



**دانشگاه الزهرا - دانشکده فنی و مهندسی**

**پایان نامه جهت اخذ درجه کارشناسی ارشد**

**رشته مهندسی کامپیوتر- گرایش هوش مصنوعی**

**عنوان**

**کشف تقلب در سیستم‌های مراقبت سلامت با رویکرد تحلیل گراف**

**استاد راهنما**

**دکتر محمدرضا کیوان پور**

**دانشجو**

**روناک نمکی**

**بهار 1399**

****

کلیه دستاوردهای این تحقیق متعلق به دانشگاه الزهرا(س) است.

سپاس بی کران پروردگار یکتا را که هستی مان بخشید و به طریق علم و دانش رهنمونمان شد و به همنشینی رهروان علم و دانش مفتخرمان نمود و خوشه چینی از علم و معرفت را روزیمان ساخت. جان ما را صفای خود ده و دل ما را هوای خود ده، و چشم ما را ضیای خود ده، و ما را از فضل و کرم خود آن ده که آن به.خداوندا به ما توفیق تلاش در شکست، صبر در نومیدی، رفتن بی همراه، جهاد بی سلاح، کار بی پاداش، فداکاری در سکوت، دین بی دنیا، مذهب بی عوام، عظمت بی نام، خدمت بی نان، ایمان بی ریا، خوبی بی نمود، گستاخی بی خامی، مناعت بی غرور، عشق بی هوس، تنهایی در انبوه جمعیت و دوست داشتن بی آنکه دوستت بدارند، را عنایت فرما.

یارب دل ما را تو به رحمت جان ده درد         همه را به صابری درمان ده

این بنده چه داند که چه می باید جست     داننده تویی هر آنچه دانی آن ده

چکیده

واژگان کلیدی:

فهرست مطالب

[فصل اول 1](#_Toc44395255)

[مقدمه 1](#_Toc44395256)

[1 – 1 مقدمه 2](#_Toc44395257)

[1 – 2 طرح مساله 3](#_Toc44395258)

[1 - 3اهداف و نوآوریها 4](#_Toc44395259)

[1 – 3 –1 سوالات تحقیق 4](#_Toc44395260)

[1 – 3 – 2 محدودیت های تحقیق 4](#_Toc44395261)

[1 – 4 بازیگران نظام سلامت 4](#_Toc44395262)

[1 –5 ساختار پایان نامه 4](#_Toc44395263)

[1 –6 جمع بندی 5](#_Toc44395264)

[فصل دوم 6](#_Toc44395265)

[پیشینه پژوهش 6](#_Toc44395266)

[2 – 1 تعریف تقلب 7](#_Toc44395267)

[1 – 2 – 1 نقش داده‌کاوی در تشخیص تقلب 7](#_Toc44395268)

[1 – 2 – 2 انواع تقلب 7](#_Toc44395269)

[2 – 2 دسته بندی چالش های کشف تقلب 9](#_Toc44395270)

[2 – 2 – 1 چالش های کشف تقلب از منظر سازمان بیمه 9](#_Toc44395271)

[2 – 2 - 2 چالشهای کشف تقلب از منظر داده 9](#_Toc44395272)

[2 – 2 – 3 چالشهای کشف تقلب از منظر مقالات 9](#_Toc44395273)

[2 – 2 – 3 – 1 مفهوم رانش 10](#_Toc44395275)

[2 – 2 – 3 – 2 مفهوم توزیع اریب کلاس‌ها 11](#_Toc44395276)

[2 – 2 – 3 – 3 مفهوم کاهش داده ها 12](#_Toc44395277)

[2 – 2 – 3- 4 مفهوم پشتیبانی تشخیص انلاین 14](#_Toc44395278)

[2 – 2 – 3 – 5 ادغامدرنظاممراقبتهای بهداشتی 14](#_Toc44395279)

[2 – 2 – 3 – 6 خطای پیش بینی و اثر "Black Swan" 14](#_Toc44395280)

[2 – 2 – 3 – 7 ازدستدادناطلاعاتدرپیش پردازش 15](#_Toc44395281)

[2 – 2- 3 – 8 خودکارسازی فرآیند داده‌کاوی برای کاربران غیرمتخصص 15](#_Toc44395282)

[2 – 2- 3 – 9 ماهیت بین رشته ای تحقیق و دانش متخصص حوزه 15](#_Toc44395283)

[2 – 3 انواع ناهنجاری در تشخیص تقلب 15](#_Toc44395284)

[2 – 3 – 1 تشخیص ناهنجاری گراف ایستا 15](#_Toc44395285)

[2 – 3 - 1 – 1 تشخیص ناهنجاری مبتنی بر ساختار 15](#_Toc44395286)

[2 – 3 – 1 – 2 تشخیص ناهنجاری مبتنی بر اجتماع 16](#_Toc44395287)

[2 – 3 – 2 تشخیص ناهنجاری گراف پویا 17](#_Toc44395288)

[2 – 3 – 2 – 1 تشخیص ناهنجاری مبتنی بر فاصله 17](#_Toc44395289)

[2 – 3- 2 – 2 تشخیص ناهنجاری مبتنی بر فشرده سازی 17](#_Toc44395290)

[2 – 3 – 2 – 3 تشخیص ناهنجاری مبتنی بر تجزیه 17](#_Toc44395291)

[2 – 3 – 2 – 4 تشخیص ناهنجاری مبتنی بر خوشه یا اجتماع 17](#_Toc44395292)

[2 – 3 – 2- 5 تشخیص ناهنجاری مبتنی بر مدل‌های احتمالاتی 17](#_Toc44395293)

[2 – 3 – 2 – 6 تشخیص ناهنجاری مبتنی بر پنجره 18](#_Toc44395294)

[بدون تیتر 18](#_Toc44395295)

[بدون تیتر اصلی: برخی معیارهای تحلیل شبکه‌ی پزشکان 19](#_Toc44395296)

[بدون تیتر اصلی: کلان داده‌هادر تشخیص تقلب 20](#_Toc44395297)

[بدون تیتر اصلی: جدول 20](#_Toc44395298)

[بدون تیتر اصلی: رویکردهای کلی روش‌های کشف تقلب 21](#_Toc44395299)

[بدون تیتر اصلی:روش‌های با ناظر 21](#_Toc44395300)

[بدون تیتر اصلی:روش‌های بدون ناظر 22](#_Toc44395301)

[بدون تیتر 25](#_Toc44395302)

[فصل سوم 31](#_Toc44395303)

[روش پیشنهادی 31](#_Toc44395304)

[1-3 تشکیل گراف / ترکیب دیتاست ها 33](#_Toc44395305)

[3– 2 تولید بردار رفتار برای ارایه دهندگان 33](#_Toc44395306)

[3 – 3 معرفی معیار های مشابهت/ ویژگی های مبتنی بر گراف 33](#_Toc44395307)

[3 – 4 محاسبه مشابهت بردارهای رفتار دو ارایه دهنده 33](#_Toc44395308)

[فصل چهارم 38](#_Toc44395309)

[پیاده سازی و ارزیابی 38](#_Toc44395310)

[1-4 پیاده سازی سیستم در قسمت تشکیل گراف / بردار رفتار / ترکیب دیتاست ها 39](#_Toc44395311)

[Algorithm: 39](#_Toc44395312)

[4 - 2 پیاده سازی سیستم در قسمت محاسبه مشابهت بردار های رفتار 39](#_Toc44395313)

[Algorithm: 39](#_Toc44395314)

[4 -4آزمون و تحلیل نتایج 40](#_Toc44395315)

[4 – 4 – 1 معیارهای ارزیابی سیستمهای تشخیص تقلب 40](#_Toc44395316)

[4 – 4- 2 روش آزمون 40](#_Toc44395317)

[4 – 4 – 3 ارزیابی روش پیشنهادی 40](#_Toc44395318)

[*4 – 4- 4-1 ارزیابی روش پیشنهادی با استفاده از []* 40](#_Toc44395319)

[4 – 5 جمعبندی 43](#_Toc44395320)

[فصل پنجم 46](#_Toc44395321)

[نتیجه گیری و توسعه های آتی 46](#_Toc44395322)

[5 – 1نتیجه گیری 47](#_Toc44395323)

[5 – 2 توسعههای آتی 47](#_Toc44395324)

[مراجع 48](#_Toc44395325)

فهرست اشکال

فهرست جداول

# فصل اول

# مقدمه

## 1 – 1 مقدمه

هزینه بهداشت و درمان با توجه به جمعیت، اقتصاد، جامعه، و تغییرات قانون به سرعت در حال افزایش است.این افزایش در هزینه‌های بهداشت و درمان بر دولت و سیستم‌های بیمه سلامت خصوصی تأثیر می‌گذارد. رفتارهای متقلبانه ی ارايه دهندگان بهداشت و درمان و بیماران با تحمیل هزینه‌های غیرضروری به مشکلی جدی برای سیستم‌های بیمه تبدیل شده است.بنابراین شرکت های بیمه روش‌هایی را برای تشخیص تقلب ایجاد می‌کنند که عمدتا برگرفته از تجارب خبرگان بوده و کمتر به روش‌های مبتنی بر تحلیل داده متکی است.صنعت بیمه و در راس آن بیمه سلامت با هزینه ای بالغ بر ۵۰ هزار میلیارد تومان یکی از کلیدی‌ترین هزینه‌های تحت مدیریت و نظارت دولت است که با تخمینی معادل ۳ تا ۱۰ درصد یعنی ۱۵ تا ۵۰ هزار میلیارد ریال تقلب مواجه است.

سازمان‌های بیمه‌گر پس از دریافت اسناد هزینه از ارایه‌دهندگان خدمات سلامت اعم از مراکز درمانی، پزشکان، داروخانه‌ها،ازمایشگاه‌ها فرایند بررسی هزینه‌ها و تطبیق آن‌ها با معیارها و جداول هزینه‌شده توسط آن را که به آن رسیدگی به اسناد می‌گویند را آغاز می‌نمایند. این رسیدگی هم‌اکنون به‌صورت دستی و توسط افراد خبره صورت می‌گیرد. استفاده از روش‌های تحلیل داده‌های بزرگ نظیر داده کاوی به ذی‌نفعان کمک می‌کند تا بتوانند ضمن تعمیم و بهره‌برداری از الگوهای شناخته شده جهت به‌کارگیری الگوریتم‌های همراه با ناظر نسبت به کشف الگوهای ناشناخته از طریق به‌کارگیری الگوریتم‌ها و مدل‌های نظارت نشده بپردازند.

سامانه‌های مراقبت سلامت در سراسر جهان مجموعه ای از افراد، سازمان‌ها و منابع هستند که برای رفع نیازهای درمانی جمعیت هدف تشکیل می‌شوند و درحال تغییر و توسعه هستند و اغلب از سه بخش به نام‌های ارایه دهندگان خدمات، بیمار و شرکت‌های بیمه تشکیل شده‌اند [1].

مراقبت سلامت در آمریکا از سال ۱۹۸۰ یک هزینه‌ی بزرگ ایجاد نموده است .بر اساس گزارشی از GAO[[1]](#footnote-2)به کنگره در سال ۲۰۰۴، هزینه‌های سالانه مراقبت سلامت به دو بیلیون دلار نزدیک شد که ۱۵/۳٪ تولید ناخالص داخلی برآورد شد. اندازه بخش مراقبت سلامت و حجم زیاد پول شامل آن،آنرا برای اهداف تقلب جذاب می‌سازد.تقلب مراقبت سلامت بر اساس تعریفNHCAA[[2]](#footnote-3)یک فریب عمدی یا ارایه اطلاعات نادرست است که توسط یک شخص یا یک موجودیت با علم به اینکه این فریب می‌تواند منجربه مقداری سود غیرمجاز برای آن فرد یا موجودیت شود انجام می‌شود. NHCAAمحتاطانه تخمین می‌زند که حداقل ۳٪ یا بیش از ۶۰ بیلیون دلار از هزینه‌های مراقبت سلامت سالانه‌ی آمریکا به دلیل تقلب آشکار از بین رفته است.دولت و سازمان های اجرای قانون این خسارت را ۱۰٪ یا ۱۷۰ بیلیون دلار تخمین می‌زنند.فقط خسارت مالی نگرانی عمده نیست بلکه تقلب به شدت مانع از ارایه مراقبت با کیفیت و امن سیستم مراقبت سلامت آمریکا از بیماران مشروع می‌شود.بنابراین تشخیص تقلب مؤثر برای بهبود کیفیت و کاهش هزینه‌ ی خدمات مراقبت بهداشت مهم است.در سال‌های اخیر سیستم‌ها برای پردازش مطالبات الکترونیکی بطور فزاینده ای پیاده‌سازی شده‌اند تا بازرسی و بازبینی داده‌ ی مطالبات خودکار انجام شود. این سیستم‌ها برای شناسایی مناطقی که نیاز به توجه خاص دارند مانند داده ی ورودی اشتباه یا ناقص،تکرار مطالبات و خدمات پزشکی که تحت پوشش نیستند، طراحی شده‌اند.اگرچه این سیستم‌ها ممکن است برای تشخیص انواع خاصی از تقلب بکار روند، قابلیت تشخیص تقلب آن‌ها معمولاً محدود است از آنجایی که تشخیص عمدتا برپایه ی قوانین ساده ی از پیش تعیین شده است که توسط متخصصان حوزه مشخص می‌شوند[2].

## 1 – 2 طرح مساله

## 1 – 3اهداف و نوآوری­ها

### 1 – 3 –1 سوالات تحقیق

* چگونه می‌توان تکنیک های داده‌کاوی و سایر تکنیک‌های استفاده شده در این حوزه را جهت مقابله با پدیده تقلب در سیستم‌های مراقبت سلامت به‌کار گرفت.
* کدام یک از الگوریتم‌های مورد بررسی دقت بیشتری در پیش‌بینی تقلب دارند.
* با توجه به میزان دقت و صحت الگوریتم‌های بررسی شده برای کشف تقلب، در چه مرحله‌ای میتوان از این الگوریتم ها استفاده نمود.

### 1 – 3 – 2 محدودیت های تحقیق

* اکثر فرم‌ها و پرونده ها بصورت غیرسیستمی و دستی ذخیره شده‌اند و تعداد کمی از آن‌ها بصورت الکترونیکی ذخیره و نگه داری می‌شوند و جستجوی دستی در این اسناد بسیار زمان‌بر است.
* شرکت های بیمه اطلاعات و پرونده های مشتریان خود را به آسانی در اختیار افراد خارج از سازمان قرار نمی‌دهند و اخذ مجوز برای دسترسی به این اطلاعات فرآیندی زمانبر و مشکل است.
* دسترسی به تعداد اندکی پرونده که وقوع تقلب در آن‌ها محرز شده است بسیار مشکل است زیرا این پرونده ها اغلب محرمانه هستند.
* کسب دانش مورد نیاز از افراد خبره و کارشناسان کشف تقلبات بیمه‌ای مشکل است زیرا اکثر آن‌ها مدیران بخش بیمه هستند و به ندرت وقت آزادی برای قبول و انجام مصاحبه دارند[3] .

## 1 – 4 بازیگران[[3]](#footnote-4) نظام سلامت

ارتباط میان بازیگران مختلف سیستم سلامت به واضح‌ترین شکل ممکن در شکل ۱-۱ ارایه شده است که موارد سواستفاده بین دو یا تعداد بیشتری بازیگران را شامل می‌شود. برای مثال با دیدن بیمارستان و تعیین اینکه چه نوع سواستفاده‌هایی می‌تواند در ارتباط میان آن و سایر ذی‌نفعان، بیماران و سایر بیمارستان‌ها صورت گیرد، یک تحلیل‌ می‌تواند انجام شود [4] .

## 1 –5 ساختار پایان نامه

## 1 –6 جمع بندی

# فصل دوم

# پیشینه پژوهش

## 2 – 1 تعریف تقلب

تقلبدربیمهیکمسالهمهمو پرهزینهبرایبیمهگذارانو شرکتهایبیمهدرتمامبخشهایصنعتبیمهاست.درسال­هایاخیرتشخیصتقلبتوجهاتو نگرانیهایبسیاریرابهخودجلبکرده­است.دیکشنریآکسفوردتقلبرابعنوان"ارتکابفریبعمدییااشتباهیکهمنجربهسودشخصییامالیمی‌شود"تعریفمی‌کند. تقلببهشکل­هایبسیارگوناگونیرخمی‌دهدو همانطورکهتکنولوژیهایجدیدو سیستمهایاقتصادیو اجتماعیجدیدفرصتهایجدیدیرابرایفعالیتهایتقلب‌آمیزفراهممی‌کنند،تغییرمی‌کند. تعیینخسارتکلیمشاغلبهدلیلفعالیت­هایتقلب‌آمیزدشواراست. تقلبرااین‌گونهتوصیفمی­کندکهمنجربهسواستفادهازمنافعیکسازمانمی‌شودو نهلزومامنجربهیکپیامدقانونیمستقیم![5]

اگرچهکهیکتعریفپذیرفته­شدهجهانیازتقلبمالیوجودندارد،[6] آن­رابعنوانیکعملعمدیکهدرتضادباقوانینو قاعدههاو سیاستو باهدفکسبمنافعمالیغیرمجازاستتعریفمی­کند.

