به نام خدا

شماره: .....

تاریخ: .....

کاربرگ درخواست تصویب موضوع پیشنهادی پایان نامه در دوره کارشناسی ارشد

نام و نام خانوادگی: تخصص: محل خدمت: شماره تماس: امضا: تاریخ:

استاد راهنمای اول:

دوم:

استاد مشاور اول:

دوم:

موضوع پیشنهادی پایاننامه:

فارسی: تشخیص تقلب بیمه سلامت از طریق تحلیل گراف

انگلیسی: Healthcare Fraud Detection using Graph Analysis

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].

هزینه بهداشت و درمان با توجه به جمعیت و تغییرات قانون به سرعت در حال افزایش است .این افزایش در هزینههای بهداشت و درمان بر دولت و سیستم های سلامت تأثیر میگذارد . رفتارهای متقلبانه ارایه دهندگان بهداشت و درمان و بیماران با تحمیل هزینههای غیرضروری به مشکلی جدی برای سیستم های بیمه ای تبدیل شدهاست صنعت بیمه و در راس آن بیمه سلامت با هزینهای بالغ بر۵۰هزارمیلیارد تومان یکی از کلیدیترین هزینههای تحت نظارت و مدیریت دولت است که با تخمینی معادل ۳تا۱۰ درصد یعنی ۱۵تا۵۰ هزارمیلیارد ریال تقلب مواجه است. باتوجه به حجم اسناد و انبوه بیمه شدگان و ارایه دهندگان خدمات، کشف تقلب به صورت سنتی غیر ممکن است .کاهش هزینههای سلامت از طریق کاهش تقلب میتواند منجر به افزایش کیفیت و کمیت خدمات سلامت به افراد باشد. شرکت های بیمه می توانند با آگاهی از انواع تقلبات و فرایندهایی که احتمال بروز تقلب در آنها وجود دارد سیستم هشدار دهنده و پیشگیرانهای را طراحی کنند و با آگاهی از میزان آسیب پذیری خود استراتژیهای موثرتری را بهکار گیرند [2].

*مهلت تکمیل و تصویب این کاربرگ تا پایان نیم‌سال دوم پس از شروع تحصیل میباشد.

**طبق مصوبه شورای تحصیلات تکمیلی اصلاح نهایی عنوان پایان نامه در جلسه دفاعیه طبق نظر هیئت داوران انجام خواهد گرفت .

رویکردهای این حوزه را می توان از مناظر مختلف دسته بندی کرد:

۱.روشهای مبتنی بر یادگیری ماشین

۱.۱باناظر

۲.۱بدون ناظر

۲ .روشهای مبتنی بر روشهای آماری

۱.۲نمایهسازی(profiling)

۲.۲قانونBenford

۳.۲بصری سازی

۳.روشهای مبتنی بر تشخیص ناهنجاری

۱.۳تحلیل گراف

۱.۱.۳ایستا

-مبتنی برساختار

- مبتنی براجتماع

۲.۱.۳پویا

-مبتنی برفاصله

-مبتنی برفشردهسازی

-مبتنی برتجزیه

-مبتنی برمدل احتمالاتی

-مبتنی برپنجره

۲.۳استنتاج قواعد

۳.۳سایرروشها

از چالش های تشخیص تقلب مراقبت سلامت نیاز به دادهی برچسبدار و عدم ارائه بهبود چندان مناسب در نتایج حاصل از اعمال الگوریتم های باناظر و بدون ناظر است. تفسیر دشوار نتایج و عدم پوشش دهی همه موارد تقلب ، پیچیدگی محاسباتی بالا در تحلیل مبتنی بر شبکه و وقوع False Positiveدر مواردیکه اشتراک در ساختار گراف هست.همچنین رویکرد مبتنی بر قاعده نیاز به متخصصینی برای تعریف قوانین دارد. [14][7]برخی دیگر از چالشهای پیشروی این حوزه:

- کاهش حجم ابعاد داده
- پنهان بودن ماهیت تقلب [12]
- پویایی و حساسیت به تغییر در تقلب [12]

