شناسایی صورت حساب جعلی مودیان مالیاتی به کمک تکنیک‌های داده‌کاوی

ابوالفضل قربانی[۱]
کامران لایقی[۲]
فاطمه داوودی[۳]

تاریخ دریافت: ۹۵/۶/۷، تاریخ پذیرش: ۹۶/۳/۳

چکیده

در این مقاله شاخص‌هایی ارائه می‌شود که با استفاده از آن‌ها می‌توان آن دسته از کاربرانی که مشکوک به ارائه صورت‌حساب‌های جعلی هستند را شناسایی نمود. به کمک روش‌های داده‌کاوی می‌توان به صورت‌سازی‌هایی که مربوط به درج اطلاعات نادرست مالی و عملیاتی برای فرار از پرداخت مالیات یا کاهش آن است، پی برد. در این پژوهش، ابتدا با استفاده از الگوریتم‌های خوشه‌بندی[۴] مانند شبکه‌های خودسازمانده[۵] و شبکه‌های عصبی گازی[۶]، گروه‌هایی از مودیان مالیاتی را که رفتار مشابهی دارند، شناسایی و سپس با استفاده از الگوریتم‌های درخت تصمیم‌گیری، شبکه‌های عصبی[۷] و شبکه‌های بیزی[۸] به شناسایی متغیرهای مربوط به رفتارهای متقلبانه، الگوهای رفتاری مرتبط و تشخیص موارد تقلب مبادرت نمود. جامعه آماری این پژوهش بنگاه‌های اقتصادی اعم از شرکت‌ها، کارخانه‌ها، کارگاه‌ها در شهر و استان تهران می‌باشد. نتایج روش‌های داده‌کاوی این پژوهش، متغیرهایی که در مورد بنگاه‌های اقتصادی کوچک و بزرگ و متوسط جهت ممیزی باید مد نظر قرار بگیرد را متمایز کرد و مدل شبکه عصبی با درصد صحت ۹۲٪ بر روی داده‌های آموزش با درصد صحت ۸۸٪ بر روی داده‌های اعتبار سنجی و با درصد صحت ۸۹٪ بر روی داده‌های آزمون توانسته موفق به کشف فرار مالیاتی گردد.

واژه‌های کلیدی: فاکتورهای جعلی، کشف تقلب، داده کاوی، خوشه‌بندی، شبکه‌های عصبی، شبکه‌های عصبی گازی

---

۱. دانشجوی کارشناسی ارشد مهندسی نرم افزار دانشگاه آزاد اسلامی، واحد تهران شمال
۲. عضو هیات علمی دانشگاه آزاد اسلامی، واحد تهران شمال (نویسنده مسئول) k_layeghi@iau-tnb.ac.ir
۳. دانشجوی دکتری مخابرات دانشگاه صنعتی خواجه نصیرالدین طوسی، تهران

4. Clustering
5. Self - Organizing Map (SOM)
6. Neural Gas Network
7. Neural Network
8. Bayesian Network

۱- مقدمه

فرار و تقلب مالیاتی همواره یکی از دغدغه‌های دائمی سازمان‌های مالیاتی دولتی در تمام کشورها، خصوصاً در کشورهای در حال توسعه بوده است (دایوا و همکاران، ۲۰۰۰). مالیات تنها منبع درآمدزایی دولت‌ها نیست، اما میزان درآمد مالیاتی یک کشور، شاخص مهمی برای نشان دادن تعهد و اثربخشی دولت برای انجام وظایفش و محدودسازی دسترسی به سایر منابع دیگر درآمدی از جمله منابع محدود طبیعی، به‌حساب می‌آید. به‌طور خاص، مالیات بر ارزش افزوده، که در بیش از ۱۳۰ کشور در مراحل مختلف توسعه اقتصادی اجرا شده، امروزه به یک جزء کلیدی از درآمدهای مالیاتی دولت‌ها تبدیل شده‌است و حدود ۲۵ درصد از درآمد مالیاتی جهان را به‌خود اختصاص می‌دهد (هریسون و کرلاو، ۲۰۰۵). به صورت متوسط در مورد کشورهای در حال توسعه، مالیات، حدود ۷۵ درصد از منابعی را که دولت‌ها سالیانه صرف هزینه‌ها و سرمایه گذاری‌های خود می‌کند، تشکیل داده است. به‌صورت تقریبی می‌توان گفت مالیات بر ارزش افزوده در مجموع بر روی بیش از ۴۰۰ میلیون صورت‌حساب در یک سال اعمال می‌شود که ۵۶٪ آنها به‌صورت فاکتور مکتوب و ۴۴٪ در قالب الکترونیکی صادر می‌شوند (برگمن، ۲۰۱۰).

پدیده فاکتورسازی در رابطه با مالیات بر ارزش افزوده، با استفاده از فرآیندهای تعیین مالیات قابل پرداخت را می‌توان این‌گونه توضیح داد: هنگامی که یک شرکت، صورتحساب نادرست ارائه می‌کند، در واقع خریدی را شبیه سازی می‌کند که هرگز وجود نداشته‌است، بنابراین، اعتبار مالیاتی شرکت مذکور به اشتباه افزایش یافته و مقدار قابل پرداخت مالیات بر ارزش افزوده آن کاهش می‌یابد. به این ترتیب، در اظهارات مالی سندی برای افزایش هزینه‌ها و مخارج ارائه می‌شود که کاهش در پرداخت مالیات بر درآمد را به همراه خواهد داشت. به این ترتیب یک تقلب و سند سازی رخ داده‌است. سندسازی‌ها بر دو نوع هستند:

- سندسازی فیزیکی: اگر عناصر فیزیکی تشکیل‌دهنده صورتحساب تقلبی باشند؛
- سندسازی ایدئولوژیک: اگر ماهیت سند تغییر نکرده اما عملیات ثبت شده در آن تقلبی یا غیر واقعی باشد.

کشف حالت دوم نسبت به حالت اول پیچیده‌تر و تشخیص آن دشوارتر است. سندسازی ایدئولوژیک حاوی معاملات ساختگی فراوانی است که برای کشف آن نیاز به ممیزی و بررسی دفترچه فروش و اصلاحات، ارجاع به اطلاعات تأمین‌کنندگان، خواهد بود. کشف این نوع از سندسازی‌ها، به دلیل نیاز به صرف زمان بسیار زیاد جهت جمع‌آوری و بررسی شواهد، بسیار پرهزینه و دشوار است. از جمله رایج‌ترین موارد جعل فیزیکی اسناد حسابداری، می‌توان به استفاده از فاکتورهای غیر قطعی، استفاده از فاکتور تقلبی

برای معرفی مالیات‌دهنده به‌عنوان یک مودی خوش‌حساب، و همچنین، استفاده از مجموعه‌های دوتایی از صورتحساب‌های مالیاتی که دارای فاکتورهایی با دو شماره یکسان که در واقع یکی از آنها جعلی و حاوی مبلغ بالاتری است، اشاره کرد.

در جعل ایدئولوژیک، فاکتورها برای ثبت یک عملیات غیر واقعی یا مغشوش ساختن محتویات یک عملیات واقعی استفاده می‌شوند. طبق روش مورد استفاده در اداره مالیات شیلی جهت برآورد فرار از پرداخت مالیات بر ارزش افزوده در مورد فاکتورهای جعلی و سایر بزرگنمایی‌های اعتباری مورد استفاده در بازه زمانی ۱۹۹۶-۲۰۰۴، فرار از پرداخت مالیات بواسطه فاکتورسازی در این بازه زمانی بین ۱۵٪ و ۲۵٪ از کل فرار از پرداخت مالیات بر ارزش افزوده را در بر گرفته است (اشنایدر و انست، ۲۰۰۰). با افزایش بحران اقتصادی در کشورها این نسبت افزایش قابل توجهی خواهد داشت. کشف و جلوگیری از این مهم مستلزم سرمایه‌گذاری منابع در بخش نظارت متمرکز بر مالیات‌دهندگان و شناسایی آن‌دسته از مالیات‌دهندگانی است که ریسک پذیری بیشتری دارند. صرف منابع و زمان بر تمامی مودیان مالیاتی حتی افرادی که هیچ‌گاه دست به این روش نمی‌زنند، اتلاف وقت است (اسلمورد و ایتزاکی، ۲۰۰۲). بنابراین باید راهی وجود داشته باشد که بتوان این دو گروه را از هم تمیز داد. از این منظر، تکنیک‌های داده کاوی کاربرد بسیار زیادی دارند، چرا که استخراج و تولید دانش از حجم زیادی از داده‌ها جهت تشخیص و توصیف رفتار جعلی و عدم پرداخت مالیات و در نهایت بهبود استفاده از منابع را میسر می‌سازند (فیاد و همکاران، ۱۹۹۶). مقاله حاضر به شرح زیر سازماندهی شده است. در بخش دوم به مطالعات مشابه‌ای که در کشورهای دیگر انجام شده می‌پردازیم. در بخش سوم تکنیک‌های هوش مصنوعی و چگونگی تشخیص فرار مالیاتی را بیان می‌کنیم. در بخش چهارم روش‌ها و تکنیک‌های داده کاوی این پژوهش را توضیح می‌دهیم. در بخش پنجم انواع اطلاعات مورد استفاده و بهترین نتایج بدست آمده درباره تشخیص و ردیابی تقلب را در صدور فاکتورها توصیف کرده، و در بخش ششم اصلی‌ترین نتایج و خطوط پژوهش‌های آتی را بیان خواهیم کرد.

