مجله علوم آماری، بهار و تابستان ۱۴۰۰
جلد ۱۵، شماره ۱، ص ۱۱۹ – ۱۴۶
DOI: 10.29252/jss.15.1.121
مقاله پژوهشی

استفاده از الگوریتم‌های یادگیری آماری رده‌بندی در آمار رسمی

زهرا رضائی قهرودی[۱]، حسن رنجی[۲] و علیرضا رضایی[۲]

[۱]دانشکده ریاضی، آمار و علوم کامپیوتر، دانشگاه تهران
[۲]مرکز آمار ایران

تاریخ دریافت: ۱۳۹۹/۰۱/۰۵ تاریخ پذیرش و انتشار: ۱۳۹۹/۰۸/۰۳

چکیده: در اکثر آمارگیری‌ها، پرسش مشاغل و فعالیت‌ها از طریق پرسش‌های باز سوال می‌شود و کدگذاری این اطلاعات به هزاران رده به روش دستی صورت می‌گیرد که بسیار زمان‌بر و پرهزینه است. با توجه به ضروریات مدرن‌سازی نظام آماری کشورها، امروزه استفاده از روش‌های یادگیری آماری در آمار رسمی برای داده‌های اولیه و ثانویه ضروری است. همچنین، روش‌های رده بندی یادگیری آماری در فرایند تولید آمار رسمی بسیار کاربرد دارد. هدف این مقاله، کدگذاری برخی فرایندهای آمارگیری ها با روش‌های یادگیری آماری و آشنایی مدیران در مورد امکان استفاده از روش‌های یادگیری آماری در تولید آمارهای رسمی است. دو کاربرد از روش‌های یادگیری آماری رده بندی شامل کدگذاری خودکار رشته فعالیت‌های اقتصادی و کدگذاری پرسش های باز پرسشنامه های مراکز آماری با چهار روش تکرار، روش ترکیبی ماشین بردار پشتیبان با ترکیب مدل‌ها در سطوح مختلف تجمیع، ترکیب روش تکرار و ماشین بردار پشتیبان و روش نزدیکترین همسایه روی داده‌های آمارگیری از کارگاه‌های صنعتی ایران انجام شده است.

واژه‌های کلیدی: کدگذاری خودکار، متن‌کاوی، یادگیری آماری، آمار رسمی.

آدرس الکترونیکی نویسنده مسئول مقاله: زهرا رضائی قهرودی، z.rezaeigh@ut.ac.ir
کد موضوع‌بندی ریاضی (۲۰۱۰): 62G05, 62D05, 6205.

۱ مقدمه

یادگیری آماری، مجموعه وسیعی از ابزارها برای درک داده‌ها است که به دو دسته راهنماییده[1] و ناراهنماییده[2] تقسیم می‌شوند. از طرفی یادگیری آماری شاخه‌ای از آمار کاربردی است که در پاسخ به یادگیری ماشین ظاهر شده است و بر مدل‌های آماری و ارزیابی عدم‌حتمیت تأکید دارد. یادگیری راهنماییده شامل ساختن یک مدل آماری برای پیش‌گویی و برآورد خروجی براساس یک یا چند ورودی است. از دیدگاه نظری، یادگیری راهنماییده شامل یادگیری از یک مجموعه‌داده آموزشی است که در آن هر نقطه در مجموعه‌داده آموزشی، یک جفت ورودی/ خروجی است، به طوری که در آن ورودی به یک خروجی نگاشت می‌شود. پس از یادگیری یک تابع بر اساس مجموعه‌داده‌های آموزشی، آن تابع بر روی یک مجموعه‌داده آزمایشی، که در مجموعه داده‌های آموزشی ظاهر نشده است، اعمال می‌شود. بسته به نوع خروجی، مسائل یادگیری راهنماییده، به رگرسیون یا رده‌بندی تقسیم می‌شوند. اگر خروجی مقادیر پیوسته را اخذ کند، راه حل، رگرسیون است و اگر خروجی مقادیر گسسته باشد، راه حل، روش رده‌بندی[3] است. یادگیری ناراهنماییده نوعی الگوریتم یادگیری آماری است که برای به دست آوردن استنباط از مجموعه‌داده‌های ورودی بوجود آمده است. به عبارت دیگر در یادگیری ناراهنماییده، تنها داده ورودی وجود دارد و متغیرهای خروجی متناظر وجود ندارند. در این روش، یادگیری در خصوص ارتباط و ساختار داده‌ها است. معمول‌ترین روش‌های یادگیری ناراهنماییده، کاهش بُعد و خوشه‌بندی است که برای تحلیل کاوشگرانه داده‌ها و پیدا کردن الگوهای پنهان یا گروه‌بندی داده‌ها مورد استفاده قرار می‌گیرد (هستی و همکاران ، ۲۰۰۹؛ جیمز و همکاران ، ۲۰۱۴).

روش‌های یادگیری آماری و یادگیری ماشین در آمار رسمی هم برای نمونه‌های احتمالی (روش‌های سنتی گرداوری داده‌ها) و هم در نمونه‌های نااحتمالی (داده‌های ثبتی و مِه‌داده‌ها) و داده‌های آمیخته‌مُد[4] کاربرد دارد. اخیراً روش‌های یادگیری ماشین یا یادگیری آماری در فرایند تولید آمار رسمی کاربردهای زیادی دارد. به عنوان مثال یک مدل روی داده‌های آموزشی که به صورت دستی کدگذاری شده است، آموزش داده می‌شود و سپس برای پیش‌بینی کدی با بیشترین احتمال روی داده‌های جدید استفاده می‌شود. این رویکرد در اداره آمار استرالیا برای کدگذاری استفاده می‌شود (کلارک و بروکر ، ۲۰۱۱). کدگذاری خودکار مبتنی بر روش‌های یادگیری آماری در ایالات متحده (دی ، ۲۰۱۴) و آلمان (بثمان و همکاران ، ۲۰۱۴) نیز استفاده می‌شود.

---

[1] Supervised
[2] Unsupervised
[3] Classification
[4] Mixed mode data

در اداره آمار هلند، سامانه‌هایی مانند $COBS$[1] و $PART$[2] براساس کلمات اختصاری هلندی به صورت خودکار و نیمه‌خودکار به کدگذاری متغیرهای مربوط به تحصیلات، رشته فعالیت‌های اقتصادی، شغل‌ها، بیماری‌ها و دلایل مرگ‌و میر، طبقه‌بندی کالاها و خدمات و غیره می‌پردازد (اداره آمار هلند ، ۲۰۱۲). در پرسشنامه‌های کاغذی، پرسش‌های شغل، رشته فعالیت اقتصادی، بیماری‌ها و غیره به صورت پرسش باز مطرح می‌شود در حالی که در آمارگیری‌های وبی، این پرسش‌ها به صورت درخت جستجو است. تخصیص کد صحیح به پرسش‌های باز براساس دیکشنری یا رده‌بندی‌های بین‌المللی به روش‌های مختلف انجام می‌گیرد. روش‌های سنتی کدگذاری، کدگذاری دستی است که بسیار زمان‌بر، هزینه‌بر و نیاز به دانش حرفه‌ای در این زمینه دارد زیرا تعداد عناوین کدها زیاد است. از طرفی کدگذاری توسط افراد مختلف با مهارت‌های مختلف انجام می‌شود که خود منجر به خطا می‌شود. کدگذاری خودکار براساس کدگذاری قاعده‌مبنا انجام می‌شود. به عنوان مثال، اگر متن پاسخ حاوی کلمه‌ای باشد که بتواند در فرهنگ لغتِ از پیش تعریف شده، با یکی از کدها جور شود، کد مربوطه در فرهنگ لغت، به متن پاسخ تخصیص داده می‌شود. اگرچه روش‌های کدگذاری خودکار، هزینه‌ها را کاهش می‌دهد، اما کدگذاری‌های خودکارِ کامل نیز چالش‌برانگیز است. با کدگذاری خودکار جزئی، پاسخ‌های آسان‌کدگذاری به روش خودکار کدگذاری می‌شود و پاسخ‌های سخت‌کدگذاری به صورت دستی کدگذاری می‌شوند. در این حالت، یک نمره عددی (عدد آستانه) برای تشخیص بین لغات آسان‌کدگذاری و سخت‌کدگذاری استفاده می‌شود.

در کدگذاری براساس روش‌های یادگیری آماری، مدل‌های آماری بر اساس داده‌های آموزشی رده‌بندی شده آموزش دیده می‌شوند. چنین روش‌هایی نه تنها برای کدگذاری شغل بلکه برای رده‌بندی‌های عمومی دیگر یا تقسیم‌بندی پرسش‌های باز نیز استفاده می‌شوند. در این روش هنگامی که مدل آموزش دیده می‌شود، مشاهدات جدید می‌تواند به طور خودکار رده‌بندی شوند. برای ساخت یک مدل، متن به داده عددی تبدیل می‌شود. در روش استاندارد متن‌کاوی، یک متغیر برای هر کلمه در متون مختلف ایجاد می‌شود. برای هر یک از متغیرهای تک‌نگاشت[3]، فراوانی کلماتی که در یک متن وجود دارد شمارش می‌شود یا وجود یا عدم وجود کلمه در متن موردنظر مشخص می‌شود. تغییرات زیادی در روش‌های متن‌کاوی مانند اضافه کردن متغیرها برای حضور یا عدم حضور یک دنباله چندکلمه‌ای (متغیرهای n-نگاشت[4])، حذف کلماتی که زیاد استفاده

[1] Computer Ondersteund Bewerkings Systeem
[2] Programma Anders Typeren
[3] Unigram
[4] Ngram

می‌شوند و احتمالاً مفید نیستند (کلمات بی‌اثر[1])، ریشه‌دار کردن کلمات[2] به ریشه گرامری آن‌ها وجود دارد. تعداد زیادی از این متغیرها با الگوریتم‌های یادگیری آماری مانند ماشین‌های بردار پشتیبان[3] (SVM) مدل‌بندی می‌شوند (واپنیک ، ۲۰۰۰). الگوریتم‌های مختلفی برای یادگیری کدگذاری شغل استفاده می‌شود. اداره آمار استرالیا رده‌بندی خودکار را با استفاده از روش SVM روی داده‌های سرشماری استرالیا از سال ۲۰۰۶ به بعد انجام می‌دهد. در آمارگیری اجتماعی امریکا تغییراتی در روش‌های متن‌کاوی ایجاد شده است (تامسون و همکاران ، ۲۰۱۲)، به طوری که متغیرهای دنباله‌ای یک و دو کلمه‌ای از متن ایجاد می‌شود و برای محدود کردن تعداد متغیرها برای تحلیل، از آستانه ۳۰ استفاده می‌شود، یعنی در صورتی متغیر ایجاد می‌شود که حداقل ۳۰ بار در متون داده‌های آموزش استفاده شده باشد.

برخی از نویسندگان روش‌های نزدیکترین همسایه را بررسی کرده‌اند که در آن کدی را به پاسخ یا متن در داده‌های آموزشی منتسب می‌کنند که بیشترین شباهت به پاسخ مورد نظر را داشته باشد. برای اندازه‌گیری نزدیکی یا شباهت بین دو پاسخ، معیارهای تشابه مختلف استفاده شده است. سامانه PACE از روش k-نزدیکترین همسایه روی داده‌های اداره سرشماری ایالات متحده استفاده می‌کند (کرسی و همکاران ، ۱۹۹۲). راس و همکاران (۲۰۱۴) روش نزدیک‌ترین همسایه را با معیار تشابه ژاکارد برای رده‌بندی استاندارد شغل[4] استفاده می‌کنند. دقت یا نسبت مشاهدات به درستی رده‌بندی شده برای کدگذاری کاملاً خودکار، برای کدهای شش رقمی ۵۱٪ و برای کدهای سه رقمی ۶۴٪ در نظر گرفته شده است.

