عقیده‌کاوی نظرات مرتبط با دارو با استفاده از یادگیری چندنمونه­ای مبتنی بر ماشین­های بردار پشتیبان

زهره کریمی1\*

1 دانشکده فنی و مهندسی، دانشگاه دامغان، دامغان، ایران

z.karimi@du.ac.ir

چکیده: افزایش روزافزون نظرات کاربران در مورد محصول یا رخداد خاص در فضای مجازی، منجر به بروز زمینه تحقیقاتی عقیده­کاوی شده است که کاربردهای زیادی در اقتصاد، سیاست، پزشکی و غیره دارد. هدف عقیده­کاوی، تحلیل خودکار نظرات کاربران است و از کاربردهای مهم آن، تحلیل خودکار نظرات کاربران در مورد کارایی و اثرات جانبی داروها است. عقیده کاوی با استفاده از روش­های یادگیری ماشین بسیار مورد توجه قرار گرفته است و تمرکز اکثر پژوهش­ها روی یادگیری ماشین نظارتی است. از روش های رو به رشد در یادگیری ماشین، روش یادگیری چندنمونه­ای (MIL) است که نیاز به داده­های برچسب­دار کمتری در مقایسه با روش­های باناظر دارد. درMIL، داده ها در کیسه هایی گروه بندی شده و در فرایند آموزش از برچسب کیسه به جای برچسب هر نمونه برای یادگیری استفاده می­شود. هدف این پژوهش بررسی نتیجه­ی اعمال روش­های یادگیری چندنمونه­ای روی نظرات مرتبط با داروها است. در این مقاله ابتدا، بردار فرکانس کلمات-معکوس فرکانس سند از نظرات استخراج شده و سپس، روش توسعه یافته­ی ماشین بردار پشتیبان استاندارد برای یادگیری چندنمونه­ای جهت پیش­بینی مثبت یا منفی بودن نظرات مرتبط با دارو اعمال می­شود. ارزیابی های انجام شده روی مجموعه استاندارد Drug، بر اساس معیارهایF1 و دقت، کارایی روش پیشنهادی را در مقایسه با روش استاندارد ماشین بردار پشتیبان تایید می­کند.

**کلید واژه‌ها:** عقیده‌کاوی، نظرات مرتبط با دارو، یادگیری چندنمونه­ای، ماشین بردار پشتیبان، سیستم­های دارویی.

1- مقدمه

با پیشرفت روز افزون تکنولوژی، امکانات زیادی جهت ایجاد نظرات توسط افراد در مورد رخدادها و اشیای گوناگون و در دسترس عموم قرار گرفتن آن­ها پدید آمده است. شبکه­های اجتماعی، بلاگ ها و ...نمونه هایی از این امکانات هستند. رشد روزافزون نظرات تولید شده توسط کاربران، منجر به ایجاد زمینه­ی تحقیقاتی جدید و رو به گسترشی در پردازش زبان طبیعی به نام عقیده کاوی شده است که هدف آن، تحلیل خودکار نظرات کاربران است. این زمینه­ی جدید پژوهشی، عقیده‌کاوی، نظرکاوی یا تحلیل احساس نیز نامیده می شود و دامنه­ی گسترده­ای از کاربردها از جمله، اقتصاد، سیاست، پزشکی، بازاریابی و ... را تحت پوشش خود قرار می دهد [1]. در زمینه ی علم دارو، تحلیل خودکار نظرات کاربران، امکان تحلیل تعداد زیادی نظرات افراد در خصوص کارایی و اثرات جانبی داروها را فراهم می کند که جهت بهبود سیستم های مراقبت دارویی بکار می رود.

در سال­های اخیر، روش هایی برای عقیده­کاوی با استفاده از تکنیک­های پردازش زبان طبیعی و یادگیری ماشین ارائه شده است. عقیده­کاوی در سه سطح سند، جمله یا جنبه می­تواند انجام شود. عقیده کاوی در سطح سند یا جمله، یک مساله دسته بندی متن است که هدف آن، پیش­بینی مثبت، منفی یا خنثی بودن کل نظر (در سطح سند) یا هر جمله ی نظر (در سطح جمله) است. عقیده کاوی در سطح جنبه، در سطح ریزدانگی بیشتری انجام می­شود. عقیده­کاوی در مقایسه با روش­های دسته­بندی متن رایج دارای چالش­های بیشتری است. کوتاه بودن متون، تنک بودن ماتریس داده استخراج شده از متن و وجود کلمات عامیانه از جمله چالش­های این مساله هستند. از چالش­های پیش­بینی تمایل موجود در نظرات، هزینه­ی بدست آوردن داده­های برچسب­دار در روش­های باناظر است. لذا پیشنهاد می­شود به جای رویکردهای باناظر که به صورت رایج در این زمینه اعمال شده­اند، از روش­هایی استفاده شود که نیاز آن­ها به برچسب داده­ها در مقایسه با روش­های باناظر کمتر باشد. روش یادگیری چندنمونه­ای (MIL) از رویکردهایی است که بدین منظور در ادبیات یادگیری ماشین پیشنهاد شده و بسیار مورد توجه قرار گرفته است [2].

