دانشگاه صنعتی شریف
دانشکده مهندسی کامپیوتر
سمینار کارشناسی ارشد گرایش هوش مصنوعی

عنوان:
یادگیری بدون برد با شبکه‌های عمیق
Deep Zero-Shot Learning

نگارش:
سید محسن شجاعی
۹۳۲۰۷۹۷۹

استاد راهنما:
دکتر مهدیه سلیمانی

استاد ممتحن داخلی:
دکتر حمیدرضا ربیعی

بهمن ۹۴

چکیده: این گزارش، به مسئله دسته‌بندی جویبار داده می‌پردازد. دسته‌بندی جویبار داده، یک مسئله دسته‌بندی برخط است که در آن، در طی زمان تغییر مفهوم رخ می‌دهد. این دو ویژگی سبب می‌شود که دسته‌بندی با چالش‌های زیادی مواجه شود. در این نوشتار، پس از معرفی چالش‌های این مسئله، سیستمی معرفی می‌کنیم که با استفاده از دسته‌بندهای برخط وفقی و انتخاب فعال داده‌ها برای برچسب‌گذاری و هم‌چنین استفاده از جمع‌سپاری، هزینه لازم برای دسته‌بندی جویبار داده را کاهش دهد. پس از آن، به بررسی قسمت‌های تشکیل‌دهنده این سیستم پرداخته و کارهای پیشین انجام شده در هر یک از این حوزه‌ها را مورد بررسی قرار می‌دهیم. در ادامه، با توجه به گستردگی مباحث مرتبط با این سیستم، بر روی قسمت دسته‌بند این سیستم تمرکز کرده و روشی نظام‌مند بر مبنای مدل‌های احتمالاتی ناپارامتری برای دسته‌بندی جویبار داده در حالت نظارتی ارائه می‌دهیم. این روش، مسئله تشخیص مفهوم را به صورت یک مسئله خوشه‌بندی پویا مدل کرده و با استفاده از مدل‌های مخلوط ناپارامتری، مدلی کارا برای این مسئله ارائه می‌دهد. در پایان نیز نتایج روش‌های پایه و روش پیشنهادی را گزارش و با یکدیگر مقایسه می‌کنیم.

واژه‌های کلیدی: دسته‌بندی جویبار داده، تغییر مفهوم، یادگیری فعال، محاسبات جمعی.

۱ مقدمه

پیشرفت‌های گسترده در حوزه سخت‌افزار و شبکه‌های ارتباطی در سال‌های اخیر، باعث فراگیر شدن حسگرهای اطلاعاتی شده است که حجم بسیار زیادی از اطلاعات را تولید می‌کنند و از طریق اینترنت با سرعت زیادی منتشر می‌کنند. هم‌چنین، فراهم شدن ابزارهای ذخیره‌سازی داده با قابلیت ذخیره‌سازی بالا و هزینه کم سبب شده است تا با ذخیره‌سازی داده‌های تولید شده در هر لحظه، مجموعه داده‌های بسیار عظیمی به وجود آیند که از آن‌ها به داده‌های حجیم[۱] یاد می‌کنند. داده‌های حجیم، دو ویژگی اصلی دارند. یکی حجم بالای آن‌ها و دیگری سرعت بالای تولید آن‌هاست. حجم بالای این داده‌ها سبب می‌شود که با استفاده از روش‌های قدیمی ذخیره‌سازی داده مانند پایگاه داده‌های رابطه‌ای[۲] نتوان آن‌ها را ذخیره‌سازی و مدیریت کرد. از طرفی، سرعت بالای تولید آن‌ها و نیاز به پردازش برخط داده‌ها در بسیاری از کاربردها مانند تشخیص تراکنش تقلبی در سیستم‌های بانکی، تشخیص نفوذ و حمله به سرویس‌دهندگان شبکه [؟]، تشخیص هرزنامه [؟]، تبلیغات برخط[؟] و شخصی‌سازی اطلاعات تولید شده بر روی رسانه‌های برخط [؟] سبب شده است که در بسیاری از موارد، این مجموعه داده‌های حجیم را به صورت یک جویبار داده در نظر بگیرند. در جویبار داده، فرض بر این است که در هر لحظه دسته‌ای از داده‌ها برای پردازش می‌آیند که باید با سرعت بالایی مورد پردازش قرار گیرند و نتیجه پردازش آن‌ها قبل از آمدن دسته بعدی تعیین گردد، و پس از آن نیز از دسترس خارج می‌شوند. این فرض از آن جهت به ما کمک می‌کند که می‌توان با استفاده از آن، مجموعه داده‌های خیلی بزرگ که امکان پردازش آن‌ها به صورت دسته‌ای وجود ندارد یا مجموعه داده‌هایی که با سرعت بالا در حال تولید هستند را به صورت برخط پردازش کرد.

یکی از پردازش‌هایی که برای استخراج اطلاعات بر روی جویبارهای داده انجام می‌پذیرد و در تمامی کاربردهای ذکر شده به کار می‌رود، دسته‌بندی[۳] جویبار داده است. در مسئله دسته‌بندی جویبار داده فرض می‌شود که هر کدام از داده‌های جریان ورودی، متعلق به یکی از کلاس‌های از پیش تعریف شده است که دسته‌بند باید آن را تشخیص دهد. مسئله دسته‌بندی جویبار داده ویژگی‌های خاصی دارد که آن را از سایر مسائل دسته‌بندی متمایز می‌کند. یکی از مهم‌ترین ویژگی‌های مسئله دسته‌بندی جویبار داده، ناایستان[۴] بودن محیط و تغییر مفهوم[۵] کلاس‌ها در طی زمان است که باعث ایجاد چالش‌های زیادی در حل این مسئله می‌شود.

در ادامه ابتدا چالش‌های مسئله دسته‌بندی جویبار داده را مورد بررسی قرار می‌دهیم. پس از آن، به هدف این پژوهش که ایجاد سیستمی یکپارچه برای حل کارای این چالش‌هاست می‌پردازیم. در بخش ۲، قسمت‌های مختلف سیستم ارائه شده را مورد بررسی قرار داده و روش‌های ارائه شده در هر کدام را با ذکر دلایل اصلی پیدایش و نقاط قوت و ضعف آن‌ها بر می‌شماریم. در بخش ۳، ایده پیشنهادی خود را مطرح نموده و در بخش ۴ به مقایسه الگوریتم‌های پایه و روش پیشنهادی پرداخته‌ایم. در پایان نیز ضمن بیان کارهای آتی در بخش ۵، مباحث شرح داده شده را در بخش ۶ جمع‌بندی می‌نماییم.

۱.۱ چالش‌های موجود در دسته‌بندی جویبار داده

به طور کلی، مسئله دسته‌بندی جویبار داده، یک مسئله دسته‌بندی برخط با توزیع داده‌ها و برچسب‌های ناایستان است. این دو ویژگی سبب ایجاد چالش‌های بسیاری در مسئله دسته‌بندی جویبار داده شده است. در ادامه، به تشریح هر یک از این ویژگی‌ها و چالش‌های به وجود آمده توسط آن‌ها می‌پردازیم.

به طور کلی، سناریوهای دسته‌بندی جویبار داده را می‌توان به دو دسته کلی برخط و برخط دسته‌ای[۶] تقسیم کرد. در سناریوی برخط، داده‌های جویبار به صورت تک‌تک وارد سیستم دسته‌بندی شده و برچسب آن‌ها تخمین زده می‌شود و پس از آن، برچسب واقعی داده در اختیار دسته‌بند قرار می‌گیرد و دسته‌بند نیز مدل خود را به روز می‌کند[؟]. در روش‌های برخط دسته‌ای، داده‌ها به صورت گروهی یا تک‌تک وارد سیستم شده و برچسب آن‌ها تخمین زده می‌شود، اما برچسب آن‌ها به صورت گروهی در اختیار دسته‌بند قرار می‌گیرد[؟]. به عنوان مثال، در تشخیص تراکنش‌های بانکی تقلبی، معمولا روزانه تراکنش‌ها توسط گروهی از افراد خبره بررسی شده و متقلبانه بودن آن‌ها را تشخیص داده و برچسب‌های واقعی را در اختیار دسته‌بند قرار می‌دهند. البته با توجه به حجم زیاد داده در جویبار داده و ممتد بودن ایجاد آن‌ها، معمولا امکان برچسب‌گذاری تمامی داده‌ها وجود ندارد و در هر دو روش برخط و برخط دسته‌ای، معمولا برچسب زیرمجموعه‌ای از داده‌ها به صورت تصادفی یا به انتخاب دسته‌بند در اختیار دسته‌بند قرار می‌گیرد که این موضوع سبب به وجود آمدن روش‌های نیمه‌نظارتی[۷] و فعال[۸] دسته‌بندی جویبار داده شده است که در فصل ۲ به آن‌ها می‌پردازیم. بنابراین، در تمامی سناریوهای دسته‌بندی

جویبار داده، دسته‌بند باید به طور هم‌زمان داده‌های جدید را برچسب‌گذاری کند و با استفاده از برچسب داده‌های پیشین، مدل خود را به روزرسانی کند. علاوه بر این، با توجه به حجم زیاد داده‌ها و ممتد بودن تولید آن‌ها، امکان ذخیره‌سازی تمام داده‌ها وجود نداشته و بنابراین، تنها حجم محدودی از داده‌ها را می‌توان نگه داشت. با توجه به این دو خاصیت، دسته‌بندی جویبار داده، مسئله‌ای برخط است و در نتیجه باید دسته‌بندی برای انجام این کار طراحی کنیم، که قابلیت به‌روزرسانی مدل به صورت برخط را داشته باشد.

در اکثر مسائل دسته‌بندی، فرض ایستایی محیط وجود دارد. به این معنی که فرض می‌کنیم که داده‌ها و برچسب‌های آن‌ها از یک توزیع ثابت تولید شده‌اند که ما از آن آگاه نیستیم و در نتیجه آن را تخمین می‌زنیم و هر چه داده‌های بیشتری به دست آوریم، می‌توانیم تخمین خود را از آن توزیع بهبود دهیم. در مسئله دسته‌بندی جویبار داده، فرض ایستایی محیط برقرار نیست. به این معنی که با گذر زمان، توزیعی که داده‌ها و یا برچسب‌های آن‌ها از آن پیروی می‌کنند تغییر می‌کند. به بیان ریاضی، فرض کنید که جویباری از داده‌ها به شکل

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_t, y_t), \ldots\} \quad (۱)$$

داریم که در آن، y_i برچسب متناظر با x_i است. بهترین تصمیم برای برچسب‌گذاری داده x_t، انتخاب کلاسی است که تابع احتمال پسین بیشتری داشته باشد[؟]:

$$\widehat{y_t} = \arg\max_c P(y_t = c|x_t) \quad (۲)$$

در محیط ایستان، فرض بر این است که این تابع پسین در طی زمان ثابت است، اما در جریان داده فرض بر این است که این تابع به ازای داده‌های مختلف می‌تواند متفاوت باشد. فرآیند تغییر تابع احتمال پسین در طی زمان را تغییر مفهوم می‌گویند. یک روش مدل کردن تغییر مفهوم این است که فرض کنیم که داده‌های جریان داده از تعدادی منبع تولید می‌شوند که داده‌های تولید شده توسط هر کدام از این منابع و برچسب‌هایشان از توزیع ثابتی پیروی می‌کنند. با توجه به نامحدود بودن تعداد حالات توابع تصمیم‌گیرنده، تعداد این منابع نیز می‌تواند نامحدود باشد.

یکی از چالش‌هایی که تغییر مفهوم ایجاد می‌کند، این است که برخلاف مسئله‌های دسته‌بندی ایستان، داشتن داده‌های برچسب‌دار بیشتر، لزوما دقت ما را افزایش نمی‌دهد. زیرا داده‌ها از منابع مختلفی تولید می‌شوند و توزیع داده‌های تولید شده از منابع مختلف با یکدیگر متفاوت است. بنابراین، بهترین جداساز داده‌های منابع مختلف با یکدیگر متفاوت است. و ممکن است داده‌های یک منبع نه تنها به یادگیری بهترین جداساز داده‌های منبع دیگر کمک نکند، بلکه سبب کاهش دقت آن نیز بشود. از این رو، در صورتی یک مدل دسته‌بند جویبار داده کاراست که اولا بتواند داده‌های هم‌مفهوم را به درستی از یکدیگر تمیز دهد، و ثانیا بتواند با استفاده از داده‌های هم‌مفهوم دسته‌بندهای برخط کارا ایجاد کند. علاوه بر این، با توجه به برخط بودن مسئله، دسته‌بندهای جویبار داده، باید این کار را به صورت برخط انجام دهند. در بخش ۲، به بررسی روش‌های مختلف حل این چالش‌ها می‌پردازیم.