روش کدگذاری در آمارگیری ALWA موسسه تحقیقات شغل آلمان پیش‌پردازش کامل عین متن به جای روش متن‌کاوی با استفاده از متغیرهای n-نگاشت است (اسچرهواتز ، ۲۰۱۴). با استفاده از پاسخ‌های عین‌متن، به جای متغیرهای n-نگاشت، تعداد متغیرهای یادگیری به شدت کاهش می‌یابد. اسچرهواتز (۲۰۱۴) از روش‌های مختلف از جمله بیز ساده[5] و مدل گرادیان تقویتی[6] (فریدمن، ۲۰۰۱) برای کدگذاری استفاده کردند. این مقاله نشان داد که رویکردهای بیز ساده و مدل گرادیان تقویتی با افزایش دقت، عملکرد یکسانی را نشان می‌دهند. در این مقاله در بخش ۲ به معرفی روش‌های یادگیری آماری در آمارگیری‌ها شامل روش‌های یادگیری آماری در داده‌های اولیه و داده‌های ثانویه پرداخته می‌شود. همچنین روش‌های پردازش زبان‌های طبیعی و متن‌کاوی معرفی می‌شود. در بخش ۳ به معرفی چهار روش یادگیری

[1] Stopwords
[2] Stemming words
[3] Support vector machines
[4] Standard Occupational Classification
[5] Naïve Bayes
[6] Gradient boosting model

آماری در آمار رسمی برای رده‌بندی پرداخته می‌شود. در بخش ۴ به معرفی کاربرد روش‌های یادگیری آماری در فعالیت‌های مرکز آمار ایران شامل دو کاربرد از روش‌های یادگیری آماری در متن‌کاوی شامل کدگذاری خودکار رشته فعالیت‌های اقتصادی و تخصیص کد واجد شرایط بودن یا نبودن به پرسش باز عدم تکمیل پرسشنامه با استفاده از نرم‌افزار R پرداخته شده است و در نهایت نتیجه‌گیری در بخش ۵ ارائه خواهد شد. با توجه به اینکه حجم داده‌های آمار رسمی زیاد است و اجرای روش‌های یادگیری آماری در نرم‌افزار R برای داده‌های حجیم، امکان‌پذیر نیست، کدهای نوشته شده در نرم‌افزار R ، در نرم‌افزار SAS نیز پیاده شده است تا این روش‌ها در مراکز آماری قابل اجرا شوند (رضائی قهرودی و همکاران ، ۱۳۹۹).

۲ یادگیری آماری در آمارگیری‌ها

با توجه به اینکه فرایند تولید آمار در سازمان‌های مختلف به شیوه‌های گوناگون طراحی و اجرا می‌شود، این امر باعث می‌شود که تبادل دانش، بهره‌گیری از تجربیات موفق و همکاری بین سازمان‌های ملی و بین‌المللی به‌سادگی صورت نگیرد. به‌منظور حل این مشکل و کمک به سازمان‌های آماری برای بحث و تبادل نظر در زمینه توسعه نظام داده و فراداده‌های آماری، مدل عمومی فرایند کسب و کار آماری[1] توسط کمیسیون اقتصادی سازمان ملل برای اروپا معرفی شده است (برفی پور و قادری ، ۱۳۹۹). ویرایش پنجم این مدل در سال ۲۰۱۳ توسط این کمیسیون ارائه شده است. در مدل عمومی فرایند کسب و کار آماری، تمام فرآیندهای اصلی تولید آمار در یک سازمان آماری به طور یکپارچه، توصیف و ارتباط بین آن‌ها به روشنی بیان می‌شود. این مدل به منظور بهینگی و کارآمدی نظام تولید محصولات و ارائه خدمات در یک سازمان آماری طراحی شده است. مدل عمومی فرایند کسب و کار آماری تمرکز خود را بر تولید آمارهای رسمی با استفاده از منابع داده اولیه مانند آمارگیری‌های نمونه‌ای و منابع داده ثانویه مانند داده‌های ثبتی یا مِه‌داده‌ها دارد. به‌عنوان نمونه اداره آمار استرالیا و اداره آمار کانادا سازمان‌هایی هستند که مدل عمومی فرایند کسب و کار آماری را همان‌گونه که هست، پذیرفته‌اند.

با توجه به ضرورت مدرن‌سازی نظام آماری کشورها، استفاده از روش‌های یادگیری آماری در فرایند تولید آمار رسمی شامل فرایند تولید داده‌های اولیه (داده‌های حاصل از آمارگیری‌های نمونه‌ای و سرشماری‌ها) و داده‌های ثانویه (مِه‌داده، داده‌های ثبتی و آمارهای چندمنبعی) در مدل عمومی فرایند کسب و کار آماری امری ضروری است. از طرفی پردازش زبان‌های طبیعی و متن‌کاوی، نقش مهمی در پردازش نیمه‌خودکار

[1] Generic Statistical Business Process Model

متون در پرسش‌نامه‌ها دارد. در ادامه به کاربرد روش‌های یادگیری آماری در داده‌های اولیه و ثانویه در فرایند تولید آمار رسمی پرداخته می‌شود.

۲.۱ کاربرد روش‌های یادگیری آماری در داده‌های اولیه

از آنجا که منابع ساخت چارچوب‌های نمونه‌گیری شامل ثبت‌ها، داده‌های اداری، سرشماری‌ها و سایر آمارگیری‌ها است، این منابع می‌توانند از طریق فرآیندهای اتصال رکوردها با استفاده از الگوریتم‌های خوشه‌بندی، ترکیب شوند (هارون و همکاران، ۲۰۱۵). از طرفی هنگام تهیه و آماده‌سازی چارچوب‌های نمونه‌گیری، کیفیت اطلاعات چارچوب نمونه‌گیری مانند اطلاعات رشته فعالیت‌های صنعتی، اطلاعات جغرافیایی، اطلاعات مشاغل و ...از اهمیت بالایی برخوردار است. در این حوزه نیز الگوریتم‌های رده‌بندی یادگیری آماری امکان کدگذاری نیمه‌خودکار و با کیفیت بالا را به جای کدگذاری دستی که زمان‌بر و پر هزینه است، فراهم می‌کند. علاوه بر کدگذاری رشته فعالیت‌ها، از روش‌های یادگیری ماشین برای اعتبارسنجی کیفیت اطلاعات چارچوب‌های آماری نیز استفاده می‌شود. به عنوان مثال، می‌توان از روش‌های خوشه‌بندی برای شناسایی مقادیر دورافتاده در اطلاعات چارچوب استفاده کرد.

اگر چه ممکن است روش‌های یادگیری آماری در عملیات گردآوری داده‌ها کمک‌کننده نباشد، اما در مدیریت فعالیت گردآوری می‌تواند کمک‌کننده باشد. به عنوان مثال از احتمالات پاسخ برای برآورد یا پیش‌بینی احتمال پاسخ هر واحد نمونه استفاده می‌شود. از هر دو روش سنتی (به‌عنوان مثال رگرسیون لوژستیک) و روش‌های یادگیری ماشین (به‌عنوان مثال الگوریتم‌های رگرسیون) می‌توان برای پیش‌بینی احتمال پاسخ هر واحد نمونه با استفاده از اطلاعات موجود، استفاده کرد. به منظور بهبود برازش مدل‌های تمایل به پاسخگویی که به برآورد احتمال پاسخ می‌پردازد، معمولاً احتمالات براساس گروه‌هایی از واحدهای نمونه که رفتارهای مشابهی دارند، تخمین زده می‌شوند. روش‌های یادگیری آماری از اطلاعات کمکی یا پارادادههای مربوط به واحدها در الگوریتم‌های خوشه‌بندی، رده‌بندی یا رگرسیون برای رده‌بندی واحدها به زیر گروه‌ها یا زیرجامعه‌های مختلف با هدف افزایش نرخ پاسخگویی استفاده می‌کنند. همچنین در حین گردآوری داده‌ها، به ویژه با دستگاه‌های الکترونیکی، داده‌ها در حین ثبت، بررسی و تأیید می‌شوند. در این مرحله می‌توان از روش‌های خوشه‌بندی برای شناسایی داده‌های دورافتاده نادرست در حین گردآوری داده‌ها استفاده کرد تا حین اجرای آمارگیری، اطلاعات نادرست توسط پاسخگو یا با استفاده از روش‌های یادگیری راهنماییده اصلاح شود. از طرفی بسیاری از مراکز آماری به پاسخگویان این اجازه را می‌دهند که نظر یا پرسش‌هایشان را به‌صورت پرسش باز در پرسش‌نامه ثبت کنند. در این موارد می‌توان از ابزارهای پردازش

زبان طبیعی برای پردازش نظرات یا پرسش‌های مطرح‌شده بهره برد و با استفاده از روش‌های متن‌کاوی هر یک از نظرات را در کمترین زمان و با هزینه کم رده‌بندی کرد.

در فعالیت‌های مربوط به تجمیع داده‌ها از چندین منبع آماری، در صورت عدم وجود شناسه یکتا در همه منابع داده، می‌توان از تکنیک‌های احتمالی اتصال رکوردها استفاده کرد (هارون و همکاران، ۲۰۱۵). در مرحله یکپارچه‌سازی و تجمیع داده‌ها، می‌توان از روش‌های یادگیری آماری برای پاک‌سازی داده‌ها شامل شناسایی داده‌های دورافتاده، خطاها، رکوردهای ناسازگار و غیره استفاده کرد. همچنین برای رده‌بندی و کدگذاری داده‌ها مطابق استانداردهای بین‌المللی می‌توان از روش‌های یادگیری آماری رده‌بندی بهره برد. از روش‌های یادگیری آماری راهنماییده می‌توان برای جانهی داده‌های گم‌شده یا نادرست استفاده کرد. از آنجا که عملکرد روش جانهی در گروه‌های همگن بهبود می‌یابد، بنا بر این می‌توان از روش‌های یادگیری آماری برای تعیین گروه‌های همگن نیز بهره برد. در جدول ۱ به‌طور خلاصه، امکان استفاده از روش‌های یادگیری آماری در برخی فعالیت‌های آماری مراکز آماری معرفی شده است.