### 1 – 2 – 1 نقش داده‌کاوی در تشخیص تقلب

ازنظراقتصادیتقلبدربیمهبهیکمسالهجدیدرحالافزایشتبدیلشده­است.براساسگزارشخبرBBCدرسال۲۰۰۷مطالبههایبیمهتقلبی1.6بیلیونپونددرسالبرایبیمه­گذارانانگلستانهزینهدارد.خساراتکلیناشیازتقلبتوسطبیمهغیرقابلمحاسبهاست.تشخیصتقلببیمهبرایجلوگیریازنتایجمخربتقلببیمهمهماست.تشخیص تقلب بیمهشاملبررسیمطالبههایجعلیازمطالبههایاصلیاست.بهاینترتیبافشایرفتاریافعالیتجعلیتصمیمگیرندگانرابرایتوسعهاستراتژی­هایمناسببرایکاهشاثرتقلبقادرمی‌سازد.داده­کاوییکنقشمهمدرتشخیص تقلب بیمهدارد،همانطورکهاغلببرایاستخراجو پرده­برداریازحقایقپنهانپشتمقدارزیادیدادهبه‌کارمی‌رود.داده­کاویدرباره­ی پیداکردنروشهاییاستکهقابلاعتمادند،قبلاناشناختهبودندو ازداده­هاقابلاجراهستند.ایندادهبایددردسترسو مرتبطو کافیو تمیزباشد.هم‌چنینمسالهداده­کاویبایدبه‌خوبیقابلتعریفباشدو باابزارهایپرس­وجوقابلحلنباشدو بایکمدلپردازشداده‌کاویتعلیمداده­شود.[7]داده‌کاویرابهاینصورتتعریفمی‌کنند:پروسهشناساییالگوهایموردعلاقهدرپایگاهدادهبه­طوریکهبعدابتواننددرتصمیم­گیریاستفادهشوند.[8] داده‌کاویراپروسهایتعربفمی‌کندکهازریاضیاتاماریو هوشمصنوعیو تکنیکهاییادگیریماشینبرایاستخراجو شناساییاطلاعاتمفیداستفادهمی‌کندو بطورپیوستهازیکپایگاهدادهبزرگتجربهکسبمی‌کند.[9]بیانمی‌کندکههدفداده­کاویبدستاوردناطلاعاتمفیدو غیراشکارازدادههایذخیرهشدهدرمخازنبزرگاست.[10] مشخصمی­کندکهیکیازمزایایمهمداده‌کاویایناستکهمی­تواندبرایتوسعهیککلاسجدیدازمدل‌هابرایشناساییجملاتقبلازاینکهتوسطمتخصصان شانتشخیصدادهشونداستفادهشود. [11] اشارهداردکهتشخیصتقلبیکیازبهترینکاربرد­هایداده­کاویدرصنعتو دولتاست. تکنیکهایدادهکاویگوناگونیدرتشخیص تقلب بیمهبه‌کارگرفتهشده‌اندمانندشبکههایعصبی،مدلهایرگرسیونمنطقی،روشهاینایوبیزینو درختتصمیم.

### 1 – 2 – 2 انواع تقلب

بر اساس اینکه چه کسی مرتکب تقلب میشود، رفتارهای تقلب آمیز به صورتی که در ادامه آمده اند دسته بندی میشوند[12]:

1. **تقلب ارایه دهنده ی خدمات[[4]](#footnote-5)**
2. **تقلب مشترکان بیمه**
3. **تقلب حاملان بیمه**
4. **تقلب دسیسه ای [[5]](#footnote-6)**

## 2 – 2 دسته بندی چالش های کشف تقلب

### 2 – 2 – 1 چالش های کشف تقلب از منظر سازمان بیمه

شرکت های بیمه می توانند با آگاهی از انواع تقلبات و فرایندهایی که احتمال بروز تقلب در آن‌ها وجود دارد سیستم هشدار دهنده و پیشگیرانه ای را طراحی کنند و با آگاهی از میزان آسیب‌پذیری خود استراتژی های موثرتری را به‌کار گیرند، اما برای تحقق این امور شرکت های بیمه با محدودیت‌ها و پیچیدگی های زیادی مواجه اند [15]:

* پنهان بودن ماهیت تقلب
* پویایی و حساسیت به تغییر در تقلب( هنگام شناسایی یک سبک تقلب، کلاهبرداری با سبک دیگری در حال شکل‌گیری است)
* عدم توافق اجماع در برخی موارد بر آنچه که واقعاً به منزله ی تقلب در بیمه استاندارد.
* نگاه جامعه به شرکت های بیمه ای و ارایه دهندگان خدمات درمانی بعنوان نهادهای حمایتی
* ضعف سیستم‌های کنترلی مبتنی بر فناوری اطلاعات

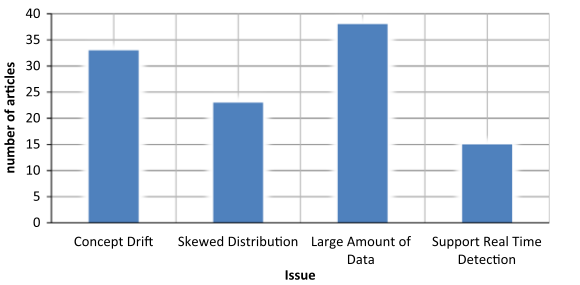
### 2 – 2 - 2 چالش های کشف تقلب از منظر داده

از چالش های موجود در کشف تقلب عدم وجود داده‌های با برچسب سالم و وجود تعداد بسیار کم داده‌های تقلبی جهت یادگیریاست.

### 2 – 2 – 3 چالش های کشف تقلب از منظر مقالات

نمودار 2 - 1 توزیع مطالعات سیستم‌های تشخیص تقلب را بر اساس چالش ها بر اساس تعداد مقالات منتشر شده در سال‌های ۱۹۹۴ تا ۲۰۱۴ نشان میدهد و به رایج ترین انواع تقلب های الکترونیکی مانند کارت های اعتباری، بیمه مراقبت سلامت، مخابرات، بیمه اتومبیل متمرکز است[16].

### 

  
نمودار 2 - 1توزیع مقالات FDSبراساس مسائل و چالش های بین سال های 1994تا 2014[16]

2 – 2 – 3 – 1 مفهوم رانش[[6]](#footnote-7)

تعاریف مختلفی برای مفهوم مساله رانش وجود دارد.در داده کاوی رانش به پدیده‌ای که مدل پایه ی آن در طول زمان درحال تغییر است اشاره دارد.کار سیستم‌های تشخیص تقلب در محیط پویا که رفتار کاربران قانونی/غیرقانونی بطور پیوسته در حال تغییر است مفهوم پدیده رانش گفته می شود.[17]

برای مثال در حوزه کارت اعتباری رفتار صاحب کارت ممکن است به دلیل برخی عوامل خارجی تغییر کند. برای مثال مقدار تراکنش و تکرار به عادات خرج یک فرد وابستگی نزدیکی داشته باشد که در‌واقع تحت تأثیر شیوه زندگی، منبع درآمد فرد و … است که در طول زمان می‌تواند تغییر کند.[18]

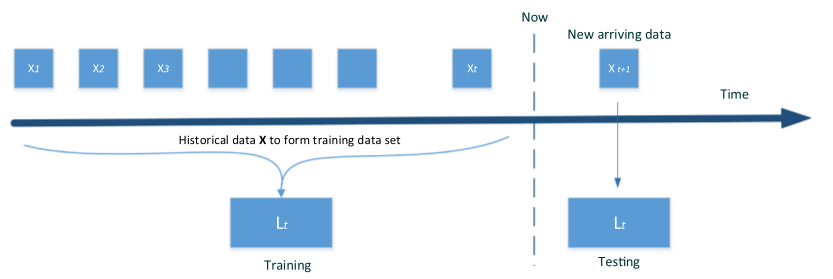
به‌علاوه مفهوم رانش سابقاً به یک سناریوی یادگیری با ناظر زمانی که رابطه بین داده ورودی و متغییر هدف در طول زمان تغییر می‌کند ارجاع داده می‌شد.اگرچه در یادگیری با ناظر هدف پیش‌بینی یک متغییر هدفyبا استفاده ازمجموعه ای از ویژگی‌های ورودی xاست.در نمونه یادگیری که برای ساخت مدل استفاده می‌شود هردوی xو yدر زمان پیش گویی ناشناخته اندو رابطه بین داده ورودی و متغییر هدف ممکن است تغییر کند.[19]

مفهوم رانش یک نگرانی بزرگ است، مخصوصاً در یادگیری آنلاین که مدل تشخیص فوراً به روز می‌شود، اما براساس داده‌های خروجی. بنابراین وقتی داده‌های جدید می‌رسند، مدل ممکن است گمراه شود و اخطار اشتباه دهد. توجهات در تحقیقات به مقابله با رفتار غیرایستا و بطور پویا به روز رسانی مدل تشخیص تقلب، اختصاص یافته است و در نتیجه استفاده از الگوریتم های یادگیری تطبیقی[[7]](#footnote-8) برای مقابله با مفهوم رانش لازم است.الگوریتم های یادگیری تطبیقی می‌توانند بعنوان الگوریتم های یادگیری افزایشی توسعه یافته دیده شوند که قادر به‌ به روزرسانی مدل تشخیص برای داده ی در جریان تکامل، در طول زمان می‌باشند.[19],[20]

[21] مفهوم یادگیری افزایشی با رانش را این‌گونه بیان می‌کند:فرآیند یادگیری افزایشی در هرزمانtکه داده‌های قبلی در دسترس هستند، یک نمونه‌ی هدف xt+1می‌رسد وظیفه‌اش این است که برچسب yt+1را پیش‌بینی می‌کند. بدین منظور یادگیرندهLtدر فاز یادگیری با استفاده از همه یا انتخاب از داده‌های قبلی برچسب دار ساخته می‌شود.

Xhistorical=(x1 , x2 , …, xt)

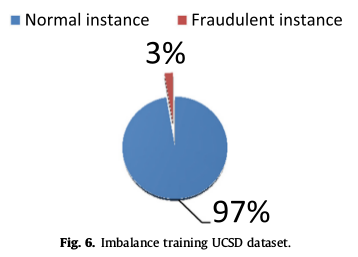
که این را در شکل 1-3می‌بینیم:



#### 2 – 2 – 3 – 2 مفهوم توزیع اریب کلاس‌ها[[8]](#footnote-9)

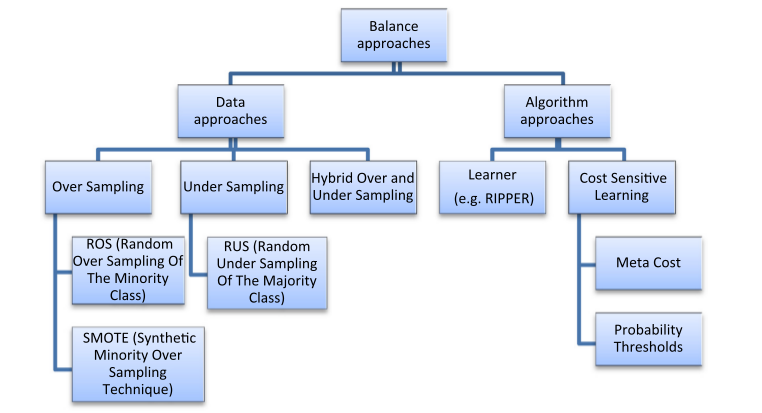
مساله‌ی داده‌های اریب یکی از مهم‌ترین مسایلی است که در سیستم‌های تشخیص تقلب با آن مواجه ایم.نامتوازن بودن داده‌ها، تأثیری جدی روی کارایی طبقه بندهایی که قرار است توسط اکثریت کلاس سراسر پوشانیده شوند و اقلیت کلاس نادیده گرفته شوند دارد[22].

داده ی رقابت‌های داده کاویUCSD[[9]](#footnote-10)برای تشریح مساله نامتوازن بودن استفاده شده و یک مجموعه داده از دنیای واقعی است که برای تشخیص تراکنش های تجارت الکترونیک غیرعادی استفاده شده‌است. مجموعه داده‌ی آموزش شامل ۱۰۰۰۰۰ تراکنش از۷۳۷۲۹ مشتری در طی ۹۸ روز است و مجموعه داده تست شامل ۵۰۰۰۰ تراکنش است. داده ی آموزشی به شدت نامتوازن است که شامل ۹۷۳۴۶ تراکنش طبیعی و فقط ۲۶۵۴ تراکنش تقلبی است.همانطور که در شکل 2-3می‌بینیم درصد تراکنش ها درحدود ۹۷-۳ ٪ قانونی و جعلی می‌باشد. بنابراین یک مکانیزم متوازن سازی نیاز‌است تا این داده‌ها را با نرخ ۱:۱ میان طبیعی و جعلی متوازن کند.روش های متوازن سازی داده‌ها در دو سطح می‌تواند طبقه بندی شود:سظح داده و سطح الگوریتم، که تکنیک‌های آن در شکل3-3نشان داده‌شده‌است.



شکل 2-3مجموعه داده آموزشی ناهمگون UCSD [16]

#### 2 – 2 – 3 – 3 مفهوم کاهش داده ها



شکل 3-3روش‌های رسیدگی به داده‌های ناهمگون [116]

مقیاس بزرگ و ابعاد زیاد مجموعه داده‌ی تقلب و حضور تعداد زیاد ویژگی ها/ورودی ها/متغییرها فرایند داده کاوی و تشخیص را بسیار دشوار و پیچیده می‌سازد.[23]. بعلاوه این شرایط فرایند تشخیص را نیز کند میکندبنابریان سیستم‌های تشخیص تقلب موجود از روش‌های کاهش داده برای کاهش حجم مجموعه داده‌ها استفاده می‌کند[24]. بعلاوه داده ی کم‌حجم مدل و در نتیجه زمان محاسبه را کاهش می‌دهد[25]. روش‌های کاهش داده شامل کاهش ابعاد(Dimension Reduction) و کاهش عددی(Numerosity Reduction) است[26].

کاهش ابعاد شامل استراتژی های بسیاری است به نام های فشرده‌سازی داده(data compression)، انتخاب داده(feature selection)،ساخت ویژگی(feature construction).مرسوم ترین و پرتکرارترین استراتژی های استفاده‌شده در سیستم‌های تشخیص تقلب هستند. استراتژی فشرده‌سازی داده، از طریق استفاده از تکنیک‌ های فشرده‌سازی مانند [27] ,[28] نمایش داده‌ی اصلی را فشرده می‌کند.‌‌‌‌‌‌‌در این میان انتخاب ویژگی یک استراتژی دیگر کاهش ابعاد است.مهم ترین و مرتبط ترین ویژگی‌های انتخاب می‌شوند تا در ساخت مدل استفاده شوند.انتخاب ویژگی توسط‌[29] نام گذاری شد.

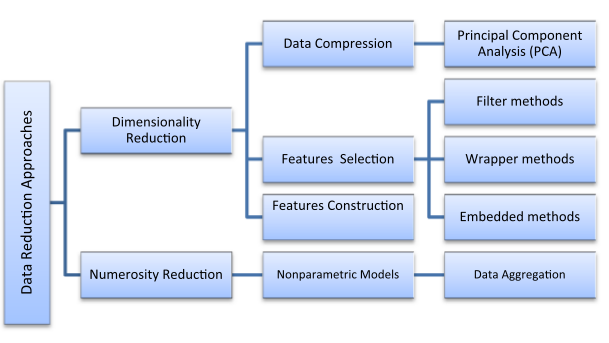
سه روش انتخاب ویژگی که در سیستم‌های تشخیص تقلب استفاده می‌شوند:

روش‌های تصفیه [[10]](#footnote-11)- روش‌های بسته‌بندی[[11]](#footnote-12)- روش‌های جاسازی [[12]](#footnote-13)

* روش تصفیه بعنوان الگوریتم پیش‌پردازش برای رتبه بندی ویژگی‌ها عمل می‌کند که ویژگی‌های با رتبه بالا انتخاب و به یک پیش‌بینی کننده اعمال می‌شوند.
* روش جاسازی شامل انتخاب متغییر بعنوان بخشی از فرایند یادگیری بدون تقسیم داده به مجموعه‌ی آموزش و تست می‌باشد [30] .ساخت ویژگی در جایی است که یک مجموعه کوچک از ویژگی‌های مفیدتر از مجموعه ی اصلی مشتق می‌شود.

در کاهش عددی نیز داده‌ها با نمایش کوچکتری جایگزین می‌شوند.مانند استفاده از تجمیع داده ها[[13]](#footnote-14)[31] , [32].

روش‌های کاهش داده شامل کاهش ابعاد و کاهش ارقام در شکل۴-۳ آورده شده‌اند:



شکل 4-3استراتژی‌های کاهش داده [117]

#### 2 – 2 – 3- 4 مفهوم پشتیبانی تشخیص انلاین

سیستم‌های تشخیص تقلب در دو حالت متفاوت کار می‌کنند که تشخیص آنلاین و آفلاین است که بر اساس انواع تقلب متفاوت است و هریک کاربرد خاص خود را دارند[16].

#### 2 – 2 – 3 – 5 ادغام در نظام مراقبت های بهداشتی

تعداد بسیار کمی از مقالات مورد بررسی تلاش برای ادغام فرایند داده کاوی را در چارچوب تصمیم گیری واقعی انجام دادند. تأثیر کشف دانش توسط داده کاوی بر میزان کاری و زمان کار حرفه ای مراقبت های بهداشتی مشخص نیست. مطالعات آینده باید یکپارچگی سیستم توسعه یافته را در نظر بگیرند و تاثیر آن بر محیط کار را بررسی کنند.[33]

#### 2 – 2 – 3 – 6 خطای پیش بینی و اثر "Black Swan"

در مراقبت های بهداشتی، پیش بینی بهتر از پیش بینی اشتباه بهتر است [34]. کمی کمتر از نیمی از ادبیات که در تجزیه و تحلیل شناسایی شده است، به پیش بینی اختصاص یافته، اما هیچ یک از مقالات درباره نتیجه خطای پیش بینی بحث نشده است. دقت پيش بيني بالا براي سرطان يا هر بيماري ديگر، برنامه صحيحي را براي تصميم گيري فراهم نمي کند.

علاوه بر این، مدل پیش بینی ممکن است در پیش بینی رویدادهای عادی بهتر از موارد نادر باشد.محققان باید مدل های پیشرفته ای را برای رسیدگی به غیر قابل پیش‌بینی”The Black Swan"توسعه دهند[35].

یک مطالعه [36] یک مسئله مشابه را در توصیه های مبتنی بر شواهد برای تجویز پزشکان مطرح کرد. نگرانی آنها این بود که چه مقدار شواهد باید برای تهیه یک توصیه کافی باشد.

بسیاری از مطالعات در این بررسی این مسائل برجسته را رفع نمی کند. پژوهش های آینده باید به چالش های پیاده سازی مدل های پیش بینی کننده بپردازد، به ویژه اینکه چگونه فرآیند تصمیم گیری باید در صورت اشتباهات و حوادث غیر قابل پیش بینی سازگار شود.