تغییر مفهوم و برخط بودن مسئله، چالش‌های دیگری را نیز به مسئله دسته‌بندی اضافه می‌کنند. به عنوان مثال، به دست آوردن داده‌های برچسب‌دار در تمامی مسائل دسته‌بندی یکی از چالش‌های حل مسئله است. اما این چالش در مسئله دسته‌بندی جویبار داده به دلیل ناایستان بودن محیط و برخط بودن مسئله بسیار پررنگ‌تر می‌شود. دلیل این امر این است که با توجه به تغییر مفهوم در طی زمان، نیاز به داده‌های برچسب‌دار برای تشخیص مفهوم کلاس‌ها در هر لحظه وجود خواهد داشت و برخلاف مسائل ایستان که پس از مدتی دسته‌بند مورد نظر می‌تواند به دسته‌بند بهینه همگرا شود، در مسئله دسته‌بندی جویبار داده، همواره نیاز به گرفتن داده‌های برچسب‌دار وجود خواهد داشت. از طرفی، به دلیل زیاد بودن حجم داده‌ها و سرعت بالای تولید آن‌ها امکان برچسب‌گذاری درصد بسیار کمی از داده‌ها وجود دارد. به عنوان مثال، در مسئله تشخیص نفوذ به سرویس‌دهندگان اینترنتی، تعداد درخواست‌های واصله در هر لحظه بسیار زیاد است، به گونه‌ای که درصد بسیار کمی از آن را می‌توان توسط عوامل انسانی برچسب‌گذاری کرد. یکی دسته ار روش‌های ارائه شده برای این مشکل، استفاده از روش‌های فعال است که سعی می‌کنند با درخواست هوشمندانه برچسب داده‌هایی که اطلاعات بیشتری در مورد مفهوم فعلی و هم‌چنین دسته‌بند آن مفهوم در اختیار ما قرار می‌دهند، تعداد داده‌های برچسب‌دار مورد نیاز را کاهش دهند [؟]. این روش‌ها به طور جامع در فصل ۲ مورد بررسی قرار خواهند گرفت. یکی دیگر از روش‌های مواجهه با کمبود داده‌های برچسب‌دار در مسئله دسته‌بندی جویبار داده، استفاده از روش‌های جمع‌سپاری برای به دست آوردن برچسب داده‌ها به تعداد بیشتر و هزینه کمتر است. در ادامه به معرفی جمع‌سپاری می‌پردازیم و در ۳.۱، راهکاری بر مبنای جمع‌سپاری فعال برای حل مسئله دسته‌بندی جویبار داده ارائه می‌دهیم.

۲.۱ جمع‌سپاری

مسائل زیادی در زمینه هوش‌مصنوعی وجود دارند که به سادگی توسط انسان قابل حل هستند، اما پیچیده‌ترین برنامه‌های کامپیوتری فعلی نیز از حل کارای آن‌ها ناتوان هستند. به عنوان مثال، دسته‌بندی تصاویر یا تعیین مکان اشیا در یک تصویر، از جمله مسائلی است که با وجود پیشرفت‌های قابل توجهی که در زمینه بینایی ماشین[۹] به وجود آمده است، اما هم‌چنان دقت برنامه‌های کامپیوتری فعلی در حل این مسائل چندان راضی کننده نیست؛ این در حالی است که انجام چنین کاری برای هر انسانی به سادگی امکان‌پذیر است. مثال‌های بسیاری از این دست وجود دارد، از جمله: فیلتر کردن محتوا، تعیین میزان تناسب یک صفحه وب با یک عبارت جست‌وجو، تبدیل گفتار به متن و هم‌چنین بسیاری از مسائل دسته‌بندی. محاسبات انسانی[۱۰]، شیوه‌ای نوین برای حل چنین مسائلی است که با استفاده از جمع‌سپاری، از هوش محاسباتی و قدرت ادراک انسان‌ها در کنار قدرت و

سرعت محاسباتی کامپیوترها برای حل مسائل استفاده می‌کند. [؟]. برای حل مسئله از طریق جمع‌سپاری، ابتدا سوالات به ریزمسئله‌های ساده شکسته می‌شود، به گونه‌ای که هر فردی قابلیت انجام آن‌ها را داشته باشد. سپس از گروهی از افراد که لزوما متخصص نیستند استفاده می‌کنند تا این ریزمسائل را حل کنند. در این روش، به دلیل سادگی مسائل و متخصص نبودن کارمندان، به ازای انجام هر کدام از ریزمسائل، پول بسیار کمتری نسبت به کارمند متخصص به آن‌ها پرداخت می‌کنند و بنابراین، مسئله را با هزینه کمتر حل می‌کنند. MTurk[11]، یکی از بازارهای معروفی است که به شرکت‌ها و افراد این امکان را می‌دهد که مسائل خود را با جمع‌سپاری حل کنند.

پاسخ‌های به دست آمده از طریق جمع‌سپاری به دلایل مختلفی مانند پایین بودن دستمزد و متخصص نبودن کارمندان یا دلایلی مانند خستگی یا عجله برای انجام کار بیشتر در زمان کمتر، همواره با درصدی از اشتباه همراه است. یکی از اساسی‌ترین چالش‌های محاسبات انسانی، چگونگی استفاده از پاسخ‌های نویزی به دست آمده از جمع‌سپاری در سامانه‌های اطلاعاتی برای بهبود دقت آن‌هاست. یکی از راه‌حل‌هایی که برای غلبه بر این نوع خطاهای انسانی ارائه شده است، استفاده از سناریوی وارسی چندگانه[12] است. در این سناریو، هر سوال به چند پاسخ‌دهنده داده می‌شود و پاسخ نهایی با تجمیع پاسخ‌های دریافت شده استنتاج می‌شود. برای تجمیع پاسخ‌های بعضا متناقض کاربران، روش‌های مختلفی ارائه شده است که در بخش ۲ به بررسی آن‌ها می‌پردازیم. یکی از چالش‌های موجود در استفاده از سناریوی وارسی چندگانه برای حل مشکل خطای عوامل انسانی این است که برای افزایش دقت پاسخ‌های به دست آمده از تجمیع، هر سوال را باید از تعداد افراد بیشتری بپرسیم که این امر باعث بالا رفتن هزینه می‌شود. اصلی‌ترین راه حلی که برای این چالش ارائه شده است، انتخاب سوالات یا فرد پاسخ‌دهنده به هر سوال توسط خود سیستم و به صورت فعال انجام می‌شود [؟].

۳.۱ هدف پژوهش

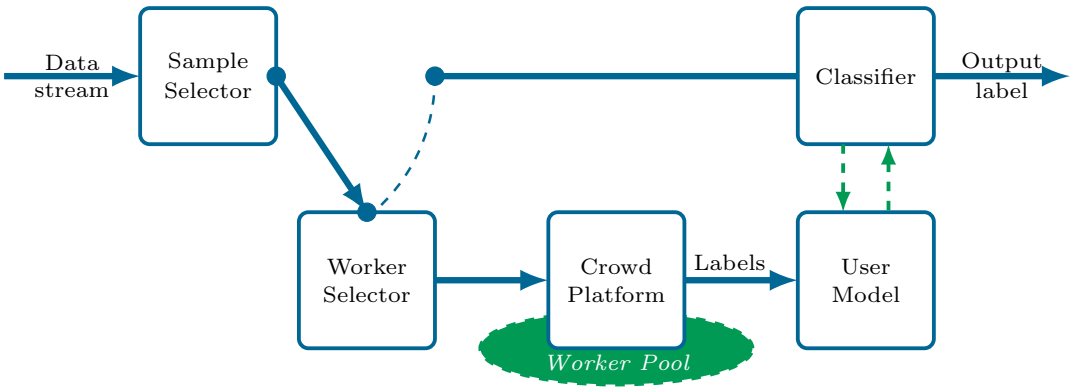
هدفی که در این پژوهش دنبال می‌شود، ارائه راهکاری مبتنی بر جمع‌سپاری فعال برای حل مسئله دسته‌بندی جویبار داده است. همان‌طور که در بخش ۱.۱ گفته شد، ناایستانی محیط در جویبارهای داده و تغییر مفهوم در طی زمان، سبب می‌شود که به طور پیوسته نیاز به گرفتن داده‌های برچسب‌دار وجود داشته باشد. از طرفی با توجه به سرعت بالای جویبارهای داده و زمان‌گیر بودن برچسب‌گذاری داده‌ها توسط عوامل انسانی، امکان برچسب‌گذاری تمامی داده‌ها وجود ندارد و تنها درصد کمی از داده‌ها را می‌توان برچسب‌گذاری کرد. علاوه بر این، برچسب‌زنی داده‌ها فرآیندی هزینه‌بر است که این امر باعث می‌شود برچسب زدن تعداد زیادی داده در طی زمان کاری بسیار پرهزینه باشد. ما برای حل این دو مشکل، راهکاری مبتنی بر جمع‌سپاری فعال ارائه می‌دهیم. شکل ۱.۱، ساختار کلی و اجزای تشکیل دهنده این سیستم را نشان می‌دهد. همان‌طور که در این شکل می‌بینید، این سیستم یک بخش دسته‌بند دارد که برچسب داده‌های ورودی را تخمین می‌زند. هر دسته از داده‌های جریان داده که وارد سیستم می‌شود، ابتدا یک واحد تصمیم‌گیرنده، برای هر کدام از داده‌های آن تصمیم می‌گیرد که آیا پرسیدن برچسب این داده از عوامل انسانی با توجه به هزینه‌ای که دارد به صرفه است یا خیر. این بخش در واقع تلاش می‌کند تا با کمترین تعداد سوالات از عوامل انسانی، به بالاترین دقت دست پیدا کند. به همین منظور، باید برچسب داده‌هایی را بپرسد که اطلاعات زیادی را برای دسته‌بندی سایر داده‌های دسته فعلی و داده‌های دسته‌های آینده در اختیار دسته‌بند قرار دهد. پس از این بخش، داده‌هایی که برای برچسب خوردن توسط کارمندان انتخاب می‌شود، به بخش انتخاب کارمند می‌رود. این بخش در تعامل با یک بستر جمع‌سپاری است. در این مدل، فرض می‌شود که در هر لحظه، مجموعه‌ای از کارمندان حاضر هستند و سیستم می‌تواند از بین آن‌ها تعدادی را انتخاب کند و آن‌ها را به کار گیرد و آن‌ها نیز جواب سیستم را در زمان بسیار کوتاهی می‌دهند. در واقع در این سیستم، فرض می‌شود که بستر جمع‌سپاری مورد استفاده، از مدل به‌کارگیری معرفی شده در [؟] استفاده می‌کند. بخش انتخاب کارمند، به ازای هر کدام از داده‌های رسیده، تصمیم می‌گیرد که با توجه به مجموعه کارمندان حاضر، هر سوال را به چه تعداد از پاسخ‌دهندگان بدهد به گونه‌ای که بتواند بیشترین میزان اطلاعات را با کمترین هزینه به دست آورد. پس از این بخش، سوالات تعیین شده از افراد مورد نظر پرسیده می‌شود و برچسب‌های نویزی به دست آمده به قسمت مدل کاربر فرستاده می‌شوند. مدل کاربر در واقع مدلی است که سعی می‌کند با مدل کردن پاسخ‌دهندگان و سوالات و در تعامل با مدل کلاسه‌بند، از برچسب‌های به دست آمده، بیشترین میزان اطلاعات را به دست آورد. بنابراین، با به دست آمدن برچسب‌های جدید، مدل کاربران و مدل دسته‌بند به روز می‌شود.