جدول ۱. کاربرد یادگیری آماری در فعالیت‌های مرتبط با داده‌های اولیه

مدل	روش‌های یادگیری آماری
اتصال رکوردها	خوشه‌بندی
کدگذاری	رده‌بندی
شناسایی داده‌های دورافتاده	خوشه‌بندی
طبقه‌بندی	رده‌بندی
برآورد	رگرسیون-رده‌بندی
جانهی	رگرسیون-رده‌بندی
کالیبدن	رگرسیون-رده‌بندی
کنترل افشا	رگرسیون-رده‌بندی

۲.۲ کاربرد روش‌های یادگیری آماری در داده‌های ثانویه

داده‌های ثانویه داده‌هایی هستند که برای اهداف آماری گردآوری نمی‌شوند اما ممکن است حاوی اطلاعات آماری موردعلاقه سازمان‌ها یا مراکز آماری باشند. این داده‌ها شامل مِه‌داده‌ها، داده‌های ثبتی و ترکیب این داده‌ها با منابع دیگر مانند داده‌های اولیه هستند. در این زیر بخش، به معرفی روش‌های یادگیری آماری برای مِه‌داده‌ها، داده‌های ثبتی و داده‌های چندمنبعی پرداخته می‌شود. مِه‌داده شامل مجموعه داده‌های با حجم بالا، سرعت بالا و تنوع زیاد است که توسط دستگاه‌ها و سازمان‌های دولتی و خصوصی در قالب

داده‌های ساختمند[1] (مانند پایگاه‌های داده، حسگرها و ...) و داده‌های ناساختمند[2] (مانند ایمیل، رسانه‌های اجتماعی، تصاویر و...) تولید و گردآوری می‌شوند و نیاز به اشکال جدید پردازش برای کشف دانش و بهینه‌سازی فرایند دارند. برای رفع چالش حجم بزرگ منابع مِه‌داده‌ها، رویکردهای مدرن یادگیری ماشین بهتر از روش‌های آماری سنتی قابل استفاده هستند. زیرا روش‌های کارای یادگیری ماشین مانند جنگل تصادفی یا یادگیری عمیق اگرچه از نظر محاسباتی و تحت شرایط خاص، بسیار پرهزینه‌تر از روش‌های سنتی هستند ولی به راحتی قابل تجزیه و تبدیل به‌صورت پردازش‌های موازی برای داده‌های بزرگ هستند. به‌عنوان مثال، در «روش جنگل تصادفی» به دلیل اینکه درختانی که جنگل را تشکیل می‌دهند، به‌طور مستقل از روی نمونه‌های مختلف ساخته و آموزش داده می‌شوند، به راحتی قابل تجزیه هستند و امکان پردازش بسیار کارآمد مِه‌داده‌ها را فراهم می‌آورد. به طور مشابه، مدل‌های یادگیری عمیق، علی‌رغم اینکه از نظر محاسباتی پیچیده هستند، اما به صورت موازی پردازش می‌شوند، زیرا نورون‌های موجود در هر لایه از شبکه‌های عصبی بطور مستقل پردازش می‌شوند. به این ترتیب، کاهش قابل توجهی در زمان محاسبات حاصل می‌شود و استفاده از مدل‌های بسیار پیچیده یادگیری عمیق روی مِه‌داده‌ها را امکان‌پذیر می‌کند.

داده‌های چندمنبعی براساس تلفیق و ترکیب چند منبع داده مانند ترکیب یک یا چند آمارگیری، ثبت‌های اداری یا مِه‌داده‌ها به‌وجود می‌آیند. کاربرد روش‌های یادگیری در داده‌های ثانویه در دو حوزه روش‌های ادغام، اتصال رکوردها و جورسازی منابع و همچنین در تحلیل داده‌ها نیز وجود دارد. در روش‌های ادغام، اتصال رکوردها و جورسازی منابع مختلف، از روش‌های احتمالی و روش‌های قطعیتی یادگیری آماری برای یکپارچه‌سازی داده‌ها استفاده می‌شود. موضوع اتصال رکوردها و جورسازی به‌خصوص اتصال منابع داده‌های اداری یا مِه‌داده‌ها توسط هارون و همکاران (۲۰۱۷) و لوهر و راگانتن (۲۰۱۷) مورد مطالعه قرار گرفته است. نکته مهم در رویکردهای یادگیری آماری، میزان ادغام منابع داده است. اتصال و یکپارچه‌سازی منابع اطلاعاتی در دو سطح خرد و کلان صورت می‌پذیرد. اتصال خرد هنگامی حاصل می‌شود که واحدهای فردی در چندین مجموعه داده با یکدیگر در ارتباط باشند که اغلب نیازمند شناسه یکتا است. هنگامی که واحدهای مشاهده‌شده در مِه‌داده بتوانند به واحدهای موجود در یک یا چند ثبت اداری مرتبط شوند، رکوردهای مِه‌داده‌ها با اضافه شدن متغیرهای کمکی موجود در ثبت‌های اداری، غنی می‌شوند و می‌توانند در برازش مدل‌ها و پیش‌بینی مورد استفاده قرار گیرند. به عنوان مثال، اجرای سرشماری نفوس و مسکن ۱۳۹۵ در ایران به روش ترکیبی با ترکیب مصاحبه حضوری و اینترنتی بوده است. پس از اجرای سرشماری به دو روش و اتصال دو پایگاه داده، شناسایی افرادی که در هر دو روش مصاحبه حضوری و اینترنتی شرکت

[1] Structured data
[2] Unstructured data

کرده‌اند نیازمند استفاده از روش‌های جورسازی آماری بود. همچنین ایجاد چارچوب کارگاهی از منابع و پایگاه‌داده‌های مختلف ثبتی در کشور و حذف کارگاه‌های تکراری، از کاربردهای دیگر روش‌های یادگیری آماری است.

وقتی اتصال رکوردها در سطح فردی برای واحدها از چندین منبع به‌صورت جداگانه امکان‌پذیر نباشد، اما در سطوح تجمیعی و انبوهشی امکان‌پذیر باشد، اتصال رکوردها در سطح کلان صورت می‌گیرد. افراد در یک منطقه شهرداری یا کسب و کارهای یک رشته فعالیت از جمله مثال‌های مربوط به اتصال رکوردها در سطح کلان است. در آمار رسمی از روش‌های یادگیری ناراهنماییده خوشه‌بندی برای کاهش بعد و درنظر گرفتن ویژگی‌های اصلی و تأثیرگذار استفاده می‌شود (گروه یادگیری ماشین UNECE ، ۲۰۱۸).

۲.۳ پردازش زبان‌های طبیعی

زبان طبیعی[1] به زبان‌هایی گفته می‌شود که انسان‌ها برای برقراری ارتباط با انسان‌های دیگر از آن استفاده می‌کنند. زبان غیرطبیعی به زبان‌هایی مانند زبان برنامه‌نویسی گفته می‌شود که پردازش آن‌ها مربوط به مترجم‌ها است. پردازش زبان‌های طبیعی کاربردهای متعددی از جمله تصحیح املاء، ترجمه ماشین، تحلیل و پردازش احساسات، خلاصه‌سازی، استخراج اطلاعات، خوشه‌بندی، رده‌بندی و غیره دارد.

متن‌کاوی که به داده‌کاوی متن، تحلیل هوشمند متن و کشف دانش در متون نیز شناخته می‌شود، فرآیند نیمه‌خودکار استخراج الگوها (اطلاعات مفید و دانش) از حجم زیادی منابع داده ناساختمند است. متن‌کاوی، هنر تبدیل متون به متغیرهای عددی و سپس تحلیل آن‌ها با استفاده از روش‌های آماری است. به دلیل اینکه تعداد متغیرها می‌تواند بسیار زیاد باشد، روش‌های یادگیری آماری و یادگیری ماشین نقش اساسی در این نوع داده‌ها دارد. متن‌کاوی و داده‌کاوی هر دو به دنبال کشف و استخراج الگوهای جدید از داده‌ها هستند و هر دو از فرآیندهای نیمه‌خودکار برای کشف دانش استفاده می‌کنند. تفاوت متن‌کاوی و داده‌کاوی در این است که داده‌کاوی داده‌های ساختمند (پایگاه داده‌ها) را مورد کاوش قرار می‌دهد در حالی که متن‌کاوی داده‌های ناساختمند (متون $Word$ ، فایل‌های PDF ، فایل‌های XML) را مورد بررسی قرار می‌دهد. از آنجا که متون و زبان‌های طبیعی بسیار متنوع و پیچیده است و از نوع داده‌های ناساختمند هستند، به منظور تحلیل آن‌ها باید از فناوری‌های جدید و فرایند متن‌کاوی برای تبدیل متون به داده‌های عددی استفاده کرد. متن‌کاوی در آمار رسمی کاربردهای مختلفی از جمله رده‌بندی و خوشه‌بندی پرسش‌های متن‌باز در آمارگیری‌ها، کدگذاری رشته فعالیت‌های اقتصادی و مشاغل دارد.

[1] Natural Language

اکثر روش‌های متن‌کاوی براساس متغیرهای n-نگاشت است. در ادامه مراحل فرایند متن‌کاوی براساس متغیرهای n-نگاشت به صورت کلی بیان می‌شود:

الف- انجام فرایند پیش‌پردازش: فرایند حذف لغات و حروف اضافه (مانند و، یک، از و ...)، تعیین ریشه لغات (برای مثال «تلفن کردن»، «تلفن شده» و «تلفن‌ها» به صورت «تلفن» نوشته شوند)، حذف نقطه‌گذاری و ویرگول و غیره توجه به کلمات مترادف، برای مثال شاگرد و دانش‌آموز در یک گروه قرار می‌گیرند.

ب- استخراج کلمات یا ویژگی‌ها از تمام متون، ایجاد متغیرهای تک‌نگاشت: برای هر کلمه در متون، یک متغیر ایجاد می‌شود. لازم به ذکر است لغاتی که به ندرت در متون ظاهر شده باشند یا از اهمیت کمتری برخوردار باشند، به منظور کاهش زمان محاسبات و پردازش، حذف می‌شوند. برای افزایش دقت پیش‌گویی، می‌توان متغیرهای n-نگاشت نیز ایجاد کرد. این کار منجر به افزایش تعداد متغیرهای کمکی و در نتیجه افزایش زمان محاسبات می‌شود.

ج- ساخت ماتریس عبارت در متن[1] (TDM) که در آن سطرها، هر یک از متون، ستون‌ها متغیرهای ایجادشده از تمام متون و W_{ij}، شمارش فراوانی کلماتی که در یک متن وجود دارد یا کد صفر یا یک به منزله وجود یا عدم وجود کلمه در متن موردنظر یا وزن کلمه j در متن i است (جدول ۲).

جدول ۲. ماتریس TDM

کلمات				
n	...	۲	۱	متن
$W_{۱n}$	...	$W_{۱۲}$	$W_{۱۱}$	۱
$W_{۲n}$	...	$W_{۲۲}$	$W_{۲۱}$	۲
⋮	⋮	⋮	⋮	⋮
W_{sn}	...	$W_{s۲}$	$W_{s۱}$	s

د- استفاده از الگوریتم‌های مختلف یادگیری آماری روی داده‌های آموزشی و آزمایشی که متغیر پاسخ آن یک‌بار به‌صورت دستی کدگذاری شده است.

ه- پیش‌بینی متغیر پاسخ (کدها) برای متون جدید

[1] Term by Document Matrix

۳ روش‌های یادگیری آماری در آمار رسمی

برای کدگذاری رشته فعالیت‌های اقتصادی، مشاغل یا هر یک از رده‌بندی‌های بین‌المللی می‌توان از روش‌های یادگیری آماری استفاده کرد. در ادامه چهار الگوریتم یادگیری آماری که توسط جیون و همکاران (۲۰۱۷) ارائه شده است، معرفی می‌شود.

