استفاده از خوشه‌بندی BIRCH و الگوریتم بهینه‌سازی واکنش شیمیایی جهت کشف تقلب در حوزه سلامت

مجید عبدالرزاق‌نژاد و مهدی خرد

چکیده: حوزه سلامت به علت وسعت عملکرد مالی و همچنین وسعت کاربرد آن، یکی از سیستم‌های ایده‌آل برای تقلب است و با وجود راهکارهای مختلف در این زمینه، شناسایی داده‌های تقلب هنوز یکی از چالش‌ها برای ارائه‌دهندگان خدمات سلامت می‌باشد. در این مقاله برای اولین بار الگوریتم BIRCH به عنوان یک الگوریتم خوشه‌بندی سلسله‌مراتبی با الگوریتم بهینه‌سازی واکنش شیمیایی (CRO) ترکیب شده است. الگوریتم BIRCH با پیچیدگی زمانی خطی قابلیت کار با حجم بالای داده‌ها و شناسایی داده‌های پرت را دارد و CRO یکی از الگوریتم‌های فرابتکاری جدید الهام‌گرفته از واکنش شیمیایی در دنیای واقعی است که با یک جمعیت پویا از مولکول‌ها توسط چهار عملگر برخورد به دیواره، تجزیه، برخورد بین مولکولی و ترکیب فضای جستجو را مورد کاوش قرار می‌دهند. الگوریتم خوشه‌بندی بهبودیافته BIRCH-CRO با حذف فرایند خوشه‌بندی سراسری داخلی نسخه کلاسیک BIRCH و تعیین بهینه پارامترهای اصلی آن باعث بهبود سرعت و دقت تشخیص داده‌های تقلب در حوزه سلامت نسبت به سایر الگوریتم‌های بدون نظارت ارائه‌شده در این حوزه گردیده است. همچنین الگوریتم پیشنهادی توانایی کار با داده‌های آنلاین و حجم بالا را دارد و با توجه به نتایج به دست آمده، عملکرد مناسبی را فراهم می‌کند.

کلیدواژه: **الگوریتم بهینه‌سازی واکنش شیمیایی، حوزه سلامت، خوشه‌بندی BIRCH، کشف تقلب.**

۱- مقدمه

کشف تقلب، یک مرحله و فرایند مهم در مواجهه با داده‌ها به خصوص داده‌های مالی است. تقلب به عنوان یک فریب عمده شناخته می‌شود که می‌تواند صورت‌های مختلفی را به خود بگیرد. تقلب در حوزه سلامت، سالانه موجب زیان مالی قابل توجهی می‌شود، به گونه‌ای که تقلب و سوء استفاده در درخواست‌های پزشکی تبدیل به نگرانی عمده برای شرکت‌های بیمه سلامت در دهه‌های گذشته شده است [۱]. پرداخت نامناسب توسط شرکت‌های بیمه یا پرداخت‌کننده شخص ثالث، معمولاً به علت خطا، سوء استفاده یا تقلب رخ می‌دهد. مقیاس این پرداخت‌های نامناسب به قدری بزرگ است که آن را به عنوان یک چالش اصلی در سیستم سلامت تبدیل کرده است [۲].

بسیاری از سیستم‌های بیمه سلامت، وابسته به متخصصان انسانی برای بازبینی درخواست‌ها و شناسایی موارد مشکوک هستند. این امر سبب زمان‌برشدن فرایند بازبینی درخواست‌ها می‌شود [۳].

این مقاله در تاریخ ۱۴ شهریور ماه ۱۳۹۷ دریافت و در تاریخ ۷ خرداد ماه ۱۳۹۸ بازنگری شد.

مجید عبدالرزاق‌نژاد (نویسنده مسئول)، گروه مهندسی کامپیوتر، دانشکده فنی مهندسی، دانشگاه بزرگمهر قائنات، قائن، ایران، (email: abdolrazzagh@buqaen.ac.ir).

مهدی خرد، گروه مهندسی کامپیوتر، دانشکده فنی مهندسی، دانشگاه قم، قم، ایران، (email: ma.kherad2@gmail.com).

ظهور فناوری‌های مختلف، فرصت‌هایی برای کشف بهتر تقلب را فراهم ساخته است. در این بین، نگرش داده‌کاوی به عنوان روشی مؤثر و کارامد مطرح می‌شود. تکنیک‌های داده‌کاوی امکان شناسایی الگوهای رفتاری را در بین حجم زیادی از داده فراهم می‌آورد که این امر امکان شناسایی و کشف الگوهای مشکوک و یا حتی تقلب را میسر می‌سازد [۳].

خوشه‌بندی به عنوان یک تکنیک داده‌کاوی که اقدام به گروه‌بندی داده‌های مشابه در یک خوشه می‌نماید، کاربرد گسترده در شناسایی داده‌های پرت دارد. خوشه‌بندی BIRCH به عنوان یک الگوریتم خوشه‌بندی سلسله‌مراتبی برای پایگاه‌های داده خیلی بزرگ مناسب است. هزینه I/O آن خطی است که متناسب با سایز مجموعه داده می‌باشد. در این روش با تنها یک بار بررسی مجموعه داده‌ها می‌توان به نتیجه خوشه‌بندی نسبتاً خوبی دست یافت. خوشه‌بندی BIRCH امکان موازی‌سازی را نیز فراهم می‌سازد و نسبت به نویز نیز حساس است [۴]. این الگوریتم خوشه‌بندی یک الگوریتم مستقل نیست. در واقع نیاز به یک الگوریتم خوشه‌بندی سراسری دارد تا عملیات خوشه‌بندی را خاتمه دهد. در بیشتر موارد الگوریتم خوشه‌بندی k-means جهت همکاری با این الگوریتم انتخاب شده است. استفاده از یک الگوریتم فرابتکاری جهت خوشه‌بندی می‌تواند نیاز به تعیین تعداد خوشه اولیه را از بین ببرد. الگوریتم بهینه‌سازی واکنش شیمیایی یکی از الگوریتم‌های فرابتکاری جدید با جمعیت پویا است که می‌توان در این راستا از آن بهره جست. ویژگی‌های مختلفی را می‌توان برای هر مولکول در نظر گرفت، امکان اجرای موازی را نیز دارد که می‌تواند سبب تسریع در پاسخگویی شود [۵].

استفاده از الگوریتم خوشه‌بندی به عنوان یک راهکار جهت کشف تقلب به دلیل عدم نیاز به دانش پیشین در خصوص تقلبی‌بودن یا نبودن داده می‌تواند عملکرد مناسبی نه تنها در کشف تقلب در داده‌ها و الگوهای جاری داشته باشد، بلکه می‌تواند سبب شناسایی الگوهای تقلبی جدید نیز شود. اکثر مطالعات مرورشده از روش‌های ترکیبی در این خصوص استفاده می‌کنند. در واقع آنها با ادغام روش با نظارت و بدون نظارت اقدام به کشف تقلب می‌کنند. این امر می‌تواند سبب بروز مشکل در مواجهه با الگوهای جدید شود و از این رو ضروری است تا مکانیزمی مناسب برای این چالش طراحی گردد. با توجه به موارد مطرح‌شده، در این مقاله سعی بر ارائه مکانیزیمی جدید برای کشف داده تقلب در بیمه سلامت توسط ادغام الگوریتم خوشه‌بندی BIRCH و الگوریتم بهینه‌سازی واکنش شیمیایی داریم. در فرایند الگوریتم خوشه‌بندی BIRCH کلاسیک، ابتدا باید درخت به طور کامل ساخته شود و بعد به بهبود برگ‌ها و عناصر موجود در زیر خوشه‌ها پرداخته شود که این امر منجر به افزایش زمان می‌گردد. از سوی دیگر پارامترهایی مانند فاکتور شاخه‌بندی و حد آستانه در ساخت درخت تأثیرگذار هستند.

