بهبود کارایی الگوریتم‌های تشخیص تقلب مالیاتی با استفاده از الگوهای پردازش موازی

مهدی سامعی‌راد[1]

اسدالله شاه بهرامی[2]

تاریخ دریافت: ۱۳۹۴/۸/۸ تاریخ پذیرش: ۱۳۹۴/۱۱/۱۴

چکیده

تقلب مالیاتی شامل طیف وسیعی از شیوه‌های کتمان حقایق، اظهار اطلاعات نادرست و انجام معاملات مالی خارج از چهارچوب‌های قانونی است. امروزه با گسترش سیستم‌های مالیاتی و حجم بالای داده‌های ذخیره شده در آن، نیاز به ابزاری است تا بتوان داده‌های ذخیره شده را پالایش و پردازش کرده و اطلاعات و دانش مورد نظر را استخراج نمود. با توجه به سیاست‌های مالیاتی به‌ویژه در مالیات بر ارزش افزوده، نرخ تقلب مالیاتی رو به رشد است. اخیراً پژوهشگران از روش‌های مختلفی از قبیل قوانین همبستگی، خوشه‌بندی، شبکه‌های عصبی، درخت‌های تصمیم، شبکه‌های بیزین، رگرسیون و ژنتیک در جهت کشف تقلب مالیاتی استفاده کرده‌اند. ولی به دلیل حجم بالای داده‌های مالیاتی، اکثر الگوریتم‌ها در تشخیص تقلب، دارای زمان اجرای زیادی هستند. در ابتدا از الگوریتم Apriori که از قوانین همبستگی و مدل‌های یادگیری بدون ناظر است، جهت کشف رفتارهای مشکوک متقلبان مالیاتی استفاده می‌شود و همچنین در مرحله بعد، یک سیستم تشخیص تقلب مالیاتی مبتنی بر شبکه‌های بیزین ارائه می‌شود و با توجه به کارایی پایین آن از نظر سرعت، کارایی آن با استفاده از تکنیک‌های پردازش موازی افزایش داده می‌شود. نتایج پیاده‌سازی بر روی پایگاه داده‌های مختلف مالیاتی نشان داد که با استفاده از الگوهای پردازش موازی، می‌توان کارایی برنامه‌های کشف تقلب‌های مالیاتی را به‌طور قابل ملاحظه‌ای بهبود بخشید.

واژه‌های کلیدی: داده‌کاوی، تشخیص تقلب مالیاتی، شبکه‌های بیزین، موازی‌سازی

---

۱. دانشجوی دکتری تخصصی سیستم‌های نرم‌افزاری، دانشگاه آزاد اسلامی واحد رشت (نویسنده مسئول) SameeRad@iauRasht.ac.ir

۲. عضو هیات علمی گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه گیلان Shahbahrami@guilan.ac.ir

۱- مقدمه

مالیات محوری‌ترین و اساسی‌ترین منبع درآمدی دولت‌ها و کشورهای در حال توسعه محسوب می‌شود (چی هانگ لین و همکاران، ۲۰۱۲: ۱). به دلیل استفاده کارشناسان مالیاتی از استراتژی‌های حسابرسی سنتی، مقدار قابل توجهی از درآمد مالیاتی دولت‌ها از دست می‌رود (گونزالس و ولاسکوئز، ۲۰۱۳: ۱) . با تغییرات جدید در قانون مالیاتهای مستقیم مورخ ۱۳۹۴/۴/۳۱ و یکسان شدن دیدگاه سازمان امور مالیاتی نسبت به طبقه‌بندی مشاغل بند الف، ب و ج همچنین در جهت استقرار دستورالعمل ماده ۱۶۹ مکرر ق.م.م. و در راستای تحقق وصولی در بخش مالیاتهای غیرمستقیم مانند ارزش افزوده و الزام فعالان اقتصادی از بابت تسلیم اظهارنامه ارزش افزوده و ارائه منظم صورت معاملات به‌صورت فصلی و همچنین سایر تکالیف مالیاتی، طبق مطالعات صورت پذیرفته، آمار تقلب مؤدیان مالیاتی در ایران رو به افزایش است (سایت روزنامه رسمی، ۱۳۹۴ به آدرس www.rrk.ir).

از طرفی به دلیل حجم بالای اطلاعات، تنوع داده‌ها و افزایش سرعت تولید داده‌ها در سیستم‌های مالیاتی، شاید بتوان گفت که دیگر سیستم‌های سنتی پاسخگوی مدیریت این حجم بزرگ اطلاعات نیست و نیاز به سیستم‌های هوشمند در جهت مدیریت این حجم از داده‌های عظیم است که در این سیستم‌های هوشمند می‌توان از انواع الگوریتم‌های هوش مصنوعی و فرآیند داده–کاوی در جهت کشف الگوهای مورد نیاز و پنهان استفاده کرد.

در جهت پردازش این حجم از داده‌ها، مطمئناً به سیستم‌های پردازش موازی نیاز است که بتواند روابط موجود بین داده‌ها را کشف و استخراج نماید. امروزه سیستم‌های سخت‌افزاری موجود دارای قابلیت‌های پردازش موازی از قبیل پردازش‌های چند نخی بر روی سیستم‌های چند هسته‌ای هستند و دستورات مورد نیاز در جهت استفاده از آنها در زبان‌های سطح بالا فراهم شده است (استرافسکی، ۲۰۱۰: ۱۰و۲۷). همچنین یقیناً با در دست داشتن یک سیستم هوشمند، می‌توان به برداشت اطلاعات از بانک‌های مختلف مالیاتی با استفاده از روش‌های هوشمند داده‌کاوی اقدام نموده و با تشخیص متقلبان مالیاتی و اطلاع‌رسانی به ادارات مالیاتی از این سیستم هوشمند برای تشخیص مالیات به‌صورت صحیح و دقیق جهت افزایش وصول مالیات استفاده نمود (شیان وو و همکاران، ۲۰۱۲: ۲).

الگوریتم‌های هوشمند مختلفی در جهت کشف تقلب مالیاتی مورد استفاده قرار می‌گیرد، ولی با توجه به حجم و تنوع داده‌های مالیاتی، سرعت این الگوریتم‌ها پایین و بر خط نیست. هدف این مقاله، افزایش کارایی الگوریتم‌های داده‌کاوی مانند الگوریتم Apriori و یا شبکه‌های بیزین در جهت کشف تقلب در

اظهارنامه‌های مالیات‌دهندگان، اطلاعات پرونده ایشان و تراکنش‌های مالی آنان در طول آخرین سال فعالیت اقتصادی با استفاده از تکنیک‌های پردازش موازی است.

برای پیاده‌سازی موازی، از تکنولوژی‌های خاصی[1] استفاده شده است (واگاتا، ۲۰۱۰: ۲؛ ورنکار، ۲۰۱۰: ۱۱؛ استرافسکی، ۲۰۱۰: ۵؛ اوکور، ۲۰۱۴: ۸). نتایج پیاده‌سازی الگوریتم بیزین با الگوهای پردازش موازی نشان داد که می‌توان سرعت پردازش آن‌را نسبت به حالت سریال بهبود بخشید.

هدف اصلی از این پیاده‌سازی، ساخت مدلی برای مالیات‌دهندگان، تشخیص متقلبان مالیاتی، پیش‌بینی تقلب مالیات‌دهندگانی که احتمال تقلب آنان در سال‌های آتی وجود دارد و افزایش میزان مالیات وصولی در ادارات دارایی می‌باشد.

ساختار مقاله به این صورت است: برخی از کارهای مرتبط در بخش دو مورد بحث قرار گرفته و کشف رفتارهای مشکوک مالیاتی با استفاده از قوانین همبستگی در بخش سه و راه حل پیشنهادی و نحوهٔ استفاده از الگوهای پردازش موازی در بخش چهار توضیح داده شده و نتایج پیاده‌سازی در بخش پنجم بیان می‌گردد و نهایتاً نتیجه‌گیری در بخش ششم مطرح می‌گردد.