۲- مطالعات مشابه

تقلب در شکل‌ها و انواع مختلف، پدیده‌ای است که هیچ جامعه مدرنی از آن مستثنی نیست. همه دولت‌ها، اعم از بزرگ و کوچک، عمومی یا خصوصی، محلی و یا چند ملیتی، تحت تأثیر این واقعیت هستند؛ واقعیتی که نقش مهمی در تضعیف همبستگی و برابری شهروندان در مقابل قانون داشته و مشاغل را به‌طور جدی مورد تهدید قرار می‌دهد. بسیاری از حوزه‌ها و صنایع تحت تاثیر این پدیده جهانی هستند. طبق مطالعه‌ای که توسط چینا در سال ۲۰۰۶ انجام شد، از ۱۵۰ شرکت بزرگ و متوسط در کشور شیلی درباره

این موضوع نظرسنجی شده‌است. نتایج نشان می دهد که ۴۱ درصد از آنها طی دو سال گذشته قربانی سندسازی بوده اند. این امر چالش بزرگی را در زمینه پیشگیری و فرصت‌های تشخیص تقلب و جعل ایجاد می‌کند (پدرسچی و همکاران، ۱۹۹۹)، از آنجایی که تقلب تأثیر مخربی بر وجهه شرکت نزد مشتریان و تامین کنندگان دارد، اثرات به مراتب بدتری از میزان گزارش شده به همراه دارد. در بسیاری از موارد، حتی شرکت‌هایی که قربانیان تقلب بوده‌اند، هنوز مشخص نشده‌اند.

منشا بسیاری از مشکلات کشف تقلب، فراوانی بیش از اندازه اطلاعات است (لوندین و همکاران، ۲۰۰۳). پردازش این حجم وسیع از داده‌ها در راستای شناسایی معاملات جعلی، نیازمند انجام تجزیه و تحلیل آماری است که به الگوریتم‌های سریع و کارآمد نیاز دارد؛ در این میان، الگوریتم‌های داده کاوی، تکنیک–های مرتبطی جهت تفسیر داده‌ها و بهبود درک فرایندهای نهفته در پس داده‌ها فراهم ساخته‌است (میات گلن، ۲۰۰۷). این تکنیک‌ها، تشخیص فرار از پرداخت مالیات و رفتارهای مالی غیرقانونی را در سایر حوزه‌ها از جمله بانکداری، مخابرات، بیمه، فناوری اطلاعات، پول‌شویی، و حتی زمینه‌های پزشکی و علمی، و غیره را تسهیل نموده‌است (پاتاک و همکاران، ۲۰۱۰).

برای تشخیص هرچه دقیق‌تر تقلب‌های مالیاتی، موسسات مالیاتی بتدریج استفاده از ممیزی‌های گزینشی تصادفی و یا تمرکز بر روی مالیات دهندگانی که در سال‌های اخیر ممیزی نشده بودند، و انتخاب موارد بر اساس تجربه و دانش حسابرسان را در دستور کار خود قرار داده‌اند. سایر روش‌ها بر اساس تجزیه و تحلیل آماری و تهیه و تنظیم نسبت‌های مالی یا مالیاتی توسعه یافتند که به ایجاد سیستم‌های کاملاً قانونی و مدل های ریسک منجر گردید (سازمان توسعه و همکاری‌های اقتصادی[1]، ۱۹۹۹). این موارد، اطلاعات مالیاتی را به شاخص‌های رتبه بندی تبدیل کرد که رده‌بندی مالیات دهندگان را براساس انطباق ریسک میسر ساخت. در سال‌های اخیر، تکنیک‌های داده کاوی و هوش مصنوعی در فعالیت‌های برنامه‌ریزی حسابرسی گنجانیده شدند که عمدتاً برای تشخیص الگوهای تقلب یا فرار از مالیات، و توسط مقامات مالیاتی به مقاصد خاص مورد استفاده قرار می‌گیرند (دفتر پاسخگویی ایالات متحده[2]، ۲۰۰۴؛ سازمان توسعه و همکاری‌های اقتصادی، ۲۰۰۴). این موسسه، از تکنیک‌های داده کاوی برای مقاصد مختلف استفاده می‌کند که از آن جمله، سنجش خطر پذیر بودن مالیات دهندگان، شناسایی موارد فرار از پرداخت مالیات و فعالیت‌های مجرمانه مالی (دوبین، ۲۰۰۷)، تشخیص تقلب‌های الکترونیکی، تشخیص تقلب در مالیات مسکن، تشخیص تقلب مالیات‌دهندگانی که از محل اعتبارات مالیات پولشویی درآمد کسب می‌کنند،

1. OECD
2. US Government Accountability Office

می‌باشند (سازمان توسعه و همکاری‌های اقتصادی، ۲۰۰۴؛ واتکینسا و همکاران، ۲۰۰۳). جدول (۱) بخشی از تکنیک‌های داده کاوی را که نهادهای مالیاتی سراسر جهان از آن استفاده می‌کنند را نشان می‌دهد.

در اداره مالیات استرالیا، برنامه پذیرش بر اساس یک مدل شناسایی ریسک تهیه و تدوین شده است که از تکنیک‌های آماری و داده کاوی جهت مقایسه، پیدا کردن ارتباط‌ها و الگوهای رگرسیون لجستیک[1]، درخت‌های تصمیم‌گیری و ماشین‌برداری پشتیبان[2] استفاده می‌کند (اداره پاسخگویی دولت آمریکا، ۲۰۰۴؛ دفتر پاسخگویی دولت آمریکا، ۲۰۰۸).

یکی از قابل توجه‌ترین موارد، رویکرد مورد استفاده توسط دنی و کریستن است که برای کشف خوشه‌های کوچک یا زیرگروه‌های جمعیتی غیرمعمول بنام نقاط داغ[3] با استفاده از تکنیک‌هایی مانند شبکه‌های خودسازمانده جهت کشف ویژگی‌ها و الگوریتم‌های خوشه‌بندی نظیر کی‌مینز[4] و تصاویری که برای کاربران غیرفنی هم به آسانی قابل درک هستند، مورد استفاده قرار می‌گیرد. در نیوزیلند، مدل موجود با درجه‌ای از انطباق ممیزی همراه است که به‌طور همزمان توسط همتای استرالیایی آن نیز بکار می‌رود (سازمان توسعه و همکاری‌های اقتصادی، ۲۰۰۴a). این طرح شامل تجزیه و تحلیل اقتصادی، بین‌المللی، جمعیتی، تنوع قومی و ساختار خانوادگی است. کشور کانادا نیز به نوبه خود از شبکه‌های عصبی و درخت‌های تصمیم‌گیری برای تشخیص ویژگی‌های مالیات‌دهندگانی که از پرداخت مالیات فرار کرده یا مرتکب تقلب می‌شوند، استفاده می‌کند که براساس نتایج حاصل از ممیزی‌های گذشته و در راستای تشخیص الگوهای عدم پذیرش یا فرار از مالیات شکل گرفته‌اند (سازمان توسعه و همکاری‌های اقتصادی، ۲۰۰۴b).

پرو یکی از اولین کشورهایی بود که از این تکنیک‌ها برای تشخیص فرار از پرداخت مالیات استفاده کرد و یک ابزار هوش مصنوعی را بر اساس شبکه‌های عصبی به سیستم گزینش قوانین کالو[5] اضافه کرد. در سال ۲۰۰۴، این مدل در سطح کاربرد قوانین فازی[6] و ارتباط با متغیرهای پیش‌پردازش[7] و طبقه‌بندی و رگرسیون[8] بهبود یافته و برای انتخاب مناسب‌ترین متغیرها مورد استفاده قرار گرفت (گارسیا و والدراما، ۲۰۰۷؛ تورگلر، ۲۰۰۵).

1. Logistic Regression
2. Support Vector Machines (SVM)
3. Hot Spots
4. K-means
5. Maritime Customs of Callao
6. Fuzzy Rules
7. Pre-processing Variables
8. Classification and Regression Trees (CART)

برزیل از تجزیه و تحلیل ریسک پروژه استفاده کرده و هوش مصنوعی را به‌صورت مشترک با اداره درآمد فدرال برزیل[1] و دانشگاه‌های این کشور مورد استفاده قرار داد (دیجیامپیتری و همکاران، ۲۰۰۸). این پروژه شامل تهیه و تنظیم یک سیستم تشخیص جهت شناسایی معاملات مشکوک بر اساس یک صفحه نمایش گرافیکی از اطلاعات و سوابق واردات و صادرات و یک سیستم اطلاعات محصول صادراتی بر اساس زنجیره‌های مارکوف[2]، برای کمک به واردکنندگان جهت ثبت نام و طبقه‌بندی محصولات، جلوگیری از تکرار و محاسبه احتمال اعتبار یک رشته در یک دامنه می‌شود. در مورد شیلی، اولین آزمایش در سال ۲۰۰۷، و با استفاده از شبکه‌های خودسازمانده و کی‌مینز جهت بخش‌بندی مالیات‌دهندگان ارزش افزوده، با توجه به اظهارنامه‌ها شکل گرفت (لوکدی و همکاران، ۲۰۰۷). پس از آن، در سال ۲۰۰۹، و در پی یک رویه بین‌المللی، مدل‌های ریسک در مراحل مختلف چرخه حیات مالیات‌دهندگان ساخته شدند، و شبکه‌های عصبی، درخت‌های تصمیم‌گیری و تکنیک‌های رگرسیون لجستیک نیز در آن استفاده شدند. در ادامه، آزمایش اول به‌منظور شناسایی کاربران بالقوه فاکتورسازی‌ها از طریق شبکه‌های عصبی مصنوعی و درخت‌های تصمیم‌گیری عمدتاً با استفاده از اطلاعات اظهارنامه‌های مالیاتی و درآمدی در بنگاه‌های خرد و کوچک توسعه پیدا کرده است.