در روش پیشنهادی سعی شده تا با ترکیب الگوریتم بهینه‌سازی واکنش شیمیایی و خوشه‌بندی BIRCH، این چالش رفع شود. در این

جدول ۱: تحقیقات انجام‌شده در حوزه سلامت با روش نظارت‌نشده.

مقاله	سال	راهکار	وجه تمایز
[۶]	۲۰۱۰	ادغام خوشه‌بندی و رگرسیون	استفاده از فاصله جغرافیایی برای کشف تقلب
[۷]	۲۰۱۱	استفاده از خوشه‌بندی k-means	-
[۸]	۲۰۱۱	خوشه‌بندی، استخراج ویژگی، حذف داده پرت، مدل‌سازی HMM	-
[۹]	۲۰۱۱	استفاده از خوشه‌بندی تکراری	قابل استفاده برای خوشه‌بندی آنلاین داده‌ها
[۳]	۲۰۱۳	خوشه‌بندی مبتنی Geo-location	استفاده از فاصله جغرافیایی برای کشف تقلب و استفاده از خوشه‌بندی بر طبق اطلاعات شخصی، مبلغ بیمه علاوه بر درخواست‌ها
[۱۰]	۲۰۱۳	ادغام co-clustering و مدل بیزین	توانایی کار با دیتاست‌های بزرگ
[۱۱]	۲۰۱۳	ادغام یادگیری ماشین و طبقه‌بندی	کار با داده‌های متغیر و پویا، توانایی شناسایی الگوهای جدید
[۱۲]	۲۰۱۵	ادغام الگوریتم خوشه‌بندی تکاملی و الگوریتم SVM	توانایی کار با وجود تغییر در داده‌ها و کار با داده‌های پویا
[۱۳]	۲۰۱۶	خوشه‌بندی مبتنی بر تراکم و درخت تصمیم	خوشه‌بندی بر طبق اطلاعات شخصی، مبلغ بیمه
[۱۴]	۲۰۱۶	الگوریتم شبکه عصبی بهبودیافته و خوشه‌بندی	-
[۱۵]	۲۰۱۶	استفاده از کشف اجتماع برای کشف تقلب	در نظر گرفتن رابطه بین پزشکان در کشف تقلب علاوه بر رابطه بین پزشک و بیمار- توانایی کار با داده‌های بزرگ
[۱۶]	۲۰۱۶	استفاده از k-means	عملکرد بلادرنگ
[۱۷]	۲۰۱۶	استفاده از k-means و ماتریس برخورد	-
[۱۸]	۲۰۱۷	استفاده از k-means	-
[۱۹]	۲۰۱۹	استفاده از PCA برای کاهش ابعاد داده و شبکه عصبی خودسازمانده SOFM برای کشف تقلب	کاهش ابعاد داده
[۲۰]	۲۰۱۸	استفاده از روش‌های آماری و بیزین	-
[۲۱]	۲۰۱۸	استفاده از الگوریتم داده‌کاوی بدون نظارت بر اساس متدولوژی CRISP	-

روش، الگوریتم بهینه‌سازی واکنش شیمیایی به جای حد آستانه و فاکتور خوشه‌بندی عمل می‌کند. هر بار که داده‌ای از مجموعه داده خوانده می‌شود، الگوریتم واکنش شیمیایی محل مناسب برای آن را مشخص می‌کند. در صورتی که داده تفاوت زیادی با سایر داده‌ها داشته باشد، الگوریتم واکنش شیمیایی دستور ایجاد یک برگ جدید را می‌دهد.

در ادامه این مقاله در بخش ۲، پیشینه تحقیق آمده و در بخش ۳، خوشه‌بندی BIRCH تشریح شده است. ساختار الگوریتم بهینه‌سازی واکنش شیمیایی در بخش ۴ و روش پیشنهادی در بخش ۵ توضیح داده شده‌اند. نتایج پیاده‌سازی روش پیشنهادی در بخش ۶ ارائه گردیده و نهایتاً خلاصه و نتیجه‌گیری رویکرد پیشنهادی در بخش ۷ جمع‌بندی شده است.

۲- پیشینه تحقیق

در این بخش به بررسی مطالعات و تحقیقات انجام‌شده در حوزه کشف تقلب خواهیم پرداخت. محققان بسیاری سعی در استفاده از تکنیک‌های داده‌کاوی داشتند و از این رو تمامی مطالعات در این خصوص را می‌توان به سه دسته تقسیم‌بندی کرد:

- **استفاده از روش‌های نظارت‌شده:** که با برچسب از قبل داده شده به داده‌ها در خصوص تقلبی‌بودن یا مشروع‌بودن آنها کار می‌کنند.
- **استفاده از روش‌های نظارت‌نشده:** که بدون توجه به برچسب داده‌شده به داده‌ها و با توجه به الگوی داده‌ها، آنها را تقسیم‌بندی کرده و سپس یک کارشناس با توجه به داده‌ها تقلبی‌بودن یا نبودن آنها را تعیین می‌کند.
- **استفاده از روش‌های نیمه‌نظارت‌شده:** که ترکیبی از روش‌های نظارت‌شده و نظارت‌نشده است.

از آنجا که موضوع این تحقیق استفاده از روش‌های نظارت‌نشده در کشف تقلب است در جدول ۱ به طور خلاصه مطالعات موجود در این زمینه آورده شده است.

به طور کلی به علت این که روش‌های بانظارت احتیاج به داده‌های برچسب‌دار جهت کشف تقلب دارند و اغلب یافتن این نوع داده‌ها دشوار است، استفاده از این نوع روش‌ها مناسب نیستند.

از سوی دیگر ممکن است صحت برچسب داده‌ها نیز به طور کامل تأیید نشود و به همین دلیل اغلب از روش بدون نظارت در این حوزه استفاده می‌شود. علاوه بر این در روش‌های بانظارت، انتخاب مجموعه تست و مجموعه آزمایش یک چالش جدی است. اندازه درست مجموعه آموزش یک پارامتر مهم در طبقه‌بندی است. با افزایش اندازه مجموعه آموزش، پیچیدگی مدل نیز افزایش می‌یابد و از تعداد خطاها کاسته می‌شود. اما این بدان معنی نیست که مجموعه آموزش بزرگ بهتر است زیرا یک مدل پیچیده با خطای کم ممکن است در مورد درخواست‌های جدید ضعیف عمل کند.

با توجه به این مطالعات و تحقیقات می‌توان گفت که در اکثر موارد از روش‌های نیمه‌نظارت‌شده جهت کشف تقلب استفاده شده است. این امر سبب بروز مشکلات و چالش‌هایی در خصوص کشف تقلب می‌گردد. اولین چالش در خصوص مفهوم drift است. داده‌های مربوط به سلامت می‌توانند تغییر کنند و سیستم کشف تقلب باید توانایی برخورد با تغییرات در داده‌ها را داشته باشد. این تغییرات می‌توانند در نوع داده‌ها و ویژگی‌های آنها رخ دهد. کشف تقلب در سیستم بلادرنگ معضل دیگری است که سیستم‌های کشف تقلب با آن مواجه هستند. همچنین اکثر سیستم‌های کشف تقلب توانایی کار با حجم زیاد داده را ندارد.

۳- خوشه‌بندی BIRCH

خوشه‌بندی یکی از تکنیک‌های پرکاربرد در داده‌کاوی و یک تکنیک یادگیری بدون نظارت است. خوشه‌بندی تقسیم داده‌ها به گروه‌هایی حاوی اشیای مشابه است. هر گروهی که خوشه نامیده می‌شود، شامل اشیایی

است که با هم متشابه اما متفاوت از اشیای گروه‌های دیگر هستند [۲۲]. انواع مختلف الگوریتم‌های خوشه‌بندی را می‌توان در ۴ گروه خوشه‌بندی سلسله‌مراتبی، خوشه‌بندی جزءبندی، خوشه‌بندی گرید و خوشه‌بندی مبتنی بر تراکم تقسیم‌بندی نمود [۲۲].