۲- پیشینه تحقیق

در زمینه کشف تقلب در داده، کارهای مختلفی با استفاده از الگوریتم‌های مختلف داده‌کاوی انجام شده است. برای مثال از شبکه‌های عصبی چند لایه، ماشین بردار پشتیبان (SVM)، برنامه نویسی ژنتیک (GP)، گروه‌بندی در پردازش داده‌ها (GMDH)، رگرسیون لجستیک (LR) و شبکه‌های عصبی احتمالاتی (PNN) اقدام به شناسایی شرکت‌هایی نموده‌اند که در جهت حرکت به سمت تقلب مالیاتی، کتمان حقایق تراکنش‌های مالی و فرار مالیاتی بوده‌اند. ولی با توجه به حجم بالای اطلاعات مالیاتی موجود، پردازش آنها زمان‌گیر بوده و دلیل استفاده از موازی‌سازی در این مطالعه همین مطلب می‌باشد.

1. Net Parallel Loops and Parallel LINQ

جدول (۱)- برخی از تحقیقات انجام شده در زمینه تقلب مالیاتی

نویسندگان	مفروضات	معلومات	رویکرد	محیط	ابزار (تکنیک)
گونزالس و همکاران	فعالیت اشخاص مبتنی بر فاکتور است	تعداد و جمع مبلغ فاکتورها	خوشه‌بندی و شبکه عصبی	کانادا، شیلی و آمریکا	کلمنتاین و SPSS (غیرموازی)
وو و همکاران	همه اظهارنامه ارزش افزوده تسلیم کنند	اطلاعات اظهار شده در اظهارنامه‌ها	داده‌کاوی و قوانین همبستگی	ارزش افزوده آلمان	DM Miner (غیرموازی)
راویس آنکار	بررسی سوابق مؤدی برای تشخیص تقلب	فیلدهای موجود در اظهارنامه	ژنتیک و رگرسیون	داده‌های آزمایشی	t-statistic (غیرموازی)
نگای	تقلب مالیاتی محل نگرانی مصرف‌کننده	مقالات منتشر شده در این زمینه	رگرسیون و شبکه عصبی	اطلاعات ۴۹ مقاله مرتبط	کلمنتاین SPSS (غیرموازی)

منبع: یافته‌های تحقیق

۳- کشف رفتارهای مشکوک مالیاتی با استفاده از قوانین همبستگی[1]

مجموعه‌ای مشتمل بر n شیء را در نظر می‌گیریم: $O=\{o_1,o_2,o_3,...,o_n\}$ ، زیرمجموعه $L \subseteq 0$ را یک مجموعه اقلام[2] می‌گوییم. هر مجموعه اقلام که شامل K شیء باشد یک مجموعه اقلام K تایی نامیده می‌شود. اطلاعات رکورد هر مؤدی مبین یک مجموعه اقلام است که در ارتباط با انجام یا عدم انجام تکالیف مالیاتی مالیات‌دهنده می‌باشد. لذا بانک D شامل m تراکنش (رکورد) T_i است که هرکدام توسط یک شناسه یگانه t_i مشخص می‌شود. این مطلب در جدول (۲) مشخص گردیده است همچنین لیست کامل و ماتریسی در جدول (۳) می‌باشد.

1. Association Rules
2. Item Set

جدول (۲)- نمونه جامعه آماری تراکنشی تکالیف مؤدیان

شماره مؤدی	تراکنش t_i
۰۰۱	{تسلیم اظهارنامه، برگ تشخیص، مؤدی کوچک، عدم تسلیم دفاتر، تمکین، برگ قطعی و واریز}
۰۰۲	{عدم تسلیم اظهارنامه، عدم تسلیم دفاتر، برگ تشخیص، مؤدی بزرگ، اعتراض و توافق ماده ۲۳۸، برگ قطعی و واریز}
۰۰۳	{اعلامیه متمم، تسلیم اظهارنامه، برگ تشخیص، مؤدی کوچک، عدم تسلیم دفاتر، اعتراض، برگ قطعی و عدم واریز}
۰۰۴	{اعلامیه متمم، عدم تسلیم اظهارنامه، صدور برگ تشخیص، مؤدی کوچک، تسلیم دفاتر، تمکین، برگ قطعی و واریز}
۰۰۵	{تسلیم اظهارنامه، صدور برگ تشخیص، مؤدی کوچک، عدم تسلیم دفاتر، اعتراض و کمیسیون، برگ قطعی و عدم واریز}

منبع: پایگاه داده مشاغل آزمایشی و آموزشی نگارنده، ۲۰۱۵

جدول (۳)- ماتریس دو بعدی برای جدول (۲) و برای نمونه جامعه آماری ۴۰ نفر مؤدی مالیاتی

مؤدی	اعلامیه متمم	تسلیم اظهارنامه	تسلیم دفاتر	عدم تسلیم دفاتر یا اظهارنامه	صدور برگ تشخیص اولیه	صدور برگ متمم	مؤدی کوچک	مؤدی متوسط	مؤدی بزرگ	تمکین مؤدی	اعتراض و توافق ماده ۲۳۸	اعتراض و عدم توافق	برگ قطعی و واریز	برگ قطعی و عدم واریز
۱	خیر	بلی	خیر	بلی	بلی	خیر	بلی	خیر	خیر	بلی	خیر	خیر	بلی	خیر
۲	خیر	خیر	خیر	بلی	بلی	خیر	خیر	خیر	بلی	خیر	بلی	خیر	بلی	خیر
۳	بلی	بلی	خیر	بلی	بلی	بلی	بلی	خیر	خیر	خیر	خیر	بلی	خیر	بلی
۴	خیر	بلی	خیر	بلی	بلی	خیر	بلی	خیر	خیر	بلی	خیر	خیر	بلی	خیر
۵	بلی	خیر	خیر	بلی	خیر	بلی	خیر	بلی	خیر	خیر	بلی	خیر	بلی	خیر
۶	خیر	خیر	خیر	بلی	بلی	خیر	بلی	خیر	خیر	خیر	بلی	خیر	بلی	خیر
۷	خیر	بلی	بلی	خیر	بلی	خیر	بلی	خیر	خیر	بلی	خیر	خیر	بلی	خیر