طراحی چنین سامانه‌ای، چالش‌های زیادی دارد که در ادامه به بررسی آن‌ها پرداخته و چالش‌های هدف این پژوهش را مشخص می‌کنیم.

۴.۱ چالش‌های هدف

همان‌طور که در بخش قبل بیان شد، یک سامانه دسته‌بند جویبار داده مبتنی بر جمع‌سپاری، از چهار قسمت اصلیِ انتخاب داده، انتخاب کاربر، مدل کاربر و مدل دسته‌بند است. وظیفه مدل کاربر، مدل کردن رفتار پاسخ‌دهندگان و استخراج بیشترین میزان اطلاعات از بین پاسخ‌های نویزی آن‌ها برای یافتن مفهوم هر داده و تعیین تابع جداساز هر یک از مفاهیم توسط دسته‌بند است. قسمت انتخاب داده، زیرمجموعه‌ای از داده‌ها را برای برچسب‌گذاری توسط که بیشترین میزان اطلاعات را درباره توزیع تمامی داده‌ها داشته باشد. علاوه بر این، اساسی‌ترین چالش، طراحی مدل دسته‌بندی برخط و وفقی است که بتواند با این دو قسمت تعامل داشته باشد.

در بخش ۲، بررسی جامعی از کارهایی که در هر یک از این زمینه‌ها انجام شده است ارائه می‌دهیم. با توجه به گستردگی بسیار زیاد مسئله و عدم وجود یک راه‌حل نظام‌مند و غیرمکاشفه‌ای برای دسته‌بندی وفقی و برخط جویبار داده، تمرکز کار خود را معطوف به طراحی مدلی برخط و وفقی برای دسته‌بندی جویبار داده می‌کنیم. برای این منظور، در بخش ۳، الگوریتمی کارا و انعطاف‌پذیر[13]، مقاوم[14] نسبت به بیش‌برازش[15] و بر مبنای مدل‌های غیرپارامتری بیزی[16] ارائه می‌کنیم.



شکل ۱.۱: ساختار کلی یک سامانه دسته‌بند جویبار داده مبتنی بر جمع‌سپاری فعال

۵.۱ معیارهای ارزیابی

هدف اصلی در مسئله دسته‌بندی جویبار داده، مانند هر مسئله دسته‌بندی دیگری، یافتن مدلی با قدرت تعمیم[۱۷] زیاد است. معیارهای مختلفی برای سنجش قدرت تعمیم مدلهای دسته‌بندی ارائه شده است که ساده‌ترین آن‌ها، معیار درستی[۱۸] است که برابر نسبت تعداد برچسب‌های درست تخمین زده شده به تعداد کل داده‌هاست. این معیار در صورتی می‌تواند تخمین زننده قدرت تعمیم دسته‌بند باشد که اولا داده‌ها متوازن [۱۹] باشند و ثانیا ارزش تشخیص صحیح برچسب‌های هر کلاس با یکدیگر برابر باشد. منظور از متوازن بودن داده‌ها، یکسان بودن تعداد داده‌های کلاس‌های مختلف است. در بسیاری از کاربردهای دسته‌بندی جویبار داده، از جمله تشخیص تقلب در تراکنش‌های بانکی، یا تعیین تبلیغاتی که یک کاربر بر روی آن‌ها کلیک خواهد کرد در تبلیغات برخط، داده‌ها کاملا نامتوازن هستند و تعداد داده‌های کلاس مثبت بسیار کمتر از کلاس دیگر است. از این رو، در دسته‌بندی جویبار داده نیازمند معیارهایی هستیم که فرض نامتوازن بودن داده‌ها را در نظر بگیرند. برای این منظور، معیار F_1 ارائه شده است. این معیار، در واقع میانگین توافقی[۲۰] دقت[۲۱] و بازخوانی[۲۲] است و از این رو می‌توان آن را به صورت میانگینی وزن‌دار از دقت و بازخوانی در نظر گرفت که به کمینه این دو، وزن بیشتری اختصاص می‌دهد.

۲ کارهای پیشین

چنانچه در بخش ۳.۱ بیان شد، هدف از این پژوهش، ارائه راهکاری مبتنی بر جمع‌سپاری فعال برای دسته‌بندی جویبار داده است. از این رو، پژوهش‌های پیشین را در سه بخش روش‌های دسته‌بندی جویبار داده، روش‌های انتخاب فعال داده در دسته‌بندی و روش‌های دسته‌بندی با استفاده از جمع‌سپاری تشریح می‌کنیم و ضمن بیان برتری‌ها و کاستی‌های هر یک، به تشریح پژوهش‌های نزدیک‌تر به اهداف پژوهش حاضر می‌پردازیم.

۱.۲ دسته‌بندی جویبار داده

به طور کلی، دسته‌بندها را می‌توان بر حسب روش مدیریت تغییر مفهوم به سه دسته انتخاب داده، وزن‌دهی به داده و روش‌های گروهی تقسیم کرد. روش‌های مبتنی بر انتخاب داده و وزن‌دهی به داده، غالبا روش‌های تک‌مدله هستند. در روش‌های تک مدله، از ابتدا، تنها یک فرضیه[۲۳] را از فضای فرضیه[۲۴] انتخاب می‌کنند و در هر مرحله با آمدن داده‌های جدید، آن مدل را تغییر و به‌روزرسانی می‌کنند. روش‌های تک‌مدله در واقع بر مبنای این فرض طراحی شده‌اند که مفاهیم به طور پیوسته در حال تغییر هستند و مفاهیم تکرار شونده در جویبار داده وجود ندارد و تغییر مفهوم در فضای مفاهیم به صورت نرم[۲۵] صورت می‌پذیرد.

برخلاف روش‌های تک‌مدله، روش‌های گروهی در هر لحظه مجموعه‌ای از فرضیات را نگه می‌دارند و به ازای هر داده جدید، از نظرات تمامی آن‌ها استفاده می‌کنند و نتیجه نهایی را از تجمیع نظرات تمامی این دسته‌بندها استخراج می‌کنند. این روش‌ها، دقت بالاتری در تشخیص تغییر مفهوم دارند و به راحتی می‌توانند با وزن‌دهی به دسته‌بندهای پایه، این تغییر مفهوم را مدیریت کنند. علاوه بر این، این روش‌ها می‌توانند به گونه‌ای طراحی شوند که قابلیت استفاده از تکرارشوندگی مفاهیم را داشته باشند. از این رو، اکثر روش‌های ارائه شده برای دسته‌بندی جویبار داده، گروهی هستند و روش‌های تک‌کلاسه‌بند، به صورت محدود ارائه شده‌اند. روش‌های گروهی، بر حسب نحوه ایجاد دسته‌بندهای جدید و به‌روزرسانی آن‌ها و هم‌چنین حذف دسته‌بندهای ناکارآمد و هم‌چنین نحوه تجمیع نظرات دسته‌بندها، به دسته‌های مختلفی تقسیم می‌شوند که در ادامه به آن‌ها می‌پردازیم.

۱.۱.۲ روش‌های تک‌کلاسه‌بند

همان‌طور که در فصل اول بیان شد، یکی از چالش‌هایی که تغییر مفهوم به وجود می‌آورد این است که توزیع داده‌هایی که از مفاهیم مختلف پیروی می‌کنند، با یکدیگر متفاوت است و استفاده از داده‌های مفاهیم دیگر ممکن است نه تنها سبب افزایش دقت دسته‌بند نشود، بلکه دقت آن را کاهش نیز بدهد. روش‌های گروهی، با استفاده از چند کلاسه‌بند مختلف، سعی می‌کنند داده‌های مربوط به مفاهیم مختلف را در دسته‌بندهای متفاوتی قرار دهند تا از این طریق، از یک طرف تا حد امکان از اطلاعات موجود در داده‌های پیشین استفاده کنند و از طرف دیگر از تداخل داده‌های مربوط به مفاهیم مختلف جلوگیری کنند؛ اما در روش‌های تک‌کلاسه‌بند، تنها یک کلاسه‌بند برای تخمین برچسب تمامی داده‌های یک جویبار داده استفاده می‌شود. بنابراین، این روش‌ها برای مدیریت تغییر مفهوم نیاز به مکانیزمی برای فراموشی دارند تا از این طریق، داده‌های غیرمرتبط به داده‌های فعلی را از یاد ببرند.

دو روش کلی برای این کار وجود دارد. در دسته اول که به روش‌های انتخاب داده معروفند، در هر لحظه، سعی می‌شود که با استفاده از یک پنجره متحرک، مجموعه‌ای از آخرین داده‌ها را که به مفهوم فعلی مرتبط است، نگه‌داری کنند و یک دسته‌بند را با استفاده از آن‌ها آموزش داده و برچسب داده فعلی را با استفاده از آن دسته‌بند تعیین کنند [؟]. روش‌های مبتنی بر وزن‌دهی نمونه، دسته دیگری از روش‌ها هستند که در آن‌ها، با دادن وزن بیشتر به داده‌های اخیر، به صورت تدریجی باعث فراموشی داده‌های قدیمی می‌شوند [؟، ؟]. برای استفاده از این روش، دسته‌بند پایه مورد استفاده باید قابل به‌روزرسانی باشد و از یادگیری وزن‌دار پشتیبانی کند. به عنوان مثال، دسته‌بند نایو بیز[۲۶]، از این دو ویژگی پشتیبانی می‌کند. روش دیگری برای این کار، روش ارائه شده در [؟] است. در این روش، مدلی احتمالاتی برای دسته‌بند در نظر گرفته شده است که در آن، فرض می‌شود که دسته‌بند مورد نظر یک دسته‌بند خطی با پارامتر w است که از توزیع نرمال پیروی می‌کند:

$$w \sim N(w; \mu_t, \Sigma_t) \tag{۳}$$

با آمدن داده جدید، طبق رابطه بیز، توزیع احتمال w به صورت زیر به‌روز می‌شود:

$$P(w_{t+1}|x_t, y_t) \propto P(y_t|x_t, w)N(w; \mu_t, \Sigma_t) \tag{۴}$$

برای وزن‌دهی بیشتر به داده‌های اخیر و کم کردن تاثیر داده‌های گذشته، تاثیر تابع درستنمایی داده‌های پیشین در توزیع احتمال پسین w با استفاه از یک توان کمتر از یک، کم می‌شود:

$$P(w_{t+1}|x_t, y_t) \propto P(y_t|x_t, w)N(w; \mu_t, \Sigma_t)^{\gamma} \quad 0 \ll \gamma < 1 \tag{۵}$$

با این روش، هر چه داده‌ها قدیمی‌تر باشند تاثیر آن‌ها در توزیع احتمال w کمتر خواهد بود. مزیت روش‌های وزن‌دهی به داده‌ها نسبت به روش‌های انتخاب داده این است که تاثیر یک داده را به صورت ناگهانی صفر نمی‌کنند. از این رو، کارآیی آن‌ها در محیط‌هایی که تغییر مفهوم به صورت تدریجی رخ می‌دهد بهتر است. از طرفی، روش‌های مبتنی بر پنجره، راحت‌تر می‌توانند تغییر مفهوم‌های ناگهانی را مدیریت کنند.

۲.۱.۲ روش‌های گروهی

همان‌طور که در بخش گذشته بیان شد، در صورتی که بتوانیم داده‌های تولید شده توسط مفاهیم مختلف را در طی زمان از یکدیگر تفکیک کنیم، می‌توانیم از داده‌های هر مفهوم استفاده کرده و دسته‌بندی مناسب برای آن‌ها بسازیم. روش‌های گروهی نیز بر همین مبنا ساخته شده‌اند. به این معنی که در هر لحظه مجموعه‌ای پویا از دسته‌بندها را نگه‌داری می‌کنند و با مشاهده کاهش کارآیی، این مجموعه را با به‌روزرسانی دسته‌بندهای پایه یا حذف دسته‌بندهای ناکارآمد و اضافه کردن دسته‌بندهای جدید به‌روزرسانی می‌کنند. دسته‌بندی جویبار داده با استفاده از دسته‌بندهای گروهی، شامل دو گام اصلی است. یکی تجمیع نظرات دسته‌بندها برای تخمین برچسب یک داده تازه وارد شده و دیگری، به‌روزرسانی دسته‌بند که در ادامه، به روش‌های مختلف انجام این دو کار می‌پردازیم.