تکنیک خوشه‌بندی سلسله‌مراتبی یکی از تکنیک‌های پرکاربرد در خوشه‌بندی داده‌های حجم بالا مانند داده‌های سری زمانی است. خوشه‌بندی سلسله‌مراتبی فاصله جفت‌ها را از یکدیگر محاسبه و سپس خوشه‌های مشابه را به روش پایین به بالا، بدون نیاز به تأمین تعداد خوشه‌ها ادغام می‌کند [۲۳]. خوشه‌بندی سلسله‌مراتبی یک درخت از خوشه‌ها می‌سازد که به عنوان دندروگرام شناخته می‌شود. هر نود خوشه، شامل چند خوشه فرزند است.

الگوریتم خوشه‌بندی BIRCH یکی از الگوریتم‌های مطرح در خوشه‌بندی سلسله‌مراتبی است. این روش با کمک تکنیک خوشه‌بندی سلسله‌مراتبی و تکنیک دیگر (معمولاً k-means) برای کار بر روی داده‌های عددی با حجم بالا طراحی شده است. این الگوریتم از ویژگی خوشه یا به اختصار CF جهت انجام خوشه‌بندی استفاده می‌کند و همین امر سبب شده تا الگوریتم از سرعت و قابلیت مقیاس‌پذیری بالایی برخوردار باشد [۲۴]. ویژگی خوشه، یک روش خلاصه‌سازی اطلاعات است که نقاط داده برای هر زیرخوشه را نگهداری می‌نماید. در الگوریتم BIRCH از درخت متوازن استفاده می‌شود. درخت متوازن CFها را جهت اجرای الگوریتم ذخیره می‌کند. هر گره از این درخت که دارای گره فرزند است، در خود مجموع CFهای فرزندانش را نگهداری می‌نماید. بدین ترتیب اطلاعات خلاصه‌شده‌ای از فرزندان خود را خواهد داشت. درخت دارای دو پارامتر به نام‌های فاکتور شاخه‌بندی B و حد آستانه T است. حداکثر تعداد فرزندان گره‌های غیر پایانی با فاکتور شاخه‌بندی B مشخص می‌شوند و حد آستانه T به حداکثر فاصله‌ای که میان دو نمونه از خوشه‌های برگ قرار دارد اطلاق می‌شود [۲۴].

پس از درج و یا حذف یک CF در درخت و یا در صورت افزایش مقادیر B و T از مقدار تعیین‌شده توسط کاربر، گره‌های درخت متوازن تقسیم و یا ادغام می‌شوند. واضح است که مقادیر B و T در اندازه درخت تولیدشده نهایی، نقش بسزایی ایفا می‌کنند.

٤- الگوریتم بهینه‌سازی واکنش شیمیایی

بهینه‌سازی واکنش شیمیایی (CRO) روش جدیدی برای بهینه‌سازی است که از ماهیت واکنش‌های شیمیایی الهام گرفته است. یک واکنش شیمیایی، یک فرایند طبیعی تبدیل مواد ناپایدار به مواد پایدار است. در دیدگاه میکروسکوپی، یک واکنش شیمیایی با برخی مولکول‌های ناپایدار با انرژی بیش از حد آغاز می‌شود. این مولکول‌ها از طریق دنباله‌ای از واکنش‌های ابتدایی با یکدیگر تعامل دارند. در پایان، آنها به مولکول‌هایی با حداقل انرژی تبدیل می‌شوند. این ویژگی در CRO برای حل مشکلات بهینه‌سازی تعبیه شده است [۵].

CRO تمام جزئیات واکنش شیمیایی را در نظر نمی‌گیرد و از اصول واکنش‌های شیمیایی تنها دو قانون اول ترمودینامیک را در نظر می‌گیرد. اولین قانون حفاظت از انرژی است. انرژی نمی‌تواند ایجاد یا نابود شود و تنها از یک فرم به دیگری تبدیل می‌شود و از یک نهاد به دیگری انتقال می‌یابد. یک سیستم واکنش شیمیایی شامل مواد شیمیایی و محیط اطراف آن است. هر ماده شیمیایی دارای انرژی پتانسیل و انرژی جنبشی است و انرژی محیط اطراف به وضوح توسط بافر انرژی مرکزی در CRO نشان داده می‌شود. واکنش گرماگیر، به حرارت به دست آمده از محیط اطراف برای راه‌اندازی واکنش نیاز دارد. واکنش گرماده به واکنش‌هایی اطلاق می‌شود که مواد شیمیایی گرمای واکنش خود را به محیط می‌دهند. این دو نوع واکنش‌ها می‌تواند با اندازه بافر اولیه مشخص شود: زمانی که مثبت باشد، واکنش آن گرماگیر است و هنگامی که آن صفر است، واکنش گرمازا است [۵].

قانون دوم می‌گوید که آنتروپی سیستم گرایش به افزایش دارد. آنتروپی اندازه‌گیری میزان اختلال است. انرژی بالقوه، انرژی ذخیره‌شده در یک مولکول با توجه به پیکربندی مولکولی آن است. هنگامی که مولکول به شکل‌های دیگر تبدیل می‌شود سیستم آشفته‌تر می‌گردد. به عنوان مثال هنگامی که مولکول‌هایی با انرژی جنبشی بیشتر (که از انرژی پتانسیل تبدیل می‌شوند) سریع‌تر حرکت می‌کنند، سیستم بیشتر اختلال ایجاد می‌کند و آنتروپی آن افزایش می‌یابد. بدین ترتیب تمام سیستم‌های واکنشی تمایل به رسیدن به حالت تعادل دارند که انرژی پتانسیل آن به حداقل برسد. در CRO این پدیده را با تبدیل انرژی پتانسیل به انرژی جنبشی و به تدریج از دست دادن انرژی مولکول‌های شیمیایی به محیط اطراف در نظر می‌گیریم [۵].

CRO یک الگوریتم فراابتکاری مبتنی برجمعیت با جمعیت پویا است. به عبارت دیگر، تعداد کل مولکول‌ها (جمعیت) در تکرارهای مختلف می‌تواند متغیر باشد. هنگامی که یک برخورد غیرمستقیم (دیوار یا بین مولکولی) اتفاق می‌افتد، تعداد مولکول‌ها قبل و بعد از واکنش یکسان باقی می‌مانند. از سوی دیگر یک برخورد مستقیم مانند تجزیه باعث افزایش تعداد مولکول‌ها و ترکیب باعث کاهش تعداد مولکول‌ها بعد از واکنش‌های یادشده می‌شوند که این تغییرات تحت تاثیر مقادیر متغیرهای α و β می‌باشد.

نکات ضروری در CRO عبارتند از [۵]:

۱) ساختار مولکول ω: یک جواب مسئله است، نیاز به فرمت خاصی ندارد و می‌تواند عدد، بردار یا ماتریس باشد.

۲) انرژی پتانسیل PE: انرژی پتانسیل به عنوان مقدار تابع هدف (برازندگی) مربوط به جواب ω است. اگر f به عنوان تابع هدف مشخص شود داریم

$$PE_{\omega} = f(\omega) \qquad (۱)$$

۳) انرژی جنبشی KE: یک مقدار عددی غیر منفی است که تحمل سیستم را از پذیرش بدترین جواب موجود مشخص می‌کند.

