مروری بر الگوریتم‌های یادگیری دسته جمعی جهت تشخیص هرزنامه

*علی حلافی[۱]، علی هارون آبادی[۲]، علی چوبین[۳]
[۱]دانشجوی کارشناسی ارشد، دانشگاه آزاد اسلامی واحد اهواز، Hallafi_a@yahoo.com
[۲]استادیار، گروه نرم افزار، دانشگاه آزاد اسلامی واحد تهران مرکزی ، تهران، a.haronabadi@gmail.com
[۳] دانشجوی کارشناسی ارشد، دانشگاه آزاد اسلامی واحد اهواز، alichoobin@gmail.com

چکیده

دردنیای امروز اینترنت یکی از مهمترین روش‌های ارتباطی است، ونقش مهمی در زندگی امروز ایفا می‌کند. پست الکترونیک نیز یکی از مهمترین وگسترده ترین خدماتی است که اینترنت در اختیار کاربرانش قرار می‌دهد. این خدمت بنا به دلایلی از جمله سادگی . هزینه‌ی کم بسیار مورد توجه قرار گرفته است. همچنین عده‌ای عوامل سودجو از محیط مجازی و این خدمت برای تبلیغ کالا و یا مسائلی دیگر به فکر سوء استفاده هستند،بطوری که در راستای این عمل یکسری نامه‌های نامعتبر به روزانه به طور ناخواسته وارد پست الکترونیکی ما می‌شوند که از آنها به نام هرزنامه یا نامه‌های تجاری ناخواسته یاد می‌شود که در حال افزایش هستند.راه‌های متعدد وفراوانی جهت مقابله با این پدیده ارائه شده است که هر کدام یکسری مزایا ومعایب دارند، شناخت صحیح عملکرد هر کدام از این الگوریتم‌ها و میزان دقت در تشخیص و فیلترکردن نامه‌های الکترونیکی نامعتبرکاری ضروری می باشد.ما در این جستار به الگوریتم‌های یادگیری جمعی(به جای الگوریتم‌های منفرد) و معرفی آنها پرداخته‌ایم.

کلمات کلیدی: یادگیری، الگوریتم های دسته جمعی، تکنیک‌های فیلترینگ و هرزنامه

مقدمه

هرزنامه[۱] یا نامه‌های ناخواسته[۲] به نامه‌های الکترونیکی نامعتبری اطلاق می‌شود که به شکل چشمگیری در حال افزایش بوده و به جرأت می‌توان گفت که بیش از نیمی از فضای پست‌های الکترونیکی ما را به خود جای داده است که مسائل و مشکلات زیادی را برای کاربران محیط مجازی به وجود آورده که عبارتند از: اتلاف فضای دخیره سازی، هدر رفت پهنای باند، ضایع شدن حداقل زمان ممکن کاربران جهت پاکسازی پست الکرونیکی که معولاً به صورت مرتب این عمل انجام می‌شود، از جنبه‌های دیگر نیاز مشکلات ناخواسته‌ای را منجر شده که به صورت اشاره‌ای عبارتند از کلاهبرداری اینترنتی، لو رفتن اطلاعات حساب بانکی و از بـُعد دیگر داشتن ضمیمه‌های همچون ویروس،تروجان و موارد این چنین را شامل می‌شود که می‌بایست نکات وهشدارهای امنیتی در این حوزه را شناخت و سعی در بهره‌گیری از امکانات موجود(نرم‌افزاری وسخت‌افزاری) جهت داشتن پست الکترونیکی عاری از هرزنامه یا با حداقل آن را داشت. دراین جستار ما می‌خواهیم مروری بر روش‌های یادگیری دسته‌جمعی که یکی از کاربردهای آن جهت تشخیص و فیلترکردن هرزنامه به صورت خودکار را مورد بررسی و بازبینی قرار دهیم.

تاریخچه

نامه‌های الکترونیکی ناخواسته تا کنون ۳ دوره تاریخی را پشت سرگذاشته‌اند:

دوره نخست: این دوره که از سال ۱۹۸۷ تا اواسط ۱۹۹۰ به طول انجامید سال‌های اولیه حضور نامه‌های نامعتبر در دنیای اینترنت بوده است. دراین دوره هرزنامه‌ها به صورت دستی فرستاده می‌شدند از این رو ارسال توده‌ای عظیمی از نامه‌های نامعتبر کاری دشوار ونیاز به نیروی انسانی نسبتاً زیاد بود.

