تحقیقات مالی
Journal of Financial Research

دانشکدهٔ مدیریت دانشگاه تهران
دورهٔ ۱۷، شمارهٔ ۲
پاییز و زمستان ۱۳۹۴
ص. ۲۳۸- ۲۱۹

ارزیابی مالیات عملکرد شرکت‌ها و تحلیل روندهای مالیاتی با استفاده از الگوریتم‌های داده‌کاوی

بابک سهرابی[۱]، ایمان رئیسی وانانی[۲]، وحیده قانونی شیشوان[۳]

چکیده: همواره فاصلهٔ قابل‌توجهی میان مالیات ابرازی شرکت‌ها و مالیات تشخیصی آن‌ها وجود دارد که منجر به عدم رعایت عدالت میان مؤدیان شده است. یکی از علت‌های دشواربودن رعایت عدالت، شناسایی مؤدیان بر مبنای رفتار مالیاتی و برخورد مناسب با آن‌هاست. هدف اصلی پژوهش حاضر طراحی سیستم پیش‌بینی و تحلیل رفتار مالیاتی شرکت‌هاست. این سیستم کمک می‌کند تا با بهره‌گیری از متغیرهای کلیدی ارزیابی عملکرد مالیاتی، رفتار مالیاتی شرکت‌ها شناسایی و تحلیل شود. این سیستم برای سازمان امور مالیاتی کشور به‌منظور ارزیابی ریسک مالیاتی شرکت‌ها طراحی شده است و بر مبنای آن، ریسک مالیاتی شرکت‌ها به سه گروه پرریسک، با ریسک مالیاتی متوسط و کم‌ریسک تقسیم‌بندی شده است. همچنین، به کمک الگوریتم‌های خوشه‌بندی و طبقه‌بندی، خوشه‌های مالیاتی مشتریان شناسایی و درخت تصمیمی با دقت ۸۰٪ طراحی شد که رفتار مالیاتی هر یک از خوشه‌ها را بررسی و تحلیل می‌کند و با اضافه‌شدن شرکت‌های جدید به فهرست شرکت‌های مالیات‌دهنده، رفتار مالیاتی آن‌ها را نیز پیش‌بینی می‌نماید.

واژه‌های کلیدی: ارزیابی مالیاتی، خوشه‌بندی، پیش‌بینی، تحلیل روند، داده‌کاوی.

۱. استاد گروه مدیریت فناوری اطلاعات، دانشکدهٔ مدیریت، دانشگاه تهران، تهران، ایران
۲. استادیار گروه مدیریت صنعتی، دانشکدهٔ مدیریت و حسابداری، دانشگاه علامه طباطبایی، تهران، ایران
۳. دانشجوی کارشناسی ارشد مدیریت فناوری اطلاعات، دانشکدهٔ مدیریت و حسابداری، دانشگاه تهران، تهران، ایران

تاریخ دریافت مقاله: ۱۳۹٤/۰۲/۰۷
تاریخ پذیرش نهایی مقاله: ۱۳۹٤/۰٦/۱۸
نویسندهٔ مسئول مقاله: بابک سهرابی
E-mail: bsohrabi@ut.ac.ir

مقدمه

مالیات مهم‌ترین منبع مالی برای تأمین درآمدهای عمومی و یکی از کاراترین و مؤثرترین ابزارهای سیاست مالی در دنیاست که دولت به واسطهٔ آن بسیاری از خدمات اجتماعی و رفاهی را در خدمت مردم قرارمی‌دهد و به بسیاری از فعالیت‌ها و جریانات اقتصادی و اجتماعی سمت‌وسو می‌بخشد (وو، او، چانگ و ین، ۲۰۱۲). اتکا به درآمدهای مالیاتی منبع درآمدی پایدار و درون‌زای ملی را فراروی دولت قرارمی‌دهد (عباسیان، محمودی و شاکر، ۱۳۹۱).

همواره تلاش برای بهره‌گیری از ابزارهایی که سود یا درآمد شرکت را کم و هزینه‌های شرکت را بیش از واقع نمایش می‌دهد وجود دارد تا مالیات کمتری پرداخت شود (مهدعیسی، یوسف و مهدلی، ۲۰۱۴). اجتناب مالیاتی از جمله مشکلاتی است که از قوانین و مقررات مالیاتی نشأت می‌گیرد و این امر در میان مؤدیان مالیات بر درآمد شرکت‌ها متداول‌تر است، زیرا بخش بزرگی از درآمدهای مالیاتی دولت را تشکیل می‌دهد (پوپا، ۲۰۱۴).

در حال حاضر ممیزان مالیاتی با چالش شناسایی و جمع‌آوری مالیات از افرادی روبه‌رویند که به طور موفقیت‌آمیزی از پرداخت مالیات صحیح فرارمی‌کنند. به منظور رویارویی با این چالش ممیزان مالیاتی به منابع محدود و راهبردهای حسابرسی سنتی تجهیز شده‌اند که زمانبر و کسالت‌آور است (شیون و سی اس، ۲۰۱۲).

فقدان پیش‌بینی علمی و برآورد توان مالیاتی شرکت‌ها موجب شده است تا پیش‌بینی بر اساس عادت سنواتی صورت گیرد (حسنی، شبان، مختاری‌مسینایی و مودی، ۱۳۹۱). استفاده از ابزارها و روش‌های پیش‌بینی سنتی خطای بالایی دارد و به عملکرد ضعیف‌تری می‌انجامد (فلاح‌پور،گل‌ارضی و فتوره‌چیان، ۱۳۹۲). با پیشرفت فناوری، ابزارهای مدرنی به منظور یافتن موارد عدم رعایت و عدم تطبیق درآمدهای گزارش‌شده با سایر منابع به‌وجود آمده است (شیون و اس‌سی، ۲۰۱۲). داده‌کاوی ابزارهای مفیدی جهت ارزیابی مالیاتی شرکت‌ها ارائه می‌دهد. در این مقاله هدف این است که پس از پیش‌پردازش داده‌ها با استفاده از تکنیک‌های داده‌کاوی مدل داده‌ای طراحی شود تا در ابتدا متغیرهای کلیدی مؤثر بر رفتارهای مؤدیان مالیاتی شناسایی شود و با بهره‌گیری از آن، مؤدیان را طبقه‌بندی کرد تا تحلیل رفتار آن‌ها را سهولت بخشد.

پیشینهٔ نظری پژوهش

تعریف مالیات در طول زمان دستخوش تغییراتی شده و هر صاحب‌نظری سعی کرده است مفهوم مالیات را براساس دیدگاه خویش تعریف کند (عبدالاسلام و عبدمناف، ۲۰۱۴). بیستریسیو[1] اخذ

1. Gheorghe Bistriceanu

مالیات را اجباری و جبران‌ناپذیر می‌داند که از درآمد افراد بدون تأثیر مستقیم برای تأمین بودجهٔ دولت‌ها گرفته می‌شود (عبدالاسلام و عبدمناف، ۲۰۱۴).

براساس نظر بانس، مالیات روش دریافت بخشی از درآمد یا ثروت خصوصی اشخاص و کسب‌وکارهاست که جبران‌ناپذیر است و تأثیر مستقیم و فوری بر وظایف و کارکرد دولت‌ها ندارد (عبدالاسلام و عبدمناف، ۲۰۱۴). در تعریف علمی و از جنبهٔ اقتصاددانان مالیات بر حسب تعریف سازمان همکاری و توسعهٔ اقتصادی[1] پرداختی است الزامی و بلاعوض. اطلاق صفت بلاعوض به‌مالیات‌ها از این جهت است که دولت در مقابل دریافت مالیات از واحدی نهادی، خدمتی را به صورت فردی به آن واحد عرضه نمی‌کند، هر چند با دریافت مالیات منابع مالی خود را افزایش دهد و با استفاده از این منابع، کالاها و خدماتی را برای سایر واحدها اعم از تک تک افراد یا کل جامعه فراهم آورد[2].