مؤدی	اعلامیه متمم	تسلیم اظهارنامه	تسلیم دفاتر	عدم تسلیم دفاتر یا اظهارنامه	صدور برگ تشخیص اولیه	صدور برگ متمم	مؤدی کوچک	مؤدی متوسط	مؤدی بزرگ	تمکین مؤدی	اعتراض و توافق ماده ۲۳۸	اعتراض و عدم توافق	برگ قطعی و وارینز	برگ قطعی و عدم وارینز
۸	بلی	خیر	خیر	بلی	خیر	بلی	بلی	خیر	خیر	خیر	بلی	خیر	بلی	خیر
۹	خیر	بلی	خیر	بلی	بلی	خیر	خیر	خیر	بلی	خیر	بلی	خیر	بلی	خیر
۱۰	خیر	بلی	بلی	خیر	بلی	خیر	بلی	خیر	خیر	بلی	خیر	خیر	بلی	خیر
۱۱	خیر	خیر	خیر	بلی	بلی	خیر	خیر	بلی	خیر	خیر	بلی	خیر	بلی	خیر
۱۲	بلی	خیر	خیر	بلی	خیر	بلی	بلی	خیر	خیر	خیر	خیر	بلی	خیر	بلی
۱۳	خیر	بلی	خیر	بلی	بلی	خیر	بلی	خیر	خیر	خیر	بلی	خیر	بلی	خیر
۱۴	بلی	خیر	خیر	بلی	بلی	بلی	بلی	خیر	خیر	خیر	بلی	خیر	بلی	خیر
۱۵	خیر	بلی	خیر	بلی	بلی	خیر	بلی	خیر	خیر	خیر	بلی	خیر	بلی	خیر
۱۶	خیر	بلی	خیر	بلی	بلی	خیر	بلی	خیر	خیر	خیر	بلی	خیر	بلی	خیر
۱۷	خیر	خیر	خیر	بلی	بلی	خیر	بلی	خیر	خیر	خیر	بلی	خیر	بلی	خیر
۱۸	بلی	بلی	خیر	بلی	بلی	بلی	خیر	بلی	خیر	خیر	خیر	بلی	خیر	بلی
۱۹	خیر	خیر	خیر	بلی	بلی	خیر	بلی	خیر	خیر	خیر	بلی	خیر	بلی	خیر
۲۰	بلی	بلی	خیر	بلی	خیر	بلی	بلی	خیر	خیر	خیر	بلی	خیر	بلی	خیر
۲۱	خیر	خیر	خیر	بلی	بلی	خیر	خیر	بلی	خیر	خیر	بلی	خیر	بلی	خیر
۲۲	خیر	بلی	خیر	بلی	بلی	خیر	بلی	خیر	خیر	خیر	بلی	خیر	بلی	خیر
۲۳	خیر	بلی	خیر	بلی	بلی	خیر	بلی	خیر	خیر	خیر	بلی	خیر	بلی	خیر
۲۴	خیر	خیر	خیر	بلی	بلی	خیر	بلی	خیر	خیر	خیر	بلی	خیر	بلی	خیر
۲۵	خیر	بلی	بلی	خیر	بلی	خیر	خیر	بلی	خیر	بلی	خیر	خیر	بلی	خیر
۲۶	خیر	بلی	خیر	بلی	بلی	خیر	بلی	خیر	خیر	خیر	بلی	خیر	بلی	خیر
۲۷	خیر	بلی	خیر	بلی	بلی	خیر	بلی	خیر	خیر	خیر	بلی	خیر	بلی	خیر
۲۸	خیر	خیر	خیر	بلی	بلی	خیر	خیر	بلی	خیر	خیر	بلی	خیر	بلی	خیر
۲۹	خیر	بلی	خیر	بلی	بلی	خیر	خیر	بلی	خیر	خیر	بلی	خیر	بلی	خیر
۳۰	بلی	خیر	خیر	بلی	بلی	بلی	بلی	خیر	خیر	خیر	خیر	بلی	خیر	بلی
۳۱	خیر	بلی	بلی	خیر	بلی	خیر	خیر	خیر	بلی	بلی	خیر	خیر	بلی	خیر

مؤدی	اعلامیه متمم	تسلیم اظهارنامه	تسلیم دفاتر	عدم تسلیم دفاتر یا اظهارنامه	صدور برگ تشخیص اولیه	صدور برگ متمم	مؤدی کوچک	مؤدی متوسط	مؤدی بزرگ	تمکین مؤدی	اعتراض و توافق ماده ۲۳۸	اعتراض و عدم توافق	برگ قطعی و واریز	برگ قطعی و عدم واریز
۳۲	خیر	بلی	خیر	بلی	بلی	خیر	بلی	خیر	خیر	خیر	بلی	خیر	بلی	خیر
۳۳	بلی	خیر	خیر	بلی	بلی	بلی	بلی	خیر	خیر	خیر	بلی	خیر	بلی	خیر
۳۴	بلی	بلی	خیر	بلی	خیر	بلی	بلی	خیر	خیر	خیر	بلی	خیر	بلی	خیر
۳۵	خیر	بلی	بلی	بلی	بلی	خیر	بلی	خیر	خیر	بلی	خیر	خیر	بلی	خیر
۳۶	خیر	خیر	خیر	بلی	بلی	خیر	بلی	خیر	خیر	خیر	بلی	خیر	بلی	خیر
۳۷	خیر	بلی	خیر	بلی	بلی	خیر	بلی	خیر	خیر	خیر	بلی	خیر	بلی	خیر
۳۸	خیر	بلی	خیر	بلی	بلی	خیر	خیر	بلی	خیر	خیر	خیر	بلی	بلی	خیر
۳۹	خیر	بلی	بلی	بلی	بلی	خیر	بلی	خیر	خیر	خیر	بلی	خیر	بلی	خیر
۴۰	بلی	خیر	خیر	بلی	بلی	بلی	خیر	خیر	بلی	خیر	بلی	خیر	بلی	خیر

منبع: پایگاه داده مشاغل آزمایشی و آموزشی نگارنده، ۲۰۱۵ و جدول اطلاعات مالیاتی

تعداد تراکنش‌های T_i موجود در پایگاه داده D که دارای مجموعه L هستند را فراوانی تجمعی مجموعه L نامیده و آنرا با f(L) نشان می‌دهیم. همچنین نسبت (f(L)/m) فراوانی انباشته تعداد کل تراکنش‌هاست که بیانگر احتمال رخداد مجموعه اقلام L یعنی Pr(L) می‌باشد.

هدف: یافتن الگوهای منظم مستتر در داده‌های مالیاتی است تا بتوان همبستگی بین یک یا چند مؤدی از بابت شباهت در تقلب و مدل رفتاری را نتیجه گرفت. یک روش برای یافتن قوانین قوی، آن است که ابتدا همه قوانین را تولید کرده و مقدار پشتیبان و اطمینان هرکدام را محاسبه نموده و نهایتاً قوانین ضعیف را کنار گذاشت. ولی این روش به عملیات محاسباتی زیادی نیاز دارد لذا از الگوریتم Apriori استفاده کرده و در دو فاز، فقط قوانین قوی را تولید می‌کنیم. در فاز اول تولید مجموعه اقلام مکرر و در فاز دوم تولید قوانین قوی و در نهایت بلندی قانون[1] را به دست می‌آوریم (ورسلیز، ۲۰۰۹: ۲۹۳).

به‌عنوان مثال می‌خواهیم رابطه بین «صدور برگ متمم» و «عدم واریز مالیات بعد از ابلاغ برگ قطعی» را محاسبه کنیم که طبق جدول شماره ۳ تعداد کل مؤدیان جامعه نمونه آماری ۴۰ نفر می‌باشد که تعداد

1. Lift of a Rule

صدور برگ متمم ۱۱ فقره و تعداد عدم تمکین به برگ قطعی ۴ فقره و همچنین تعداد هر دو مورد همزمان ۴ فقره است. با فرض مقادیر آستانه S_{min}=۰,۶ و P_{min}=۰,۴ قانون {عدم واریز} ← {صدور متمم} به‌عنوان قانون ضعیف انتخاب می‌شود چرا که پشتیبان آن $S=\frac{۴}{۴۰}$=۰,۱ و اطمینان آن $P=\frac{۴}{۱۱}$=۰,۳۶ است که حداقل‌ها را رعایت نکرده است؛ یعنی $S<S_{min}$ و $P<P_{min}$ است. این قانون به معنای آن است که صدور برگ متمم، عدم تمکین به برگ تشخیص و در پی آن عدم واریز مالیات را به همراه نداشته است.

در یک مجموعه داده بزرگ با مؤدیان زیاد، تعداد زیادی تکالیف مالیاتی و قانون وجود دارد، اما بسیاری از اینها به دلیل این که مقادیر لازم برای پشتیبان و اطمینان را ندارند قوی نیستند. حال با استفاده از الگوریتم Apriori طبق جدول شماره (۴) اقدام به تولید قوانین قوی و بلند می‌نماییم.

جدول (۴)- تولید بلندی قانون $(Lift\{L \Rightarrow H\}=\frac{Conf\{L \Rightarrow H\}}{f(H)})=\frac{f(L \cup H)}{f(L)*f(H)}$

قانون	Lift	بلند
عدم تسلیم اظهارنامه یا دفاتر ⟵ {صدور برگ تشخیص اولیه و برگ قطعی و واریز و اعتراض و ۲۳۸}	$\frac{25}{25*36}$=0.028	بلند
برگ قطعی و واریز ⟵ {عدم تسلیم اظهارنامه یا دفاتر و صدور برگ تشخیص اولیه و اعتراض و ۲۳۸}	$\frac{25}{25*36}$=0.028	بلند

منبع: یافته‌های تحقیق

* قانون اول به معنای آن است که اثر همزمان {صدور برگ تشخیص اولیه و برگ قطعی و واریز و اعتراض و توافق ماده ۲۳۸} عدم تسلیم اظهارنامه یا دفاتر را به همراه داشته است.
* قانون دوم به این معنی است که {عدم تسلیم اظهارنامه یا دفاتر و صدور برگ تشخیص اولیه و اعتراض و توافق ماده ۲۳۸} موجب واریز مالیات توسط مؤدی بعد از صدور برگ قطعی شده است.