به طور کلی، روش‌های گروهی را به دو دسته روش‌های مبتنی بر ترکیب مدل[۲۷] و روش‌های مبتنی بر انتخاب مدل[۲۸] می‌توان تقسیم کرد[؟]. مبنای روش‌های مبتنی بر ترکیب مدل این است که فرض می‌کنند که هر داده، از ترکیب خطی از مدل‌های پایه ساخته شده‌اند و به این ترتیب، فضای فرضیه را غنی می‌کنند[؟]. روش‌های مختلفی بر مبنای ترکیب مدل در زمینه دسته‌بندی جویبار داده به وجود آمده است که از آن جمله می‌توان به [؟، ؟] اشاره کرد. این روش‌ها در هر لحظه، مجموعه‌ای از دسته‌بندهای پایه را در اختیار دارند و به ازای هر کدام از آنها، یک وزن در نظر گرفته‌اند و با استفاده از رابطه زیر، نظرات دسته‌بندهای پایه را تجمیع کرده و برچسب هر داده را تخمین می‌زنند:

$$\hat{y}_i^t = \arg\max_c \sum_k W_k^t I_{[h_k(x_i^t)=c]} \tag{۶}$$

که در آن، W_k^t وزن دسته‌بند kام در لحظه t است. این مدل‌ها که عمدتا بر مبنای روش‌هایی همچون بگینگ[۲۹] و بوستینگ[۳۰] هستند، پس از مشاهده برچسب واقعی داده‌ها، در صورتی که کاهشی در کارآیی مشاهده کنند، با تغییر وزن دسته‌بندها یا به‌روزرسانی دسته‌بندهای پایه، کارآیی مدل را افزایش می‌دهند اما مشکلی که در تمامی این روش‌ها وجود دارد این است که روابطی که بر مبنای آن وزن‌ها را به‌روزرسانی می‌کنند معمولا مکاشفه‌ای هستند و از این رو احتمال بیش‌برازش وجود دارد.

در روش‌های مبتنی بر انتخاب مدل، فرض بر این است که هر داده، توسط یکی از مدل‌های پایه تولید شده است [؟، ؟]. با توجه به این فرض، برای دسته‌بندی یک داده، کافی است که دسته‌بند متناظر با مفهوم آن داده را یافته و آن داده را دسته‌بندی کنیم. مشکلی که برای این کار وجود دارد، این است که همواره در مورد مفهوم یک داده عدم‌قطعیت وجود دارد و هم‌چنین یافتن مفهوم یک داده، مسئله‌ای بدون‌نظارت[۳۱] است. روش‌های مختلفی برای حل این روش ارائه شده است. فرض ساده کننده‌ای که تقریبا در تمامی روش‌های موجود به کار می‌برند این است که داده‌ها را به صورت دسته‌هایی از داده‌های متوالی در نظر می‌گیرند و فرض می‌کنند که مفهوم تمامی داده‌های یک دسته یکسان است و از این فرض استفاده کرده و به تخمین مفهوم این داده‌ها می‌پردازند. به عنوان مثال، در روش ارائه شده در [؟]، به ازای هر دسته از داده‌ها، یک بردار از ویژگی‌های دسته به نام بردار مفهومی به دست می‌آورد و با خوشه‌بندی[۳۲] این بردارهای مفهومی، دسته‌های با مفهوم یکسان را در یک خوشه قرار می‌دهند و به ازای هر خوشه، دسته‌بندی با داده‌های دسته‌های آن تولید می‌کند. برای آن که داده‌های هم‌مفهوم در یک خوشه قرار بگیرند، ویژگی‌های زیر را از دسته‌های داده استخراج می‌کند:

$$z_i = \begin{cases} \{p(f_i = v|c = j) : j = 1 : m, v \in V_i\}, & \text{if} f_i \text{is nominal} \\ \{\mu_{i,j}, \sigma_{i,j} : j = 1 : m\}, & \text{if} f_i \text{is numeric} \end{cases} \quad (۷)$$

در عبارت۷، V_i مجموعه مقادیری است که ویژگی iام یک داده می‌تواند به خود بگیرد و m تعداد کلاس‌های مختلف است. در صورتی که از فرض نایو بیز استفاده کنیم، به ازای یک دسته از داده‌ها خواهیم داشت:

$$p(f_i = v|c = j) = \frac{n_{v,j}}{n_j} \quad (۸)$$

که در آن، $n_{v,j}$ تعداد داده‌های عضو دسته است که ویژگی iام آن‌ها مقدار v دارد و عضو کلاس j نیز هستند و n_j نیز تعداد کل داده‌های عضو کلاس jام را نشان می‌دهد. بردار Z از این نظر می‌تواند نشان دهنده مفهوم یک دسته باشد که در صورتی که فرض نایو بیز برقرار باشد و ویژگی‌های پیوسته در هر کلاس از یک توزیع نرمال پیروی کنند، این بردار در واقع یک آماره کافی کمینه[۳۳] برای یادگیری دسته‌بند است. در این روش، با آمدن برچسب هر داده، ابتدا بردار مفهومی را ایجاد کرده و با خوشه‌بندی آن، مفهوم مربوط به این داده‌ها را تعیین می‌کند و دسته‌بند مربوط به آن خوشه را با این داده‌ها به‌روزرسانی می‌کند. در صورتی که بردار مفهومی این دسته جزو هیچکدام از خوشه‌های پیشین قرار نگیرد، یک دسته‌بند جدید به مجموعه دسته‌بندهای پایه اضافه می‌شود و با داده‌های این دسته آموزش داده می‌شود. روش ارائه شده در [؟]، از نوعی فرض همواری مفهوم در طی زمان استفاده شده است و بنابراین، برای دسته‌بندی داده‌های بدون برچسب، از دسته‌بندی که توسط داده‌های دسته قبل به‌روزرسانی شده است استفاده می‌کند. با وجود این که این روش، یکی از روش‌های مرجع برای مدیریت تغییر مفهوم در محیط‌هایی است که مفاهیم تکرارشونده وجود دارد، اما فرض‌های ساده‌کننده‌ی بسیاری استفاده کرده است که سبب کاهش کارآیی آن می‌شود. به عنوان مثال، این روش فرض می‌کند که مفهوم تمامی داده‌های یک دسته یکسان است و قابل بیان توسط بردار مفهومی تعریف شده است. هم‌چنین در این روش، مکانیزمی برای حذف یک دسته‌بند وجود ندارد و بنابراین، تعداد دسته‌بندها می‌تواند بیش از حافظه موجود شود و ایجاد مشکل کند.

با توجه به این که مفهوم داده‌ها متغیری است که در مورد آن نایقینی داریم، بنابراین، یکی از روش‌های مدل کردن این نایقینی استفاده از مدل‌های احتمالاتی است [؟]. این مدل‌ها، سعی می‌کنند با مدل کردن مفهوم به صورت یک متغیر پنهان[۳۴] و بیان رابطه آن با سایر متغیرهای مسئله مانند بردار ویژگی داده و برچسب آن، مدلی احتمالاتی برای مسئله تهیه کنند و با استنتاج بر روی این مدل یا استفاده از روش‌های تخمین مانند بیشینه‌سازی احتمال پسین[۳۵] یا بیشینه‌سازی درستنمایی[۳۶]، برچسب داده‌ها را تخمین بزنند. با این دید، در صورتی که کلیه داده‌هایی که تا قبل از لحظه t وارد سیستم شده‌اند را X نامیده و برچسب‌های متناظر آن‌ها را Y در نظر گرفته و دسته داده‌های رسیده در لحظه t را با X^t و iامین عضو از این دسته را با x_i^t و برچسب متناظر آن را با y_i^t نشان دهیم، ما به دنبال یافتن $P(y_i^t|x_i^t, X^t, X, Y)$ هستیم که از رابطه زیر به دست می‌آید:

$$P(y_i^t|x_i^t, X^t, X, Y) = \sum_k P(c^t = k|X^t, X, Y)P(y_i^t|x_i^t, c = k) \quad (۹)$$

در عبارت بالا، تنها فرضی که استفاده شده است، این است که در صورت دانستن مفهوم فعلی و مدل متناظر آن، داده فعلی از سایر داده‌ها مستقل می‌شود، که با توجه به این که فرض کرده‌ایم که هر داده توسط یکی از منابع تولید می‌شود کاملا درست است. عبارت ۹ در واقع نمونه‌ای از عبارت ۶ است که در آن، وزن مربوط به دسته‌بند kام، احتمال این است که مفهوم فعلی همان مفهوم دسته‌بند kام باشد. همان‌طور که دیده می‌شود، این احتمال علاوه بر داده‌های برچسب‌دار دسته‌های پیشین، می‌تواند به داده‌های بدون برچسب دسته فعلی نیز وابسته باشد. در صورتی وابسته کردن این احتمال به X^t می‌تواند مفید باشد که دسته‌بندهای پایه ساده باشند و هر کدام از آن‌ها تنها در یک ناحیه از فضای ویژگی خوب عمل کنند. با وجود این که در بسیاری از روش‌های ارائه شده، از دسته‌بندهای پایه ساده مانند دسته‌بند نایو بیز استفاده می‌کنند، اما از X^t برای تخمین مفهوم مورد نظر استفاده نمی‌کنند [؟، ؟]. در روش‌های مبتنی بر انتخاب مدل، معمولا به جای دخیل کردن عدم‌قطعیتی که در مورد مفهوم داده‌ها وجود دارد و تخمین برچسب

داده به صورت جمع وزن‌داری از نظرات دسته‌بندهای پایه، تنها نظر یکی از دسته‌بندها که بیشترین احتمال را دارد در نظر می‌گیرند [؟، ؟] که می‌تواند سبب بیش‌برازش و کاهش دقت دسته‌بند شود.

همان‌طور که گفته شد، فرض اصلی در این روش‌ها این است که برچسب‌های درست داده‌ها پس از این که توسط سیستم تخمین زده شدند، در اختیار دسته‌بند قرار می‌گیرد. هر چند این سناریو در برخی از کاربردها، مانند تخمین احتمال کلیک کردن یک تبلیغ از سوی کاربر در تبلیغات برخط، درست است اما در برخی از کاربردهای دسته‌بندی جویبار داده، چنین فرضی صحیح نیست. به عنوان مثال، در تشخیص تراکنش‌های متقلبانه در سیستم بانکی، تعداد بسیار زیادی تراکنش در هر لحظه انجام می‌شود و متقلبانه بودن یا نبودن آن تنها توسط کارشناس تعیین می‌شود و در نتیجه با توجه به حجم زیاد تراکنش‌ها، با وجود تعداد زیادی کارشناس نیز، باز هم نمی‌توان به درصد زیادی از داده‌ها برچسب زد. بنابراین، نیازمند روشی هستیم که داده‌هایی را برای برچسب‌زنی انتخاب کند که بیشترین میزان اطلاعات را در بر داشته باشد.

۲.۲ روش‌های فعال انتخاب داده

همان‌طور که در فصل ۱ بیان شد، یکی از اصلی‌ترین راه‌های کاهش هزینه دسته‌بندی، انتخاب هوشمندانه داده‌هایی است که توسط عوامل انسانی برچسب‌دهی می‌شوند. معیارها و روش‌های مختلفی برای انتخاب بهترین داده ارائه شده است. در ادامه، به بررسی رویکردهای اصلی انتخاب فعال در مسئله دسته‌بندی پرداخته و در هر قسمت، روش‌هایی که با آن رویکرد برای دسته‌بندی جویبار داده ارائه شده‌اند را مورد بررسی بیشتر قرار می‌دهیم.