#### 2 – 2 – 3 – 7 از دست دادن اطلاعات در پیش پردازش

پیش پردازش داده ها، از جمله دستکاری داده های از دست رفته، پرهزینه‌ترین و مهم‌ترین بخش داده کاوی است. شایع ترین روش مورد استفاده در مقالات مورد بررسی حذف یا حذف داده های از دست رفته است. در یک مطالعه، حدود 46.5٪ از داده ها و 363از 410ویژگی به دلیل مقادیر گم شده حذف شدند[37]. در یکی دیگر، محققان [38] تنها قادر به استفاده از 2064از 4948مشاهدات (42٪)بودند. با حذف مقادیر از دست رفته و داده‌های پرت، ما مقدار قابل توجهی از اطلاعات را از دست می دهیم. پژوهش های آینده باید بر روی یافتن یک روش بهتر تخمین مقادیر از دست‌رفتهنسبت به حذف تمرکز نمایند. علاوه بر این، تکنیک های جمع آوری داده ها باید به منظور جلوگیری از این موضوع توسعه یابند یا اصلاح شوند.

#### 2 – 2- 3 – 8 خودکارسازی فرآیند داده‌کاوی برای کاربران غیرمتخصص

کاربران نهایی داده کاوی در مراقبت های بهداشتی، پزشکان، پرستاران و متخصصین مراقبت های بهداشتی هستند که آموزش های محدودی در زمینه تحلیلی دارند. یک راه حل برای این مشکل این است که یک سیستم خودکار (یعنی بدون نظارت انسان) برای کاربران نهایی ایجاد شود [39]. یک ساختار خودکار مبتنی بر ابر برای جلوگیری ازخطاهای پزشکی نیز می تواند توسعه یابد [40]؛ اما این کار چالش برانگیز خواهد بود زیرا در آن زمینه های کاربردی مختلف وجود دارد و یک الگوریتم دقت مشابهی برای تمام برنامه های کاربردی ندارد [39].

#### 2 – 2- 3 – 9 ماهیت بین رشته ای تحقیق و دانش متخصص حوزه

تجزیه و تحلیل بهداشت و درمان یک زمینه تحقیقاتی بین رشته ای است [39]. به عنوان یک روش تجزیه و تحلیل، داده کاوی باید ازترکیبی از نظر کارشناس از حوزه های خاص مراقبت های بهداشتی و مشکل مشخص(به عنوان مثال، انکولوژی برای تحقیقات سرطان و متخصص قلب برای CVD) استفاده کند[41]. تقریبا 32٪ از مقالات در تجزیه و تحلیل از نظر متخصص به هیچ شکلی شکل استفاده نمی شود. پژوهش های آینده باید شامل اعضای از رشته های مختلف از جمله مراقبت های بهداشتی باشد[33].

## 2 – 3 انواع ناهنجاری در تشخیص تقلب

### 2 – 3 – 1 تشخیص ناهنجاری گراف ایستا

#### 2 – 3 - 1 – 1 تشخیص ناهنجاری مبتنی بر ساختار

دو نوع روش تشخیص ناهنجاری مبتنی بر ساختار به نام‌های ناهنجاری در گراف‌های ساده‌ی ایستا[[14]](#footnote-15) و ناهنجاری در گراف ویژگی ایستا[[15]](#footnote-16)وجود دارد که بصورت زیر تشریح می‌شوند:

در طرح تشخیص ناهنجاری در گراف ساده‌ی ایستا، ویژگی‌های مرکزی گراف‌های مختلف مانند درجه گره، مرکزیت egonetو … استخراج می‌شوند و یک فضای ویژگی با بقیه ویژگی‌هایی که از منابع اطلاعاتی اضافی برای تشخیص تقلب استخراج شده‌اند ساخته می‌شود. در [42]یک روش تشخیص ناهنجاری ارایه کرده‌اند که از شاخص‌های گراف برای شناسایی کاربران با روابط غیرعادی نسبت به سایر کاربران در شبکه اجتماعی آنلاین استفاده می‌کنند.آن‌ها از ویژگی‌های مختلف نظریه گراف مانند تعداد گره‌های همسایه و یال‌ها،betweenness centralityوcommunity cohesivenessبرای تمایز رفتارهای آنلاین افراد توسط الگوهای مصرف آن‌ها استفاده نمودند. به‌علاوه دنبال کردن ارتباطات کاربران می‌تواند الگوهای معناداری را آشکار سازد.زیرا کاربران می‌توانند هویت خود را با اطلاعات اشتباه پنهان سازند اما ارتباطات میان یکدیگر را نمی‌توانند پنهان کنند. آن‌هااز شاخص‌های محلی مانند single node (ego)و one-level neighborhood (an egonet)و two-level neighborhood (a super egonet)و betweenness centralityو average betweenness user’s egonetبرای شناسایی کاربران با ساختارهای ارتباطی ناهنجار استفاده کردند.

در طرح تشخیص ناهنجاری در گراف ویژگی ایستا در [43] یک روش تشخیص مبتنی بر گراف به‌نام GBADارایه دادند که اساساً مبتنی بر این نظریه است که یک فرد سعی در ارتکاب یک عمل غیرقانونی یا غیرعادی را دارد، بنابراین از رفتارهای شناخته شده‌ای پیروی و قصد واقعی خود را پنهان می‌کند. این روش شامل سه الگوریتم مختلف GBAD-MDLو GBAD-MPLو GBAD-Pاست.الگوریتمGBAD-MDLزیرساخت هنجاری را با استفاده از اصل بیشینه طول توصیف (MDL[[16]](#footnote-17)) پیدا می‌کند و زیرساختارهای مشابه را با سطح قابل پذیرشی از تغییر از زیرساختار طبیعی جستجو می‌کند.الگوریتم GBAD-MPLنیز بهترین زیرساختار را با جستجو در یال‌ها و راس‌هایی که گم شده‌اند تعیین می‌کند.الگوریتم GBAD-Pاز روش ارزیابی MDLبرای کشف بهترین زیرساختار در گراف استفاده می‌کند اما به‌جای امتحان کردن همه نمونه‌ها برای مشابهت، این روش همه‌ی بسط‌ها برای زیرساختارهای طبیعی را در جستجوی بسط با کمترین احتمال، امتحان می‌کند. نویسنده از این روش برای کشف کارمندان مشکوک و اعمال آن‌ها به‌عنوان یک ابزار برای پشتیبانی تحقیقات جرم استفاده نموده است.

#### 2 – 3 – 1 – 2 تشخیص ناهنجاری مبتنی بر اجتماع

در [44]یک روش تشخیص ناهنجاری مبتنی بر اجتماع ارایه دادند، با شناسایی اجتماعاتی که برای مرزهای اجتماعی اهمیتی قایل نیستند. این کار بر اساس یک نظریه تعلق گره‌های دارای سورفتار متمایل به چندین اجتماع است.نویسنده جداسازی اجتماعات را بهبود داده است که هر گره فقط به یک اجتماع تنها تعلق داشته باشد. در [45] یک روش خوشه بندی متمرکز و تشخیص ناهنجاری در گراف‌ها به نام FocusCoارایه نمودند.الگوریتم شامل سه گام است۱. استنتاج وزن‌های ویژگی‌ها ۲. استخراج خوشه‌های گراف‌های ویژگی متمرکز ایستا ۳. تشخیص ناهنجاری.

به‌طور مختصر هدف این است که خروجی مجموعه‌ای از گره‌های ارایه شده توسط کاربر که مربوط به ویژگی‌های متمرکز هستند توافق کنند. در این روش یک خوشه از گره‌های متصل به هم به نام خوشه های متمرکز، با توجه به ویژگی‌های متمرکز یافت می‌شود و بر اساس خوشه‌های متمرکز یک ناهنجاری به‌عنوان گره‌ای که از نظر ساختاری متعلق به خوشه هست اما انحراف زیادی در ویژگی‌های متمرکز دارد. آن‌ها همچنین نشان دادند که این روش برای گراف‌های ساختگی و واقعی بسیار مؤثر و مقیاس پذیر است.

### 2 – 3 – 2 تشخیص ناهنجاری گراف پویا

#### 2 – 3 – 2 – 1 تشخیص ناهنجاری مبتنی بر فاصله

گراف‌های دنیای واقعی به‌طور مداوم در حال تغییرند. تشخیص ناهنجاری در این نوع از گراف‌های پویا کاری بسیار چالش برانگیز است.معیار مبتنی بر فاصله می‌تواند برای اندازه‌گیری تغییر بین دو شی به‌کار رود. دو شی که در معیار اندازه اختلاف کمی دارند،یکسان نامیده می‌شوند.معیارهای مختلفی برای تشخیص ناهنجاری وجود دارد.فاصله‌ی خطای اصلاح تطابق گراف[[17]](#footnote-18) ،بیشینه زیرگراف مشترک[[18]](#footnote-19)، فاصله ماتریس همسایگی[[19]](#footnote-20)، فاصله ویرایش گراف[[20]](#footnote-21) ، فاصله همینگ برای ماتریس‌های همسایگی گراف‌ها و … [46]

#### 2 – 3- 2 – 2 تشخیص ناهنجاری مبتنی بر فشرده سازی

در این فرآیند یک نمایش گراف فشرده با استفاده ازحداقل طول توصیفی و روش فشرده‌سازی با بهره گیری از الگوها و تنظیمات داده‌ها با کمترین هزینه‌ی رمزگذاری به‌دست می‌آید.سپس ناهنجاری‌ها به‌عنوان گراف‌هایی که مانع فشرده‌سازی هستند تعریف می‌شوند[46].

#### 2 – 3 – 2 – 3 تشخیص ناهنجاری مبتنی بر تجزیه

این روش ناهنجاری‌های موقتی را با نمایش مجموعه‌ای از گراف‌های تکامل زمانی[[21]](#footnote-22)به‌عنوان یک تنسور یا آرایه‌ی چند بعدی تشخیص می‌دهد و factorizationیا کاهش بعد انجام می‌دهد. یک روش جدید تجزیه ماتریس فشرده[[22]](#footnote-23) برای محاسبه تقریب‌های کم مرتبه خلوت در [47] ارایه شده است. خطای بازسازی هرگراف خلوت در طول زمان پی‌گیری شد و درجایی که تغییر زیاد باشد، گراف منطبق ناهنجاری خواهد بود.

#### 2 – 3 – 2 – 4 تشخیص ناهنجاری مبتنی بر خوشه یا اجتماع

در مورد روش مبتنی بر اجتماع یا خوشه، به جای زیر نظر گرفتن تغییرات در کل شبکه، یک اجتماع در هر زمان، برای هر حادثه‌ی غیرعادی ای زیر نظر گرفته می‌شود. در [48] یک برنامه‌ی تشخیص داده‌ی پرت ساختار یافته در جریان‌های شبکه‌ای وسیع ارایه دادند که با تقسیم پویای شبکه برای ساخت مدل‌های آماری مقاوم در برابر رفتار ارتباطی است.

#### 2 – 3 – 2- 5 تشخیص ناهنجاری مبتنی بر مدل‌های احتمالاتی

نظریه احتمال مدلی می‌سازد که می‌تواند طبیعی تلقی شود و انحراف از این مدل به‌عنوان ناهنجاری علامت گذاری شود. یک روش دو مرحله‌ای در [49] ارایه شد که گام اول شامل مدل‌های بیزین مزدوج برای فرآیندهای شمارش زمان گسسته برای ردیابی جفت پیوندهای میان همه‌ی گره‌ها در گراف برای ارزیابی طبیعی بودن رفتار است. و در گام دوم استنتاج شبکه استاندارد روی کاهش زیرمجموعه‌ی گره‌های بالقوه غیرعادی اعمال می‌شود.

#### 2 – 3 – 2 – 6 تشخیص ناهنجاری مبتنی بر پنجره

الگوریتم‌های تشخیص ناهنجاری برخی روش‌ها را ارایه می‌کنند که محدود به یک چارچوب زمانی است. در [50]روشی ارایه دادند به نام یادگیری الگو و تشخیص ناهنجاری روی جریان‌ها[[23]](#footnote-24)که یک روش تقسیم‌بندی و پنجره‌گذاری است که گراف را همانطور که در طول زمان در جریان است تقسیم می‌کند و ناهنجاری‌ها و الگوهای غیرقانونی که متعلق به مجموعه‌ای از الگوهای قانونی یافت شده در پنجره زمانی جاری است را حفظ می‌کند.

**[59]تا[66] حذف شد!**

## 2 –4 تعریفBig Dataو کاربرد آن در تقلب

تعریف Big Data در تحقیقات مورد توافقی جهانی نیست. بنابراین ، از تعریف فراگیر دیمچنکو و همکاران استفاده می کنیم. که داده های بزرگ رابا پنجV تعریف می کنند: حجم (Volume)و سرعت (Velocity)و تنوع(Variety)، صحت(Veracity) و ارزش(Value).

حجم وابسته است به مقدار زیاد داده ها، سرعت مربوط است به سرعت بالایی که در آن داده های جدید تولید می شود ، تنوع مربوط به سطح پیچیدگی داده ها (به عنوان مثال ترکیبداده ها از منابع مختلف) ، صحت بیانگر اصالت داده ها ، ومقدار نشان دهنده ی اینکه چقدر کیفیت داده ها با توجه به نتایجمورد نظر، خوب است.

مجموعه داده های منتشر شده توسط CMS بسیاری از این خصوصیات Big Data را به معرض نمایش می گذارد. این مجموعه داده ها

واجد شرایط Big Volume هستند زیرا در برگیرنده سوابق سالانه مطالبات پزشکان ارائه دهنده خدمات پزشکی در کل ایالت متحده می باشند.

هر ساله CMS داده های سال گذشته را منتشر می کند که حجم وسیعی از داده های موجود را افزایش می­ دهد.مجموعه داده ها حاویحدود 30 ویژگی است. هر کدام از 30 ویژگی ، از مشخصات جمعیت ارائه دهنده و انواع پروسه های درمانی، تامبالغ پرداخت و تعداد خدمات انجام شده می باشد ، بنابراین واجد شرایطBig Variety می باشندعلاوه بر این ، مجموعه داده ترکیبی مورد استفاده در این تحقیق، ذاتاً شرط داده هایBig Variety را فراهم می کند ، زیرا سه منبع اصلی (اما متفاوت) Medicare را با هم ترکیب می کند. از آنجایی کهCMSیک برنامه دولتی با کنترل کیفیت شفاف و مستندات دقیق است، برای هر مجموعه داده ، این مجموعه داده ها قابل اعتماد ، معتبر و نمایانگر کلیه مطالبات ارائه دهندگان شناخته شده ی Medicareهستند که نشان از صحت Big Veracity داده ها دارد.مجموعه داده LEIE می تواند به عنوان Big Valueدر نظر گرفته شود زیرا شامل بزرگترین مخزن کلاهبرداران ارائه دهنده خدمات پزشکی شناخته شده دنیای واقعیدر ایالت متحده می باشد.

## بدون تیتر اصلی: برخی معیارهای تحلیل شبکه‌ی پزشکان

معیارهای مرکزی برای تجزیه و تحلیل اهمیت نسبی پزشکان در شبکه هستند، برای بررسی اینکه همکاری میان آنها در درمان بیماران وجود دارد یا خیر و عبارتند از:

1. درجه: اهمیت پزشک متناسب با تعداد بیماران مشترک با پزشکان دیگر است.
2. مقدارخاص[[24]](#footnote-25): هرچه تعداد بیماران مشترک با سایر پزشکان مهم بیش‌تر باشد، پزشک مهم‌تر درنظر گرفته می‌شود.. اگر پزشک تعداد زیادی از بیماران را به اشتراک بگذارد، اما با پزشکانی که در شبکه مهم نیستند، پزشک مهم در نظر گرفته نمی‌شود[67].
3. بینابینی[[25]](#footnote-26): با توجه به اینکه پزشکان می‌توانند برای دیگر پزشکانی که نزدیکترند تاثیرگذار باشند، مثلا زمانی که پزشک یک پزشک دیگر را به بیمار خود نشان می‌دهد،تعریف معیاری که این مجاورت را نشان دهد امکان‌پذیر است[68].
4. نزدیکی[[26]](#footnote-27): پزشکان می‌توانند توسط مقدار سایر پزشکان که در میان یکدیگر شناخته شده‌اند،غیرمستقیم و بدون اطلاع آن‌ها، مرتبط باشند.بنابراین، با توجه به تعداد پزشکان مورد نیاز برای اتصال دو پزشک، می توان میزان جدایی پزشکان را محاسبه کرد. [69].

## بدون تیتر اصلی: کلان داده‌ها[[27]](#footnote-28)در تشخیص تقلب

در داده‌های کلان، فقط موضوع حجم [[28]](#footnote-29) مطرح نیست و باید سایر موارد از قبیل تنوع[[29]](#footnote-30) داده‌ها و سرعت [[30]](#footnote-31), برقرار باشد تا بتوان به دنیای یادگیری عمیق وارد شد. در موضوع کشف تقلب حوزه درمان نه‌تنها با حجم انبوهی از داده‌های متنوع مواجه‌ایم، بلکه این داده‌ها و الگوهای مربوط روز‌به روز درحال تغییرند.

در سال‌های اخیر علاقه رو به افزایش در کاوش داده‌های مراقبت سلامت برای تشخیص تقلب شکل گرفته است.سیستم‌ها برای پردازش مطالبات الکترونیک پیاده‌سازی شده‌اند تا بصورت اتوماتیک بازرسی و مرور از داده‌ های مطالبات را انجام دهند.این سیستم‌ها برای تشخیص اعمال تقلبی، صورتحساب اشتباه، مطالبات تکراری و سرویس‌هایی که تحت پوشش درمانی نیستند، طراحی شده‌اند.قابلیت های تشخیص تقلب این سیستم‌ها معمولاً محدود است زیرا تشخیص بطور عمده مبتنی بر قوانین ساده از پیش تعریف‌شده توسط متخصصان امر است.برای رسیدن به تشخیص موثرتر، بسیاری از محققین روش‌های پیچیده‌تر مقابله با تقلب را توسعه داده‌اند که بر اساس داده کاوی، یادگیری ماشین و سایر روش‌های تحلیلی است. روش‌های جدید ارایه شده برخی مزیت‌های اصلی مانند یادگیری خودکار الگوهای تقلب از داده‌ها و تعیین احتمال تقلب برای هر مورد و شناسایی انواع جدید تقلب که قبلاً ثبت نشده‌اند را دارا هستند[70],[71].