شرحی مختصر بر روش‌های داده‌کاوی

هدف داده‌کاوی،کشف دانش جدید، معتبر و قابل‌پیگیری با استفاده از ابزارهای هوش مصنوعی و آماری در حجم بالایی از داده است. داده‌کاوی، استخراج یا اقتباس دانش از مجموعهٔ داده‌هاست و به فرایندی گفته می‌شود که دانش را از داده‌ها استخراج می‌کند و این دانش در قالب الگوها و مدل‌ها بیان می‌شود. دو هدف اصلی داده‌کاوی، پیشگویی و توصیف است (رادفر، نظافتی و یوسف اصلی، ۱۳۹۳). به منظور ارزیابی شرکت‌ها و با توجه به تنوع حوزهٔ فعالیت، اندازهٔ شرکت، تفاوت آن‌ها در مالیات ابرازی و تشخیصی تصمیم بر آن شد که ابتدا شرکت‌ها بر مبنای رفتار مالیاتی آن‌ها طبقه‌بندی شوند و با توجه به طبقه‌بندی صورت‌گرفته و تفاوت رفتاری هر گروه، نسبت به آن‌ها تصمیم‌گیری شود. با توجه به عدم‌شناخت متغیرهایی که مستقیماً بر رفتار مالیاتی شرکت‌ها تأثیر می‌گذارد و پیچیدگی موجود در داده‌ها و عدم توانایی هوش انسانی در طبقه‌بندی شرکت‌ها روش‌های داده‌کاوی برای طبقه‌بندی و پیش‌بینی داده‌ها انتخاب شد. از میان روش‌های داده‌کاوی با توجه به هدف طبقه‌بندی شرکت‌ها بر مبنای رفتار مالیاتی آن‌ها، الگوریتم‌های خوشه‌بندی انتخاب شد، از جمله روش‌های یادگیری نظارت‌نشده.

همچنین، از الگوریتم‌های مختلف خوشه‌بندی می‌توان نام برد که معمولاً در چهار گروه الگوریتم‌های تقسیم‌بندی، سلسله‌مراتبی، مبتنی بر توزیع و مبتنی بر شبکه قرارمی‌گیرد (کرمی و جانسون، ۲۰۱۴). در ادامه به توصیف برخی از این الگوریتم‌ها می‌پردازیم.

---

1. Organization for Economic Cooperation and Development
2. Amar.org.ir, 2015

الگوریتم K-Means

روش K-means روش خوشه‌بندی است که برای تحلیل داده‌ها و بررسی مشاهدات داده‌ای بر اساس مکان یا فاصله میان نقاط دادهٔ ورودی به‌کار می‌رود(گوش و کومادوبی، ۲۰۱۳). هر مرکز خوشه از طریق محاسبهٔ میانگین مختصات نقاط هر خوشه حاصل می‌شود (گوش و کومادوبی، ۲۰۱۳).

مراحل پیاده‌سازی K-Means عبارت است از ۱. انتخاب تعداد خوشه‌های مطلوب، ۲. انتخاب نقاطی به عنوان حدس اولیه از مراکز خوشه‌ها، ۳. بررسی هر نقطه در مجموع داده و نسبت‌دادن آن به خوشه‌ای که با مرکز آن کمترین فاصله را دارد، ۴. پس از قرارگیری هر نقطه در یک خوشه دوباره مراکز خوشهٔ جدید را محاسبه می‌کنیم. مراحل ۳ تا ۴ دوباره تکرار می‌شود تا زمانی که خوشهٔ نقطه‌ای تغییر نکند یا مراکز خوشه تغییری نداشته باشد (گوش و کومادوبی، ۲۰۱۳؛ وو، او، چنج و ین، ۲۰۱۲).

الگوریتم K-Medoids

الگوریتم خوشه‌بندی K-Medoids از روش‌های تقسیم‌بندی در مبحث خوشه‌بندی است که داده‌ها در هر خوشه بیشترین شباهت را با یکدیگر و بیشترین تفاوت را با خوشه‌های دیگر دارد (ونتاین، زونگ‌شنگ و ان، ۲۰۱۳). در الگوریتم K-Medoids قبل از اینکه فاصلهٔ داده‌های دیگر از هر مرکز خوشه محاسبه شود، K نقطه به صورت تصادفی از nداده به عنوان مرکز خوشه انتخاب می‌شود که مرکز مشخص‌شده میانه است (ونتاین، زونگ‌شنگ و ان، ۲۰۱۳). سپس، هر نقطه به نزدیک‌ترین خوشه نسبت داده می‌شود. این روش تکرارشونده برای تغییر مراکز خوشه ادامه می‌یابد تا بهترین خوشه‌بندی حاصل شود (ونتاین، زونگ شنگ و ان، ۲۰۱۳).

الگوریتم DBSCAN

روش DBSCAN[1] مخفف خوشه‌بندی مکانی بر مبنای چگالی در کاربردهای نویز است که خوشه‌بندی مبتنی بر توزیع است (کرمی و جانسون، ۲۰۱۴؛ اندرید و همکاران، ۲۰۱۳). الگوریتم DBSCAN به دو پارامتر ورودی نیاز دارد: شعاع هر خوشه[2] و حداقل نقطه در درون هر خوشه[3] (کرمی و جانسون، ۲۰۱۴). بر مبنای این پارامترها، داده‌کاوی انجام و خوشه‌ها تفکیک می‌شود.

---

1. Density Based Spatial Clustering of Applications with Noise
2. EPS
3. MinPts

الگوریتم Linkage

در روش خوشه‌بندی سلسله‌مراتبی از بالا به پایین[1] ابتدا هر دادهٔ خوشه‌ای مجزا در نظر گرفته می‌شود و طی فرایندی تکراری در هر مرحله خوشه‌هایی که شباهت بیشتری با یکدیگر دارد به صورت بازگشتی ترکیب می‌شود. در نهایت، یک خوشه یا تعداد مشخصی خوشه حاصل می‌شود (لویس، ملو و وایت، ۲۰۱۲).

طبقه‌بندی

طبقه‌بندی فرایند یافتن مدلی برای توضیح گروه‌ها یا مفاهیم متمایز است. هدف مدل طبقه‌بندی تعیین گروه داده‌ها در مجموعهٔ داده‌ای جدید است. مدل طبقه‌بندی شامل دو گام آموزش و طبقه‌بندی است. در گام آموزش داده‌ها به همراه گروهی که به آن تعلق دارد آموزش داده می‌شود تا مدل را خلق کند. مدل با بهره‌گیری داده‌های تست آزمون می‌شود تا دقت مدل اندازه‌گیری شود. اگر دقت مدل پذیرفتنی باشد، مدل برای پیش‌بینی گروه داده‌ها در مجموعهٔ داده‌ای جدید به کار گرفته می‌شود (روکاچ و مایمون، ۲۰۰۸؛ نارپراتمی و سیتانگانگ، ۲۰۱۵).

پیشینهٔ تجربی

جدول ۱ شامل خلاصه‌ای از مطالعات صورت‌گرفته در حوزهٔ امور مالیاتی است که در ترکیب با روش‌های داده‌کاوی انجام پذیرفته است.

