بهبود کارایی سیستم‌های توصیه‌گر در مواجه با مساله شروع سرد با استفاده از تحلیل رفتار کاربران در شبکه‌های اجتماعی

مهدیه رفیعی[۱] بهروز شاهرخ‌زاده [۲]

۱- دانش آموخته کارشناسی ارشد- گروه مهندسی کامپیوتر و فناوری اطلاعات، واحد قزوین، دانشگاه آزاد اسلامی، قزوین، ایران
m.rafiei@qiau.ac.ir

۲- استادیار- گروه مهندسی کامپیوتر و فناوری اطلاعات، واحد قزوین، دانشگاه آزاد اسلامی، قزوین، ایران
bshahrokhzadeh@qiau.ac.ir

چکیده: هدف سیستم‌های توصیه‌گر معرفی آیتم‌هایی به کاربران است که می‌تواند موردعلاقه آنها باشد. یکی از چالش‌های اصلی در عملکرد سیستم‌های توصیه‌گر، مشکل شروع سرد است. زمانی که کاربر یا آیتم جدیدی به مجموعه اضافه می‌شود، سیستم به دلیل عدم وجود اطلاعات کافی نمی‌تواند پیشنهادهای مناسبی را ارائه کند. در این مقاله رویکردی ارائه می‌شود که در آن از داده-های رسانه‌های اجتماعی مانند توئیتر برای ایجاد یک پروفایلِ رفتاری استفاده می‌شود. سپس با استفاده از تکنیک‌های یادگیری ماشین، پروفایل‌های کاربران خوشه‌بندی می‌شوند. براساس این خوشه‌بندی‌ها، پیش‌بینی‌هایی با استفاده از الگوریتم جنگل تصادفی و ارتقای گرادیان ایجاد می‌شود. در این فرآیند، لازم نیست کاربر هیچ نوع داده‌ای را به طور صریح ارائه دهد. درنتیجه با کمک اطلاعات شبکه‌های اجتماعی کاربران، مشکل شروع سرد کاهش می‌یابد. بدین ترتیب که با این داده‌ها، یک پروفایل کاربری ایجاد شده و به عنوان ورودی سیستم توصیه‌گر استفاده می‌شود. آزمایش‌های متعددی انجام شد و در مقایسه با برخی از الگوریتم‌های جدید شروع سرد، نتایج رضایت‌بخشی حاصل شد. در این مقاله به این نتیجه رسیده‌ایم که فرایند خوشه‌بندی میزان دقت عملکرد مدل‌ها را بسیار بالا می‌برد و میانگین خطای مطلق را کاهش می‌دهد و همچنین الگوریتم ارتقای گرادیان نسبت به الگوریتم جنگل تصادفی از کارایی بهتری برخوردار است.

واژه‌های کلیدی: سیستم‌های توصیه‌گر، مساله شروع سرد، رسانه اجتماعی، خوشه‌بندی، جنگل تصادفی، ارتقای گرادیان.

نوع مقاله: پژوهشی

DOI: 10.52547/jiaeee.20.1.59

تاریخ ارسال مقاله: ۱۴۰۰/۷/۲۴

تاریخ پذیرش مشروط مقاله: ۱۴۰۱/۰۱/۲۶

تاریخ پذیرش مقاله: ۱۴۰۱/۶/۵

نام نویسنده‌ی مسئول: دکتر بهروز شاهرخ زاده

نشانی نویسنده‌ی مسئول: ایران – قزوین – میدان جانبازان – بلوار نخبگان – دانشگاه آزاد اسلامی واحد قزوین – گروه مهندسی کامپیوتر و فناوری اطلاعات

۱- مقدمه

با گذشت سال‌ها، مقدار محتوایی که در اینترنت یافت می‌شود، درحال رشد است. بنابراین سیستم‌های توصیه‌گر با کمک در یافتن آنچه دنبال آن هستیم این مشکل را تسکین می‌دهند. سیستم‌های توصیه‌گر در زمینه‌های متعددی به ویژه تجارت الکترونیک اجرا شده‌اند. در واقع، این سیستم‌ها نقش حیاتی در وب سایت‌های رتبه‌بندی شده مانند آمازون، یوتوب، نت‌فلیکس و .. دارند [۱]. با این‌حال، این روش‌ها از مشکل شروع سرد رنج می‌برند. مشکل شروع سرد به این معناست که اگر کاربر جدید یا آیتم جدیدی به سیستم اضافه شد، سیستم نتواند پیشنهادات خوب در این زمینه ارائه دهد [۲]. برای مثال از مشکل شروع سرد، می‌توانیم یک وبسایت فیلم را درنظر بگیریم که فاقد هرگونه اطلاعات از کاربرانی است که تازه به آن پیوستند. بنابراین مشخص نیست که چه فیلمی برای توصیه به کاربر جدید مناسب است. اینگونه سایت‌ها برای حل این مشکل از کاربران می‌خواهند اطلاعاتی درمورد علایق یا سرگرمی‌هایشان بدهند. برخی از آنها از کاربران می‌خواهند در رای‌گیری شرکت کنند [۳]. اشکال این روش‌ها این است که کاربران تمایل ندارند برای شرکت در این رای‌گیری‌ها زمان صرف کنند یا اطلاعاتشان را در اختیار این سایت‌ها قرار بدهند. در چنین شرایطی می‌توان از اطلاعاتی که کاربر در شبکه‌های اجتماعی خود قرار داده است استفاده کرد. در واقع می‌توان مقدار زیادی اطلاعات از شبکه‌های اجتماعی کاربر مانند توئیتر، فیسبوک و .. استخراج کرد و آن‌ها را به دانش مفید تبدیل کرد و از آن‌ها در سیستم‌های توصیه‌گر استفاده کرد [۴, ۵]. این روش به طور قابل توجهی تعامل با کاربران را کاهش می‌دهد. زیرا نیاز به اقدامات زیادی از جانب آن‌ها نیست.