الف) **نمونه‌برداری بر مبنای نایقینی:** در روش‌های نمونه‌برداری بر مبنای نایقینی، معیارهایی برای عدم‌قطعیت در مورد برچسب یک داده تعریف شده و برچسب داده‌ای درخواست می‌شود که بیشترین میزان نایقینی را در مورد برچسب آن داشته باشیم. در صورتی که پارامترهای مربوط به دسته‌بند را θ در نظر بگیریم، یکی از روش‌های نمونه‌برداری بر اساس نایقینی این است که داده‌ای که کم‌ترین میزان اطمینان را به برچسب تخمینی آن داریم را بپرسیم. به عبارت دیگر:

$$\begin{aligned} x^*_{LC} &= \arg\min_x P(\hat{y}|x) \\ where \quad \hat{y} &= \arg\max_y P(y|x,\theta) \end{aligned} \tag{۱۰}$$

روش‌های دیگر انتخاب بر مبنای نایقینی، انتخاب بر اساس حاشیه[۳۷] و آنتروپی[۳۸] هستند که به ترتیب به صورت زیر تعریف می‌شوند:

$$x^*_M = \arg\min_x \left[P(\hat{y}_1|x,\theta) - P(\hat{y}_2|x,\theta)\right] \tag{۱۱}$$

$$x^*_M = \arg\max_x H(y|x,\theta) = \arg\max_x -\sum_y P(y|x,\theta)\log P(y|x,\theta) \tag{۱۲}$$

در رابطه ۱۱، $\hat{y}_1$ و $\hat{y}_2$ به ترتیب اولین و دومین محتمل‌ترین برچسب‌ها برای داده x است. دو روش اول، تنها اطلاعات در مورد محتمل‌ترین برچسب‌ها را در نظر می‌گیرند. این در حالی است که در معیار آنتروپی، میزان اطلاعات در مورد تمامی برچسب‌ها در نظر گرفته می‌شوند. بنابراین، در مسئله دسته‌بندی که تنها محتمل‌ترین برچسب برای ما مهم است، استفاده از دو معیار اول مناسب‌تر است و معیار آنتروپی در شرایطی مناسب است که تابع هدف ما، از جنس لگاریتمی باشد[؟].

ب) **نمونه‌برداری بر مبنای جست‌وجوی فضای فرضیه:** روش‌های نمونه‌برداری مبتنی بر جست‌وجوی فضای حالت، به دنبال یافتن نمونه‌هایی هستند که به دست آوردن برچسب آن‌ها، بیشترین میزان اطلاعات را در مورد فضای دسته‌بندهای سازگار[۳۹](VS) به ما بدهد. VS، مجموعه تمام دسته‌بندهایی است که داده‌های برچسب‌دار آموزشی را به درستی دسته‌بندی کند. یکی از روش‌های نمونه‌برداری مبتنی بر جست‌وجوی فضای حالت، درخواست بر مبنای عدم توافق[۴۰](QBD) است. در این روش، در صورتی که به ازای یک داده حداقل دو دسته‌بند عضو VS فعلی وجود داشته باشند که برچسب‌های متفاوتی به آن اختصاص بدهند، آن سوال از کاربر پرسیده می‌شود. این روش، دو مشکل اصلی دارد: اول این که این روش هیچ تفاوتی بین داده‌هایی که می‌توانند پرسیده شوند قائل نمی‌شود و دوم نیز اینکه نگه داشتن VS در بسیاری از موارد کار بسیار مشکلی است. از این رو، روش‌های درخواست بر اساس کمیته[۴۱](QBC) ارائه شده‌اند. روش‌های QBC، روش‌هایی مبتنی بر QBD هستند که به جای نگه داشتن کل VS، تعدادی دسته‌بند را به عنوان اعضای کمیته در نظر گرفته و بر اساس میزان عدم توافق اعضای این گروه از دسته‌بندها بر روی برچسب یک داده خاص، پرسیدن برچسب آن داده را ارزش‌گذاری می‌کنند. دو ویژگی روش‌های بر مبنای QBD را از یکدیگر متمایز می‌کند. یکی نحوه انتخاب فرضیه‌های عضو کمیته و دیگری نحوه سنجش میزان عدم توافق آن‌ها بر روی یک نمونه. روش‌های مختلفی برای ساخت یک گروه از دسته‌بندها ارائه شده است، مانند بگینگ یا بوستینگ. روش دیگر ساخت گروه دسته‌بندها که مبنای احتمالاتی دارد، این است که تابع احتمال پسین پارامتر دسته‌بند را به دست آوریم و با نمونه‌برداری از آن، گروه دسته‌بندها را مشخص کنیم. معیارهای مختلفی نیز برای سنجش عدم توافق یک گروه از دسته‌بندها بر روی یک داده ارائه شده است. از آن جمله می‌توان به آنتروپی آرا[۴۲] و فاصله KL[۴۳] اشاره کرد.

روش انتخاب نمونه بر اساس معیار فاصله KL، نمونه‌هایی را انتخاب می‌کند که دسته‌بندهای کمیته نسبت به آن عدم قطعیت بالایی داشته و علاوه بر این، نسبت به آن توافق نظر نیز نداشته باشند[؟]. روش‌های مبتنی بر جست‌وجوی فضای فرضیه از این جهت می‌توانند در مسائل انتخاب نمونه جهت دسته‌بندی جویبار داده مفید باشد که در جویبار داده، به دلیل تغییر دائمی مفهوم، در مورد فضای فرضیه عدم قطعیت بسیار بالایی وجود داشته و از آن جا که هدف روش‌های انتخاب نمونه، محدود کردن VS است، بنابراین، استفاده از این روش‌ها می‌تواند در یافتن داده‌هایی که داشتن برچسب آن‌ها به ما در تشخیص مفهوم کمک زیادی می‌کند، مفید باشد.

ج) **نمونه‌برداری بر مبنای میانگین کاهش خطا:** همان‌طور که گفته شد، هدف روش‌های استقرایی در مسائل دسته‌بندی، یافتن مدلی برای دسته‌بندی است که قدرت تعمیم بالایی داشته باشند. روش‌هایی که تا به این جا بررسی شد، رسیدن به این هدف را به صورت غیرمستقیم دنبال می‌کردند. در روش نمونه‌برداری بر مبنای میانگین کاهش خطا، سعی می‌شود برچسب داده‌ای تقاضا شود که میانگین خطای دسته‌بندی روی تمامی داده‌ها را بیش از سایرین کاهش دهد. برای این منظور، نیاز است که توزیع احتمال داده‌ها را بدانیم تا با استفاده از آن، میانگین خطای دسته‌بند روی کل داده‌ها را به دست آوریم. برای حل این مشکل، یک مجموعه نسبتا بزرگ از داده‌های بدون برچسب را در نظر گرفته و فرض می‌کنند داده‌های آن مجموعه، به خوبی تابع احتمال $P(x)$ را مدل می‌کنند. از طرف دیگر، تا زمانی که برچسب یک داده را نداشته باشیم، نمی‌توانیم در مورد میزان کاهش میانگین خطای دسته‌بند پس از دیدن آن برچسب صحبت کنیم. برای حل این مشکل، از خطای دسته‌بند، بر حسب توزیعی که برای برچسب آن داده تا به حال به دست آورده‌ایم، میانگین می‌گیریم. این روش از آن جهت که به صورت صریح، هدف را کم کردن خطای دسته‌بندی روی کل داده‌ها قرار می‌دهد، دقت بهتری نسبت به روش‌های پیشین دارد، اما چون به ازای هر بار انتخاب یک داده باید یک بار دسته‌بند را به ازای تمامی داده‌های کاندید و تمامی برچسب‌های ممکن آموزش دهد و میزان خطای آن را بر روی کل فضا به دست آورد، از پیچیدگی محاسباتی بالایی برخوردار بوده و از این جهت کاربرد آن محدود شده است.

۳.۲ دسته‌بندی با استفاده از جمع‌سپاری

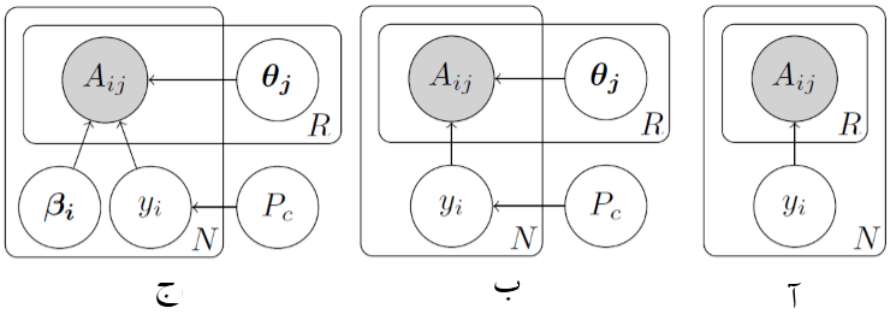
همان‌طور که در بخش ۲.۱ بیان شد، جمع‌سپاری یکی از روش‌های جدید حل مسئله است که با استفاده از آن، می‌توان با به‌کارگیری هوشمندانه خردجمعی انسان‌ها، کارهایی را که الگوریتم‌های فعلی هوش مصنوعی از حل آن‌ها عاجز هستند، در زمان کوتاه و هزینه کم به انجام رساند. یک دسته اصلی از مسائلی که توسط این روش حل می‌شود، مسائل دسته‌بندی است. دلیل کاربرد فراوان جمع‌سپاری در حل مسائل دسته‌بندی، سادگی آن‌ها برای انسان و هم‌چنین وجود روش‌های کارا برای سنجش کیفیت پاسخ‌های به دست آمده است. به همین دلیل، برای حل بسیاری از مسائل توسط جمع‌سپاری، آن‌ها را به مسائل دسته‌بندی تبدیل می‌کنند.

دو روش کلی برای حل مسئله دسته‌بندی توسط جمع‌سپاری وجود دارد. دسته اول روش‌ها که مبتنی بر تجمیع نظرات هستند، به ازای هر کدام از داده‌ها، تعدادی برچسب از عوامل انسانی غیرمتخصص جمع‌آوری می‌کنند و پس از آن، با استفاده از روش‌هایی که در ۱.۳.۲ به تشریح آن‌ها خواهیم پرداخت، برچسب‌های واقعی را تخمین می‌زنند. دسته دیگری از روش‌ها که به آن‌ها روش‌های استقرایی[۴۴] گفته می‌شود، برای تعداد کمی از داده‌ها برچسب گرفته شده و با استفاده از این داده‌های برچسب‌دار، یک دسته‌بند آموزش داده می‌شود. پس از آن، برچسب داده‌های بدون برچسب، توسط آن دسته‌بند مشخص می‌شود. در ادامه، به بررسی این دو روش می‌پردازیم.

۱.۳.۲ مدل‌های مبتنی بر تجمیع نظرات

همان‌طور که گفته شد، در روش‌های مبتنی بر تجمیع نظرات، به ازای هر داده تعدادی برچسب وجود دارد که ممکن است با یکدیگر متفاوت باشند و باید با تجمیع آن‌ها، برچسب واقعی را تخمین بزنیم. ساده‌ترین روش تجمیع نظرات، روش رأی اکثریت است. در این روش، برچسب یک داده را کلاسی در نظر می‌گیرند که بیش‌ترین تعداد رأی‌ها را داشته باشد. مدل گرافیکی این روش در شکل ۲.۲-آ نشان داده شده است. همان‌طور که در این شکل مشاهده می‌کنید، فرضی که در این روش وجود دارد، این است که پاسخ‌های داده شده، همگی از یکدیگر مستقل هستند. این در حالی است که دقت پاسخ‌های یک فرد در مسائل مختلف، کاملا با یکدیگر هم‌بستگی دارند، هم‌چنین سختی سوالات نیز بر روی دقت کلی افراد تاثیرگذار است. در مدل نمایش داده شده در شکل ۲.۲-آ، y_i، متغیر تصادفی متناظر با برچسب درست داده‌ی iام است. $A_{i,j}$ نیز نشان‌دهنده پاسخ فرد jام به سوال iام است.