جدول ۱. مطالعات صورت‌گرفته در داخل و خارج از کشور

عنوان مقاله	محقق	روش تحقیق	نتیجهٔ مقاله
مدل یکپارچه در پیش‌بینی تقلب مالیاتی شرکت‌ها[2]	مهد یوسف، لینگ لایی (۲۰۱۴)	درخت تصمیم	طراحی چارچوبی برای فرار مالیاتی بر مبنای فاکتورهای شناختی، قانونی، کنترل درونی و بیرونی و چهار بعد فرار مالیاتی
شناسایی مؤدیان مالیاتی با صورتحساب‌های نادرست از طریق تکنیک‌های داده‌کاوی[3]	گنزالس، ولسکواز (۲۰۱۳)	خوشه‌بندی و طبقه‌بندی	بهره‌گیری از تکنیک‌های خوشه‌بندی و طبقه‌بندی در شناسایی متغیرهای کلیدی مؤثر در شرکت‌های کوچک و شرکت‌های بزرگ و متوسط و ارائهٔ مدلی برای پیش‌بینی تقلب مالیاتی

1. Bottom-up
2. An integrative model in predicting corporate tax fraud
3. Characterization and detection of taxpayers with false invoices using data mining techniques

ادامهٔ جدول ۱

عنوان مقاله	محقق	روش تحقیق	نتیجهٔ مقاله
داده‌کاوی در مدیریت مالیات، به کارگیری تحلیل در افزایش قبول مالیاتی[1]	مارتیکاینن (۲۰۱۲)	تکنیک‌های داده‌کاوی	بهره‌گیری از روش‌های داده‌کاوی در ارائهٔ چارچوبی برای مدیریت فرایند مالیات و ارائهٔ مدل پیش‌بینی مالیات
مدل‌سازی رفتار متقلبانهٔ مؤدیان مالیاتی با استفاده از داده‌کاوی در شناسایی تقلب مالیاتی؛ مطالعهٔ موردی کشور مراکش[2]	امور، تکیاوت (۲۰۱۲)	تکنیک‌های داده‌کاوی	ارائهٔ مدلی برای تشخیص وجود تقلب مالیاتی یا نبود آن با توجه به متغیرهای کلیدی حیاتی. راهنمایی برای حسابرس مالیاتی در شناسایی متغیرهای مؤثر بر تقلب مالیاتی
بررسی تفاوت درآمد مشمول مالیات طبق گزارش حسابرس مالیاتی و درآمد مشمول مالیات تشخیصی ادارهٔ امور مالیاتی	فضل‌زاده، نبی نجفی (۱۳۹۲)	مدل‌های آماری	بررسی وجود اختلاف بین درآمد مشمول ابرازی و درآمد تشخیصی و علل آن
توانایی نسبت‌های مالی در کشف تقلب در گزارش‌های مالی؛ تحلیل لاجیت	صفرزاده (۱۳۸۹)	تکنیک تحلیل لاجیت و مدل‌های آماری	ارائهٔ الگو قابل‌اتکا برای کشف تقلب در گزارش مالی، ارائهٔ شاخص‌های مؤثر در کشف تقلب مالیاتی
مدل‌سازی غیرخطی و پیش‌بینی درآمدهای مالیاتی کشور به تفکیک منابع مالیاتی	خالوزاده، حمیدی علمداری، زایر (۱۳۸۷)	مدل‌های آماری و شبکهٔ عصبی	پیش‌بینی مالیات با استفاده از روش‌های آماری و شبکهٔ عصبی

با توجه به مطالعات صورت‌گرفته و مدل‌های ارائه‌شده در مقالات بین‌المللی، داده‌های لازم گردآوری و مدلی برای مجموعه دادهٔ مورد بررسی طراحی شد که در ادامه به جزئیات آن می‌پردازیم.

روش‌شناسی پژوهش

جامعهٔ تحقیق حاضر شامل تمامی شرکت‌های پذیرفته‌شده در بورس است که به ارائهٔ گزارش حسابرسی مستقل پرداخته‌اند.

1. Data mining in tax administration –using analytics to enhance tax compliance
2. Taxpayers Fraudulent Behavior Modeling the Use of Data mining in Fiscal Fraud Detecting Moroccan

داده‌های ۶۹۰ شرکت طی سال‌های ۱۳۸۴ تا ۱۳۸۸ در قالب فایل Excel جمع‌آوری شد. با توجه به تعدد شرکت‌ها و تفاوت در فعالیت آن‌ها سعی شده است شرکت‌هایی به عنوان نمونه انتخاب شود که گزارش‌های مالی حداقل سه ساله داشته باشند که میزان مالیات تشخیصی آن‌ها را سازمان امور مالیاتی اعلام کرده باشد. بنابراین، نمونه‌گیری با توجه به داده‌های در دسترس است. متغیرها از طریق گزارش حسابرس مستقل شرکت‌های نمونه استخراج شد. تمامی متغیرها کمّی و پیوسته است. در مرحلهٔ پیش‌پردازش، داده‌های پرت[1] با به‌کارگیری تکنیک مربوط بررسی شد. در آغاز با مطالعهٔ پژوهش‌های صورت‌گرفته معیارهای کلیدی عملکرد (KPI) شناسایی شد. سپس، متغیرهای مورد نظر از گزارش‌های حسابرسان مستقل استخراج و با به‌کارگیری تکنیک‌های داده‌کاوی سعی در طراحی مدلی در ارزیابی و پیش‌بینی مالیات عملکرد سال‌های بعد و تحلیل روند مالیاتی شرکت‌ها شد. اعتبارسنجی مدل با به‌کارگیری ماتریس شباهت در ارزیابی خوشه‌بندی صورت گرفت. این روش‌ها به منظور انتخاب روش برتر خوشه‌بندی بررسی شد. همچنین، به منظور ارزیابی طبقه‌بندی صورت‌گرفته با الگوریتم‌های طبقه‌بندی، از روش شناسایی میزان امکان جایگزینی در درخت تصمیم استفاده شد. در نهایت، با توجه به مدل به‌دست‌آمده به تحلیل عملکرد شرکت‌ها و ارزیابی خوشه‌های مالیاتی آن‌ها پرداختیم. با توجه به بررسی مقالات انجام‌گرفته در داخل و خارج از کشور و مشاوره از رئیس گروه‌های مالیاتی و داده‌های در دسترس، متغیرهای جدول ۲ از بخش ترازنامه، سود و زیان، و توضیحات پیوست صورت‌های مالی استخراج شد. نکته اینکه در اظهارنامه‌ای که مؤدی تسلیم واحد مالیاتی می‌کند، همچنین گزارش حسابرس مستقل، درآمد ضایعات در سود و زیان آورده می‌شود.

در مقایسهٔ صورت‌های مالی شرکت‌ها با یکدیگر مشکلاتی همچون تفاوت در اندازهٔ شرکت‌ها از لحاظ دارایی، بدهی و سرمایه وجود داشت. به همین دلیل با بهره‌گیری از تجزیه و تحلیل نسبت‌های مالی سعی شد این مشکل برطرف شود. از این‌رو، متغیرهای جدول ۳ نیز در کنار متغیرهای جدول ۲ برای خوشه‌بندی جامعهٔ مالیاتی بررسی شد.