در روش یادگیری چندنمونه ای، داده­های آموزشی در مجموعه­هایی که کیسه نامیده می شوند، قرار داده می­شوند و نظارت مورد نیاز در مرحله­ی آموزش از طریق برچسب کیسه­ها و نه تمام نمونه ها فراهم می­شود. نیاز به برچسب­گذاری تعداد زیادی داده در روش های باناظر، صرف هزینه و زمان زیادی را می­طلبد، روش­های MIL با استفاده از کیسه­هایی که نمونه­های برخی از داده­ها در آن­ها می­تواند نامشخص باشد، یادگیری را انجام می­دهند. بدین صورت، زمان و هزینه­ی مورد نیاز برای برچسب­گذاری داده­ها کاهش می­یابد. این روش­ها در سال­های اخیر در کاربردهای زیادی مورد توجه قرار گرفته­اند [3, 4]. تمرکز این مقاله روی یادگیری چندنمونه­ای نظرات کاربران در خصوص داروها در سطح سند است. از آن­جا که در رویکردهای باناظر عقیده­کاوی، ماشین بردار پشتیبان (SVM) به موفقیت خوبی دست یافته است [5, 6, 7] ، بهره­گیری از توسعه یافته­ی SVM برای MIL (mi-SVM) مدنظر خواهد بود.

 در ادامه، در ابتدا، دانش زمینه و پژوهش­های مرتبط به ترتیب در بخش­های 2 و3 بررسی شده و سپس در بخش 4 روش پیشنهادی شرح داده می­شود. روش ارزیابی و نتایج در بخش 5 آمده است و نهایتاً در بخش 6 جمع­بندی و نتیجه­گیری ارائه می­شود.

2- دانش زمینه

در این بخش روش یادگیری چندنمونه ای شرح داده شده و سپس مساله­ی بهینه­سازی SVM بیان می­گردد.

2-1- یادگیری چندنمونه ای

یادگیری چندنمونه­ای، نوعی یادگیری باناظر ضعیف است که در آن، داده های آموزشی در مجموعه هایی که کیسه نامیده می شوند قرار می­گیرند. اطلاعات برچسب کیسه­ها موجود بوده و برچسب نمونه ها در دسترس نیست. این شیوه یادگیری در سال های اخیر بسیار مورد توجه قرار گرفته است، از انجا که برچسب­گذاری کیسه­ها در مقایسه با برچسب­گذاری نمونه­ها هزینه­ی کمتری دارد. کاربردهای تشخیص شی، الگوریتم های تشخیص بیماری به کمک رایانه، پیش بینی تاثیر دارو، دسته بندی صدا و متن از کاربردهای مهم یادگیری چندنمونه­ای هستند [2].

در یادگیری به شیوه­ی رایج، فرض بر آن است که مجموعه داده­های برچسب­دار $(x\_{i},y\_{i})\in R^{d}×Y$ که به صورت مستقل از توزیع ناشناخته نمونه­برداری شده­اند در دسترس هستند و هدف، یادگیری تابعی $f:R^{d}\rightarrow Y$ است که دسته­بند نامیده می­شود. در این­جا تمرکز روی دسته­های دودویی است یعنی $Y=\{+1, -1\}$. یادگیری چندنمونه­ای تعمیم این شیوه­ی یادگیری است که فرض ضعیف­تری در مورد داده­های برچسب­دار دارد. در MIL، الگوهای ورودی در کیسه­هایی گروه­بندی شده و برچسب به هر کیسه (و نه به هر نمونه) نسبت داده می­شود. به بیان دقیق­تر، الگوهای $x\_{1},…x\_{n}$ ، در کیسه­های $B\_{1},…,B\_{m}$ ، $B\_{I}=\{x\_{i}:i\in I\}$ گروه­بندی شده­اند، $I$ مجموعه­ اندیس­های هر کیسه را نشان می­دهد و در حالت کلی، کیسه­های گوناگون با یکدیگر اشتراک ندارند ($I⊆\{1,…, n\}$). به هر کیسه­ی $B\_{I}$ یک برچسب $Y\_{I}$ نسبمت داده شده است. برچسب کیسه­ها به این صورت مشخص می­شود که کیسه­های منفی فقط شامل نمونه­های منفی هستند و در کیسه­های مثبت، حداقل یک نمونه با برچسب مثبت وجود دارد. در حالت کلی، رابطه­ی بین $y\_{i}$ و کیسه­ی دربرگیرنده­ی $x\_{i}$، $B\_{I}$، به صورت $Y\_{I}=\max\_{i\in I}y\_{i}$ یا با مجموعه­ای از قیود خطی به صورت زیر بیان می­شود:

|  |  |
| --- | --- |
| (1) | $$\sum\_{i\in I}^{}\frac{y\_{i}+1}{2}\geq 1, ∀I s.t Y\_{I}=1 , y\_{i}=-1, ∀I s.t Y\_{I}=-1$$ |