الف- روش تکرار با استفاده از تعریف تکرار n-نگاشت‌مبنا: تکرار در عبارت‌های متنی اشاره به دو رشته یا دو متن دارد که شبیه هم هستند. معمولاً پیش‌پردازش متن‌ها منجر به بهبود عملکرد متن‌کاوی و تشخیص تکرار در متن‌هایی می‌شود که پیش‌پردازش روی آن‌ها انجام شده است. تعاریف مختلفی از تکرار براساس متغیرهای n-نگاشت وجود دارد. یک تعریف می‌تواند اشاره به تشابه متغیرهای n-نگاشت ساخته شده از متون داده‌های آزمایش به متغیرهای ساخته شده از متون پاسخ در مشاهدات آموزشی داشته باشد. تعریف دیگر می‌تواند برا ساس پاسخ‌های متنی مشابه باشد. به عنوان مثال پاسخ «تعمیر تایر خودرو» و « تعمیر تایر در خودرو » دو متن یکسان نیستند اما اگر پیش‌پردازش‌هایی روی متن‌ها انجام شود و حروف اضافه مانند «در» یا فاصله خالی آخر عبارت دوم حذف شوند، دو متن رشته‌ای کاملاً شبیه هم هستند. در دیدگاه تکرار براساس n-نگاشت‌ها نیز پس از حذف حروف اضافه، دو متنِ رشته‌ای شامل تک‌نگاشت‌های مشابه «تعمیر»، «تایر» و «خودرو» هستند. در ادامه نحوه تشخیص و انتساب کد به یک متن جدید ارائه می‌شود. در روش متن‌کاوی براساس متغیرهای n-نگاشت، ابتدا تمام متون در دیکشنری (مجموعه‌ای از متون که قبلا به روش دستی کدگذاری شده باشند) به متغیرهای تک‌نگاشت یا چندنگاشت مشابه جدول ۳ تبدیل می‌شوند که در آن فرایند حذف لغات اضافه، تعیین ریشه لغات، حذف نقطه‌گذاری و ویرگول و ...انجام شده است. این پیش‌پردازش‌ها باعث می‌شود تعداد متغیرهای تک‌نگاشت کاهش یابد. جدول ۳ برخی از سطرها و ستون‌های ماتریس TDM داده‌های آموزشی است.

جدول ۳. ماتریس TDM داده‌های آموزشی

متن	کد رشته فعالیت اقتصادی (ISIC)	اسباب	دفتر	باغ	زمین	محصور	مسافر	فروش
اسباب فروشی	۴۷۶۴	۱	۰	۰	۰	۰	۰	۱
دفترمسافربری	۴۹۲۲	۰	۱	۰	۰	۰	۱	۰
باغ محصور	۱۲۱۰	۰	۰	۱	۰	۱	۰	۰
باغ و زمین	۱۲۱۰	۰	۰	۱	۱	۰	۰	۰

در مرحله بعد باید اولین متن (سطر اول) از داده‌های آزمایشی با تمام سطرهای داده‌های آموزشی مقایسه شوند. مقایسه به این صورت است که تمام مقادیر صفر یا یک منتسب به تک‌نگاشت‌های سطر اول داده‌های آزمایش $Var۱_i$، برای $i = ۱, \ldots, \ell$ که در آن ℓ تعداد تک‌نگاشت‌های ساخته شده از

دیکشنری است (جدول ۴) با تمام مقادیر صفر یا یک منتسب به تک‌نگاشت‌های تمام سطرهای داده‌های آموزشی مقایسه شوند. به عنوان مثال، مقایسه تک‌نگاشت‌های سطر اول داده‌های آزمایش، $Var۱_i$، با سطر اول داده‌های آموزش، $Var۲_i$، به صورت $\sum_{i=۱}^{\ell} |Var۱_i - Var۲_i|$ است. در مقایسات هر سطر داده آزمایش با تمام سطرهای داده آموزش، برای متونی از داده‌های آموزش که فرمول بالا مقدار صفر را اخذ کند، کد منتسب به آن متن (C_i) برای متن داده آزمایش در نظر گرفته می‌شود. به عنوان مثال اگر پس از مقایسه سطر دوم ماتریس TDM از داده‌های آزمایشی با تمام سطرهای ماتریس TDM داده‌های آموزشی (جدول ۳)، سطر دوم داده آزمایشی با ۵ سطر داده آموزشی که کد ۴۹۲۲ دارد و ۲ سطر از داده آموزشی که کد ۷۹۱۱ دارد، مشابه باشد، یعنی $\sum_{i=۱}^{\ell} |Var۱_i - Var۲_i| = ۰$، آنگاه در سطر دوم جدول ۵، اعداد ۵ و ۲ به ترتیب برای کدهای ۴۹۲۲ و ۷۹۱۱ درج می‌شود. تعداد تکرارهای آموزشی کد C_i، $m_i(x)$ در نظر گرفته می‌شود و در روش تکرار از آن استفاده می‌شود. فرض کنید C_i برای $i = ۱, \ldots, \ell$، i امین

جدول ۴. ماتریس TDM داده‌های آزمایشی

داده‌های آزمایشی		متغیرهای تک‌نگاشت ساخته شده از دیکشنری									
ردیف	متن	مغازه	اسباب	دفتر	باغ	زمین	محصور	مسافر	فروش	-	گیلاس
۱	مغازه اسباب فروشی	۱	۱	۰	۰	۰	۰	۰	۱	۰	۰
۲	دفترمسافربری	۰	۰	۱	۰	۰	۰	۱	۰	۰	۰
⋮	⋮										
۲۰	باغ بادام										

کد رشته فعالیت اقتصادی در داده‌های دیکشنری، L تعداد کد رشته فعالیت اقتصادی در داده‌های آموزشی یا دیکشنری و $m_i(x)$ تعداد تکرارهای آموزشی کد C_i باشد. احتمال تخصیص کد C_i به رشته فعالیت متنی (x) ثبت شده توسط مأمور براساس روش تکرار به صورت

$$\hat{p}_d(c_i x) = \begin{cases} \frac{m_i(x)}{M(x)} & M(x) > ۰ \\ \frac{۱}{L} & \text{جاهای دیگر} \end{cases}$$

است، که در آن $M(x) = \sum_{i=۱}^{L} m_i(x)$ تعداد کل تکرارها در هر سطر داده آزمایش (تعداد متن‌های مشابه در داده آموزشی با متن موردنظر در داده آزمایش) براساس یک معیار فاصله و تشابه. برای انجام روش تکرار ابتدا ماتریس جدول ۵ ایجاد می‌شود که در آن ستون‌ها نام کد رشته فعالیت اقتصادی یا کد شغل در فایل داده‌های آموزشی و سطرها تعداد رکوردها در داده‌های آزمایش است (مثلاً ۲۰ متن جدید).

اگر تعداد کل تکرارها در هر سطر داده آزمایش، $M(x)$، بیش از یک باشد، احتمال اینکه کد C_i به

جدول ۵. تعداد تکرارهای داده‌های آزمایش با داده‌های آموزشی

ردیف	کد ISIC						تعداد کل تکرارها
	۱۲۴۰	۴۹۲۲	۱۱۹۰	۷۹۱۱	۱۲۱۰ ...	۴۷۶۴	
۱	۰	۰	۰	۰	۰ ...	۲	۲
۲	۰	۵	۰	۲	۰ ...	۰	۷
⋮							
۲۰							

متن سطر اول اختصاص داده شود، برابر تقسیم تعداد تکرارهای آموزشی کد C_i، $m_i(x)$ بر کل تکرارهای سطر اول. در صورتی که $M(x) = ۰$، احتمال اختصاص هر کد به متن مورد نظر برابر $\frac{۱}{L}$ است که در آن L تعداد کد رشته فعالیت اقتصادی در داده‌های آموزشی یا دیکشنری است. براساس نتایج جدول ۵ احتمال انتساب کد ۴۷۶۴ به سطر اول داده آزمایشی برابر ۱ است. همچنین احتمال انتساب کد ۴۹۲۲ به سطر دوم داده آزمایشی برابر $\frac{۵}{۷}$ است.

ب- مدل‌های ترکیبی با سطوح مختلف تجمیع: در این روش با استفاده از روش ماشین بردار پشتیبان به ترکیب مدل حاصل از سطوح مختلف تجمیع مانند ترکیب مدل حاصل از کد فعالیت اقتصادی ۴ رقمی و کد فعالیت اقتصادی ۳ رقمی پرداخته می‌شود. ساختار کدهای مربوط به مشاغل، رشته فعالیت اقتصادی و سایر رده‌بندی‌های بین‌المللی، ساختار سلسله مراتبی دارد. به عنوان مثال کد چهار رقمی مشاغل (گروه واحد) ۷۱۳۱ مربوط به «نقاشان ساختمانی و کارکنان مشاغل مرتبط» یکی از کدهای گروه فرعی (کدهای سه رقمی) ۷۱۳ مربوط به «نقاشان، تمیزکارهای سازه‌های ساختمانی و کارکنان حرفه‌های مرتبط» است. به عبارت دیگر کد سه رقمی ۷۱۳ همه کدهای چهار رقمی مرتبط با این شغل را در بردارد.

یکی از روش‌های یادگیری آماری برای کدگذاری مشاغل یا رشته فعالیت‌ها، استفاده از روش‌های یادگیری آماری برای کدهای مشاغل ۴ رقمی و کدهای ۳ رقمی به طور مجزا و ترکیب احتمال تخصیص هر کد به شغل متنی نوشته شده است. یکی از روش‌هایی که در حال حاضر به صورت گسترده برای مسئله رده‌بندی مورد استفاده قرار می‌گیرد، روش ماشین بردار پشتیبان SVM است (جاشیمز ، ۱۹۹۸). ماشین بردار پشتیبان یکی از روش‌های یادگیری راهنماییده است که از آن برای رده‌بندی و رگرسیون استفاده می‌شود. این روش بسیار مشابه رگرسیون لوژستیک است اما ایده بسیار جذابی در این روش نهفته است. الگوریتم ماشین بردار پشتیبان با کمک یک نگاشت غیرخطی، فضای داده‌های آموزشی را به یک بعد بالاتر تبدیل می‌کند و سپس در این بعد جدید به دنبال ابرصفحه‌ای است که نمونه‌های یک رده را از رده‌های دیگر جدا کند. با یک نگاشت غیرخطی مناسب، مجموعه داده‌های دو رده می‌توانند توسط یک ابرصفحه جدا شوند. الگوریتم های ماشین بردار پشتیبان جهت یافتن ابرصفحه از مفاهیمی چون بردارهای پشتیبان و حاشیه‌ها

استفاده می‌کنند. یک ماشین بردار پشتیبان از طریق یافتن ابرصفحه‌ای که تفاوت بین دو رده را بیشینه کند، رده‌بندی را انجام می‌دهد. بردارهایی که این ابرصفحه را تعریف می‌کنند، بردار پشتیبان نامیده می‌شوند. به عبارت دیگر، ماشین بردار پشتیبان یک رده‌بندی‌کننده دودوئی است که ابتدا کار خود را برای جداکردن داده‌هایی که به صورت خطی قابل تفکیک است، آغاز می‌کند. ایده اصلی این روش در کمینه کردن ریسک ساختاری و تعمیم‌دادن صفحه رده‌بندی‌کننده به صفحه‌ای با بیشترین حاشیه امنیت برای ایجاد تعمیم‌پذیری بالا و کمینه کردن ریسک عدم رده‌بندی صحیح برای نقاطی است که در آینده با آنها روبرو خواهیم شد. بنا بر این، در ساده‌ترین فرم، ماشین بردار پشتیبان عبارت است از یک ابرصفحه که مجموعه نمونه‌های مثبت و منفی را با حداکثر فاصله از هم جدا می‌کند.