دسته‌ای از روش‌ها که برای بهبود رأی اکثریت ارائه شده‌اند، با مدل کردن توانایی افراد در پاسخ‌گویی به سوالات و میزان دقت آن‌ها در زمان‌های مختلف، میزان درستی نظر آن‌ها را تشخیص داده و از این طریق، با دقت بیشتری جواب صحیح را تخمین می‌زنند. برای این کار، در تمامی این مدل‌ها، میزان دقت و توانایی افراد در پاسخ‌گویی به سوالات را به صورت متغیرهای تصادفی پنهان در نظر گرفته و با طراحی مدل‌های گرافیکی احتمالاتی مناسب، رابطه بین آن‌ها و پاسخ‌های به دست آمده را مدل می‌کنند. پس از آن، یا با استفاده از روش‌هایی مانند بیشینه‌سازی انتظار[۴۵] به صورت هم‌زمان مقادیر آن‌ها و برچسب‌های صحیح را تخمین می‌زنند و یا با استفاده از روش‌های دقیق و تقریبی استنتاج بیزی مانند انتشار باور[۴۶][؟]، انتشار انتظار[۴۷][؟]، و انتقال پیام وردشی[۴۸][؟] توزیع احتمال پسین آن‌ها را به همراه توزیع پسین برچسب داده‌ها به دست می‌آورند. مدل گرافیکی مربوط به این روش‌ها در شکل ۲.۲-ب نشان داده شده است. همان‌طور که در این شکل مشاهده می‌کنید، در این مدل، پاسخ کاربر jام به سوال iام، علاوه بر پاسخ صحیح آن سوال، به توانایی فرد در پاسخ‌گویی به سوالات نیز بستگی دارد. در این نمودار، $\theta_{\mathbf{j}}$ نشان‌دهنده پارامترهایی است که برای مدل کردن دقت افراد در نظر گرفته می‌شود. روش‌های مختلف، توانایی افراد را با روش‌های متفاوتی مدل کرده‌اند.



شکل ۲.۲: مدل‌های گرافیکی روش‌های تجمیع آرا. آ-رأی اکثریت ب-روش‌های مبتنی بر مدل کردن توانایی افراد ج-روش‌های مبتنی بر مدل کردن توانایی افراد و سختی سوالات [؟]

نام روش	پارامترهای مدل	الگوریتم استنتاج
[؟]	$\boldsymbol{\pi^j}$	EM (Likelihood)
[؟]	$1/\beta_j$ و r_j	EM (Joint Prob.)
[؟]	$\alpha_j, \boldsymbol{\pi^j}$	EM/ MF/ BP
[؟]	$\{d_i, \delta_i\}$ و r_j	EP

جدول ۱.۲: خلاصه‌ای از روش‌های تجمیع نظرات در جمع‌سپاری[؟]

در دسته دیگری از روش‌ها که از پیچیدگی بیشتری نیز برخوردار هستند، علاوه بر مدل کردن توانایی کاربران در پاسخ‌گویی به سوالات، میزان سختی سوالات را نیز برای به دست آوردن برچسب‌های درست در نظر می‌گیرند. این مدل‌ها در مواقعی می‌توانند مفید باشند که توانایی نسبی کاربران در دسته‌بندی مجموعه داده‌های مختلف، تابعی از سختی سوالات باشد. مدل کلی این روش‌ها را در شکل ۲.۲-ج مشاهده می‌کنید. در این مدل، به ازای سوال iام، پارامترهایی را که مشخص کننده میزان سختی آن سوال است را تحت عنوان $\beta_\mathbf{i}$ مطرح کرده است. طبق این مدل، در صورتی که توانایی فرد iام در پاسخ‌گویی به سوالات را بدانیم و دشواری سوال j را نیز بدانیم، مستقل از روش‌های دیگر، می‌توان احتمال درست بودن آن پاسخ را به دست آورد.

روش‌های مختلفی در هر یک از دسته‌های گفته شده وجود دارند که دقت افراد و سختی سوالات را با استفاده از پارامترهای مختلفی مانند ماتریس آشفتگی[۴۹] مدل می‌کنند و با استفاده از روش‌های مختلف استنتاج، بر روی این مدل‌ها استنتاج انجام می‌دهند. جدول ۱.۲، لیستی از روش‌های ارائه شده برای تجمیع نظرات را به همراه پارامترهای مورد استفاده در این روش‌ها و روش‌های استنتاج مورد استفاده را نشان می‌دهد[؟]. برای توضیح بیشتر در مورد این روش‌ها به [؟] مراجعه کنید.

۲.۳.۲ مدل‌های استقرایی

همان‌طور که گفته شد، استفاده از روش‌هایی که تنها از تجمیع آرا برای تخمین برچسب‌ها در مسئله دسته‌بندی به کمک جمع‌سپاری استفاده می‌کنند، نیاز به داده‌های برچسب‌دار زیادی دارد. هدف از روش‌های استقرایی این است که با آموزش یک دسته‌بند، تعداد برچسب‌های مورد نیاز را کاهش دهیم. در این روش، به ازای درصد کمی از داده‌ها، تعدادی برچسب گرفته و با استخراج ویژگی[۵۰] و بردن داده‌ها در فضای ویژگی، یک دسته‌بند با استفاده از این داده‌های برچسب‌دار آموزش می‌دهیم. از آن پس، به جای گرفتن برچسب‌های متعدد به ازای هر سوال و تجمیع آن‌ها، از دسته‌بند آموزش داده شده برای برچسب‌زنی داده‌های بدون برچسب استفاده می‌کنیم.

روش‌های استقرایی جمع‌سپاری، نیازمند روشی برای آموزش دسته‌بند با استفاده از داده‌های با برچسب نویزی هستند. تفاوت اصلی روش‌های مختلفی که در این زمینه ارائه شده‌اند، نحوه روبرو شدن آن‌ها با این مسئله است. ساده‌ترین روش برای آموزش دسته‌بند با استفاده از برچسب‌های نویزی، این است که با استفاده از یکی از روش‌های مبتنی بر تجمیع نظرات، ابتدا برچسب‌های واقعی را تخمین زده و از برچسب‌های تخمین زده شده برای آموزش دسته‌بند استفاده کنیم. اصلی‌ترین مشکل این روش این است که نایقینی موجود در برچسب‌های تخمین زده شده را در نظر نگرفته‌ایم.

روش ارائه شده در [؟]، از یک دسته‌بند مبتنی بر برازش لگاریتمی[۵۱] به عنوان مدل دسته‌بندی استفاده کرده و هم‌چنین از مدل [؟] برای مدل کردن دقت کاربران استفاده می‌کند. این روش، با استفاده از روش بیشینه‌سازی انتظار، به صورت تکرارشونده[۵۲] در مرحله انتظار، با استفاده از مقادیر فعلی، پارامتر دسته‌بند و هم‌چنین حساسیت و اختصاصی بودن کاربران و میانگین برچسب‌های داده‌ها را تخمین زده و در مرحله بیشینه‌سازی، با استفاده از برچسب‌های تخمین زده شده در مرحله انتظار، تابع درستنمایی را به صورت هم‌زمان بر حسب پارامترهای دسته‌بند و پارامترهای مدل کاربران بیشینه می‌کند. مزیت این روش نسبت به روش قبل این است که نایقینی برچسب‌های داده را در تخمین پارامتر دسته‌بند دخیل کرده و از این طریق، دقت دسته‌بند را بالا می‌برد.

روش دیگری که در [؟] ارائه شده است، از دسته‌بند خطی مبتنی بر برازش پروبیت[۵۳] [؟] استفاده می‌کند و توانایی افراد در پاسخ‌گویی به سوالات را با یک متغیر تصادفی که از توزیع بتا پیروی می‌کند مدل می‌کند. این روش، با استفاده از قابلیت دسته‌بند پروبیت در بروزرسانی برخط، با استفاده از الگوریتم انتشار انتظار مدل خود را به صورت افزایشی به‌روزرسانی می‌کند و از این رو قابلیت استفاده جهت دسته‌بندی جویبار داده را دارد.

۳ روش ارائه شده

همان‌طور که در بخش ۱.۲ گفته شد، در جویبار داده فرض بر این است که هر داده از یک منبع تولید شده است و در صورتی که بتوانیم منبع هر داده را به درستی تخمین بزنیم، می‌توانیم با یادگیری دسته‌بند از روی داده‌های هم‌مفهوم، هر داده را با دقت بالایی دسته‌بندی کنیم. به طور کلی، به داده‌هایی هم‌مفهوم گفته می‌شود که تابع احتمال پسین $p(y|x)$ آن‌ها یکسان باشد. با توجه به این که مفهوم داده‌ها در اختیار ما قرار نمی‌گیرد، بنابراین، تشخیص مفهوم داده‌ها مسئله‌ای بدون نظارت است. همان‌طور که در بخش۲ گفته شد، تمامی روش‌های ارائه شده برای تشخیص مفهوم، از فرض‌های محدودکننده‌ای مانند یکسان بودن منبع تولید کننده داده‌های یک دسته و مستقل بودن آن‌ها از یکدیگر، یا یکسان بودن مفهوم دسته‌های هم‌خوشه در فضای ویژگی‌های استخراج شده از دسته‌ها [؟] استفاده می‌کنند و بر مبنای این فرضیات، از روش‌های مکاشفه‌ای برای تشخیص و انتخاب مفهوم استفاده می‌کنند. این در حالی است که در بسیاری از کاربردها، چنین فرضیاتی درست نیست. به عنوان مثال، در مسئله مدل کردن رفتار کاربران یک سامانه، در هر لحظه تعداد زیادی کاربر از سامانه استفاده می‌کنند که رفتار آن‌ها را نمی‌توان با استفاده از یک مدل واحد مدل نمود. در این بخش، با مدل کردن مسئله تشخیص مفهوم در جویبار داده به صورت یک مسئله خوشه‌بندی پویا[۵۴] و ترکیب فرآیند رستوران چینی تکرارشونده[۵۵][؟] با مدل‌های دسته‌بندی احتمالاتی، یک چارچوب احتمالاتی کاملا اصولی و نظام‌مند برای دسته‌بندی جویبار داده ارائه می‌دهیم.

مدل فرآیند رستوران چینی تکرار شونده ، یک مدل مخلوط[۵۶] است و از دسته مدل‌های ناپارامتری بیزی[۵۷] است. ویژگی مدل‌های ناپارامتری بیزی این است که برخلاف مدل‌های پارامتری، ساختار متغیرهای پنهان، با آمدن داده‌های بیشتر بزرگ‌تر می‌شود [؟]. مدل‌های مخلوط که معروف‌ترین آن‌ها نیز مدل مخلوط گاوسی[۵۸] است، روش‌هایی هستند برای مدل کردن مجموعه داده‌هایی که ساختار خوشه‌ای دارند و داده‌های هر خوشه از توزیعی متفاوت از داده‌های خوشه‌های دیگر تولید شده است[؟] و از این رو برای مدل کردن جویبار داده مناسب هستند.

در جویبار داده، با توجه به نامشخص بودن تعداد مفهوم‌ها، نیازمند مدل مخلوطی هستیم که تعداد مفاهیم را با توجه به داده‌ها تعیین کند. از این رو، از مدل‌های مخلوط ناپارامتری استفاده می‌کنیم. یکی از مدل مخلوط ناپارامتری بیزی، مدل مخلوط فرآیند رستوران چینی[۵۹] است. این مدل از این جهت که تعداد خوشه‌ها را به صورت خودکار و بر اساس داده‌ها تعیین می‌کند، مورد اهمیت است؛ اما این مدل، بر مبنای فرض جایجایی‌پذیری[۶۰] داده‌ها طراحی شده است. منظور از فرض جایجایی‌پذیری در یک مدل، این است که به ازای یک جویبار داده، در صورتی که ترتیب داده‌ها در جویبار داده تغییر کند، احتمال تولید آن جویبار توسط آن مدل تغییر نکند. همان‌طور که می‌دانیم، در جویبار داده، بین داده‌های نزدیک به هم همبستگی[۶۱] وجود دارد و هر چه این فاصله کمتر باشد، این همبستگی بیشتر می‌شود. به همین جهت، باید مدلی برای داده‌ها داشته باشیم که علاوه بر این که تعداد خوشه‌ها را به صورت خودکار تعیین می‌کند، وابستگی بین داده‌های نزدیک به یکدیگر را مدل کند. از این رو، از مدل مخلوط رستوران چینی تکرارشونده استفاده می‌کنیم. فرضی که در این مدل استفاده می‌شود، این است که داده‌های جویبار داده به صورت دسته‌ای وارد می‌شوند و داده‌های هر دسته جایجایی‌پذیر هستند اما بین داده‌های دسته‌های مختلف، این خاصیت برقرار نمی‌باشد. برخلاف فرض محدودکننده‌ای که تمامی داده‌های یک دسته از یک مفهوم تولید شده‌اند، فرض جایجایی‌پذیری داده‌های یک دسته، فرض بسیار ضعیفی است و کاملا مطابق با واقعیت است. مدل ارائه شده در [؟] با فرض یکسان بودن مفهوم داده‌های یک دسته، دسته‌ها را خوشه‌بندی می‌کرد، اما در این مدل، داده‌های یک دسته می‌توانند از مفاهیم مختلفی تولید شده باشند. در واقع، در این مدل، فرضی که وجود دارد، این است که مفهوم داده‌های یک مدل، از یک توزیع یکسان تولید شده است. فرض دیگری که در این مدل وجود دارد، این است که تغییر توزیع مفهوم‌های داده‌های یک دسته در طی زمان به صورت هموار صورت می‌پذیرد. بدین معنا که تابع توزیع روی مفاهیم مختلف، در طی زمان به صورت هموار تغییر می‌کند. با استفاده از این دو فرض، در ادامه روشی احتمالاتی برای دسته‌بندی جویبار داده ارائه می‌کنیم و روشی برخط و مبتنی بر نمونه‌برداری برای استنتاج بر روی این مدل ارائه می‌کنیم.