*یک تابع* $f:R^{d}\rightarrow Y$*، تابع جداکننده­ی MIL نامیده می­شود اگر با توجه به داده­های چندنمونه­ای، برای تمام کیسه­ی* $B\_{I}$ *شرط زیر برقرار باشد:*

|  |  |
| --- | --- |
| (2) | $$Sgn(max\_{iϵI}f\left(x\_{i}\right))=Y\_{I}$$ |

2-2- SVM

*SVM یک روش دسته­بندی باناظر مبتنی بر تئوری یادگیری آماری است که کاربرد موفقی در مسئله­ی عقیده­کاوی داشته است* [6, 8] *مساله­ی بهینه­سازی SVM استاندارد به صورت زیر است:*

|  |  |
| --- | --- |
| *(3)* | $\min\_{w,b,ξ}\frac{1}{2}\left‖ω\right‖^{2}+C\sum\_{i=1}^{m}ξ\_{i}$$$such that ∀i, 1\leq i\leq n : y\_{i}\left(<ω,ϕ\left(x\_{i}\right)>+b\right)\geq 1-ξ\_{i} , ξ\_{i}\geq 0 $$$, y\_{i}\in \{+1,-1\}$ |

راه­حل مساله­ی مذکور به صورت $f\left(x\right)=<ω,ϕ\left(x\right)>+b$ است، $<.,.>$ نشان­دهنده­ی ضرب داخلی است، $ω\in R^{d}$ بردار وزن و $ϕ\left(.\right):R^{d}\rightarrow R^{D}$ تابع کرنل[[1]](#footnote-1) است که جهت انتقال داده به فضای با ابعاد بالاتر، $D$، به امید یافتن یک ابرصفحه­ی خطی در آن فضا بکار می­رود. $b$ بایاس و $ξ\_{i}$ متغیر slack است. در این مقاله مجال پرداخت به جزئیات مربوط به کرنل و قدرت تعمیم SVM نیست، برای مطالعه­ی جزئیات به [9] مراجعه شود.

3-پژوهش­های مرتبط

در این بخش، ابتدا مرور مختصری روی روش های عقیده­کاوی در سطح سند داریم و سپس روش­های عقیده­کاوی اعمال شده روی نظرات مرتبط با دارو با تمرکز بیشتر بررسی می­شود. SVM، دسته­بند بیزین و بیشینه آنتروپی از روش­های اعمال شده روی عقیده­کاوی هستند [10].مشخصه­های تک گرام و دو گرام از جمله ویژگی­هایی هستند که جهت استخراج بردار از متن بکاررفته­اند. شبکه­های عصبی و SVM نیز روی دو مجموعه داده­ی متعادل و نامتعادل اعمال شده­اند و نتیجه­ی آن­ها از جنبه­های گوناگون با یکدیگر مقایسه شده است [11]. Tang یک بازنمایی معنایی مخصوص نظرات کاربران را پیشنهاد داده است و عقیده­کاوی در سطح سند را انجام داده است [12]. با توجه به ویژگی تنک بودن ماتریس­های بازنمایی­های رایج عقیده­کاوی بازنمایی تنک[[2]](#footnote-2) برای عقیده­کاوی پیشنهاد شده است [7]. پژوهش­هایی نیز در زمینه­ی یادگیری چندنمونه­ای جهت عقیده­کاوی انجام شده است [13, 14, 15].