براساس ملاحظات بالا، در این مقاله هدف ساختن یک پروفایل رفتاری از جریان اجتماعی کاربر در شبکه‌های اجتماعی وی است. این پروفایل شامل اطلاعاتی از قبیل سلیقه‌ها، ترجیحات و شخصیت کاربران است. به این منظور، دو منبع داده‌ی توئیتر و مووی توئیتینگ[1] را با هم ادغام می‌کنیم و مجموعه داده جدید فراهم می‌کنیم تا بتوانیم داده‌های جریان اجتماعی کاربر را با داده‌های رتبه‌بندی تلفیق کنیم. استخراج اطلاعات کاربران از شبکه اجتماعیِ توئیتر صورت می‌گیرد. سپس با استفاده از تکتیک‌های یادگیری ماشین کاربران را براساس پروفایل آن‌ها خوشه‌بندی می‌کنیم. یک مدل پیش‌بینی برای هریک از خوشه‌ها ایجاد خواهد شد تا پیش‌بینی کند که آیا یک مورد خاص به کاربر تعیین‌شده توصیه شود یا خیر. برای بررسی عملکرد سیستم، آن را در حوزه فیلم به کار بردیم [۶]. زیرا این موضوعی است که مردم زیادی به آن علاقمند هستند و همچنین فیلم‌ها یک موضوع بسیار رایج و متداول در شبکه‌های اجتماعی هستند. نوآوری روش ما بصورت زیر خلاصه می‌شود:

۱- تلفیق دو منبع داده‌ی رسانه‌های اجتماعی کاربران و رتبه‌هایی که کاربران به فیلم‌ها می‌دهند برای ساخت پروفایل رفتاری.

۲- استفاده از متد خوشه‌بندی برای دسته‌بندی کاربران براساس سلیقه، شخصیت و رفتار آنها.

۳- پیش‌بینی مدل با استفاده از الگوریتم ارتقای گرادیان. هدف نشانه‌ای خواهد بود که نشان می‌دهد آیا یک فیلم به کاربر توصیه شده است یا خیر.

نتایج بدست‌آمده پس از تایید و مرحله آزمایش، رفتار مطلوبی از مدل پیشنهادی را نشان می‌دهد، بنابراین برای به کاربردن آن در یک محیط واقعی بسیار مناسب است.

بقیه مقاله بصورت زیر ساختاربندی شده است. در بخش ۲ مقدمات لازم و کارهای پیشین بررسی می‌شود. در بخش ۳ پیشنهاد خود را توضیح می‌دهیم. در بخش ۴ آزمایش‌ها و ارزیابی سیستم را توصیف می‌کنیم و در نهایت در بخش ۵ به نتایج و اظهارات پایانی و کارهای آینده اشاره شده است.

۲- کارهای گذشته

ما می‌توانیم در ادبیات طرح‌های مختلفی را بیابیم که هدف آن‌ها حل مشکل شروع سرد است. درسال ۲۰۱۶ سان [۷] مطالعه‌ی مقایسه‌ای از طرح‌های مختلف پیشنهاد داده است که برخی از آنها در اینجا ذکر خواهد شد. در سال ۲۰۲۱ طهماسبی و همکاران [۸] از داده‌های جمعیتی کاربر در کنار تکنیک‌های گسترش پروفایل به منظور غنی‌سازی مجموعه کاربران همسایه استفاده می‌کردند. برای این منظور، یک تابع شباهت ترکیبی در نظر گرفته می‌شود که براساس داده‌های جمعیت شناختی و ماتریس رتبه‌بندی آیتم کاربر است. همچنین، پروفایل های رتبه‌بندی کاربران با استفاده از دو تکنیک مختلف برای کاهش مشکل شروع سرد در سیستم‌های توصیه‌گر توسعه داده می‌شوند. در سال ۲۰۲۰ ژانگ و همکارانش [۹] با استفاده از مدل زنجیره مارکوف برای تعبیه کاربر در شبکه‌های همسایه استفاده می‌کردند به طوریکه تازه واردان بدون ارتباط هم می‌توانند توسط کاربران مشابه نشان داده شوند. این روش از اطلاعات زمانی و روابط اجتماعی برای بهبود مشکل شروع سرد استفاده می‌کند. در سال ۲۰۱۷ هرناندو و همکارانش [۱۰] یک روش جدید برای کاربران غیر ثبت‌نام شده در یک سیستم توصیه‌گر ارائه دادند که با استفاده از قوانین استنتاج و یک مدل احتمالی مشکل شروع سرد را بهبود می‌بخشد. در سال ۲۰۱۳ چن و همکاران [۱۱] یک روش توصیه شروع سرد برای کاربران جدید پیشنهاد دادند که با استفاده از یک مدل، کاربر را با شبکه‌های اعتماد و بی‌اعتمادی ادغام می‌کند تا کاربران قابل اعتماد را شناسایی کند. در سال ۲۰۱۲ بابادیلا و همکاران [۱۲] یک معیار تشابه جدید را با استفاده از بهینه‌سازی مبتنی بر یادگیری عصبی ارائه دادند که نشان می‌دهد چگونه می‌توان معیارهای کیفیت اصلی یک سیستم توصیه‌گر را با استفاده از اعتبارسنجی متقابل بدست آوریم. آلمازرو و همکارانش در سال ۲۰۱۰ [۱۳] رویکردی ترکیبی مبتنی بر جمعیت و فیلتر مشارکتی در حوزه فیلم با استفاده از داده