۱.۳ مدل پیشنهادی

همان‌طور که گفته شد، فرض می‌کنیم که داده‌ها به صورت دسته‌ای وارد سیستم می‌شوند که n_t نشان‌دهنده تعداد داده‌های دسته tام است و $x_{t,i}$ و $y_{t,i}$ به ترتیب نشان‌دهنده بردار ویژگی و برچسب این دسته باشد. در مدل پیشنهادی، فرض بر این است که مفاهیم می‌توانند در میانه جویبار ظاهر شوند و پس از مدتی غیرفعال شوند. در این مدل، ϕ_k نشان‌دهنده پارامترهای دسته‌بند kام است و $n_{k,t}$ نشان‌دهنده تعداد داده‌های دسته tام است که توسط مفهوم kام تولید شده است. علاوه بر این، $n_{k,t}^{-i}$ همان کمیت را قبل از این که داده‌ی iام وارد شود را نشان می‌دهد. هر کدام از داده‌های $x_{t,i}$ و برچسبش، از یکی از مدل‌های پایه با پارامتر $\theta_{t,i}$ تولید شده است که در صورتی که شماره مدلی که این داده توسط آن تولید شده است را با $z_{t,i}$ نمایش دهیم، خواهیم داشت، $\theta_{t,i} = \phi_{z_{t,i}}$. علاوه بر این، برای سادگی، از نمادگذاری‌های زیر نیز استفاده می‌کنیم (z یک متغیر

عمومی است.):

$$z_{t,1:i} = \{z_{t,1}, z_{t,2}, \ldots, z_{t,i}\}$$

در روش پیشنهادی، هر یک از دسته‌بندهای پایه یک دسته‌بند نایو بیز است با چهار مجموعه پارامتر $\rho_{k,1:m_1,1:C}$ و $\mu_{k,1:m_2,1:C}$ و $\sigma_{k,1:m_2,1:C}$ و β_k که m_1 و m_2 و C به ترتیب تعداد ویژگی‌های گسسته و پیوسته است. در شکل ؟؟ ، تمامی پارامترهای دسته‌بند kام، با نماد ϕ_k نمایش داده شده است و پارامترهای توزیع احتمال پیشین این پارامترها را که $\gamma_{1:m_1,1:C}$ و $\eta_{1:m_2,1:C}$ و $\nu_{1:m_2,1:C}$ و $a_{m_2,1:C}$ و $b_{1:m_2,1:C}$ و π که m_1 و m_2 هستند، با نماد ϕ_0 نمایش داده شده است. مدل گرافیکی روش پیشنهادی، در شکل ؟؟ آمده است. علاوه بر این، مدل مولّد[۶۲] این روش نیز بدین شرح است:

For each batch $t \in \{1, 2, \ldots\}$

For each data $i \in \{1, \ldots, n_t\}$

1. Draw the concept indicator

$z_{t,i}|z_{1:t-1}, z_t^{-i} \sim RCRP(\alpha, \lambda, \Delta)$

2. If $z_{t,i}$ is a new concept,

a) Draw $\beta_{znew}|G_0 \sim Dir(\pi)$

b) for each $c \in \{1, \ldots, C\}$

i) for each $j \in \{1, \ldots, m_1\}$

Draw $\rho_{znew,j,c}|G_0 \sim Dir(\gamma_{j,c})$

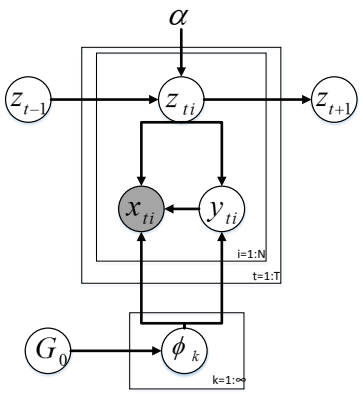
ii) for each $j \in \{1, \ldots, m_2\}$

Draw $\mu_{znew,j,c}|G_0 \sim N(\eta_{j,c}, \nu_{j,c})$

Draw $\sigma_{znew,j,c}|G_0 \sim Gam(a_{j,c}, b_{j,c})$

3. Draw $y_{t,i} \sim Mult(\beta_{z_{t,i}})$

4. Draw $x_{t,i} \sim \prod_{j=1}^{m_1} Mult(x_{t,i}^j; \rho_{z_{t,i},j,y_{t,i}}) \times \prod_{j=1}^{m_2} N(x_{t,i}^{j+m_1}; \mu_{z_{t,i},j,y_{t,i}}, \sigma_{z_{t,i},j,y_{t,i}})$



شکل ۳.۳: مدل گرافیکی روش پیشنهادی

در این مدل مولد، داده iام دسته tام، این گونه ساخته می‌شود که ابتدا شماره مدل پایه‌ای که این داده را تولید می‌کند را با نمونه‌گیری از یک فرآیند رستوران چینی تکرارشونده تعیین می‌کند و در صورتی که تا به حال، هیچ داده‌ای توسط این مدل ساخته نشده است، پارامترهای آن را با نمونه‌برداری از G_0 تعیین می‌کند و پس از آن، $x_{t,i}$ و $y_{t,i}$ این داده را با استفاده از آن مدل پایه ایجاد می‌کند. فرآیند رستوران چینی تکرار شونده به صورت زیر تعریف می‌شود:

$$p(z_{t,i} = k|z_{1:t-1}, z_t^{-i}) \propto \begin{cases} \sum_{\tau=1}^{\Delta} e^{-\frac{\tau}{\lambda}} n_{k,t-\tau} + n_{k,t}^{-i} & \text{existing concept} \\ \alpha & \text{new concept} \end{cases} \quad (۱۳)$$

بنابراین، طبق این مدل، هر چه یک مدل تعداد داده بیشتری را تا به این لحظه تولید کرده باشد، احتمال این که داده بعدی را نیز تولید کند بیشتر است و همواره نیز به یک احتمال متغیری، یک مدل پایه جدید به مجموعه مدل‌ها اضافه می‌شود. طبق رابطه ۱۳، در تعیین تعداد داده‌های تولید شده، از یک تابع کاهشی استفاده شده است تا از این طریق، تاثیر داده‌های اخیر در تصمیم‌گیری بیشتر از داده‌های قدیمی‌تر باشد و از طرفی، تنها داده‌های Δ دسته آخر را در نظر گرفته است. در واقع این مدل، یک روش مولد بسیار مناسب برای مدل کردن روش تغییر مفهوم است. زیرا از یک سو، به داده‌های یک دسته اجازه می‌دهد که از مدل‌های متفاوتی تولید شده باشند و از طرف دیگر، با استفاده از روش‌های انتخاب داده و وزن‌دهی به داده‌ها، توزیع احتمال روی مفاهیم مختلف را در بین دسته‌های مختلف تعیین می‌کند. برای استفاده از این مدل مولد، نیازمند روشی برای استنتاج بر روی این مدل هستیم که در ادامه به آن می‌پردازیم.

۲.۳ الگوریتم استنتاج

به طور کلی، به دلیل پیچیدگی مدل‌های مخلوط ناپارامتری، امکان استنتاج دقیق بیزی بر روی این مدل‌ها وجود ندارد و از این رو، در این مدل‌ها، از روش‌های تقریبی استنتاج مانند الگوریتم‌های وردشی و روش‌های نمونه‌برداری استفاده می‌کنند. در استنتاج بر روی مدل‌های مبتنی بر رستوران چینی، به دلیل خاصیت جابجایی‌پذیری داده‌ها، روش‌های نمونه‌برداری گیبس[۶۳] کاربرد فراوان دارند. روش نمونه‌برداری گیبس، روشی دسته‌ای است و از مجموعه روش‌های نمونه‌برداری بر مبنای زنجیره مارکوف[۶۴] است که برای استنتاج

بر روی یک مدل که دارای متغیرهای پنهان است، یک زنجیره مارکوف تشکیل می‌دهد که فضای حالت آن، مجموعه حالاتی است که این متغیرهای پنهان می‌توانند به خود مقدار بگیرند و توزیع احتمال حالات در وضعیت پایداری، تابع احتمال پسین متغیرهای پنهان است[؟]. برای این که از این روش برای استنتاج بر روی مدل پیشنهادی استفاده کنیم، دو تغییر اصلی در آن ایجاد می‌کنیم. اول این که این الگوریتم را با استفاده از نمونه‌برداری رو به جلو[۶۵] [؟]، به الگوریتمی برخط تبدیل می‌کنیم و ثانیا با حاشیه راندن[۶۶] پارامترهای دسته‌بندهای پایه، فضای حالت زنجیره مارکوف آن را بسیار کوچک می‌کنیم[؟].

الگوریتم استنتاج ما به این صورت عمل می‌کند که یک زنجیره مارکوف تشکیل می‌دهد که فضای حالات آن در لحظه t، مجموعه مقادیر ممکن $z_{1:t}$ باشد. برای این که بتوان تابع احتمال گذار بین حالات مختلف را به روش گیبس به دست آورد، باید با به حاشیه راندن ϕ_kها، تابع احتمال زیر را به دست آورد:

$$p(z_{t,i} = k|z_{1:t}^{-(t,i)}, x_{1:t}, y_{1:t}) \propto P(x_{t,i}, y_{t,i}|z_{t,i} = k, x_{1:t}^{-(t,i)}, y_{1:t}^{-(t,i)})p(z_{t,i} = k|z_{1:t}^{-(t,i)}) \quad (۱۴)$$

$$P(x_{t,i}, y_{t,i}|z_{t,i} = k, x_{1:t}^{-(t,i)}, y_{1:t}^{-(t,i)}) = \int_{\phi_k} p(y_{t,i}|\phi_k)p(x_{t,i}|y_{t,i}, \phi_k)p(\phi_k|x_{1:t}^{-(t,i)}, y_{1:t}^{-(t,i)})d\phi_k \quad (۱۵)$$

$$p(z_{t,i} = k|z_{1:t}^{-(t,i)}) = RCRP(\alpha, \lambda, \Delta) \quad (۱۶)$$

با توجه به مزدوج[۶۷] بودن تابع احتمال پیشین G_0 با تابع درستنمایی $p(x_{t,i}, y_{t,i}|\phi_k)$، تابع احتمال پسین پارامترهای دسته‌بندهای پایه و انتگرال ۱۵ به صورت تحلیلی قابل محاسبه است. برای برخط کردن الگوریتم، به این صورت عمل می‌کنیم که در هر مرحله N نمونه نگه می‌داریم که نشان‌دهنده تابع احتمال پسین $P(z_{1:t})$ باشد و در گام بعد، به ازای هر کدام از این نمونه‌ها، با فرض آن که مقادیری که برای $z_{t-1,1:n_{t-1}}$ در مرحله قبل به دست آمده است، درست است، استنتاج کرده و مقادیر $z_{t,i}$ها را به دست می‌آوریم[؟]. با توجه به این که فرآیند استنتاج به ازای هر کدام از این نمونه‌ها فرآیندی کاملا مستقل از دیگران است، بنابراین، این روش قابلیت موازی شدن بسیار بالایی دارد.