عقیده­کاوی در کاربردهای گوناگون تاکنون اعمال شده هر چند، در زمینه­ی نظرات مرتبط با دارو پژوهش­های اندکی انجام شده است. اولین پژوهش در زمینه تحلیل خودکار نظرات در مورد داروها با استفاده از قوانین [16] و منابع واژگان احساسی [17, 18] از جمله [19] SentiWordNet است. منابع واژگان احساسی مجموعه­ای شامل لغات با بار احساسی مثبت یا منفی و امتیاز این لغات است. این امتیاز با توجه به نقش گرامری کلمات می­تواند متفاوت باشد. الگوریتم­های دسته­بندی باناظر موفق در زمینه­ی عقیده­کاوی شامل دسته­بندی بیزین ساده، درخت­های تصمیم، دسته­بند k-نزدیکترین همسایه و SVM از دسته­بندهایی هستند که جهت تخمین میزان قطبیت (مثبت، منفی یا خنثی بودن) پست­های توییت های مرتبط با سلامتی بررسی و مقایسه شده است [20]. Tanveer و همکاران از کیسه کلمات[[3]](#footnote-3) (BoW) برای بازنمایی نظرات استفاده کرده­اند [21]. Mishra و همکاران یک سیستم مبتنی بر ماشین بردار پشتیبان را جهت تعیین قطبیت نظرات مرتبط با دارو پیشنهاد داده­اند که از منبع واژگان SentiWordNet جهت امتیازدهی به توکن ها بهره برده است [22].

Gräßer و همکاران یک مجموعه داده از نظرات در زمینه­ی داروها را از سایت Drugs.com جمع آوری کرده­اند که شامل نظرات بیماران و متخصصین سلامتی در مورد داروها است [23]. آن­ها از رگرسیون لجستیک برای دسته­بندی این نظرات استفاده کرده­اند. یادگیری عمیق نیز در سال های اخیر، جهت دسته بندی نظرات مرتبط با دارو بکاررفته است [24]. ارائه­ی کدگذاری شده­ی دو مسیره از انتقال دهنده­ها[[4]](#footnote-4) از مدل­های معروف یادگیری عمیق است که کاربردهای زیادی در پردازش متن داشته است. این مدل جهت دسته­بندی نظرات مرتبط با داروها بکار رفته است [25]. معماری­های گوناگون یادگیری عمیق از جمله شبکه­های عصبی کانولوشنال، شبکه­های عصبی بازگشتی[[5]](#footnote-5) با حافظه­ی کوتاه مدت[[6]](#footnote-6) روی نظرات مرتبط با دارو اعمال شده و با یکدیگر مقایسه شده­اند [24]. بر اساس دانش نویسنده، تاکنون تاثیر یادگیری چندنمونه­ای بر این مجموعه داده بررسی نشده است.

4-روش پیشنهادی

هدف این مقاله، اعمال روش یادگیری چندنمونه­ای mi-SVM به داده­های نظرات مرتبط با داروها است. گام­های روش اعمال شده در شکل 1 آمده است. ابتدا توکن­ها که همان کلمات هستند از نظرات استخراج شده، سپس کلمات توقف[[7]](#footnote-7) از لیست کلمات حذف می­شود، این کلمات، کلماتی همچون to، the و ... هستند که زیاد تکرار می­شوند اما حاوی بارمعنایی خاصی نیستند. سپس، بردار ویژگی مربوط به هر نظر استخراج می­شود. برای این­کار از مشخصه­ی رایج فراوانی کلمات-معکوس فراوانی سند (tf-idf) استفاده می­شود که رابطه­ی آن به صورت زیر است:

|  |  |
| --- | --- |
| (4) | $$tf-idf\_{i,j}=tf\_{i,j}×log\frac{n}{df\_{i}}$$ |

$tf\_{i,j}$ *فراوانی عبارت (کلمه­) iم در نظر j م است و* $df\_{i}$ *فراوانی کلمه­ی iم در تمام اسناد (نظرات) است. به این ترتیب نظر iم به صورت بردار* $x\_{i}=(x\_{i1}, …,x\_{id})$ *بازنمایی می­شود که در آن* $x\_{ij}=tf-idf\_{i,j}$*، d، تعداد کلمات یا ابعاد هر نظر است. این داده­ها در کیسه­هایی که با یکدیگر اشتراک ندارند قرار می­گیرند و برچسب هر کیسه به این صورت مشخص می­شود که کیسه­هایی که تمام نمونه­های آن­ها برچسب منفی داشته باشند، برچسب منفی و کیسه­هایی که حاوی حداقل یک برچسب مثبت باشند، برچسب مثبت می­گیرند. نهایتاً روش mi-SVM روی این داده­ها اعمال می­شود. در ادامه، جزئیات این روش، ارائه می­شود.*



**شکل 1-** مراحل آموزش روش پیشنهادی

4-1- mi-SVM

*مساله­ی بهینه­سازی* SVM *برای* MIL*، به صورت زیر بیان می­شود:*

|  |  |
| --- | --- |
| (5) | $$\min\_{\left\{y\_{i}\right\}}\min\_{w,b,ξ}\frac{1}{2}\left‖ω\right‖^{2}+C\sum\_{i=1}^{m}ξ\_{i}$$$$such that ∀i, 1\leq i\leq n : y\_{i}\left(<ω,ϕ\left(x\_{i}\right)>+b\right)\geq 1-ξ\_{i} ,$$$ ξ\_{i}\geq 0$$ y\_{i}\in \{+1,-1\}$$\sum\_{i\in I}^{}\frac{y\_{i}+1}{2}\geq 1, ∀I s.t Y\_{I}=1 , y\_{i}=-1, ∀I s.t Y\_{I}=-1$  |