دوره دوم: این دوره مربوط به شروع دوران حضور ماشین‌ها و نرم‌افزارها در فرستادن و پالایش هرزنامه‌ها است. این دوره با ایجاد یک ماشین هرزنامه‌نویس در سال ۱۹۹۴ آغاز شد، به دنبال آن درسال ۱۹۹۷ اولین نرم‌افزار پالایش هرزنامه‌ها طراحی شد[۱].

دوره سوم: این دوره مربوط به جنگ نرم‌افزارها برای فرستادن و پالایش هرزنامه‌ها می‌باشد. این دوره تقریباً از سال ۲۰۰۲ و با انتشار مقاله‌ی "تدبیری برای هرزنامه" توسط پل گراهام آغاز شد، دراین مقاله استفاده از یادگیری ماشین وتفکیک‌کننده‌های آماری برای پالایش هرزنامه‌ها پشنهادشد. این دوره تاکنون نیز ادامه دارد که در طول آن نرم‌افزارهای ارسال هرزنامه‌ها قدرتمند شدند، وهرزنامه‌ها را با سرعت وگوناگونی فراوان تولید می‌کنند، ودر مقابل نرم-

[1] Spam
[2] Unwanted Email

افزارهای پالایش هرزنامه‌ها به رویکردهای یادگیری ماشین واستفاده از دسته‌بندهای آماری وترکیبی روی آورده است. این دوره اهمیت بیشتری نسبت به بقیه دوره دارد[۲و۳].

انواع هرزنامه

به جهت وسعت وگستردگی که هرزنامه‌ها دارند، طبقه بندی خاصی ندارند و می‌توان را برحسب معیارهای متفاوتی دسته بندی کرد. برای مثال فرستادن هرزنامه دلایل متعددی می‌تواند داشه باشد از جمله دلایل عقیدتی، کلاهبرداری اطلاعاتی وتبلیغاتی. برهمین اساس هرزنامه‌ها به ۳ دسته زیر تقسیم می-شوند[۴].

- هرزنامه تبلیغاتی[۳]: این نوع هرزنامه‌ها برای تبلیغ و معرفی کالا و خدمات به کار می‌روند و یکی از مهمترین علل گستردگی آن هزینه پایین تبلیغات نسبت به روش‌های دیگر است.
- هرزنامه‌های مالی[۴]: این نوع هرزنامه‌ها با سناریوهای مختلف سعی در فریب کاربران و دریافت پول از آن‌ها را دارد، بعنوان مثال به کاربر گفته می‌شود که در فلان قرعه‌کشی برنده شده‌اید ومی‌بایست برای دریافت جایزه هزینه‌ی را بپردازید.
- هرزنامه‌های کلاهبرداری[۵]: هدف این هرزنامه‌ها دستزسی به اطلاعات شخصی و محرمانه کاربران مانند نام کاربری و رمزعبور کارت اعتباری آن‌ها است.

نیاز به فیلترینگ هرزنامه

گرچه اینترنت روش‌های ارتباطی جدیدی را عرضه کرده است ولی سیل هرزنامه‌ها و هرزنامه‌نویس روز به روز در حال افزایش و تغییر ماهیت هستند. که نیازمند راه‌های فیلترینگ هوشمند و خودکار همچنان ضروری بنظر می‌رسد.

یادگیری ماشین: یادگیری ماشین یکی از حوزه‌های مهم هوش مصنوعی است. الگوریتم‌هایی در این حوزه قرار دارند که قابلیت یادگیری را داشته باشند، که باعث افزایش کارایی خود در واحد زمان را داشته باشند[۵].

انواع فیلترینگ‌های هرزنامه : در حال کلی فیلترینگ هرزنامه را می‌توان به دو دسته تقسیم کرد:

۱. مبتنی بر روش‌های یادگیری ماشین: در پالایش(فیلترینگ) هرزنامه‌ها از الگوریتم‌های یادگیری ماشین نیز می‌توان بهره برد. بدین صورت که هدف این الگوریتم‌ها ایجاد تمایز بین هرزنامه‌ها و نامه‌های معتبر می‌باشد. این روش‌ها توانایی استخراج دانش از یک مجموعه داده یا مجموعه مستندات را دارا می‌باشند و در حقیقت دانشی را که از یک مجموعه داده استخراج می‌کنند برای شناسایی هرزنامه‌های جدید مورد استفاده قرار می-دهند.

