مروری بر روش‌های مواجهه با تغییرات زمانی در دسته‌بندی متن

مرضیه سپهر[۱]، مرتضی براری[۲]،سمیه کافی[۳]

[۱] دانشجوی کارشناسی ارشد مهندسی کامپیوتر، دانشگاه صنعتی مالک اشتر

[۲] استادیار مجتمع فناوری اطلاعات و ارتباطات، دانشگاه صنعتی مالک‌اشتر

[۲] دانشجوی دکترای مهندسی کامپیوتر، دانشگاه صنعتی مالک‌اشتر

مسئول مکاتبات :مرضیه سپهر

چکیده

رشد سریع اطلاعات به خصوص رشد سریع وب، دسته‌بندی متون را به صورت دستی غیرممکن می‌کند. از این جهت دسته‌بندی خودکارمتن به عنوان یکی از زمینه‌های تحقیقاتی مهم مورد توجه محققین قرار گرفته است. در دسته بندی متن عموما داده ها به صورت آفلاین جمع آوری شده و آموزش بر روی این آن‌ها انجام می‌شود سپس مدل آموزش دیده به این روش، می‌تواند برای پیش بینی داده‌های دیده نشده‌ی ورودی مورد استفاده قرار بگیرد. اما در دنیای واقعی متون مدام در حال به روز رسانی هستند و توزیع داده ها و مفاهیم متنی به مرور زمان دستخوش تغییر می‌شوند به طوری که موجب اثرگذاری بر روی عملکرد دسته بند می‌شوند. دراین تحقیق، ابتدا به دسته بندی انواع تغییرات ممکن در طول زمان، روشهای تشخیص آن و سپس انواع روش‌های مواجه با تغییرات زمانی در دسته‌بندی متون پرداخته شده است و سپس به مقایسه‌ی روش‌ها با توجه به نوع تغییرات پرداخته شده است.

کلمات کلیدی: دسته بندی متن، زمان در دسته‌بندی متن، اثرات زمانی

۱. مقدمه

رشد سریع اطلاعات به خصوص رشد سریع وب، دسته‌بندی متون موجود را به صورت دستی غیرممکن می‌کند. از این جهت دسته‌بندی خودکارمتن به عنوان یکی از زمینه‌های تحقیقاتی مهم مورد توجه محققین قرار گرفته است. هدف از دسته‌بندی خودکار متن اختصاص خودکار متون، به دسته‌های از پیش تعیین‌شده، با توجه به محتوای آن‌هاست(Sebastiani۲۰۰۲). در دسته بندی متن عموما داده ها ابتدا به صورت آفلاین جمع آوری شده و آموزش بر روی آن‌ها انجام می‌شود سپس مدل آموزش دیده به این روش، می‌تواند برای پیش بینی داده‌های دیده نشده‌ی ورودی مورد استفاده قرار بگیرد. در شرایطی که داده‌ها مدام در حال به روز رسانی هستند، توزیع آنها می‌تواند در طول زمان تغییر کند به طوری که موجب اثرگذاری بر روی عملکرد دسته بند شود. لذا مفهوم زمان را می‌توان به عنوان یک بعد از اطلاعات در نظر گرفت.

دو تعریفی که مرجع (Alonso, Gertz et al.۲۰۰۷) از مفهوم زمان در بازیابی اطلاعات دارد بدین شرح است:۱) یک فاصله که دو نقطه‌ی تاریخ را در یک زنجیره‌ی تاریخی از هم جدا می‌کند.۲) یک زنجیره که در آن وقایع در یک توالی تغییر ناپذیر از گذشته، حال و تا آینده رخ می‌دهد. در تعریف دوم منظور از واقعه یا رخداد[۱]یک اتفاق است که به اندازه‌ی کفایت رخ می‌دهد به نحوی که در داده‌ها تاثیر گذار باشد، مثل یک گردهمایی یا فعالیت اجتماعی مانند انتخابات. بر این اساس ،زمان، یک پارامتر تاثیر گذار در کاربردهای بازیابی اطلاعات به حساب آمده و تغییرات و نحوه‌ی مواجه با آن در تحقیقات به عناوین مختلف مورد بررسی قرار گرفته است.

دسته‌ای از تحقیقات از این تغییرات با عنوان رانش مفهوم[۲](Forman ۲۰۰۶, Nishida and Yamauchi ۲۰۰۹, Šilić and Bašić ,۲۰۱۲ D'hondt, Verberne et al.۲۰۱۴)و دسته ای دیگر با تغییر زمانی[۳](Mourão, Rocha et al. ۲۰۰۸, Rocha, Mourão et al. ۲۰۰۸, Salles, Rocha et al. ۲۰۱۰, Rocha, Mourão et al. ۲۰۱۳) یاد کرده اند. بر اساس تحقیقات انجام شده در این مقاله،۳ تفاوت اصلی بین دسته‌ی اول و دسته‌ی دوم مشاهده شد. فرض اصلی دسته‌ی اول این است که داده‌ها به ترتیب زمانی ذخیره نشده‌اند و داده‌های آزمایشی همیشه جدیدتر از داده‌های آموزشی هستند(Rocha, Mourão et al.۲۰۱۳) بر خلاف دسته‌ی دوم که فرض می‌کند داده‌های آزمایشی می‌توانند از هر بازه‌ی زمانی باشند. در تحقیقات دسته‌ی اول، داده‌ی مورد بررسی اغب جریان داده‌ی متنی[۴] هستند. در حالی‌که تحقیقات دسته‌ی دوم اغلب بر روی داده‌های متنی معمولی کار

[1] Event
[2] Concept drift
[3] Temporal variation

[۴] جریان داده دنباله ای پیوسته از داده‌ها است که به همان ترتیب دریافت مورد دسترسی قرار می گیرند.