های جمعیت شناختی برای بهبود روند توصیه معرفی کردند. منگ و همکارانش در سال ۲۰۱۳ [۱۴] از اطلاعات اضافی، مانند زیرمجموعه اجتماعی و مدل تصمیم‌گیری آنتولوژی، برای کمک به توصیه در مسئله شروع سرد استفاده کردند. در سال ۲۰۱۳ سافوری و همکاران [۱۵] چارچوبی را برای ارزیابی تاثیر ویژگی‌های جمعیتی بر درجه‌بندی کاربر ارایه دادند. این چارچوب با استفاده از مجموعه داده فیلم برای ارزیابی دقت توصیه‌های تولید شده مورد بررسی قرار گرفت. در سال ۲۰۱۵ روسلی و همکارانش [۱۶] یک معیار جدید را با ترکیب مقادیر شباهت به‌دست‌آمده از یک صفحه فیس بوک طراحی کردند. تمام مقادیر شباهت برای تولید مقدار شباهت کاربر جدید ادغام شدند. سان و همکاران در سال ۲۰۱۱ [۱۷] کاربران را براساس ماتریس رتبه-بندی کاربر-آیتم خوشه‌بندی کردند و سپس از نتایج خوشه‌بندی و اطلاعات جمعیت شناختی کاربران برای ساخت یک الگوریتم درخت تصمیم به منظور دستیابی به ارتباط بین کاربران موجود و کاربران جدید استفاده شد. پیش‌بینی برای کاربران جدید با ترکیب الگوریتم درخت تصمیم‌گیری با الگوریتم فیلترینگ مشترک انجام شد.

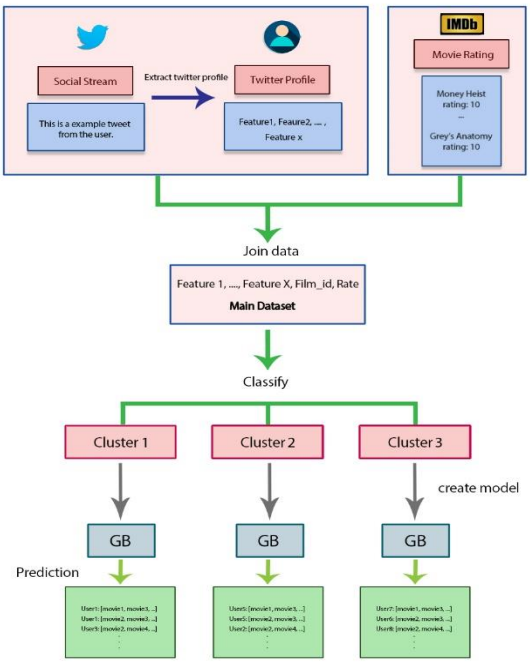
۳- روش پیشنهادی

در این بخش یک سیستم توصیه‌گر پیشنهاد می‌شود که در آن ورودی فقط جریان اجتماعی کاربران در شبکه اجتماعی توئیتر است. پس از استخراج اطلاعات لازم از پروفایل توئیتر کاربران، آنها را با استفاده از الگوریتم K-means خوشه‌بندی می‌کنیم. سپس به ازای هر خوشه با استفاده از الگوریتم جنگل تصادفی و ارتقای گرادیان، یک مدل طراحی می‌کنیم و با استفاده از آن مدل، پیش‌بینی می‌کنیم که آیا چنین آیتمی برای توصیه به آن کاربر مناسب است یا خیر. مدل پیشنهادی برای توصیه فیلم است. هدف این خواهد بود که آیا یک فیلم خاص مناسب است یا نه و برای کاربر خاصی توصیه شود یا خیر. در شکل (۱) می‌توانیم یک دیاگرام از روش پیشنهادی را ببینیم.

۳-۱- جمع‌آوری داده‌ها

اولین چالش یافتن داده‌های مناسب است که به ما اجازه دهد طرح پیشنهادی خود را توسعه داده و اجرا کنیم. باید کاربرانی که دارای داده‌های رتبه‌بندی فیلم‌ها و همچنین داده‌های جریان اجتماعی هستند، پیدا کنیم. توئیتر برای تعریف کاربران و مووی‌توئیتینگ [۱۸] برای به دست آوردن رتبه‌هایی که کاربران به فیلم‌ها می‌دهند.

مووی‌توئیتینگ یک مجموعه داده متشکل از رتبه‌بندی فیلم‌ها است. در این مجموعه داده از دسترسی گسترده رسانه‌های اجتماعی استفاده کردند و یک مجموعه داده رتبه‌بندی فیلم‌های جدید را بر اساس توئیت‌های عمومی و ساختاریافته آماده کردند. این مجموعه داده همیشه به‌روز است تا زمانیکه توئیتر و IMDb همیشه به روز باشد. زیرا این مجموعه داده از داده‌های توئیتر[۲] و فیلم‌های سایت IMDb[۳] استفاده می‌کند. برای جمع‌آوری داده‌های رسانه‌های اجتماعی از API توئیتر[۴] استفاده کردیم و تمام توئیت‌هایی که برای آنها اطلاعاتی درمووی‌توئیتینگ داریم، جمع‌آوری کردیم. بعد از جمع‌آوری توئیت‌ها آنها را پردازش کردیم و برای هر کاربر یک پروفایل ایجاد کردیم. ویژگی‌های استخراج شده را می‌توان در جدول (۱) مشاهده کرد. از طرف دیگر لیستی از رتبه‌بندی‌های فیلم برای هر کاربر خواهیم داشت که شامل این اطلاعات است: شناسه فیلم[۵] (شناسه فیلم در مووی-توئیتینگ) و رتبه‌بندی[۶] (رتبه‌ای که کاربر خاص با شناسه[۷] به فیلم داده است).