۴- افزایش کارایی با موازی‌سازی

یکی از مراحل داده‌کاوی استفاده از الگوریتم‌های داده‌کاوی در جهت یادگیری الگوهای موجود در داده‌های آموزشی است. به عبارتی داده‌های موجود یک سیستم عملیاتی در فرآیند دسته‌بندی به دو دسته آموزشی و تست تقسیم می‌شود. با استفاده از داده‌های آموزشی، یادگیری الگوها و آموزش صورت می‌گیرد و نحوهٔ عملکرد یادگیری سیستم نیز با مرحله تست انجام می‌شود و هر چقدر داده‌های آموزشی بیشتر و جامع‌تر باشد، الگوریتم مورد نظر عملکرد بهتری خواهد داشت. از میان تکنیک‌های موجود در روش دسته‌بندی،

تکنیک بیزین ساده[1] با استفاده از پردازش موازی مورد استفاده قرار گرفته است.

شبکه‌های بیزین

شبکه‌های بیزین جزو دسته‌بندی کننده‌های آماری هستند و می‌توانند ردیف تعلق داده‌ها را مشخص کنند و در آن صفات داده‌های ورودی از هم مستقل بوده و بر روی هم تاثیر ندارند. از این روش تحت عنوان تکنیک نظارت شده[2] یاد می‌شود.

شبکه بیزین، یک گراف مستقیم است که گره‌ها در آن نشان‌دهنده متغیرها هستند $(X_1,\ldots,X_n)$ که طبق فرمول (۱) از آنها برای محاسبه احتمال وقوع X_i به شرطی که والدین آن اتفاق افتاده باشند استفاده می‌کنیم. X_i احتمال وقوع هر گره و Parent(Xi) احتمال وقوع والدین آن گره می‌باشد و کل احتمال از حاصل‌ضرب احتمالات مثل رابطه ۱ بدست می‌آید.

$$P(x_1,\ldots,x_n)=\prod_{i=1}^{n}P(x_i\mid Parent(x_i)) \qquad (1)$$

به عنوان نمونه توزیع احتمال شرطی مربوط به تقلب، با بررسی میزان«مالیات بر درآمد» و همچنین میزان «خرید»، «ارزش منطقه‌ای محل درآمد» (یا همان سرقفلی) و میزان احتمال «فروش» بر روی یک جامعه نمونه آماری برای مشاغل ثبت شده در سازمان امور مالیاتی در شکل (۱) قابل مشاهده است. نتایج بررسی‌ها نشان داد مالیات‌دهندگان دارایِ برگ متمم، کلاً ۵۷٫۹٪ تقلب را شامل می‌شوند. برخی از تجزیه و تحلیل‌های شکل (۱) (کارکِرا، ۲۰۱۴: ۲۵) عبارتند از:

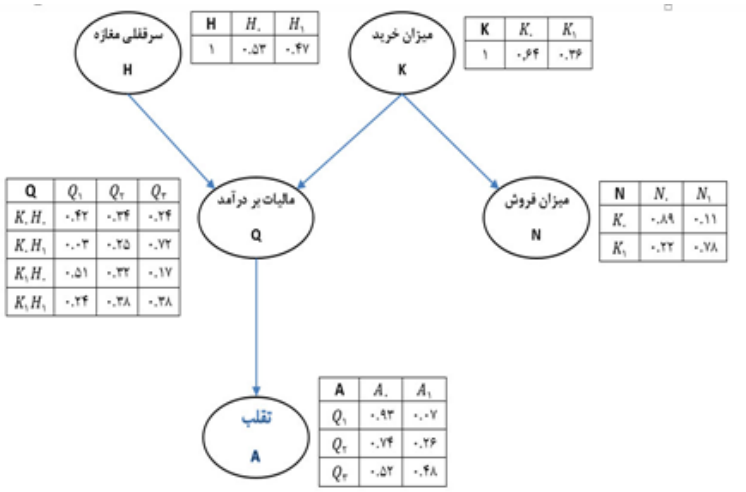
روش پیشگویی[3]: از علت به معلول رسیدن است، در این حالت شرایط موجود است.

روش استدلال شهودی[4]: نتیجه موجود است، علت و دلایل رخ دادن این نتیجه بررسی و جست‌وجو می‌شود.

روش استدلال تعاملی[5]: در این روش به‌صورت افقی، تاثیر پارامترهای مختلف بر روی هم بررسی می‌شود.

1. Naïve Bayesian
2. Supervised Method
3. Casual Reasoning or Prediction
4. Evidential Reasoning
5. Intercausal Reasoning

شکل (۱)- درخت بیزین جهت تشخیص تقلب مالیاتی



H	H_0	H_1
۱	۰.۵۳	۰.۴۷

K	K_0	K_1
۱	۰.۶۴	۰.۳۶

Q	Q_1	Q_2	Q_3
K_0H_0	۰.۴۲	۰.۳۴	۰.۲۴
K_0H_1	۰.۰۳	۰.۲۵	۰.۷۲
K_1H_0	۰.۵۱	۰.۳۲	۰.۱۷
K_1H_1	۰.۲۴	۰.۳۸	۰.۳۸

N	N_0	N_1
K_0	۰.۸۹	۰.۱۱
K_1	۰.۲۲	۰.۷۸

A	A_0	A_1
Q_1	۰.۹۳	۰.۰۷
Q_2	۰.۷۴	۰.۲۶
Q_3	۰.۵۲	۰.۴۸

$$P(A_1) = \sum P(K_{0-1}, H_{0-1}, N_{0-1}, Q_{1-3}, A_1) \quad (۲)$$

$$= \sum P(K_{0-1}) * P(H_{0-1}) * P(N_{0-1}|K_{0-1}) * P(Q_{1-3}|H_{0-1}, K_{0-1}) * P(A_1|Q_{1-3})$$

(۲ حالت) (۲ حالت) (۲ حالت) (۳ حالت) (۱ حالت)

$$P(x_1, \dots, x_n|C_i) = \prod_{i=1}^{n} P(x_{1\dots n}|C_i) = P(x_1|C_i) * P(x_2|C_i) * \dots * P(x_n|C_i) = ۰/۵۷۹$$

بیزین، مجموعه صفات ورودی را به‌صورت یک بردار نشان می‌دهد. فرمول (۲) محاسبه احتمال تقلب مالیاتی در بین مالیات‌دهندگان است. یعنی هر صفت، مؤلفه‌ای از این بردار می‌باشد.

$$P(C_i|x) = \frac{P(x|C_i).P(C_i)}{P(x)} \quad (۳)$$

محاسبه احتمال به روش بیزین ساده در فرمول (۳) قابل ملاحظه است که در آن بردار X مجموعه صفات ورودی مثل سن فرد، شغل، میزان درآمد و ... می‌باشد و M تا کلاس خروجی داریم. در حال حاضر در تقلب مالیاتی ۲ تا کلاس (تقلب و عدم تقلب) داریم و نیز احتمال تک‌تک کلاس‌هایی را که روی داده باشد را محاسبه می‌نماییم.

