بسمه تعالی

دانشگاه اصفهان

دانشکده مهندسی کامپیوتر

گروه آموزشی هوش مصنوعی و رباتیکز

نام و نام خانوادگی : فاطمه وهابی

شماره دانشجویی : 4013614052

موضوع گزارش : تمرین صفرم درس پردازش زبان طبیعی

استاد : دکتر حمیدرضا برادران کاشانی

دستیار استاد : امیرمسعود سلطانی

زمستان 1401

پرسش‌ها

فرآیند ریشه‌یابی با لم‌سازی چه تفاوت‌هایی دارد؟ مثال بزنید.

در فرایند لم‌سازی ما یک عبارت یا کلمه را به فرم عادی در می‌آوریم و این به این معنا نیست که چیزی از آن حذف می‌کنیم.

مانند : extra-linguistic, extralinguistic, extra linguistic

همه آن‌ها یکسان هستند اما به گونه‌ای متفاوت نوشته شده‌اند. در لم‌سازی همه این گونه‌های متفاوت را به یک شکل در می‌آوریم.

اما در فرایند ریشه‌یابی، کلمات را به محتوای کلی آن تغییر می‌دهیم و این به این معناست که در صورت لزوم چیزی از آن حذف می‌شود.

مانند : automate(s), automatic, automation ---> automat

هر کدام با توجه به کاربردی که در جمله دارد ترجمه می‌شود. اما محتوای کلی کلمه یکی است. همه این کلمات به automat اشاره می‌کنند.

برنامه نویسی

تحلیل کدها در فایل کد نوشته شده است.

بهترین نتیجه مربوط به تصحیح خطا است. و بدترین مربوط به بخش ریشه‌یابی می‌باشد.

اما باید به این نکته توجه کرد که عملیات تصحیح خطا زمان زیادی طول می‌کشد که بهینه نیست. همچنین اعمال پیش پردازش بسیاری را لازم بود در نظر بگیریم. برخی از انها در جهت بهبود نتیجه در نظر گرفته شدند و برخی نیز اگر اعمال نمی‌شد، برنامه در مرحله تصحیح خطا ارور می‌داد.

عملیات ریشه یابی و لم سازی با دو عملیات پیش پردازش به نتیجه مطلوبی رسیدند.

نتیجه عملیات تصحیح خطا 0.89 ، نتیجه عملیات لم‌سازی 0.88 ، نتیجه عملیات ریشه یابی 0.85 می‌باشد.

با توجه به اختلاف کم نتیجه عملیات تصحیح خطا و عملیات لم‌سازی، و و همچنین زمان اجرای زیاد عملیات تصحیح خطا، میتوان گفت که عملیات لم‌سازی در الویت است. پس از ان به ترتیب عملیات تصحیح خطا و ریشه‌یابی را معرفی می‌کنیم.

اگر زمان و تعداد عملیات پیش پردازش اهمیتی نداشته باشد، می‌توان از عملیات تصحیح خطا استفاده کرد. این عملیات از این جهت حائز اهمیت است که تراکم هر کلمه را در بردار کلمات، دقیق‌تر می‌کند و از به حساب نیاوردن برخی کلمات به علت غلط املایی، جلوگیری می‌کند.

خوانش مقاله

1. هدف مقاله را در یک یا دو جمله توضیح دهید.

این مقاله تأثیر نرخ عدم تعادل[1] و دسته‌بند را بر عملکرد چندین استراتژی نمونه‌گیری مجدد[2] برای مقابله با مجموعه داده‌های نامتعادل بررسی می‌کند. این مطالعه متمرکز بر ارزیابی چگونگی تأثیر یادگیری، زمانی که الگوریتم‌های نمونه‌گیری مجدد مختلف، داده‌های نامتعادل اولیه را به توزیع‌های کلاسی متعادل مصنوعی تبدیل می‌کنند، است. به عبارتی، این مقاله یک تحلیل تجربی کامل از تأثیر نسبت عدم تعادل (نسبت بین کلاس های اقلیت و اکثریت) و دسته‌بند بر اثربخشی استراتژی‌های نمونه‌گیری مجدد در هنگام برخورد با داده‌های نامتعادل ارائه کرده است.

2. تفاوت این مقاله با مقالات پیش از خود در چیست؟

این مقاله به طور قابل توجهی کارهای قبلی را با افزایش دامنه[3] و جزئیاتی که در آن تأثیر نسبت عدم تعادل و دسته‌بند بر اثربخشی محبوب‌ترین استراتژی‌های نمونه‌گیری مجدد (تحت بیش نمونه برداری و کم نمونه برداری) مورد مطالعه قرار می‌گیرد، گسترش داده است.

3. روش های بیش نمونه برداری و کم نمونه برداری که در مقاله آمده است را توضیح دهید. از هر کدام یک مثال بیاورید و به صورت خلاصه بگویید چگونه عمل می‌کنند.