این الگوریتم‌ها دسته‌بندی مختلفی دارند که به اختصار توضیح می‌دهیم:

۱.۱یادگیری با مربی: منبع اصلی این دسته‌بندها داده‌های آموزشی برچسب دار می‌باشد.که دارای دو فاز است فاز آموزش[۶] و فاز آزمون[۷].

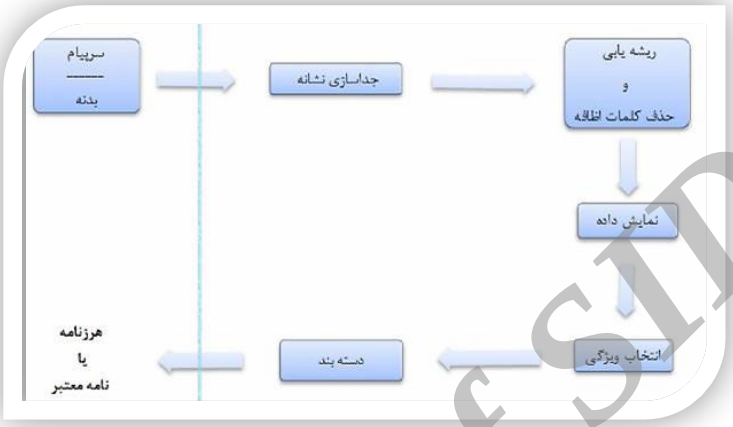
۱.۲یادگیری بدون مربی: تفاوت اصلی این الگوریتم‌ها با الگوریتم‌های یادگیری با مربی در فاز آموزش است، در این الگوریتم‌ها داده‌ای برچسب‌داری برای آموزش وجود ندارد تا با استفاده از برچسب آن‌ها یک مدل دسته بندی بسازد، بلکه این الگوریتم‌ها با توجه به شباهت داده‌ای جدید آن‌ها را به گروه‌های تقسیم می‌کند. مسائل خوشه بندی[۸] از جمله زیر مجموعه‌های یادگیری بدون مربی محسوب می‌شود.

البته علاوه بر دو دسته فوق، دسته‌ی دیگری وجود دارد که باعنوان یادگیری نیمه نظارتی[۹] شناخته می‌شود.

۲. مبتنی بر روش‌های عدم یادگیری ماشین: این فیلترینگ‌ها همانند فیلترینگ‌های مبتنی بر قانون، از الگوها و قوانین ثابتی جهت تشخیص هرزنامه‌ها استفاده می‌کنند. پیاده‌سازی اینگونه فیلترینگ‌ها اصولاً کاری ساده است و توانایی تشخیص هرزنامه‌ها پایین است. روش‌هایی مانند لیست سیاه وسفید از دسته فیلترینگ‌ها می‌باشند[۶].

[3] Advertisement
[4] Financial
[5] Phishing
[6] Train
[7] Test
[8] Clustering
[9] Semi Supervised

مراحل اصلی فیلترینگ هرزنامه: مراحل اصلی فیلترینگ هرزنامه که مبتنی بر محتوا است در شکل (۱) آمده است.



شکل(۱) مراحل اصلی فیلترینگ

۱. جداسازی نشانه[10]: استخراج کلمات از بدنه‌ی نامه
۲. حذف حروف اضافه: حذف حروفی که اغلب در بسیاری از متون و به طور مکرر دیده می‌شوند از این رو دارای ارزش محتوایی نیستند.
۳. ریشه‌یابی: باز گرداندن کلمات به ریشه‌ی اصلی خود.
۴. نمایش: تبدیل کردن مجموعه کلمات به فرم مشخص و مورد نیاز برای الگوریتم مورد نظر، مانند ساخت بردار ویژگی با استفاده از روش کیسه‌ی از کلمات.
۵. انتخاب ویژگی: انتخاب زیر مجموعه‌ی از کلمات نشان داده شده، که حاوی اطلاعات مفیدتری هستند. به عبارت دیگر حذف ویژگی‌های نامناسب از بردار ویژگی نشان داده شده در مرحله قبل.