چنانچه مجموعه داده‌های آموزشی دارای دومتغیر باشد، می‌توان تصور نمود که داده‌ها در یک فضای دوبعدی قرار دارند و بنابراین به دنبال خطی جهت جداسازی رده‌ها هستیم. هنگامی که فضا سه‌بعدی می‌شود، جداکننده یک صفحه است و بطور عام چون تعداد متغیرها در مجموعه داده‌های اولیه بیش از سه است، از کلمه ابرصفحه برای جداساز میان رده‌ها استفاده می‌شود. بنا بر این ابرصفحه یک مفهوم در هندسه است که تعمیمی از یک صفحه در ابعاد مختلف است. زمانی که با ماشین بردار پشتیبان چندرده‌ای روبرو هستیم (مانند رده‌بندی مشاغل)، دو رویکردِ یک رده در مقابل بقیه رده‌ها و یک رده در مقابل رده دیگر وجود دارد. با توجه به تعداد زیاد کدهای مشاغل، تعداد مشاهدات در سطح کدهای چهار رقمی کم است ولی تعداد مشاهدات در سطح کدهای سه رقمی نسبتاً بیشتر خواهد بود. اگر رده‌بندی یک متن براساس کدهای چهار رقمی منجر به احتمال‌های نزدیک به هم در چند کد چهار رقمی شود، به عنوان مثال احتمال ۰٫۴۱ و ۰٫۴۲ برای کدهای ۱۴۱۰ و ۱۴۱۸، ولی رده‌بندی متن مورد نظر براساس کدهای سه رقمی منجر به کدهای سه رقمی متفاوت شود، میانگین‌گیری احتمال‌ها در سطح کد سه و چهار رقمی باعث می‌شود برابری احتمال‌ها در سطح کد چهار رقمی از بین رود و انتساب یک کد به متن، تعیین تکلیف شود.

فرض کنید C_i برای $i = ۱, \ldots, L$، کدهای چهار رقمی متعلق به کدهای سه رقمی m_j رشته فعالیت اقتصادی یا مشاغل برای $j = ۱, \ldots, \ell$ باشد که در آن L و ℓ به ترتیب تعداد کدهای چهار رقمی و سه رقمی باشند. احتمال‌های حاصل از مدل‌بندی یادگیری آماری برای کدهای سه رقمی و چهار رقمی به ترتیب با $\hat{p}_{\text{۳digit}}(m_j|x)$ و $\hat{p}_{\text{۴digit}}(c_i|x)$ برای یک متن یا رکورد x نمایش داده می‌شود. در این روش، از میانگین هر دو احتمال به صورت

$$\hat{p}_{\frac{\text{۳}}{\text{۴}}\text{digit}}(c_i|x) = \frac{\hat{p}_{\text{۳digit}}(m_j|x) + \hat{p}_{\text{۴digit}}(c_i|x)}{\text{۲}}$$

استفاده می‌شود. این میانگین‌گیری، احتمال‌های مشابه در سطح کد چهار رقمی را از بین می‌برد مگر مواردی که کدهای چهار رقمی دارای احتمال برابر، دارای کدهای سه رقمی مشابه باشد.

ج- روش هیبرید: ترکیب روش تکراری و یادگیری آماری: در این بخش از روش ماشین‌های بردار پشتیبان SVM با هسته خطی به عنوان روش یادگیری آماری استفاده شده است که عملکرد خوبی در رده‌بندی متون نشان داده است. هسته خطی نیازمند یک پارامتر کوک (C) است که تعادلی بین خطای آموزش و پیچیدگی مدل بوجود می‌آورد. در این بخش $C = ۱$ در نظر گرفته شده است. در روش SVM با استفاده از روش پلات (پلات ، ۱۹۹۹)، نمرات به احتمال تبدیل می‌شوند. در روش ترکیبی پیشنهادی جیون و همکاران (۲۰۱۷)، روش مبتنی بر تکرار در داده‌های آموزشی با روش یادگیری آماری SVM ترکیب می‌شوند. فرض کنید $\hat{p}_s(c_i|x)$ احتمال برآوردشده در روش یادگیری آماری و $\hat{p}_d(c_i|x)$ احتمال برآوردشده در روش تکراری باشد. در روش ترکیبی، نمره ترکیبی $\theta(c_i|x)$ به صورت

$$\theta(c_i|x) = \frac{M(x)}{M(x)+۱}\hat{p}_d(c_i|x) + \frac{۱}{M(x)+۱}\hat{p}_s(c_i|x)$$

تعریف می‌شود. اگر تکرار وجود نداشته باشد $(M(x) = ۰)$ ، $\theta(c_i|x)$ احتمال محاسبه شده به روش یادگیری آماری است $\theta(c_i|x) = \hat{p}_s(c_i|x)$. زمانی که تکرار وجود داشته باشد، کدگذاری به روش تکرار مناسب‌تر است و منجر به دقت بالاتر می‌شود. در روش ترکیبی، الگوریتم یادگیری آماری زمانی روی پیش‌بینی تأثیر دقیق‌تر می‌گذارد که $\hat{p}_d(c_i|x)$ برای کدهای مختلف برابر باشد. در این حالت به $\hat{p}_s(c_i|x)$ وزنی معادل یک تکرار اختصاص داده می‌شود که تأثیر آن را کاهش می‌دهد.

د- روش نزدیکترین همسایه تعدیل یافته: روش نزدیکترین همسایه NN (فیکس و هاج ، ۱۹۵۱) ، روش دیگری است که در کدگذاری مشاغل، رشته فعالیت‌ها یا هر کدگذاری استاندارد دیگری استفاده می‌شود. در روش رده‌بندی NN ، یک مشاهده در داده‌های آموزشی با نزدیکترین فاصله به رکورد جدید در داده‌های آزمایشی انتخاب و منتسب می‌شود. روش‌های نزدیک‌ترین همسایه چندگانه نیز توسط یو (۲۰۰۲) پیشنهاد شده است. روش نزدیک‌ترین همسایه به عنوان تعمیمی از روش تکراری است به گونه‌ای که روش تکرار، نزدیک‌ترین همسایه با فاصله صفر است. برای تعریف نزدیکی، لازم است یک معیار فاصله یا یک معیار تشابه معرفی شود که برای رده‌بندی متون و نوشته‌ها، معیار تشابه کسینوسی بسیار استفاده

می‌شود (کناس ، ۱۹۸۷؛ میترا و راملر ، ۲۰۱۰). معیار تشابه بین دو بردار u و v به صورت

$$\text{Cosine}(u,v) = \frac{u.v}{|u||v|} = \frac{\sum u_i v_j}{\sqrt{\sum u_i^۲ \sum v_j^۲}}$$

تعریف می‌شود، که در آن بردارهای u و v نشان‌دهنده حضور یا عدم حضور n-نگاشت‌ها در متن است. همچنین u مربوط به هر سطر یا متن داده‌های آزمایشی و v مربوط به هر سطر در داده آموزشی است. برای انتساب کد صحیح به یک متن در داده آزمایش لازم است معیار تشابه متن داده آزمایش با تک تک متن‌های داده آموزش به دست آید و کد متنی که تشابه بیشتر داشته باشد، به متن داده آزمایش منتسب شود.

مشابه قبل، ممکن است بخواهیم فقط کد متن‌های آسان‌کدگذاری را کدگذاری کنیم و برای متون سخت، کدگذاری به روش دستی باشد. بنابراین، پیشنهاد می‌شود مواردی که با احتمال یا دقت بالاتری به روش NN پیش‌بینی شود، به کدها منتسب شود. برای یک متن جدید x ، فرض کنید $K(x)$ تعداد نزدیکترین همسایگان در داده‌های آموزشی و $s(x) = \text{Cosine}$ معیار تشابه نزدیکترین همسایه باشد. وقتی چند مشاهده نزدیک‌ترین همسایه وجود دارد اغلب $K(x) > ۱$ است. فرض کنید $k_i(x)$ تا از $K(x)$ رکورد، کدهای c_i را اخذ کنند. مشابه روش تکرار، احتمال تخصیص کد c_i به یک متن جدید به روش نزدیکترین همسایه به صورت $\hat{p}_{NN}(c_i|x) = \frac{k_i(x)}{K(x)}$ برآورد می‌شود و نمره برای هر پاسخ متنی به صورت

$$\gamma(c_i|x) = \hat{p}_{NN}(c_i|x)s(x)\frac{K(x)}{K(x)+ ۰/۱}, \quad i = ۱,\ldots,L,$$

تعریف می‌شود. کد پیش‌بینی‌شده تنها بستگی به $\hat{p}_{NN}(c_i|x)$ دارد زیرا $s(x)$ و $K(x)$ برای هر پاسخ متنی ثابت است. بنا بر این نقش $s(x)$ و $\frac{K(x)}{K(x)+ ۰/۱}$ برای ترتیب دادن به مشاهدات براساس اینکه پاسخ‌های ساده‌تر، نمره بالاتری بگیرند، است. ضریب $s(x)$ این مفهوم را تداعی می‌کند که شباهت بیشتر یک متن جدید و نزدیکترین همسایگی آن منجر به رده‌بندی دقیقتر می‌شود. آخرین عبارت در رابطه بالا نیز این مفهوم را منعکس می‌کند که اگر همه چیز برابر باشد، رده‌بندی بر اساس تعداد بیشتری از همسایگان نزدیک، احتمالا نسبت به تعداد کمتر همسایگان نزدیک، دقیقتر باشد. عبارت $\frac{K(x)}{K(x)+ ۰/۱}$ زمانی که $K(x) = ۱$ باشد برابر ۰/۹۱ است و زمانی که $K(x)$ افزیش یابد، به یک همگرا است. در ادامه، این روش در یک مثال محاسبه می‌شود. فرض کنید با متن جدیدی به صورت «مغازه اسباب‌بازی فروشی» روبرو باشیم. این متن شامل سه متغیر تک‌نگاشتی (ریشه‌یابی شده) « مغازه»، «اسباب‌بازی»

و «فروش» است. فرض کنید هیچ تکراری از این عبارت در داده‌های آموزشی وجود نداشته باشد ولی چهار رکورد در داده‌های آموزشی وجود داشته باشد که شامل یکی از سه متغیر تک‌نگاشتی باشد. جدول ۶ نشان می‌دهد که سه متن از چهار متن داده آموزشی پاسخ «فروش» با کد ۴۵۱۰ و یک متن از چهار متن داده آموزشی شامل پاسخ «اسباب‌بازی» با کد ۴۷۶۴ است. معیار تشابه کسینوسی $s(x)$ بین متن داده آزمایشی و به عنوان مثال رکورد داده آموزشی با کد ۴۵۱۰ برابر است با

$$s(x) = \frac{۱ \times ۰ + ۱ \times ۰ + ۱ \times ۱}{\sqrt{۱ + ۱ + ۱} \times \sqrt{۰ + ۰ + ۱}} = ۰٫۵۷۷۴$$

در این مثال، تعداد نزدیکترین همسایگان در داده‌های آموزشی $K(x) = ۴$ است. بنابراین $\frac{K(x)}{K(x)+۰٫۱} = \frac{۴}{۴+۰٫۱} = ۰٫۹۷۵۶$. مقادیر نمره γ برای دو کد فعالیت ۴۵۱۰ و ۴۷۶۴ به دلیل برآورد احتمال‌های مختلف دو کد فعالیت $\hat{p}_{NN}(c_{۴۵۱۰}|x) = \frac{۳}{۴}$ و $\hat{p}_{NN}(c_{۴۷۶۴}|x) = \frac{۱}{۴}$ است. متن داده آزمایش کد ۴۵۱۰ را اخذ می‌کند زیرا این کد بالاترین نمره $\gamma = ۰٫۴۲۲۵$ را دارد.