کردند. تحقیقات دسته‌ی اول اغلب سعی در شناسایی دقیق زمان تغییر و نوع تغییر داشته و سپس برای مواجهه با تغییر روشی را پیشنهاد داده اند درحالی‌که در تحقیقات دسته‌ی دوم عموما از روش‌های ابتکاری و یا پیشفرض‌ها برای تعیین زمان مواجهه با تغییر استفاده کرده‌اند.

در این تحقیق سعی شده تا حد امکان یک دید کلی و همه جانبه در این زمینه‌ی اثر زمان در دسته‌بندی خودکار متن و بررسی روش‌های مواجهه با آن ارائه شود. باید توجه داشته که پیش از بررسی روش‌های مواجهه با تغییر، ابتدا به سه سوال مهم باید پاسخ داده شود:

- اثرگذاری تغییرات زمانی به چه نحوی قابل مشاهده هستند؟
- تغییرات زمانی به چه تناوبی رخ می‌دهند؟
- چه زمانی باید نسبت به تغییرات عکس‌العمل نشان داد؟

بر این اساس، در تحقیق حاضر ابتدا به سه سوال مذکور به ترتیب در بخش‌های دوم، سوم و چهارم پاسخ داده می‌شود و سپس در بخش پنجم به بررسی روش‌های مواجهه با تغییر پرداخته خواهد شد. در بخش پنجم به مقایسه‌ی روش‌ها با توجه به نوع تغییرات پرداخته شده است و در نهایت نتیجه‌گیری ارائه می‌گردد.

۲. اثرات زمانی

از یک دیدگاه تغییرات زمانی را می‌توان بر اساس نحوه‌ی اثرگذاری آن به ۳ دسته تقسیم کرد(Rocha, Mourão et al. ۲۰۱۳, D'hondt, Verberne et al.).

- تغییر در توزیع کلاسها[۱]
- تغییر در توزیع کلمات[۲]
- تغییر در شباهت کلاسها[۳]

در دسته‌ی اول از این نوع تغییر، گذر زمان موجب تغییر در توزیع کلاس‌ها می‌شود. یعنی ممکن است دسته‌هایی از بین بروند یا دسته‌های جدیدی به وجود آیند. علت این مسئله تقسیم شدن یک دسته به دسته های کوچکتر و یا ادغام چند دسته با یکدیگر است. برای مثال در ساختار سلسله‌مراتبی موضوعات زیرمجموعه‌ی علوم کامپیوتر در پایگاه ACM ، در سال ۱۹۶۴ دو موضوع بازیابی اطلاعات و هوش مصنوعی متعلق به یک کلاس برنامه‌های کاربردی بودند ولی در حال حاضر هر کدام به ترتیب متعلق به کلاس‌های سیستمهای اطلاعاتی[۴] و روشهای محاسباتی[۵] هستند.

گاهی اوقات با گذر زمان، کلمات ممکن است در دسته‌ی خاصی از متون وارد شوند، از یک دسته خارج شوند، از دسته‌ای به دسته‌ی دیگر مهاجرت کرده و یا اینکه وزن آن‌ها در یک دسته تغییر کند. در مرجع (Wenerstrom and Giraud-Carrier ۲۰۰۶) از این نوع تغییر با نام تغییر متنی[۶] نام برده شده است. انتقال کلماتی مانند انرژی هسته‌ای، غنی سازی، سانتیفیوژها وغیره از دامنه‌ی فیزیک به دامنه‌ی اخبار سیاسی در چند سال اخیر مثالی از مهاجرت کلمات از دسته‌ای به دسته‌ی دیگر است و ورود کلماتی مانند یارانه در دامنه‌ی اخبار سیاسی در زبان فارسی را می‌توان مثالی از ورود کلمات جدید به یک کلاس دانست.

تغییر در شباهت کلاس‌ها اشاره دارد به تکامل تدریجی شباهت بین کلاس‌ها در طول زمان. برای مثال در گذشته، دسته‌ی جرم‌شناسی و بیولوژی به هم شباهتی نداشتند. اما اخیرا به علت استفاده از علم بیولوژی در شناسایی مجرمین این دو دسته به هم ارتباط بیشتری پیدا کرده اند. لازم به ذکر است که تغییر در شباهت کلاس‌ها به معنی تغییر در توزیع کلمات بین دو کلاس متفاوت است درحالی که تغییر در توزیع کلمات به تغییر کلمات درون یک کلاس اشاره دارد که می‌تواند منجر به افزایش شباهت بین دو کلاس نیز بشود.