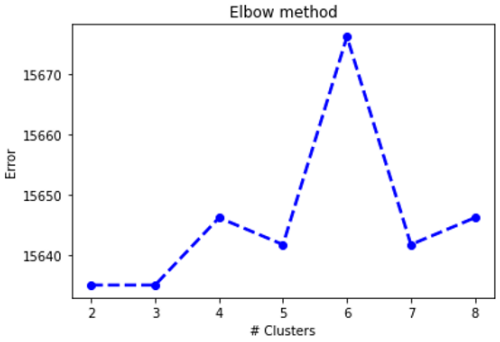


شکل (۱): دیاگرام اصلی روش پیشنهادی

۳-۲- ادغام کردن داده های فیلم و پروفایل توئیتر

همانطور که گفته شد و در جدول (۱) می‌بینیم، از ویژگی‌های استخراج‌شده از پروفایل توئیتر، پیش‌بینی‌هایی برای رتبه‌بندی فیلم‌ها ایجاد شد. رتبه‌بندی‌ها اعداد صحیح از ۱ تا ۱۰ هستند. از آنجا که امتیازبندی کاربرها به‌طور دقیق مهم نیست و تنها علاقه یا عدم علاقه کاربر به آن فیلم اهمیت دارد، با برچسب زدن به این داده‌ها در دو برچسب ممکن، این بازه مبتنی بر فاصله را به درجه‌بندی باینری تبدیل خواهیم کرد: رتبه‌بندی ۰ تا ۶ به ۰ نگاشت می شود، یعنی برای کاربر مناسب نیست و رتبه بندی ۷ تا ۱۰ به ۱ نگاشت شده یعنی برای وی مناسب است. برای این منظور، از مجموعه‌ی مشخصی فیلم برای شبیه‌سازی یک کاتولوگ استفاده می‌کنیم. کاربران براساس سلیقه، شخصیت و رفتار آن‌ها (پروفایل توئیتر[۸]) با روش خوشه‌بندی[۹] و الگوریتم K-Means به ۳ خوشه تقسیم می‌شوند. بعد از محاسبات باتوجه به شکل (۲) طبق روش elbow [۱۹] چون خطا در سه تا خوشه از همه کمتر است، بنابراین تعداد خوشه بهینه ۳ است. در مزیت انتخاب ۳ خوشه نسبت به ۲ خوشه، بحث کم برازش[۱۰] مدل پیش

می‌آید. با انتخاب ۲ خوشه مدل کم‌برازش می‌شود یعنی الگوریتم یک مدل خیلی کلی از مجموعه آموزشی به دست می‌آورد، که حتی اگر داده‌های مجموعه‌ی آموزشی را نیز به این الگوریتم بدهیم، الگوریتم خطای قابل توجهی خواهد داشت. سپس از الگوریتم‌های جنگل تصادفی و ارتقای گرادیان برای ساخت مدل در هر خوشه استفاده می‌شود، که هدف آن‌ها نشانه‌ای (پرچم) خواهد بود که نشان می‌دهد آیا یک فیلم توصیه شده است یا خیر که مقادیر آن می‌تواند ۰ یا ۱ باشد. در نهایت نتایج را با هم مقایسه می‌کنیم. در این مقاله، فیلم‌هایی که انتخاب شده، فیلم‌هایی هستند که امتیازهای بالای ۳۰۰ دریافت کردند (کاربرانی که پروفایل توئیتر آنها را داریم).



شکل (۲): روش elbow برای تعیین K بهینه در خوشه‌بندی

این یک فرآیند آموزشی خواهد بود که در آن، از بخشی از داده‌ها (۸۰ درصدِ داده‌ها) برای آموزش مدل خود استفاده خواهیم کرد. در این فرآیند، ارتباط بین ویژگی‌های مختلف از پروفایل توئیتر و برچسبِ[۱۱] هدف ایجاد خواهد شد. از آنجایی که می خواهیم پیش‌بینی‌های مربوط به چندین فیلم و نه فقط یک فیلم را بدانیم، مجبور خواهیم بود برای هر فیلم از یک مدل استفاده کنیم. تفاوت بین این مدل‌ها فقط در برچسب‌ها (برچسب پیشنهادی برای یک فیلم مشخص) خواهد بود. همه موارد دیگر (پروفایل کاربری توئیتر) در هر مدل یکسان خواهند ماند.

۳-۳- پیش‌بینی

باید پیش‌بینی شود کدام محصولات برای توصیه به کاربر مناسب‌تر هستند. به منظور دستیابی به این هدف، یک کاتالوگ محصول خواهیم داشت (کاتالوگ فیلم)، فهرستی از کاربران که برای آن، هم پروفایل توئیتر و هم پروفایل مووی‌توئیتینگ داریم. از این کاربران لیستی از تمام توئیت‌های موجود در توئیتر و فهرستی از رتبه‌بندی فیلم‌ها را داریم. هنگامی که مدل‌ها آموزش داده شوند، از داده‌های آزمایش (۲۰ درصدِ داده‌ها) به منظور ایجاد پیش‌بینی استفاده خواهیم کرد. از آنجایی که داده‌های واقعی کاربران (رتبه‌بندی فیلم‌ها) را داریم می‌توانیم پیش‌بینی‌ها و ارزیابی داده‌های واقعی را با هم مقایسه کنیم بنابراین دقت پیش‌بینی‌های خود را خواهیم داشت. در پایان این مرحله، لیستی از فیلم‌ها (آن‌هایی که در فهرست فیلم‌ها به کار می‌روند) و یک نشانه متناظر برای هر فیلم خواهیم داشت که نشان می‌دهد آیا سیستم، فیلم را برای کاربر توصیه می‌کند یا خیر.