*در SVM استاندارد، برچسب نمونه­ها داده شده در حالی که در مساله­ی بالا، برچسب داده­های موجود در کیسه­های مثبت، ناشناخته است و بنابراین بهینه­سازی علاوه بر پارامترهای SVM استاندارد، روی برچسب داده­های آموزشی نیز انجام می­شود. قید دوم در مساله­ی بهینه­سازی مذکور، قید نشان دهنده­ی ارتباط برچسب کیسه­ها با برچسب نمونه­ها است. مساله­ی بهینه­سازی (4)، یک مساله­ی بهینه­سازی برنامه­ریزی عدد صحیح مختلط[[8]](#footnote-8) است. حل این مساله در مقیاس متوسط هم کارا نبوده و نیاز به استفاده از روش­های اکتشافی در حل آن است. یک روش رایج در حل این مسائل، بهینه­سازی متناوب در دو گام است: (1) با داشتن برچسب­های داده­ها، مساله­ی بهینه­سازی حل شود که در این صورت، مساله به یادگیری یک SVM استاندارد و بنابراین یک برنامه­نویسی چندجمله­ای[[9]](#footnote-9) کاهش می­یابد که به صورت کارا قابل حل است. برچسب داده­ها یا همان متغیرهای مخفی[[10]](#footnote-10)، برچسب داده­های کیسه­ی مثبت است که در این گام فرض شده آن­ها را داریم و (2) با داشتن پارامترهای ابرسطح SVM، مقادیر برچسب داده­ها جهت کمینه کردن تابع هدف محاسبه شود که شامل بروزرسانی برچسب* $y\_{i}$ *یک داده است. این دو گام تکرار می­شوند. برای اجرای گام (1) در اولین مرتبه، نیاز به مقداردهی اولیه برچسب داده­های کیسه­های مثبت است. برای این مقداردهی اولیه، می توان برچسب کیسه­های مثبت را مثبت در نظر گرفت. گام­های شبه­کد در شکل 1 نشان داده شده است. همانند SVM استاندارد، امکان بهره­گیری از کرنل روی داده­های ورودی است که می­تواند باعث قدرت تعمیم بهتر مدل شود.*

|  |
| --- |
| 1. *مقداردهی اولیه* $y\_{i}=Y\_{I}$ *به ازای هر* $i\in I$
2. *گام های زیر را تا زمانی که برچسب کیسه­های مثبت تغییر می­کند تکرار کن:*
* *مساله­ی بهینه­سازی SVM را با برچسب­های مشخص شده برای داده­ها حل کن.*
* *خروجی* $f\_{i}=<w,ϕ(x\_{i})>+b$ *را برای تمام* $x\_{i}$ *ها در کیسه­های مثبت مشخص کن.*
* *برای هر* $i\in I$ *در کیسه­های مثبت،* $y\_{i}=sgn\left(f\_{i}\right)$
* *برای هر کیسه­ی مثبت* $B\_{I}$*: اگر* $\sum\_{i\in I}^{}\frac{1+y\_{i}}{2}==0$ *،* $i^{\*}=argmax\_{i\in I}f\_{i}$ *و* $y\_{i}^{\*}=1$ *قرار بده.*
 |
| شکل 2- شبه کد mi-SVM |

5- ارزیابی

مجموعه داده­ی بکاررفته در این مقاله، مجموعه­ای از نظرات در مورد دارو است که از سایت Drugs.com بدست آمده است [23]. به هر نظر، امتیازی بین 0 تا 9 نسبت داده شده است که میزان رضایت بیمار از دارو را مشخص می­کند. مجموعه داده شامل 215063 نظر است. ایجاد کننده مجموعه­ی داده، نظرات را بر اساس نرخ رضایت کاربر به سه دسته منفی (دسته­ی صفر، نرخ کوچکتر مساوی 4، دسته­ی خنثی (دسته­ی 1، نرخ رضایت بین 5 و 6) و دسته­ی مثبت (دسته­ی 2، نرخ بزرگتر مساوی 7) تقسیم کرده است. ما زیرمجموعه­ای از این داده­ها شامل 3000 داده را به صورت تصادفی حاوی داده­های مثبت و منفی انتخاب کرده­ایم به نحوی که نسبت داده­های هر دسته برابر با نسبت آن­ها در مجموعه داده­ی اصلی است. بعد از توکنایز کردن و حذف واژگان توقف، کلمات متن (تک گرام ها) استخراج شده­اند و سپس tf-idf کلمات محاسبه شده­اند. جهت کاهش ابعاد ورودی و کارایی بیشتر، کلماتی که کمتر از 50 مرتبه در نظرات ظاهر شده باشند، حذف شده­اند. نهایتاً داده­ها در کیسه­هایی تقسیم شده و برچسب کیسه به عنوان ورودی به الگوریتم داده می­شود.