۳. انواع تغییر

از دیدگاهی دیگر می‌توان تغییرات زمانی را بر اساس دوره‌ی تناوب آن به ۳ دسته‌ی زیر، تقسیم کرد (D'hondt, Verberne et al.2014):

- تغییر ناگهانی[۷]
- تغییر تدریجی[۸]
- تغییر بازگشت کننده[۹]

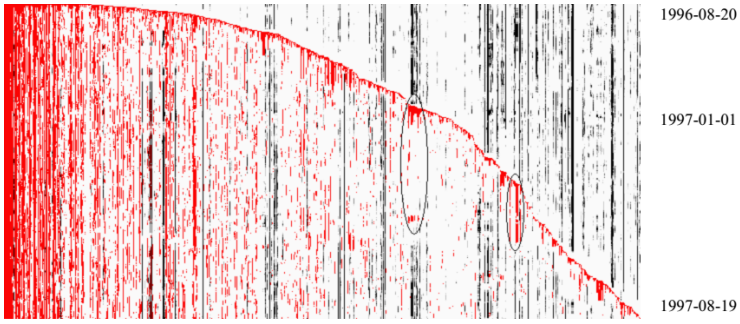
---

[1] Class distribution
[2] Term distribution
[3] Class similarity
[4] Information system
[5] Computing Methodologies
[6] Contextual drift
[7] Sudden drift
[8] Gradual
[9] Recurrent

در تغییرات ناگهانی لحظاتی وجود دارد که درآن توزیع داده در پیکره (توزیع کلاس یا کلمات) به صورت قابل ملاحظه‌ای تغییر می‌کند. این نوع تغییرات در سیستم‌هایی که اطلاعات را متناسب با علاقه‌ی کاربر فیلتر می‌کنند، رایج است. در چنین سیستم‌هایی تغییر ناگهانی علاقه‌ی کاربر موجب کاهش عملکرد سیستم می‌شود. در بعضی از تحقیقات از این نوع تغییر با نام تغییر جهت مفهوم[1] یاد شده است (Nishida and Yamauchi۲۰۰۹) .

در تغییرات تدریجی نرخ تغییر پایین است. درساده ترین حالت تغییر تدریجی، نمونه‌های آموزشی از ۲ توزیع Pj و Pj+1 تشکیل شده‌اند و به مرور زمان، احتمال مشاهده‌ی مثال‌هایی با توزیع Pj کاهش و احتمال مشاهده‌ی مثال‌هایی با توزیع Pj+1 افزایش میابد. نوع دیگری از تغییر تدریجی، تغییر افزایشی[2] است که در آن داده‌ها از بیش از دو نوع توزیع تشکیل می‌شوند. در این نوع از تغییر تفاوت بین توزیع‌ها بسیار کم است و تغییرات در دوره‌ی طولانی‌تری از زمان دیده می‌شوند(Brzezinski and Stefanowski۲۰۱۴). بعضی از مراجع تغییرات افزایشی را دسته‌ای جدا از تغییرات تدریجی در نظر گرفته اند(Ditzler۲۰۱۲).تغییرات بازگشت‌کننده، هر چند وقت یکبار رخ داده و موجب تغییر در مفاهیم ویا توزیع آن‌ها می‌شوند ولی پس از چندی مجددا مفاهیم قدیمی تکرار می‌شوند و این سیر ادامه پیدا می‌کند(Gama, Žliobaitė et al.۲۰۱۴). به عقیده‌ی(Forman۲۰۰۶) این نوع تغییر بیشتر درمتون خبری رخ می‌دهد. هر چند بررسی‌های حاصل از تحقیق (Šilić and Bašić۲۰۱۲) نشان از تغییرات تدریجی در متون خبری است. لذا تشخیص نوع تغییر به پیکره‌ی مورد بررسی وابسته است و نیاز به انجام آزمایشات بر روی پیکره‌ی مورد نظر دارد.

برای تشخیص این نوع تغییرات معمولا از روش های بصری استفاده می‌شود. برای مثال در مرجع (Forman۲۰۰۶) برای این منظور یک نمایش بصری جدید از فضای ویژگی ارائه شده است. روش مذکور به این صورت است که با داشتن مجموعه‌ی کل کلمات در تمام روزها (مثلا برای مدت یک سال)، کلمات بر اساس تاریخی که اولین بار در ۱۰۰ ویژگی برتر دیده شدند مرتب می‌شوند، سپس تصویری تشکیل داده می‌شود که در آن ستون‌ها نماینده‌ی کلمات و سطرها نماینده‌ی روزها هستند و ترتیب روزها به این نحو است که از بالای تصویر به سمت پایین تصویر، به تاریخ روزها افزوده می‌شود. یک پیکسل به شرطی به رنگ قرمز درمی‌آید که کلمه‌ی مربوط به ستون آن پیکسل در آن روز ،جزء ۱۰۰ کلمه ی برتر باشد. وگرنه بر اساس میزان اهمیت، به رنگی در طیف سیاه و سفید درمی‌آید به طوری که هر چقدر مشکی‌تر باشد بی‌اهمیت‌تر و هر چه سفیدتر باشد با اهمیت‌تر است. شکل ۱ یکی از تصاویر حاصل از روش مذکور در مرجع (Forman۲۰۰۶) را نشان می‌دهد.



شکل ۱:نمایش تغییر توزیع کلمات در مرجع (Forman۲۰۰۶)

در این شکل، مرز قرمز رنگی که از پیکسل‌ها به وجود آمده زمانی را نشان می‌دهد که هر یک از کلمات برای اولین بار در ۱۰۰ کلمه‌ی برتر رویت شده‌اند.این نمایش بصری کمک می‌کند تا نرخ تغییر و نوع تغییر به وضوع مشاهده شود. برای مثال قسمت‌هایی از تصویر که در آن، در یک ستون بعد از چند پیکسل قرمز پیکسل های سیاه رویت شده اند نشان دهنده‌ی کاهش اهمیت کلمات مهم بعد از چند روز است و بر عکس، قسمت‌هایی که چندین پیکسل عمودی پیوسته ازرنگ قرمز دیده می‌شود یعنی کلمه‌ی مربوط به ستون مذکور در طی چند هفته مدام استفاده شه است(بیضی سمت راست در شکل). همینطور برای تشخیص نوع تغییر نیز با مشاهده‌ی رنگ قرمز در یک ستون به صورتی که در بیضی سمت چپ در تصویر نشان داده شده است میی‌توان به یک تغییر بازگشت‌کننده پی برد.