جدول ۶. محاسبه $(c_i|x)$. مقدار متغیرهای تک‌نگاشت در صورت وجود لغت در متن ۱ و در بقیه موارد صفر است.

	متغیرهای تک‌نگاشت								
رکورد	مغازه	اسباب‌بازی	فروش	کد ISIC	$\hat{p}_{NN}(c_i\|x)$	$s(x)$	$\frac{K(x)}{K(x)+۰٫۱}$	$\gamma(c_i\|x)$
متن ۱ از داده آموزشی	۰	۰	۱					
متن ۲ از داده آموزشی	۰	۰	۱	۴۵۱۰	۰٫۷۵	۰٫۵۷۷۴	۰٫۹۷۶	۰٫۴۲۲
متن ۳ از داده آموزشی	۰	۰	۱					
متن ۴ از داده آموزشی	۰	۱	۱	۴۷۶۴	۰٫۲۵	۰٫۵۷۷	۰٫۹۷۶	۰٫۴۲۲
متن داده آزمایشی	۱	۱	۱	$\hat{c}_i = ۴۵۱۰$				

۴ کاربرد یادگیری آماری در فعالیت‌های مرکز آمار ایران

تحلیل و پردازش احساسات و عقاید مردم در نظرسنجی‌ها، تعیین بار معنایی مثبت یا منفی پرسش‌های باز در آمارگیری‌های نمونه‌ای و رده‌بندی پرسش‌های باز از جمله کاربردهای روش‌های یادگیری آماری با استفاده از متن‌کاوی در آمار رسمی است. از آنجا که روش‌های سنتی یا دستی کدگذاری بسیار زمان‌بر، هزینه‌بر و نیاز به دانش حرفه‌ای در این زمینه دارد و از طرفی فرایند کدگذاری توسط افراد مختلف با مهارت‌های مختلف انجام می‌شود، ضرورت استفاده از روش‌های خودکار یا نیمه‌خودکار کدگذاری متون و پرسش‌های باز در فرایند تولید داده بسیار حائز اهمیت است. در ادامه چند مثال از کاربردهای متن‌کاوی در فرایند تولید

داده با استفاده از روش‌های یادگیری آماری رده‌بندی معرفی می‌شود.

مثال ۱. تشخیص خودکار یا نیمه‌خودکار واجد شرایط بودن یا نبودن کارگاه‌ها از متن‌کاوی پرسش باز «سایر دلایل عدم تکمیل پرسشنامه: برای انجام کار و استفاده از روش‌های یادگیری آماری لازم است دیکشنری‌ای براساس کدگذاری دستی دست‌نوشته‌های مأموران از چند آمارگیری قبلی از طریق اختصاص یکی از دو کد واجد شرایط و غیرواجد شرایط به هر متن صورت پذیرد. سپس دیکشنری به ماتریس TDM تبدیل شود. با استفاده از روش‌های یادگیری آماری و تبدیل داده‌ها به داده‌های آموزشی و آزمایشی، امکان برآورد و اختصاص کد به هر متن جدید به صورت خودکار به‌وجود می‌آید.

مثال ۲. تشخیص بار مثبت یا منفی جملات یا پرسش‌های باز: برای انجام کار و استفاده از روش‌های یادگیری آماری لازم است مشابه مثال قبل، دیکشنری‌ای براساس متون مختلف یا دست‌نوشته‌های آمارگیری‌های مختلف تهیه شود به‌گونه‌ای که به هر متن براساس کلمات مثبت یا منفی، کد مثبت یا منفی به صورت دستی اختصاص داده شود. تشکیل ماتریس TDM ، تفکیک ماتریس به داده‌های آموزشی و آزمایشی، استفاده از روش‌های یادگیری آماری ماشین بردار پشتیبان SVM روش‌های نزدیک‌ترین همسایه و غیره روی داده‌های آموزشی، ارزیابی براساس داده‌های آزمایش و پیش‌بینی رده مناسب برای هر متن، فرایندهای مورد نیاز برای تبدیل یک متن به یک رده از قبل تعیین شده با استفاده از روش‌های یادگیری آماری است.

مثال ۳. تخصیص کد صحیح ISIC یا ISCO یا هر کد دیگر به پرسش‌های باز به صورت خودکار. روش کار تشکیل یک دیکشنری جامع و کامل از طریق کدگذاری دستی کتابچه‌های رده‌بندی‌های بین‌المللی مانند رده‌بندی رشته فعالیت‌های اقتصادی و دست‌نوشته‌های مأموران آمارگیری از چند آمارگیری قبلی و انجام فرایندهای ذکر شده در مثال ۲. این رویکرد کدگذاری همانطور که قبلاً بیان شد، در اداره آمار استرالیا (کلارک و بروکر ، ۲۰۱۱) ، ایالات متحده (کلارک و بروکر ، ۲۰۱۱) و آلمان (بشمان و همکاران ، ۲۰۱۴) استفاده شده است.

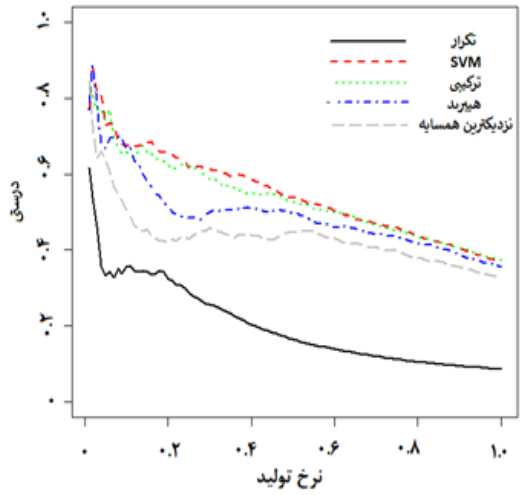
۴.۱ کدگذاری نیمه‌خودکار رشته فعالیت‌های اقتصادی

برای کدگذاری رشته فعالیت اقتصادی، دیکشنری شامل ۱۳۲۶۰ رشته فعالیت کدگذاری شده از دست‌نوشته‌های ماموران در چند دوره آمارگیری تهیه شد و پس از انجام برخی پیش‌پردازها، داده‌ها با نرخ ۰/۹ و ۰/۱ به داده‌های آموزش و آزمایش تقسیم شده است. سپس ۵ روش کدگذاری خودکار شامل تکرار، ماشین بردار پشتیبان برای کد ۴ رقمی $ISIC$ ، روش ترکیبی ماشین بردار پشتیبان (ترکیب مدل‌ها از سطوح مختلف

تجمیع در کد ۳ رقمی و ۴ رقمی)، روش هیبرید (ترکیب روش تکرار و ماشین بردار پشتیبان) و روش نزدیکترین همسایه با آموزش روی ۰٫۹ داده‌ها و اجرا روی ۰٫۱ داده‌های آزمایش انجام شده است. نتایج میزان دقت برای نرخ‌های تولید مشخص به تفکیک پنج روش در شکل ۱ نمایش داده شده است. همانطور که مشخص است روش ماشین بردار پشتیبان و روش ترکیبی ماشین بردار پشتیبان برای کدهای ۳ و ۴ رقمی رشته فعالیت‌های اقتصادی بهترین نتیجه را داده است. هر چه دیکشنری تهیه‌شده جامع‌تر باشد و همه مصادیق را پوشش دهد، همچنین پیش‌پردازهای کافی روی داده‌ها صورت پذیرد، نتایج اختصاص کد به متون با دقت بالاتری صورت می‌گیرد. جدول ۷ میزان درستی پنج روش برای داده‌های آزمایشی رشته فعالیت اقتصادی کارگاه‌های صنعتی به تفکیک احتمال انتساب کد ISIC به رشته فعالیت اقتصادی را نشان می‌دهد. گروههای ۱ تا ۶ به ترتیب احتمال انتساب ۰٫۲۹ – ۰٫۰۰، ۰٫۳۹ – ۰٫۳۰، ۰٫۴۹ – ۰٫۴۰، ۰٫۵۹ – ۰٫۵۰، ۰٫۶۹ – ۰٫۶۰ و ۱٫۰۰ – ۰٫۷۰ را نشان می‌دهد. نتایج شکل ۱ و جدول ۷ نشان می‌دهد که روش تکرار برای این نوع داده‌ها، روش مناسبی نیست. علت آن نیز این است که متون مربوط به کدهای مختلف ISIC شبیه هم نیست که روش تکرار بتواند خوب جواب دهد. میزان درستی چهار روش دیگر غیر از تکرار برای روش نزدیکترین همسایه، ۳۳ درصد و برای سه روش دیگر حدود ۴۰ درصد است. تنها دلیل پایین بودن میزان درستی، این است که دیکشنری تهیه شده برای آموزش باید کامل‌تر شود و مرکز امار ایران باید دست‌نوشته‌های مختلف ماموران آمارگیری را به این دیکشنری اضافه کنند تا در مرحله آموزش، رایانه خوب آموزش داده شود. در این مقاله دیکشنری ساخته شده شامل دست نوشته‌های ۴ سال ماموران آمارگیری بوده است. نتایج روش ماشین بردار پشتیبان در جدول ۷ نشان می‌دهد از ۱۳۲۶ داده آزمایشی، ۴۸۷ انطباق درست از طریق این روش اتفاق افتاده است. این اعداد منجر به میزان درستی این روش معادل ۰٫۳۷ شده است. همچنین نتایج این جدول نشان می‌دهد برای ۳۶۱ متن نوشته شده در داده آزمایشی با احتمال انتساب بین ۰٫۳۹ – ۰٫۳۰ به کد ISIC مشخص، ۴۰ متن به درستی به کد ISIC منتسب شده است. همچنین برای ۱۲۵ متن نوشته شده در داده آزمایشی با احتمال انتساب بین ۱٫۰۰ – ۰٫۷۰ به کد ISIC مشخص، ۸۴ متن به درستی به کد ISIC منتسب شده است.