وجود تعداد بسیار کم داده جهت یادگیری داده با برچسب سالم[13]

- وجود پدیده رانش که در داده کاوی به پدیدهای که مدل پایه ی آن در طول زمان درحال تغییر است اشاره دارد.سیستمهای تشخیص تقلب در محیط پویا که رفتار کاربران قانونی/غیرقانونی بطور پیوسته در حال تغییر است مفهوم پدیده رانش گفته میشود. [13].
- توزیع اریب کلاسها[1] کهاشاره بر نامتوازن بودن دادهها دارد[13].
- تفاوت در قوانین سیستم سلامت و درمانی هر کشور

کار اصلی پژوهش من:

با توجه به کمبود راه حلهای مفهومی کلان داده و کاربرد آن با ابزارها، کتابخانهها و بسترهای کلان داده ی جایگزین در تشخیص تقلب مراقبت سلامت در [15]روی مجموعه دادهی CMSتحقیقاتی انجام شده است که روشهای Gradient Boosted Trees,RandomForest,Linear Regressionروی قسمت

های مختلف این مجموعه داده آزمایش شده و درنتایج حاصل از Linear Regression روی مجموعه دادهی ترکیبی متشکل از مجموعه داده part B , part D, DMEPOSنسبت به نتایج اعمال سایرروشها روی سایر تک تک مجموعه دادهها بهبود حاصل شده است اما کاری است با حجم محاسبات بسیار بالا که باید در سطح آزمایشگاه های محاسبات ابری پیاده سازی گردد

استفاده از روش پیشپردازش در [17]که با استفاده از Appache pig scriptمرحلهی پیش پردازش داده ها شامل پاکسازی دادهها و تولید Provider similarity Graphانجام شده و از یک الگوریتم personalize page rank برای یافتن ناهنجاری و از ویژگی تخصص پزشک داخلی، دندانپزشک، چشم پزشک و جراح پلاستیک استفاده شده است . می توان از الگوریتم Page Rank استفاده شده در این تحقیق بهره گرفته و هر پزشک را یک نود گراف در نظر گرفته و ارتباطات میان پزشکان تنها بر اساس تخصص ان ها باشد، به این ترتیب حجم محاسبات گراف به جای n^2 به mn کاهش می یابد.

برای تحلیل مبتنی بر گراف باید از یک پایگاه داده گرافی استفاده نمود، مانند هدوپ یا Neo4j و... .

می توان مشابه روشی که در [18] استفاده شده، از دیتابیس گرافی neo4j استفاده کنیم و نتایج حاصل را با معیارهای F-measure و ROC گزارش شده در آن مقایسه نماییم. البته نتایج گزارش شده در [16] نیز می تواند بعنوان بنچ مارک برای مقایسه نتایج مورد استفاده قرار گیرد.

با توجه به اینکه تحقیقات زیادی روی تشخیص تقلب دسیسهای (conspiracy)صورت نگرفته، در[16] شبکهای از پزشکان تبانیگر باروشهای بدون ناظر Isolation Forest,Local Outlier Factor,Unsupervised Random Forest,K Nearest Neighbor,Auto Encoder یافت شده اند و از نظر معیارهای ارزیابی مختلف با یکدیگر مقایسه شده اند.

هم چنین دیتاست استفاده شده می تواند مشابه دیتاست استفاده شده در [18] باشد و یا ترکیبی ازدیتاست استفاده شده در [17] با ستون Excluded در دیتاست [19] باشد در واقع پزشکان متناظر در دیتاست[17] را که در [19]هستند را یافته و ستون Exclude که شامل پزشکان مرتکب تقلب شده را به دیتاست [17] اضافه کرد تا بتوان مقادیر F measure یا ROC و ... را محاسبه نمود.

با توجه به اینکه داده ی حوزه سلامت به شدت نامتوازن است و تنها درصد بسیار کمی از ان شامل برچسب تقلب است، انجام متوازن سازی روی داده ها ضرورت دارد و تا حد زیادی از وقوع overfittingجلوگیری می نماید.