در مرجع (D’hondt, Verberne et al ۲۰۱۴.) برای تشخیص تغییر بازگشت‌کننده و تدریجی از دو آزمایش متفاوت استفاده شده است. برای تشخیص تغییر تدریجی، توزیع دسته‌های مختلف در طول سال‌های متوالی بررسی می‌شود. منطق نویسنده این است که اگر بین دو سال متوالی اندازه‌ی یک دسته تغییر چندانی نداشته باشد، ولی در طول یک دوره‌ی طولانی از زمان، اندازه‌ی آن دسته به طور قابل ملاحظه ای رشد داشته باشد، تغییر داده‌های مربوط به آن دسته تدریجی است. برای تشخیص تغییر بازگشت کننده، توزیع کلمات در سال‌های مختلف و برای هر کلاس به صورت جداگانه بررسی می‌شود.

در مقاله ی (Šilić and Bašić۲۰۱۲) برای شناسایی انواع تغییرات ابتدا انواع روش‌های نمونه‌برداری داده‌های آموزشی و آزمایشی از کل داده‌ها ارائه شده است و سپس با بررسی میزان عملکرد دسته‌بندهای ساخته شده بر روی انواع نمونه برداری‌ها، نوع تغییر موجود در پیکره‌ی مورد بررسی تشخیص

[1] Concept shift
[2] Incremental

داده شده است. بیشتر روش‌های مواجهه با تغییرات محدود به نوع خاصی از تغییر می‌شوند، در حالی که در واقعیت اغلب ترکیبی از چنین تغییراتی رخ می‌دهد.

۴. انواع روش‌های تشخیص تغییر

بسیاری از الگوریتم‌هایی که برای تطبیق با تغییرات زمانی ارائه شده اند، از یک مدل به روز رسانی ساده استفاده می‌کنند(Katakis, Tsoumakas et al. ۲۰۰۶). به این ترتیب که به محض ورود یک داده به سیستم، به روز رسانی مدل انجام می‌شود. در حالی که این روش‌ها هزینه‌ی بالایی دارند. راه دیگر، تشخیص زمان تغییرات و به‌روزرسانی مدل در صورتی است که این تغییرات قابل توجه باشند(Tsymbal۲۰۰۴). در حقیقت هدف از تشخیص تغییر، شناسایی نقاط و یا بازه هایی است که در طول آنها تغییر رخ می‌دهد. در حالت کلی تشخیص تغییر می‌تواند با نظارت و بررسی داده‌ی خام صورت گیرد یا اینکه بر اساس عملکرد حاصل از دسته‌بند انجام شود. بر اساس دسته بندی انجام شده در (Gama, Žliobaitė et al.۲۰۱۴) به طور کلی تشخیص‌دهنده‌های تغییر به سه دسته تقسیم می‌شوند :

۴.۱. تشخیص دهنده های مبتنی بر آنالیز ترتیبی

در این روش ابتدا فرض می‌شود که X_1^n یک توالی از نمونه‌هاست که در آن زیرمجموعه ای از نمونه ها مثل X_1^w $(1<w<n)$ از یک توزیع مانند P0 و زیرمجموعه‌ی دیگر از نمونه ها X_w^n (یعنی سایر نمونه‌ها) از توزیعی مانند P1 تولید شده اند. زمانی که توزیع مثال‌ها در نقطه‌ای مانند W از P0 به P1 تغییر می‌کند، انتظار می‌رود احتمال مشاهده‌ی زیردنباله‌ای با توزیع P1 به شدت بالاتر از زیردنباله‌ای با توزیع P0 شود. میزان شدت باید طوری باشد که نسبت دو احتمال از یک حد آستانه کمتر نباشد. با فرض اینکه نمونه های Xi از هم مستقل باشند، یک شاخص آماری برای آزمودن این فرضیه که آیا W نقطه‌ی تغییر است یا نه بر اساس رابطه‌ی(۱) محاسبه می‌شود:

$$T_w^n = log\frac{P(x_w \dots x_n|P1)}{P(x_w \dots x_n|P0)} = \sum_{i=w}^{n} log\frac{P1[x_i]}{P0[x_i]} = T_w^{n-1} + log\frac{P1[x_n]}{P0[x_n]} \qquad (۱)$$

در این رابطه اگر $T_w^n > L$ باشد آنگاه تغییر تشخیص داده می‌شود.(L یک حد آستانه تعریف شده توسط کاربر است.)