۴.۲ تشخیص خودکار واجد شرایط بودن کارگاه‌ها از پرسش‌های باز

در پرسش‌نامه‌های مراکز آماری برای برخی از پرسش‌ها، رده سایر با ذکر نام وجود دارد. به عنوان مثال در پرسش‌نامه طرح صنعت مرکز آمار ایران، یکی از پرسش‌ها، نحوه تکمیل پرسشنامه است که شامل یک گزینه تکمیل و عدم تکمیل است. برای پرسش‌نامه‌هایی که تکمیل نشده است، علت عدم تکمیل پرسشنامه



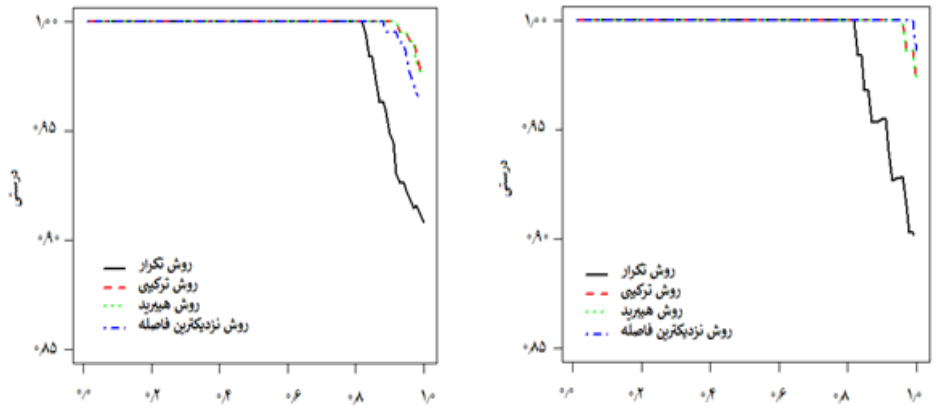
شکل ۱. میزان دقت نرخ‌های تولید مشخص به تفکیک روش های کدگذاری رشته فعالیت اقتصادی

جدول ۷. تعداد تکرارهای داده‌های آزمایش با داده‌های آموزشی میزان درستی پنج روش برای داده‌های آزمایشی رشته فعالیت اقتصادی کارگاه‌های صنعتی

		احتمال انتساب (کد ISIC به رشته فعالیت اقتصادی)						
روش		کل	گروه ۱	گروه ۲	گروه ۳	گروه ۴	گروه ۵	گروه ۶
تکرار	کل	۱۳۲۶	۹۱۴	۵۹	۷	۱۱۱	۷	۲۲۸
	انطباق درست	۱۱۳	۹	۶	۲	۱۷	۳	۷۶
	درستی	۰٫۰۹	۰٫۰۰	۰٫۱	۰٫۲۹	۰٫۱۵	۰٫۴۳	۰٫۳۳
ترکیبی	کل	۱۳۲۶		۳۰۰	۵۴۰	۲۵۳	۱۱۱	۱۲۲
	انطباق درست	۴۹۴	–	۳۹	۱۸۲	۱۲۵	۷۰	۷۸
	درستی	۰٫۳۸		۰٫۱۳	۰٫۳۴	۰٫۴۹	۰٫۶۳	۰٫۶۴
ماشین بردار پشتیبان	کل	۱۳۲۶		۳۶۱	۵۸۶	۱۵۲	۱۰۲	۱۲۵
	انطباق درست	۴۸۷	–	۴۰	۲۱۳	۸۳	۶۷	۸۴
	درستی	۰٫۳۷		۰٫۱۱	۰٫۳۶	۰٫۵۵	۰٫۶۶	۰٫۶۷
هیبرید	کل	۱۳۲۶		۳۴۲	۴۸۴	۱۲۸	۱۳۹	۲۳۳
	انطباق درست	۴۷۲	–	۴۳	۱۷۴	۷۶	۵۳	۱۲۶
	درستی	۰٫۳۶		۰٫۱۳	۰٫۳۶	۰٫۵۹	۰٫۳۸	۰٫۵۴
نزدیکترین همسایه	کل	۱۳۲۶	۲۱۲	۱۳۵	۲۷	۲۷۰	۴۶	۶۳۶
	انطباق درست	۴۳۱	۲۰	۱۹	۱۳	۷۴	۲۳	۲۸۲
	درستی	۰٫۳۳	۰٫۰۹	۰٫۱۴	۰٫۴۸	۰٫۲۷	۰٫۵۰	۰٫۴۴

شامل سه گزینه عدم دسترسی به پاسخگو، عدم همکاری پاسخگو، عدم دسترسی به کارگاه است. از طرفی دلایل عدم دسترسی به کارگاه شامل گزینه‌های نقص اطلاعات آدرسی، تغییر مکان کارگاه، تغییر کاربری محل کارگاه، تعطیل دائم، تخریب و ساخت و ساز، تکراری و سایر با ذکر علت است.

با توجه به اینکه رده «سایر با ذکر علت» نیاز به بررسی و درج دو کد واجد شرایط (۱) و غیر واجد شرایط (۲) به منظور جانهی واحدهای واجدشرایط برای اصلاح وزن‌های نمونه‌گیری دارد، تاکنون این کار به صورت دستی در مرکز آمار ایران انجام می‌شده است که زمان‌بر است. با استفاده از روش‌های یادگیری آماری، امکان اختصاص کد به هر متن نوشته شده در سایر به صورت خودکار وجود دارد. روش کار به این صورت است که با ساخت یک دیکشنری براساس کدگذاری دست‌نوشته‌های (اختصاص یکی از دو کد واجد شرایط (۱) و غیر واجد شرایط (۲) به هر متن) کارشناسان کدگذاری برای چند سری از داده‌های آمارگیری از کارگاه‌های صنعتی به صورت دستی و استفاده از روش‌های یادگیری آماری، امکان برآورد و اختصاص کد به هر متن جدید به صورت خودکار فراهم می‌شود. بنا براین برای انجام این کار، مشابه مثال کدگذاری $ISIC$ ، دیکشنری‌ای توسط نویسندگان مقاله ایجاد شد. تعداد رکوردهای فایل دیکشنری شامل ۷۲۵ متن از دلایل عدم دسترسی است که به صورت دستی کدگذاری شده است. با تشکیل ماتریس TDM ، چهار الگوریتم معرفی شده در زیربخش ۲-۱ روی داده‌های آموزشی و آزمایشی با دو نرخ ۰/۹ و ۰/۷ اجرا شده است. نتایج درستی رده‌بندی روی داده‌های آزمایش برای چهار روش معرفی‌شده و نرخ‌های تولید مختلف در شکل ۲ ارائه شده است. همان‌طور که ملاحظه می‌شود با بیش از ۹۵ درصد درستی، کدگذاری متون



شکل ۲. میزان دقت برای نرخ‌های تولید مشخص به چهار روش الف- نرخ داده آموزشی ۰/۹، ب- نرخ داده آموزشی ۰/۷

به صورت خودکار انجام شده است. مسلماً چون دیکشنری تهیه‌شده در این مثال، جامع بوده است و اکثر مصادیق را پوشش داده است، نتایج اختصاص کد به متون با دقت بالاتری صورت گرفته است. همچنین نتایج شکل ۲ و جداول ۱۱ و ۱۲ نشان می‌دهد با نرخ داده‌های آموزشی ۰/۹ و ۰/۷ تفاوتی در میزان دقت

روش‌ها برای نرخ‌های مختلف داده‌های آموزشی وجود ندارد و پایداری نتایج نسبت به حجم نمونه‌های مختلف داده‌های آموزشی دیده می‌شود. به منظور بررسی تاثیر تعداد یا درصد نمونه داده‌های آموزشی، میزان تکراری بودن متون و همچنین تاثیر میزان تمیز بودن داده‌ها، می‌توان در مطالعات بعدی شبیه‌سازی‌هایی در این زمینه انجام داد و میزان درستی نتایج را استخراج کرد. لازم به ذکر است برای ارزیابی عملکرد الگوریتم‌های یادگیری آماری راهنماییده از ماتریس درهم‌ریختگی استفاده می‌شود. دو شاخص نرخ مثبت درست و نرخ منفی درست برای ارزیابی نتیجه رده‌بندی پاسخ‌های دوحالتی استفاده می‌شود. شاخص‌های

$$\text{Sensitvity} = \frac{TP}{TP+FN} = ۱ - \beta$$
$$\text{Specificity} = \frac{TN}{TN+FP} = ۱ - \alpha$$
$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = ۱ - \beta$$
$$\text{Precision} = \frac{TP}{TP+FP}$$
$$\text{Negative Predictive Value} = \frac{TN}{TN+FN}$$
$$\text{Prevalence} = \frac{TP+FN}{TP+TN+FP+FN}$$
$$\text{Detection Rate} = \frac{TP}{TP+TN+FP+FN}$$
$$\text{Detection Prevalence} = \frac{TP+FP}{TTP+TN+FP+FN}$$
$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{۲}$$

از ماتریس درهم‌ریختگی جدول ۸ به دست می‌آیند، که در آنها α و β به ترتیب خطای نوع اول و دوم هستند. برای ارزیابی عملکرد الگوریتم‌های یادگیری آماری راهنماییده با بیش از دو رده نیز از ماتریس

جدول ۸. نمونه‌ای از ماتریس درهم‌ریختگی یک ماتریس ۲×۲.

	مقادیر واقعی	
پیش‌بینی	مثبت (P)	منفی (N)
درست (T)	TP	FP
غلط (F)	FN	TN

درهم‌ریختگی استفاده می‌شود. هر ستون از ماتریس، مقادیر واقعی را در بردارد در صورتی که هر سطر، مقادیر پیش‌بینی را نشان می‌دهد. همانطورکه برای جدول ۸ بیان شد، می‌توان از شاخص‌های ارزیابی مانند دقت[1] و نرخ مثبت درست یا یادآوری[2] استفاده کرد. یادآوری عبارت است از اینکه «از کل مقادیر واقعی رده موردنظر، میزانی که به درستی رده موردنظر را پیش‌بینی کرده است» و دقت بر این مفهوم که «از کل مقادیر پیش‌بینی‌شده رده موردنظر، میزانی که به درستی رده موردنظر را پیش‌بینی کرده است» دلالت دارند. درستی[3] بر میزان نمونه‌هایی اشاره دارد که سیستم یادگیری در تشخیص آن‌ها موفق بوده است. شاخص‌های

جدول ۹. جدول توافقی مقادیر واقعی و مقادیر پیش‌بینی

	مقادیر واقعی		
پیش‌بینی	۱	۲	۳
۱	$n_{۱۱}$	$n_{۱۲}$	$n_{۱۳}$
۲	$n_{۲۱}$	$n_{۲۲}$	$n_{۲۳}$
۳	$n_{۳۱}$	$n_{۳۲}$	$n_{۳۳}$

دیگری مانند حساسیت[4] به عنوان مثال در یک مسئله با ۳ رده مشابه جدول ۹، درستی، یادآوری و دقت به صورت

$$\text{Accuracy} = \frac{n_{۱۱} + n_{۲۲} + n_{۳۳}}{N}$$
$$\text{Precision of first class} = \frac{n_{۱۱}}{n_{۱۱} + n_{۱۲} + n_{۱۳}}$$
$$\text{Recall of first class} = \frac{n_{۱۱}}{n_{۱۱} + n_{۲۱} + n_{۳۱}}$$

محاسبه می‌شود. در مثال کاربردی داده‌های تشخیص واجد شرایط بودن کارگاه‌های صنعتی نتایج به شرح زیر است. جدول ۱۰ ماتریس‌های در هم‌ریختگی مربوط به نرخ داده آموزشی ۰٫۹ به ترتیب برای روش‌های تکرار، ترکیبی، هیبرید و نزدیک‌ترین همسایه را نشان می‌دهد. جدول ۱۱ نیز به مقایسه شاخص‌های ارزیابی عملکرد الگوریتم‌های مختلف با نرخ داده آموزشی ۰٫۹ پرداخته است. نتایج نشان می دهد درستی روش تکرار ۹۳ درصد و درستی سه روش دیگر بیش از ۹۸ درصد است. جدول ۱۲ نیز به مقایسه شاخص‌های ارزیابی عملکرد الگوریتم‌های مختلف با نرخ داده آموزشی ۰٫۷ پرداخته است. نتایج نشان می‌دهد برای نرخ

[1] Precision
[2] Recall
[3] Accuracy
[4] Sensitivity

داده آموزشی ۰٫۷ درستی روش تکرار ۹۲ درصد و درستی سه روش دیگر بیش از ۹۸ درصد است.