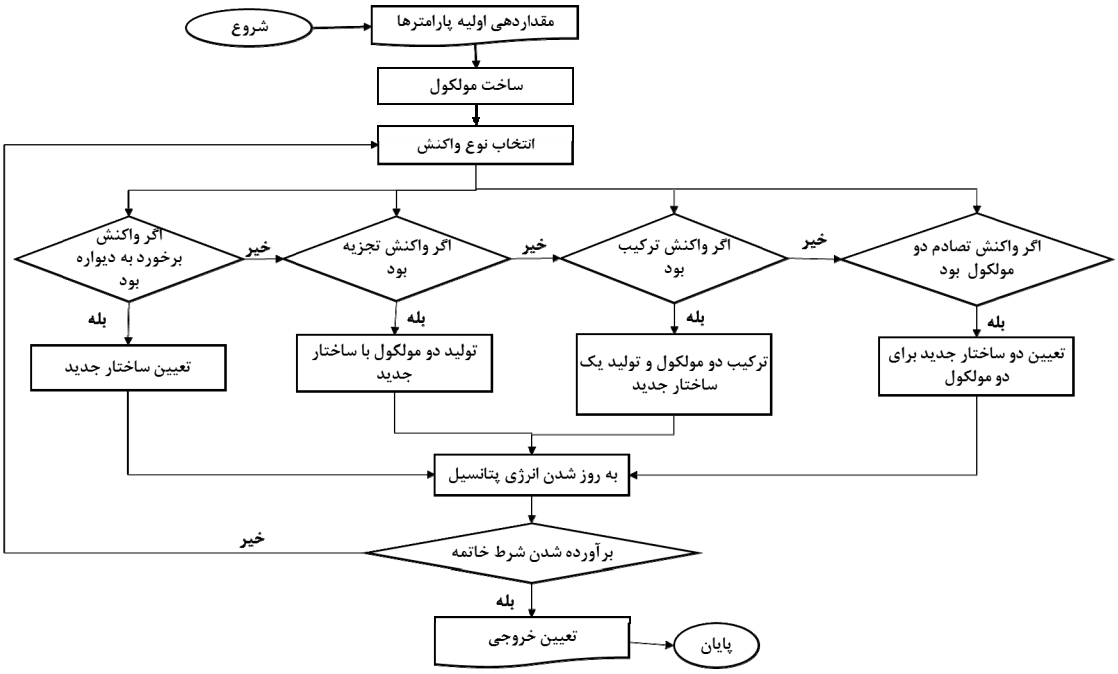
در ادامه مراحل الگوریتم و نحوه استفاده از این واکنش‌ها که در شکل ۱ نمایش داده شده است تشریح می‌شود.

مرحله اول (مقداردهی اولیه)

اندازه جمعیت، بافر، انرژی جنبشی اولیه، α و β مقداردهی می‌شود. همچنین مولکول ساخته و ساختار آن مشخص می‌شود.

مرحله دوم (انتخاب نوع واکنش)

در این گام یکی از واکنش‌های پایه، انتخاب می‌شود. جهت انتخاب نوع واکنش یک عدد تصادفی در بازه صفر تا یک مشخص و اگر این عدد از حد آستانه MoleColl بیشتر بود، واکنش تک‌مولکولی داریم وگرنه نوع واکنش، بین مولکولی خواهد بود. سپس با توجه به نوع واکنش، یک مولکول برای واکنش‌های تک‌مولکولی یا دو مولکول برای واکنش‌های بین مولکولی انتخاب می‌شوند. انتخاب مولکول(ها) نیز به تصادف صورت می‌گیرد.



شکل ۱: روندنمای الگوریتم بهینه‌سازی واکنش شیمیایی.

مرحله سوم (اجرای واکنش)

اگر واکنش انتخاب‌شده، برخورد به دیواره باشد، ابتدا انرژی پتانسیل جدید برای آن تعیین می‌شود. اگر انرژی پتانسیل و انرژی جنبشی در حالت پیش از برخورد، بیشتر از انرژی پتانسیل جدید باشد، یعنی امکان وجود چنین واکنشی باشد، انرژی جنبشی و انرژی پتانسیل و همچنین بافر به روز می‌شوند. سپس انرژی پتانسیل کمینه با انرژی پتانسیل جدید برای مولکول مقایسه می‌شود و در صورتی که انرژی پتانسیل جدید، کمتر باشد، به عنوان انرژی پتانسیل بهینه ذخیره می‌شود.

اگر نوع واکنش انتخاب‌شده، تجزیه باشد ابتدا دو ساختار مولکولی جدید تولید می‌شود که مقادیر انرژی جنبشی و پتانسیل آن بر طبق مقدار اولیه، مقداردهی می‌شوند. اگر مجموع انرژی‌های پتانسیل و جنبشی مولکول بیشتر از مجموع انرژی‌های پتانسیل برای دو مولکول جدید باشد، یعنی واکنش انجام‌پذیر است و در نتیجه، مولکول به دو مولکول جدید شکسته می‌شود.

اگر نوع واکنش، برخورد بین مولکولی باشد، بعد از تعیین دو مولکول برای برخورد، انرژی پتانسیل جدید برای هر دو تعیین می‌شود. تفاوت بین انرژی پتانسیل جدید با مجموع انرژی پتانسیل و انرژی جنبشی مولکول قبل از واکنش برای هر دو مولکول محاسبه می‌شود. اگر این مقدار بیشتر از بافر باشد یعنی واکنش قابل انجام است و در نتیجه، انرژی جنبشی و انرژی پتانسیل برای هر دو مولکول به روز می‌شود. همچنین میزان انرژی پتانسیلِ مینیمم نیز در صورتی که مقدار انرژی پتانسیل جدید کمتر باشد به روز می‌شود.

اگر نوع واکنش، ترکیب باشد، ابتدا یک ساختار مولکولی جدید تولید می‌شود و سپس انرژی پتانسیل و جنبشی آن بر طبق انرژی‌های دو مولکول منتخب برای اجرای واکنش محاسبه می‌شوند.

مرحله چهارم (شرط خاتمه)

در این گام بررسی می‌شود که شرط خاتمه برآورده شده است یا خیر. شرط خاتمه بر طبق نوع مسئله تعیین می‌گردد و می‌تواند رسیدن به یک تکرار مشخص یا رسیدن به یک مقدار مشخص برای انرژی پتانسیل باشد.