الگوریتم‌های یادگیری ماشین در فیلترینگ هرزنامه

از آنجا که فیلترینگ هرزنامه را می‌توان گونه‌ای از دسته‌بندی متون به حساب آورد، بسیاری از الگوریتم‌های یادگیری ماشین که در دسته‌بندی متون کاربرد دارند، در شناسایی هرزنامه نیز می‌توانند مفید باشند. برخی از این الگوریتم‌ها عبارتند از درخت تصمیم، شبکه عصبی، بیزین، ماشین بردار پشتیبان، یادگیری جمعی و غیره می‌باشند. که در ادامه به معرفی الگوریتم‌های یادگیری جمعی می‌پردازیم.

الگوریتم‌های مبتنی بر یادگیری جمعی

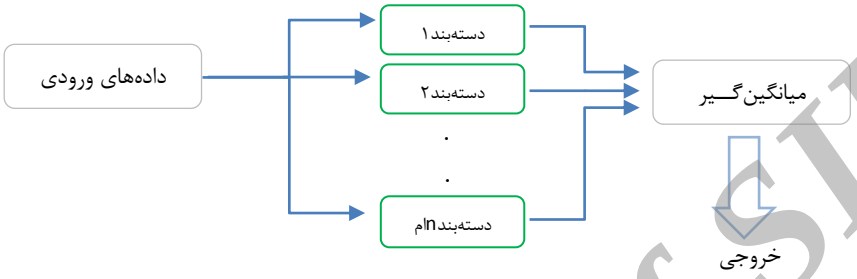
الگوریتم‌های یادگیری جمعی[11] که به نام‌های مختلفی چون ماشین‌های رایزن[12]، ترکیبی از چند دسته‌بند، مخلوطی از کارشناسان نیز شناخته می‌شوند. از اوایل دهه‌ی ۷۰ وارد پژوهش‌های پیرامون دسته‌بندی شدند ای الگوریتم‌ها جزء دسته‌بندی "یادگیری با مربی" هستند. تفاوت اصلی این الگوریتم‌ها با روش-های دیگر این است که به جای استفاده از یک دسته‌بند[13]، از چند دسته‌بند[14] که به آن‌ها دسته‌بندهای پایه گفته می‌شود، برای دسته‌بندی استفاده می‌کنند. که در نهایت این دسته‌بندها با یکدیگر به روشی خاص مانند رأی‌گیری، میانگین‌گیری، انتخاب بهترین فرد ترکیب می‌شوند[۷].

ساختار یادگیری جمعی

10
11 Ensemble Learning
12 Committee Machine
13 Single Classifier
14 Multiple Classifier

الگوریتم‌های یادگیری جمعی ساختارهای گوناگونی دارند و می‌توان آن‌ها را براساس معیارهای گوناگون به دسته‌های مختلفی تقسیم کرد، این الگوریتم‌ها از لحاظ تنوع الگوریتم‌های یادگیری پایه به دو دسته‌ی همگن و ناهمگن تقسیم می‌شوند، دریادگیری جمعی همگن تمامی دسته‌بندهای پایه انواع الگوریتم-های مبتنی بر یادگیری یکسان، مانند بیزین ساده، استفاده می‌کنند. در مقابل، یادگیری جمعی ناهمگن شامل تعدادی دسته‌بند پایه است که از الگوریتم‌های یادگیری متفاوتی بهره می‌برند. به عنوان مثال یکی از دسته بندهای پایه از ماشین بردارپشتیبان، یکی از K-نزدیکترین همسایه و دیگری از بیزین ساده استفاده می‌کنند[۸].