۳- تکنیک‌های داده کاوی مورد استفاده

با توجه به مطالعات به‌عمل آمده در خصوص تکنیک‌های استفاده شده جهت شناسایی تقلب مالیاتی در کشورهای مختلف (جدول ۱)، به منظور توصیف و شناسایی الگوها، اغلب موارد از سه تکنیک داده کاوی شبکه‌های خودسازمانده، شبکه عصبی گازی و درخت تصمیم‌گیری مورد استفاده قرار گرفته و برای تشخیص تقلب و یا فقدان تقلب از شبکه‌های عصبی پس انتشار و شبکه‌های بیزی استفاده شده که در ادامه نحوه عملکرد تکنیک‌ها توضیح داده شده‌است.

1. Brazilian Federal Revenue
2. Markov

جدول (۱)- تکنیک‌های داده کاوی مورد استفاده در اداره‌های مالیاتی کشورهای مختلف جهت شناسایی تقلب مالیاتی (کاستلون و همکاران، ۲۰۱۳)

تکنیک مورد استفاده	ایالات متحده	کانادا	استرالیا	بریتانیا	بلغارستان	برزیل	پرو	شیلی
شبکه‌های عصبی	√	√		√	√		√	√
درخت تصمیم‌گیری	√	√	√				√	√
رگرسیون لجستیک	√		√	√	√		√	√
SOM	√	√	√				√	√
k-means			√					√
دستگاههای بردار پشتیبانی	√	√						√
تکنیک‌های تجسم	√					√		
شبکه‌های بیزی			√					
نزدیکترین مجاورت با K[1]			√					
قوانین ارتباط[2]							√	
قوانین فازی							√	
زنجیره‌های مارکوف							√	
سری‌های زمانی[3]		√						
رگرسیون				√				
شبیه سازی	√							

1. K-Nearest Neighbour
2. Assosiation Rules
3. Times Series

منبع: یافته‌های تحقیق

۳-۱- شبکه‌های خودسازمانده

شبکه‌های خودسازمانده یکی از مدل‌های رایج در شبکه‌های عصبی مصنوعی است که برای تجزیه و تحلیل و تجسم داده‌های چندبعدی، براساس یادگیری رقابتی بدون نظارت[1]، مورد استفاده قرار می‌گیرد (وسانتو، ۲۰۰۰). این شبکه به‌طور خاص، مجموعه‌ای از نورون‌های[2] ترتیب‌بندی شده در یک بُعد شبکه ای «a» را شامل می‌شود که معمولاً به شکل مستطیل، استوانه‌ای یا حلقوی است و یک فضای خروجی در بُعد «d» را ایجاد می‌کند که «a» کمتر یا برابر با آن است و رابطه‌های اطراف روی آن تعریف شده، هدفش کشف ساختار زیربنایی داده‌های وارد شده به آن است. بر اساس ساختار، همه نورون‌های مشابه، ورودی‌ها را به‌طور همزمان دریافت می‌کنند. در طول آموزش، نورون‌های شبکه فعالیت‌هایی را تحت‌تأثیر تحریک داده‌های ورودی ایجاد می‌کنند که شناسایی هرچه بهتر مناطق شکل‌گیری متغیرهای ورودی خاص توسط نورون‌ها را امکان‌پذیر می‌سازند. الگوهای فعالیت تولید شده در هر منطقه، دارای ویژگی‌های مشابهی بوده و می توانند در قالب یک گروه یا خوشه واحد و بر اساس سنجش فاصله (معمولاً اقلیدسی[3]) گروه‌بندی شوند. لایه خروجی نورون برنده یا بهترین واحد تطبیق لایه ایست که بردار وزن آن شباهت زیادی به اطلاعات ورودی دارد. این ابزار معمولاً برای خوشه‌بندی و تقسیم‌بندی، و ایجاد گروه‌هایی از اشیاء با رفتار شبیه به یکدیگر، اما متفاوت از اشیاء سایر گروه‌ها، استفاده می‌شود.

۳-۲- شبکه‌های عصبی گازی

شبکه‌های عصبی گازی یک الگوریتم نسبتاً جدید برای شبکه‌های عصبی بدون نظارت است که تمرکز اصلی‌اش بر کوانتیزاسیون برداریِ[4] سازه‌های دلخواه می‌باشد (فیلیپون و همکاران، ۲۰۰۸). تفاوت عمده آن با شبکه خودسازمانده در این است که این روش شبکه‌ای را تعریف نمی‌کند که روابط توپولوژیک[5] بین واحدهای شبکه را تحمیل کند و هر نورون می‌تواند آزادانه در سراسر فضای داده‌ها حرکت کند. این آزادی در جابجایی، امکان جایگذاری بهتر توزیع داده ها را در فضای ورودی به الگوریتم می دهد چرا که لازم نیست سلول های عصبی روابط خاصی باهم برقرار کنند. با این حال، بنظر می‌رسد که داشتن پیش زمینه درباره تعداد گروه‌ها لازم و ضروری باشد. در طول آموزش شبکه، نورون‌ها موقعیت خود را تغییر داده و خودشان را با ابر داده ها تطبیق می دهند. در این الگوریتم، هر متغیر ورودی یک تحریک در هر واحد از

1. Unsupervised Competitive Learning
2. Neurons
3. Euclidean
4. Vector Quantization
5. Topological Relationships

شبکه ایجاد می‌کند. در هر تکرار، یک بردار داده تصادفی به همه سلول های عصبی وارد می‌شود. برای هر بردار داده، نزدیکترین نورون براساس فاصله اقلیدسی پیدا می‌شود. این نورون برنده[1] نامیده می‌شود. در مرحله بعد، اطراف (قطر) نورون برنده ایجاد می شود، که باعث کاهش چشمگیر تعداد تکرارها خواهد شد.

۳-۳- درخت‌های طبقه‌بندی

درخت‌های طبقه بندی[2] یکی از روش‌های یادگیری غیرپارامتری نظارتی[3] رایج است که بخاطر سادگی و کاربردشان در حوزه های مختلف مورد توجه می‌باشند(مورتی، ۱۹۹۸). به‌طور کلی، الگوریتم‌های ساخت و ساز درخت‌ها از نظر استراتژی‌های مورداستفاده برای پارتیشن‌بندی گره‌ها[4] و هرس آنها باهم متفاوت هستند. در برخی موارد مطالعاتی، از درخت تصمیم براساس روش کای[5] استفاده شده است که تعداد متفاوتی از شاخه‌ها را از یک گره ایجاد کرده و متغیرهای پیوسته[6] و متغیرردهای[7] را در نظر می‌گیرد. در واقع، این الگوریتم شامل تشکیل تمام جفت های احتمالی و ترکیبات دسته‌ها، و گروه بندی مقوله‌هایی می شود که رفتار همگنی نسبت به متغیر پاسخ در یک گروه نشان داده و آنها را از دسته بندی‌هایی که رفتار متفاوتی دارند، دور نگاه می‌دارد. جفتی که کمترین ارزش را براساس این شاخص دارد، یک طبقه بندی جدید از دو ارزش ادغام شده را تشکیل می دهد. البته در صورتی‌که از نظر آماری معنی دار نباشد. برای دسته بندی های ادغام شده، تحکیم هرچه بیشتر ارزش های شاخص – با یک دسته بندی کمتر – انجام شده و فرایند زمانی پایان خواهد یافت که بدلیل دستیابی به نتایج معنادار از نظر آماری، ادغام بیشتری امکانپذیر نباشد.