جدول۲. متغیرهای استخراج‌شده از صورت‌های مالی

نوع متغیر	نوع صورت مالی
وجوه نقد، دارایی جاری، دارایی، بدهی جاری، بدهی، سرمایه	ترازنامه
فروش خالص، بهای تمام‌شده، هزینه، درآمد، ضایعات، سودخالص	سود و زیان
مالیات ابرازی، مالیات تشخیص داده شده، استهلاک، مواد مصرفی	توضیحات تکمیلی

1. Outlier Detection

جدول ۳. متغیرهای استخراج‌شده با بهره‌گیری از تجزیه‌وتحلیل نسبت‌های مالی

متغیر	نام اختصاری	نسبت مالی	توضیحات
نسبت جاری	nesbat jari	نسبت‌های نقدینگی	دارایی جاری/ بدهی جاری (وضعیت مطلوب نقدینگی را نشان می‌دهد)
سرمایهٔ در گردش	sarmaye gardesh	نسبت‌های نقدینگی	سرمایه/ کل دارایی‌ها (توان نقدینگی را نشان می‌دهد)
موجودی نقد به دارایی جاری	M/DJ	نسبت‌های نقدینگی	اندازه‌گیری توان نقدینگی نسبت به دارایی
موجودی نقد به دارایی	M/D	نسبت‌های نقدینگی	
مواد مصرف به فروش	MM/F	نسبت‌های فعالیت	نشان‌دهندهٔ تعداد دفعات خرید و فروش در سال
مواد مصرف به بهای تمام‌شده	MM/BT	نسبت‌های فعالیت	
فروش به بهای تمام‌شده	F/BT	نسبت‌های فعالیت	نشان‌دهندهٔ میزان اختصاص فروش به بهای تمام شده، شامل مواد مصرفی، سربار و دستمزد مستقیم
بهای تمام‌شده به فروش	BT/F	نسبت‌های فعالیت	
فروش به دارایی	F/D	نسبت‌های فعالیت	نشان‌دهندهٔ کارایی دارایی
نسبت بدهی	B/D	نسبت‌های اهرمی	کل بدهی/ کل دارایی (ریسک مالی را به نمایش می‌گذارد)
بدهی جاری به بدهی	BJ/B	نسبت‌های اهرمی	
سود به فروش	S/F	نسبت‌های سودآوری	سود خالص (پس از کسر مالیات)/ فروش
نرخ بازده دارایی	bazdeh daraee	نسبت‌های سودآوری	سود خالص (پس از کسر مالیات)/ کل دارایی (نشان‌دهندهٔ بهره‌گیری از دارایی برای رسیدن به سود)
قدرت کسب سود دارایی	GD	نسبت‌های سودآوری	سود ناخالص (قبل از کسر مالیات)/ دارایی (نشان‌دهندهٔ بهره‌گیری از دارایی برای رسیدن به سود)
نرخ رشد مالیات تشخیصی به ابرازی	roshd	سایر	(درآمد مشمول مالیات تشخیصی، درآمد مشمول مالیات ابرازی)/ درآمد مشمول مالیات ابرازی
ضایعات به موجودی مواد	Z/MM	سایر	ضایعات یکی از مهم‌ترین منابع درآمد شرکت‌های تولیدی و یکی از مواردی است که ممیزان مالیاتی کمترین توجه را به آن دارند.
ضایعات به سایر درآمدها	Z/SD	سایر	
استهلاک به دارایی	E/D	سایر	
سایر درآمدها به هزینه	D/H	سایر	

پس از معرفی داده، پالایش داده‌ها در چند مرحله صورت پذیرفت. ابتدا، ردیف‌هایی که در آن‌ها برخی متغیرها فاقد مقدار[1] بودند حذف شد. در گام دوم، تمامی ردیف‌های مغایر با استانداردهای حسابداری حذف شد. این ردیف‌ها به شرح زیر پالایش شد (موارد زیر شامل رکوردهای اطلاعاتی است که به دلیل اشتباه حسابرسان مالیاتی، به غلط درج شده است. به همین سبب نیاز است که از داده‌های اصلی جدا و حذف شود):

- ردیف‌هایی که در آن‌ها دارایی جاری بیش از دارایی، بدهی جاری بیش از بدهی، بدهی بیش از دارایی است.
- ردیف‌هایی که مجموع بدهی و سرمایه بیش از دارایی و مواد مصرفی بیش از بهای تمام‌شده است.

و در آخر، با حذف ردیف‌هایی که در آن‌ها درصد رشد مالیات بیش از ۶ بود سعی شد تا داده‌های دارای انحراف زیاد (به اصطلاح داده‌های خارج از محدوده[2] در متون داده‌کاوی) حذف شود. در نتیجه، تعداد داده‌ها از ۶۹۰ به ۳۸۶ رکورد داده کاهش یافت. علت انتخاب رشد شش برابری، گستردگی دامنه و تعداد کم شرکت‌هایی بود که مالیات آن‌ها بیش از شش برابر رشد کرده است.

اعتبارسنجی متغیرها

به منظور بررسی میزان اثرگذاری متغیرها و سنجش اعتبار و روایی آن‌ها در ارزیابی موضوع تحقیق، آزمون تحلیل عاملی شامل آزمون‌های KMO[3] و بارتلت[4] روی داده‌ها انجام شد. با توجه به نتایج حاصل و KMO= ۰/۵۴۲ و تأیید فرض مخالف آزمون بارتلت، همبستگی لازم میان داده‌ها وجود دارد. مجموعهٔ داده‌های مورد بررسی، هفت عنصر کلیدی در مجموعه متغیرها نزدیک به ۷۴ درصد از واریانس مؤثر بر رشد مالیات را تبیین می‌کند که نشان‌دهندهٔ روایی و توانایی متغیرها در سنجش موضوع هدف است. با توجه به اینکه هدف این تحقیق یافتن عوامل جدید از طریق تحلیل عاملی نیست، به همین علت به ماتریس‌های چرخش به‌دست‌آمده از تحلیل عاملی توجه نشده است.

پس از تأیید اعتبار و توانایی تمامی متغیرهای شناسایی‌شده در سنجش هدف تحقیق، همبستگی متغیرهای کلیدی و اثرگذار بر مدل داده‌کاوی با بررسی تک‌تک متغیرهای تصمیم،

---

1. Missing Values
2. Outlier Data
3. Kaiser-Meyer-Olkin
4. Bartlett's Test

مرور پژوهش‌های صورت‌گرفته، تأیید روایی توسط ریاست گروه‌های فعال در این حوزه انجام شد. در نهایت سه متغیر قدرت کسب سود دارایی، نسبت سود به فروش و نسبت بهای تمام‌شده به فروش به عنوان متغیرهای ورودی و اثرگذار بر مالیات در مدل انتخاب شد. این متغیرهای مهم، قادر به تبیین اثر بقیهٔ متغیرهای مالی ذکرشده در تحقیق است.

یافته‌های پژوهش

مدل‌سازی

بسیاری از مسائل داده‌کاوی را می‌توان به صورت مسئلهٔ خوشه‌بندی بیان کرد، که در آن یک عامل هوشمند یا نیمه‌هوشمند باید بتواند بدون دردست‌داشتن هیچ اطلاعات زمینه‌ای، طبقه‌بندی منطقی از یک سری موارد در دسترس داشته باشد. نخست، با استفاده از تکنیک‌های مختلف خوشه‌بندی، به خوشه‌بندی شرکت‌ها پرداختیم.

با توجه به داده‌ها و پس از پیاده‌سازی ارزیابی کلینسکی هرباز[1] که یکی از تکنیک‌های ارزیابی به منظور تعیین تعداد خوشهٔ بهینه در نرم‌افزار متلب است، تعداد بهینهٔ سه خوشه برای الگوریتم‌های k-means, K-medoids, linkage انتخاب شد. در الگوریتم DBSCAN تعداد خوشه به‌عنوان ورودی الگوریتم مورد نیاز نیست.

اساس کار الگوریتم k-means میانگین است؛ یعنی، جمع میانگین همهٔ نقاط از مرکز خوشه‌ای که به آن تعلق دارد حداقل شود (کرمی و جانسون، ۲۰۱۴). پس از پیاده‌سازی، داده‌ها در سه خوشه به تعداد عناصر داده‌ای ۲۶۳، ۳۵ و ۸۸ تقسیم شد.