جدول (۱): ویژگی های استخراج شده از پروفایل توئیتر کاربران

Id	user id	f7	preferred weekday
f1	account year of creation	f8	friends count
f2	early bird	f9	followers count
f3	night owl	f10	favorite's count
f4	preferred hour	f11	geo enabled
f5	weekend tweeter	f12	number of tweets
f6	week tweeter		

جدول (۲): نمونه پروفایل توئیتر با برچسب

id	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	rate	rec
35252	2008	FALSE	FALSE	18	FALSE	TRUE	6	87	124	21	TRUE	575	5	0
35253	2014	FALSE	FALSE	21	FALSE	TRUE	6	67	522	314	TRUE	1089	7	1
35257	2011	FALSE	FALSE	20	FALSE	TRUE	6	399	72	194	TRUE	718	6	0
35257	2011	FALSE	FALSE	20	FALSE	TRUE	6	399	72	194	TRUE	718	6	0
35257	2011	FALSE	FALSE	20	FALSE	TRUE	6	399	72	194	TRUE	718	6	0
35258	2010	FALSE	FALSE	6	FALSE	TRUE	0	302	17	705	TRUE	1162	10	1
35259	2010	FALSE	FALSE	4	FALSE	TRUE	3	121	119	9	TRUE	3251	7	1
35259	2010	FALSE	FALSE	4	FALSE	TRUE	3	121	119	9	TRUE	3251	6	0

۳-۴- انتخاب

تمام فیلم‌هایی را انتخاب می‌کنیم که برای کاربر قابل توصیه هستند و سپس فیلم‌هایی را انتخاب می‌کنیم که احتمال بالاتری دارند. این بدان معنی است که فیلم‌هایی که از مدل پیش‌بینی انتخاب می‌شود، با قطعیت بیشتری مشخص شده‌اند.

۴- نتایج ارزیابی

زمانی که الگوریتم جنگل تصادفی و الگوریتم ارتقای گرادیان، پیش‌بینی‌ها را ایجاد کردند، به ارزیابی کیفیت این پیش‌بینی‌ها خواهیم پرداخت و نتایج را نشان خواهیم داد. دقت پیش‌بینی‌ها را با شاخص‌های زیر اندازه گرفتیم. این شاخص‌ها به این دلیل انتخاب شده‌اند که آنها رایج‌ترین معیارها برای ارزیابی الگوریتم‌های طبقه‌بندی هستند، به خصوص دقت[۱۲]، میانگین خطای مطلق[۱۳] و معیارِ F[۱۴].

- دقت (ارزش پیش‌بینیِ مثبت). درصد موفقیت از توصیه را در مقابل شکست مشخص می‌کند.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (۱)$$

TP: نشان می‌دهد که پیش‌بینی مثبت است و به طور معمول درست است.

TN: این بدان معناست که پیش‌بینی منفی است و به طور معمول درست است.

FP: نشان می‌دهد که پیش‌بینی مثبت است و معمولاً نادرست است.

FN: نشان‌دهنده پیش‌بینی منفی و غلط است. جایی که TP مثبت واقعی را نشان می دهد، TN منفی واقعی را نشان می‌دهد، FP، مثبت کاذب را نشان می دهد و FN منفی کاذب را نشان می‌دهد. در مطالعه

واقعی، مقادیر واقعی و غلط نشان داده می‌شوند، در حالی‌که مقادیر پیش‌بینی با مثبت و منفی نشان داده می‌شوند [۲۰].

- معیار F. معیار F معیاری برای اندازه‌گیری دقت یک تست است. این آزمون به صورت میانگین هارمونیک وزن‌دار دقت و فراخوان آزمون F1 تعریف می‌شود.

$$F = \frac{2 * precision}{precision + recall} \quad (۲)$$

معیار precision: (معیار صحت) معیاری است که به ما می‌گوید الگوریتم چند درصدِ مثبت‌هایش درست بوده است.

$$precision = \frac{TP}{TP + FP} \quad (۳)$$

معیار recall: به دنبال محاسبه‌ی پوشش بر روی کل داده‌ها است. تمرکز اصلی این معیار برخلاف معیار صحت بر روی داده‌هایی است که واقعاً درست بوده‌اند.

$$recall = \frac{TP}{TP + FN} \quad (۴)$$

- میانگین خطای مطلق.

$$MAE = \frac{\sum |y_i - \hat{y}_i|}{n} \quad (۵)$$

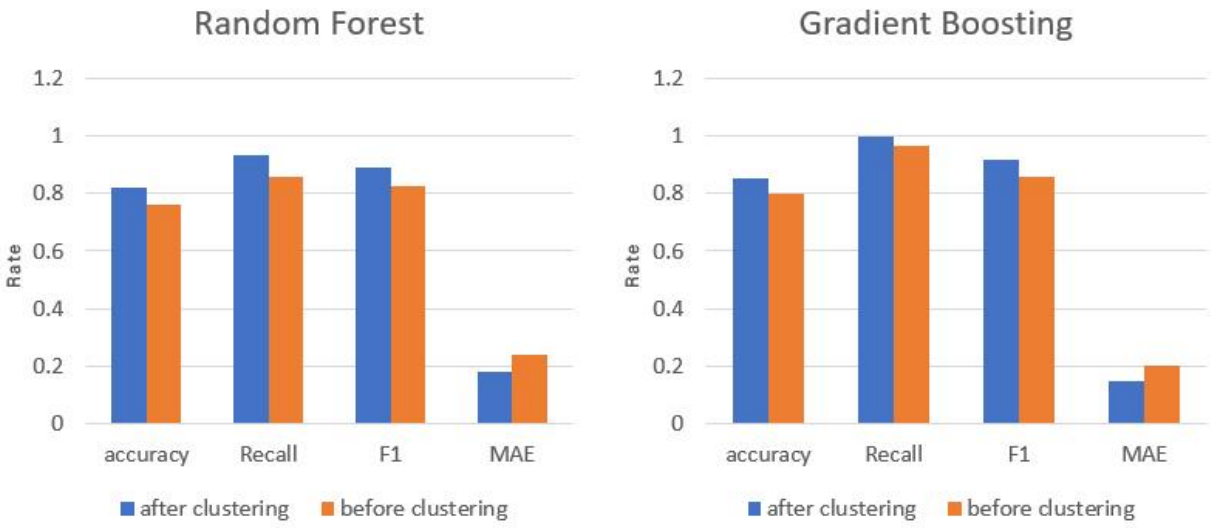
با استفاده از مجموعه داده مووی‌توئیتینگ، تعداد کل کاربرانی که به فیلم‌ها رای دادند و پروفایل کاربری توئیتر دارند، ۷۰۷۳۴ کاربر بود. از این کاربران در مجموع ۹۰۶۱۸۳ رای[۱۵] داریم تا آزمایش‌های خود را انجام دهیم. تعداد کل فیلم‌های رای داده‌شده ۳۷۳۱۸ است پس از ادغام مجموعه داده مووی‌توئیتینگ و داده های توئیتر، دیتاست نهایی را تهیه کردیم و تعداد ۴۸۶۲۶۰ داده در مجموع جمع‌آوری شده است. ورودی‌های ما شبیه نمونه جدول (۲) خواهد بود. می‌توانیم ببینیم که هیچ کدام از ستون‌های رتبه‌بندی را نداریم اما ستونی را داریم که به طور مستقیم از رتبه محاسبه می‌شود و در صورت مناسب بودن فیلم برای کاربر، دارای مقدار ۱ خواهد بود. در غیر این صورت مقدار صفر خواهد بود. سپس با استفاده از الگوریتم k-means تمام داده‌هایی که از توئیتر استخراج شده است را خوشه‌بندی می‌کنیم. روش elbow، مجموع فواصل درون خوشه‌ای داده‌ها را به عنوان تابعی از تعداد خوشه‌ها در نظر می‌گیرد. به این ترتیب تعداد خوشه‌ها به نحوی انتخاب می‌شوند که افزودن یک خوشه دیگر، بهبودی در حداقل‌سازی WSS[۱۶] ایجاد نکند. باتوجه به اینکه محدوده‌ی جستجو کوچکتر شده است و هر کاربر با کاربر شبیه خودش سنجیده می‌شود، میزان دقت الگوریتم تا حد خوبی افزایش پیدا می‌کند.