جهت ارزیابی مدل، ماتریس اغتشاش را به صورت زیر درنظر بگیرید:

|  |  |
| --- | --- |
|  | برچسب واقعی |
| برچسب پیش­بینی شده | منفی | مثبت |
| منفی | TN | FN  |
| مثبت | FP | TP |

ملاک­های ارزیابی عبارتند از: دقت (Accuracy) و F1 هستند که در رابطه­های (6) و (7) تعریف شده­اند.

|  |  |
| --- | --- |
| (6) | $$Accuracy=\frac{TP+TN}{TP+TN+FP+FN}$$ |
| (7) | $F1=\frac{TP}{TP+\frac{FP+FN}{2}}$ |

جهت مقایسه، روش پیشنهادی با روش SI-SVM که در واقع روش مبتنی بر یادگیری چندنمونه­ای نیست مقایسه شده است. این روش، مبنای مقایسه­ی روش­های یادگیری چندنمونه­ای در بسیاری از مقالات است [26, 27, 28]؛ بدین معنا که در صورتی که روش یادگیری چندنمونه­ای روی مجموعه داده­ای نتیجه­ای بهتر از این روش داشته باشد می­توان به این نتیجه رسید که یادگیری چندنمونه­ای برای مجموعه­ی مورد نظر مناسب است. با توجه به این­که براساس دانش نویسنده، تاکنون روش یادگیری چندنمونه­ای جهت دسته­بندی نظرات مرتبط با دارو بکار نرفته است، از الگوریتم SI-SVM به عنوان روش مبنا جهت ارزیابی روش پیشنهادی استفاده می­شود. در الگوریتم SI-SVM برچسب نسبت داده شده به هر نمونه، برچسب کیسه­ی آن است. سپس، اطلاعات کیسه در مراحل بعد در نظر گرفته نمی­شود و دسته­بند ماشین بردار پشتیبان استاندارد به نمونه­ها اعمال شده و برچسب نمونه­ها نهایتا مشخص می­شود. در صورتی که پیش­بینی برچسب کیسه مدنظر باشد، برچسب هر کیسه در صورتی که مثبت است که حداقل حاوی یک نمونه با برچسب مثبت باشد.

آزمایشات با دو اعتبارسنجی متقاطع[[11]](#footnote-11) تودرتو انجام شده است. اعتبارسنجی متقاطع درونی، مقادیر ابرپارامترها را از طریق جستجوی توری[[12]](#footnote-12) می­یابد و اعتبار سنجی متقاطع بیرونی، با بهره­گیری از پارامترهای مشخص شده، مدل را ارزیابی می­کند. کل این مراحل، 10 بار تکرار شده و نهایتاً میانگین و انحراف معیار نتایج در جدول 1 به ازای شش مقدار متفاوت برای تعداد نمونه­های هر کیسه گزارش شده است. همان­طور که مشخص است، نتایج بدست آمده بر اساس هر دو ملاک دقت و F1 کارایی روش یادگیری چندنمونه­ای را به ازای کیسه­های با اندازه­ی متفاوت تایید می­کند.

جدول 1- مقایسه­ی روش SI-SVM و mi-SVM روی مجموعه داده نظرات دارو

|  |  |  |
| --- | --- | --- |
| اندازه کیسه | SI-SVM | mi-SVM |
|  | Accuracy(%) | F1(%) | Accuracy (%) | F1(%) |
| 10 | 38/75(0.64) | 16/85(13/0) | **63/75****(63/75)** | **30/85****(30/0)** |
| 20 | 30/76(0.59) | 55/85(17/0) | **39/76****(73/0)** | **61/85****(5/0)** |
| 30 | 89/75(0.79) | 32/85(31/0) | **48/76****(65/0)** | **40/85****(26/0)** |
| 40 | 92/75(06/1) | 13/85(50/0) | **38/76****(22/1)** | **22/85****(51/0)** |
| 50 | 11/76(95/0) | 30/85(28/0) | **98/76****(89/0)** | **43/85****(26/0)** |
| 60 | 81/76(47/1) | 81/85(55/0) | **48/77****(46/1)** | **89/85****(52/0)** |