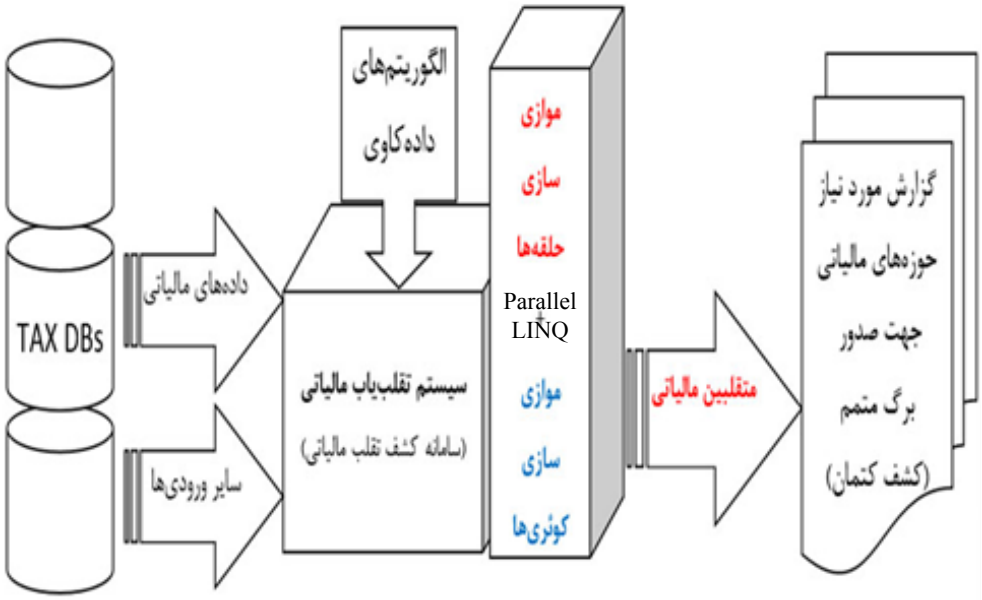
همچنین چون در فرمول بالا P(x) ثابت است آن را مقدار یک در نظر می‌گیریم. پس برای همه حالات، احتمال را محاسبه و بزرگترین احتمال، نشان‌دهنده تعلق بردارمان به آن کلاس است. برای مثال احتمال

تقلب و احتمال عدم تقلب را محاسبه می‌کنیم، هرکدام بزرگتر بود بردار X به آن کلاس C_i تعلق دارد. مثلاً اگر i=۱ را تقلب در نظر بگیریم، همکلاسی‌های متقلبان مشخص می‌شوند که برای این منظور برنامه تقلب‌یاب طبق شکل (۲) پیاده‌سازی می‌شود. با تحلیل خروجی برنامه تقلب‌یاب و بدون در نظر گرفتن ورودی داریم:

$$P\left(\text{تقلب} = Y\right) = \frac{4}{40} = 0.1 = 10\% \quad , \quad P\left(\text{تقلب} = N\right) = \frac{36}{40} = 0.9 = 90\%$$

مهمترین فیلدهای مؤثر در پارامترها، فیلدهای سری اول: «واصل شدن اعلامیه»، «تسلیم اظهارنامه»، «تسلیم دفاتر» و یا «عدم تسلیم یکی از آنها» هستند. با استفاده از برنامه طراحی شده با ساختار شکل (۲) و (۳) برای این منظور و پردازش موازی، نرخ تقلب را محاسبه کردیم.

شکل (۲)- ساختار سیستم تقلب‌یاب و موازی‌سازی با LinQ و Net.



منبع: سامعی‌راد و شاه‌بهرامی، ۲۰۱۵

۵- الگوریتم تشخیص تقلب مالیاتی بر روی بستر موازی

مراحل اجرای موازی الگوریتم تشخیص تقلب مالیاتی به‌شرح زیر است:

۱) واکشی جداول جهت تشخیص تقلب مشاغل (بند الف، ب و ج)

۲) انتخاب نمونه داده از جداول (۵٪ کل اطلاعات اظهارنامه مشاغل)

۳) ساخت مدل رفتاری متقلبین مالیاتی با استفاده از بیزین چندلایه

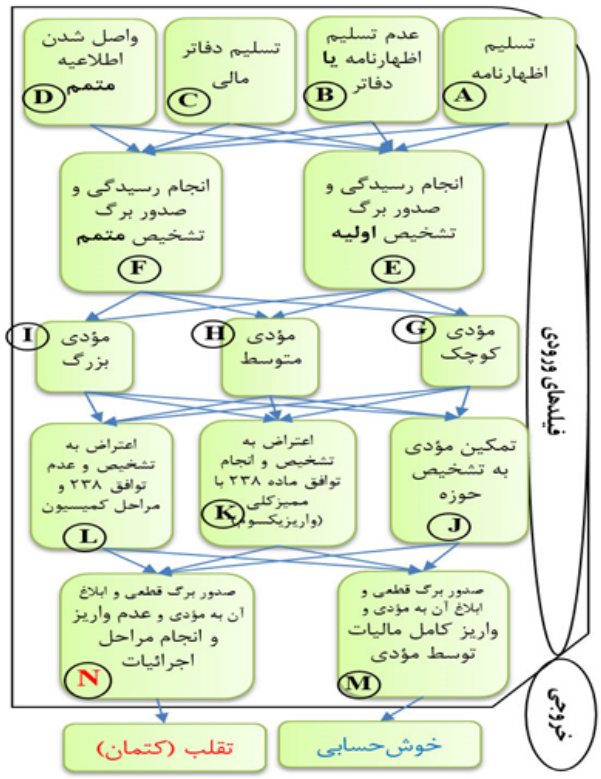
۴) تطبیق مدل رفتار متقلب با سایر اطلاعات به تفکیک حوزه مالیاتی

۵) ارزیابی نتایج تطبیق مدل رفتار متقلب با کل داده (نتایج مرحله ۴)

۶) برگشت به مرحله ۳ در صورت عدم رعایت حد آستانه خطا

۷) تشخیص تقلب، تهیه گزارش و ارائه آن به حوزه‌های مالیاتی

شکل (۳)- برچسب‌زنی لایه‌های بیزین موازی، جهت درج در فرمول



منبع: سامعی‌راد و شاه‌بهرامی، ۲۰۱۵

با توجه به شکل (۳) برای هر مشخصه‌ای از مؤدی که در ساخت مدل استفاده شده است، یک برچسب زده می‌شود که در فرمول (۴) از آن استفاده و با شروط زیر اقدام به محاسبه نرخ تقلب مالیاتی می‌گردد.

$$X = (\text{اعلامیه} = 1 \ , \ \text{اظهارنامه} = 0 \ , \ \text{دفاتر} = 0 \ , \ \text{عدم} = 1)$$

$$P(\text{تقلب} = 1 \mid X) = P(N = 1 \mid A = 0, B = 1, C = 0, D = 1)$$

$$= \sum \frac{P(A,B,C,D,E_{0-1},F_{0-1},G_{0-1},H_{0-1},I_{0-1},J_{0-1},K_{0-1},L_{0-1},N=1)}{P(A,B,C,D)} \qquad (۴)$$

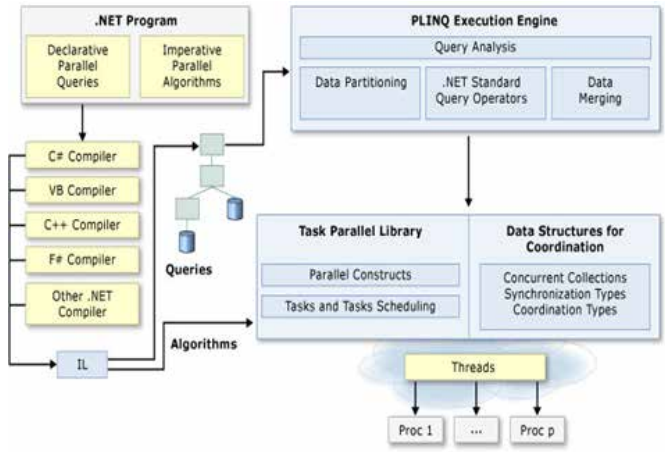
$$= P(A,B,C,D) \sum \frac{P(E_{0-1},F_{0-1},G_{0-1},H_{0-1},I_{0-1},J_{0-1},K_{0-1},L_{0-1},N=1)}{P(A,B,C,D)}$$

$$= \sum P(E_{0-1},F_{0-1},G_{0-1},H_{0-1},I_{0-1},J_{0-1},K_{0-1},L_{0-1},N=1)$$

۶- موازی‌سازی الگوریتم تشخیص تقلب مالیاتی

تفکر موازی و تحلیل و طراحی موازی می‌تواند راه‌گشای مشکلات چشم‌گیر نرم‌افزارهای امروزی باشد. در این مرحله و طبق بررسی انجام شده در سه سطح، امکان موازی‌سازی سیستم تقلب‌یاب وجود دارد که عبارتند از: موازی‌سازی در سطح حلقه‌های استفاده شده در برنامه، موازی‌سازی در سطح پرس‌وجوهای استفاده شده در برنامه، موازی‌سازی در سطح زیرساخت، شبکه و سیستم عامل. تکنولوژی مورد استفاده موازی‌سازی پلتفرم مایکروسافت است که یک شمای کلی از آن در شکل (۴) نشان داده شده است.