۳-۴- شبکه عصبی پرسپترون چندلایه

مدل پرسپترون چندلایه[8] یک مدل شبکه ای عصبی مصنوعی از لایه‌ها است که برای طبقه بندی و گروه بندی براساس عملکرد مغز انسان – از طریق مجموعه بهم پیوسته‌ای از بردارها – مورداستفاده قرار می‌گیرد (پارلوس، ۱۹۹۴). این شبکه باید ارتباط بین ویژگی‌های ورودی و خروجی مطلوب هر مورد را پیدا کند. این کار از طریق یک روش یادگیری موسوم به پس–انتشار[9] صورت می‌گیرد که با تنظیم وزن شبکه، خطای پیشگویی را به حداقل می رساند. این روش دو مرحله دارد. ابتدا، جابجایی‌ها براساس ورودی محاسبه شده

1. Winning Neuron
2. Classification Trees
3. Non-parametric Supervised Learning
4. Partition Nodes
5. CHAID Methodology
6. Continuous
7. Categorical
8. Multilayer Perceptron Model
9. Back Propagation

و وزن شبکه اولیه مشخص می‌شود و خطای پیش‌بینی آن محاسبه می‌گردد. در مرحله دوم، خطا مجدداً از واحدهای خروجی به واحدهای ورودی محاسبه شده، خطای هر واحد مشخص خواهد شد. بدین صورت، وزن‌ها بوسیله روش گرادیان[1] نزولی به‌روزرسانی خواهند شد. این فرآیند تکرار شونده است، به‌طوری‌که پس از تکرار چندباره الگوریتم، شبکه آنقدر همگرایی پیدا خواهد کرد که طبقه‌بندی همه متغیرهای آموزش را امکانپذیر ساخته، خطا را به حداقل خواهد رساند.

۳-۵- شبکه‌های بیزی

شبکه‌های بیزی نمودارهای بدون دور را برای پیش‌بینی احتمال دستیابی به نتایج مختلف، براساس مجموعه‌ای از حقایق بسط و توسعه داده‌اند (فریدمن و همکاران، ۱۹۹۷؛ هکرمن و همکاران، ۱۹۹۵). این شبکه از مجموعه ای از گره‌ها که نشان‌دهنده متغیرهای مسأله هستند و مجموعه ای از کمان‌های متصل کننده گره‌ها تشکیل شده‌اند که رابطه وابستگی بین صفات داده‌های مشاهده شده را مشخص می‌کنند. شبکه‌های بیزی، توزیع احتمالی را توصیف می‌کنند که بر مجموعه‌ای از متغیرهای مشخص حاکم است و فرضیات استقلال شرطی با احتمالات شرطی را مشخص می‌نماید. معمولاً مسأله به دو بخش تقسیم می‌شود: یادگیری ساختاری[2] که ساختار شبکه را نشان می‌دهد، و یادگیری پارامتری[3]، که در آن، از طریق ساختار نمودار شناخته شده، احتمال هر گره بدست می‌آید. اصلی ترین مزیت این روش‌ها آن است که امکان بدست آوردن احتمال وقوع یک رویداد خاص بر اساس مجموعه ای از اقدامات وجود دارد و از طریق وب گراف[4]، دید روشنی از روابط حاصل خواهد شد.

۴- داده‌ها، تجزیه و تحلیل و نتایج

هدف مقاله حاضر، پیدا کردن متغیرهای تاثیرگذار در خصوص سنوات گذشته شرکت‌ها است که منجر به رسیدگی و تعیین درآمد مشمول مالیات می‌شود نظیر تغییر مدیر عامل و اعضای هیات مدیره، تغییر حسابرس، نوع حسابرسی (اگر حسابرسی به وسیله سازمان حسابرسی صورت پذیرد)، سن شرکت، و بسیاری فاکتورهای دیگر.

۴-۱- گزینش داده‌ها و ویژگی‌ها

نمونه برگرفته از جامعه آماری شامل مودیانی می‌باشد که لااقل یک دوره نسبت به تنظیم و تسلیم اظهارنامه مالیات بر ارزش افزوده در دوره‌های سنوات ۹۰ و ۹۱ اقدام نموده اند، که این نمونه انتخاب شده

1. Gradient
2. Structural Learning
3. Parametric Learning
4. Webgraph

شامل۵۸۱ بنگاه اقتصادی (شرکت‌ها، کارخانه‌ها،کارگاه‌ها و ...) در شهر و استان تهران می‌باشد. جدول ۲ طبقه‌بندی مالیات دهندگان در تحلیل ما را نشان می‌دهد. موارد تقلب یا فاقد تقلب در سه گروه مجزا طبقه‌بندی می‌شوند: «۰» یعنی مالیات دهنده حسابرسی شده و هیچیک از فاکتورهای وی در هیچ‌یک از دوره‌های مورد بررسی، اشتباه نبوده است؛ «۱» یعنی مالیات دهنده در سال تجزیه و تحلیل از فاکتورهای جعلی استفاده نکرده اما در سایر دوره‌های بررسی تقلب داشته است (معمولاً سال گذشته یا آینده)؛ و «۲» یعنی مالیات‌دهنده در سال مورد بررسی از فاکتورهای نادرست استفاده کرده است. منظور از شرکت‌های حسابرسی شده شرکت‌هایی می‌باشند که در سنوات سال‌های قبل توسط ممیزان سازمان امور مالیاتی مورد حسابرسی و رسیدگی قرار گرفته‌اند.

جدول (۲)- تعداد مالیات‌دهندگان مورد استفاده در تجزیه و تحلیل

مالیات دهندگان	خرد-کوچک	متوسط-بزرگ	تعداد کل بنگاه‌ها
بنگاه‌های فعال در دوره زمانی ۱۳۹۰-۱۳۹۱	۵۵۸ (٪۹۶)	۲۳ (٪۴)	۵۸۱ (٪۱۰۰)
شرکت‌های حسابرسی شده بر اساس صورتحساب‌ها در سال ۱۳۹۰-۱۳۹۱ دارای تقلب یا فاقد تقلب	۱۲۸ (٪۷۶)	۴۱ (٪۲۴)	۱۶۹ (٪۱۰۰)

منبع: یافته‌های تحقیق

جهت ساخت بردار ویژگی‌ها، ۲۰ متغیر از فرم سه ماهه پرداخت مالیات بر ارزش افزوده مربوط به پرداخت اجرایی مالیات بر ارزش افزوده، ۳۱ متغیر از فرم سالانه مالیات بر درآمد در ارتباط با داده‌های مالی کسب و کارها و طبقه درآمدی مشمول مالیات، و ۳۱ نسبت مالیاتی در ارتباط با اطلاعات مالیات بر ارزش افزوده و مالیات بر درآمد بر اساس سوددهی و نقدینگی شرکت، و سایر فاکتورها انتخاب شد. از لحاظ رفتار و ویژگی‌های شرکت، این رویکرد به شکل‌گیری ۸۲ شاخص منجر شد که نشان‌دهنده رفتار تقلب و فاقد تقلب در طول چندین دوره، و براساس سابقه عملکرد، ویژگی‌های خاص آن و اطلاعات تولید شده در مراحل مختلف است (جدول ۳).

جدول (۳)- نوع اطلاعات مورد استفاده برای ساخت بردار مشخصات

مفهوم	نوع اطلاعات
پرداخت مالیات	اظهارنامه مالیات بر ارزش افزوده، اظهارنامه مالیات بر درآمد، نرخ‌های مالیاتی و مالیات بر درآمد
ویژگی‌های فردی	سن مالیات دهنده، سن شرکت، تغییر حسابرس، نوع حسابرس، تغییر مدیر عامل، صورتحساب الکترونیک، حسابرسی رایانه‌ای، فعالیت‌های اقتصادی، اطلاعات مکانی اعم از اجاره یا مالکیت شعبات
تاریخچه	ممیزی گزینشی، تخلفات قبلی، رسیدگی به مشکلات
رفتار و گستره سال مطالعه	شکست در فعالیت، رفع اتهامات، ضرر و زیان، اشتباه در اسناد، قوانین بدهی، نداشتن صورتحساب، فاکتورهای بررسی شده، هشدارها
چرخه حیات	استارت آپ‌ها، تأیید فعالیت‌ها، مهر و امضای اسناد، تغییر اطلاعات، انقضای تعلیق قبلی فعالیت‌ها
روابط	عوامل، نمایندگان قانونی، شرکا، اعضای خانواده، تأمین کنندگان، حسابدارها، ارتباط‌ها و نمایندگان (دارایی‌ها، سابقه تخلفات، بازرسی‌ها، همپوشانی‌ها)

منبع: یافته‌های تحقیق

در پیش پردازش داده‌ها که با استفاده از یک قاعده کلی در ارتباط با شفاف‌سازی داده‌ها صورت گرفت، مواردی که بالاتر از پنج برابر میانگین انحراف معیار بودند، به‌عنوان مقادیر خارج از محدوده شناخته شدند و از نمونه آماری حذف و الباقی ، جهت مطالعه باقی ماندند. در اغلب موارد، توزیع متغیرهای مورد استفاده سیر نزولی داشت، یعنی درصد زیادی از مالیات دهندگان مقادیر کمی مالیات پرداخت می‌کردند و تنها یک گروه کوچک از آنها پرداخت‌های بالا داشتند. همین امر در مورد متغیرهای رفتاری صادق است، چراکه سوء رفتار گروه کوچکی از مالیات دهندگان را نشان می‌دهد. بنابراین، نمونه هایی که ارزش‌های بالاتری از جمع مالیات‌دهندگان داشتند، گروه تمرکز مطالعه را تشکیل داد. از آنجایی‌که اظهارنامه مالیات بر ارزش افزوده به صورت سه ماهانه و اظهارنامه مالیات بر درآمد به صورت سالانه انجام می‌شود، لذا باید مجموع سالیانه مقادیر سه ماهه اظهارنامه ارزش افزوده را در طول سال مدنظر قرار داد و آن‌را با اطلاعات مالیات بر درآمد مقایسه کرد. در خصوص داده‌های پوچ، اطلاعات مالیات بر ارزش افزوده کامل‌تر از اطلاعات

مالیات بر درآمد است. بنابراین، به‌واسطه ارتباط مستقیم بین آن‌ها، اطلاعات بدهی و اعتبار مالیات بر ارزش افزوده باید برای تکمیل داده‌های درآمد و هزینه در هر دوره مورد استفاده قرار گیرد. در مورد سایر حوزه‌های درآمدی، همانند بخش کد فروش، میانه (میزان متوسط) برای مالیات دهندگان استفاده شد. به‌منظور ممانعت از ورود متغیرهایی با دامنه‌های ارزشی بزرگتر جهت جلوگیری از اثر سوء بر روی متغیرهایی که طیف کوچکتری دارند، نرمال‌سازی[1] متغیرها به‌صورتی که قابل مقایسه با یکدیگر باشند – با استفاده از انحراف معیار حداقل–حداکثر[2] در این طیف [۱و۰] مورد استفاده قرار گرفت.