خلاصه‌ای از برخی پژوهش‌های انجام شده در حوزه مساله:

[3] محاسبه ریسک بر اساس فاصله‌ی مهالنوبیس و چگالی Likelihood مقدار مطالبات و مقایسه‌ی ریسک با یک آستانه از پیش تعیین شده و ساخت درخت تصمیم آن و آزمایش روی ۴ تخصص انتخابی شامل چشم، اعصاب، حلق و عمومی و ارایه‌ی دقت بالا در مقایسه با روش نیمه نظارتی و بدون ناظر.

[4] ابتدا برای رفتارهای غیرعادی سناریوهایی توسط متخصصان و پزشکان تولید می‌شود. سپس actor ها و ویژگی‌ها با روش‌های وزن‌دهی binary pairwise comparison وزن‌دهی می‌شوند. انبار داده‌ی دو مرحله‌ای شامل پاکسازی داده و محاسبه‌ی امتیاز z ویژگی‌هاست که از آن موتور تخصیص خطا که امتیاز خطای actor ها و مطالبات را محاسبه می‌کند. در این مقاله از ابزار visualization توسعه یافته QlikView هم برای تحلیل proactive و هم reactive به‌کار می‌رود استفاده کرده‌اند که بر اساس مقدار ورودی ویژگی‌ها یا امتیاز ریسک‌های نتیجه‌گیری شده مطالبات یا actor ها به تحلیل می‌پردازد و کاربر را قادر می‌سازد تا با ابزار آموزش تعمیم و تغییر پارامترها و معرفی روابط جدید به‌عنوان شاخص‌های خطا برای تراکنش‌ها تعامل داشته باشد.

[5] بر اساس اطلاعات نسخ دارویی طی ۵ گام نسبت به شناسایی پزشکان متقلب اقدام کرده‌اند.یک مجموعه داده از پزشکان شامل ۱۶۴ پزشک عمومی و ۴۷۴۸۹۷ نسخه داروی تهیه و رکوردهایی که داده‌های ناشناس زیادی داشتند از مجموعه داده حذف شدند و از روش‌های آماری برای پر کردن داده‌های مفقود استفاده نشده است.برای شناسایی رفتار متقلبانه‌ی پزشکان ۱۵ مصاحبه با افراد متخصص صورت گرفته و راه‌های تقلب پزشکان را بررسی نموده اند. چون ارزیاب‌های نسخه به نسخ بالای ۴ دارو حساس هستند، پزشکان متقلب ۳ یا کمتر دارو را در یک نسخه‌ی جعلی قرار می‌دهند. برای هرکدام از گروه‌ها هم میانگین و هم انحراف معیار محاسبه شده است. در آخرین گام توسط روش خوشه‌بندی ۹۲٪ صورتحساب‌ها که مربوط به ۱۱ ماه است جدا شدند و برای هر پزشک مقادیر شاخص ها محاسبه و با استفاده از امتیاز z نرمال‌سازی شد. سپس بر اساس hierachical cluster method عمل خوشه‌بندی انجام و پزشکان به دو گروه عادی و متقلب تقسیم‌بندی شدند و بر اساس معیار فاصله اقلیدسی، تعداد بهینه خوشه‌ها با استفاده از شاخص اعتباری بیشینه مقدار ضریب سیلوعت محاسبه گردید.

[6] برای تشخیص حلقه‌های جرم و شبکه‌های تبانی مانند پزشکان و داروخانه‌هایی که از اطلاعات افراد بی‌خانمان به‌عنوان بیمار استفاده می‌کنند، به تحلیل گراف برای آزمایش نقاط داده در ارتباط با یکدیگر می پردازند. با رویکرد ego-net که به گره‌های فردی و ویژگی‌های چکیده‌ی همسایه‌های محلی گره متمرکزند مانند درجه و آنتروپی ارتباطات محلی و سپس تحلیل ساختار کلی ارتباطات شبکه‌ای مراقبت سلامت و جستجو برای اجتماعاتی که ناهنجار هستند و استفاده از Fruchterman-Reingold که یک layout مبتنی بر فیزیک است برای آشکارسازی خوشه‌های پزشکان و داروخانه‌هایی که از طریق تراکنش‌های مُسکن به‌هم مرتبط هستند.