6- نتیجه­گیری

در این مقاله، روش یادگیری چندنمونه­ای مبتنی بر ماشین بردار پشتیبان روی مجموعه داده­های نظرات مرتبط با تاثیرات مثبت و منفی داروها اعمال شد. این روش، تعمیمی از ماشین بردار پشتیبان است که در آن برچسب داده­ها در مرحله­ی آموزش به صورت ضعیف­تر و در قالب برچسب کیسه مشخص شده است. جهت اعمال این روش به داده­های نظرات مرتبط با دارو، ابتدا نظرات که در قالب متن­های کوتاه و به صورت داده­ی بدون ساختار هستند به صورت بردار بازنمایی می­شوند. بدین منظور، مشخصه­ی آماری فراوانی عبارت-معکوس فراوانی سند از نظرات استخراج می­شود. سپس داده­ها در قالب کیسه­هایی گروه­بندی شده و داده­ها همراه با برچسب کیسه­ها به عنوان ورودی روش یادگیری چندنمونه­ای درنظر گرفته می­شوند. این روش با معیارهای دقت و F1 مورد ارزیابی قرار گرفته و نتایج بدست آمده، کارایی روش یادگیری چندنمونه­ای را در مقایسه با روش استاندارد یادگیری تایید می­کند. جهت پژوهش­های آتی، بهره­گیری از یادگیری چندنمونه­ای جهت عقیده­کاوی در سطح جمله و جنبه پیشنهاد می­گردد.

**مراجع:**

|  |  |
| --- | --- |
| [1]  | K. Ravi and R. Vadlamani, "A survey on opinion mining and sentiment analysis: tasks, approaches and applications," Knowledge-based systems 89, pp. 14-46, 2015.  |
| [2]  | M.-A. Carbonneau, C. Veronika, G. Eric and G. Ghyslain, "Multiple instance learning: A survey of problem characteristics and applications," Pattern Recognition 77 , pp. 329-353, 2018.  |
| [3]  | P. J. Sudharshan, P. Caroline, S. Fabio, E. O. Luiz, H. Laurent and H. Paul, "Multiple instance learning for histopathological breast cancer image classification," Expert Systems with Applications 117, pp. 103-111, 2019.  |
| [4]  | K. Bhattacharjee, P. Millie, Z. Yu-Dong and C. S. Suresh, "Multiple Instance Learning with Genetic Pooling for medical data analysis," Pattern Recognition Letters 133 , pp. 247-255, 2020.  |
| [5]  | J.-R. Chang, L. Hsin-Ying, C. Long-Sheng and C. Chia-Wei, "Novel feature selection approaches for improving the performance of sentiment classification," Journal of Ambient Intelligence and Humanized Computing, pp. 1-14, 2020.  |
| [6]  | A. P. Gopi, S. J. R. Naga, N. V. Lakshman and S. K. Satya, "Classification of tweets data based on polarity using improved RBF kernel of SVM," International Journal of Information Technology, pp. 1-16, 2020.  |
| [7]  | Z. Karimi and R. Ramezani, "Sparse Representation for Sentiment Analysis," in In 2020 6th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS), 2020.  |
| [8]  | L. Wang, N. Jianwei and Y. Shui, "Sentidiff: Combining textual information and sentiment diffusion patterns for twitter sentiment analysis," IEEE Transactions on Knowledge and Data Engineering 32, no. 10, pp. 2026-2039, 2019.  |
| [9]  | C. Cortes and V. Vladimir, "Support vector networks," Machine Learning, 20, p. 273 – 297, 1995.  |
| [10]  | B. Pang, L. Lee and V. Shivakumar, "Thumbs up? Sentiment classification using machine learning techniques," in Proceedings of the conference on empirical methods in natural language processing (EMNLP), Prague, 2002.  |
| [11]  | R. Moraes, J. F. Valiati and W. P. G. Neto.‏, "Document-level sentiment classification: An empirical comparison between SVM and ANN," Expert Systems with Applications 40.2, pp. 621-633, 2013.  |
| [12]  | D. Tang, "Sentiment-specific representation learning for document-level sentiment analysis," in Proceedings of the eighth ACM international conference on web search and data mining, 2015.  |
| [13]  | N. Pappas and P.-B. Andrei, "Explaining the stars: Weighted multiple-instance learning for aspect-based sentiment analysis," in In Proceedings of the 2014 Conference on Empirical Methods In Natural Language Processing (EMNLP), 2014.  |
| [14]  | D. Kotzias, D. Misha, B. Phil and d. F. Nando, "Deep multi-instance transfer learning," arXiv preprint arXiv:1411.3128, 2014.  |
| [15]  | D. Kotzias, D. Misha, D. F. Nando and S. Padhraic, "From group to individual labels using deep features," in In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, 2015.  |
| [16]  | J.-C. Na, M. K. Wai Yan, S. K. Christopher, F. Schubert, C. Yun-Ke and T. Yin-Leng, "Sentiment classification of drug reviews using a rule-based linguistic approach," in In International conference on asian digital libraries.  |
| [17]  | L. Goeuriot, N. Jin-Cheon, M. K. Wai Yan, K. Christopher, C. Yun-Ke, T. Yin-Leng and K. Jung-Jae, "Sentiment lexicons for health-related opinion mining," in In Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, 2012.  |
| [18]  | M. Wiley, C. Jin, V. Hristidis and K. Esterling, "Pharmaceutical drugs chatter on online social networks," Journal of biomedical informatics 49, pp. 245-254, 2014.  |
| [19]  | F. Sebastiani and E. Andrea, "Sentiwordnet: A publicly available lexical resource for opinion mining," in In Proceedings of the 5th International Conference on Language Resources and Evaluation, 2006.  |
| [20]  | V. Bobicev, S. Marina, J. Yasser and S. David, "Learning sentiments from tweets with personal health information," in In Canadian conference on artificial intelligence, Berlin, Heidelberg,, 2012.  |
| [21]  | A. Tanveer, S. M. S. David and I. Diana, "Can i hear you? sentiment analysis on medical forums," in InProceedings of the sixth international joint conference on natural language processing 2013, 2013.  |
| [22]  | A. Mishra, M. Ankit and A. Sanchit, "Towards automatic pharmacovigilance: analysing patient reviews and sentiment on oncological drugs," in In 2015 IEEE International Conference on Data Mining Workshop (ICDMW), 2015.  |
| [23]  | F. Gräßer, K. Surya, M. Hagen and Z. Sebastian, "Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning," In Proceedings of the 2018 International Conference on Digital Health , pp. 121-125, 2018.  |
| [24]  | C. Colón-Ruiz and S.-B. Isabel, "Comparing deep learning architectures for sentiment analysis on drug reviews," Journal of Biomedical Informatics 110, 2020.  |
| [25]  | B. Biseda and M. Katie, "Enhancing Pharmacovigilance with Drug Reviews and Social Media," arXiv preprint arXiv:2004.08731, 2020.  |
| [26]  | S. Ray and C. Mark, "Supervised versus multiple instance learning: An empirical comparison," in In Proceedings of the 22nd international conference on Machine learning, 2005.  |
| [27]  | E. Alpaydın, C. Veronika, L. Marco and M. T. David, "Single-vs. multiple-instance classification," Pattern recognition 48, no. 9, pp. 2831-2838.  |
| [28]  | R. C. Bunescu and J. M. Raymond, "Multiple instance learning for sparse positive bags," In Proceedings of the 24th international conference on Machine learning, pp. 105-112, 2007.  |