علاوه بر این، قبل از انتخاب متغیرها جهت استفاده در مدل‌های رفتاری، لازم بود آنها را تجزیه و تحلیل نموده تا به متغیرهای تاثیرگذار تقلیل دهیم. در نهایت، ۱۵ مولفه اصلی در بنگاه‌های خرد و کوچک انتخاب شد که ۶۱/۳۱ درصد از واریانس در داده‌ها را پوشش می‌دهد و همچنین، ۱۶ مولفه اصلی برای بنگاه‌های متوسط و بزرگ انتخاب شد که ۵۹/۹ درصد از واریانس در داده‌ها را پوشش می‌دهد. از آنجایی‌که هدف ما، تولید متغیرهای رفتاری در ارتباط با کاربرد و فروش فاکتورهای جعلی بود و سایر رفتارها مدنظر نبودند، متغیرهایی انتخاب شدند که همبستگی[3] متوسط به بالایی با کاربرد متغیر و فاکتورهای جعلی داشتند. متغیرهایی که بیش از ۱۰٪ شانس داشتند که همبستگی پیرسون[4] شان ضریب صفر داشته باشد، حذف شدند – به جز برخی از متغرهای مهم مانند ساختار سرمایه، سودآوری، عملکرد، نقدینگی، جریان نقد عملیاتی و مقادیر مالیات بر ارزش افزوده. بدین صورت، در این روش، ۴۲ متغیر در بخش خرد و کوچک و ۳۶ متغیر در بخش متوسط و بزرگ انتخاب شدند. در گروه اول، ۳۵ درصد از متغیرها با مالیات بر ارزش افزوده مطابقت داشتند، ۳۵ درصد از متغیرها با مالیات بر درآمد و ۳۰ درصد با رفتار در ارتباط بودند. از سوی دیگر، در گروه دوم، این درصدها به‌ترتیب تا ۳۱، ۳۸، و ۳۱ در ارتباط بودند. بعد از حذف موارد غیرمرتبط و متناقض، مجموعه نهایی داده‌ها از تعداد ۵۳۲ مالیات دهنده خرد و کوچک و ۲۲۶ بنگاه متوسط و بزرگ تشکیل گردید که ۶/ ۴ درصد در گروه اول و ۴/ ۳ درصد در گروه دوم کاهش یافت.

۴-۲- مدل‌سازی

به منظور شناسایی و انتخاب متغیرهای تاثیرگذار، در مرحله اول، تکنیک‌های داده کاوی در نمونه آماری اعمال شد تا روابط بین پرداخت مالیات و متغیرهای رفتاری آنها در ارتباط با استفاده از فاکتورهای جعلی شناسایی شود. سپس تکنیک‌های طبقه بندی در موارد بروز تقلب و عدم بروز آن اعمال شد تا الگوهای

1. Normaliation
2. Min–max Standard Deviation
3. Correlation
4. Pearson Correlation

خاص این گروه از مالیات دهندگان شناسایی شود. سرانجام، ابزارهای طبقه بندی به‌منظور تشخیص موارد وجود تقلب یا عدم وجود تقلب در اطلاعات حاصل، بکار گرفته شد.

۴-۲-۱- توصیف نمونه آماری/تنوع شرکت‌ها

در ابتدا، الگوریتم خوشه بندی از جمله شبکه خودسازمانده به‌منظور شناسایی خوشه‌ها یا گروه‌های مالیات دهندگانی که رفتار مشابه داشتند، در نمونه آماری مالیات دهندگان اعمال گردید. فرضیه اصلی عبارت است از این‌که، تنها متغیرهای رفتاری و مالیاتی، و مالیات دهندگانی که منجر به رفتار مالی متقلبانه یا غیر متقلبانه می‌شوند، شناسایی و تشخیص داده شوند. برای انجام آزمایش‌ها از بسته R-SOM استفاده شد که بر اساس یک توپولوژی شبکه مستطیل شکل، با سه نورون ورودی و نورون‌های خروجی ۲۴×۲۴ در بنگاه‌های خرد و کوچک، و نورون‌های خروجی ۳۶×۳۶ در شرکت‌های متوسط و بزرگ، با حداکثر ۱۰۰ بار تکرار اعمال می شود. برای آموزش سیستم، به دلیل محدودیت‌های محاسباتی و همچنین مشکل بیش برازش[1]، اطلاعات مربوط به ۱۰۰۰ بنگاه اقتصادی به‌صورت تصادفی به ورودی مدل اضافه شد. در نتیجه، ۵ خوشه در بخش بنگاه‌های خرد و کوچک و ۵ خوشه در بنگاه‌های متوسط و بزرگ مطابق جدول ۴ و ۵ تولید شدند. در جدول ۴ منظور از دوره t بازه زمانی سال‌های ۱۳۹۰-۱۳۹۱ است.

جدول (۴)- خوشه‌ها در شرکت‌های خرد-کوچک

ردیف	متغیر	دوره	مفهوم	فقدان تقلب	تقلب
۱	بستانکاری فاکتور فروش	t	ارزش افزوده	√	
	پرداخت ارزش افزوده			√	×
	اعتبارات			×	√
	ترازنامه‌های اعتبار مالیاتی			×	√
۲	نسبت بدهی ها/اعتبارات	t	نسبت درآمد ارزش افزوده		
	نسبت درآمد/دارایی‌ها				
۴	بررسی فعالیت	<t	چرخه حیات	×	√

1. Overfitting

۵	تخلفات و بی نظمی‌ها	<t	سابقه رفتار	×	
	تخلفات غیرمستقیم			×	
	ممیزی‌های مثبت قبلی				√

منبع: یافته‌های تحقیق

جدول (۵)- خوشه‌ها در شرکت‌های متوسط و بزرگ

ردیف	متغیر	دوره	مفهوم	فقدان تقلب	تقلب
۱	هزینه‌ها و مخارج	t	درآمد	√	
	دارایی‌ها			√	
	مسئولیت‌ها			√	×
۲	اعتبارات	t	مالیات بر ارزش افزوده	√	
	تزارنامه اعتبار مالیاتی			√	
	تعداد اشتباهات فروش				×
۳	نسبت هزینه‌ها/دارایی‌ها	t	نسبت درآمد مالیات بر ارزش افزوده		√
	نسبت درآمد/دارایی‌ها				√
	نسبت فاکتور بستانکاری/کل بستانکاریها				√
۴	رسمی سازی حسابرسی	t	ویژگی‌ها	√	×
	نوع حسابرسی			√	×
	تغییر حسابرس			√	

ردیف	متغیر	دوره	مفهوم	فقدان تقلب	تقلب
۵	نوع حسابرسی قبلی	<t	سابقه رفتار		√
	محدودیت‌های مهرزنی				√
	اتهامات و تخلفات				√
	شکست در فعالیت				√
	ممیزی‌ها			√	×

منبع: یافته‌های تحقیق

خوشه بدست آمده در گروه اول تفاوت‌های مهمی از نظر تقلب‌های فروش و/یا فاکتورسازی، سطح پرداخت مالیات بر ارزش افزوده، سطح هزینه‌های گزارش شده، سطح رسمیت شرکت، مشارکت در دیگر شرکت‌ها و برخی از مسائل ردیابی باهم داشتند. گروه متوسط و بزرگ از نظر اشتباهات در ثبت فروش و/یا فاکتورسازی، سطح استفاده از ترازنامه‌های اعتبار مالیاتی، یادداشت‌های اعتباری و فاکتورهای دارایی‌های ثابت، بدهکاری‌ها و دارایی‌ها و همچنین نتایج ممیزی‌های قبلی و سطح رسمیت شرکت باهم تفاوت داشتند. اگرچه متغیرهای رفتاری خاصی شناسایی شد، اما هیچ‌یک ارتباط معناداری با صدور فاکتورهای نادرست نداشتند. علاوه بر این، متغیرهای رفتاری در ارتباط با ویژگی‌های بلندمدت و سابقه بی نظمی، از یک گروه به گروه دیگر چندان متفاوت نبودند.