۵- روش پیشنهادی BIRCH-CRO

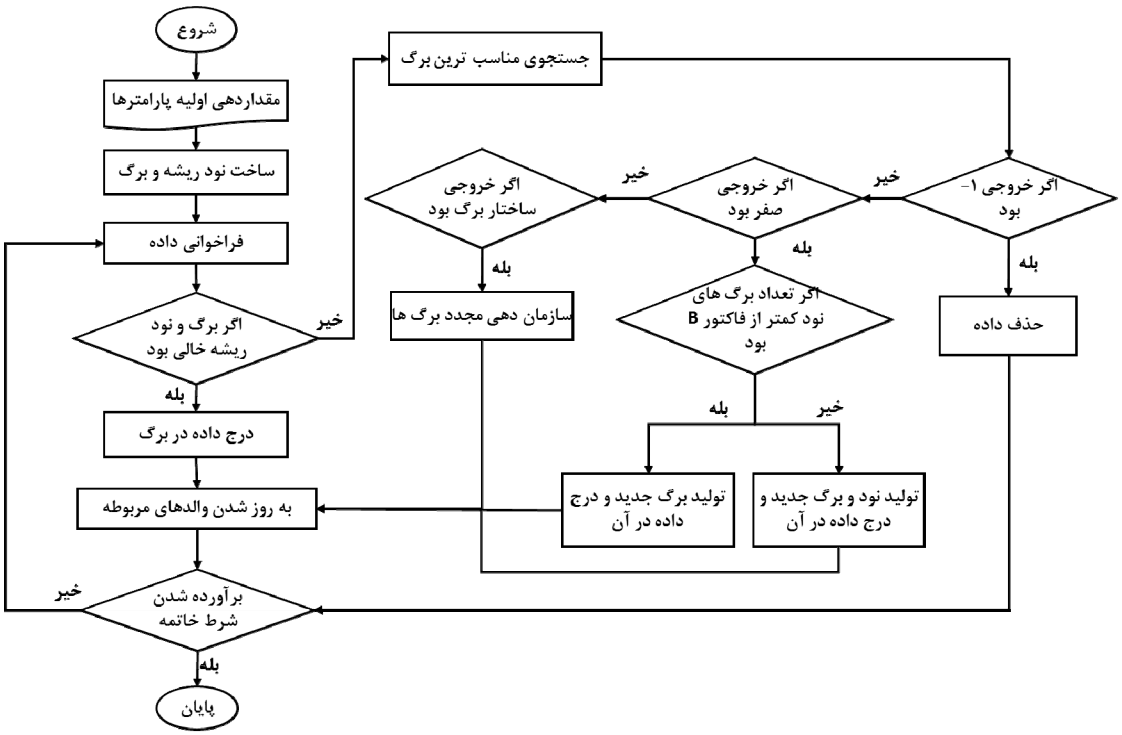
در این بخش به ارائه کلیات الگوریتم پیشنهادی خواهیم پرداخت. الگوریتم پیشنهادی، وظیفه کشف تقلب در درخواست‌های بیمه را بر عهده دارد. داده تقلب می‌تواند به عنوان داده‌های نامتعارف و ناهنجار و به طور رسمی‌تر به عنوان داده پرت در نظر گرفته شود. در روش پیشنهادی از ادغام الگوریتم خوشه‌بندی BIRCH و الگوریتم بهینه‌سازی واکنش شیمیایی استفاده می‌کنیم. الگوریتم BIRCH همان طور که قبلاً نیز ذکر گردید، از درخت متوازن برای خوشه‌بندی استفاده می‌کند. بعد از تولید درخت با توجه به ابعاد داده‌ها و معیار تشابه برای داده‌ها، درخت می‌تواند دارای تعداد زیادی برگ و در نتیجه خوشه باشد، پس نیاز به مکانیزمی برای اصلاح و ادغام این خوشه‌ها است. در این الگوریتم CRO به جای حد آستانه و فاکتور خوشه‌بندی عمل می‌کند. هر بار که داده‌ای از مجموعه داده خوانده می‌شود، CRO محل مناسب برای آن را مشخص می‌کند. در صورتی که داده تفاوت زیادی با سایر داده‌ها داشته باشد، CRO دستور ایجاد یک برگ جدید را می‌دهد. در ادامه مراحل کار در روش پیشنهادی که در شکل ۲ نیز روندنمای کلی آن ترسیم شده است آمده است:

گام اول: مقداردهی اولیه به پارامترها

در این قسمت پارامترهای اولیه برای الگوریتم واکنش شیمیایی مقداردهی می‌شوند. در الگوریتم CRO، تعداد اولیه مولکول‌ها و همچنین تعداد تکرار مقداردهی اولیه می‌شود. همان طور که گفته شد، درخواست‌ها دارای مقادیر عددی و غیر عددی هستند. در این بخش، صفات عددی و غیر عددی مشخص می‌شوند.

گام دوم: پیش‌پردازش داده

در این مرحله ابتدا داده خوانده شده و سپس اطلاعات تمامی فیلدها باتوجه به اطلاعات فراداده (متا دیتا) مورد بررسی قرار می‌گیرد. لذا داده‌هایی که فاقد فیلدهای معتبر باشند، حذف می‌شوند.



شکل ۲: روندنمای روش پیشنهادی.

شکل ۳: ساختار مولکول.

گام سوم: ساخت درخت

در این مرحله درخت متوازن با توجه به کمک الگوریتم CRO ساخته می‌شود. دو نوع عنصر در درخت وجود دارد: نود و برگ. هر نود حاوی CF یا ویژگی خوشه و هر برگ حاوی اطلاعات مربوط به هر درخواست است. در اینجا CF حاوی چهار بخش است. بخش اول تعداد عناصر موجود در خوشه، بخش دوم میانگین مبالغ هر درخواست، بخش سوم، شامل نام بیماری و بخش آخر شامل لیستی از صفات غیر عددی است. معیار شباهت در اینجا برابر با یکسان‌بودن نام بیماری و همچنین کم‌بودن میزان تفاوت در مبلغ بیمه است. اما جهت ساخت درخت مراحل زیر انجام می‌پذیرد:

۱) تولید یک نود ریشه و یک نود برگ
۲) فراخوانی داده
۳) اگر برگ و در نتیجه نود ریشه خالی بود داده را به برگ اضافه کن.
۴) در غیر این صورت، مناسب‌ترین برگ برای درج داده در درخت را پیدا کن. یافتن برگ مناسب توسط الگوریتم CRO انجام می‌پذیرد.
۵) اگر خروجی الگوریتم CRO، ساختار جدید برای برگ بود، برگ‌ها را دوباره سازمان‌دهی کن و عمل به روز رسانی والدهای برگ را انجام بده و در غیر این صورت اگر خروجی آن ۰ بود، یک برگ جدید در نود مورد نظر تولید کن. در غیر این صورت اگر خروجی ۱- بود، یعنی داده تقلبی یا تکراری بوده و از مجموعه داده و عملیات حذف می‌شود.
۶) اطلاعات CF مربوط به تمامی والدهای مربوط به برگی را که داده در آن درج شده است به روز کن.

همان طور که اشاره شد برای جستجو و یافتن گره مناسب در درخت از الگوریتم CRO استفاده می‌شود. ابتدا توسط ریشه و اطلاعات CF موجود در آن مناسب‌ترین و نزدیک‌ترین فرزند ریشه انتخاب می‌شود. سپس داده‌های موجود در این فرزند، جهت انجام محاسبات استخراج می‌شوند. سپس الگوریتم CRO وارد عمل می‌گردد و ساختار مولکول‌ها شکل داده می‌شود. ساختار هر مولکول حاوی چند مرکز خوشه است. به عبارت بهتر هر مولکول یک راه‌حل کامل را نشان می‌دهد. تعداد مراکز خوشه یا برابر با تعداد برگ‌های نود والد و یا یک واحد بیشتر از آن است. سپس داده‌های موجود برای هر مولکول تعیین و عمل خوشه‌بندی انجام می‌شود. یعنی داده‌ها به نزدیک‌ترین مرکز خوشه تخصیص داده می‌شوند. آن گاه با استفاده از تابع برازندگی، میزان مطلوبیت هر نود مشخص می‌شود. در این الگوریتم به دنبال حداقل‌کردن فاصله داده‌های موجود در خوشه با مراکز خوشه هستیم. البته تعداد کمینه خوشه نیز معیار دیگری است که باید برآورده شود. در نتیجه تابع برازندگی شامل میزان فاصله در هر خوشه یا به عبارت بهتر شعاع هر خوشه در تعداد خوشه است. رابطه (۲) تابع برازندگی را به صورت فرموله نمایش می‌دهد. سپس واکنش‌های شیمیایی به تصادف انجام می‌پذیرد تا زمانی که به شرط خاتمه که تعداد تکرارها است نزدیک شویم، این عملیات صورت می‌پذیرد

$$fit = k \times \sum_{i=1}^{k} \frac{d_i}{Num\ data\ in\ cluster} \qquad (۲)$$

که در آن k تعداد مراکز خوشه در هر مولکول و d_i میزان فاصله خوشه با داده‌هایش است. در شکل ۳ ساختار مولکول ترسیم شده است.

در الگوریتم پیشنهادی، برخلاف الگوریتم BIRCH معمولی، تنها

جدول ۲: مجموعه داده مورد استفاده برای ارزیابی.