جدول ۱۰. ماتریس های در هم‌ریختگی با روش‌های مختلف

روش	پیش‌بینی	مقادیر واقعی ۱	مقادیر واقعی ۲
تکرار	۱	۵۴	۲
	۲	۳	۱۴
ترکیبی، هیبرید و نزدیک‌ترین همسایه	۱	۵۵	۱
	۲	۰	۱۷

جدول ۱۱. مقایسه شاخص‌های ارزیابی عملکرد الگوریتم‌های مختلف با نرخ داده آموزشی ۰٫۹

شاخص	روش تکراری	روش‌های ترکیبی، هیبرید و نزدیک‌ترین همسایه
Accuracy	۰٫۹۳۱	۰٫۹۸۶
NIR	۰٫۷۸۱	۰٫۷۵۳
$P\_Value[Acc > NIR]$	۰٫۰۰۰	۰٫۰۰۰
Kappa	۰٫۸۰۴	۰٫۹۶۲
Mcnemar's Test P_Value	۱٫۰۰۰	۱٫۰۰۰
Sensitvity	۰٫۹۴۷	۱٫۰۰۰
Specificity	۰٫۸۷۵	۰٫۹۴۴
Precision	۰٫۹۶۴	۰٫۹۸۲
Value Pred. Negative	۰٫۸۲۳	۱٫۰۰۰
Prevalence	۰٫۷۸۱	۰٫۷۵۳
Rate Detection	۰٫۷۴۰	۰٫۷۵۳
Prevalence Detection	۰٫۷۶۷	۰٫۷۶۷
Accuracy Balanced	۰٫۹۱۱	۰٫۹۷۲

جدول ۱۲. مقایسه شاخص‌های ارزیابی عملکرد الگوریتم‌های مختلف با نرخ داده آموزشی ۰٫۷

شاخص	روش تکراری	روش‌های ترکیبی و هیبرید	روش نزدیک‌ترین همسایه
Accuracy	۰٫۹۲۷	۰٫۹۸۶	۰٫۹۸۲
NIR	۰٫۸۱۶	۰٫۷۶۰	۰٫۷۷۰۶
$P\_Value[Acc > NIR]$	۰٫۰۰۰	۰٫۰۰۰	۰٫۰۰۰
Kappa	۰٫۷۷۷	۰٫۹۶۲	۰٫۹۴۸
Mcnemar's Test P_Value	۰٫۲۴۴	۰٫۲۴۸	۱٫۰۰۰
Sensitvity	۰٫۹۲۷	۱٫۰۰۰	۰٫۹۸۸۱
Specificity	۰٫۷۴۰	۱٫۰۰۰	۰٫۹۶۰۰
Precision	۰٫۹۸۲	۰٫۹۸۲	۰٫۹۸۸۱
Value Pred. Negative	۰٫۷۴۰	۱٫۰۰۰	۰٫۹۶۰۰
Prevalence	۰٫۸۱۶	۰٫۷۵۷	۰٫۷۷۱
Rate Detection	۰٫۷۵۷	۰٫۷۵۷	۰٫۷۶۱
Prevalence Detection	۰٫۷۷۱	۰٫۷۷۱	۰٫۷۷۱
Accuracy Balanced	۰٫۹۲۶۰	۰٫۹۷۲	۰٫۹۷۴

۵ بحث و نتیجه‌گیری

در این مقاله با استفاده از چهار روش یادگیری آماری، دو کاربرد از روش‌های یادگیری آماری در متن‌کاوی شامل کدگذاری خودکار رشته فعالیت‌های اقتصادی و تخصیص کد واجد شرایط بودن یا نبودن به پرسش باز عدم تکمیل پرسشنامه انجام شده است. تحلیل‌های این بخش با استفاده از نرم‌افزار R و بسته $e۱۰۷۱$ انجام شده است. با توجه به اینکه حجم داده‌های آمار رسمی زیاد است و اجرای روش‌های یادگیری آماری در نرم‌افزار R برای داده‌های حجیم، امکان‌پذیر نیست، با برنامه‌نویسی در نرم‌افزار SAS امکان اجرای این روش‌ها در مراکز آماری نیز قابل اجرا است. با توجه به ضرورت مدرن‌سازی نظام آماری کشورها، از آن‌جا که استفاده از روش‌های یادگیری آماری در مدل عمومی فرایند کسب و کار آماری هم در داده‌های اولیه و هم در داده‌های ثانویه کاربرد دارد، امید است مرکز آمار ایران از روش‌های یادگیری آماری در هر یک از مراحل مدل عمومی فرایند کسب و کار آماری استفاده کند. همچنین در کارهای آتی، ساخت ماتریس عبارت در متن می‌تواند به گونه‌ای باشد که اعداد داخل ماتریس TDM، W_{ij} (وزن کلمه j در متن i) یا شمارش فراوانی کلماتی که در یک متن وجود دارد باشد. این رویکرد ممکن است عملکرد بهتری نیز داشته باشد.

تقدیر و تشکر

بدینوسیله از حمایت مالی پژوهشکده آمار در طرح تحقیقاتی که مقاله حاضر بخشی از نتایج آن طرح است، کمال تشکر و قدردانی را داریم. از داوران و اعضای محترم هیئت تحریریه مجله نیز به خاطر پیشنهادهای ارزنده‌ای که ارائه نمودند کمال تشکر را داریم.

مراجع

رضائی قهرودی، ز.، رنجی، ح. و رضایی، ع. (۱۳۹۹)، یادگیری آماری و کاربردهای آن در آمار رسمی. پژوهشکده آمار.

برفی پور، آ.، قادری، س. (۱۳۹۵)، مدل عمومی فرایند کسب و کار آماری. مجله بررسی های آمار رسمی ایران، ۸۸، ۵۳-۸۲.

Bethmann, A., Schierholz, M. Wenzig, K. and Zielonka, M. (2014), Auto-

matic Coding of Occupations, *In Proceedings of Statistics Canada Symposium*, August 29–31, 2014, Que´bec, Canada, http://www.statcan.gc.ca/sites/default/files/media/14291-eng.pdf.

Clarke, F. R. and S. J. Brooker. (2011), Use of Machine Learning for Automated Survey Coding, *In Proceedings of the 58th ISI World Statistics Congress*, August 21–26, Dublin, Ireland.

Creecy, R. H., B. M. Masand, S. J. Smith, and D. L. Waltz. (1992), Trading MIPS and Memory for Knowledge Engineering, *Communications of the ACM*, **35**: 48–64.

Day, J. (2014), Using an Autocoder to Code Industry and Occupation in the American Community Survey, Presentation for the Federal Economic Statistics Advisory Committee Meeting, http://www2.census.gov/adrm/fesac/2014-06-13_day.pdf.

Fix, E. and J. L. Hodges (1951), Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties, Technical Report, USAF School of Aviation Medivine, Randolph Field, Texas. Project 21-49-004, Rept. 4, Contract AF41(128)-31.

Friedman, J. H. (2001), Greedy Function Approximation: A Gradient Boosting Machine, *The Annals of Statistics*, **29**, 1189–1232.

Gweon H., Schonlau M., Steiner S., Kaczmirek L. and Blohm M. (2017), Three Methods for Occupation Coding Based on Statistical Learning, *Journal of Official Statistics*, **33**, 1, 101–122.

Harron, K., Dibben, C. and Goldstein, H. (2015), *Methodological Developments in Data Linkage*, London: Wiley.

Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, M. and Goldstein, H. (2017), Challenges in Administrative Data Linkage for Research, *Big Data & Society*, **4**, https://doi.org/10.1177/2053951717745678.

Hastie T., Tibshirani R., Friedman J. H. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed, New York: Springer.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2014), *An Introduction to Statistical Learning with Applications in R*, Springer.

Joachims, T. (1998), Text Categorization with Support Vector Machines: Learning with Many Relevant Features, *In Proceedings of the 10th European Conference on Machine Learning*, 21–23, Chemnitz, Germany, 137–142, https://doi.org/10.1007/BFb0026683.

Knaus, R. (1987), Methods and Problems in Coding Natural Language Survey Data, *Journal of Official Statistics*, **3**, 45–67.

Lohr, S. L. and Raghunathan, T. E. (2017), Combining Survey Data with Other Data Sources, *Statistical Science*, **32**, 293–312.

Maitra, R. and Ramler, I. P. (2010), A K-mean-Directions Algorithm for Fast Clustering of Data on the Sphere, *Journal of Computational and Graphical Statistics*, **19**, 377–396.

Platt, J. (1999), Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *In Advances in Large Margin Classifiers*, edited by A. J. Smola, P. Bartlett, B. Scho¨lkopf, and D. Schuurmans, 61–74. Cambridge, Massachusetts: MIT Press.

Russ, D. E., Ho, K. Y., Johnson C. A., and Fries en M. C. (2014), Computer-Based Coding of Occupation Codes for Epidemiological Analyses, *In Proceedings of the 27th IEEE International Symposium on Computer-Based Medical Systems*, May 27–29, New York, USA, 347–350, https://doi.org/10.1109/CBMS.2014.79.

Schierholz, M. (2014), Automating Survey Coding for Occupation, Master's thesis, Ludwig-Maximilians-Universita¨t Munich, http://doku.iab.de/fdz/reporte/2014/MR_10-14_EN.pdf.

Statistics Netherlands. Hague/Heerlen (2012), Method Series Theme: Coding; Interpreting Short Descriptions Using a Classification.

Thompson, M., Kornbau, M. E. and Vesely, J. (2012), Creating an Automated Industry and Occupation Coding Process for the American Community Survey, http://ftp.census.gov/adrm/fesac/2014-06-13-thompson-kornbau-vesely.pdf.

UNECE Machine Learning Team. (2018), The Use of Machine Learning in Official Statistics.

Vapnik, V.N. (2000), *The Nature of Statistical Learning Theory*, 2nd edition, New York: Springer.

Yu, C. (2002), *High-Dimensional Indexing: Transformational Approaches to High-Dimensional Range and Similarity Searches*, 2341, Berlin: Springer, https://doi.org/10.1007/3-540-45770-4.

Journal of Statistical Sciences, Spring and Summer, 2021
Vol. 15, No. 1, pp 119-146
DOI: 10.29252/jss.15.1.121

Using Machine Learning Classification Algorithms in Official Statistics

Rezaei Ghahroodi[1], Z., Ranji, H.[2] and Rezaei, A.[2]
[1]School of Mathematics, Statistics and Computer Science, University of Tehran, Tehran, Iran.
[2]Statistical Center of Iran, Tehran, Iran.

Abstract: In most surveys, the occupation and job-industry related questions are asked through open-ended questions, and the coding of this information into thousands of categories is done manually. This is very time consuming and costly. Given the requirement of modernizing the statistical system of countries, it is necessary to use statistical learning methods in official statistics for primary and secondary data analysis. Statistical learning classification methods are also useful in the process of producing official statistics. The purpose of this article is to code some statistical processes using statistical learning methods and familiarize executive managers about the possibility of using statistical learning methods in the production of official statistics. Two applications of classification statistical learning methods, including automatic coding of economic activities and open-ended coding of statistical centres questionnaires using four iterative methods, are investigated. The studied methods include duplication, support vector machine (SVM) with multi-level aggregation methods, a combination of the duplication method and SVM, and the nearest neighbour method.

Keywords: Automated coding, Text mining, Statistical learning, Official statistics.

Mathematics Subject Classification (2010): 62G05, 62D05, 6205.