اکنون برای هر خوشه یک جنگل تصادفی و یک الگوریتم ارتقای گرادیان ایجاد خواهیم کرد. برای این کار از ۸۰٪ داده‌ها برای آموزش مدل‌های مختلف و از ۲۰٪ باقی‌مانده برای ارزیابی پیش‌بینی‌ها استفاده می‌کنیم. براساس مدل‌های آموزش دیده‌شده و براساس ویژگی‌های ذکرشده، یک مدل طبقه‌بندی شده ایجاد خواهد شد. مدل‌ها را با داده‌های آموزشی خود آموزش می‌دهیم و دو محاسبه موازی انجام می‌دهیم. جنگل تصادفی و الگوریتم ارتقای گرادیان. بعد از چندین اجرای پیش‌بینی‌هایمان، از چندین معیار استفاده کردیم. مقادیر خطا برای ۲۰ اجرا انجام شده است اما صرفا یک اجرا برای نمونه در جدول (۳) نشان داده شده است. برای هر اجرا، ما دو معیار ذکرشده را برای مدل ساخته‌شده با جنگل تصادفی و همچنین مدل ساخته‌شده با الگوریتم ارتقای گرادیان محاسبه می‌کنیم. به این ترتیب می‌توانیم بین هر دو مدل مقایسه‌ای انجام دهیم.

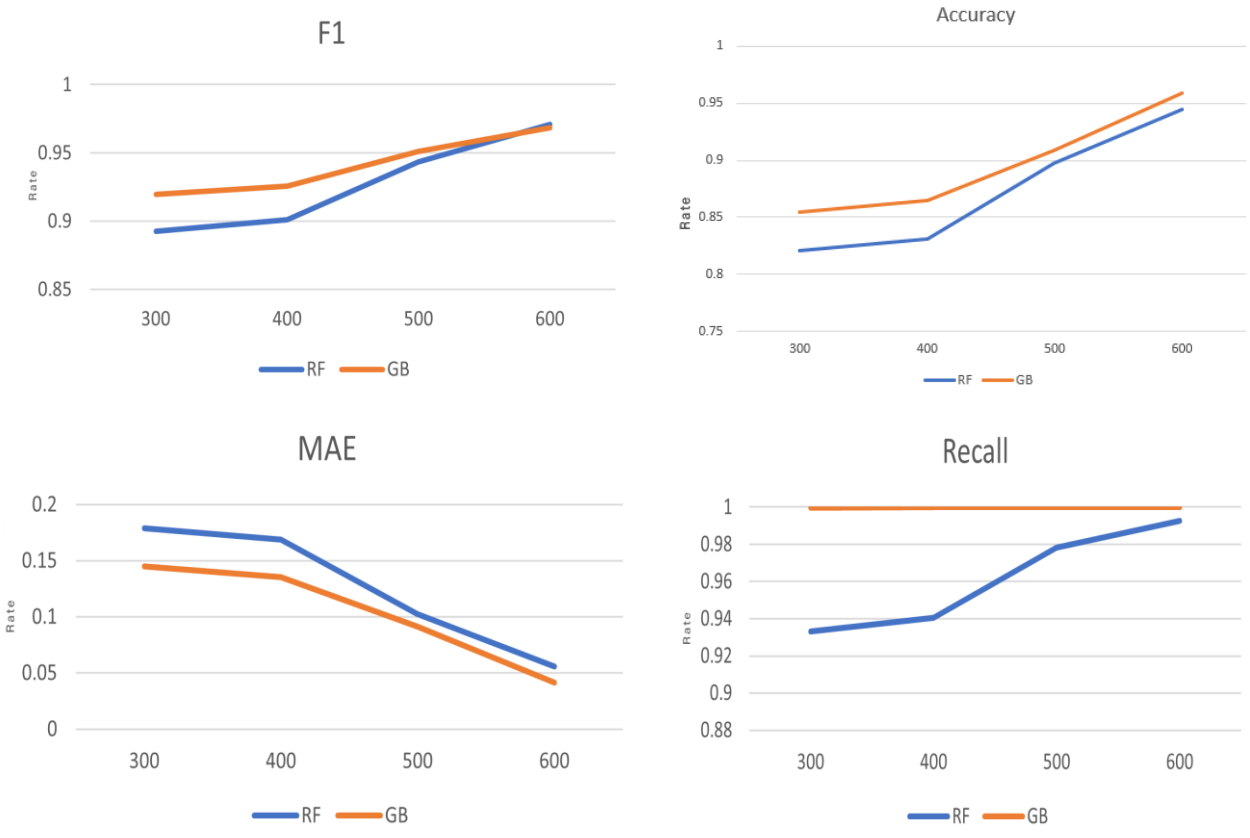
جدول (۳): اعتبارسنجی پیش بینی ها برای ۵ اجرا