شکل (۴)- تکنولوژی مورد استفاده .Net برای موازی‌سازی‌ها



منبع: سایت مایکروسافت،۲۰۱۳

سطح اول: موازی‌سازی در سطح حلقه‌ها

حلقه‌های استفاده شده در بدنه برنامه در حالت سریال با توجه به حجم پردازشی، بسیار وقت‌گیر بودند که با موازی‌سازی، ساختار آن نیز دگرگون شده است. در اینجا شمایی از برنامه که دارای دو حلقه تو در تو است نشان داده شده است.

```
for (i=0, i< Code_Hoze.Capacity; i++)
 for (j=0, j<TaxPaery.length; j++)
 calculateMultiLevelBaysianFormules();
```

با تکنولوژی موازی‌سازی در دات‌نت، به‌صورت ذیل کد مربوط به حلقه‌های استفاده شده در برنامه، برای اخذ تسریع، موازی‌سازی شد:

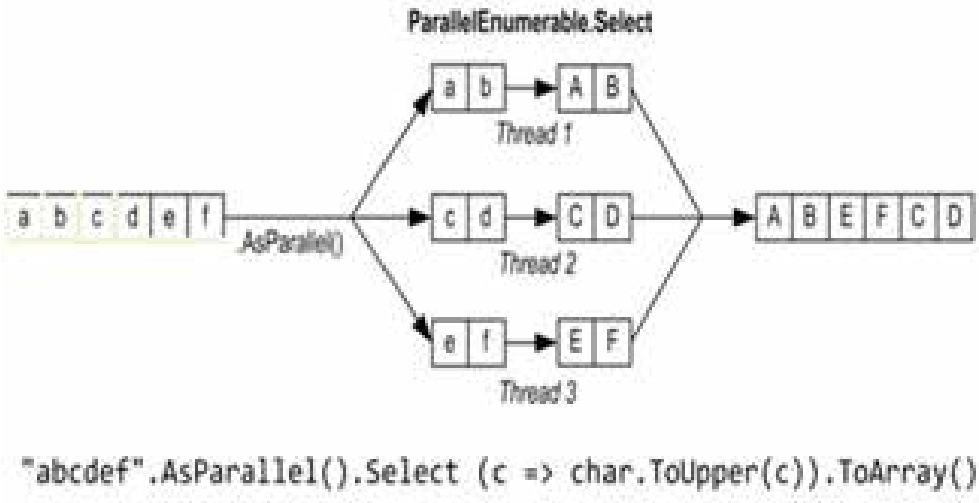
```
for (i=0, i< Code_Hoze.Capacity; i++)
 Parallel.For (0, TaxPaery.length, j =>
 { ret=calculateMultiLevelBaysianFormules
 (f1Arr[j], f2Arr[j], f3Arr[j], f4Arr[j]); }
```

سطح دوم: موازی‌سازی در سطح پرس‌وجو

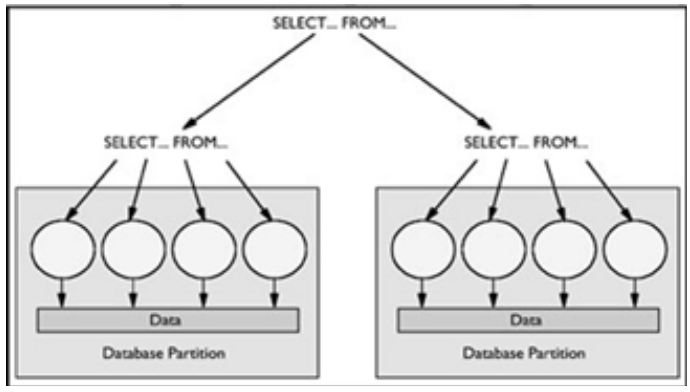
داده‌های مالیاتی جهت پردازش در برنامه، باید به‌طور مداوم از پایگاه داده‌ها واکشی شوند، که کد قسمتی از این عملیات در اینجا ذکر شده است. با توجه به وقت‌گیر بودن این فرآیند، از الگوی PLINQ به صورت شکل ۶ برای موازی‌سازی استفاده شده است.

```
for (i=0, i< dgvTaxPayer.RowCount; i++)
 for (j=0, j< ghatee_inf_RowsCount; j++)
 KernelComDo.Excute (QueryString);
```

```
for (int i = 0; i < dgvTaxPayer.RowCount; i++)
 { Int32 result1=ghatee_inf_Rows.Where (p => p.Field <string>
 («K_Parvand») == dgvMoadi.Rows[i].Cells[2].Value).Count(); }
```



شکل (۵)- ساختار PLINQ برای موازی‌سازی پرس‌وجوهای برنامه



منبع: سایت مایکروسافت، ۲۰۱۳

سطح سوم: موازی‌سازی در سطح زیرساخت

بعد از تحلیل، طراحی و پیاده‌سازی سیستم تقلب‌یاب مالیاتی، برنامه مزبور بر روی سرور قدرتمند Intel Xeon با ۱۶ پردازنده ۲ هسته‌ای X۵۵۷۰ و ۶۴GB حافظه رم اجراء شده است.

۷- نتایج پیاده‌سازی

جدول (۵) مجموع اطلاعات، داده‌ها و ابزارهایی که در این پیاده‌سازی مورد استفاده قرار گرفته است را نشان می‌دهد. همچنین مفروضات، معلومات و رویکرد و محیط را نشان می‌دهد.

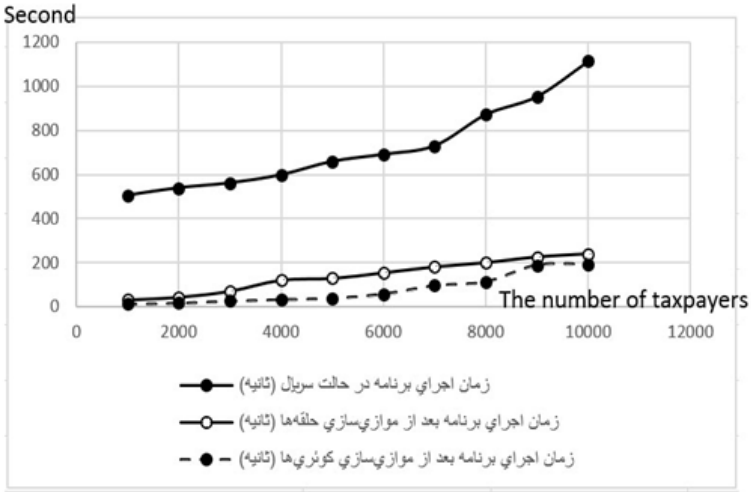
جدول (۵)- اطلاعات و ابزارهای استفاده شده