Opinion mining of Drug Reviews using Support Vector Machine for Multiple Instance Learning

\*Zohre Karimi1

1 School of Engineering, Damghan University, Damghan, Iran

z.karimi@du.ac.ir

*Abstract*— **An increasing growth in generating users' comments on the product or special event in social media and internet has led to the new field of research called opinion mining. The opinion mining has many applications in economics, politics, medicine, and so on. The purpose of the opinion mining is to analyze user comments, automatically. In the pharmaceutical field, automated analysis of user opinions about the efficiency and side effects of drugs are an important knowledge for enhancing the performance of pharmacovigilance systems. The prior works on opinion mining of drug sentiments are focused on learning supervised classifiers. Multiple instance learning (MIL) is a weakly supervised method, where data is grouped in bags and the bag label is used in the training instead of the label of each sample. The aim of this study is to apply MIL method to drug-related comments. In this paper, first, the term frequency-inverse of document frequency is extracted from comments and then the generalized support vector machine for MIL is applied to predict the label of the drug reviews. The evaluations performed on the benchmark dataset, Drug, based on F1 criteria and accuracy, confirm the performance of the proposed method compared to the standard method of support vector machine.**

Keywords—Opinion mining, Drug reviews, Multiple instance learning, Support vector machine, pharmacovigilance system.

1. Kernel [↑](#footnote-ref-1)
2. Sparse [↑](#footnote-ref-2)
3. Bag Of Words [↑](#footnote-ref-3)
4. Bidirectional Encoder Representation from Transformer [↑](#footnote-ref-4)
5. Recurrent [↑](#footnote-ref-5)
6. Short-term [↑](#footnote-ref-6)
7. Stopwords [↑](#footnote-ref-7)
8. Mixed integer programming problem [↑](#footnote-ref-8)
9. Quadratic Programming [↑](#footnote-ref-9)
10. Hidden [↑](#footnote-ref-10)
11. Cross-validation [↑](#footnote-ref-11)
12. Grid search [↑](#footnote-ref-12)