داده	تعداد
داده تکراری	۲۰۰
درخواست با مقادیر نامتعارف	۲۰۰
قیمت نامتعارف	۲۰
تعداد داده‌های صحیح	۱۰۰۰
جمع کل	۱۴۲۰

زمانی یک خوشه جدید تولید می‌شود که معیار شباهت برآورده نشود و در واقع سبب خوشه‌بندی مناسب داده‌ها شود. در حالی که در BIRCH معمولی، در صورت برآورده‌نشدن حد آستانه یا زیادبودن تعداد داده‌ها در یک برگ، برگ جدید تولید می‌شود که این امر نه تنها منجر به افزایش بی‌رویه داده‌ها می‌شود بلکه برای تعیین حد آستانه نیاز به شناخت دقیق ازداده‌ها داریم که در روش خوشه‌بندی بدون نظارت این عمل دشوار و گاه، ناممکن است. همچنین در BIRCH معمولی، در مرحله آخر بعد از ساخت درخت، نیاز به استفاده از الگوریتم خوشه‌بندی سراسری است در حالی که در الگوریتم BIRCH-CRO این نیاز در زمان اضافه‌کردن یک داده به درخت رفع می‌شود و در نتیجه نیاز به استفاده از الگوریتم خوشه‌بندی سراسری نیست.

٦- پیاده‌سازی و ارزیابی

جهت پیاده‌سازی الگوریتم BIRCH-CRO از نرم‌افزار Matlab ۲۰۱۷ استفاده شده و Heart Attack Payment-Hospital [۲۵] مجموعه داده مورد استفاده است. این مجموعه داده، مبلغ پرداخت‌شده برای بیماران را با توجه به سطح مراقبت تعیین نموده است. این مجموعه حاوی حدود ۲۰۰۰۰ رکورد و ۲۳ ویژگی است. تقلب در حوزه سلامت می‌تواند ابعاد مختلفی داشته باشد. در این مقاله با توجه به مجموعه داده‌های در دسترس به موارد زیر رسیدگی می‌شود:

- **درخواست‌های تکراری:** برای مثال دو درخواست با شناسه یکسان وجود داشته باشد.
- **ارائه درخواست‌های نامتعارف:** برای مثال نام و اطلاعات مرکز ارائه‌دهنده خدمات سلامت وجود ندارد و یا اشتباه است. یا شناسه درخواست از فرمت مجاز پیروی نمی‌کند.
- **قیمت نامتعارف برای خدمات:** این نوع از تقلب مربوط به درخواست‌هایی می‌شود که دارای مبلغی غیر متعارف هستند. برای مثال درخواست‌های دیگر با نوع بیماری و خدمات یکسان، دارای مبلغی کمتر از مبلغ قیدشده در یک درخواست باشند.

در این بخش هر سه نوع تقلب در نظر گرفته شده و سپس عملیات ارزیابی و مقایسه صورت گرفته است. در این راستا از مجموعه داده یک زیرمجموعه در نظر گرفته شده و عملیات مربوط به کشف تقلب انجام گرفت. جدول ۲ مجموعه داده مورد نظر را شرح می‌دهد.

به دلیل این که روش پیشنهادی یک روش بدون نظارت مبتنی بر خوشه‌بندی است، نتایج به دست آمده توسط روش پیشنهادی BIRCH-CRO با روش‌های مرسوم خوشه‌بندی استفاده‌شده برای کشف تقلب [۷]، [۸] و [۱۶] تا [۱۸] یعنی الگوریتم k-means، DBSCAN و BIRCH مقایسه شده است. شایان ذکر است که در کارهای گذشته پایه تعریف تقلب و نیز داده‌های محک مورد استفاده بسیار متنوع بوده است. مثلاً در برخی مقالات فاصله مکانی کاربر تا فراهم‌کننده بیمه به عنوان شاخص شناسایی تقلب مورد توجه قرار گرفته است. در برخی از مقالات رابطه نوع بیماری با تشخیص بیماری و هزینه درمان و در برخی رابطه بین پزشک و بیمار به عنوان پایه تعریف تقلب در حوزه سلامت در نظر گرفته شده است. همچنین داده‌های مورد استفاده در اکثر مقالات نیز به سبب داده دنیای واقعی بودن مثلاً داده جمع‌آوری شده از مرکز پزشکی و بیمه آمریکا (Medicare and Medicaid Services) دارای دسترسی عمومی نیستند و لذا در این مقاله الگوریتم‌های مطرح‌شده در فوق برای داده Heart Attack Payment-Hospital که در UCI دسترسی عمومی دارد، پیاده‌سازی شده و بهترین نتایج به دست آمده از ارزیابی کارایی این الگوریتم‌ها در شناسایی تقلب‌های سه‌گانه تعریف‌شده در جدول ۳ ارائه گردیده است. مطابق با جدول ۳، الگوریتم پیشنهادی توانسته پاسخ بهتری را نسبت به سایر الگوریتم‌ها از لحاظ معیارهای ارزیابی مختلف از جمله زمان به دست آورد.

در بررسی پیچیدگی زمانی الگوریتم‌های خوشه‌بندی BIRCH، K-means و DBSCAN مشاهده می‌شود که اگر m تعداد نقاط، n تعداد ویژگی‌ها، k تعداد خوشه‌ها و l تعداد تکرارهای لازم در فرایند k-means باشد، آن گاه پیچیدگی زمانی این الگوریتم $O(n\times m\times k\times l)$ خواهد بود. این پیچیدگی برای DBSCAN برابر با $O(m^2)$ و برای BIRCH نیز برابر با $O(n\times m\times f\times r)$ است که در آن f تعداد گره‌ها و r تعداد ورودی‌های هر گره می‌باشد [۲۶]. این بدان معناست که الگوریتم‌های k-means و BIRCH دارای پیچیدگی زمانی خطی و DBSCAN دارای پیچیدگی زمانی درجه دوم است. همچنین k-means به واسطه چالش در تعیین تعداد بهینه خوشه‌ها و مراکز مناسب آنها و نیز نقص در یافتن خوشه‌های نامحدب در پیاده‌سازی به زمان بیشتری نسبت به خوشه‌بندی BIRCH نیاز دارد. این مهم در نتایج به دست آمده جدول ۳ نیز آشکار است.

با توجه به این نتایج، الگوریتم پیشنهادی به علت عدم نیاز به مقایسه تک‌تک داده‌ها با مرکز خوشه یا یکدیگر، زمان کمتری برای خوشه‌بندی داده‌ها نیاز دارد و از سوی دیگر به علت عدم نیاز به تعیین تعداد خوشه‌ها، به راحتی می‌تواند تعداد خوشه‌های مربوط را تعیین کند. عدم تعیین خوشه و توانایی کار با تعداد داده زیاد و همچنین زمان خوشه‌بندی کوتاه سبب برتری الگوریتم BIRCH شده است. با این حال الگوریتم BIRCH-CRO، دارای مزیتی است که سبب برتری آن از الگوریتم BIRCH می‌شود و آن عدم نیاز به تعیین حد آستانه است. در الگوریتم BIRCH معمولی نیاز به تعیین یک حد آستانه است که شعاع خوشه را تعیین می‌کند. اگر این حد آستانه کم باشد سبب می‌شود تا داده‌هایی که فاصله بسیار کمی از هم دارند در خوشه قرار گیرند و در نتیجه خوشه تعداد کمی داده در خود جای می‌دهد که این امر سبب زیادشدن تعداد برگ‌ها می‌شود. ازدیاد تعداد برگ‌ها سبب بزرگ‌شدن درخت و در نتیجه کاهش توانایی مدیریت درخت و افزایش زمان پاسخ می‌شود. از سوی دیگر تعیین مقدار حد آستانه زیاد نیز باعث می‌شود که شعاع خوشه زیاد شود و چه بسا داده‌های پرت نیز به خوشه راه یابند. در الگوریتم پیشنهادی، شعاع و حد آستانه توسط CRO تعیین می‌شود و یک مقدار ثابت و پایدار نیست. از سوی دیگر حد آستانه با توجه به مقدار داده‌ها تعیین می‌شود. برای مثال داده‌های بزرگ که دارای تفاوت دوهزارتایی هستند، نمی‌توان حد آستانه ۰/۱ را برای آن قرار داد. لذا تعیین حد آستانه تنها باید توسط یک فرد خبره و برای هر ویژگی در مجموعه داده صورت گیرد. اما در روش پیشنهادی سیستم بدون نیاز به حد آستانه می‌تواند برگ مناسب را جهت درج داده پیدا کند. این امر می‌تواند سبب بهبود عملکرد سیستم شود و همچنین در مواجهه با الگوها و درخواست‌های جدید نیز به خوبی کار کند.