## بدون تیتر اصلی: جدول

مواردی که در جدول زیر باید مورد بررسی قرار داد عبارتند از[12] :

* داده یک مساله مهم در زمینه مراقبت سلامت است.عمده داده شامل داده‌های مطالبات از منابع دولتی و شرکت های بیمه خصوصی هستند.
* سیستم‌های مراقبت سلامت هر کشور متفاوتند و بطور مداوم در حال تغییر و توسعه اند.
* چندین تحقیق روی برخی کشورهای توسعه یافته مانند آمریکا و استرالیا انجام شده است. به عبارت دیگر کشورهای مختلف باید بعنوان منابع داده ی جدید در نظر گرفته شوند.
* تشخیص تقلب مراقبت سلامت عمدتا با استفاده از یادگیری ماشین و داده کاوی انجام شده است. روش‌های یادگیری ماشین به سه دسته تقسیم می‌شوند با ناظر، بدون ناظر و نیمه نظارتی.
* بیشتر مطالعات از روش‌های یادگیری بدون ناظر استفاده کردند. در برخی موارد روش‌های یادگیری نیمه نظارتی ارایه شده نیز می‌توانند در تشخیص تقلب مراقبت سلامت مفید باشند.
* تحقیقات بررسی شده نشان می‌دهند الگوریتم های شناخته شده مانندSVMو KNNو بیزین برای کلاس بندی، خوشه بندی و تشخیص موارد غیرعادی(anomaly) در تشخیص تقلب مراقبت سلامت استفاده شده اند.
* اگرچه الگوریتم های متفاوتی برای تشخیص تقلب مراقبت سلامت به کار گرفته می‌شوند اما الگو یا روش استانداردی که همه موارد را پوشش دهد وجود ندارد.
* با توجه به انواع تقلب، اکثریت تحقیقات روی تشخیص تقلب ارایه دهندگان خدمات انجام شده زیرا تقلب ارایه دهندگان خدمات یک مساله مهم برای بهبود کیفیت و امنیت یک سیستم مراقبت سلامت است، محققان زیادی به آن‌ها توجه نشان می‌دهند.
* بطور ویژه تحقیق زیادی روی تشخیص تقلب دسیسه ای [[31]](#footnote-32) صورت نگرفته ، اگرچه که چنین تحقیقاتی می‌توانند برای عواقب سخت تقلب دسیسه گران و کاهش هزینه‌های مراقبت سلامت بسیار مفید واقع شوند.
* مرسوم ترین منبع داده ی استفاده شده در آمریکا HFCA، .در استرالیا HCIو در تایوان NHIاست.

در [12]برای تشخیص تقلب مراقبت سلامت روش یادگیری بدون ناظر پرتکرارترین روش استفاده شده است زیرا به‌دست آوردن داده ی برچسب دار در تشخیص تقلب حوزه سلامت بسیار دشوار و پرهزینه است.

- داده ی مراقیت سلامت در حال حاضر به‌عنوان مجموعه‌ای از داده‌های بزرگ از بسیاری از انواع داده‌ها درنظر گرفته می‌شود.این شرایط مفهوم کلان داده را در‌پی دارد.

- کلان داده در تحلیل‌های مراقبت سلامت یک زمینه تحقیقاتی جدید است و مطالعات کمی در این زمینه گزارش شده اند.

- بیشتر مقالات موجود هنوز روش‌های سنتی داده کاوی را پوشش می‌دهند و حمایت زیادی برای تحلیل کلان داده وجود ندارد.

- کمبود راه حل‌های مفهومی کلان داده و کاربرد آن با ابزارها، کتابخانه‌ها و بسترهای کلان داده ی جایگزین در تشخیص تقلب مراقبت سلامت

- ترکیباتیازانواعدادهدرزمینه‌هایمختلفبایدبررسیشوندتاروش‌هایمؤثربدستبیایند.

## بدون تیتر اصلی: رویکردهای کلی روش‌های کشف تقلب

## بدون تیتر اصلی:روش‌های با ناظر

در [76] مقایسه‌ای از درخت‌های تصمیم،شبکه عصبی و رگرسیون منطقی به ترتیب نرخ صحیح شناسایی تقلب پزشکی آن‌ها ارایه داده شده‌است. شبکه‌های عصبی می‌توانند به مجموعه‌داده‌های پیچیده و بزرگ و ارتباطات غیرخطی متغییرها رسیدگی کند. اگرچه کاربرد شبکه عصبی به‌طور کلی نیازمند متخصصان آماری برای مثلاً تنظیم پارامترها می‌باشد. به‌علاوه [77] بر این نکته اشاره دارد که کلاس‌بندی ممکن است کارایی کم و بیش از حد وابستگی[[32]](#footnote-33) به مجموعه داده‌ی اریب داشته باشد .برای جلوگیری از مشکل بیش از حد وابستگی به مجموعه داده‌ها [78] یک روش توقف اولیه[[33]](#footnote-34) در تشخیص تقلب پزشکی مبتنی بر شبکه عصبی‌شان پیاده‌سازی نمودند، که بر اساس استفاده از یک مجموعه آموزشی برای به‌روز‌رسانی وزن‌ها و بایاس‌ها و یک مجموعه آموزشی دیگر برای توقف آموزش زمانی‌که شبکه شروع به بیش از حد وابستگی به داده‌ها می‌کند، می‌باشد. آن‌ها همچنین به حل مساله‌ی اختلاف زیاد پیش‌بینی به دلیل سایز کم نمونه‌های با تعداد زیاد ویژگی پرداختند.

در مقایسه، درخت‌های تصمیم قوانین کلی‌ای دارند که تفسیر راحتی دارند . درخت تصمیم می‌تواند داده‌های خلوت را نیز بررسی کند اما ممکن است با افزایش سایز داده منجر به وابستگی بیش از حد به داده و کاهش قابلیت تفسیر نتایج شود. برای‌مثال [79] یک مدل امتیازدهی برای احتمال سواستفاده ارایه داد و سپس ارایه دهندگان را با استفاده از درخت تصمیم، کلاس‌بندی نمود. [80] یک شبکه بیزین پویا از شاخص‌های تقلب ارایه داد که وزن‌ها توسط قدرت پیش‌گویی تقلب هر ویژگی تعیین می‌گردد. طبقه‌بندهای بیزین زمان آموزش کوتاهتری دارند و در بررسی داده‌های خلوت مؤثرتر دیده شده‌اند. [81] از الگوریتم kنزدیکترین همسایه برای طبقه‌بندی نمایه‌ی پزشکان استفاده نموده است.[82] یک روش مبتنی بر بردار ماشین پشتیبان استفاده نموده است. تعدادی از تلاش‌های تشخیص پزشکی شامل ترکیب روش‌های با ناظر می‌باشد.

به‌طور کلی روش‌های باناظر برای تشخیص الگوهای قبلاً کشف شده‌ی تقلب مناسب است . به‌دلیل اینکه این روش‌ها براساس طبقه‌بندی مطالبات قبلی هستند ، یک نفر باید از مساله‌ی تخمین بالقوه بیش ازحد[[34]](#footnote-35) آگاه باشد[76].

وجود کلاس‌های با سایز نامتعادل درداده‌های مطالباتنیز می‌تواند منجر به وابستگی بیش‌از حد به داده‌ها گردد.. این مدل ها برای مقابله با الگوهای تقلب جدید و تغییرات در قوانین باید به طور مرتب به روز شوند. عدم توانایی روش های نظارت شده برای تشخیص تقلب های پویا و انطباق، توجه به روش های نا منظم را افزایش داده است[83].

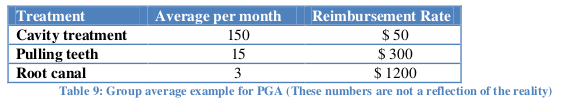
## بدون تیتر اصلی:روش‌های بدون ناظر

خوشه‌بندی اولین بار روی داده‌های پزشکی برای بخش‌بندی درمان پزشکان عمومی توسط [84] اعمال شد. در [85] و [86] از داده های جغرافیایی در یک رویکرد مبتنی بر خوشه بندی استفاده نمودند.الگوریتم گروه‌خوشه‌ای بیزین برنولی[[35]](#footnote-36)[87] با تمرکز بر وقوع ویزیت میان ارایه‌دهندگان و ذی‌نفعان، داده‌های دوتایی را مدل می‌کند.این به طور بالقوه می تواند یک نوع تقلب در حال ظهور به نام "تقلب توطئه"را نشان دهد که شامل ویژگی های بیش از یک عضو از سیستم پزشکی است. این الگوریتم های خوشه بندی به بازرسان کمک می کند که صورتحساب و متغیر مطلوبشان را گروه‌بندی کنند. [88] یک مرور کلی از روش های تشخیص داده پرتدر برخی از آزمایش ها برای ارزیابی اثربخشی آن ارایه می‌کند. این روش های تجزیه تحلیل شامل مدل های خطی، طرح جعبه‌ای[[36]](#footnote-37)، تحلیل قله[[37]](#footnote-38)، خوشه بندی چند متغیره و ارزیابی متخصص می باشد.[89] یک روش تشخیص داده پرت مبتنی بر چگالی محلی برای شناسایی الگوهای پرداخت نامناسب در سیستم پزشکی استرالیا ارایه می‌دهد. [90]یک رویکرد یکپارچه که ترکیبی از انتخاب ویژگی، خوشه بندی، تشخیص الگو و تشخیص بیرونی است برای شناسایی تقلب در سیستم پزشکی استرالیا ارایه نمودند.[91] یک روش تشخیص ناهنجاری دو مرحله ای برای شناسایی بیمارستان های جعلی در سیستم مراقبت بهداشت عمومی برزیل ارایه می‌کند. همچنین شامل مطالعات تشخیص داده پرت با داده های تجویزی است. [92] یک الگوریتم بدون ناظر مبتنی بر فاصله برای ارزیابی خطر تقلب نسخه‌ها ارایه نموده است. [93] یک مدل رفتاری پایه نرمال را برای شناسایی ناهنجاری ها برای شناسایی ناهنجاری‌های مربوط به هر نسخه ایجاد می‌کند.[94] یک مدل تشخیص داده پرت مبتنی بر استنتاج بیزی است که با استفاده از توزیع احتمالات و فواصل قابل قبول برای ارزیابی ارجاعات ارائه می‌دهد.[95] استفاده ازیک تابع غلظت[[38]](#footnote-39) را به عنوان یک ابزار تشخیص پیش نمایش داده پرت برای کمک به ارزیابی تقلب پزشکی ارائه می دهد.علاوه بر این، ابزارهای صنعتی مبتنی بر تجزیه و تحلیل گراف، تجزیه و تحلیل ارتباطات و انجمن‌ها، ممکن است به بازرسان کمک کند تا روابط، پیوندها و الگوهای پنهان به اشتراک گذاری اطلاعات و تعاملات در گروه های بالقوه جعلی ارائه دهندگان و بیماران را آشکار سازند تعداد و کیفیت ارتباط بین مشاغل را می توان با استفاده از شباهت در اطلاعات ارتباطی آن‌ها، مکان، ارائه دهندگان خدمات، دارایی‌ها و وابستگی‌ها تجزیه و تحلیل نمود. ارتباطات بالقوه با بازیکنان درگیر در تقلب ممکن است پرچم های قرمز را به ارمغان بیاورند و منجر به تحقیقات آتی گردند. این به طور خاص می‌تواند برای آشکارسازی شبکه های سازمان یافته، پیچیده و هماهنگ ارائه دهندگان و بیماران مفید باشد. رویکردهای بدون ناظر به طور کلی مورد استفاده قرار می گیرد تا قبل از اینکه متخصصان حوزه را به تحقیق بفرستند فعالیت های جعلی را به طور بالقوه برچسب بزنند. بنابراین، یک همکاری نزدیک بین پزشکان، آمارگیران و افرادی که در تصمیم گیری شرکت دارند، در مراحل تعیین و تنظیم مدل و تجزیه و تحلیل و تفسیر نتایج سودمند خواهد بود[83].

الگوریتم Aprioriیکی دیگر از تکنیک‌هایی است که در تشخیص تقلب استفاده می‌گردد. این الگوریتم(Agrawalو همکاران ۱۹۹۳)، مهم‌ترین الگوریتم کلاسیک برای کاوشاقلاممکرر است.Aprioriبرای یافتن همه اقلام مکرر در پایگاه داده داده شده DBاستفاده می شود. بر اساس اصل Aprioriهر زیر مجموعه ای از اقلام مکرر نیز باید مکرر باشد. به عنوان مثال: اگر {XY} مجموعه اقلام مکرر است، هر دو {A} و {B} باید مجموعه های مکرر باشند.ایده کلیدی الگوریتم Aprioriاین است که چند گذر از پایگاه داده را انجام دهیم.که یک رویکرد تکراری که به نامجستجوی اول-پهنا[[39]](#footnote-40)(جستجوی سطح هوشمندانه) شناخته می شود که در آن k-آیتمبرای کشف (k + 1)آیتمبه کار می روند. در ابتدا، مجموعه اقلام ۱-تکراری یافت می‌شود که آستانه پشتیبانی را برآورده می کند، توسط L1نشان داده می‌شود. در هر گذر بعدی، ما با یک مجموعه بذر از اقلام موجود در گذر قبلی که بزرگ بوده است، شروع می‌کنیم . این مجموعه بذر برای تولید مجموعه های جدید بالقوه بزرگ استفاده می شود که به نامهای مجموعه اقلام کاندید شناخته می‌شود.در پایان گذر، تعیین می‌شود که کدامیک از اقلام نامزدها واقعا بزرگ (مکرر) هستند، و آنها تبدیل به دانه برای گذر بعدی می شوند. بنابراین، L1برای پیدا کردن L2استفاده می شود، مجموعه ای از مجموعه های مکرر 2-آیتم که برای پیدا کردن L3و غیره استفاده می شود، تا زمانی که هیچ مجموعه مکرر k-آیتمی موجود نباشد [96]

روش های مختلفی برای بهبود کارایی الگوریتم Aprioriمانند جدول هش، کاهش تراکنش، تقسیم بندی و … استفاده می‌شود [97],[98]. در[99]، نويسندگان روشي را براي تفسير ويژگي هایی که مقادیر پیوسته دارند با استفاده از فاصله مساوي عرض باند داخلی[[40]](#footnote-41)ارایه نمودند كه براساس نظر متخصصين پزشكي انتخاب شده است.یک تحقیق دیگر[100] ، صورتحساب پزشکی را با استفاده از الگوریتم Aprioriتحلیل می کند.، برخی از اصلاحات را در الگوریتم Aprioriموجود پیشنهاد دادند و سپس از اثربخشی آن در اطلاعات مفید ساخته شده در صورتحساب پزشکی استفاده کردند. همچنین از الگوریتم Aprioriبرای کشف بیماری های مکرر در اطلاعات پزشکی استفاده می کند. در [101]روشي براي تشخيص وقوع بيماري با استفاده از الگوريتم Aprioriدر نقاط خاص جغرافيايي در دوره زماني خاص ارائه شده است.

بولتن و هاند[102] در سال ۲۰۰۱، PGA[[41]](#footnote-42)را به عنوان یک روش نامزد برای یک تکنیک تشخیص تقلب بدون ناظر ارایه نمودند.که ترکیبی است از تحلیل خوشه بندی و نمایه سازی. تحلیل خوشه یک کار توصیفی مرسوم برای شناسایی یک مجموعه محدود از دسته ها یا خوشه ها برای توصیف مجموعه داده است[103]. برای کشف اینکه آیا یک ارایه دهنده یک رفتار مطالبه‌ای مشکوک دارد یا نه باید با سایر متخصصان زمینه مشابه مقایسه شود. PGAابزاری محبوب است برای فهم اینکه چگونه رفتار یک پزشک خاص با رفتار سایر پزشکان در یک گروه خاص مرتبط می‌شود. یک جنبه مهم از این تحلیل این است که چگونه ارایه دهندگان با یکدیگر گروه می‌شوند و چرا با هم گروه می‌شوند. به‌طور خاص در صنعت پزشکی متخصصان فوق تخصص بسیاری وجود دارد و گروه‌بندی تقریبی برخی از آن‌ها باهم کار ساده‌ای نیست. بنابراین استفاده از متخصصان برای تشکیل گروه برای اعمال تحلیل PGAنیاز است. فرض بر این است که گروه‌بندی پزشکان ممکن است و سپس مقایسه رفتار می‌تواند آغاز گردد. به‌عنوان مثال ۲۰ دندانپزشک در کد‌پستی خاصی باهم گروه بندی شده‌اند و توزیع معالجات بررسی شده است، نتیجه میانگین گروهی تعداد دفعات معالجات ارایه شده در هر گروه از بیماران است. خلاصه‌سازی‌های میانگین‌های گروهی در جدول نشان داده شده‌اند و این مثالی است از اینکه چگونه PGAمی‌تواند به‌کار رود. حال نمایه‌های یک دندانپزشک خاص می‌تواند با میانگین گروه مقایسه شود. اگر دندانپزشکی در طول یک دوره زمانی خاص به‌طور غیرعادی تعداد بالایی کانال ریشه انجام داده باشد ، به این معنی نیست که مرتکب تقلب شده است و می‌تواند بر تحقیقات بیشتر روی آن دندانپزشک دلالت داشته باشد. از آنجایی که کانال ریشه یک جراحی دهانی در نظر گرفته می‌شود نرخ بازپرداخت آن به‌طور قابل ملاحظه ای بالاتر از سایر معالجات است و بنابراین محتمل است که دندانپزشک با قصد منحرف از این معالجه برای صورتحساب بیشتر استفاده کرده باشد.