۴ نتایج پیاده‌سازی

در این بخش، بررسی و مقایسه نتایج حاصل از پیاده‌سازی روش‌های پایه و روش پیشنهادی می‌پردازیم. در این آزمایش‌ها از مجموعه داده Spam [؟] که در این حوزه مورد توجه است، بهره گرفته‌ایم. این مجموعه داده دربردارنده ۹۳۲۴ ایمیل است که حدود ۲۵ درصد آن را ایمیل‌های هرز و بقیه را ایمیل‌های عادی تشکیل می‌دهند. به طور کلی، یکی از مشکلات اساسی در زمینه دسته‌بندی جویبار داده این است که اکثر مجموعه داده‌های واقعی به صورت عمومی وجود ندارند[؟، ؟] و به همین دلیل اکثر مجموعه داده‌های مورد استفاده در این حوزه، مجموعه داده‌های مصنوعی هستند و معمولا به گونه‌ای طراحی شده‌اند که روش‌های پیشنهادی پیشین روی آن‌ها به خوبی جواب دهد. بنابراین، این مجموعه داده به دلیل واقعی بودن آن از اهمیت برخوردار است.

با توجه به چارچوب احتمالاتی روش ارائه شده، دو روش تک دسته‌بند احتمالاتی را از بین روش‌های تک‌مدله انتخاب کردیم. روش نایو بیز که به عنوان روش پایه در اکثر روش‌های ارائه شده استفاده می‌شود و روش پروبیت نیز از این جهت حائز اهمیت است که به تازگی برای دسته‌بندی جویبار داده از آن استفاده شده است [؟]. هم‌چنین با توجه به این که روش پیشنهادی ما در دسته روش‌های انتخاب مدل قرار می‌گیرد، دو روش [؟، ؟] که روش‌های جدیدی در زمینه انتخاب مدل هستند را انتخاب کردیم. نتایج حاصل از مقایسه این روش‌ها در جدول ۲.۴ آورده شده است. در این جدول، CCP و PASC و NPSC به ترتیب روش‌های ارائه شده در [؟] و [؟] و روش پیشنهادی را نشان می‌دهند. همان‌طور که ملاحظه می‌شود، روش پیشنهادی، نتایج بهتری نسبت به سایر روش‌ها به دست آورده است. یکی از دلایل این امر این است که روش پیشنهادی، داده‌های با مفهوم

جدول ۲.۴: مقایسه روش پیشنهادی با سایر روش‌ها

NPSC	Probit	PASC	CCP	NB	معیار ارزیابی	نام روش / مجموعه داده
۹۴.۵	۹۲.۴	۹۱.۲	۹۱.۶	۹۰.۷	Accuracy	Spam
۹۵.۴	۹۵.۱	۹۲.۱	۹۲.۳	۹۴.۹	Precision	
۹۷.۴	۹۴.۸	۹۱.۱	۹۶.۹	۹۲.۵	Recall	
۹۶.۴	۹۴.۹	۹۴.۲	۹۴.۵	۹۳.۷	F_1	

یکسان را بهتر تشخیص داده است و به همین جهت به دسته‌بندهای پایه‌ای با داده‌هایی همگن‌تر دست یافته است که این امر سبب افزایش دقت دسته‌بندی شده است. دلیل این امر می‌تواند این باشد که انواع مختلفی از هرزنامه وجود داشته است و در هر بازه زمانی، از تعدادی از این مدل‌ها داده تولید شده است و با توجه به این که روش پیشنهادی بر خلاف سایر روش‌های انتخاب مدل، متفاوت بودن مفهوم داده‌های یک دسته را پشتیبانی می‌کند، به دقت بالاتری دست یافته است.

۵ کارهای آتی

در ادامه مسیر تحقیق، ایده مطرح شده در زمینه مدل کردن مسئله دسته‌بندی جویبار داده با استفاده از مدل‌های مخلوط دسته‌بندی ناپارامتری را تکمیل‌تر می‌کنیم. همان‌طور که در بخش ۴ مشاهده شد، دقت دسته‌بند پروبیت که یک دسته‌بند جداکننده است نسبت به روش نایو بیز که یک دسته‌بند مولد است بهتر است. به همین دلیل، در ادامه، یکی از اصلی‌ترین کارها، استفاده از مدل‌های جداکننده به عنوان دسته‌بند پایه است. مشکلی که در استفاده از این روش‌ها در مدل‌های مخلوط ناپارامتری وجود دارد، نامزدوج بودن تابع درستنمایی این مدل‌ها با تابع احتمال پیشین است که این چالشی است که ما باید برای حل آن یک الگوریتم استنتاج جدید ارائه کنیم.

علاوه بر تغییر مدل دسته‌بندهای پایه به دسته‌بندهای جداکننده و ارائه الگوریتم استنتاج برای آن‌ها، برای ارزیابی کارایی دسته‌بند ارائه شده، نیازمند آزمایش‌های بیشتری بر روی مجموعه داده‌های متفاوت هستیم. بر این اساس، یکی دیگر از کارهای آتی، یافتن مجموعه داده‌های واقعی و تولید مجموعه داده‌های مصنوعی مناسب و آزمایش بر روی آن‌هاست. یکی دیگر از کارهای آتی، بررسی میزان حساسیت روش پیشنهادی بر روی پارامترهای مدل ارائه شده و استفاده از روش‌هایی مانند بیز تجربی[۶۸] برای تخمین این پارامترهاست.

جدول ۳.۵: جدول زمان‌بندی

عنوان فعالیت	مدت زمان لازم	درصد پیشرفت	زمان اتمام
مطالعه روش‌های پیشین	۳ ماه	۱۰۰	شهریور ۹۲
پیاده‌سازی روش‌های پایه و بررسی تأثیر پارامترهای مختلف بر آن‌ها	۱ ماه	۱۰۰	مهر ۹۲
طرح ایده پیشنهادی	۲ ماه	۸۰	آذر ۹۲
پیاده‌سازی روش پیشنهادی، بررسی و مقایسه با سایر روش‌ها	۳ ماه	۶۰	اسفند ۹۲
نگارش مقاله	۱ ماه	۰	فروردین ۹۳
جمع‌بندی و نگارش پایان‌نامه	۲ ماه	۰	خرداد ۹۳

۶ جمع‌بندی

در این گزارش، سیستمی برای دسته‌بندی جویبار داده ارائه کردیم که در آن، برای کاهش هزینه مربوط به برچسب زدن داده‌ها، زیرمجموعه‌ای از داده‌ها را با استفاده از روش انتخاب فعال برای برچسب‌زنی انتخاب می‌کرد و برای برچسب‌زنی آن‌ها، به جای استفاده از افراد متخصص، از جمع‌سپاری استفاده می‌کند. در بخش ۱، به بررسی ساختار این سیستم و اجزای آن پرداختیم و آن را به سه بخش اصلی سیستم انتخاب فعال، سیستم دسته‌بندی جویبار داده و بخش تجمیع نظرات افراد غیرمتخصص تقسیم کردیم و چالش‌های موجود در هر قسمت را بیان کردیم. در بخش ۲، به بررسی روش‌های ارائه شده در هر یک از این حوزه‌ها پرداختیم و دلایل قوت و ضعف هر یک را مورد بررسی قرار دادیم. در قسمت ۳، با تمرکز بر روی بخش دسته‌بند نظارتی جویبار داده، چارچوبی کاملا نظام‌مند و احتمالاتی برای دسته‌بندی جویبار داده ارائه کردیم که با مدل کردن مسئله تغییر مفهوم به عنوان یک مسئله خوشه‌بندی پویا و استفاده از مدل‌های ناپارامتری بیزی به یک روش کارا برای دسته‌بندی دست یافتیم. سپس با مقایسه روش‌های پایه و روش پیشنهادی، موفقیت آن را بر روی یک مجموعه داده واقعی نشان دادیم. نقاط قوت و ضعف برخی روش‌های ارائه‌شده در حوزه دسته‌بندی جویبار داده نیز در جدول ۴.۶، به اختصار شرح داده شده است.

۷ واژه‌نامه

[۱] Big Data
[۲] Relational Database
[۳] Classification
[۴] Non-Stationary
[۵] Concept Drift
[۶] Online-Batch
[۷] Semi-Supervised
[۸] Active
[۹] Machine Vision
[۱۰] Human Computation
[۱۱] Amazon Mechanical Turk, https://www.mturk.com
[۱۲] Multiple Checking
[۱۳] Flexible
[۱۴] Robust
[۱۵] Overfitting
[۱۶] Non-Parametric Bayesian
[۱۷] Generalization
[۱۸] Accuracy
[۱۹] Balanced
[۲۰] Harmonic Mean
[۲۱] Precision
[۲۲] Recall
[۲۳] Hypothesis
[۲۴] Hypothesis Space
[۲۵] Smooth
[۲۶] Naive Bayes Classifier
[۲۷] Model Combination
[۲۸] Model Selection
[۲۹] Bagging
[۳۰] Boosting
[۳۱] Unsupervised
[۳۲] Clustering
[۳۳] Minimal Sufficient Statistic

جدول ۴.۶: مقایسه روش‌های ارائه شده در حوزه دسته‌بندی جویبار داده

نام روش	سال ارائه	مزایا و معایب
وزن‌دهی به درستنمایی [؟]	۲۰۱۱	+مدلی احتمالاتی و نظام‌مند - استفاده از تنها یک مدل - عدم قابلیت یادگیری مفاهیم تکرارشونده - سرعت پایین در بازیابی دقت پس از تغییر مفهوم
ترکیب دسته‌بندهای پایه [؟]	۲۰۱۱	+غنی کردن فضای فرضیه با ترکیب مدل‌های ساده + پشتیبانی از مفاهیم تکرارشونده - قوانین به‌روزرسانی مکاشفه‌ای - تعداد زیاد دسته‌بندهای پایه - نداشتن مکانیسمی برای محدود کردن تعداد دسته‌بندها
خوشه‌بندی بردار ویژگی استخراج شده از دسته‌ها [؟]	۲۰۱۰	+پشتیبانی از مفاهیم تکرارشونده - نداشتن مکانیسمی برای محدود کردن تعداد دسته‌بندها - حساسیت زیاد به پارامترها - یکسان فرض کردن مفهوم تمامی داده‌های یک دسته
روش گروهی مبتنی بر دقت [؟]	۲۰۱۳	+پشتیبانی از مفاهیم تکراشونده +روشی احتمالاتی برای انتخاب دسته‌بندی که باید به‌روز شود - روشی مکاشفه‌ای برای انتخاب مفهوم یک داده - یکسان فرض کردن مفهوم تمامی داده‌های یک دسته
روش دسته‌بندی مبتنی بر مدل مخلوط فرآیند دیریکله [؟]	۲۰۰۹	+ارائه مدلی غیرخطی بر مبنای ترکیب دسته‌بندهای ساده خطی + تعیین تعداد دسته‌بندهای مورد نیاز بر حسب پیچیدگی مدل - عدم پشتیبانی از جویبار داده

[۳۴] Latent Variable
[۳۵] Maximum A Posteriori
[۳۶] Maximum Likelihood
[۳۷] Margin
[۳۸] Entropy
[۳۹] Version Space
[۴۰] Query By Disagreement
[۴۱] Query By Committee
[۴۲] Vote Entropy
[۴۳] Kullback-Leibler
[۴۴] Inductive
[۴۵] Expectation Maximization (EM)
[۴۶] Belief Propagation (BP)
[۴۷] Expectation Propagation (EP)
[۴۸] Variational Message Passing
[۴۹] Confusion Matrix
[۵۰] Feature Extraction
[۵۱] Logistic Regression Classifier
[۵۲] Iterative
[۵۳] Probit Regression
[۵۴] Dynamic Clustering
[۵۵] Recurrent Chinese Restaurant Process
[۵۶] Mixture Model
[۵۷] Bayesian Nonparametric
[۵۸] Gaussian Mixture Model
[۵۹] Chinese Restaurant Process Mixture Model
[۶۰] Exchangeability
[۶۱] Correlation
[۶۲] Generative Model
[۶۳] Gibbs Sampling
[۶۴] Markov Chain Monte Carlo
[۶۵] Forward Sampling
[۶۶] Marginalizing
[۶۷] Conjugate
[۶۸] Empirical Bayes