اساس کار الگوریتم K-medoids بر مبنای میانه است؛ یعنی، مجموع میانهٔ هر نقطه از مرکز خوشه‌ای که به آن تعلق دارد حداقل شود (ونتاین، زونگ‌شنگ و ان، ۲۰۱۴). نتیجهٔ پیاده‌سازی این الگوریتم به سه خوشه با تعداد ۲۶۱، ۹۰، ۳۵ رکورد مالیاتی منجر شد.

در گام سوم، روش‌های خوشه‌بندی سلسله‌مراتبی انتخاب شد که از میان آن‌ها الگوریتم linkage و روش فاصلهٔ درون‌مربعی[2] (الگوریتم واریانس حداقل) و معیار minkowski روی داده‌ها پیاده‌سازی شد. در این روش داده‌ها به صورت سلسله‌مراتبی درختی است که در محل تعداد خوشهٔ مورد نظر برش انجام می‌گیرد (آنیل و ریچارد، ۱۹۸۸). خوشه‌های حاصل شامل تعداد داده‌های ۲۷۳، ۷۸، ۳۵ در هر خوشه است.

---

1. Calinski Harabasz Evaluation
2. Ward

در گام آخر، از روش‌های خوشه‌بندی بر مبنـای چگـالی، از روش خوشـه‌بنـدی DBSCAN استفاده شد. روش کار بدین صورت است که در بررسـی داده بـه بررسـی همسـایگی آن داده در شعاع epsilon می‌پردازیم. داده‌ها در سه حالت قرارمی‌گیرد: کمتر از minpts که نویز است؛ دور و بر خلوتی اما در همسایگی نقطهٔ هسته‌ای که مرز شناخته می‌شود؛ یا بیش از minpts است که هسته نامیده می‌شود (اندریـد و همکـاران، ۲۰۱۳) کـه بـا epsilon=۱۰۰، minpts=۰/۶ چهـار خوشه با تعداد شرکت‌های ۲۴۳، ۸۷، ۴۸، ۸ حاصل شد.

در روش DBSCAN، از آنجا که خوشه‌های با تعداد هشت بـه خوشـه‌هـای بـا تعـداد ۲۴۳ نزدیک بود، در این خوشه ادغام و تعداد داده‌های موجود در خوشـه‌هـا بـه ترتیـب ۲۵۱، ۸۷، ۴۸ می‌شود. خلاصه‌ای از روش‌های خوشه‌بندی اعمال‌شده و نتایج آن‌ها در جدول ۴ آمده است.

جدول ٤. خوشه‌بندی‌های حاصل از روش‌های مختلف خوشه‌بندی

الگوریتم خوشه‌بندی	تعداد داده‌ها در هر خوشه	تنظیمات مربوط به هر روش		
K-means	۳۵,۸۸,۲۶۳	nCluster=۳	DistanceMetric='cityblock'	
K-medoids	۳۵,۹۰,۲۶۱	k=۳		
Linkage	۳۵,۷۸,۲۷۳	Metrics='minkowski'	Method= 'ward'	Ncluster=۳
DBSCAN	۸,۴۸,۸۷,۲۴۳	epsilon=۰/۶	MinPts=۱۰۰	

اعتبارسنجی خوشه‌ها

به منظور انتخاب خوشه‌بندی مناسب، ماتریس شباهت یکی از روش‌هـای ارزیـابی خوشـه‌بنـدی انتخاب شد. هرچه این ماتریس به شکل بلوکی- قطری نزدیک‌تر باشد، نشان‌دهندهٔ این است که داده‌های نزدیک به یکدیگر در یک خوشه قرارگرفته است و اعتبار تحقیق در سطح مطلوب‌تـری قراردارد.

ماتریس خوب ساختار بلوکی مربعی قوی، به‌خصوص در قطر اصلی، دارد. در ماتریس شباهت میانگین شباهت هر خوشه را وقتی در نظر می‌گیریم که شباهت درون‌خوشه‌ای مقایسه می‌شـود. این امتیاز میزان تمایز خوشه‌ها از یکدیگر را اندازه‌گیری می‌کنـد (لوسـون و فلـوش، ۲۰۱۲) کـه نسبت مقادیر ماتریس در بلوک‌های قطر اصلی به کل ماتریس است. این مقدار بین ۰ و ۱ خواهد بود. هرچه امتیاز به ۱ نزدیک باشد، خوشه‌بندی بهتر خواهد بـود (Clusterevaluation,۲۰۱۵). نتایج در جدول ۵ آمده است.

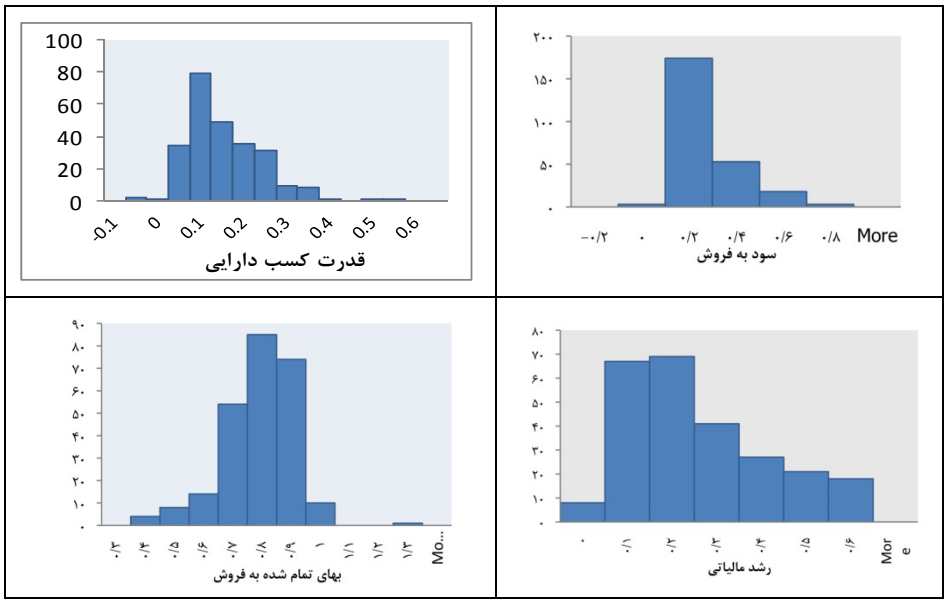
جدول ۵. ارزیابی خوشه‌بندی الگوریتم‌های مختلف خوشه‌بندی

الگوریتم خوشه‌بندی	امتیاز ماتریس شباهت	توضیحات
k-means	Similarity matrix= ۰/۵۶۰۷	عدم توانایی در شناسایی داده‌های با ریسک مالیاتی متوسط از داده‌های پرریسک مالیاتی
k-medoids	Similarity matrix= ۰/۵۵۵۷	وضعیت نسبت به k-means نامناسب‌تر است و قادر به جداسازی شرکت‌های پرریسک از ریسک متوسط نیست.
Linkage	Similarity matrix= ۰/۵۸۷۰	با وجود اشکال‌های موجود در روش‌های بالا، همچنین اختصاص دامنهٔ بزرگی از ردیف‌های داده‌ای به شرکت‌های کم‌ریسک، نسبت به دو روش بالا در شناسایی شرکت‌های پرریسک از ریسک متوسط بهتر عمل می‌کند.
DBSCAN	Similarity matrix= ۰/۶۹۵۵	این روش خوشه‌بندی به خوبی توانسته است داده‌های با ریسک متوسط را از داده‌های پرریسک جداسازی کند.