۴.۲. تشخیص دهنده های مبتنی بر جداول کنترلی

جداول کنترل، یادگیری را به عنوان یک فرآیند در نظر می‌گیرند و تکامل این فرآیند را بر اساس میزان رخداد خطا در پیشگویی‌های انجام شده توسط یادگیرنده نظارت می‌کنند (Nishida and Yamauchi۲۰۰۹). براساس این نظارت سیستم در یکی از ۳ حالت تحت کنترل، خارج از کنترل و وضعیت هشدار قرار خواهد گرفت. وضعیت سیستم تا زمانی تحت کنترل است که خطای سیستم ثابت باشد؛ و این یعنی نمونه‌ی آزمایشی ورودی از همان توزیعی آمده است که نمونه‌های قبلی آمده‌اند. وضعیت خارج از کنترل وضعیتی است که خطا به طور قابل ملاحظه‌ای در مقایسه با نمونه‌های اخیر افزایش داشته است. لذا به احتمال ۹۹ درصد نمونه‌ی آخر با توزیع متفاوتی نسبت به نمونه‌های قبلی آمده است. وضعیت هشدار حالتی بین دو وضعیت قبلی است. خطا در این حالت در حال افزایش است ولی هنوز خارج از کنترل نشده است. افزایش خطا در این حالت ممکن است به علت نویز، کاهش عملکرد مدل پیشگویانه به صورت مقطعی و یا تغییر واقعی باشد. در این وضعیت تعداد بیشتری مثال نیاز است تا بتوان به یک نتیجه‌ی قطعی رسید و تشخیص داد افزایش خطا حاصل از نویز است یا تغییر واقعی. با توجه به توضیحات مذکور، با توجه به فاصله‌ی زمانی بین وضعیت هشدار و وضعیت خارج از کنترل می‌توان نرخ تغییر را مشخص کرد. اگر فاصله‌ی زمانی بین وضعیت هشدار و وضعیت خارج از کنترل کم باشد نشان‌دهنده‌ی سرعت بالای تغییر است و اگر این فاصله زیاد باشد نشان از سرعت پایین تغییر دارد.

۴.۳. تشخیص دهنده‌های مبتنی بر نظارت بر توزیع پنجره‌های زمانی

این روش‌ها عموما از یک پنجره‌ی مرجع با اندازه‌ی ثابت استفاده می‌کنند که شامل داده های قدیمی می‌شود و از یک پنجره‌ی لغزنده استفاده می‌کنند که حاوی جدیدترین نمونه ها هستند. مقایسه‌ی بین دو پنجره از مقایسه‌ی بین شاخص‌های آماری توزیع دو پنجره انجام می‌شود. اگر هر دو پنجره دارای شاخص‌های آماری یکسانی نباشند، شروع پنجره‌ی لغزنده به عنوان محل تغییر تعیین می‌شود. شاخص آماری مذکور می‌تواند از روی داده‌های خام و یا عملکرد مدل آموزشی تعریف شود. دو پنجره‌ی مذکور می‌توانند اندازه‌ی یکسانی داشته باشند و یا اینکه اندازه ی پنجره‌ی لغزان می‌تواند به صورت تصاعدی رشد کند. نویسندگان مرجع (Kifer, Ben-David et al.۲۰۰۴) از معیار Chernoff bound برای مقایسه ی آماری توزیع دو پنجره استفاده کرده اند. تحقیق (Vorburger and Bernstein۲۰۰۶) یک معیار مبتنی بر آنتروپی ارائه کرده است تا تفاوت بین توزیع دو پنجره را نشان دهد. در این روش اگر توزیع دو پنجره مساوی باشد، آنگاه آنتروپی برابر ۱ می‌شود و اگر توزیع‌ها مساوی نباشند مقدار آتروپی برابر صفر می‌شود. در این روش شاخص آنتروپی به صورت مداوم و در طول زمان نظارت می‌شود و اگر این شاخص از یک حد آستانه‌ی تعریف شده توسط کاربر پایین‌تر برود نشان‌دهنده‌ی یک تغییر است. در تحقیق (Bach and Maloof۲۰۰۸) از دو مدل با نام‌های مدل پایدار و مدل واکنشی استفاده شده است. مدل پایدار بر اساس داده‌های آموزشی در طول یک بازه‌ی طولانی از زمان و مدل واکنشی بر اساس یکی بازه‌ی زمانی کوتاه از داده‌های آموزشی ساخته شده است. این تکنیک از مدل واکنشی به عنوان

شاخصی برای تشخیص تغییر استفاده می‌کند. به این صورت که وقتی مدل واکنشی برای یک نمونه نسبت به مدل پایدار عملکرد بهتری از خود نشان می‌دهد نشان‌دهنده‌ی رخداد یک تغییر در بازه‌ی زمانی کوتاه مربوط به مدل واکنشی است.

در مقایسه‌ی بین این سه روش می‌توان گفت محدودیت تشخیص‌دهنده‌های دسته‌ی سوم نسبت به دسته‌ی اول، نیاز به حافظه است. تشخیص‌دهنده‌های دسته‌ی اول نیازی به ذخیره‌ی داده‌ی ورودی ندارند ولی تشخیص‌دهنده‌های دسته‌ی دوم باید داده‌های موجود در هر دو پنجره‌ی مرجع و لغزنده را در حافظه نگه دارند. مزیت تشخیص‌دهنده‌های دسته‌ی اول دقت بالاتر آنها نسبت به سایر تشخیص‌دهنده ها است.