جدول ۳: نتایج به دست آمده برای کارایی شناسایی تقلب.

روش/ معیار	Accuracy	Sensitivity	Specificity	Precision	F-score	time
BIRCH	۰/۹۹۴	۰/۹۹۸	۰/۹۸۵	۰/۹۹۴	۰/۹۹۵	۰/۶۳
K-means	۰/۹۵۲	۱	۰/۸۳۸	۰/۹۳۶	۰/۹۶۶	۰/۷۹
DBSCAN	۰/۹۷۹	۱	۰/۹۳۰	۰/۹۷۱	۰/۹۸۵	۱/۲
BIRCH-CRO	۰/۹۹۶	۱	۰/۹۸۸	۰/۹۹۵	۰/۹۹۷	۰/۳۸

۷- نتیجه‌گیری و پیشنهادها

حوزه سلامت به علت رونق زیاد آن دارای وسعت مالی زیادی است و همین امر باعث شده تا متقلبان و سوءاستفاده‌کنندگان به سمت حوزه سلامت جذب شوند. در این راستا ضروری است تا روشی جهت کشف تقلب در حوزه سلامت طراحی گردد.

این مقاله توسط روش خوشه‌بندی که یک راهکار داده‌کاوی بدون نظارت است، یک مکانیزم برای کشف تقلب ایجاد نموده است. روش پیشنهادی در این پژوهش، حاصل ادغام الگوریتم خوشه‌بندی BIRCH و الگوریتم بهینه‌سازی واکنش شیمیایی است. ابتدا صفات مربوط به داده‌ها به دو دسته عددی و غیر عددی تقسیم می‌شوند که این کار از قبل توسط یک متخصص صورت می‌گیرد. سپس داده به الگوریتم BIRCH، ارسال و در این الگوریتم درخت CF ساخت می‌شود. در هنگام ساخت درخت از الگوریتم CRO جهت جستجوی بهترین برگ استفاده می‌گردد و در الگوریتم CRO به تعداد برگ‌های موجود خوشه تولید می‌شود.

جهت پیاده‌سازی الگوریتم پیشنهادی از برنامه Matlab استفاده شد و سپس برای ارزیابی و تعیین عملکرد مکانیزم پیشنهادی، با الگوریتم‌های خوشه‌بندی دیگر شامل k-means، DBSCAN و الگوریتم BIRCH مورد مقایسه قرار گرفت.

این الگوریتم، مزیت‌های الگوریتم BIRCH را برای خوشه‌بندی حجم زیاد داده در زمان کوتاه و همچنین مزیت سرعت بالا برای جستجوی مناسب در بین داده‌ها برای یافتن داده مشابه را از الگوریتم CRO به ارث می‌برد. در واقع ترکیب این دو الگوریتم سبب شده تا بتوان در زمان کوتاه درخت متوازن را شکل و عملیات خوشه‌بندی را انجام داد. نتایج به دست آمده نیز همین امر را نشان می‌دهند.

الگوریتم پیشنهادی به علت عدم نیاز به تعیین حد آستانه می‌تواند طیف وسیعی از داده‌ها را خوشه‌بندی کند. در الگوریتم BIRCH معمولی نیاز به تعیین یک حد آستانه است که شعاع خوشه را تعیین می‌کند. اگر این حد آستانه کم باشد، سبب کوچک‌شدن خوشه و در نتیجه قرارگرفتن تعداد کمی داده در خوشه می‌شود که این امر سبب زیادشدن تعداد برگ‌ها، بزرگ‌شدن درخت و در نتیجه کاهش امکان مدیریت و افزایش زمان پاسخ می‌شود. از سوی دیگر تعیین مقدار حد آستانه زیاد نیز باعث افزایش شعاع خوشه و قرارگرفتن داده‌های پرت و نامربوط در خوشه می‌گردد. تعیین مناسب حد آستانه تنها توسط دانستن محدوده تغییرات داده‌ها امکان‌پذیر است و از این جهت یک چالش بزرگ در به کارگیری الگوریتم BIRCH وجود دارد.

در الگوریتم پیشنهادی، شعاع و حد آستانه توسط CRO تعیین می‌شود و یک مقدار ثابت و پایدار نیست. از سوی دیگر حد آستانه با توجه به مقدار هر ویژگی از داده‌ها در مجموعه داده تعیین می‌شود که این امر می‌تواند سبب بهبود عملکرد سیستم و افزایش آن در مواجهه با الگوها و درخواست‌های جدید شود.

با وجود عملکرد مناسب مکانیزم پیشنهادی اما لازم است تا کارایی آن بر روی مجموعه داده‌های بیشتری و یا حتی در حوزه‌های دیگری به جز سلامت مورد ارزیابی قرار گیرد. علاوه بر این مکانیزم پیشنهادی، ابعاد کمی از حالت‌های تقلب را در نظر می‌گیرد. بررسی عملکرد چنین راهکاری با دیگر حالات تقلب می‌تواند منجر به بهبود مکانیزم کشف تقلب شود. همچنین در روش پیشنهادی نیاز است تا یک فرد خبره ویژگی‌ها اصلی و مهم هر درخواست را استخراج و عددی‌بودن یا نبودن آن را تعیین کند. محول‌کردن این عملیات به سیستم کامپیوتری می‌تواند منجر به کشف بیشتری از داده‌های تقلب و الگوهای مورد استفاده در تقلب شود.

مراجع

[1] S. Roglaski, "Business intelligence: 360 insight: the intelligence challenge," *DM Review Magazine*, vol. 68, pp. 90-113, Jun. 2016.

[2] B. H. Pilon, J. J. Murillo-Fuentes, J. P. C. L. da Costa, R. T. de Sousa Junior, and A. M. R. Serrano, "Gaussian process for regression in business intelligence: a fraud detection application," in *Proc. of the 7th Int. Joint Conf. on Knowledge Discovery, Knowledge Engineering, and Knowledge Management*, vol.3, pp. 39-49, Nov. 2015.

[3] Q. Liu and M. Vasarhelyi, "Healthcare fraud detection: a survey and a clustering model incorporating geo-location information," in *Proc.* 29th *World Continuous Auditing and Reporting Symp.*, 10, pp., Brisbane, Australia, 21-22 Nov. 2013.

[4] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: an efficient data clustering method for very large databases," *ACM Sigmod Record*, vol. 25, no. 2, pp. 103-114, Jun. 1996.

[5] A. Y. Lam and V. O. Li, "Chemical reaction optimization: a tutorial," *Memetic Computing*, vol. 4, no. 1, pp. 3-17, Mar. 2012.

[6] R. M. Musal, "Two models to investigate medicare fraud within unsupervised databases," *Expert Systems with Applications*, vol. 37, no. 12, pp. 8628-8633, Dec. 2010.

[7] S. Thiprungsri and M. A. Vasarhelyi, "Cluster analysis for anomaly detection in accounting data: an audit approach," 2011.

