکثریت در آن پیاده سازی شده است.

جدول ۱:جزئیات مجموعه داده‌های استفاده شده برای ارزیابی طبقه‌بندی روش پیشنهادی. (ستون‌ها از چپ به راست شامل تعداد داده‌های آموزش/ تعداد داده‌های آزمون، تعداد ویژگی‌ها، تعداد کلاس‌ها و تعداد نظردهندگان شبیه‌سازی‌شده است. مجموعه داده‌های مشخص شده با *، سنتزی هستند.)

dataset	#instance	#feature	#class	#annotator
Pid	768	8	2	5
breast cancer	683	10	2	5
ionosphere	351	34	2	4
balance	625	4	3	7
TAE	151	5	3	8
cmc	1473	9	3	8
ecoli	336	7	8	10
Glass	214	10	7	7
leaves	384	65	6	5
seeds	210	7	3	10
thyroid	215	5	3	10
LabelMe	1000/1500	300	8	77
Reuters-21578	1800/5216	1200	8	38

جدول ۲:نتایج مربوط به پیاده‌سازی روش پیشنهادی و مقایسه آن بر روی مجموعه داده‌های UCI

Data set	Raykar Model	Proposed Model
	Accuracy (%)	Accuracy (%) EICS
BreastFVE GGF cancer	95.122	96.098
Pid	76.623	80.221
Balance	84.043	87.234
Seeds	71.111	75.873
Thyroid	86.154	90.769
Cmc	50.971	49.593
Ecoli	51.43	50.74
Glass	61.538	63.231

TAE	45.304	45.058
Ionosphere	82.075	83.962
Leaves6	71.241	77.843

نتیجه‌گیری و پیشنهادات

یادگیری مبتنی بر جمع‌سپاری یکی از انواع مهم یادگیری است که سعی در کاهش هزینه‌های برچسب‌زنی دارد. در این نوع یادگیری هر نمونه دارای چندین برچسب توسط نظردهندگان مختلف بوده و برچسب واقعی نمونه‌ها موجود نیست. مدل رأی اکثریت ساده بدون درنظر گرفتن ویژگی نمونه‌ها و دقت نظردهندگان به تخمین برچسب نمونه‌ها می‌پردازد. روش ریکار به‌عنوان یک روش پایه برای حل مسائل جمع‌سپاری و غلبه بر مشکلات روش رأی اکثریت مطرح شده است. این روش و سایر روش های طبقه بندی مبتنی جمع سپاری رایج، از الگوریتم EM و روش گرادیانی برای تخمین پارامترها استفاده میکنند. استفاده از الگوریتم EM به این دلیل است که برچسب اصلی داده ها مخفی (پنهان) است. همانطور که در (V.C. Raykar et al, 2010) یاد شده است، الگوریتم EM دارای گیرافتادن مشکل مینیمم های محلی است.

در این مقاله روش طبقه بندی مبتنی بر جمع سپاری به منظور غلبه بر مشکلات موجود در روش های مبتنی بر EM پیشنهاد شد. روشهای پیشنهاد که مبتنی بر یادگیری تکاملی هستند، از الگوریتم PSO به دو منظورِ ارائه نقطه شروع مناسب الگوریتم EM استفاده شده است. نتایج آزمایش‌ها بر روی داده‌های مصنوعی، داده‌های واقعی با نظردهندگان شبیه‌سازی‌شده و داده‌های واقعی با نظردهندگان واقعی نشان از برتری مدل پیشنهادی نسبت به روش ریکار و سایر روش ها دارد.

به‌عنوان پیشنهادهایی برای کارهای آینده که در ادامه روش ذکرشده می‌توان انجام داد موارد زیر مطرح می‌شود.

استفاده از سایر روش های تکاملی بجای PSO در فرایند جمع سپاری و تحلیل نتایج حاصله.

استفاده از مدل های تکاملی بر روی روش های جمع سپاری مبتنی بر خوشه بندی مانند روش کارتیک (Kartik Audhkhasi and Shrikanth (Shri) Narayanan, 2013).

استفاده از بستر تکاملی ارائه شده بر روی مسائل رگرسیون مبتنی بر جمع سپاری.

مراجع

F. mansori, "online semi-supervised music recommender," Master Of Science Thesis, Computer Department Engineering, Ferdowsi University of Mashhad, 2013.

R. Jin , Z. Ghahramani, “Learning with Multiple Labels,” Neural Information Processing Systems, pp. 921-928, 2003.

Q. Liu , M. Steyvers , J.W. Fisher , A. Ihler , On reliable crowdsourcing and the use of ground truth information, Technical Report, Workshop on Crowdsourcing at Scale, HCOMP, 2013.

V.C. Raykar, Shipeng Yu. “Eliminating Spammers and Ranking Annotators for Crowdsourced Labeling Tasks.” J. Machine Learning Research, vol. 13, pp. 491-518, 2012.

Kartik Audhkhasi, Shrikanth (Shri) Narayanan, “A Globally-Variant Locally-Constant Model for Fusion of Labels from Multiple Diverse Experts without Using Reference Labels,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 4, April 2013.

V.C. Raykar, S. Yu, L.S. Zhao, G.H. Valadez, C. Florin, L. Bogoni, and L. Moy, “Learning from Crowds,” J. Machine Learning Research, vol. 11, pp. 1297-1322, Mar. 2010.

Amazon Mechanical Turk http://mturk.com

A.P. Dempster, N.M. Liard, and D.B. Rubin, “Maximum Likelihood from Incomplete Data via the EM Algorithm,” J. Royal Statistical Soc.: Series B, vol. 39, pp. 1-38, 1977.

Kennedy, James. "Particle swarm optimization." Encyclopedia of machine learning. Springer US, 2011.

Rodrigues, Filipe, et al. "Learning supervised topic models for classification and regression from crowds." IEEE transactions on pattern analysis and machine intelligence, 2017.

A.P. Dawid and A.M. Skene, “Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm,” J. Royal Statistical Soc. Series C (Applied Statistics), vol. 28, no. 1, pp. 20-28, 1979.

Filipe Rodrigues, Francisco Pereira, Bernardete Ribeiro. "Learning from multiple annotators: Distinguishing good from random labelers." Pattern Recognition Letters, vol.34 pp.1428–1436, 2013

G. Demartini, D. E. Difallah, and P. Cudré-Mauroux, “Zen-Crowd: Leveraging Probabilistic Reasoning and Crowdsourcing Techniques for Large-Scale Entity Linking,” Proc. 21st Interna-tional Conference on World Wide Web (WWW), pp. 469-478, 2012.

Karger, David R., Sewoong Oh, and Devavrat Shah. "Iterative learning for reliable crowdsourcing systems." Advances in neural information processing systems. 2011.