بیش نمونه برداری[4]:

ساده‌ترین روش برای افزایش اندازه کلاس اقلیت مربوط به نمونه‌گیری بیش از حد تصادفی است، یعنی یک روش غیر اکتشافی که توزیع کلاس را از طریق تکرار تصادفی مثال‌های مثبت متعادل می‌کند. این به تعادل توزیع کلاس بدون افزودن اطلاعات جدید به مجموعه داده کمک می کند. با این وجود، از آنجایی که این روش نمونه‌های مثبت موجود را تکرار می‌کند، احتمال بیش برازش[5] بیشتر است.

---

[1] imbalance ratio
[2] Resampling
[3] scope
[4] Over-sampling
[5] overfit

به عنوان مثال، SMOTEBoost رویکردی است که Chawla و همکاران معرفی کرده‌اند، که SMOTE را با روش استاندارد تقویت ترکیب می‌کند.

این روش که SMOTE (تکنیک بیش نمونه برداری اقلیت مصنوعی) نامیده می‌شود، به دسته‌بند اجازه می‌دهد تا مناطق تصمیم‌گیری بزرگ‌تری بسازد که شامل نمونه‌های نزدیک از کلاس اقلیت است.

گارسیا و همکاران سه نوع را بر اساس مفهوم همسایگی با هدف در نظر گرفتن مجاورت و توزیع فضایی نمونه‌ها ایجاد کرد.

هان و همکاران الگوریتم Borderline-SMOTE را ارائه کردند که فقط نمونه‌های اقلیت جدیدی را بر اساس نمونه‌های موجود که نزدیک مرز تصمیم هستند ایجاد می‌کند.

هونگیو و هرنا روش DataBoost-IM را معرفی کردند که ترکیبی از تقویت و تولید داده است.

کم نمونه برداری[6]:

هدف کم‌نمونه‌گیری تصادفی، متعادل کردن مجموعه داده‌ها از طریق حذف تصادفی نمونه‌های منفی است. علیرغم اینکه اطلاعات مهم ممکن است زمانی که نمونه ها به طور تصادفی کنار گذاشته می‌شوند از بین بروند، به طور تجربی نشان داده شده است که یکی از موثرترین روش‌های نمونه‌گیری مجدد است.

برخلاف رویکرد تصادفی، پیشنهادات دیگری مبتنی بر انتخاب هوشمندانه‌تر نمونه‌های منفی وجود دارد که باید حذف شوند.

به عنوان مثال، کوبات و متوین تکنیک انتخاب یک طرفه (OSS) را پیشنهاد کردند، که به طور انتخابی تنها موارد منفی را حذف می‌کند که یا اضافی هستند یا با نمونه‌های کلاس اقلیت هم‌مرز هستند (نویسندگان فرض می‌کنند که این موارد مرزی نویز هستند). نمونه‌های مرزی با استفاده از مفهوم پیوندهای Tomek شناسایی شدند، در حالی که نمونه‌های اضافی با استفاده از چگالش هارت حذف شدند.

برخلاف روش انتخاب یک طرفه، قانون به اصطلاح تمیز کردن همسایگی بر پاکسازی داده‌ها بیشتر از کاهش داده‌ها تأکید دارد. برای این منظور، ویرایش Wilson برای شناسایی و حذف موارد منفی پر سر و صدا استفاده می شود.

[6] Under-sampling

به همین ترتیب، باراندلا و همکاران روشی را معرفی کردند که نه تنها نمونه‌های پر سر و صدا از کلاس اکثریت را با استفاده از ویرایش ویلسون (WE) حذف می‌کند، بلکه نمونه‌های اضافی را نیز از طریق الگوریتم متراکم زیرمجموعه انتخابی اصلاح شده (MSS) حذف می‌کند.

ین و همکاران یک الگوریتم زیر نمونه‌برداری مبتنی بر خوشه ارائه کرد. ابتدا تمام نمونه‌های اصلی را در چند خوشه جمع می‌کند و سپس با در نظر گرفتن نسبت تعداد نمونه‌های کلاس اکثریت به تعداد نمونه‌های کلاس اقلیت در خوشه، تعداد مناسبی از نمونه‌های کلاس اکثریت را از هر خوشه انتخاب می‌کند.

گارسیا و هررا استفاده از الگوریتم‌های محاسباتی تکاملی را برای نمونه‌گیری کمتر از کلاس اکثریت پیشنهاد کردند.

چن و همکاران روشی مبتنی بر هرس بردارهای پشتیبانی کلاس اکثریت معرفی کرد.