شکل 1-4مثالی از میانگین گروهی برای PGAد [14]

روش‌های کلاس بندی که برای تشخیص اختلاف بین مطالبات جعلی و قانونی آموزش داده شده‌اند فرصتی را برای استفاده از تشخیص تقلب در حوزه پزشکی فراهم می‌کنند. روش‌های با ناظر در تشخیص تقلب در کارت‌های اعتباری و در حوزه مخابرات نسبت به این روش‌ها در بخش‌های پزشکی خاص مانند مراقبت‌های بیمارستانی که تشخیص اینکه آیا فرایند ارایه شده واقعاً رخ داده یا ضروری بوده یا خیر ، ساده‌تر هستند. زمانی‌که تقلب کننده، مطالبات جعلی که مشابه قانونی هستند را ارایه می‌کند یک روش شناسایی که تشخیص دهد آیا درمانی صورت گرفته یا نه، نیاز است. در این‌گونه موارد زمانی‌که هیچ رفتار صورتحسابی افراطی‌ای وجود ندارد، تشخیص روش‌هایی مانند Profilingو تشخیص outlier با شکست مواجه خواهند شد. روش‌های بدون ناظر مانند تشخیص داده پرت بر غیرعادی بودن و پرت بودن متمرکز هستند که اگر صورتحساب طبیعی باشد رخ نمی‌دهند. Profilingوmonitoringمتمرکز بر تشخیص تغییرات در رفتار هستند و ابزار موثری برای مقابله با این نوع تقلب به نظر نمی‌رسند[14].

## بدون تیتر

در [118]یک روش تشخیص تقلب موثر هیبریدیSSIsomapو SimLOFپیشنهاد شده است.SSIsomap، در‌واقع isomapرا برای رفتار خوشه‌ها در رفتار کلاس‌ها بهبود می بخشد و SimLOFکه LOFرا بهبود می بخشد تا تشخیص داده پرت را بهبود بخشد، سپس از شواهد تئوری DempsterShaferبرای ترکیب شواهد الگوی رفتاری و شواهد بیرونی استفاده می‌شود، که درجه اعتقاد به تقلب برای مطالبات جدیدی که از راه می‌رسند فراهم می‌کند. نتیجه آزمایش نشان می دهد که روش آن‌ها دقت بیشتری نسبت به روش های موجود در تشخیص تقلب بیمه های پزشکی دارد[118].

در جدول2-4 ,3-4هریک از رویکردها و روش‌هایی که پیش‌تر بحث شد به تفکیک مزایا و معایب بیان شده است:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ردیف | نام روش | ایده | مزایا | معایب | مرجع | پارامترها |
| ۱ | ‌Medical Insurance Fraud Recognition Based On Improved Outlier Detection Algorithm | استفاده از یک الگوریتم تشخیصOutlierبهبود یافته بر اساس خوشه بندیk-means | * استفاده از مجموعه داده واقعی * کاهش زمان اجرای الگوریتم با یافتن مقدار بهینه kبا پیچیدگی زمانی از مرتبه‌o(I\*k\*m\*(no))   و پیچیدگی مکانی از مرتبهo((n-o)+k)\*m  که نسبت به الگوریتمCBLOFبهبود داشت. | * ویژگی‌های استفاده شده فقط مربوط به بازپرداخت بیمه سلامت هستند و شامل اطلاعات بیمار و نسخه نمی‌باشند | [108] | * بازپرداخت مربوط به برونشیت مزمن * بازپرداخت مربوط به بیماری‌های قلبی-ریوی * بازپرداخت مربوط به ذات الریه |
| ۲ | Outlier Detection In HealthCare Fraud-A Case Study In The Medicaid Dental Domain | اعمال فیلترینگ ویژگی‌ها برای جداسازی بازپرداخت های کم، تعداد بیماران کم، تعداد مطالبات کم و استفاده از تکنیک های تحلیل و آنالیز و استفاده از تکنیک های تشخیص Outlierشامل انحراف از مدل خطی،انحراف خوشه،انحراف انحراف از خوشه تکی،انحراف گرایشی،حداکثر انحراف | * تست روی مجموعه داده واقعی * استخراج ویژگی‌ها و ارزیابی با استفاده از انجام مصاحبه با متخصصان و پزشکان * تفسیر نتایج توسط تحلیل گران | * اعتبارسنجی دشوار اثربخشی * تکنیک تشخیص ‌‌‌‌‌‌‌‌outlierنیاز به همکاری متخصصان امر برای طراحی معیارها و به‌خصوص تفسیر نتایج دارد * تکنولوژی outlierهنوز در مرحله آزمایشگاهی است و خود را در عمل و در اجرای طولانی ثابت نکرده | [109] | * تعداد مطالبات بازپرداخت هر ذینفع * مقدار مطالبات بازپرداخت هر ذینفع * مقدار مطالبات در ایام تعطیل * میانگین تعداد مطالبات بازپرداختی هر ذینفع * میانگین مقدار مطالبات بازپرداختی هر ذینفع * کد نسخه * کد دندان * هزینه نسخه * میانگین تعداد نسخه های بازپرداختی هر مطالبه |
| ۳ | A Novel Page Rank-Based Algorithm To Identify Anomalies | استفاده از یک الگوریتم personalized page rankو محاسبه یک specialty centric personalized page rankبرای هر نود و سپس اتصال نودها براساس آن برای بدست آوردن آنومالی | * Page rankکه سابقاً در حوزه تحلیل شبکه به کار رفته بود در حوزه تقلب در مطالبات بیمه نیز مؤثر عمل می‌کند | * وقوع برخی FPها به دلیل اشتراک مشخصه های کلی درCPT (نسخه)ی پزشکان با تخصص های مختلف | 110]] | * Medical procedure code * national provider identifier * specialty * ‌‌تعداد procedureها در هر سال |
| ۴ | Community Detection Algorithm To Find Suspicious Group Of Provider Community | ساخت ماتریس ارتباط میان دو پزشک و شناسایی پزشکانی که در یک شبکه به‌هم متصلند و با سایر پزشکان ارتباطی ندارند،بعنوان شبکه تقلب آمیز، و همچنین استفاده از اپراتورهای DBبه جای برای loopبیش از دو پزشک | * تخصیص احتمال (Likelihood/ احتمال تشکلیل شبکه انحصاری) به هر پزشک * سرعت تشخیص بالا | * استفاده از داده‌های ساختگی * استفاده از برخی پکیج های نرم افزاری برای محاسبه ماتریس روابط، که برای مجموعه داده‌های بزرگ بهینه نیست * عدم تفکیک پزشکان طبق تخصص آن‌ها * post payment | [111] | * procedure code * صورتحساب هر procedure code * health claim data * تاریخ ارایه خدمات * کد پزشک * تعداد ویزیت های هر بیمار به ازای هر پزشک |
| ۵ | ‌Multi Stage Method To Detect Provider And Patient Fraud | محاسبه یک معیار ریسک بر اساس فاصله مهالنوبیس و چگالی ها و محاسبه ریسک و ساخت درخت تصمیم آن | * دقت بالا در مقایسه با روش * دقت بالا روی همه ۴ تخصص | * عدم آزمایش روش پیشنهادی با تعداد متخصصان بیشتر | [112] | * تخصص پزشکان(چشم، اعصاب، حلق، عمومی) * نرخ شکایت از پزشکان * مدت زمان هر ویزیت * تعداد ویزیت ها * تعداد تشخیص(نسخه) * تعداد سرویس ها و خدمات * تعداد دارو |
| ۶ | An Interactive Machine Learning Based Electronic Fraud And Abuse Detection System In HealthCare Insurance | تولید سناریو توسط متخصصان و پزشکان برای رفتارهای غیرطبیعی و سپس وزن دهی actorها با روش وزن دهی binary pairwise comparison‌و محاسبه امتیاز خطای actorها و مطالبات، استفاده از ابزارvisualizationتوسعه یافته تحتQlikViewکه هم برای تحلیل و هم به‌کار می‌رود | * سازگار در یک محیط پویا * قابل استفاده برای تحلیلproactive , reactive * کاهش زمان تحلیل نتایج خروجی توسط کاربران به دلیل استفاده از ابزار visualization * استفاده از بیش از یکactor | * تغییر مداوم وزن ها توسط متخصصان برای شناسایی انواع جدید تقلب * تغییر ویژگی‌های فیلد ورودی توسط متخصصان برای شناسایی انواع جدید تقلب | [113] | * نسبت تعداد نسخه ها به تعداد مشخصی از بیمه شدگان * نسبت تعدادنسخه ها به تعداد مشخصی از پزشکان * تعداد کل نسخه ها |
| ۷ | Improving Fraud and Abuse Detection in General Physician  Claims: A Data Mining Study | انجام عمل خوشه بندی بر اساس hierachical clustering methodو محاسبه تعداد بهینه خوشه ها بر اساس معیار فاصله Euclidian distance measuresبا استفاده از شاخص اعتباری بیشینه مقدار ضریب همبستگی سیلوعت | * انجام تحقیق روی پزشکان هردو بخش عمومی و خصوصی * استفاده از مجموعه داده واقعی | * حذف داده‌های ناشناس * عدم استفاده از روش‌های آماری برای پرکردن داده‌های از‌دست رفته | [114] | * درصد بیمارانی که بیش از یکبار درماه ویزیت شده‌اند * میانگین اقلام دارو در یک نسخه * میانگین هزینه نسخه دارویی پزشک * تعداد نسخ تزریقی/حاوی آنتی بیوتیک * تعداد نسخ تزریقی/حاوی آنتی بیوتیکس |
| ۸ | A Social Network Analysis Framework For Modeling Health Insurance Claims Data | به‌کار‌گیری تکنیک‌های برای تحلیل مطالبات بیمه سلامت از طریق نگاشت پزشکان با استفاده از بیماران مشترک بعنوان یک پروکسی برای ارتباط میان آن‌‌ها | * در نظر گرفتن رابطه میان پزشکان، پزشک و بیمار، پزشک و ارایه دهندگان خدمات * استفاده از مجموعه داده واقعی * بهبود درک اهمیت ویژگی‌های میان افراد * ارزیابی مدل توسط تحلیلگران فرآیند و پزشکان و متخصصان | داده ی استفاده شده فقط مرتبط به خدماتی است که توسط پزشک ارایه شده و شامل مطالبات مرتبط با تحلیل‌های کلینیکی، آزمایش‌ها و عکس برداری یا بستری در بیمارستان نیست | [73] | * IDبیمار * نوع نسخه * اطلاعات زمانی مرتبط به حادثه‌ * IDمشاغل درگیر در فرآیند * هزینه نسخه * تخصص پزشک |
| ۹ | HealthCare Fraud Detection Methods And New Approaches | استفاده از رویکرد کلان داده در تشخیص تقلب مراقبت سلامت | * کاهش هزینه‌های سیستم درمانی * بهبود تشخیص تقلب * یادگیری خودکار الگوهای رفتاری | نیاز به حجم داده زیاد | [12] | - |

جدول ۳-۴دستهبندیبهتفکیکرویکردهایکلیکشفتقلب

|  |  |  |
| --- | --- | --- |
| رویکردها | مزایا | معایب |
| روش‌های آماری | * به سرعت ارائه دهندگان مشکوک را شناسایی می کند. * می تواند نوع جدیدی از تقلب را شناسایی کند. | * ممکن است مطالبات قانونی را به عنوان جعلی شناسایی کند * نیاز به بررسی مطالبات پس از ارزیابی آماری. * نیاز به دانش از روش های آماری. * نیاز به تخصص مقدم بر تشخیص تقلب |
| روش‌های یادگیری ماشین با ناظر | * سادگی و عدم نیاز به پردازشگر قوی | * نیاز به متخصص برای داده‌ی برچسب دار * overfitting |
| روش‌های یادگیری ماشین بدون ناظر | * هزینه ی کم‌تر به دلیل عدم نیاز به داده‌ی برچسب‌دار | * تغییر مداوم وزن ها و پارامترها توسط متخصصان برای شناسایی انواع جدید تقلب * تغییر ویژگی‌های فیلد ورودی توسط متخصصان برای شناسایی انواع جدید تقلب * هزینه‌ی بالای داده‌ی برچسب‌دار |
| روش‌های یادگیری ماشین ترکیبی | .   * در مقایسه با روش‌های کاوش فرایند دارای نرخ کشف کاذب[[42]](#footnote-43) کمتر است. * کاهش هزینه‌ی برچسب داده‌ها | * پوشش کم موارد تشخیص تقلب |

# فصل سوم

# روش پیشنهادی

در این فصل، به بیان اجزاء مختلف سیستم پیشنهادی و روابط میان آنها پرداخته شده است. در این پژوهش، هدف استفاده از Big data برای تشخیص تقلب مطالبات پزشکیاست . شکل 3-1 شمای کلی سیستم پیشنهادی را نشان میدهد که در آن ورودی، مطالبات پزشکی؟؟؟؟؟؟ و خروجی آن، ؟؟؟؟؟؟؟ برچسب­گذاری شده میباشد. در این فصل، نشان داده می­شود که سیستم از چه واحدهایی تشکیل شده است و ورودی و خروجی هر واحد نیز مشخص خواهد شد. پس از پرداختن به هر واحد تشکیل دهندهی سیستم، جزئیات و واحدهای تشکیل دهنده آن شرح داده می­ شوند و نهایتا فلوچارت الگوریتم به کار رفته ارائه خواهد شد.جدول 3-1 نیز شرح اختصارات به کار رفته در این پژوهش میباشد.

مطالبات پزشکی

سیستم تشخیص تقلب پزشکی

مطالبه ی برچسب­گذاری شده

شکل 3 - 1 شمایکلیسیستمپیشنهادی

شکل 3 - 1 شمایکلیسیستمپیشنهادی

در نمودار شکل 3-2 ،اجزای سیستم پیشنهادی مشخص شده است. این نمودار نشان میدهد که ابتدا ویژگیهای مطلوب با استفاده از الگوریتم استخراج ویژگی مبتنی بر گراف که در ادامه در مورد آن بیشتر توضیح داده خواهد شد، استخراج می­گردند و بردارهای ویژگی ایجاد شده در اختیار واحد طبقه بند قرار میگیرند. سپس الگوریتم طبقه بند پیشنهادی بر روی بردارهای ویژگی اعمال می­شود و در نتیجه پرس و نتایج برچسبگذاری میگردند. در مرحله بعد، هر مطالبه جدیدی که وارد سیستم گردد، مجددا با استفاده از الگوریتم استخراج ویژگی مبتنی بر گراف به بردار ویژگی تبدیل میگردد و با بهره­گیری از نمونه هایبرچسب دار و همچنین الگوریتم طبقه بندی مشخص می کند که به کدام نمونه های آموزشی شباهت بیشتری دارد، برچسب نمونه جدید را ایجاد می­کند.

مطالبات پزشکی

تشکیل محموعه مرجع pposو pneg  و اتصال به عناصر محور

تولید ویژگی های مبتنی بر گراف

تولید بردار رفتار برای هر p

محاسبه مشابهت

تعیین برچسب

P unknown

شکل 3 -2معماری سیستم پیشنهادی

مطالبات پزشکی، شامل سه/چهار دیتاست می­باشند. در قسمت استخراج ویژگی، جزئیات کامل مطالبات و ویژگیهایی که استخراج میشوند آورده شده است.

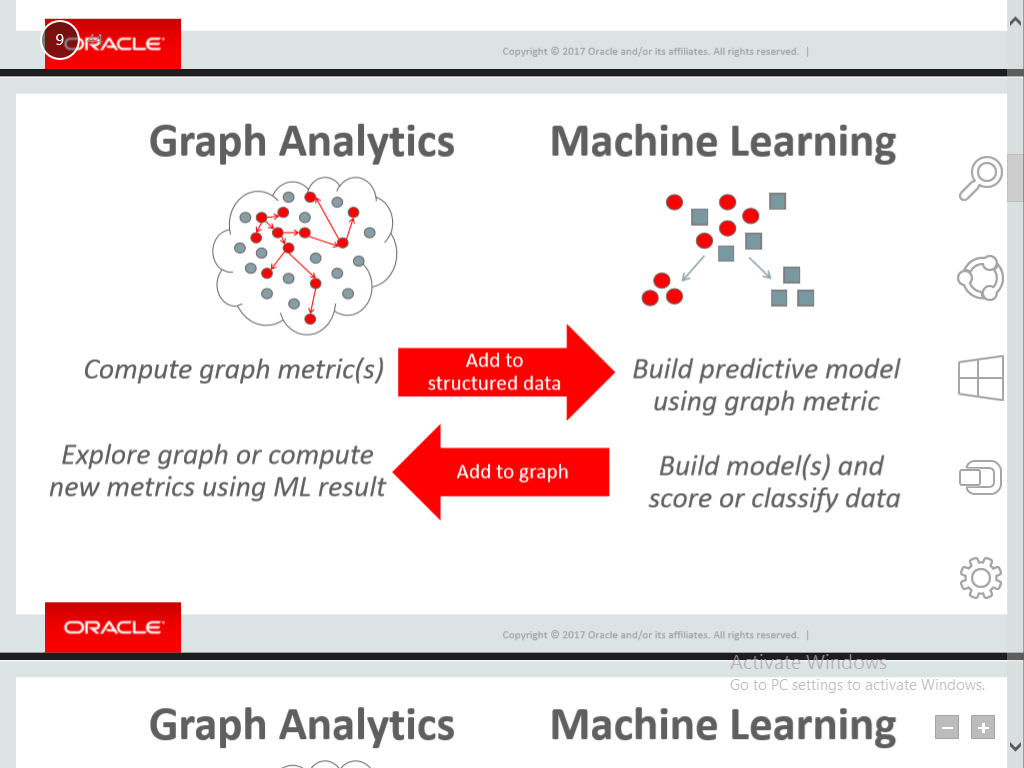
## 1-3 تشکیل گراف / ترکیب دیتاست ها

در نمودار شکل 3-3 نشان داده شده است که پس از ورودمطالبات به پایگاه داده اسپارک ؛ گرافی تشکیل میشود و سپس با استفاده از ؟؟؟ ویژگیهای آنها استخراج می شوند و به این ترتیب بردار ویژگی آن ها تشکیل میشود. در دیتاست موجود، اگر ویژگی مورد نظر وجود داشته باشد خانه­ی مربوط به آن به وسیله­ی آن ویژگی پر خواهد شد، در غیر این صورت، خانه­ی مربوط به آن خالی خواهد ماند و مقدار null خواهد گرفت. قابل توجه است که پیچیدگی زمانی الگوریتم به کار رفته در این قسمت برابر باO (???) است که در آن n بیانگر تعداد مطالبات است**.**