۵. روش‌های مواجهه با تغییرات زمانی

در تحقیقات مطالعه شده، چهار شیوه در مواجه با اثرات حاصل از تغییرات زمانی در داده‌ها گزارش شده است(Nishida, Hoshide et al. ۲۰۱۲ , Fukumoto, Ushiyama et al. ۲۰۱۴):۱) انتخاب نمونه ۲) وزن دهی‌به نمونه‌ها ۳)یادگیری دسته جمعی[1](Tsymbal , Gama, Žliobaitė et al. ۲۰۰۴) ۴)انتخاب ویژگی[2](Nishida, Hoshide et al. ۲۰۱۲). در ادامه هر یک از روش‌ها به طور کامل توضیح داده می‌شوند.

۵.۱. انتخاب نمونه

در روش انتخاب نمونه، هدف انتخاب نمونه‌های مرتبط با داده‌هایی است که از سایر داده‌ها به نمونه‌های آزمایشی نزدیک‌ترند. این شیوه که رایج‌ترین روش برای مواجه با تغییرات زمانی است، از یک پنجره لغزان تشکیل شده است که بر روی آخرین نمونه ها حرکت کرده و از مدل ساخته شده‌ی حاصل از آن‌ها برای پیشگویی نمونه‌ها درمثال‌هایی که در آینده‌ای بسیار نزدیک وجود دارند استفاده می‌کند(Widmer and Kubat ۱۹۹۶, Tsymbal ۲۰۰۴). در این الگوریتم‌ها می‌توان از پنجره‌هایی با اندازه‌ی ثابت استفاده کرد یا اینکه با روش‌های اکتشافی اندازه‌ی پنجره را متناسب با آخرین بازه‌ی تغییر در داده‌ها تنظیم کرد. روش پیشنهادی مرجع (D’hondt, Verberne et al) سعی در انتخاب پنجره به نحوی دارد که یک تعادل بین جدید بودن داده و اندازه‌ی پنجره برقرار شود. در مرجع (Šilić and Bašić ۲۰۱۲) نشان داده شده است که در صورت وجود رانش مفهوم در داده‌های آموزشی، حتی استفاده از انتخاب نمونه رندم به جای کل داده‌ها به عنوان داده‌ی آموزشی، می‌تواند میزان عملکرد دسته‌بند را افزایش دهد. در مرجع (Rocha, Mourão et al. ۲۰۱۳) الگوریتمی به نام Chronos پیشنهاد شده است که به ازای هر نمونه‌ی آزمایشی بخشی از مجموعه‌ی آموزشی به نام محتوای زمانی[3] را انتخاب می‌کند. این الگوریتم کمک می‌کند تا بزرگترین محتوای زمانی انتخاب شود به طوری که اثرات ناشی از تغییرات زمانی در عملکرد دسته بند به حداقل برسد.

۵.۲. وزن دهی به نمونه

در این روش به بعضی از نمونه‌ها در طول آموزش وزن بیشتری اختصاص داده می‌شود. نمونه ها می‌توانند بر اساس میزان قدمت و میزان ارتباط به مفاهیم جدید وزن‌دهی شوند(Tsymbal ۲۰۰۴). روش وزن‌دهی به نمونه‌ها نیز، فرض می‌کند که نمونه‌های جدیدتر آموزشی برای دسته‌بندی مثال‌های آزمایشی جدید بهتر عمل می‌کنند(Nishida, Hoshide et al. ۲۰۱۲).

۵.۳. یادگیری دسته جمعی

روش یادگیری دسته جمعی مجموعه‌ای از مدل‌های پیشگویی را نگه‌داری کرده و از رای‌گیری بین مدل‌ها یا وزن‌دهی به آنها در زمان پیشگویی نمونه‌های آزمایشی استفاده می‌کند.معمولا وقتی مدل‌های جدید اضافه می‌شوند، مدل‌های ضعیف یا قدیمی، حذف می‌شوند. حذف شدن مدل‌های قدیمی یا به صورت دوره‌ای است و یا پس از مشاهده‌ی کاهش صحت عملکرد دسته بند انجام می‌شود. هر چند در بعضی از مراجع (Nishida and Yamauchi ۲۰۰۹)، هیچ دسته‌بندی تحت هیچ شرایطی از سیستم حذف نمی‌شود. سیستم پیشنهادی این مرجع (Nishida and Yamauchi ۲۰۰۹) از چندین دسته بند آفلاین و آنلاین برای مواجهه با تغییرات زمانی استفاده کرده است. با شروع کار سیستم با ورود هر مثال جدید دسته‌بند آنلاین به‌روزرسانی می‌شود، برخلاف دسته‌بند آفلاین که با ورود مثال‌های جدید به‌روزرسانی نمی‌شوند. در حقیقت دسته‌بندهای آنلاین برای تطبیق با تغییرات تدریجی، مدام در حال یادگیری مثال‌ها هستند و دسته‌بندهای آفلاین برای پاسخ به تغییرات بازگشت کننده استفاده می‌شوند. در(Kolter and Maloof ۲۰۰۳) مجموعه‌ای از دسته‌بندها با قدمت‌های متفات ساخته می‌شود به طوری که هر یک از دسته‌بندها شامل نمونه‌هایی مربوط به دوره‌ی خاصی از زمان هستند.