مفروضات	معلومات	رویکرد	محیط	ابزار (تکنیک)
حداقل ۳ بانک از ۱۸ بانک مالیاتی در دسترس است و باید انتخاب ویژگی انجام شده باشد	فیلدهای اظهارنامه تسلیمی و اطلاعات رسیدگی پرونده‌ها	از Baysian در دسته‌بندی و از Apriori در همبستگی و از K-Means در خوشه‌بندی	اطلاعات مالیات دهندگان سازمان امور مالیاتی	Sql Server, Rapid mainer, Oracle, Mat-lab (موازی)

منبع: سامعی‌راد و شاه‌بهرامی، ۲۰۱۵

قسمتی از حلقه‌های برنامه در بخش استخراج جداول بیزین و همچنین بخشی از پرس‌وجوهای برنامه در بخش تشخیص تقلب موازی‌سازی شد و برنامه نهایی بر روی بستر پردازش سریع به اجراء درآمد. شکل (۷) مقایسه موازی‌سازی حلقه‌ها و پرس‌وجوها را نسبت به حالت سریال نشان می‌دهد. همان‌طوری‌که در این شکل‌ها قابل مشاهده است با افزایش حجم داده‌ها با توجه به افزایش تعداد مالیات‌دهندگان، کارایی الگوریتم‌ها افزایش می‌یابد؛ یعنی مدت زمان اجرای برنامه کمتر می‌شود.

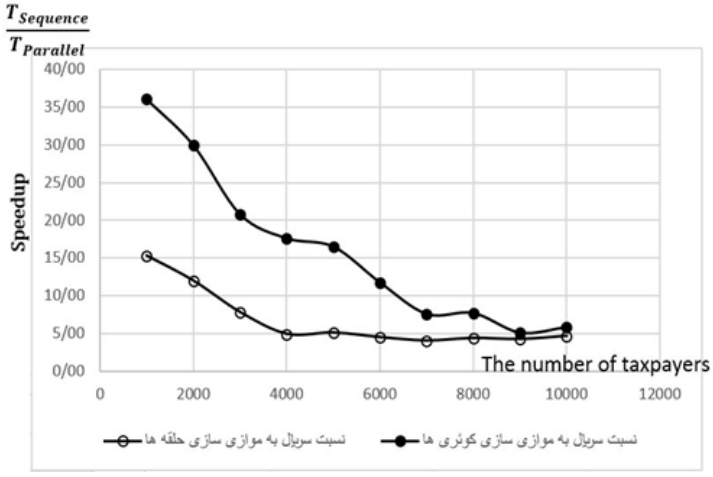
شکل (۶)- سرعت اجرای برنامه با تعداد مؤدیان مختلف در سه روش مزبور



منبع: سامعی‌راد، ۲۰۱۵

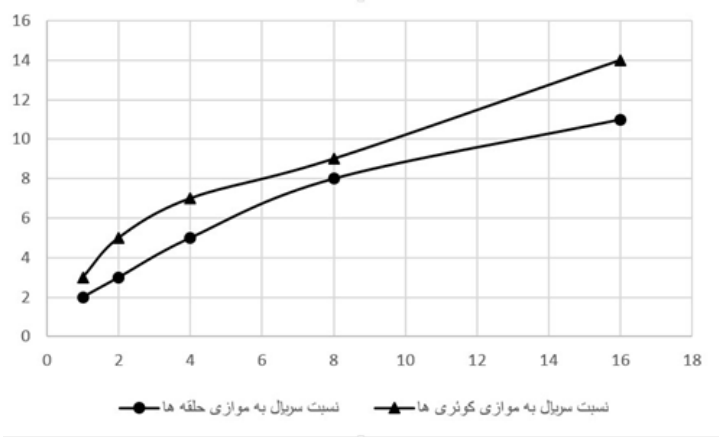
در نهایت مقایسه کارایی دو روش موازی‌سازی حلقه‌ها و موازی‌سازی پرس‌وجوها در شکل (۸) و کارایی به دست آمده بین این دو روش در شکل (۹) نشان داده شده است، که علت بالا بودن کارایی تکنیک موازی‌سازی پرس‌وجوها (PlinQ) به روش موازی‌سازی حلقه به این دلیل است که برنامه تشخیص و پیش‌بینی تقلب مالیاتی، اساساً داده–محور[1] می‌باشد.

شکل(۷)- مقایسه کارایی روش‌های موازی‌سازی پرس‌وجو و حلقه



منبع: سامعی‌راد، ۲۰۱۵

شکل (۸)- مقایسه کارایی روش‌های موازی‌سازی پرس‌وجو و حلقه روی چندپردازنده‌ها



منبع: سامعی‌راد، ۲۰۱۵

1. Data-driven

۸- نتایج تجربی

در این مطالعه، موازی‌سازی حلقه‌ها با استفاده از Parallel.For و موازی‌سازی پرس‌وجوها با استفاده از Parallel LINQ در دات‌نت و موازی‌سازی بستر اجرایی با استفاده از پردازش موازی و سیستم‌های سخت‌افزاری چندهسته‌ای موجود و استفاده مفید از قابلیت‌های پردازنده‌های چندهسته‌ای انجام پذیرفته است.

با برداشت داده‌های موجود در حوزه‌های مالیاتی، شامل ۱۵ حوزه مالیاتی در سطح مشاغل دارایی؛ در حدود ده هزار و بیست و هشت آزمایش صورت پذیرفت که از این تعداد مالیات‌دهنده، برای تعداد ۹۹۵ نفر هشدار متقلب بودن داده می‌شد که نتایج خروجی تشخیص تقلب این بخش در جدول (۶) قابل مشاهده است.

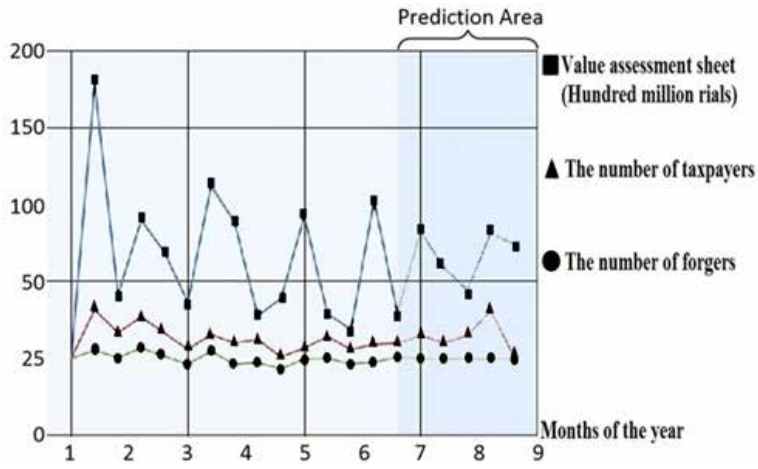
جدول (۶)- خروجی سیستم تقلب‌یاب در اثر آزمایش بر روی داده‌های مشاغل دارایی

خروجی سیستم با استفاده از موازی‌سازی حلقه، پرس‌وجو و بستر رایانش موازی			
کد حوزه	تعداد مالیات‌دهندگان	هشدار متقلب بودن	تعداد تاپل‌های پردازش شده مؤدیان از سابقه پرونده ده سال آنان، جهت تشخیص
۱	۳۰۳	۹	۲,۸۷۷,۰۹۴
۲	۱۲۶۴	۱۷۰	۲۰۳,۰۹۱,۶۸۳
۳	۸۰۶	۴۶	۱۳,۰۳۸,۸۸۸
۴	۱۱۱۷	۹۰	۷,۲۶۰,۲۷۷
۵	۸۰۲	۷۱	۶,۹۷۳,۱۱۷
۶	۴۷۶	۴۲	۱۰,۴۲۰,۶۳۰
۷	۷۷۱	۱۱۰	۱۲,۲۴۹,۹۵۶
۸	۶۰۳	۸۹	۴۵,۴۹۸,۸۸۳
۹	۶۷۰	۳۴	۱۹,۹۵۶,۶۴۷
۱۰	۳۳۳	۴۴	۴,۰۴۸,۶۴۱
۱۱	۵۰۱	۹۲	۵۷,۱۶۷,۹۰۸
۱۲	۷۳۳	۳۵	۲۹,۹۶۷,۰۹۲
۱۳	۴۷۳	۲۶	۹,۶۹۲,۵۴۶
۱۴	۵۷۸	۱۰۱	۲۷,۱۲۹,۴۲۴
۱۵	۵۹۸	۳۶	۹,۵۹۸,۹۰۵
جمع	۱۰۰۲۸	۹۹۵	۴۵۸,۹۷۱,۶۹۰

منبع: سامعی‌راد، ۲۰۱۵

با استفاده از اطلاعات موجود در جدول (۶) و مقداردهی متغیر مربوط به مرحله پیش‌بینی[1] به عدد یازده محدودهٔ پیش‌بینی را گسترش داده‌ایم. قسمت نقطه‌چین روی نمودار شکل (۱۰) نمایانگر محدوده پیش‌بینی تقلب می‌باشد.