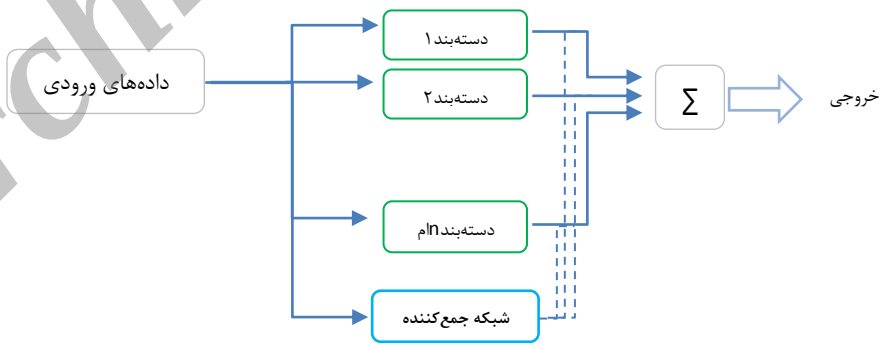
الگوریتم‌های یادگیری جمعی همچنین بر اساس چگونگی ترکیب دسته‌بندهای پایه به دوسته ایستا و پویا تقسیم می‌شوند. در الگوریتم‌های ایستا هیچ-گونه سیگنال ورودی در ترکیب دسته‌بندها استفاده نمی‌شود. از جمله معمول‌ترین روش‌ها روش میانگین‌گیری[۱۵] اشاره کرد. در این روش خروجی نهایی دسته‌بند به صورت ترکیب خطی از خروجی عوامل یادگیر به دست می‌آید. شکل (۳) ساختار میانگین‌گیری دسته‌جمعی را نشان می‌دهد.



شکل(۳) ساختار میانگین‌گیری دسته‌جمعی[۸]

سازوکارهای پویا شامل الگوریتم‌های هستند که یک سگنال ورودی در فرآیند ترکیب خروجی دسته‌بندهای پایه تأثیر دارد. در این الگوریتم‌ها علاوه بر عامل یادگیر قسمتی وجود دارد که درمورد چگونگی ترکیب خروجی عوامل یادگیر می‌باشد، به عبارت دیگر این بخش از الگوریتم با استفاده از ورودی و نیز اطلاعات مربوط به هر دسته‌بند پایه، وزن هر دسته‌بند را درجواب نهایی مشخص می‌کند.

روش "مخلوطی از کارشناسان" یا همان "پیش‌بینی با نظر خبرگان" از جمله معروف‌ترین روش‌های یادگیری جمعی پویا محسوب می‌شود. در این روش‌ها علاوه بر عوامل یادگیرپایه، یک واحد به نام شبکه‌ی جمع‌کننده[۱۶] نیز وجود دارد که به عنوان عامل تنظیم‌کننده‌ی تأثیر هر عامل یادگیر در جواب نهایی عمل می‌کند. به عبارتی در این روش‌ها دو مدل وجود دارد: یک یادگیری با مربی که عوامل یادگیرپایه از آن بهره می‌برند، و دیگری یادگیری بدون مربی که در بخش شبکه جمع‌کننده انجام می‌شود. در شکل(۳) این ساختار نشان داده شده است.



شکل(۴) ساختار ترکیبی از کارشناسان[۸]

مراحل ایجاد یادگیری جمعی

همانطور که پیش از این نیز گفته شد، یادگیری جمعی شامل مجموعه‌ای از دسته‌بندهای پایه و یک روش برای ترکیب نتایج دسته‌بندها می‌باشد. بنابراین برای ایجاد یک یادگیر جمعی دو مرحله وجود دارد.

- طراحی دسته‌بندهای پایه

[15] Averaging
[16] Gating Network

- انتخاب روشی برای ترکیب دسته‌بندهای پایه

۱. الگوریتم‌های یادگیری جمعی

در ادامه استفاده از داده‌های آموزشی در یادگیری با ساختار همگن اشاره خواهد شد که بعنوان نمونه الگوریتم‌های بگینگ[۱۷] و بوستینگ[۱۸] مورد بررسی قرار می‌گیرند که خود پایه‌ی بسیاری از الگوریتم‌ها هستند.