[8] M. Tang, B. S. U. Mendis, D. W. Murray, Y. Hu, and A. Sutinen, "Unsupervised fraud detection in Medicare Australia," in *Proc. of the 9th Australasian Data Mining Conf., Australian Computer Society, AusDM'11,* vol. 121, pp. 103-110, Ballarat, Australia, 2011.

[9] R. Ghani and M. Kumar, "Interactive learning for efficiently detecting errors in insurance claims," in *Proc. of the 17th ACM SIGKDD Int Conf. on Knowledge Discovery and Data Mining, ACM*, pp. 325-333, San Diego, CA, USA, 21-24 Aug. 2011.

[10] T. Ekina, F. Leva, F. Ruggeri, and R. Soyer, "Application of bayesian methods in detection of healthcare fraud," *Chemical Engineering Trans.*, vol. 33, pp. 151-156, Sept. 2013.

[11] C. Ngufor and J. Wojtusiak, "Unsupervised labeling of data for supervised learning and its application to medical claims prediction," *Computer Science*, vol. 14, no. 2, p. 191-214, 2013.

[12] V. Rawte and G. Anuradha, "Fraud detection in health insurance using data mining techniques," in *Proc. IEEE Int. Conf. on Communication, Information & Computing Technology, ICCICT'15,* 5 pp., Mumbai, India, 15-17 Jan. 2015.

[13] M. E. Johnson and N. Nagarur, "Multi-stage methodology to detect health insurance claim fraud," *Health Care Management Science*, vol. 19, no. 3, pp. 249-260, Sept. 2016.

[14] H. Peng and M. You, "The health care fraud detection using the pharmacopoeia spectrum tree and neural network analytic contribution hierarchy process," in *Proc. IEEE Trustcom/BigDataSE/ISPA,* , pp. 2006-2011, Tianjin, China, 23-26 Aug. 2016.

[15] A. Gangopadhyay and S. Chen, "Health care fraud detection with community detection algorithms," in *Proc. IEEE Int. Conf. on*

Smart Computing, SMARTCOMP'16, 5 pp., St. Louis, MO, USA, 18-20 May 2016.

[16] S. G. Fashoto, *et al*., "Development of improved k-means clustering to partition health insurance claims," *Annals. Computer Science Series*, vol. 14, no. 1, pp. 51-58, 2016..

[17] H. Ahmadinejad, A. Norouzi, A. Ahmadi, and A. Yousefi, "Distance based model to detect healthcare insurance fraud within unsupervised database," *Indian J. of Science and Technology*, Indian J. of Science and Technology, vol. 9, no. 43, pp. 1-6, Nov. 2016.

[18] J. Wu, R. Zhang, X. Shang, and F. Chu, "Medical insurance fraud recognition based on improved outlier detection algorithm," in Proc. *2nd Int. Conf. on Artificial Intelligence and Engineering Applications, AIEA'17*, pp. 765-772, Guilin, China, 23-24 Sept. 2017.

[19] H. Cao and R. Zhang, "Using PCA to improve the detection of medical insurance fraud in SOFM neural networks," in *Proc. of the 3rd Int. Conf. on Management Engineering, Software Engineering and Service Sciences,* pp. 117-122, Wuhan, China, 12-14 Jan. 2019.

[20] T. Ekin, F. Ieva, F. Ruggeri, and R. Soyer, "Statistical medical fraud assessment: exposition to an emerging field," *International Statistical Review*, vol. 86, no. 3, pp. 379-402, May 2018.

[21] M. H. Soleymani, M. Yaseri, F. Farzadfar, A. Mohammadpour, F. Sharifi, and M. J. Kabir, "Detecting medical prescriptions suspected of fraud using an unsupervised data mining algorithm," *DARU J. of Pharmaceutical Sciences*, vol. 26, no. 2, pp. 209-214, Dec. 2018.

[22] D. S. Vijayarani and M. P. Jothi, "Hierarchical and partitioning clustering algorithms for detecting outliers in data streams," *International J. of Advanced Research in Computer and Communication Engineering*, vol. 3, no. 4, pp. 6205-6207, Apr. 2014.

[23] C. A. Ralanamahatana, J. Lin, D. Gunopulos, E. Keogh, M. Vlachos, and G. Das, "Mining Time Series Data," in *Data Mining and Knowledge Discovery Handbook*: Springer, pp. 1069-1103, 2005.

[۲۴] م. اسماعیلی، *داده‌کاوی و مفاهیم آن*، ناشر نیاز دانش، ۱۳۹۴ ۱۳۹۴.

[25] D. O. H. H. Services. Heart Attack Payment - Hospital [Online]. Available: https://catalog.data.gov/dataset/heart-attack-payment-hospital

[26] S. Firdaus and M. A. Uddin, "A survey on clustering algorithms and complexity analysis," *International J. of Computer Science Issues,* vol. 12, no. 2, pp. 62-85, Mar. 2015.

مجید عبدالرزاق‌نژاد در سال ۱۳۸۲ مدرک کارشناسی ریاضیات کاربردی خود را از دانشگاه بیرجند و در سال ۱۳۸۴ مدرک کارشناسی ارشد ریاضیات کاربردی خود را از دانشگاه سیستان و بلوچستان با کار برروی حوزه محاسبات نرم و شبکه‌های عصبی دریافت نمود. از سال ۱۳۸۴ الی ۱۳۸۶ نامبرده به عنوان مربی در دانشگاه صنایع و معادن شهرستان بیرجند و پس از آن تا سال ۱۳۸۸ در جهاد دانشگاهی مشهد بعنوان مدرس مدعو خدمت کرده و پس از آن به دوره دکترای علوم کامپیوتر گرایش محاسبات هوشمند در دانشگاه ملی مالزی وارد گردید. در سال ۱۳۹۲ موفق به اخذ درجه دکترا در علوم کامپیوتر از دانشگاه مذکور گردید. دکتر عبدالرزاق‌نژاد از سال ۱۳۹۲ در دانشکده فنی مهندسی قائن دانشگاه بیرجند به عضویت هیات علمی درآمد و پس از استقلال این دانشکده از دانشگاه بیرجند تحت عنوان دانشگاه بزرگمهر قائنات تا کنون به‌عنوان عضو هیات علمی دانشکده فنی و مهندسی این دانشگاه مشغول به فعالیت می‌باشد. او مسئولیتهایی همچون رئیس دانشکده علوم پایه، سرپرست معاونت دانشجویی- فرهنگی و مدیرگروه مهندسی کامپیوتر دانشگاه بزرگمهر قائنات را برعهده داشته است. زمینه‌های علمی مورد علاقه نام‌برده، انواع مسائل بهینه، مسائل زمانبندی، سیستم‌های نادقیق و منطق فازی، شبکه‌های عصبی، داده‌کاوی و کاربردهای آن، الگوریتم‌های فوق ابتکاری و ابرابتکاری می‌باشد.

مهدی خرد در سال ۱۳۹۱ مدرک کارشناسی مهندسی کامپیوتر خود را از دانشگاه بیرجند و در سال ۱۳۹۵ مدرک کارشناسی ارشد مهندسی فناوری اطلاعات خود را از دانشگاه بیرجند دریافت نمود. از سال ۱۳۸۹ تاکنون نام‌برده به عنوان دبیر در آموزش و پرورش به کار مشغول است و از سال ۱۳۹۷ به دوره دکترای مهندسی فناوری اطلاعات گرایش تجارت الکترونیک در دانشگاه قم وارد گردید و تاکنون مشغول به تحصیل است. زمینه‌های علمی مورد علاقه ایشان شامل موضوعاتی مانند محاسبات نرم، داده‌کاوی، شبکه ای عصبی، یادگیری عمیق و الگوریتم‌های فراابتکاری می‌باشد.