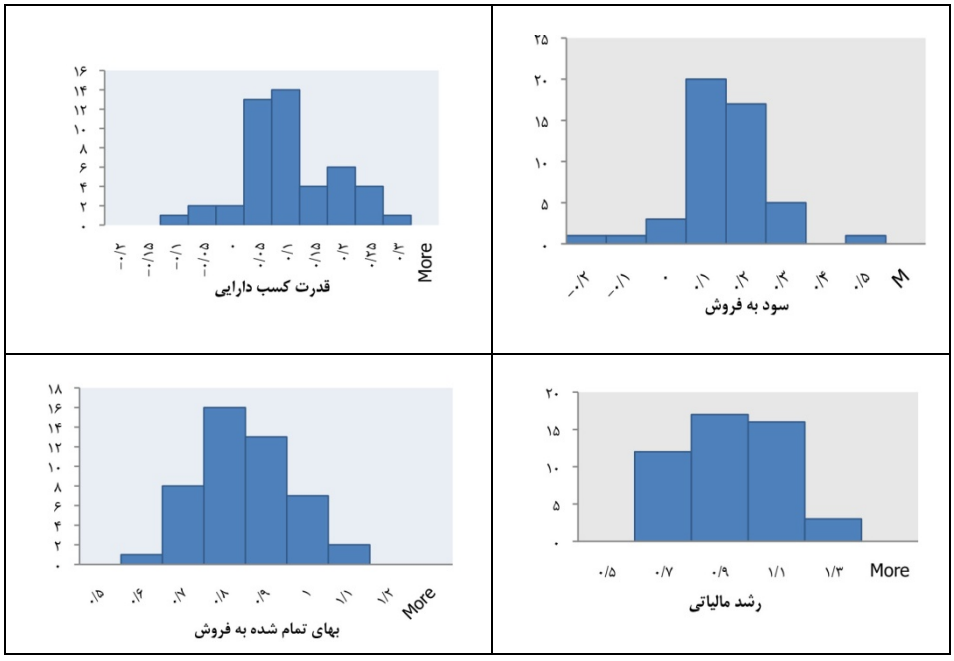
طبق جدول ۵، با توجه به امتیاز اختصاص‌داده‌شده، روش DBSCAN در خوشه‌بندی داده‌ها بهتر عمل کرده و توانسته است به خوبی شرکت‌های با ریسک متوسط را از شرکت‌های پرریسک متمایز کند؛ عملی که سایر الگوریتم‌های خوشه‌بندی در اجرای آن چندان موفق عمل نکردند.

تحلیل رفتار خوشه‌ها

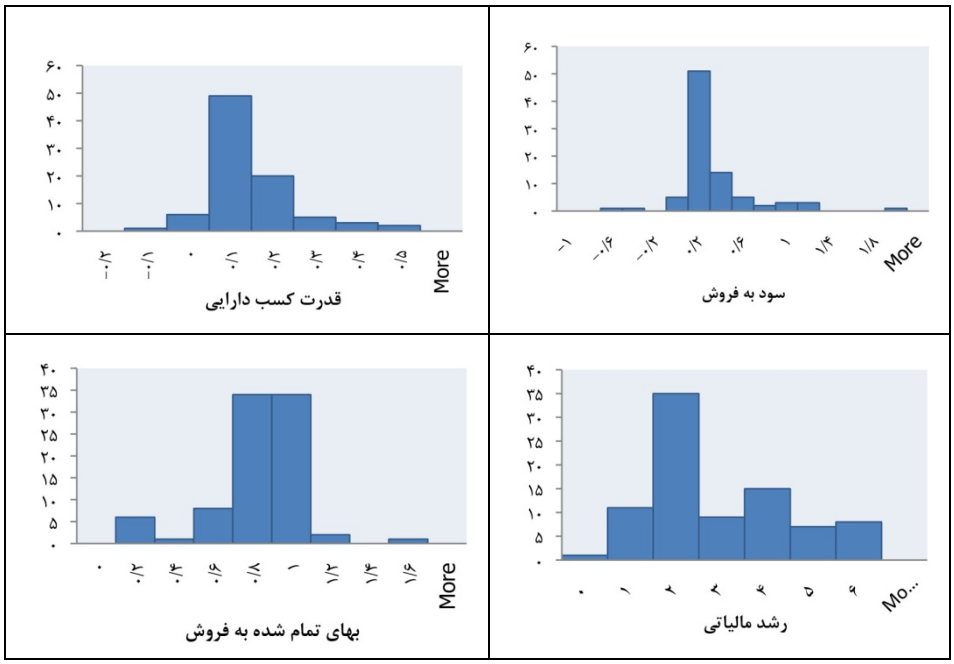
با انتخاب روش DBSCAN به عنوان روش برتر در خوشه‌بندی و با توجه به نمودارهای هیستوگرام و نمودار جعبه‌ای که روی خوشه‌بندی‌های ارائه‌شده صورت گرفت، می‌توان نتیجه گرفت که داده‌ها بر مبنای ریسک مالیاتی تقسیم‌بندی می‌شود. مدیریت ریسک مسیر ساختاریافته‌ای برای تحلیل و پاسخگویی به عدم‌اطمینان‌های آینده است (راعی، فلاح‌پور و عامری‌متین، ۱۳۹۱). در نتیجه، برنامه‌ریزی با توجه به طبقات مختلف ریسک مؤدیان انجام خواهد شد (رادفر، نظافتی و یوسف اصلی، ۱۳۹۳). بنابراین، با توجه به رشد مالیاتی شرکت‌ها در سه خوشهٔ ریسک مالیاتی بالا، ریسک مالیاتی متوسط و ریسک مالیاتی پایین تقسیم‌بندی شد. نتایج حاصل در شکل‌های ۱، ۲ و ۳ به صورت خلاصه بیان شده است. محور عمودی در تمامی نمودارها فراوانی شرکت‌هاست.



شکل ۱. شرکت‌هایی با ریسک مالیاتی پایین



شکل۲. شرکت‌هایی با ریسک مالیاتی متوسط



شکل۳. شرکت‌هایی با ریسک مالیاتی بالا

جدول ٦. تحلیل خوشه‌ها بر مبنای متغیرهای کلیدی مؤثر

نوع خوشهٔ متغیر	توضیحات	کم‌ریسک	با ریسک متوسط	پرریسک
GD	قدرت کسب دارایی	۰/۱۹ ≤ GD ≤ ۰/۰۷ میانه =۰/۱۲	۰/۱۲ ≤ GD ≤ ۰/۰۲ میانه =۰/۰۷	۰/۱۳ ≤ GD ≤ ۰/۰۴ میانه =۰/۰۷
SF	سود به فروش	۰/۲۲ ≤ SF ≤ ۰/۰۸ میانه =۰/۱۴	۰/۱۵ ≤ SF ≤ ۰/۰۳۵ میانه =۰/۰۹	۰/۲۷۵ ≤ SF ≤ ۰/۰۵ میانه =۰/۱
BTF	بهای تمام‌شده به فروش	۰/۸۳ ≤ BTF ≤ ۰/۶۸ میانه =۰/۷۷	۰/۸۷۵ ≤ BTF ≤ ۰/۷۵ میانه =۰/۸	۰/۸۴۲۵ ≤ BTF ≤ ۰/۶۷۲۵ میانه =۰/۸
RM	رشد مالیات	۰/۳۱۷ ≤ RM ≤ ۰/۰۸ میانه =۰/۱۷	۰/۹۸ ≤ RM ≤ ۰/۷ میانه =۰/۸۶	۳/۳۳۵ ≤ RM ≤ ۱/۳۳۵ میانه =۱/۹۵

با استفاده از نمودار جعبه‌ای تراکم داده‌ها، هر یک از خوشه‌بندی‌ها بررسی شد (جدول ۶). در جدول ۶، عدد اول چارک اول و عدد دوم چارک سوم را بیان می‌کند.