۱.۱. بگینگ

این الگوریتم یکی از ساده‌ترین الگوریتم‌های موجود در حوزه‌ی یادگیری جمعی محسوب می‌شود، که به آسانی پیاده‌سازی می‌شود و با این حال در آزمایش‌ها کارایی مناسبی از خود نشان داده‌است. بحث "گوناگونی" که در ایجاد دسته‌بندها به عنوان عامل اساسی در ساخت یک الگوریتم جمعی شناخته می‌شود، در این الگوریتم با استفاده از تولید زیرمجموعه‌های تصادفی همراه با جایگزینی[۱۹] از کل مجموعه داده آموزش، برای هر دسته‌بند ایجاد می‌شود. پس هر دسته‌بند بر روی یک مجموعه‌ی خاص آموزش می‌بینند، سپس این دسته‌بندها با استفاده از روش میانگین‌گیری و یا رأی اکثریت با یکدیگر ترکیب می‌شوند تا جواب نهایی را تولید کنند. این روش بیشتر زمانی مناسب است که حجم داده‌های مورد استفاده برای آموزش اندک باشد[۲]. نکته‌ی دیگری که در این الگوریتم حائز اهمیت است و در واقع یک نکته‌ی مثبت تلقی می‌شود این است که این الگوریتم توانایی ساخت دسته‌بندها به صورت با یکدیگر را دارد. چرا که دسته‌بندها به یکدیگر وابسته نیستند و هر یک بر روی زیر مجموعه‌ای تصادفی از کل فضای نمونه آموزش می‌بینند. از جمله الگوریتم‌هایی که براساس این ساختار طراحی و ارائه شده‌اند می‌توان به الگوریتم‌های Random Forest و Rotation Forest اشاره کرد.

۱.۲. بوستینگ

این الگوریتم جزء الگوریتم‌های قوی در زمینه‌ی یادگیری جمعی به حساب می‌آید. در این روش همانند روش بگینگ، تعدادی دسته بند وجود دارند که هر کدام با استفاده از یک مجموعه داده که به صورت تصادفی از مجموعه داده‌ای اصلی انتخاب می‌شوند آموزش دیده‌اند و با رأی‌گیری اکثریت با یکدیگر ترکیب شده‌اند. اما نکته‌ی اصلی در این الگوریتم این است که در این روش دسته‌بندها نمی‌توانند به صورت موازی و مستقل از هم آموزش ببینند بلکه به ترتیب ویکی پس از دیگری آموزش می‌بینند. در واقع در این روش‌ها اصولاً توزیع تصادفی که برای یک دسته‌بند استفاده می‌شود به گونه‌ای است که در آن احتمال انتخاب داده‌هایی که توسط دسته‌بندهای قبلی نادرست برچسب‌گذاری شده‌اند، بیشتر خواهد بود. بدین ترتیب می‌توان گفت که دسته‌بندهای انتهایی متخصص داده‌های سخت‌تر خواهند بود. الگوریتم AdaBoost از جمله معروف‌ترین الگوریتم‌ها با ساختار بوستینگ می‌باشد.

۱.۲.۱. Ada-Boosting مبتنی بر بوستینگ

در سال ۱۹۹۶ توسط Freund، Schapire خلق شد. دارای دو رویکرد زیر می باشد:

- ✓ انتخاب نمونه‌ها براساس خطای دسته‌بند قبلی (دارای رواج بیشتر)
- ✓ خطای وزن‌دار از مواردی که به اشتباه دسته‌بندی شده‌اند بالا است اما نه برای کلیه‌ی الگوریتم‌ها(تمام موارد تعبیه شده‌اند، اما وزن‌ها متفاوت است)

مراحل Ada-Boosting

- تعریف ε_k بعنوان مجموع احتمالات برای نمونه‌های که به اشتباه دسته‌بندی شده‌اند برای دسته‌بند C_k
- ضرب احتمالات مواردی که بصورت نادرست دسته‌بندی شده‌اند بوسیله:

$$\beta_k = (1-\varepsilon_k)/\varepsilon_k$$

- نرمال کردن مجدد احتمالات
- ترکیب دسته‌بندها $C_1,\ldots,C_k$ با استفاده از رأی‌گیری وزن‌دار مکانی که C_k بصورت $\log(\beta_k)$ وزن دهی شده است.

۱.۲.۲. Arcing مبتنی بر بوستینگ

در سال ۱۹۹۶ توسط Breiman ابداع شد.

مراحل کار:

- برای iمین نمونه در مجموعه آموزش، m_i به تعداد دفعاتی که به اشتباه دسته‌بندی بوسیله K دسته‌بند قبلی ارجاع می کند.