سپس الگوریتم شبکه عصبی گازی اعمال شده و تعداد مشابهی از خوشه‌ها تحت عنوان نقشه کونن، با استفاده از بسته R-cclust در نظر گرفته شدند که یک آرایه از ویژگی‌های مرکزی[1] را برای هر متغیر و یک طبقه بندی بردار ایجاد کرده، گروهی را که هر مالیات دهنده به آنها تعلق داشت، مشخص نمود. در این مورد، گروه‌های تولید شده نیز تحت تاثیر پرداخت مالیات بودند، اما تفاوت‌های عمده ای از نظر شرایط رفتاری باهم داشتند. این رویکرد، تمایز گروه‌ها از نظر عملکرد بهتر یا بدتر و ارتباط دهی آن با پرداخت مالیات را میسر می‌سازد، هرچند موارد مربوط به صدور فاکتور اشتباه لزوماً در تمامی گروه‌ها مشاهده نشدند. اگرچه این تکنیک‌ها می‌توانند شرح خوبی از نمونه آماری مالیات دهندگان ارائه داده و برخی از متغیرهای متمایزکننده آنها از یکدیگر را شناسایی کنند، ولی در نظر گرفتن متغیرهایی که بیشترین همبستگی را با کاربرد فاکتورهای اشتباه دارند، بیشتر با متغیرهای مالیات در ارتباط است تا متغیرهای رفتار؛ و باعث می‌شود گروه‌ها از نظر نوع معامله (فروش با فاکتور و/یا تقلب در فروش)، سطح فعالیت (سطح بالا–پایین

1. Centroids

فروش، هزینه) و پرداخت مالیات (بالا–پایین) باهم متفاوت باشند که این امر از تنوع بیشتر در این متغیرها در مقایسه با متغیرهای رفتار نشأت می‌گیرد. بر همین اساس، متغیرهای رفتاری مرتبط با عملکرد تقلب و فقدان تقلب و با توجه به نقاط مشترک بدست آمده از هر دو روش شناسایی شدند (جدول ۵).

۴-۲-۲- توصیف موارد حاوی تقلب یا فاقد تقلب

در انتخاب متغیرهای تاثیرگذار، باید توجه داشت که موارد تقلب معمولاً در میان پرونده‌های بزرگتر یافت می‌شوند. به‌همین دلیل، از درخت تصمیم‌گیری برای همه داده‌های ممیزی با نتایج شفاف استفاده شد چرا که شناسایی نقطه قطع هر متغیر را در مقایسه با زمان بروز هر رفتار نشان می‌دهد. نوع درخت مورد استفاده، تشخیص تعامل خودکار مجذور کای[1] است که طبقه بندی غیر باینری[2] و تولید چندین شاخه از یک گره را با در نظر گرفتن متغیرهای پرتکرار و قطعی امکان‌پذیر می‌سازد.

همانطور که در شکل ۱ نشان داده شده است، عواملی که بیشترین تأثیر را بر شرکت‌های خرد و کوچک دارند، نتیجه ممیزی‌های پیشین و درصد خرید آن شرکت‌ها بر اساس فاکتور است، این نشان می‌دهد که احتمال فاکتورسازی در کسانی که در گذشته بیشتر حسابرسی شده‌اند و تخلفی نداشته اند، و خریدهایشان هم در درجه اول بر اساس فاکتور نیست، کمتر از کسانی است که خریدشان عمدتاً بر اساس فاکتور ثبت شده است و تخلف‌هایی در گذشته داشته‌اند. در واقع، این دو متغیر به تنهایی، تعداد گره‌های پایانی را با غلبه موارد بدون تقلب مشخص می‌کند. علاوه بر این، متغیر نشان دهنده مزیت بیشتر قانون شکنی و بی‌نظمی در ارتباط با فاکتورهای گذشته و کثرت صدور فاکتور، زمینه ساز گره‌های پایانی با غلبه موارد صدور فاکتورها نادرست می‌باشد. به‌طور خاص، گره ۱۲ که حاوی تقریباً نیمی از موارد (۴۶٪) است، بر حسب ارزش بدست آمده از اعتبار متوسط فاکتور صادر شده (هرچه این شاخص بالاتر باشد، پتانسیل بروز تقلب بیشتر است)، به چند شاخه مختلف تقسیم شده است. به‌طور مشابه، کثرت موارد تقلب در هر شاخه به تعداد فاکتورهای صادر شده، مالیات بر ارزش افزوده پرداخت شده، بستانکاری کل در هر برگه فاکتور/ خرید و فروش، ارتباط بین هزینه‌ها و دارایی‌ها و سطح مشارکت در دیگر شرکت‌ها بستگی دارد. این تکنیک در شناسایی متغیرهای مرتبط با تقلب و عدم تقلب موثر بود. چون گره‌های پایانی اصولاً از مواردی مشخص تشکیل شده‌اند، یا با مواردی با ارزش خروجی «۱» ترکیب شده‌اند که نزدیکی بیشتری با موارد تقلب «۲» دارند. جدول ۷ با در نظر گرفتن متغیرها و قوانینی که در هر شاخه از درخت تکرار شده اند، موارد تقلب و عدم تقلب را از یکدیگر متمایز کرده و رفتارهای مربوط به هریک را نشان می‌دهد؛ متغیرهای اصلی مورد بررسی و روابطی که تولیدکننده گره‌هایی با و بدون فاکتورسازی جعلی هستند، طبق جدول ۶

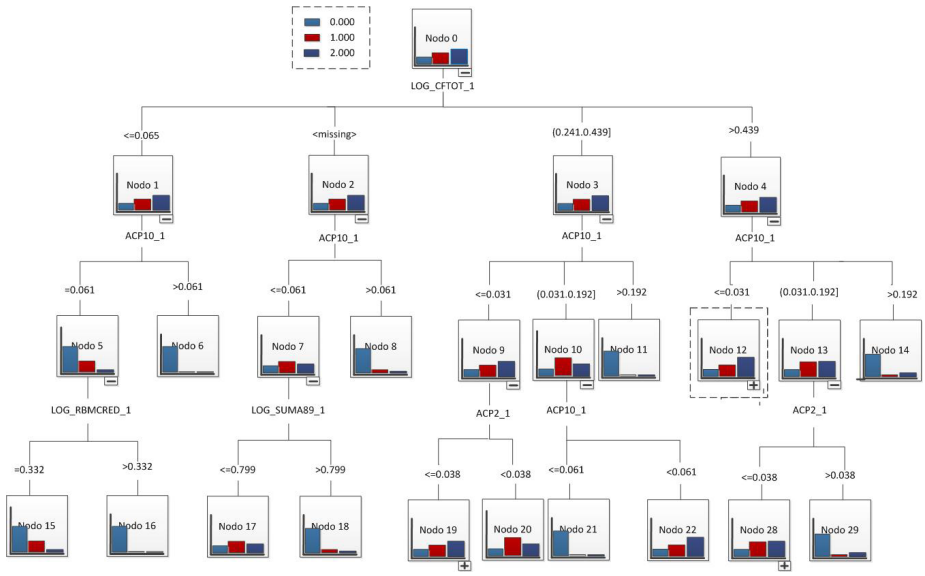
1. Chi-square Automatic Interaction Detection (CHAID)
2. Non-binary Classifications

در شکل ۱ به اختصار نشان داده شده‌اند.

جدول (۶)- معرفی متغیرهای مربوط به درخت شکل ۱

متغیر	مفهوم
ACP10-1	نتایج ممیزی‌های قبلی و کل مالیات بر ارزش افزوده تعیین شده
ACP2-1	درصد اعتبار فاکتورها
CFTOT-1	رابطه بین ترازنامه‌های اعتبار مالیاتی و اعتبارات
RBMCRED-1	بستانکاری کل بر اساس فاکتورها/فاکتورهای فروش
SUMA89-1	ارتباط بین فاکتورهای مهرشده و صادرشده

منبع: یافته‌های تحقیق



مهمترین متغیرهایی که موارد تقلب را در بنگاه‌های خرد و کوچک مشخص می‌کنند، از نتایج ممیزی‌های قبلی و کل مالیات بر ارزش افزوده تعیین شده، درصد اعتبار فاکتورها، رابطه بین ترازنامه‌های اعتبار مالیاتی و اعتبارات، بستانکاری کل براساس فاکتورها/فاکتورهای فروش و ارتباط بین فاکتورهای مهرشده و صادرشده، حاصل شده‌اند. متغیرهای متوسط و بزرگ با کل ترازنامه اعتبار مالیاتی، درصد اعتبار فاکتورها، تعداد نمایندگان قانونی، سطح رسمیت حسابداری و رابطه بین هزینه‌ها و دارایی‌ها، و غیره مطابقت دارند.