4. چرا معیار صحت در مورد مجموعه دادگان با عدم توازن ، معیار خوبی نیست؟

شواهد تجربی و نظری نشان می‌دهد که این معیارها با توجه به عدم تعادل داده‌ها و نسبت‌های دسته‌بندی درست و نادرست، به شدت مغرضانه هستند. در یک مسئله تصمیم گیری دودویی، یادگیرنده موارد را مثبت یا منفی پیش‌بینی می‌کند. اگر نمونه‌های بسیار کمی متعلق به کلاس مثبت باشند، یک سیستم یادگیری معمولی می‌تواند با دسته‌بندی همه نمونه‌ها به عنوان منفی دقت بسیار بالایی به دست آورد. با این حال، این در اکثر حوزه‌های واقعی بی‌فایده است، زیرا طبقه مورد علاقه عموماً طبقه مثبت است. بنابراین، ارزیابی‌کنندگانی مانند دقت یا میزان خطا برای داده‌های نامتعادل کلاس نامناسب به نظر می‌رسند.

5. چه معیارهایی برای بررسی نتایج الگوریتم ها روی مجموعه دادگان با عدم توازن معرفی شده است،آن‌ها را به طور خلاصه توضیح دهید؟ چرا بهتر هستند؟

- برای مقابله با عدم توازن در کلاس‌ها، حساسیت[۷] (یا یادآوری[۸]) و ویژگی خاص[۹] معمولاً برای نظارت بر عملکرد دسته‌بندی در هر کلاس به طور جداگانه، اتخاذ شده است. توجه داشته باشید که حساسیت، درصد نمونه‌های مثبتی است که به درستی دسته‌بندی شده‌اند، در حالی که ویژگی خاص، به عنوان نسبت نمونه‌های منفی که به درستی دسته‌بندی شده‌اند، تعریف می‌شود.
- در چندین مشکل ما به طور خاص می‌خواهیم به عملکرد بالا در تنها یک کلاس دست یابیم. به عنوان مثال، در تشخیص یک بیماری نادر، یکی از مهمترین موارد این است که بدانیم یک تشخیص مثبت

[7] sensitivity
[8] Recall
[9] specificity

چقدر قابل اعتماد است. برای چنین مشکلاتی، معیار دقت[10] (یا خلوص[11]) اتخاذ می‌شود، که می‌تواند به عنوان درصد نمونه‌هایی که به درستی به عنوان مثبت برچسب‌گذاری شده‌اند، تعریف شود.

- یکی از رایج‌ترین تکنیک‌ها برای ارزیابی دسته‌بندها در مسائل نامتعادل، منحنی ویژگی‌های عملیاتی گیرنده (ROC) است که ابزاری برای تجسم، سازمان‌دهی و انتخاب دسته‌بندها بر اساس مبادلات آن‌ها بین منافع (مزایای واقعی) و هزینه‌ها (مثبت‌های غلط) است. یک نمایش کمی از یک منحنی ROC، ناحیه زیر آن است که به عنوان AUC شناخته می‌شود. هنگامی که تنها یک اجرا از یک دسته‌بند در دسترس است، AUC را می‌توان به عنوان میانگین حسابی (میانگین کلان) TPrate و TNrate محاسبه کرد.
- F-Measure برای ترکیب دقت و TPrate در یک معیار ارزیابی استفاده می‌شود، که نشان‌دهنده میانگین هارمونیک وزنی بین این دو معیار ارزیابی است.
- میانگین هندسی[12]

$$Gmean = \sqrt{TPrate \cdot TNrate}$$

این معیار ارزیابی به نقطه ای از منحنی ROC مرتبط است، و ایده این است که دقت هر دو کلاس را به حداکثر برسانیم و در عین حال آن‌ها را متعادل نگه داریم. می‌توان آن را نوعی مبادله خوب بین هر دو نرخ دانست، زیرا ارزش بالا زمانی رخ می‌دهد که هر دو نیز بالا باشند، در حالی که یک مقدار پایین حداقل به یک نرخ پایین مربوط می‌شود.

- دقت بهینه شده[13]

$$OP = Acc - \frac{|TNrate - TPrate|}{TNrate + TPrate}$$

این نشان دهنده تفاوت بین دقت کلی و عبارت دوم است که میزان تعادل هر دو کلاس را محاسبه می‌کند. مقادیر بالای OP به دقت کلی بالا و دقت کلاس به خوبی متعادل نیاز دارند. با این حال، OP می‌تواند به شدت تحت تأثیر مغرضانه دقت کلی قرار گیرد.