[17] Bagging
[18] Boosting
[19] Subsampling with Replacement

- احتمال P_i انتخاب نمونه i در دسته‌بند بعدی برابر:

$$P_i = \frac{1+m_i^4}{\sum_{j=1}^{N} 1+m_j^4}$$

اندازه گیری تجربی

مقایسه تجربی:

- ۲۳ مجموعه داده از مخازن UCI
- ۱۰ بار اعتبار سنجی متقاطع
- پس انتشار شبکه عصبی
- درخت‌های طبقه‌بندی
- منفرد و ساده (شبکه‌های عصبی چندتایی با وزن‌های مقداردهی اولیه متفاوت)، بگینگ، Ada-boost و Arcing
- # از دسته‌بند بصورت جمعی
- "دقت" به معیار کارایی

جدول۱: مقایسه بر اساس دو دسته‌بند

	دسته‌بند مبتنی بر درخت			شبکه عصبی			
	بگینگ	**Arcing**	**Ada**	**ساده**	**بگینگ**	**Arcing**	Ada
شبکه عصبی ساده	**-.10**	**.38**	**.37**	1	**.88**	**.87**	**.85**
شبکه عصبی بگینگ	**-.11**	**.35**	**.35**	**.88**	**1**	**.78**	**.78**
شبکه عصبی Arcing	**.14**	**.61**	**.60**	**.87**	**.78**	**1**	**.99**
شبکه عصبیAda	**.17**	**.62**	**.63**	**.85**	**.78**	**.99**	**1**
بگینگ مبتنی بر درخت دسته‌بند	**1**	**.68**	**.69**	**-.1**	**-.11**	**.14**	**.17**
Arcing مبتنی بر درخت دسته‌بند	**.68**	**1**	**.96**	**.38**	**.35**	**.61**	**.62**
Ad مبتنی بر درخت دسته‌بند	**.69**	**.96**	**1**	**.37**	**.35**	**.60**	**.63**

نتایج :

- دسته‌بند جمعی عموماً بهتر از دسته‌بند منفرد است.
- دسته‌بند دسته جمعی با شبکه عصبی‌ها و درخت قویاً وابسته هستند.
- Adaبوستینگ و Arcing قویاً وابسته هستند حتی در مقابل الگوریتم‌های متفاوت (بوستینگ بیشتر مبتنی بر مجموعه داده است تا نوع الگوریتم دسته‌بند)
- عموماً شبکه‌های عصبی بهتر از درخت هستند.

نتیجه گیری:

- بوستینگ دارای عملکرد ضعیف با نویز بیشتر است.

منابع مورد استفاده:

[۱]Y. Yang and J. O. Pedersen, "A comparative Study on Feature selection in Text Categorization", In Proceedings of 14th International Conference on Machine Learningx, pp ۴۱۲-۴۲۰, ۲۰۰۸.

[۲]L. S. Larkey and W. B. Croft, "Combining Classifiers in Text Categorization", In Proceedings of 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieve, pp. 289-297, 1996 .

[۳]D. Opitz, "Feature Selection for Ensembles", In Proceedings of 16th National Conference on Artificial Intelligence, pp. 379-384, 1999 .

[۴]G. P. C. Fung, J. X. Yu, H. Wang, D. W. Cheung and H. Liu, "A Balanced Ensemble Approach to Weighting Classifiers for Text Classification", In Proceedings of the 6th International Conference on Data Mining (ICDM '06), pp. 869-873, 2006 .

[۵]K. Wood, W. P. Kegelmeyer and K. Bowyer, "Combination of Multiple Classifiers using Local Accuracy Estimates", IEEE Transaction on Pattern Analysis and Machine, vol. 19x, pp.۴۰۵-۴۱۰, ۱۹۹۹.

[۶] E. M. D. Santos, R. Sabourin and P. Maupin, "A Dynamic Overproduce-and-Choose Strategy for the Selection of Classifier Ensembles", Pattern Recognition, vol. 41, pp.2993-3009, 2008.

[۷] C. Schaffer, "Selection a Classification Method by Cross-Validation", Machine Learning, vol. 13, pp.135-143, 1993.

[۸] G. Brown, "Ensemble Learning", In: Encyclopedia of Machine Learning, Springer Press, Heidelberg, 2010.

[۹] N. Littlestone and M. K. Warmuth, "The Weighted Majority Algorithm. Information and Computation", 108(2):212- 261, 1994.