## 3– 2 تولید بردار رفتار برای ارایه دهندگان

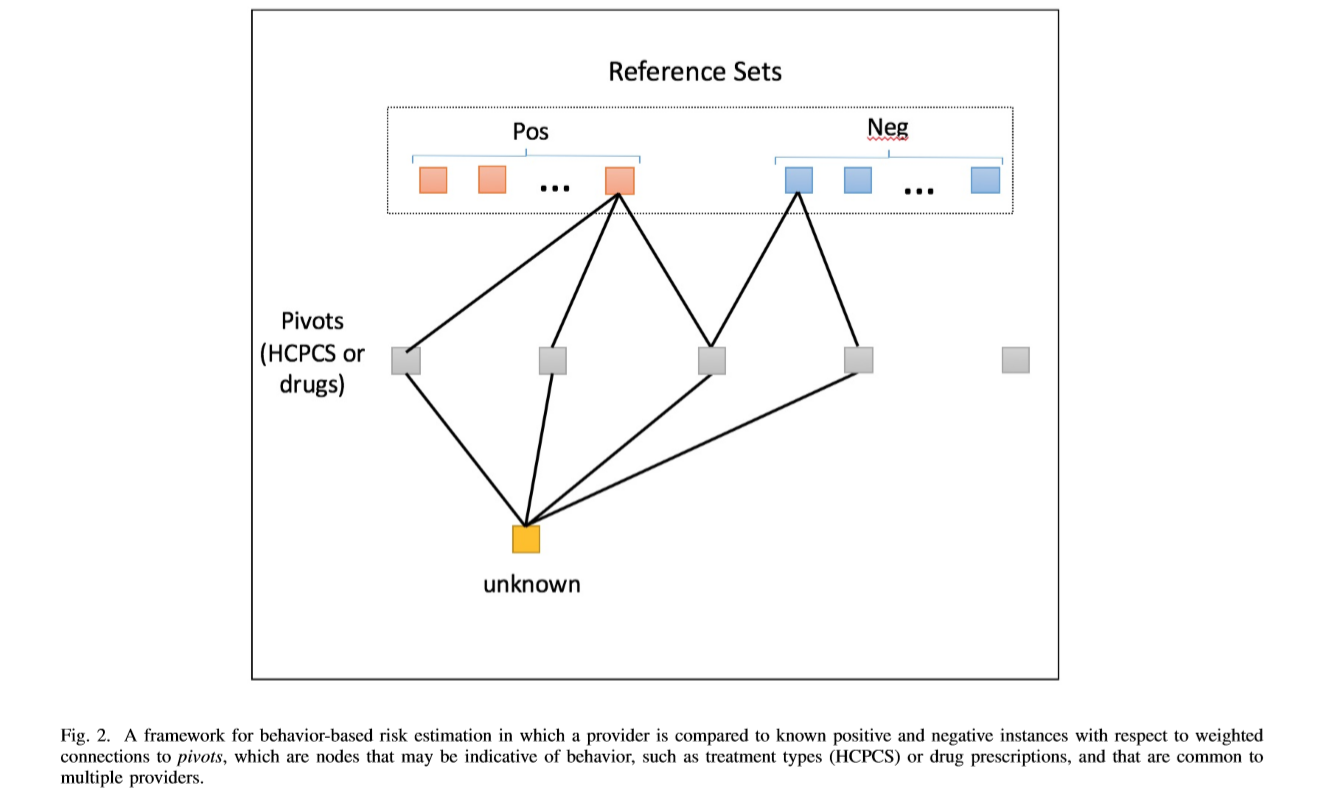
## 3 – 3 معرفی معیار های مشابهت/ ویژگی های مبتنی بر گراف

## 3 – 4 محاسبه مشابهت بردارهای رفتار دو ارایه دهنده

**3 – 5 تعیین برچسب Punknown**

هر ارائه دهنده خدمات پزشکی، هر تجویز دارویی و هرکد درمانیHealthcare Common Procedure Coding System (HCPCS)بعنوان یک نود جدا نمایش داده شد. ارائه دهندگان به هر دارویی که توسط آنها تجویز شده باشد ازطریق یک یال PRESCRIBED ( تجویز شده ) پیوند خورده اندکه صفات ( ویژگیهای ) آن، تعداد بیمارانی که توسط ارائه دهنده نسخه دارویی دریافت کرده اند ،کل هزینه تجویزها و سایر داده های Part-D را نمایش میدهد. ارائه دهندگان به هر HCPCS ( کد درمان ) از طریق یک یال CHARGE \_OF ( سرانه خدمات درمان ) پیوند خورده اند که ویژگی های آن شامل متوسط ( میانگین ) پرداخت مدیکر ( بیمه درمانی سالمندان ) ، تعداد ذینفعان منحصربفرد ( unique / غیر تکراری ) و سایر اطلاعات درمانی است که در قالب ستون ها در دیتاست PUF ظاهر میشوند.

ما هر ارائه دهنده ناشناس را با هر عضو از مجموعه های مرجع مثبت ( دارای محرومیت ) و منفی ( فاقد محرومیت ) مقایسه میکنیم. شباهت هر جفت از ارائه دهندگان ، بنوبه خود ، از طریق مقایسه نودهای محوری مجاور هریک تعیین میشود در جائیکه محورها یا مشخص کننده نوع درمان (HCPCS ) هستند یا نوع دارو .



برای هر ارائه دهنده ( ناشناس یا عضو مجموعه مرجع ) ما یک بردار رفتار تعریف میکنیم که دربرگیرنده برداری از وزن یالهای رابط به هر نود محوری مجاور است. مقدار هر عبارت متناظر با یک محور مدنظر در بردار صفر است اگر نود ارائه دهنده هیچ یالی به آن محور نداشته باشد و در غیر اینصورت ( بین دو نود یال وجود داشته باشد ) برابر ارزشی محاسبه شده به روشی خاص است که به نوع محور وابسته است. در این تحقیق ، وزن یک یال ( رابط ) به یک نود نوع دارو ، کل هزینه دارو (TotalDrugCost ) است. به معنی کل مقدار فاکتورشده (صورت حساب برای آن صادرشده) توسط این ارائه دهنده برای این داروی خاص ( این مقدار یک ستون از دیتاست Part-D است ). وزن یک یال ( رابط ) به یک نود نوع درمان (HCPCS) بصورت حاصلضرب AvgMedicarePayment و BeneUniqCount محاسبه میشود که بترتیب نمایانگر متوسط پرداخت مدیکر ( میانگین پرداخت های صورت گرفته برای بهره مندی از خدمات درمانی مدیکر به این ارائه دهنده ) و تعداد ذینفعان منحصربفرد ( تعداد افراد تحت تکفل این نوع از بیمه درمانی که مشتری این ارائه دهنده اند بدون شمارش تکرارها) است. شباهت میان هر جفت ارائه دهنده با توجه به شکلی از رفتار مدنظر ( همچون تجویز دارو یا صدرو فاکتور ( صورت حساب ) ) میتواند بوسیله اعمال یک معیار شباهت برداری استاندارد بر بردارهای رفتاری ارائه دهندگان برآورد گردد ( تخمین زده شود ).

معیار شباهت: cosine

برای تبدیل کسینوس وزن دهی شده با معکوس درجه ، بین یک ارائه دهنده و عضو مجموعه های مرجع مثبت یا منفی به ویژگی های قابل استفاده جهت تخمین ( برآورد) ریسک ، ما میانگین K نزدیکترین عضو هر مجموعه را گرفته و سه ویژگی تولید میکنیم :

مشابهت منفی (negative-similarity) : میانگین شباهت بهK نزدیکترین عضو مجموعه مرجع منفی

مشابهت مثبت (positive-similarity) : میانگین شباهت بهK نزدیکترین عضو مجموعه مرجع مثبت

نسبت مشابهت منفی و مثبت (neg-sim-ratio ) : نسبت ویژگی های 1 و 2 بهم

تکرار این محاسبات برای هر دو نوع محور ، 6 ویژگی مقداردهی شده واقعی تولید میکند :

1. مشابهت منفی HCPCS

2. مشابهت مثبت HCPCS

3. نسبت مشابهت HCPCS ( کد درمان )

4. مشابهت منفی دارو

5. مشابهت مثبت دارو

6. نسبت مشابهت دارو

:Provider attributes

7. تعداد بیماران جدید به نسبتذی نفع ( new-patients-per-bene) : نسبت سرویس های ارائه شده به بیماران جدید. (باورکلی اینست که ارائه دهنده ای که تعداد زیادی خدمات به هر بیمار جدید ارائه میدهد، ممکن است سرویس های بیش از ضرورت های درمانی ارائه دهد.)

ما بر روی 12000 ارائه دهنده محرومیت دار که هر کدام دستکم یک تجویز دارویی ، خدمت درمانی یا آدرس ( یعنی یک لینک LOCATED AT ، CHARGE OF یا PRESCRIBED ) ازمایش انجام می دهیم. ارائه دهندگان فاقد این لینک ها ( پیوندها ) از گراف جدا میشوند و بنابراین درمعرض تجزیه تحلیل های گراف قرار ندارند(حذف می شوند).

در انجام اینکار ما محورهای با درجه راس بزرگتر از 100000 را از محاسبه تشابه بردار رفتار حذف میکنیم ( با این فرض که محورهایی با اینچنین درجه بالا، توانایی تفکیک اندکی داشته و محاسبه را با کندی مواجه میسازند).

نتایج اجرای یک الگویتم درخت تصمیم با ورودی ویژگی های فوق به صورت زیر است:

|  |  |  |
| --- | --- | --- |
| ROC area | f-measure | Feature set |
| 0.718 | 0.715 | HCPCS behavior- similarity features |
| 0.666 | 0.657 | Drug behavior similarity features |
| 0.697 | 0.697 | Provider attributes |

به جای الگوریتم درخت تصمیم j48 بنا به مقاله ی خوش گفتار از الگوریتم logistic regression استفاده می کنیم:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **AUC** | **FN** | **FP** | **TN** | **TP** | **learner** |
| 0.81554 | - | - | - | - | LRخوش گفتار |
|  |  |  |  |  | LR proposed |
| 0.62985 | - | - | - | - | LOF فلوریدا |

# فصل چهارم

# پیاده سازی و ارزیابی

در این فصل ابتدا شبه کد الگوریتم­های ارائه شده در فصل قبل و توضیحات آن­ها ارائه خواهد شد. سپس در بخش ارزیابی و آزمون، مجموعه داده­ی استفاده شدهبرای ارزیابی معرفی و نتایج ارزیابی و مقایسه سیستم پیشنهادی با چندتحقیق دیگر بیان می شود.

سخت افزار مورد استفاده برای پیاده سازی این سیستم، پردازنده؟؟؟؟ با فرکانس ؟؟؟ گیگا هرتز و حافظه اصلی ؟؟؟ گیگا بایت میباشد و برنامه نویسی در ؟؟؟؟ صورت گرفته است.

## 1-4 پیاده سازی سیستم در قسمت تشکیل گراف / بردار رفتار / ترکیب دیتاست ها

|  |
| --- |
| Algorithm: Input:  Output: |
|  |

شبه کد تشکیل گراف / بردار رفتار

## 4 - 2پیاده سازی سیستم در قسمت محاسبه مشابهت بردار های رفتار

|  |
| --- |
| Algorithm: Input:  Output: |
|  |

شبه کد محاسبه مشابهت بردارهای رفتار

## 4 -4آزمون و تحلیل نتایج

در این قسمت معیارهای رایج ارزیابی سیستم­های تشخیص تقلب بیان شدهاند. در انتها، روش پیشنهادی با دو مقاله­ای که در ارزیابی خود از این مجموعه داده ها استفاده کرده اند، مقایسه می­گردد.

### 4 – 4 – 1 معیارهای ارزیابی سیستمهای تشخیص تقلب

به طور کلی برای اندازهگیری کارایی الگوریتمهای تشخیص تقلب از ؟؟؟؟؟ و ؟؟؟؟ استفاده می شود.

معیار دقت به این شکل تعریف میشود: تعداد نفوذهای درست تشخیص داده شده، تقسیم بر کل نفوذهای شناسایی شده (درست یا غلط). معیار فراخوانی نیز عبارتست از: تعداد نفوذهای درست تشخیص داده شده، تقسیم بر کل نفوذهایی که در مجموعه دادگان وجود دارد . روابط 4-1 و 4-2 نحوه محاسبه این دو معیار را نشان می­دهد.

دیتاست part D :

مجموعه داده های بخش D اطلاعات مربوط به داروهای تجویزی را كه تحت برنامه ی تجویز داروی Medicare Part D در طی یك سال خاص تجویز می شوند ، در اختیار قرار می دهد. در حال حاضر ، این داده برای سالهای تقویم 2013 تا 2015 (داده های هر سال دو سال بعد منتشر می شوند) در وب سایت CMS در دسترس هستند. پزشکان با استفاده از NPI منحصر به فرد خود در داده ها شناسایی می شوند در حالی که هریک از داروها با استفاده از نام تجاری و نام عمومیشان برچسب می خورند. سایر اطلاعات شامل متوسط پرداخت ها و هزینه ها ، متغیرهایی که مقدار داروی تجویز شده و تخصص پزشکان را توصیف می کنند، می باشند.هر سطر در مجموعه داده های بخشD ، NPI یک ارائه دهنده و نام دارو را به همراه اطلاعات خاص (مانند تعداد مطالبات) و سایر ویژگی های ثابت (مانند جنسیت) لیست می کند. تعداد کمی از پزشکان هستند که تحت تخصص های مختلفی مانند متخصص داخلی و قلب و عروق عمل می کنند.

دیتاست LEIE :

دیتاستPUF :

دیتاست NPPES:

یادگیرنده:

رگرسیون خطی:

معیارهای ارزیابی:

در ارزیابی تقلب پزشکی دو حالت در نظر می­گیریم; ارتکاب تقلب و یا عدم ارتکاب تقلب. در این تحقیق کلاس مثبت یا کلاس هدف، ارتکاب تقلب است و کلاس منفی، عدم ارتکاب تقلب است. اسپارک به ما یک ماتریس پیچیدگی برای هر مدل ارائه میدهد که معمولا برای ارزیابی کارایی یادگیرنده ها به­کار میرود. ماتریس پیچیدگی تعداد نمونه های واقعی را با تعداد نمونه های پیش بینی شده مقایسه میکند. با توجه به ماتریس نتایج، ما از AUC[67,68] برای ارزیابی کارایی تشخیص تقلب استفاده می کنیم AUC ناحیه زیرمنحنیROC(Receiver Operating Characteristic) است و ROC مقایسه بینFalse Positive وTrue Positive است. Recall از TP/TP+FN بدست می آید. تعاریف برای TP و TN و FPوFN مستقیما از ماتریس پیچیدگی به صورت زیر بدست می آیند:

* True Positive(TP) : تعداد نمونه های مثبت واقعی که به درستی مثبت پیش بینی شده اند.
* True Negative(TN) :تعداد نمونه های منفی واقعی که به درستی منفی پیش بینی شده اند.
* False Positive(FP) :تعداد نمونه های منفی واقعی که به اشتباه مثبت پیش بینی شده اند.
* False Negative(FN) : تعداد نمونه های مثبت واقعی که به اشتباه منفی پیش بینی شده اند.

منحنیAUC یک ارزیابی فراگیر از یادگیرنده است که ارزیابی را در سراسر آستانه های تصمیم گیری به تصویر می کشد. نتایج AUC در یک بازه بین 0 تا 1 هستند که یک طبقه بند خوب دارای مقدار 1 در AUC می باشد، طبقه بند تصادفی مقدار 0.5 دارد و مقادیر AUC کمتر از 0.5 نشان دهنده ی بایاس به سمت یک کلاس خاص است ثابت شده است که AUCبرای نامتوازن بودن کلاس موثر است.[200]

[200] Jeni, László A., Jeffrey F. Cohn, and Fernando De La Torre. "Facing imbalanced data--recommendations for the use of performance metrics." 2013 Humaine association conference on affective computing and intelligent interaction. IEEE, 2013.

اعتبارسنجی متقابل:

در این تحقیق اعتبارسنجی متقابل به k دسته طبقه بندی شده را در ارزیابی مدل هایمان استفاده کردیم و k=5.

طبقهبندیتضمینمیکندکههمهfold هاداراینمایشکلاسمطابقبانسبتدادههایاصلیهستند،کهاین مساله هنگاممواجههبادادههایبسیارنامتوازنمهماست. دادههایآموزشبهطورمساویبهپنجبخشتقسیممیشوندکه چهارمرحلهی آن برایآموزشمدلاستفادهمیشودوبقیه برای تست استفاده می شوند. اینفرایند 5 بارتکرارمیشودکهبه هرfold این فرصت را می دهد تابهعنوانتست باشدو در نتیجهتضمین میکندکهکلمجموعهدادهدرآموزشواعتبارسنجیاستفادهمیشود.Sparkبهطورخودکارهربارکهیادگیرندهاجراشود،foldهایمختلفیایجادمیکندوبرایاعتبارسنجینتایج،هرمدلرا 10 باربرایهریادگیرنده / مجموعهدادهاجرا می کند.استفاده از تکرارها به کاهش بایاس کمک می کند برای کاهش پراکندگی، عمل اعتبارسنجی چندین بار با افرازهای مختلف انجام و از نتایج اعتبارسنجی‌ها میانگین گرفته می‌شود.