۵.۴. انتخاب ویژگی

در اکثر تحقیقات تنها سه دسته‌ی قبلی برای مواجه با تغییرات زمانی بیان شده است ولی در تحقیق (Nishida, Hoshide et al. ۲۰۱۲, Fukumoto, Ushiyama et al. ۲۰۱۴) دسته روش انتخاب ویژگی[4] نیز برای مواجه با تغییرات زمانی ارائه شده است. این رویکرد در چند سال اخیر مورد توجه قرار گرفته و تمرکز را از وزن‌دهی به نمونه‌ها به سمت وزن‌دهی به ویژگی‌ها سوق می‌دهد. در این رویکرد برای مواجهه با تغییرات زمانی هر بار به صورت پویا ارزشمندترین ویژگی‌ها(کلمات) انتخاب می‌شوند. در تحقیق (Fukumoto, Ushiyama et al. ۲۰۱۴) ویژگی‌ها یا کلمات مورد استفاده در دسته‌بندی

[1] Ensemble learning
[2] Feature selection
[3] Temporal context
[4] Feature selection

متن به دو دسته تقسیم شده‌اند، یکی کلمات مستقل از زمان[1] و دیگری کلمات وابسته به زمان[2] است که در مرجع (Nishida, Hoshide et al. ۲۰۱۲) به ترتیب کلمات پایدار[3] و ناپایدار[4] خوانده می‌شوند. در (Fukumoto, Ushiyama et al. ۲۰۱۴) از روش TWF (Salles, Rocha et al. ۲۰۱۰) برای وزن دهی استفاده شده است با این تفاوت که به جای وزن‌دهی در در سطح نمونه، وزن‌دهی در سطح ویژگی انجام می‌شود. به این صورت که اگر یک ویژگی وابسته به زمان بود، از وزن‌دهی بر پایه‌ی TWF و اگر مستقل از زمان بود وزن دهی معمولی برای آن استفاده می‌گردد.

۶. ارتباط بین روش های مواجهه با تغییر با نوع تغییر

به طور کلی انتخاب روش مواجهه با تغییر، به نوع تغییر موجود در پیکره‌ی مورد بررسی وابسته است(D'hondt, Verberne et al). نویسندگان تحقیق (Nishida, Hoshide et al. ۲۰۱۲) معتقدند که روش وزن‌دهی به نمونه‌ها تنها در مواجهه با تغییر تدریجی خوب عمل می‌کند ولی در مواجهه با تغییر ناگهانی شکست می‌خورد. در تحقیق (Carmona-Cejudo, Baena-García et al. ۲۰۱۱) نتایج مناسبی از انتخاب نمونه برای مواجه با تغییر تدریجی موجود در دسته‌بندی جریانی از ایمیل‌ها به دست آمده است. به عقیده‌ی (D'hondt, Verberne et al) یادگیری دسته جمعی برای تغییرات تدریجی و بازگشت کننده خوب عمل می‌کند و بهترین راه مواجه با تغییر ناگهانی، انتخاب داده‌ی آموزشی تنها از نمونه‌هایی است که با توزیع جدید همخوانی دارند و زمانی که تغییر بازگشت کننده وجود ندارد بهترین راه مواجهه با تغییر این است که مدل هر چند وقت یکبار به روز رسانی شود .به این نحو که مدل‌های قدیمی آموزش دیده تخریب و مدل‌های جدید به جای مدل‌های دسته بندی قدیمی جایگزین شود.

در بعضی تحقیقات برای مواجهه با تغییر تنها به یک روش بسنده نمی‌کنند و از ترکیبی از چند روش استفاده می‌کنند. برای مثال در تحقیق (Salles, Rocha et al. ۲۰۱۰) نوعی وزن‌دهی مبتنی بر زمان به نام TWF برای دسته بندی ارائه شده است. وزن دهی TWF، تغییرات موجود در رابطه ی بین کلاس و واژگان را در دوره‌های زمانی مختلف استخراج کرده و در دسته‌بند آن را به دو روش استفاده می‌کند:TWF بر روی نمونه‌ها و TWF بر روی امتیاز به دسته‌بندها. درTWF بر روی متون، فاصله‌ی زمانی بین متون آموزشی و آزمایشی به منظور اندازه‌گیری میزان شباهت محاسبه شده و بر این اساس یک متن آموزشی که از لحاظ زمانی نزدیک‌تر به متن آزمایشی است، اهمیت بیشتری پیدا می‌کند. در TWF بر روی دسته‌بندها،امتیاز نهایی به یک کلاس از جمع وزن دار امتیازهای تولید شده توسط دسته‌بندهای عادی که در بازه‌های زمانی متفاوت ساخته شده اند به دست می‌آید.

۷. نتیجه گیری

در دسته بندی متن عموما داده ها ابتدا به صورت آفلاین جمع آوری شده و آموزش بر روی آن‌ها انجام می‌شود سپس مدل آموزش دیده به این روش، می‌تواند برای پیش بینی داده‌های دیده نشده‌ی ورودی مورد استفاده قرار بگیرد. در شرایطی که داده‌ها مدام در حال به روز رسانی هستند، توزیع آنها می‌تواند در طول زمان تغییر کند به طوری که موجب اثرگذاری بر روی عملکرد دسته بند شود. از آنجایی که انتخاب روش مواجهه با تغییر، به نوع تغییر وابسته است، لذا در این مقاله ابتدا به بررسی انواع تغییرات زمانی پرداخته شد. سپس انواع روش‌های مواجهه با تغییر بیان شد و در نهایت بهترین روش‌های مواجهه با تغییر متناسب با نوع تغییر رخ داده مورد بررسی قرار گرفت. در این مقاله سعی شد تا حد امکان یک دید کلی و همه جانبه در زمینه‌ی اثر زمان در دسته‌بندی خودکار متن ارائه شده و مسیر را برای سایر محققین در انتخاب روش مناسب برای مواجهه با تغییرات زمانی هموارتر کند.