مشخصه‌های Temporal گراف نمایش مطالبات به‌عنوان یک توالی زمانی گسسته از پزشکان و محاسبه‌ی احتمال گذار با استفاده از تخمین maximum Likelihood و مقایسه‌ی این احتمال گذار با یک مقدار پایه منجر به شناسایی رئوس source , sink , یال‌های قویاً متصل می‌گردد. مشخصه‌های Geospatial و به‌دست آوردن یک تابع توزیع تجمعی بر اساس آن و نهایتاً استفاده از الگوریتم تشخیص ناهنجاری iForest.

[7] در طرح تشخیص ناهنجاری در گراف ویژگی‌های مرکزی مختلف مانند درجه گره، مرکزیت ego-net و... استخراج می‌شوند و یک فضای ویژگی با بقیه ویژگی‌هایی که از منابع اطلاعاتی اضافی برای تشخیص تقلب استخراج شده‌اند ساخته می‌شود.الگوریتم‌های GBAD-MDL , GBAD-MPL, GBAD-P برای کشف زیرساخت‌های غیرعادی استفاده و به‌کار گرفته شده‌اند.

[8] اعمال فیلترینگ ویژگی‌ها برای جداسازی بازپرداخت‌های کم، تعداد بیماران کم و تعداد مطالبات کم و استفاده از تکنیک‌های تحلیل و آنالیز و استفاده از تکنیک‌های تشخیص outlier شامل انحراف از مدل خطی، انحراف خوشه، انحراف از خوشه تکی، انحراف گرایشی، حداکثر انحراف و ... . و ارزیابی دقیق تکنیک‌های outlier مربوط به انواع تقلب مراقبت سلامت.

[9] به‌کارگیری تکنیک‌های social network برای تحلیل مطالبات بیمه سلامت از طریق نگاشت پزشکان با استفاده از بیماران مشترک به‌عنوان یک پروکسی برای ارتباط میان آن‌ها و ارزیابی مدل توسط تحلیل‌گران فرایند و همچنین بهبود درک اهمیت ویژگی‌های مهم پزشکان و بیماران و ارتباط میان آن‌ها.

[10] تخصیص احتمال به هر پزشک، ساخت ماتریس ارتباط میان دو پزشک و شناسایی پزشکانی که در یک شبکه به‌هم متصلند و با سایر پزشکان ارتباطی ندارند، به عنوان شبکه متقلبانه و همچنین استفاده از اپراتورهای پایگاه داده به جای حلقه میان هر جفت پزشک.

[11] استفاده از روش استنتاج قوانین بیمار بر اساس روش bump hunting روی داده‌های CMS و مقایسه با طبقه‌بندهای SVM,NaiveBaysian, Random Forest, Discriminant Analysis Classifier,Logistiv Regression.

منابع:

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

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

[4] 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.

[5] 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.

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

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

[8] 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.

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

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

[11] Sadiq, Saad, et al. "Mining Anomalies in Medicare Big Data Using Patient Rule Induction Method." *Multimedia Big Data (BigMM), 2017 IEEE Third International Conference on*. IEEE, 2017.

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

[13] Aisha Abdallah, MohdAizainiMaarof, , et al.”Fraud detection system: A survey,Journal of Network and Computer Applications Journal of Network and Computer Applications.2016.

[14] Travaille, Peter, et al. "Electronic Fraud Detection in the US Medicaid Healthcare Program: Lessons Learned from other Industries." *AMCIS*. 2011.

[15] 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.

[16] da Rosa, Raquel C. *An Evaluation of Unsupervised Machine Learning Algorithms for Detecting Fraud and Abuse in the US Medicare Insurance Program*. Diss. Florida Atlantic University, 2018.

[17] Seo, Jiwon, and OferMendelevitch. "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.

[18] Branting, L. Karl, et al. "Graph analytics for healthcare fraud risk estimation." 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 2016.

[19] https://oig.hhs.gov/exclusions/exclusions_list.asp