Iteration	1	2	3	4	5
avg accuracy error (RF)	0.845763	0.851188	0.851326	0.851091	0.847843
avg accuracy error (GB)	0.875611	0.875611	0.875611	0.875611	0.875611
avg MAE error (RF)	0.154237	0.148812	0.148674	0.148909	0.152157
avg MAE error (GB)	0.124389	0.124389	0.124389	0.124389	0.124389
avg f1 error (RF)	0.911596	0.915667	0.914751	0.915391	0.912102
avg f1 error (GB)	0.932356	0.932356	0.932356	0.932356	0.932356

تمام این معیارها برای هر یک از مدل‌های مختلف محاسبه می‌شوند و سپس میانگین همه‌ی آنها محاسبه خواهد شد. این مقداری است که ما در هر سلول از جدول نشان می‌دهیم. بنابراین می‌توانیم بگوییم مدل ارتقای گرادیان طبق جدول (۳) بسیار بهتر عمل می‌کند. چیزی که مورد انتظار بود. با دیدن این نتایج می‌توان ادعا کرد که با توجه به اینکه از داده‌های رتبه‌بندی قبلی کاربر برای جستجوی موارد مشابه استفاده نمی‌کنیم، این نتایج بسیار مثبت هستند. زیرا ما تنها از داده-های جریان اجتماعی کاربر استفاده می‌کنیم. بعد از فرایند خوشه‌بندی مقادیر دقت تقریبا ۶ درصد افزایش پیدا کرده است و همچنین مقادیر میانگین خطای مطلق تقریبا با کاهش ۱۰ درصدی روبه‌رو شدند. در شکل (۳) می‌توانیم مقایسه‌ی معیارهای ارزیابی را در دو الگوریتم جنگل تصادفی و ارتقای گرادیان قبل و بعد از خوشه‌بندی بصورت یک‌جا ببینیم. همانطور که مشاهده می‌شود، نتایج بعد از خوشه‌بندی رشد بهتری داشته است. همچنین الگوریتم ارتقای گرادیان به‌قدری خوب عمل کرده است که معیار recall را به عدد ۱۰۰ درصد رسانده است. به این معنی که اگر مدل پیش بینی کند که label=1 هست یعنی صددرصد ۱ است. مقدار دقت ۸۵ درصد شده است یعنی ۸۵ درصد پیش بینی از کل داده ها درست بوده، تقریبا ۶ درصد افزایش پیدا کرده است، معیار F1 هم که میانگین هارمونیک Precision و Recall است، حدود ۹۲ درصد شده است و میانگین خطای مطلق هم که تفاضل خروجی مدل از لیبل‌ها است -که هرچقدر کمتر باشد یعنی خروجی مدل به لیبل‌ها نزدیکتر است- حدود ۱۵درصد است و تقریبا با کاهش ده درصدی روبه‌رو شده‌اند. ما الگوریتم جنگل تصادفی را با الگوریتم ارتقای گرادیان مقایسه کردیم زیرا هر دو جزو روش‌های کلاسه‌بندی جمعی[۱۷] هستند و مقایسه‌ی آن‌ها جایز است و طبقه‌بندی کننده پایه[۱۸] هم درخت تصمیم است.



شکل (۳): مقایسه معیارهای ارزیابی در الگوریتم‌های جنگل تصادفی و ارتقای گرادیان قبل و بعد از خوشه‌بندی



شکل (۴): مقایسه‌ی معیارهای ارزیابی در الگوریتم جنگل تصادفی و الگوریتم ارتقای گرادیان در فیلم‌هایی با رتبه‌های بالاتر از ۳۰۰ تا ۶۰۰

جنگل‌های تصادفی که جزو روش‌های Bagging هستند، با استفاده از یک نمونه تصادفی از داده‌ها، هر درخت را به طور مستقل آموزش می دهند. این تصادفی بودن باعث می‌شود که مدل قوی‌تر از یک درختِ تصمیمِ واحد باشد و کمتر از داده‌های آموزش استفاده کند.

ارتقای گرادیان‌ها هر بار یک درخت را آموزش می‌دهند و هر درخت جدید به اصلاح خطاهای درختان آموزش داده شده قبلی کمک می‌کند. با افزودن هر درخت، مدل حتی قوی‌تر هم می‌شود. تفاوت اصلی بین این دو الگوریتم ترتیب آموزش هر جزء درخت است. آزمایش‌ها را در سه سطح دیگر تکرار کردیم و فیلم‌هایی را درنظر گرفتیم که امتیازهای بالای ۴۰۰ و ۵۰۰ و ۶۰۰ داشتند. این عملیات برای این است که بررسی کنیم، آیا تعداد امتیازهایی که به فیلم‌ها داده شده، تاثیری در نتیجه کار دارد یا خیر؟ در شکل (۴) مشاهده می‌شود برای رسیدن به دقت بالاتر و خطای کمتر، بهتر است فیلم‌هایی را انتخاب کنیم که بیشترین امتیاز را در بین گروه کاربران دارند. پس نتیجه می‌گیریم الگوریتم ما در تمام حالات عملکرد خوبی دارد و از نتایج قابل اعتمادی برخوردار است.