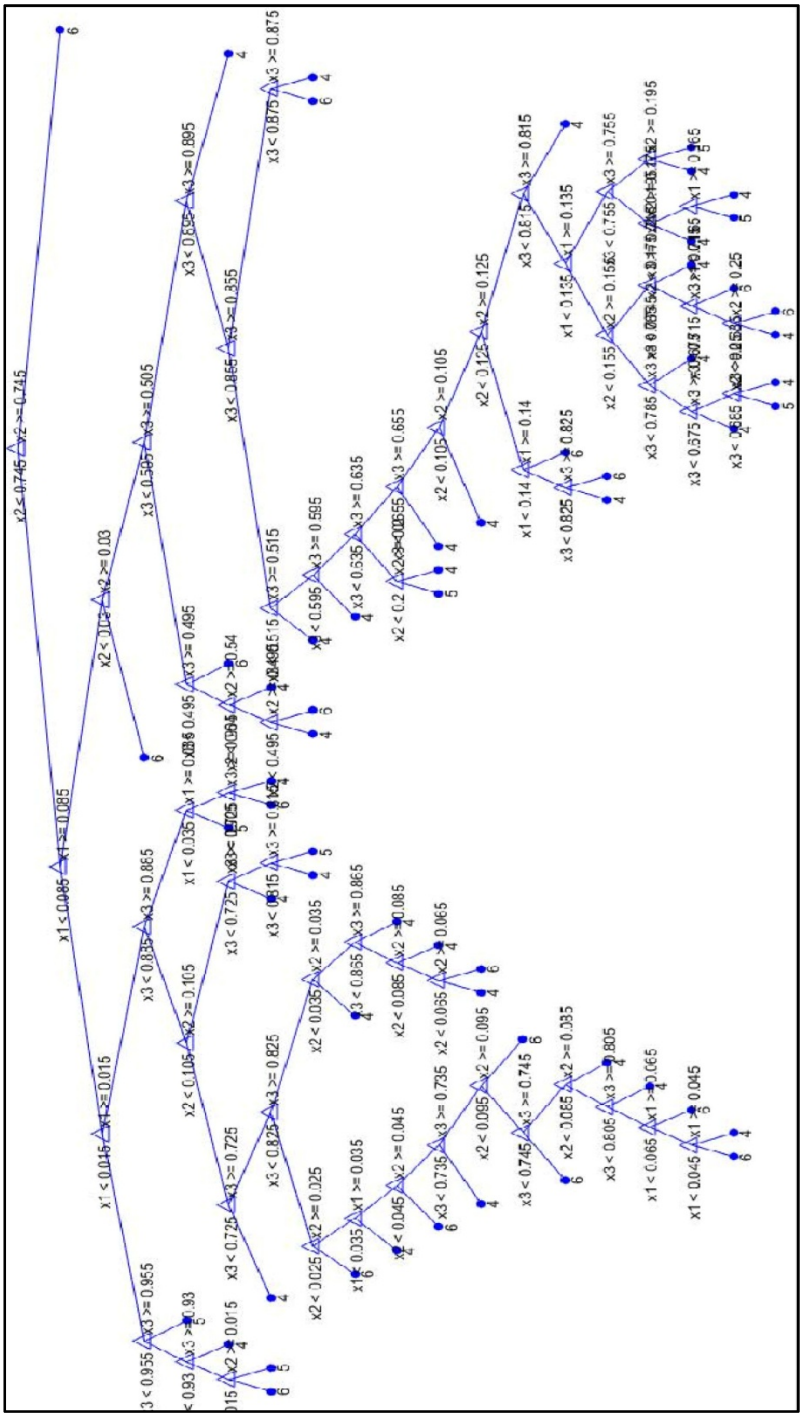
در خوشهٔ شرکت‌های با ریسک پایین که رشد مالیاتی اغلب آنها کمتر از ۵۰٪ است بهای تمام‌شده به فروش کمتر است. به این معنا که بخش کمتری از فروش به بهای تمام‌شده تعلق

گرفته که به سود ابرازی بالاتر می‌انجامد. نسبت سود به فروش بالای این خوشه نشان‌دهندهٔ این امر است که هزینه‌های عملیاتی و مالی، همچنین بهای تمام‌شده متناسب با حجم فروش و قدرت کسب دارایی بالا نشان از بهره‌گیری مناسب از دارایی در جهت سوددهی شرکت است. این خوشه را می‌توان شرکت‌هایی دانست که نیاز به بررسی و ممیزی در آن‌ها کم است. در خوشهٔ شرکت‌های با ریسک مالیاتی متوسط که رشد مالیاتی اغلب آن‌ها کمتر از ۱۰۰٪ است، قدرت کسب دارایی پایین، بهره‌گیری نامناسب از دارایی‌های شرکت در جهت سوددهی شرکت را نشان می دهد. بهای تمام‌شده نسبت به فروش بالاتر و نسبت سود به فروش کمتر نشان‌دهندهٔ متورم‌کردن بهای تمام‌شده و افزایش هزینه‌های مالی و عملیاتی شرکت به منظور کاهش سود ابرازی است. در برخورد با این خوشه باید با احتیاط بیشتری رفتار کرد و تعداد نمونهٔ بیشتری برای ممیزی و بررسی انتخاب کرد. در خوشهٔ شرکت‌های با ریسک مالیاتی بالا که رشد مالیاتی آن‌ها بیش از ۱۰۰٪ است، قدرت کسب دارایی پایین نشان از بهره‌گیری نامناسب از دارایی‌های شرکت در جهت سوددهی است. با توجه به دامنهٔ گستردهٔ بهای تمام‌شده به فروش و سود به فروش ناهماهنگی در این نسبت‌ها دیده می‌شود که نشان‌دهندهٔ عدم صحت اطلاعات واردشده در گزارش حسابرس مالیاتی است. بدین معنا که در مواردی سعی شده بهای تمام‌شده را متورم کند، اما هزینه‌های مالی و عملیاتی را کمتر نمایش دهد تا به گونه‌ای حجم فعالیت شرکت کمتر از واقع به نمایش گذاشته شود. به عبارتی، برخی شرکت‌ها سعی می‌کنند با گران نشان‌دادن مواد مصرفی، بهای تمام‌شده را متورم و با کمتر نشان‌دادن هزینه‌های عملیاتی حجم واقعی فعالیت خود را به نمایش نگذارند. اما، در کل، می‌توان گفت نسبت سود به فروش کمتر و نسبت بهای تمام‌شده به فروش بالاتر نشان از ریسک مالیاتی شرکت است. ممیزی و بررسی در این خوشه باید با جدیت صورت گیرد.

طبقه‌بندی داده‌ها

به منظور حفظ و ذخیره‌سازی دانشی که خوشه‌بندی می‌کند، می‌توان از روش‌طبقه‌بندی استفاده کرد. یکی از الگوریتم‌های مشهور در طبقه‌بندی داده‌ها، درخت تصمیم[1] است. درخت تصمیم شامل مجموعه‌ای از فیلدهای داده و روابط میان آن‌هاست[2]. از درخت تصمیم به منظور طبقه‌بندی شرکت‌های مالیاتی می‌توان استفاده کرد (شکل ۴). نتیجهٔ اعتبارسنجی درخت تصمیم در جدول ۷ ارائه شده است.