۸. منابع مورد استفاده

Alonso, O., M. Gertz and R. Baeza-Yates (۲۰۰۷). On the value of temporal information in information retrieval. ACM SIGIR Forum, ACM.

Bach, S. H. and M. Maloof (۲۰۰۸). Paired learners for concept drift. Data Mining, ۲۰۰۸. ICDM'۰۸. Eighth IEEE International Conference on, IEEE.

Brzezinski, D. and J. Stefanowski (۲۰۱۴). "Reacting to different types of concept drift: The accuracy updated ensemble algorithm." Neural Networks and Learning Systems, IEEE Transactions on ۲۵(۱): ۸۱-۹۴.

Carmona-Cejudo, J. M., M. Baena-García, R. M. Bueno, J. Gama and A. Bifet (۲۰۱۱). Using GNUsmail to Compare Data Stream Mining Methods for On-line Email Classification. WAPA.

[1] Temporal independent terms
[2] Temporal dependent terms
[3] Stationary
[4] Bursty

D'hondt, E., S. Verberne, N. Oostdijk, J. Beney, C. Koster and L. Boves (۲۰۱۴). "Dealing with temporal variation in patent categorization." Information Retrieval: ۱-۲۵.
Ditzler, G. (۲۰۱۲). Incremental Learning of Concept Drift from Imbalanced Data.
Forman, G. (۲۰۰۶). Tackling concept drift by temporal inductive transfer. Proceedings of the ۲۹th annual international ACM SIGIR conference on Research and development in information retrieval, ACM.
Fukumoto, F., S. Ushiyama, Y. Suzuki and S. Matsuyoshi (۲۰۱۴). The Effect of Temporal-based Term Selection for Text Classification. Australasian Language Technology Association Workshop ۲۰۱۴.
Gama, J., I. Žliobaitė, A. Bifet, M. Pechenizkiy and A. Bouchachia (۲۰۱۴). "A survey on concept drift adaptation." ACM Computing Surveys (CSUR) ۴۶(۴): ۴۴.
Katakis, I., G. Tsoumakas and I. Vlahavas (۲۰۰۶). "Dynamic feature space and incremental feature selection for the classification of textual data streams." Knowledge Discovery from Data Streams: ۱۰۷-۱۱۶.
Kifer, D., S. Ben-David and J. Gehrke (۲۰۰۴). Detecting change in data streams. Proceedings of the Thirtieth international conference on Very large data bases-Volume ۳۰, VLDB Endowment.
Kolter, J. Z. and M. Maloof (۲۰۰۳). Dynamic weighted majority: A new ensemble method for tracking concept drift. Data Mining, ۲۰۰۳. ICDM ۲۰۰۳. Third IEEE International Conference on, IEEE.
Mourão, F ,.L. Rocha, R. Araújo, T. Couto, M. Gonçalves and W. Meira Jr (۲۰۰۸). Understanding temporal aspects in document classification. Proceedings of the ۲۰۰۸International Conference on Web Search and Data Mining, ACM.
Nishida, K., T. Hoshide and K. Fujimura .(۲۰۱۲) Improving tweet stream classification by detecting changes in word probability. Proceedings of the ۳۵th international ACM SIGIR conference on Research and development in information retrieval, ACM.
Nishida, K. and K. Yamauchi (۲۰۰۹). Learning, detecting, understanding, and predicting concept changes. Neural Networks, ۲۰۰۹. IJCNN ۲۰۰۹. International Joint Conference on, IEEE.
Rocha, L., F. Mourão, H. Mota, T. Salles, M. A. Gonçalves and W. Meira Jr (۲۰۱۳). "Temporal contexts: Effective text classification in evolving document collections." Information Systems ۳۸(۳): ۳۸۸-۴۰۹.
Rocha, L., F. Mourão, A. Pereira, M. A. Gonçalves and W. Meira Jr (۲۰۰۸). Exploiting temporal contexts in text classification. Proceedings of the ۱۷th ACM conference on Information and knowledge management, ACM.
Salles, T., L. Rocha, G. L. Pappa, F. Mourão, W. Meira Jr and M. Gonçalves (۲۰۱۰). Temporally-aware algorithms for document classification. Proceedings of the ۳۳rd international ACM SIGIR conference on Research and development in information retrieval, ACM.
Sebastiani, F. (۲۰۰۲). "Machine learning in automated text categorization." ACM computing surveys (CSUR) ۳۴(۱): ۱-۴۷.
Šilić, A. and B. D. Bašić (۲۰۱۲). Exploring classification concept drift on a large news text corpus. Computational linguistics and intelligent text processing, Springer: ۴۲۸-۴۳۷.
Tsymbal, A. (۲۰۰۴). "The problem of concept drift: definitions and related work".
Vorburger, P. and A. Bernstein (۲۰۰۶). Entropy-based concept shift detection. Data Mining, ۲۰۰۶ICDM'۰۶. Sixth International Conference on, IEEE.
Wenerstrom, B. and C. Giraud-Carrier (۲۰۰۶). Temporal data mining in dynamic feature spaces. Data Mining, ۲۰۰۶. ICDM'۰۶. Sixth International Conference on, IEEE.
Widmer, G. and M. Kubat (۱۹۹۶). "Learning in the presence of concept drift and hidden contexts." Machine learning ۲۳(۱): ۶۹-۱۰۱.