۵- نتیجه‌گیری

در این پژوهش رویکردی را براساس یک مدل پیش‌بینی، با استفاده از اطلاعات رفتاریِ کاربران از استخراج داده‌های جریان اجتماعیِ آنها ارائه کردیم. سپس کاربران را با توجه به پروفایلِ رفتاری آنها خوشه‌بندی کردیم و برای هر خوشه با استفاده از الگوریتم جنگل تصادفی و الگوریتم ارتقای گرادیان مدلی طراحی کردیم. کاربران نیازی به ارائه صریح هیچ‌گونه اطلاعات شخصی به غیر از منبع رسانه‌های اجتماعی خود ندارند و از این طریق به کاهش مشکل شروع سرد کمک می‌شود. انجام عمل خوشه‌بندی روی داده‌ها باعث بهبود عملکرد الگوریتم ما نسبت به سایر الگوریتم‌های مشابه شد. در رویکرد پیشنهادی با استفاده از تکنیک‌های یادگیری ماشین یعنی جنگل تصادفی و الگوریتم ارتقای گرادیان مدلی را طراحی کردیم تا برای کاربرانی که پروفایلشان خوشه‌بندی شده است یک نشانه اختصاص دهد که نشان می‌دهد آیا آن فیلم مناسب است یا خیر. این پیشنهاد در محیطِ توصیه‌ی فیلم انجام شده است و نتایج به‌دست‌آمده از پیش‌بینی‌های پیشنهادی بسیار رضایت‌بخش هستند. بنابراین می‌توانیم ارزیابی کنیم که الگوریتم ما، دارای شرایط نزدیک به بهینه در مساله شروع سرد است.

برای بهبود فرایند حل مشکل شروع سرد در ادامه‌ی پژوهش حاضر، می‌توان بین پروفایل کاربر که از توئیتر کاربر استخراج شده است و ویژگی‌های آیتم (فیلم‌ها) ارتباط برقرار کرد. برای مثال بین پروفایل توئیتر و ژانر فیلم (کمدی یا درام) یا بازیگرهای فیلم می‌توان ارتباط برقرار کرد و رفتار آن‌ها را بررسی کرد. این امر می‌تواند به ما کمک کند تا روابط دقیق‌تری بدست آوریم و در نتیجه توصیه‌های بهتری به‌دست آوریم.

مراجع

[1] Herce-Zelaya, J., et al., New technique to alleviate the cold start problem in recommender systems using information from social media and random decision forests. Information Sciences, 2020. 536: p. 156-170.

[2] Bobadilla, J., et al., Recommender systems survey. Knowledge-based systems, 2013. 46: p. 109-132.

[3] Zafarani, R., M.A. Abbasi, and H. Liu, Social media mining: an introduction. 2014: Cambridge University Press.

[4] Bernabé-Moreno, J., et al., Quantifying the emotional impact of events on locations with social media. Knowledge-Based Systems, 2018. 146: p. 44-57.

[5] Bernabé-Moreno, J., et al., Leveraging localized social media insights for industry early warning systems. International Journal of Information Technology & Decision Making, 2018. 17 :(۰۱)p. 357-385.

[6] Carrer-Neto, W., et al., Social knowledge-based recommender system. Application to the movies domain. Expert Systems with applications, 2012. 39(12): p. 10990-11000.

[7] Son, L.H., Dealing with the new user cold-start problem in recommender systems: A comparative review. Information Systems, 2016. 58: p. 87-104.

[8] Tahmasebi, F., et al., A hybrid recommendation system based on profile expansion technique to alleviate cold start problem. Multimedia Tools and Applications, 2021. 80(2): p.۲۳۵۴-۲۳۳۹ .

[9] Zhang, Y., et al., Joint Personalized Markov Chains with social network embedding for cold-start recommendation. Neurocomputing, 2020. 386: p. 208-220.

[10] Hernando, A., et al., A probabilistic model for recommending to new cold-start non-registered users. Information Sciences, 2017. 376: p. 216-232.

[11] Chen, C.C., et al., An effective recommendation method for cold start new users using trust and distrust networks. Information Sciences, 2013. 224: p. 19-36.

[12] Bobadilla, J., et al., A collaborative filtering approach to mitigate the new user cold start problem. Knowledge-based systems, 2012. 26: p. 225-238.

[13] Almazro, D., et al., A survey paper on recommender systems. arXiv preprint arXiv:1006.5278, 2010.

[14] Meng, C., et al. A method to solve cold-start problem in recommendation system based on social network sub-community and ontology decision model. in 3rd International Conference on Multimedia Technology (ICMT-13). 2013. Atlantis Press.

[15] Safoury, L. and A. Salah, Exploiting user demographic attributes for solving cold-start problem in recommender system. Lecture Notes on Software Engineering, 2013. 1(3): p. 303-307.

[16] Rosli, A.N., et al., Alleviating the cold-start problem by incorporating movies facebook pages. Cluster Computing, 2015. 18(1): p. 187-197.

[17] Sun, D., C. Li, and Z. Luo. A content-enhanced approach for cold-start problem in collaborative filtering. in 2011 2nd international conference on artificial intelligence, management science and electronic commerce (AIMSEC). 2011. IEEE.

[18] Dooms, S., T. De Pessemier, and L. Martens. Movietweetings: a movie rating dataset collected from twitter.

[19] Chaipornkaew, P. and T. Banditwattanawong. A recommendation model based on user behaviors on commercial websites using TF-IDF, KMeans, and Apriori algorithms. in International Conference on Computing and Information Technology. 2021. Springer.

[20] Yassin, S.S., Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach. SN Applied Sciences, 2020. 2(9): p. 1-13.

زیرنویس‌ها

[1] Movie Tweeting

[2] https://www.twitter.com.

[3] https://www.twitter.com.

[4] https://developer.twitter.com/en/docs/api-reference-index.

[5] Movie Id

[6] Rating

[7] Film Id

[8] Twitter profile

[9] Clustering

[10] Underfit

[11] Label

[12] Precision

[13] MAE

[14] F-measure

[15] Rate

[16] Within-cluster Sum of Square

[17] Ensemble Model

[18] Base Classifier