1. Decision Tree
2. www.mathworks.com



شکل ٤. درخت تصمیم طبقه‌بندی شرکت‌های مالیاتی

جدول ۷. اعتبارسنجی درخت تصمیم

تابع خطا	میزان خطا
خطای اندازه‌گیری‌شدهٔ درخت تصمیم (Resubstitution Loss)	۰/۲۰۲۱

خطای درخت تصمیم به میزان ۰/۲ بدین مفهوم است که درخت تصمیم ایجادشده، حداقل در ۸۰ درصد موارد طبقه‌بندی صحیحی را انجام می‌دهد و شرکت‌های جدید را در خوشه‌های مالیاتی مرتبط با آن‌ها به‌صورت دقیق طبقه‌بندی و رفتار آیندهٔ آن‌ها را پیش‌بینی می‌کند. در آینده، محققان قادر خواهند بود تا با استفاده از این درخت تصمیم، وضعیت شرکت جدید را مشخص کنند و نحوهٔ رفتار مالیاتی آن شرکت را بر اساس داده‌های ورودی شبیه‌سازی کنند. در واقع، بر اساس خوشهٔ به‌دست آمده، الگوهای رفتار مالیاتی در آینده و نحوهٔ مواجه با آن شرکت را سازمان‌های مالیاتی مشخص می‌کنند.

نتیجه‌گیری و پیشنهادها

روش‌های خوشه‌بندی و طبقه‌بندی به منظور شناسایی عوامل مؤثر بر روند مالیاتی شرکت‌ها به کار گرفته شد. پس از بررسی و مشاوره با رئیس گروه‌های مالیاتی سه متغیر سود به فروش، قدرت کسب دارایی، بهای تمام‌شده به فروش به عنوان متغیرهای حیاتی استفاده شد. روش‌های مختلف خوشه‌بندی روی داده‌ها، پس از پیش‌پردازش آن‌ها، اعمال شد. در نهایت، یکی از روش‌های خوشه‌بندی مبتنی بر چگالی (DBSCAN) با توجه به ارزیابی خوشه‌بندی صورت‌گرفته به عنوان روش برگزیده در خوشه‌بندی انتخاب شد. خوشه‌بندی ارائه‌شده داده‌ها را در سه خوشهٔ شرکت‌هایی با ریسک مالیاتی بالا، متوسط و پایین تقسیم‌بندی می‌کند.

به منظور حفظ دانش ایجادشده و با توجه به پیچیدگی خوشه‌های به‌دست‌آمده، الگوریتم درخت تصمیم یکی از روش‌های طبقه‌بندی انتخاب شد تا به ارائهٔ مدلی پیش‌بینی ریسک مالیاتی شرکت‌ها بپردازد. با استفاده از درخت تصمیم حاصل از این الگوریتم، امکان طبقه‌بندی شرکت‌هایی که اخیراً شرکت مالیاتی در نظر گرفته شده‌اند و شرکت‌هایی که در این تحقیق مشارکت نداشته‌اند نیز با اطمینان ۸۰ درصد فراهم شده است.

به منظور توسعهٔ تحقیق حاضر، پیشنهاد می‌شود که از روش‌های داده‌کاوی برای پیش‌بینی مالیات تشخیصی بر مبنای متغیرهای کلیدی مالیاتی استفاده شود. همچنین، از داده‌های مالیات بر ارزش افزوده در شناسایی متغیرهای کلیدی اثرگذار بر روند مالیات‌دهی شرکت‌ها بهره گرفته

شود. در نهایت، با به‌کارگیری روش‌های داده‌کاوی همچون شبکهٔ فازی ـ عصبی و الگوریتم تکاملی، به بهینه‌سازی روند و مبلغ دریافت مالیات از شرکت‌های مختلف اهتمام گردد.

References

Abasian, E., Mahmoodi, V. & Shaker, I. (2013). Forecast Error Analysis of State Tax Revenues in Iran. *Journal of Financial Research* 13(32): 109-132. (*in Persian*)

Abdulsalam, M. & Abd Manaf, N. (2014). Do trust and power moderate each other in relation to tax compliance? *Procedia- Social and Behavioral Science,* 164: 49–54.

Andrade, G., Ramos, G., Madeira, D., Sachetto, R., Ferreira, R. & Rocha, L. (2013). G-DBSCAN: A GPU Accelerated Algorithm for Density-based Clustering. *Procedia Computer Science,* 18: 369–378.

Anil, K.J. & Richard, C.D. (1988). *Algorithms for clustering data*. Prentice-Hall.

Bernardino da Silva, B., Leitão Paes, N. & Ospina, R. (2015). The replacement of payroll tax by a tax on revenues: A study of sartorial impacts on the Brazilian economy. *Economia.* 16: 46–59.

Lawson, D.J. & Falush, D. (2012). *Similarity matrices and clustering algorithms for population identification using genetic data*. March 1, in edited.

Falahpoor, S., Gol Arzi, Q. & Fatore Chiyan, N. (2014). Predicting Stock Price Movement Using Support Vector Machine Based on Genetic Algorithm in Tehran Stock Exchange Market. *Journal of Financial Research*, 15(2): 269-288. (*in Persian*)

Ghosh, S. & Kumar Dubey, S. (2013). Comparative Analysis of K-Means and Fuzzy CMeans Algorithms. *(IJACSA) International Journal of Advanced Computer Science and Applications,* 4(4): 35-39.

Hasani, M., Shaban, M., Mokhtari Masinaee, M. & Moodi, M. (2012). Discussion effective factor on tax capacity and prediction Khorasan Jonobi tax revenues with using ARMA model. *Tax administration core research in South Khorasan state.* (*in Persian*)

http://www.mathworks.com/help/stats/classificationtree-class.html. (Seen in July 2015).

Karami, A. & Johansson, R. (2014). Choosing DBSCAN Parameters Automatically using Differential Evolution. *International Journal of Computer Applications,* 91(7): 1-11.

Lewis, R., Mello, C. & White, A. (2012). Tracking Epileptogenesis Progressions with Layered Fuzzy K-means and K-medoids Clustering. *Procedia Computer Science,* 9: 432–438.

Mohd Isa, K., Yussof, S. & Mohdali, R. (2014). The role of tax agents in sustaining the Malaysian tax system. *Procedia- Social and Behavioral Sciences,* 164: 366–371.

Nurpratami, I. & Sitanggang, I. (2015). Classification rules for hotspot occurrence using spatial entropy based Decision tree algorithm. *Procedia Environmental Sciences,* 24: 120-126.

Popa, M. (2014). Taxes, Fees and Obligations in Romania -Main Components of Companies' Fiscal Costs. *Procedia- Social and Behavioral Sciences,* 109:150-154.

Radfar, R., Nezafati, N. and YousefiAsl, Y. (2014). Classification of bank customer based on data mining algorithms. Journal of *IT management,* 1: 71-90. (*in Persian*)

Raee, R., Falahpoor, S. & Ameri matin, H. (2013). Financial Risk Assessment Model for LNG Projects, Case Study: Iran LNG Project. *Journal of Financial Research*, 14(2): 47-64. (*in Persian*)

Rokach, R. & Maimon, O. (2008). *Data Mining with Decision Trees: Theory and Applications (Series in Machine Perception and Artificial Intelligence*. 69, (USA) World Scientific Publishing Co.

Wentian, J., Zhong Sheng, G. & En, Z. (2013). Improved K-medoids Clustering Algorithm under Semantic Web. *Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013).*

Wu, R. Sh., Ou, C.S., Chang Sh. & Yen, D.C. (2012). *Using Data Mining Technique to Enhance Tax Evasion Detection Performance*. Expert Systems with Applications, 39: 8769-8777.

Clusterevaluation. Available in: http://www. uniweimar. De / medien / webis / teaching / lecturenotes /machine - learning/ unit- en- cluster- analysis-evaluation. pdf. Seen at July 2015.