جدول (۷)- متغیرهای مربوط به رفتار متقلبانه و غیرمتقلبانه با فاکتور نادرست

	شرکت‌های خرد و کوچک				
ردیف	متغیر	دوره	مفهوم	فقدان تقلب	تقلب
۱	فاکتور بستانکاری	t	مالیات بر ارزش افزوده		√
	صورتحساب صادر شده			×	
	مالیات بر ارزش افزوده				√
۲	نسبت فاکتور اعتبار/کل اعتبارات	t	نسبت درآمد به مالیات بر ارزش افزوده	×	√
	نسبت ترازنامه اعتبار مالیاتی/ میانگین اعتبارات			√	
	نسبت هزینه/دارایی‌ها				√
۳	کثرت مهرزنی	t-۲	مهرزنی		√
	نسبت فاکتورهای صادرشده/ فاکتورهای مهرشده			×	√
۴	تخلفات	≤t	سابقه رفتار		×
	ممیزی‌های منفی قبلی			√	√
	ممیزی‌های مثبت قبلی				√
	شرکت‌های متوسط و بزرگ				
	متغیر	دوره	مفهوم	فقدان تقلب	تقلب
۱	ترازنامه اعتبار مالیاتی	t	مالیات بر ارزش افزوده	√	×
۲	نسبت فاکتور اعتباری/کل اعتبارات	t	نسبت درآمد بر مالیات بر ارزش افزوده	×	√
	نسبت هزینه‌ها/دارایی‌ها			×	√

	شرکت‌های خرد و کوچک				
۳	سن شرکت	t	ویژگی ها		×
	رسمیت فرایند حسابرسی				×
	فعالیت‌های اقتصادی				√
۴	میزان سفارش‌های قابل پرداخت	<t	سابقه رفتاری		√
	شکست در پاسخ دادن به اطلاعیه‌ها				√
	بی نظمی و تخلف در تنظیم فاکتور			×	√

منبع: یافته‌های تحقیق

۴-۳- تشخیص تقلب

برای تشخیص موارد تقلب از شبکه‌های عصبی مصنوعی، درخت‌های تصمیم‌گیری و شبکه‌های بیزی استفاده شده است. برای جلوگیری از برازش بیش از اندازه مدل، داده‌ها با استفاده از قانون ۳۰/۷۰ به دو مجموعه آموزشی و مجموعه تست تقسیم شدند. یکی از پیچیدگی‌های شبکه‌های عصبی، تعیین تعداد لایه‌ها و گره‌های پنهان و تعداد بازتاب‌ها یا تکرارهاست. برای تعیین این پارامترها، شماره‌های مختلف از چرخه حیات[1] و گره‌ها در لایه‌های پنهان[2] در نظر گرفته می‌شوند تا مقادیر مناسب ارزش از طریق آزمون و خطا مشخص گردد. برای تکرارها، مقادیر ۱۰۰۰، ۵۰۰۰، ۱۰۰۰۰ و ۲۰۰۰۰ مورد استفاده قرار گرفت. در مورد گره‌ها، با استفاده از عددی که نرم افزار به صورت پیش فرض بر اساس مدل و سایر داده‌های مربوط به نیمی از تعداد گره‌های ورودی محاسبه می‌کند، به‌ترتیب ۳ و ۲۰ گره مشخص گردید.

در مورد شبکه‌های بیزی، دو روش جهت یادگیری مدل برای شناسایی مرتبط‌ترین متغیرها و بهبود زمان پردازش و عملکرد الگوریتم مورد ارزیابی قرار گرفتند. ۱- الگوریتم TAN و ۲- الگوریتم Markov Blanket (که در نرم‌افزار SPSS موجود است). به همین ترتیب، یک آزمون مستقل از حداکثر

1. Epochs
2. Hidden Nodes

احتمال و آزمون مجذور کای برای یادگیری پارامتری بکار رفت. در هر دو بخش، بهترین نتایج تشخیص موارد با فاکتورهای نادرست، با استفاده از روش شبکه عصبی بدست آمدند.

نتایج بدست آمده، در جدول ۸ نشان داده شده و شاخص‌های زیر را شامل می‌شود که در تست گروهی بدست آمده است: (۱) حساسیت نشاندهنده نسبت موارد حاوی تقلب است که به‌درستی طبقه بندی شده‌اند، (۲) ویژگی نشاندهنده نسبت موارد بدون تقلب است که طبقه‌بندی در آن درست بوده است، (۳) سازگاری نشاندهنده نسبت موارد با و بدون تقلب است که طبقه‌بندی در آنها درست بوده است و (۴) نرخ خطا نشاندهنده نسبت موارد با و بدون تقلب است که در طبقه‌بندی مناسب قرار نگرفته‌اند.

جدول (۸)- نتیجه آزمایش مربوط به تقلب و یا عدم تقلب فاکتورهای جعلی

ردیف	روش خوشه‌بندی[1]	روش	حساسیت[2] (٪)	ویژگی[3] (٪)	سازگاری[4] (٪)	نرخ خطا[5] (٪)
۱	MI-SM[6]	شبکه عصبی	۹۲/۶	۷۲/۹	۸۷/۲	۱۲/۸
۲	MI-SM	شبکه بیزین	۸۲/۳	۶۴/۱	۷۷/۹	۲۲/۱
۳	MI-SM	درخت تصمیم	۸۹/۰	۷۹/۰	۸۷/۰	۱۳/۰
۴	ME-LA[7]	شبکه عصبی	۸۸/۸	۵۹/۱	۷۲/۵	۲۷/۵
۵	ME-LA	شبکه بیزین	۷۳/۳	۶۶/۷	۷۰/۳	۲۹/۷
۶	ME-LA	درخت تصمیم	۷۹/۰	۸۵/۰	۸۲/۰	۱۸/۰

1. Segmento
2. Sensitivity
3. Specifity
4. Consistency
5. Error rate
6. Micro and Small: MI-SM
7. Medium and Large: ME-LA

منبع: یافته‌های تحقیق

در گروه بنگاه‌های خرد و کوچک نیز آزمایش ۱ نشان داد که ۹۲/۶٪ از موارد تقلب در طبقه‌بندی‌های صحیح قرار گرفته‌اند، این درحالی است که در گروه شرکت‌های متوسط و بزرگ، نسبت موارد تقلب که بدرستی طبقه‌بندی شده بودند، ۸۸/۸٪ بود. علاوه بر این، قدرت تعمیم مدل بسیار خوب بود، چراکه نتایج آزمون، مشابه نتایج آموزش شبکه بودند که در آن، تشخیص موارد بدون تقلب به‌ترتیب ۹۳/۷٪ و ۸۷/۴٪

بود. خروجی مدل شبکه عصبی برای بنگاه‌های خرد و کوچک، متغیرهای تاثیرگذار مرتبط با پرداخت مالیات بر ارزش افزوده و رفتار، مربوط به درآمد را به شرح زیر نشان می‌دهد. رابطه بین ترازنامه‌های اعتبار مالیاتی و اعتبارات میانگین، بدهی‌های کل بر اساس فاکتور صادرشده، رابطه بین درآمد پول و دارایی‌ها، و رابطه بین مالیات بر ارزش افزوده پرداخت شده و درآمد اعلام شده و در مورد شرکت‌های متوسط و بزرگ، مهم‌ترین متغیرها در ارتباط با رابطه بین توازن اعتبار مالیاتی و میانگین اعتبارات، حساب‌های قابل پرداخت به شرکت‌های مرتبط، مجموع بدهی‌ها، نسبت اعتبارات مالیاتی بر اساس فاکتورها و مالیات بر ارزش افزوده تعیین شده در دوره موردنظر بدست آمد. لازم به‌ذکر است که هر سه مدل با استفاده از ابزار IBM SPSS Modeler نسخه ۱۸,۰ اجرا گردیدند.

۵- نتیجه‌گیری

روش‌های خوشه‌بندی و طبقه‌بندی مورد استفاده برای شناسایی مالیات‌دهندگانی که رفتار مالی تقلب یا فاقد تقلب در ارتباط با استفاده از فاکتورهای جعلی داشتند، نشان می‌دهد که شناسایی برخی از ویژگی‌های متمایز بین یک یا چند گروه امکان پذیر بوده و می‌تواند با واقعیت انطباق داشته باشد. همچنین روش گاز عصبی نشان داد که شناسایی برخی متغیرهای مربوطه برای متمایزسازی رفتار تقلب و فاقد تقلب لزوماً در ارتباط با استفاده و فروش فاکتورهای جعلی امکان‌پذیر نیست. هرچند، روش کونن هیچ‌گونه متغیر رفتاری خاص در ارتباط با استفاده از فاکتورهای نادرست ارائه نداد، اما مشخص شد که خوشه‌ها در سیستم مالیاتی موثر بوده و بیشترین تأثیر را در شکل دادن به گروه‌ها دارند.

روش درخت تصمیم‌گیری که برای مواردی اعمال شد که در آن، نتایج تقلب یا عدم تقلب مشخص بود، به‌عنوان یک تکنیک خوب برای تشخیص متغیرهای متمایزکننده تقلب و درستکاری معرفی گردید. علت این امر آنست که در هنگام تجزیه و تحلیل توزیع متغیرها در هر گروه، مشخص می‌شود که موارد تقلب معمولاً حاوی مقادیر ارزش بالاتری از متغیرها هستند، بنابراین متمایزسازی طیف‌هایی که شانس بروز تقلب یا عدم تقلب در آنها وجود دارد، امکان‌پذیر است. از سوی دیگر، بنابر اظهارنظر متخصصان و کارشناسان، نتایج با مشاهدات واقعی مطابقت داشته است. بنابراین، در مورد بنگاه‌های خرد و کوچک، متغیرهایی که امکان متمایزسازی تقلب و عدم تقلب را میسر می‌سازند، عمدتاً با درصد اعتبارات مالیاتی تولیدشده توسط فاکتورها مرتبط هستند که از نظر اعتبار کل و ممیزی‌های قبلی با نتایج منفی مورد ارزیابی قرار می‌گیرند.