### 4 – 4- 2 روش آزمون

**؟؟؟؟؟؟**

### 4 – 4 – 3 ارزیابی روش پیشنهادی

همانطور که گفته شد در قسمت ارزیابی، معیارهای قابل توجه در تشخیص مبتنی بر گراف تقلب را در مورد روش پیشنهادی محاسبه و سپس نتایج حاصل را با نتایج حاصل از پژوهش[]مقایسه می­کنیم.

#### *4 – 4- 4-1 ارزیابی روش پیشنهادی با استفاده از []*

در این تحقیق برای تقسیم داده و ارزیابی از روش 10-fold cross validation لستفاده شده است.

Table :Cross Validation results of …

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| MSE | ERR | BER | Acc. | Prec. | FNR | FPR | TNR | TPR | Num |
| 1.001 | 0.006 | 0.452 | 0.994 | 0.040 | 0.901 | 0.004 | 0.996 | 0.099 | CV1 |
|  |  |  |  |  |  |  |  |  | CV2 |
|  |  |  |  |  |  |  |  |  | CV3 |
|  |  |  |  |  |  |  |  |  | CV4 |
|  |  |  |  |  |  |  |  |  | CV5 |
|  |  |  |  |  |  |  |  |  | Mean |

Table : Test result of Self-organized

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| AUC | MSE | ERR | BER | Acc. | Prec. | FNR | FPR | TNR | TPR |  |
|  |  |  |  |  |  |  |  |  |  | T1 |
|  |  |  |  |  |  |  |  |  |  | T2 |
|  |  |  |  |  |  |  |  |  |  | T3 |
|  |  |  |  |  |  |  |  |  |  | Mean |

مقایسه­نتایجبهدستآمدهازاینمقالهوروشارائهشدهدراینپژوهش،به صورتنمودارهاوجدولزیرآوردهشدهاست.

مقایسهنتایجالگوریتم­های مختلف

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Specificity | Sensivity | AUC | MSE | ERR | BER | Acc. | Prec. | FNR | FPR | TNR | TPR | Algorithm |
| - | - | 0.800 | 1.001 | 0.006 | 0.452 | 0.994 | 0.040 | 0.901 | 0.004 | 0.996 | 0.099 | SOM |
| 0.676 | 0.536 | 0.629 |  |  |  |  |  |  |  |  |  | LOF40 |
| 0.679 | 0.497 | 0.613 |  |  |  |  |  |  |  |  |  | KNN1 |
| 0.645 | 0.527 | 0.603 |  |  |  |  |  |  |  |  |  | URF100 |
| 0.650 | 0.463 | 0.555 |  |  |  |  |  |  |  |  |  | AE50\_Tanh |
| 0.436 | 0.712 | 0.554 |  |  |  |  |  |  |  |  |  | IF100 |

روش هایی که در تحقیق [] به آن پرداخته شده روشLOF نسبت به سایر روش­ها بهتر عمل کرده زیرا ؟؟؟؟

از طرفی همان­گونه که مشاهده می­شود در مقایسه روش پیشنهادی با آن­ها، روش پیشنهادی در بیشتر معیارها بهتر عمل کرده و بهبود داشته است.بنابراینروشپیشنهادی،نسبتبههردوروشاشارهشدهدراینمقاله،دارایبهبودبودهاست. کهاینبهبودمیتواندبهدلیل؟؟؟؟ باشد. همچنین در روش پیشنهادی ؟؟؟؟ نیز مورد توجه قرار گرفته است در حالیکه در[]تنها به ؟؟؟؟ توجه شده است. همچنین از مزایای روش پیشنهادی نسبت به [] میتوان ؟؟؟؟ را بیان نمود.

Fig :AUC curve –روش پیشنهادی10,20,40,80

Fig :AUC curve –All models 1

## 4 – 5 جمع بندی

دراینفصل، ابتدا پیاده­سازی مراحل مختلف الگوریتم ارائهشد. سپسبهارزیابیسیستمپرداختهشدکهبرایاینمنظور،مجموعهدادهای تهیه شد زیرا با توجه به اینکهمجموعهدادهیاستانداردیکهشاملاطلاعاتبیشترودقیقترینسبتبهدیتاست های عمومی پراکنده در دسترس باشدوامکانمقایسهوارزیابیمعتبرروشهایپیشنهادیبرایاینمسألهرافراهمکند وجود نداشت ؛چراکههمانگونهکهاشارهشد،یکیازچالشهایمهماینمسأله،عدموجودمجموعهدادهایاستانداردبرایارزیابیومقایسهیروشهایارائهشدهبود.

سپسدر ادامه معیارهایارزیابیرایججهتارزیابیسیستمهایتشخیصتقلب پزشکیبیان شد. در پایاننتایجمقایسهروشپیشنهادیویکمقالهکهازاینمجموعهدادهبرایارزیابیاستفادهکردهاند در قالب جدول و نمودارآوردهشد. نتایجدر تماممعیارهابهترعملکردهاستامادرمقایسه با روش ؟؟؟؟ تنها در معیارهای ؟؟؟؟ قوی تر عمل کرد.

table : description of PUF dataset

|  |  |  |
| --- | --- | --- |
| feature | description | Type(categorical/ numerical) |
|  |  |  |
|  |  |  |
|  |  |  |

table : description of NPPES dataset

|  |  |  |
| --- | --- | --- |
| feature | description | Type(categorical/ numerical) |
|  |  |  |
|  |  |  |
|  |  |  |

table : description of part D dataset

|  |  |  |
| --- | --- | --- |
| feature | description | Type(categorical/ numerical) |
|  |  |  |
|  |  |  |
|  |  |  |

sample of Part D dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| NPI | Provider-type | Drug-name | Total-drug-cost | Total-claim-count |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

sample of LEIE dataset

|  |  |  |
| --- | --- | --- |
| NPI | Excltype | Excludate |
|  | 1128b6 | 20091220 |
|  | 1128a1 | 20170720 |
|  | 1128b4 | 20150614 |

LEIE rules involving fraud

|  |  |
| --- | --- |
| Rule number | Description |
| 1128(a)(1) | Conviction of program-related crimes |
| 1128(a)(2) | Conviction relating to patient abuse or neglect |
| 1128(a)(3) | Felony conviction relating to health care fraud |
| 1128(b)(4) | License revocation or suspension |
| 1128(b)(7) | Fraud,kickbacks and other prohibited activities |
| 1128(c)(3)(g)(i) | Conviction of two mandatory exclusion offenses 10 years |
| 1128(c)(3)(g)(ii) | Convection of 3 mandatory exclusion offenses indefinite |

# فصل پنجم

# نتیجه گیری و توسعه های آتی

دراینفصلبهبیانخلاصهوجمع­بندیمطالعاتانجامشدهونتایجبهدستآمدهدراینپژوهشپرداختهشدهاستوهمچنینپیشنهاداتیبرایتوسعههایآتیجهتبهبودسیستمارائهگشتهاست.

## 5 – 1نتیجه گیری

درابتدایاینپژوهش،مفاهیممربوطبه تقلب و مطالبات پزشکی،پایگاهداده گرافی ومفاهیمیازایندست معرفیشدند. سپسسیستمیبرایتشخیصتقلب مبتنی بر گراف ارائهشدکهاجزاءآنبهترتیبهمراهبافلوچارتونمودارهایمربوطبهآنهاموردشرحوبررسیقرارگرفت. درروشارائهشده،ازیکیازمفاهیممهمدادهکاویبهنام استخراج ویژگی مبتنی بر گرافبهرهگرفتهشدهاستوسعیشدهبا بهره گیری از big data واستخراج ویژگی های مبتنی برگراف و همچنینطرحمعیاریبرایبیانشباهتبینمطالبات پزشکیراهکاریجهتتشخیصتقلب پزشکیارائهشود. رویکرداصلیروشارائهشدهدراینپژوهش،براینمبنابودکهدو مجموعه مرجع مثبت ( دارای محرومیت ) و منفی ( فاقد محرومیت ) ایجاد نموده و هر ارائه دهنده ناشناس را با هر عضو از این دو مجموعه مقایسه میکنیم. شباهت هر جفت از ارائه دهندگان ، بنوبه خود ، از طریق مقایسه نودهای محوری مجاور(محورها یا مشخص کننده نوع درمان (HCPCS ) هستند یا نوع دارو) تعیین میشود. . برای بدست آوردن شباهت میان هر جفت ارائه دهنده با توجه به شکلی از رفتار مدنظر ( همچون تجویز دارو یا صدرو فاکتور ( صورت حساب ) ) یک معیار شباهت برداری استانداردcosine بر بردارهای رفتاری ارائه دهندگان اعمال شد.

درادامه،نحوهپیاده­سازیسیستمپیشنهادیبااستفادهازتوضیحاتوشبهکدالگوریتم­هابیانگردید. جهتآزمونوارزیابیسیستمبیانگردیدکهمجموعهداده ی عمومی شامل برچسبی برایارزیابیروشهایاینمسألهطراحینشدهاست و از ترکیب چند دیتاست و تطبیق قوانین محرومیت پزشکان و سایر موارد به برچسب داده ها رسیدیم. درادامه،معیارهایمهمورایجیکهدرمسألهتشخیصتقلبمورداستفادهقرارمی­گیرند،بیانشدند.درنهایت،روشارائهشدهباروشمقاله ایدیگرکهازبخشی از دیتاستاشارهشدهبرایارزیابیخوداستفادهکردهاند،درمعیارهایمطرحشده،موردمقایسهوارزیابیقرارگرفت. نتایجنشاندادکهروشارائهشدهنسبتبهروش های قبلی در معیارهای؟؟؟؟ بهبود داشته است.درادامه،تعدادیازنقاطقوتونقاطضعفسیستمپیشنهادیکهبهذهنمیرسدبیانخواهدشد. ازنقاطقوتروشپیشنهادینسبتبهروشهایمشابه،میتوانبهمواردیاشارهنمود؛ازجملهاینکه؟؟؟؟. اماازمزایایروشپیشنهادینسبتبهروشهایدیگرکهازروشیغیرازتحلیل گرافاستفادهمیکنند،نیزمیتوانبهمواردیاشارهکرد؛ازجملهاینکه؟؟؟؟.دردراینروش،ازآنجاکه؟؟؟؟؛چراکهتقلب هاهمگیشناختهشدهنمیباشندوحتیممکناستمطالباتیظاهرانرمالبهنظربرسدامابانمایش گرافی و نحوه ارتباطات آن­ها بامطالباتدیگر، تقلبمحسوبگردد. درفازتشخیصنیزازآنجاکه؟؟؟؟است،بادقتوسرعتبالاتریعملمیکند.

ازنقاطضعفاینروشمیتوانبهایناشارهکردکهاگرمطالبه ایبهکراتاتفاقبیفتد،ممکناستنرمالدرنظرگرفتهشود. علاوهبراین،؟؟؟؟؟،کهبرایاینمشکل،درقسمتتوسعه­هایآتی،پیشنهادیارائهمیگردد. در اینروش به دلیل عدم انتشار دیتاست های مراکز پزشکی و عدم دسترسی به نمونه های برچسب دار، از ترکیب چند دیتاست عمومی استفاده شده ودرنتیجهحافظه­بسیاریبرای پردازشاینکار نیاز است.

## 5 – 2 توسعه های آتی

دراینقسمت،پیشنهاداتیبیانمیشوندکهممکناستدرپژوهشهایبعدی،موجببهبودراهحلگردند. یکیازاینپیشنهاداتایناستکه؟؟؟؟؟.

نکته­یمهمدیگریکهدرتوسعههایآتی،بهطورکلیدرموردمسألهیتشخیصتقلب پزشکیمیتواندموردتوجهقرارگیرد،؟؟؟؟؟.

# مراجع

[1] Johnson, Marina Evrim, and Nagen Nagarur. "Multi-stage methodology to detect health insurance claim fraud."*Health care management science* 19.3 (2016): 249-260.

[2] Sadeghian‬‬ N,Hosseini‬‬ M H (2016) .”Assessment‬‬ and‬‬ recognition‬‬ the‬‬ trueness‬‬ of‬‬ the‬‬ assurance‬‬ claims‬‬ using data‬‬ mining‬‬ techniques‬‬ based‬‬ on‬‬ the‬‬ supervised‬‬ learning” .M.Sc.‬‬ Thesis‬‬ in‬‬ Industrial‬‬ Management‬‬,Shahrood University‬‬.

[3] Li, Jing, et al. "A survey on statistical methods for health care fraud detection."*Health care management science* 11.3 (2008): 275-287.

[4] Savedoff‬‬ and‬‬ Hussmann‬‬.Chapter‬‬ 1,‬‬ The‬‬ causes‬‬ of‬‬ corruption‬‬ in‬‬ the‬‬ health‬‬ sector,‬‬ in‬‬ Transparency‬‬ International‬‬ (ed.),‬‬ Global‬‬ Corruption‬‬ Report‬‬ 2006.‬‬ London:‬‬ Pluto‬‬ Press,‬‬ p.‬‬

[5] Phua, Clifton, et al. "A comprehensive survey of data mining-based fraud detection research."*arXiv preprint arXiv:1009.6119* (2010).

[6] Wang, Jau-Hwang, et al. "Technology-based financial frauds in Taiwan: issues and approaches."*Systems, Man and Cybernetics, 2006. SMC'06. IEEE International Conference on*. Vol. 2. IEEE, 2006.

[7] Bose, Indranil, and Radha K. Mahapatra. "Business data mining—a machine learning perspective."*Information & management* 39.3 (2001): 211-225.

[8] Turban, Efraim, Ramesh Sharda, and Dursun Delen. "Decision Support and Business Intelligence Systems (required)."*Google Scholar* (2010).

[9] Frawley, William J., Gregory Piatetsky-Shapiro, and Christopher J. Matheus. "Knowledge discovery in databases: An overview."*AI magazine* 13.3 (1992): 57.

[10] Kou, Yufeng, et al. "Survey of fraud detection techniques."*Networking, sensing and control, 2004 IEEE international conference on*. Vol. 2. IEEE, 2004.

[11] Phua, Clifton, et al. "A comprehensive survey of data mining-based fraud detection research."*arXiv preprint arXiv:1009.6119* (2010).

[12] Duman, Ebru Aydoğan, and Şeref Sağıroğlu. "Heath care fraud detection methods and new approaches."*Computer Science and Engineering (UBMK), 2017 International Conference on*. IEEE, 2017.

[13] Legotlo, T. G., and A. Mutezo. "Understanding the types of fraud in claims to South African medical schemes."*South African Medical Journal* 108.4 (2018): 299-303.

[14] Travaille, Peter. *Electronic Fraud Detection in the US Medicaid Health Care Program*. MS thesis. University of Twente, 2011.

[15] Manjula, B., et al. "DFFS: Detecting Fraud in Finance Sector."*Advanced Engineering Sciences and Technologies* 9.2 (2011): 178-182.

[16] Abdallah, Aisha, Mohd Aizaini Maarof, and Anazida Zainal. "Fraud detection system: A survey."*Journal of Network and Computer Applications* 68 (2016): 90-113.

[17] Abbass, Hussein A., et al. "Online adaptation in learning classifier systems: stream data mining."*Urbana* 51 (2004): 61801.

[18] Malekian, Donia, and Mahmoud Reza Hashemi. "An adaptive profile based fraud detection framework for handling concept drift."*Information Security and Cryptology (ISCISC), 2013 10th International ISC Conference on*. IEEE, 2013.

[19] Gama, João, et al. "A survey on concept drift adaptation."*ACM computing surveys (CSUR)* 46.4 (2014): 44.

[20] Bolton, Richard J., and David J. Hand. "Unsupervised profiling methods for fraud detection."*Credit Scoring and Credit Control VII* (2001): 235-255.

[21] Zhu, Xiaojin, and Andrew B. Goldberg. "Introduction to semi-supervised learning."*Synthesis lectures on artificial intelligence and machine learning* 3.1 (2009): 1-130.

[22] Liu, Qi, and Miklos Vasarhelyi. "Healthcare fraud detection: A survey and a clustering model incorporating Geo-location information."*29th World Continuous Auditing and Reporting Symposium (29WCARS), Brisbane, Australia*. 2013.

[23] Hilas, Constantinos S., and John N. Sahalos. "An application of decision trees for rule extraction towards telecommunications fraud detection."*International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, Berlin, Heidelberg, 2007.

[24] Viaene, Stijn, Richard A. Derrig, and Guido Dedene. "A case study of applying boosting Naive Bayes to claim fraud diagnosis."*IEEE Transactions on Knowledge and Data Engineering* 16.5 (2004): 612-620.

[25] Lane, Terran, and Carla E. Brodley. "Temporal sequence learning and data reduction for anomaly detection."*ACM Transactions on Information and System Security (TISSEC)* 2.3 (1999): 295-331.

[26] Han, Jiawei, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[27] Brockett, Patrick L., et al. "A comparison of neural network, statistical methods, and variable choice for life insurers' financial distress prediction."*Journal of Risk and Insurance* 73.3 (2006): 397-419.

[28] Ai, Jing, Patrick L. Brockett, and Linda L. Golden. "Assessing consumer fraud risk in insurance claims: An unsupervised learning technique using discrete and continuous predictor variables."*North American Actuarial Journal* 13.4 (2009): 438-458.

[29] Almendra, Vinicius, and Denis Enachescu. "A supervised learning process to elicit fraud cases in online auction sites."*2011 13th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*. IEEE, 2011.

[30] Sánchez-Maroño, Noelia, Amparo Alonso-Betanzos, and María Tombilla-Sanromán. "Filter methods for feature selection–a comparative study."*International Conference on Intelligent Data Engineering and Automated Learning*. Springer, Berlin, Heidelberg, 2007.

[31] Liu, Qi, and Miklos Vasarhelyi. "Healthcare fraud detection: A survey and a clustering model incorporating Geo-location information."*29th World Continuous Auditing and Reporting Symposium (29WCARS), Brisbane, Australia*. 2013.

[32] Viaene, Stijn, Richard A. Derrig, and Guido Dedene. "A case study of applying boosting Naive Bayes to claim fraud diagnosis."*IEEE Transactions on Knowledge and Data Engineering* 16.5 (2004): 612-620.