- شاخص تعمیم یافته دقت متعادل[14]

---

[10] precision
[11] purity
[12] Gmean
[13] optimized precision
[14] generalized Index of Balanced Accuracy

که می‌تواند برای هر معیار عملکرد M به صورت زیر تعریف شود:

$$IBA_{\alpha}(\mathcal{M}) = (1 + \alpha \cdot Dom) \cdot \mathcal{M}$$

معیار ارزیابی IBA یک تبادل معین بین یک اندازه گیری از دقت کلی را مشخص می‌کند. در اینجا، از Gmean شاخصی برای میزان متعادل بودن دقت دو کلاس (شاخص غالب) استفاده خواهیم کرد. برخلاف اکثر معیارهای عملکرد، تابع IBA نه تنها از دقت کلی مراقبت می‌کند، بلکه قصد دارد دسته‌بندهایی با نتایج بهتر در کلاس مثبت (به طور کلی، مهم‌ترین کلاس) را در نظر بگیرد.

6. آیا مجموعه داده فراهم شده در بخش ۲ دارای عدم توازن است؟

بله. دو الگوریتم کم نمونه‌برداری و دو تکنیک نمونه‌برداری بیش‌ازحد، SMOTE و SMOTE مبتنی بر گراف گابریل (gg-SMOTE) ، استفاده شدند.

7. یکی از معیارهای معرفی شده را به برنامه بخش ۲ اضافه کنید.

روش میانگین هندسی را به عنوان یک معیار ارزیابی متناسب با شرایط مسائلی که کلاس‌ها توازن ندارند، معرفی می‌کنم.

$$Gmean = \sqrt{TPrate \cdot TNrate}$$

این معیار به نقطه‌ای از منحنی ROC مرتبط است، و ایده این است که دقت هر دو کلاس را به حداکثر برسانیم و در عین حال آنها را متعادل نگه داریم. می‌توان آن را نوعی مبادله خوب بین هر دو نرخ مشاهده کرد، زیرا ارزش بالا زمانی رخ می‌دهد که هر دو نیز بالا باشند، در حالی که مقدار پایین حداقل به یک نرخ پایین مربوط می‌شود.

این روش از این نظر خوب است که اگر یکی از این نرخ‌ها بالا باشد و دیگری نباشد، از نرخ به دست آمده متوجه می‌شویم که توزیع مصنوعی کلاس‌ها چه میزان مفید بوده است.

البته اگر تنها نرخ کلاسی را که در حالت طبیعی توزیع نامناسبی داشته است، بررسی کنیم، به این نتیجه می‌رسیم که نمونه گیری مجدد چگونه عمل کرده است.

همچین روش شاخص تعمیم یافته دقت متعادل نیز روش مناسبی است.

$$IBA_{\alpha}(\mathcal{M}) = (1 + \alpha \cdot Dom) \cdot \mathcal{M}$$

معیار ارزیابی IBA یک تبادل معین بین یک اندازه گیری از دقت کلی را مشخص می‌کند. در اینجا، از Gmean شاخصی برای میزان متعادل بودن دقت دو کلاس (شاخص غالب) استفاده خواهیم کرد. برخلاف اکثر معیارهای عملکرد، تابع IBA نه تنها از دقت کلی مراقبت می‌کند، بلکه قصد دارد دسته‌بندهایی با نتایج بهتر در کلاس مثبت (به طور کلی، مهم‌ترین کلاس) را در نظر بگیرد.

بدین ترتیب می‌توان به ترتیب الویت روش شاخص تعمیم یافته دقت متعادل و روش میانگین هندسی را برای افزودن به برنامه بخش 2 معرفی کرد.

8. نتیجه مقاله چیست؟

نتایج تجربی نشان داده‌اند که به‌طور کلی، بیش نمونه‌گیری برای مجموعه‌های داده‌ای با عدم تعادل کلاسی شدید، بهتر از نمونه‌گیری کم عمل می‌کند. این نتیجه را می‌توان با این واقعیت توضیح داد که نمونه‌گیری کم ممکن است نمونه‌های منفی زیادی را به منظور متعادل کردن اندازه هر دو کلاس از بین ببرد و در نتیجه باعث از دست رفتن بسیار قابل توجهی از اطلاعات بالقوه مهم برای یادگیرنده شود. هنگامی که عدم تعادل کم است، نتایج نشان می‌دهد که هم بیش نمونه برداری و هم کم نمونه برداری، عملکرد مشابهی را ارائه می‌دهند، بنابراین تجزیه و تحلیل پیچیدگی داده‌های بیشتری برای انتخاب یک تکنیک نمونه‌گیری مجدد مناسب برای یک مجموعه داده نامتعادل خاص ضروری است. علاوه بر این، استفاده از مجموعه‌های آموزشی اصلی بدون هیچ گونه پیش پردازشی به وضوح بدتر از نمونه برداری مجدد است.

از سوی دیگر، آزمایش‌ها نشان داده‌اند که ویژگی‌های دسته‌بند تأثیر کمی بر اثربخشی استراتژی‌های مختلف نمونه‌گیری مجدد دارد. از این نظر، به نظر می‌رسد روش پیش‌پردازش داده‌ها مهم‌تر از یادگیرنده برای دسته‌بندی است.