شکل (۹)- پیش‌بینی مبلغ تقلب با آزمایش بر روی داده‌های سال ۸۶



منبع: سامعی‌راد، ۲۰۱۵

۹- نتیجه‌گیری

مالیات یکی از مهمترین منابع درآمدی دولت‌ها است. با توجه به افزایش تعداد مالیات دهندگان و همچنین حجم داده‌های آنها، دیگر سیستم‌های سنتی در این زمینه کارساز نیستند و به سیستم‌های هوشمند جهت اخذ و مدیریت داده‌های مالیاتی از جمله کشف تقلب‌های مالیاتی نیاز است. الگوریتم‌های استفاده شده در این حوزه با تکنیک‌های برنامه‌نویسی سریال معمولاً وقت‌گیر هستند. در این مقاله برخی از الگوریتم‌های مورد استفاده برای کشف تقلب، با استفاده از الگوهای پردازش موازی موجود در پلتفرم‌های برنامه‌نویسی موازی شده‌اند که این کار باعث گردید سرعت اجرای برنامه‌ها به مراتب بهتر و سریعتر شود.

مشخص گردید که قوانین همبستگی و شبکه‌های بیزین بهترین ساختار را در تشخیص تقلب بر روی داده‌های مالیاتی دارند همچنین خوشه‌بندی هم در تلفیق با این دو روش بهترین نتیجه را خواهد داد که در یک مطالعه دیگر به‌طور ترکیبی باید مورد تحقیق و پژوهش قرار گیرد.

1. Prediction Steps

فهرست منابع

۱. قانون مالیات‌های مستقیم، ۱۳۹۴: روزنامه رسمی؛ www.rrk.ir/laws/ShowLaw.aspxCode=5225.

2. P.C. González, J.D. Velásquez (2013). «Characterization and detection of taxpayers with false invoices using data mining techniques,» Expert Systems with Applications vol.40, no. 5, pp. 1427-1436.

3. R.S. Wu, C.S. O.U. H. Lin, S.I. Chang, DC Yen (2012). «Using data mining technique to enhance tax evasion detection performance,» Expert Systems with Applications vol.39, no. 10, pp. 8769-8777.

4. P. Ravisankar, V. Ravi, G.R. Rao, I. Bose (2011). «Detection of financial statement fraud and feature selection using data mining technique,» Decision Support Systems vol.50, no. 2, pp. 491-500.

5. E.W.T. Ngai, Y. Hu, Y.H. Wong, Y. Chen, X. Sun (2011). «The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature,» Decision Support Systems vol.50, no. 3, pp. 559-569.

6. K. R. Karkera (2014). «Building Probabilistic Graphical Models with Python,» Packt Publishing, ISBN: 978-1-78328-900-4, Open Source community experiances distilled, www.PacktPub.com.

7. V. Ajay, D.V. Ashoka, V.N. Aradya (2015). «Application of Data Mining Techniques for Defect Detection and Classification,» In Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014, pp. 387-395. Springer International Publishing.

8. M.S. Abadeh, S. Mahmoodi, M. Taherparvar (2012). "Application Data Mining," Niyaz Danesh Press.

9. J. Shahrabi, V. ShakorNiyaz (2007). "Data mining Concepts," Metalon Press.

10. A. Ahmadi, A. Mohebbi, "Business Intelligence: data mining and optimization,"

Amirkabir University Press, 2013.

11. Bond University, Central Michigan University, Deakin University; "Computational Data Mining Techniques in Automotive Insurance Fraud Detection"; Journal of Data Science 10(2012), 537-561.
12. F. Nonyelum (2011). "Data Mining Application in credit card fraud detection system," Journal of Engineering Science and Technology Vol. 6, 311 – 322.
13. C. SAKODA, A. NAGASAKI, T. ITOH, M. ISE, K. MIYASHITA (2011). "Visualization for Assisting Rule Definition Tasks of Credit Card Fraud Detection Systems," Journal of Data Science.
14. P. Murugavel, M. Punithavalli, "Improved Hybrid Clustering and Distance-based Technique for Outlier Removal," International Journal on Computer Science and Engineering (IJCSE).
15. Carlo Vercellis (2009). "Business Intelligence: Data Mining and Optimization for Decision Making," Politecnico di Milano, Italy, WILEY.
16. V. Dheepa, R. Dhanapal (2009). "Analysis of Credit Card Fraud Detection Methods," International Journal of Recent Trends Engineering, Vol 2, No. 3, Nov.
17. A. S. SABAU (2012). "Survey of Clustering based Financial Fraud Detection Research," Informatica Economică vol. 16, no. 1.
18. A. Sharma, P.K. Panigrahi (2012). "A Review of Financial Accounting Fraud Detection based on Data Mining Techniques," International Journal of Computer Applications (0975 – 8887) Volume 39– No.1, February.
19. P. Vagata (2015). "When Should I Use PLINQ? "Parallel Computing Platform Group, Microsoft Corporation".
20. Akash Verenkar (2015). "Using .NET4 Parallel Programming Model to Achieve Data Parallelism in Multi-tier Applications," Microsoft Corporation.
21. D. Leijen, W. Schulte, S. Burckhardt (2015). "The Design of a Task Parallel Library," Microsoft Corporation.

22. S. Okur, D. Dig (2015). "How do Developers Use Parallel Libraries?" MSDN Microsoft Corporation.
23. I. Ostrovsky (2015). "Parallel Programming in .NET 4," Parallel Computing Platform Group.
24. J. Fernando Ferreira, J. Luís Sobral, "ParC#: Parallel Computing with C# in .Net," Departamento de Informática - Universidade do Minho.
25. B. George, P. Nagpal (2015). "Optimizing Parallel Applications Using Concurrency Visualizer," Parallel Computing Platform Group.
26. Chi-Hung Lin, I-Chun Lin (2012). Ching-Huei Wu, Ya-Ching Yang and Jinsheng Roan, "The application of decision tree and artificial neural network to income tax audit: the examples of profit-seeking enterprise income tax and individual income tax in Taiwan," Journal of the Chinese Institute of Engineers Vol. 35, No. 4, June, 401.
27. Dr. Ela Kumar, Arun Solanki (2010). "A Combined Mining Approach and Application in Tax Administration," International Journal of Engineering and Technology Vol.2 (2), 38-44.
28. C. PHUA, V. LEE, K. SMITH & R. GAYLER (2010). "A Comprehensive Survey of Data Mining-based Fraud Detection Research," Monash University Press.
29. Xiaoqing Liu, Ding Pan, Shihong Chen (2010). "Application of Hierarchical Clustering in Tax Inspection Case-selecting," National Natural Science Foundation of China.
30. Ying Wang (2010). "Research on Rough Sets Theory Based Tax Data Mining," International Conference on Future Information Technology and Management Engineering.
31. Stefano Basta, Fabio Fassetti, Massimo Guarascio, Giuseppe Manco (2009)."High Quality True-Positive Prediction for Fiscal Fraud Detection", IEEE International Conference on Data Mining Workshops.