[33] Islam, Md, et al. "A Systematic Review on Healthcare Analytics: Application and Theoretical Perspective of Data Mining."*Healthcare*. Vol. 6. No. 2. Multidisciplinary Digital Publishing Institute, 2018.

[34] Kusiak, Andrew, et al. "Hypoplastic left heart syndrome: knowledge discovery with a data mining approach."*Computers in biology and medicine* 36.1 (2006): 21-40.

[35] Nicholas, Taleb Nassim. "The black swan: The impact of the highly improbable."*Random: New York, NY, USA* (2007).

[36] Cercone, Nick, et al. "Finding best evidence for evidence-based best practice recommendations in health care: the initial decision support system design."*Knowledge and information systems* 29.1 (2011): 159.

[37] Huang, Yue, et al. "Feature selection and classification model construction on type 2 diabetic patients’ data."*Artificial intelligence in medicine* 41.3 (2007): 251-262.

[38] Hachesu, Peyman Rezaei, et al. "Use of data mining techniques to determine and predict length of stay of cardiac patients."*Healthcare informatics research* 19.2 (2013): 121-129.

[39] Santos, Ricardo S., et al. "A data mining system for providing analytical information on brain tumors to public health decision makers."*Computer methods and programs in biomedicine* 109.3 (2013): 269-282.

[40] Shen, Chia-Ping, et al. "A data-mining framework for transnational healthcare system."*Journal of medical systems* 36.4 (2012): 2565-2575.

[41] Duan, Lian, W. Nick Street, and E. Xu. "Healthcare information systems: data mining methods in the creation of a clinical recommender system."*Enterprise Information Systems* 5.2 (2011): 169-181.

[42] Hassanzadeh R, Nayak R, Stebila D. Analyzing the effectiveness of graph metrics for anomaly detection in online social networks. In: Web Information systems Engineering-WIsE 2012. Springer Berlin Heidelberg. 2012; 7651: p.624-30.

[43] Mookiah L, Eberle W, Holder L. Discovering Suspicious Behavior Using Graph-Based Approach. In: The Twenty-Eighth International Flairs Conference Florida. 2015 Jun: p. 428-33.

[44] Moradi F, Olovsson T, Tsigas P. Overlapping Communities for Identifying Misbehavior in Network Communications. In: Advances in Knowledge Discovery and Data Mining. Springer

International Publishing. Taiwan. 2014; 8443: 398-409.

[45] Perozzi B, Akoglu L, Sánchez IP, Müller E. Focused clustering and outlier detection in large attributed graphs. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. USA. 2014: p. 1346-55.

[46] Sensarma, Debajit, and Samar Sen Sarma. "A survey on different graph based anomaly detection techniques."*Indian Journal of Science and Technology* 8.31 (2015).

[47] Sun J, Xie Y, Zhang H, Faloutsos C. Less is more: Sparse graph mining with compact matrix decomposition. Statistical Analysis and Data Mining: The ASA Data Science Journal. 2008; 1(1):6-22.

[48] Aggarwal CC, Zhao Y, Yu PS. Outlier detection in graph streams. In: Data Engineering (ICDE). 2011 IEEE 27th International Conference on IEEE. Hannover. 2011; p. 399-409.

[49] Heard NA, Weston DJ, Platanioti KH and D J. Bayesian anomaly detection methods for social networks. The Annals of Applied Statistics. 2010; 4(2): 645-62.

[50] Eberle W, Holder L. A partitioning approach to scaling anomaly detection in graph streams. Big Data (Big Data),2014. IEEE International Conference. IEEE. Washington DC. 2014; p. 17-24.

[51]Lu, Fletcher, and J. Efrim Boritz. "Detecting fraud in health insurance data: Learning to model incomplete Benford’s law distributions."*European Conference on Machine Learning*. Springer, Berlin, Heidelberg, 2005.

[52] Hill, Theodore P. "A statistical derivation of the significant-digit law."*Statistical science* 10.4 (1995): 354-363.

[53] Carslaw, Charles APN. "Anomalies in income numbers: Evidence of goal oriented behavior."*Accounting Review* (1988): 321-327.

[54] Busta, Bruce, and Randy Weinberg. "Using Benford’s Law and neural networks as a review procedure."*Managerial Auditing Journal* 13.6 (1998): 356-366.

[55] Fawcett, Tom. *AI Approaches to Fraud Detection & Risk Management*. AAAI press, 1997.

[56] Bolton, Richard J., and David J. Hand. "Statistical fraud detection: A review."*Statistical science* (2002): 235-249.

[57] Liu, Juan, et al. "Graph analysis for detecting fraud, waste, and abuse in healthcare data."*AI Magazine* 37.2 (2016): 33-46.

[58] Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou. "Isolation forest."*2008 Eighth IEEE International Conference on Data Mining*. IEEE, 2008.

[59] Landon, Bruce E., et al. "Using administrative data to identify naturally occurring networks of physicians."*Medical care* 51.8 (2013): 715.

[60] Sauter, Simone Katja, et al. "Analyzing healthcare provider centric networks through secondary use of health claims data."*Biomedical and Health Informatics (BHI), 2014 IEEE-EMBS International Conference on*. IEEE, 2014.

[61] Mandl, Kenneth D., et al. "Provider collaboration: cohesion, constellations, and shared patients."*Journal of general internal medicine* 29.11 (2014): 1499-1505.

[62] Wang, Fei, et al. "Application of network analysis on healthcare."*Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE Press, 2014.

[63] Uddin, S., and L. Hossain. "Social networks in exploring healthcare coordination."*Asia Pacific Journal of Health Management* 9.3 (2014): 53.

[64] Uddin, Shahadat. "Exploring the impact of different multi-level measures of physician communities in patient-centric care networks on healthcare outcomes: A multi-level regression approach."*Scientific reports* 6 (2016): 20222.

[65] Almansoori, Wadhah, et al. "Applications of social network construction and analysis in the medical referral process."*Dependable, Autonomic and Secure Computing (DASC), 2011 IEEE Ninth International Conference on*. IEEE, 2011.

[66] Guo, Hao, et al. "Find referral social networks."*Security and Privacy in Social Networks and Big Data (SocialSec), 2015 International Symposium on*. IEEE, 2015.

[67] Ruhnau, Britta. "Eigenvector-centrality—a node-centrality?."*Social networks* 22.4 (2000): 357-365.

[68] Barthelemy, Marc. "Betweenness centrality in large complex networks."*The European physical journal B* 38.2 (2004): 163-168.

[69] Cunningham, Frances C., et al. "Health professional networks as a vector for improving healthcare quality and safety: a systematic review."*BMJ Qual Saf* 21.3 (2012): 239-249.

[70] Li, Jing, et al. "A survey on statistical methods for health care fraud detection."*Health care management science* 11.3 (2008): 275-287.

[71] Liu, Qi, and Miklos Vasarhelyi. "Healthcare fraud detection: A survey and a clustering model incorporating Geo-location information."*29th World Continuous Auditing and Reporting Symposium (29WCARS), Brisbane, Australia*. 2013.

[72] Cahill, M.H., Lambert, D., Pinheiro, J.C., and Sun, D.X. (2002), “Detecting Fraud in the Real World,” Handbook of massive data sets book contents, pp. 911 – 929

[73] Ferreira, P., Alves, R., Belo, O., and Cortesao, L. (2006), “Establishing Fraud Detection Patterns Based on Signatures,” New trends in Artificial Intelligence, 13 th Portuguese Conference on Artificial Intelligence EPIA 2007, pp. 428 – 440

[74] Fawcett, T. and Provost, F. (1999), “Activity monitoring: Noticing interesting changes in behavior,” Association for Computing Machinery

[75] Bolton, R.J. and Hand, D.J. (2001), “Unsupervised profiling Methods for Fraud Detection,” Technical Report, department of Mathematics, Imperial College, London

[76] Liou, Fen-May, Ying-Chan Tang, and Jean-Yi Chen. "Detecting hospital fraud and claim abuse through diabetic outpatient services."*Health care management science* 11.4 (2008): 353-358.

[77] Ortega, Pedro A., Cristián J. Figueroa, and Gonzalo A. Ruz. "A Medical Claim Fraud/Abuse Detection System based on Data Mining: A Case Study in Chile."*DMIN* 6 (2006): 26-29.

[78] Padmaja, T. Maruthi, et al. "Unbalanced data classification using extreme outlier elimination and sampling techniques for fraud detection."*Advanced Computing and Communications, 2007. ADCOM 2007. International Conference on*. IEEE, 2007.

[79] Shin, Hyunjung, et al. "A scoring model to detect abusive billing patterns in health insurance claims."*Expert Systems with Applications* 39.8 (2012): 7441-7450.

[80] Major, John A., and Dan R. Riedinger. "EFD: A hybrid knowledge/statistical‐based system for the detection of fraud."*International Journal of Intelligent Systems* 7.7 (1992): 687-703.

[80] Ormerod, Thomas, et al. "Using ethnography to design a Mass Detection Tool (MDT) for the early discovery of insurance fraud."*CHI'03 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2003.

[81] He, Hongxing, Warwick Graco, and Xin Yao. "Application of genetic algorithm and k-nearest neighbour method in medical fraud detection."*Asia-Pacific Conference on Simulated Evolution and Learning*. Springer, Berlin, Heidelberg, 1998.

[82] Kumar, Mohit, Rayid Ghani, and Zhu-Song Mei. "Data mining to predict and prevent errors in health insurance claims processing."*Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010.

[83] Ekin, Tahir, et al. "Statistical Medical Fraud Assessment: Exposition to an Emerging Field."*International Statistical Review* (2018).

[84] Lin, Chinho, et al. "Intelligent physician segmentation and management based on KDD approach."*Expert Systems with Applications* 34.3 (2008): 1963-1973.

[85] Musal, Rasim Muzaffer. "Two models to investigate Medicare fraud within unsupervised databases."*Expert Systems with Applications* 37.12 (2010): 8628-8633.

[86] Liu, Qi, and Miklos Vasarhelyi. "Healthcare fraud detection: A survey and a clustering model incorporating Geo-location information."*29th World Continuous Auditing and Reporting Symposium (29WCARS), Brisbane, Australia*. 2013.

[87] Ekina, Tahir, et al. "Application of bayesian methods in detection of healthcare fraud."*chemical engineering Transaction* 33 (2013).

[88] Capelleveen, Guido Cornelis. *Outlier based predictors for health insurance fraud detection within US Medicaid*. MS thesis. University of Twente, 2013.

[89] Shan, Yin, D. Wayne Murray, and Alison Sutinen. "Discovering inappropriate billings with local density based outlier detection method."*Proceedings of the Eighth Australasian Data Mining Conference-Volume 101*. Australian Computer Society, Inc., 2009.

[90] Tang, MingJian, et al. "Unsupervised fraud detection in Medicare Australia."*Proceedings of the Ninth Australasian Data Mining Conference-Volume 121*. Australian Computer Society, Inc., 2011.

[91] Carvalho, Luiz FM, et al. "Provider-consumer anomaly detection for healthcare systems."*2017 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 2017.

[92] Aral, Karca Duru, et al. "A prescription fraud detection model."*Computer methods and programs in biomedicine* 106.1 (2012): 37-46.

[93] Iyengar, Vijay S., Keith B. Hermiz, and Ramesh Natarajan. "Computer-aided auditing of prescription drug claims."*Health care management science* 17.3 (2014): 203-214.

[94] Bauder, Richard A., and Taghi M. Khoshgoftaar. "A probabilistic programming approach for outlier detection in healthcare claims."*Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on*. IEEE, 2016.

[95] Musal, Rasim M., and Tahir Ekin. "Medical overpayment estimation: A Bayesian approach."*Statistical Modelling* 17.3 (2017): 196-222.

[96] Rao, Sanjeev, and Priyanka Gupta. "Implementing Improved Algorithm Over APRIORI Data Mining Association Rule Algorithm 1." (2012).

[97] Agarwal, Rakesh, and Ramakrishnan Srikant. "Fast algorithms for mining association rules."*Proc. of the 20th VLDB Conference*. 1994.

[98] Ji, Yanqing, et al. "Mining Infrequent Causal Associations in Electronic Health Databases."*Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*. IEEE, 2011.

[99] Patil, B. M., R. C. Joshi, and Durga Toshniwal. "Association rule for classification of type-2 diabetic patients."*Machine Learning and Computing (ICMLC), 2010 Second International Conference on*. IEEE, 2010.

[100] Abdullah, Umair, Jamil Ahmad, and Aftab Ahmed. "Analysis of effectiveness of apriori algorithm in medical billing data mining."*Emerging Technologies, 2008. ICET 2008. 4th International Conference on*. IEEE, 2008.

[101] Ilayaraja, M., and T. Meyyappan. "Mining medical data to identify frequent diseases using Apriori algorithm."*Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013 International Conference on*. IEEE, 2013.

[102] Bolton, R.J. and Hand, D.J. (2001), “Peer Group Analysis – Local Anomaly Detection in Longitudinal Data,” Technical Report, department of Mathematics, Imperial College, London

[103] Kantardzic, M. (2002), “Data mining: concepts, models, methods, and algorithms,” Wiley IEEE Press.

[104] Williams, Graham J., and Zhexue Huang. "Mining the knowledge mine."*Australian Joint Conference on Artificial Intelligence*. Springer, Berlin, Heidelberg, 1997.

[105] Williams, Graham J. "Evolutionary hot spots data mining."*Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, Berlin, Heidelberg, 1999.

Ghaemi, Mohammad Sajjad. "Class Lecture, Topic:“Clustering and Nonnegative Matrix Factorization”. DAMAS LAB."*Computer Science and Software Engineering Department, Laval University* (2013).

[107] Major, John A., and Dan R. Riedinger. "EFD: A hybrid knowledge/statistical‐based system for the detection of fraud."*International Journal of Intelligent Systems* 7.7 (1992): 687-703.

[108] WU, JIAN, et al. "Medical Insurance Fraud Recognition Based on Improved Outlier Detection Algorithm."*DEStech Transactions on Computer Science and Engineering* aiea (2017).

[109] van Capelleveen, Guido, et al. "Outlier detection in healthcare fraud: A case study in the Medicaid dental domain."*International journal of accounting information systems* 21 (2016): 18-31.

[110] Seo, Jiwon, and Ofer Mendelevitch. "Identifying frauds and anomalies in Medicare-B dataset."*Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE*. IEEE, 2017.

[111] Gangopadhyay, Aryya, and Song Chen. "Health care fraud detection with community detection algorithms."*Smart Computing (SMARTCOMP), 2016 IEEE International Conference on*. IEEE, 2016.

[112] Aral, Karca Duru, et al. "A prescription fraud detection model."*Computer methods and programs in biomedicine* 106.1 (2012): 37-46.

[113] Kose, Ilker, Mehmet Gokturk, and Kemal Kilic. "An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance."*Applied Soft Computing* 36 (2015): 283-299.

[114] Joudaki, Hossein, et al. "Improving fraud and abuse detection in general physician claims: a data mining study."*International journal of health policy and management* 5.3 (2016): 165.

[115] Herland, Matthew, Taghi M. Khoshgoftaar, and Richard A. Bauder. "Big Data fraud detection using multiple medicare data sources."*Journal of Big Data* 5.1 (2018): 29.

[116] López, Victoria, et al. "Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics."*Expert Systems with Applications* 39.7 (2012): 6585-6608.

[117] Ceri, Stefano, et al. *Morgan Kaufmann series in data management systems: Designing data-intensive Web applications*. Morgan Kaufmann, 2003.

[118] Sun, Chenfei, et al. "An effective hybrid fraud detection method."*International Conference on Knowledge Science, Engineering and Management*. Springer, Cham, 2015.

1. General Accounting Office [↑](#footnote-ref-2)
2. National HealthCare Anti-fraud Association [↑](#footnote-ref-3)
3. Actor [↑](#footnote-ref-4)
4. 4Service Provider’s Fraud

   5 conspiracy [↑](#footnote-ref-5)
5. [↑](#footnote-ref-6)
6. Drift [↑](#footnote-ref-7)
7. Adaptive [↑](#footnote-ref-8)
8. Skewed class distribution [↑](#footnote-ref-9)
9. University of California, San Diego [↑](#footnote-ref-10)
10. filter methods [↑](#footnote-ref-11)
11. wrapper methods [↑](#footnote-ref-12)
12. Embedded methods [↑](#footnote-ref-13)
13. data aggregation [↑](#footnote-ref-14)
14. static plain graphs [↑](#footnote-ref-15)
15. static attributed graph [↑](#footnote-ref-16)
16. Maximum Description Length [↑](#footnote-ref-17)
17. Error correcting graph matching distance [↑](#footnote-ref-18)
18. Maximum Common Sub graph (MCS) [↑](#footnote-ref-19)
19. distance of adjacency matrices [↑](#footnote-ref-20)
20. Graph Edit Distance (GED) [↑](#footnote-ref-21)
21. time evolving [↑](#footnote-ref-22)
22. Compact Matrix Decomposition (CMD) [↑](#footnote-ref-23)
23. Pattern Learning and Anomaly Detection on Streams (PLADS) [↑](#footnote-ref-24)
24. eigenvalue [↑](#footnote-ref-25)
25. betweenness [↑](#footnote-ref-26)
26. closeness [↑](#footnote-ref-27)
27. Big Data [↑](#footnote-ref-28)
28. volume [↑](#footnote-ref-29)
29. variety [↑](#footnote-ref-30)
30. velocity [↑](#footnote-ref-31)
31. conspiratorial [↑](#footnote-ref-32)
32. Overfitting [↑](#footnote-ref-33)
33. Early stopping [↑](#footnote-ref-34)
34. potential overestimation issues [↑](#footnote-ref-35)
35. Bayesian Bernoulli co-clustering [↑](#footnote-ref-36)
36. boxplots [↑](#footnote-ref-37)
37. peak analysis [↑](#footnote-ref-38)
38. concentration function [↑](#footnote-ref-39)
39. breadth-first search [↑](#footnote-ref-40)
40. width bining interval [↑](#footnote-ref-41)
41. Peer Group Analysis [↑](#footnote-ref-42)
42. false discovery rate [↑](#footnote-ref-43)