احتمال عدم فریبکاری مالیات‌دهندگانی که در گذشته بارها و بارها حسابرسی شده و سوءرفتاری در

عملکرد آنها مشاهده نشده است، در آینده هم بیشتر از دیگران است. از سوی دیگر، در صورتی‌که اعتبار مالیات‌دهندگان با آیتم‌های دیگری به‌غیر از فاکتورها (دارایی‌های ثابت و غیره) گره خورده باشد، احتمال اینکه از فاکتورسازی برای حمایت از ادعاهای خود استفاده کنند، بیشتر خواهد بود. از دیگر متغیرهای مهم می‌توان به تعداد فاکتورهای صادرشده در طول یک سال و ارتباط آن با فاکتورهای مهرشده در دو سال گذشته، مقدار کل مالیات بر ارزش افزوده در طول سال اعلام شده، نسبت ترازنامه‌های میانگین اعتبار مالیاتی و ممیزی‌های مثبت قبلی و سابقه تخلفات و بی نظمی در ارتباط با فاکتورها اشاره کرد.

مهمترین متغیرها در شرکت‌های متوسط و بزرگ عبارتند از: میزان اعتبار مازاد انباشته شده در دوره‌های قبلی، درصد اعتبار مرتبط با صورتحساب‌ها، رابطه بین هزینه‌ها و دارایی‌ها، سطح عدم رسمیت در شیوه‌های حسابداری و سن شرکت، و همچنین تعداد بی‌نظمی‌ها و تخلفات مالیاتی در ارتباط با فاکتورهای قبلی و میزان سفارشات قابل پرداخت و عدم پاسخگویی به اظهارنامه‌های مالیاتی در گذشته. در رابطه با مدل‌های تشخیص، مواردی که عملکرد بهتری داشتند، مدل‌های شبکه‌ای عصبی پرسپترون چند لایه بودند که بنابر مقاصد مطالعه، دارای یک لایه ورودی شامل متغیرهای توضیحی، لایه میانی پردازش و لایه خروجی بودند. در مورد کسب و کارهای خرد و کوچک، درصد موارد تقلبی که به درستی تشخیص داده شده بود، ۹۲ درصد تعیین گردید، در حالی‌که در مورد مشاغل متوسط و بزرگ، این درصد، ۸۹٪ بود. طبق نتیجه بدست آمده و با توجه به اینکه در عمل، تنها یک گروه کوچک از شرکت‌ها را در هر سال می‌توان مورد نظارت قرار داد، استفاده از تکنیک‌های داده کاوی کمک موثر و شایانی در کشف تقلب خواهد کرد.

پیشنهادات برای پژوهش‌های آتی

توصیه می‌شود ترکیبی از نتایج بدست آمده با شبکه‌های عصبی، درخت تصمیم‌گیری و شبکه‌های بیزی جهت حسابرسی مواردی که احتمال فریبکاری شان بیشتر است، و همچنین بالاترین احتمال تقلب را دارند، انجام شود. در نهایت، برای آزمون مدل تشخیص واقعی توسعه‌یافته و سازگاری هرچه بیشتر آن با یافته‌های قبلی، اجرای آن در فعالیت‌های زمینه کاری حاضر، نقشی حیاتی برای تعیین سطح دقت طبقه‌بندی مالیات‌دهندگان منتخب در گروه نمونه دارد. اجرای یک برنامه آزمایشی که هر دو بخش اقتصادی مورد بررسی در این مطالعه را مدنظر قرار دهد، نیز پیشنهاد می‌شود که البته باید از نظر اثرگذاری واقعی مدل قطعی و قابل دفاع باشد. در راستای مطالعات آتی، توصیه می‌شود متغیرهای رفتاری و سابقه دار جدید در ارتباط با ممیزی‌های خاص و سطح پوشش آنها ایجاد شده، سایر روش‌های پیش پردازش و انتخاب متغیرها و همچنین، تکنیک‌های اعتبارسنجی متقاطع برای کشف و پیاده‌سازی سایر تکنیک‌های

داده کاوی در نظر گرفته شوند تا به بهبود تشخیص موارد با و بدون تقلب کمک شود.

فهرست منابع

1. Davia, H. R., Coggins, P., Wideman, J., & Kastantin, J. (2000). Accountant's Guide to Fraud Detection and Control (2nd Ed.).
2. Harrison, G., & Krelove, R. (2005). VAT Refunds: A Review of Country Experience. International Monetary Fund (IMF).
3. Bergman, M. (2010). Tax Evasion and the Rule of Law in Latin America: The Political Culture of Cheating and Compliance in Argentina and Chile. Penn State University Press.
4. Schneider, F., & Enste, D. (2000). Shadow Economies: Size, Causes and Consequences. Journal of Economic Literature, XXXVIII, 77–114.
5. Bonchi, F., Giannotti, F., Mainetto, G., & Pedreschi, D. (1999). A Classification-based Methodology for Planning Audit Strategies in Fraud Detection. In Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 175–184.
6. Cechhini, M., Aytug, H., Koehler, G., & Pathak, P. (2010). Detecting Management Fraud in Public Companies. Management Science, 56, 1146–1160.
7. Chena, H., Huang, S., & Kuo, C. (2009). Using the Artificial Neural Network to Predict Fraud Litigation: Some Empirical Evidence from Emerging Markets. Expert Systems with Applications, 36, 1478–1484.
8. Bonchi, F., Giannotti, F., Mainetto, G., & Pedreschi, D. (1999). A Classification-based Methodology for Planning Audit Strategies in fraud Detection. In Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 175–184.
9. Denny, W., & Christen, P. (2007). Exploratory Multilevel Hot Spot Analysis: Australian Taxation office Case Study. In Conferences in Research and Practice in Information Technology (Vol. 70, pp. 73–80). CRPIT Press.

10. Dubin, J. (2007). Criminal Investigation Enforcement Activities and Taxpayer Noncompliance. Public Finance Review, 35, 500–529.
11. Lundin, E., Kvarnstomr, H., & Jonsson, E. (2003). Synthesizing Test Data for Fraud Detection Systems. In Proceedings of the 19th Annual Computer Security Applications Conference (pp. 384–394). CSAC Press.
12. P. Castellon González, J.D. Velasquez / Expert Systems with Applications 40 (2013). 1427–1436.
13. Dubin, J. (2007). Criminal Investigation Enforcement Activities and Taxpayer Noncompliance. Public Finance Review, 35, 500–529.
14. Han J., M. Kamber, Data Mining: Concepts and Techniques (Second ed.), Morgan Kaufmann Publishers, 2006, pp. 285–464.
15. Myatt Glenn, J. (2007). Making Sense of Data, a Practical Guide to Exploratory Data Analysis and Data Mining. Wiley Interscience.
16. OECD (1999). Compliance Measurement, Practice Note. Centre for Tax Policy and Administration, Tax Guidance Series. General Administrative Principles – GAP004 Compliance Measurement. OECD Press.
17. OECD (2004a). Compliance Risk Management, Managing and Improving Tax Compliance. Forum on Tax Administration Compliance Subgroup. Centre for Tax Policy and Administration. OECD Press.
18. OECD (2004b). Compliance Risk Management, Audit Case Selection Systems. FoRum on Tax Administration Compliance Subgroup. Centre for Tax Policy and Administration. OECD Press.
19. US Government Accountability Office (2004). Data Mining: Agencies have Taken Key Steps to Protect Privacy in Selected Efforts, but Significant Compliance Issues remain.GAO Press.
20. Filippone, M., Camastra, F., Masulli, F., & Rovetta, S. (2008). A survey of Kernel and Spectral Methods for Clustering. Pattern Recognition, 41, 176–190.

21. The Application of Data Mining Techniques in Financial Fraud Detection: A Classification Framework and an Academic Review of literature; Decision Support Systems, vol. 50(3), 2010, pp. 559-569.
22. Zhou W., G. Kapoor, Detecting Evolutionary Financial Statement Fraud, Decision Support Systems, Vol. 50(3), 2011, pp. 570-576.
23. Watkinsa, R. C., Reynoldsa, K. M., Demaraa, R., Georgiopoulosa, M., Gonzaleza, A., & Eaglina, R. (2003). Tracking Dirty Proceeds: Exploring Data Mining Technologies as Tools to Investigate Money Laundering. Police Practice and Research: An International Journal, 4, 163–178.
24. Elliott R.K., and J.J. Willingham, Management Fraud: Detection and Deterrence, Petro Celli Books, NewYork, 1980, p.4.
25. Ata A., Ibrahim H. Seyrek, the Use of Data Mining Techniques in Detecting Fraudulent Financial Statements: An Application on Manufacturing Firms, Suleyman Demirel University, The Journal of Faculty of Economics and Administrative Sciences, Vol. 14(2), 2009, pp. 157-170.
26. Bose I., R.K. Mahapatra, Business Data Mining, a Machine Learning Perspective, Information Management, Vol. 39, 2001, pp. 211–225.
27. US Government Accountability Office (2008). Lessons Learned from other Countries on Compliance Risks, Administrative Costs, Compliance Burden and Transition. Report to Congressional Requesters. GAO Press.