****

**استخراج الگو و تشخیص تقلب بیمه خودرو**

**در حضور داده های نامتوازن و رانش مفهوم**

**Car Insurance fraud detection and pattern mining in the presence of unbalanced data and concept drift**

دانشجو: بهنام یوسفی مهر

شماره دانشجویی: 99112005

گرایش: داده کاوی

استاد راهنما: دکتر مهدی قطعی

1. **مقدمه**

# در طی سالیان اخیر فعالیت های کلاهبرداری با هدف قرار دادن صنعت بیمه رشد قابل توجهی داشته است ، بطوریکه طبق گفته مدیرکل دفتر امور حقوقی بیمه مرکزی ایران حدود [یک سوم تصادفات ساختگی و برای گرفتن خسارت از بیمه است](https://www.khabaronline.ir/news/1484816/%DB%8C%DA%A9-%D8%B3%D9%88%D9%85-%D8%AA%D8%B5%D8%A7%D8%AF%D9%81%D8%A7%D8%AA-%DA%A9%D8%B4%D9%88%D8%B1-%D8%B3%D8%A7%D8%AE%D8%AA%DA%AF%DB%8C-%D9%88-%D8%A8%D8%B1%D8%A7%DB%8C-%DA%AF%D8%B1%D9%81%D8%AA%D9%86-%D8%AE%D8%B3%D8%A7%D8%B1%D8%AA-%D8%A7%D8%B2-%D8%A8%DB%8C%D9%85%D9%87-%D8%A7%D8%B3%D8%AA) و این رقم  روز به روز در حال افزایش است ، این در حالی ‌است که شماری از خسارت‌ های کاذب قابل ‌تشخیص هستند؛ اما بسیاری از آن‌ ها قابل‌ اثبات نیستند و شرکت ‌های بیمه مجبور به پرداخت خسارت می ‌شوند.

# همچنین افراد متقلب سعی در یافتن شیوه های جدید و بهبود روش خود برای فریب شرکت های بیمه را دارند، علاوه بر این که تقلب شرکت های بیمه را متحمل زیان های بزرگی کرده است ، شناسایی این افراد نیز چالش بزرگی برای شرکت های بیمه بوده و هزینه و زمان زیادی را میطلبد . از این رو، استفاده از ابزارهای قدرتمندی مانند یادگیری ماشین و علوم داده کاوی برای تشخیص تقلب و استخراج الگوها نیاز است.]2[و]3[

در این پایان نامه، می­کوشیم ابتدا داده های بیمه خودرو را تحلیل، بررسی و ویژگی های مرتبط با تقلب را دریابیم و از روی آنها مدلی برای تشخیص تقلب و همچنین الگوها را استخراج ­کنیم تا سازمان­ها و شرکت های بیمه ای بتوانند با استفاده از این مدل و الگوها تصادفات ساختگی را بدون صرف هزینه و منابع تشخیص دهند. در این رویکرد، اولا در نظر داریم که به جای انتخاب ویژگی­های مورد نیاز برای تشخیص و استخراج الگوهای تقلب به صورت کاربری، از روش های استخراج ویژگی استفاده کنیم. سپس به متوازن ساختن داده ها با استفاده از ترکیب کم نمونه گیری و بیش نمونه گیری پرداخت، سپس با بررسی روش های مختلف به یک مدل مقاوم در مقابل رانش مفهوم رسید .

ترکیب ساختار این پروپوزال به این ترتیب است. چالش های اصلی مسأله در بخش 2 آمده است. در بخش 3، صورت دقیق مسأله ی تشخیص و استخراج الگوهای تقلب بیمه خودرو بیان می شود. بخش 4 بـــه خلاصه سازی و بیان کارهای گذشتگان در این زمینه و روش های قبلی حل مسأله می پردازد. در بخش 5، روش حل مناسب که در این پایان نامه مد نظر است، شرح داده می شود. بخش 6 به شرح و توصیف پایگاه داده های انتخاب شده برای مطالعه پرداخته و در بخش 7، روش ارزیابی مدل و الگوریتم مورد مطالعه در این پایان نامه بیان می شود.

1. **چالش های مسأله**

چالش های زیادی تشخیص و استخراج الگو تقلب در بیمه خودرو وجود دارد، که ایده پردازی و پرداختن به همگی آن ها در مدت زمان تعیین شده این پایان نامه امکان پذیر نمی باشد. بنابراین، در ادامه چالش هایی آورده شده است که انتظار می رود بتوان در این پایان نامه راه حلی برای آن ها مطرح کرد.

یکی از چالش های این مسأله این است که داده های مورد بررسی دارای توازن نیستند یعنی تعداد داده های عادی بسیار بیشتر از داده های تقلب اند در چنین مواردی دسته بند تعصب بیشتری بر روی داده های عادی بر خود میگیرد و مدل عملکرد مناسبی را نخواهد داشت به همین دلیل باید راهکاری برای توازن داده ها ارائه کرد .

یکی دیگر از چالش ها استخراج الگوها برای وضع قوانین و ... در بیمه خودرو است که در این مطالعه موردی تازگی خود را دارد و اکثر کارهای پیشین فقط بر روی تشخیص تقلب استوار بوده اند .

از دیگر چالش های موجود رانش مفهوم است بطوریکه عادات مشتری در طول زمان تغییر میکند و کلاهبرداران با گذشت زمان استراتژی های خود را تغییر می دهند و این امر باعث کاهش دقت مدل میشود .

در این پایان نامه، کوشش می شود تا راهکارهای مناسبی برای حل این چالش ها ارائه شود.

در انتها، زمانی که داده های ما نامتوازن است معیار دقت اثربخشی دسته بند را نمیتواند به تنهایی مورد بررسی قرار دهد که همین امر مقایسه ی راهکار پیشنهادی این پایان نامه و ارزیابی مناسب آن را دشوار می سازد. به همین جهت، معیار های مختلفی را نیز برای مقایسه در بخش ارزیابی گزارش کردیم .

1. **صورت مسأله**

مسأله ی تشخیص و استخراج الگو تقلب در بیمه خودرو به فرآیندی گفته می شود که طی آن با داشتن داده های بیمه ای خودرو به عنوان ورودی، ساختگی بودن تصادفات و همچنین الگوها استخراج می شود.

فرض کنیم وبه ترتیب نمایانگر تعداد نمونه ها و ویژگی ها باشند. ورودی و خروجی دقیق مسأله در زیر آورده شده است:

**ورودی:** ورودی ماتریس تشکیل شده از نمونه در سطرها و ویژگی در ستون ها است، بطوریکه:

مقدار نمونه ی ام در ویژگی ام

در این مسأله پایگاه داده ی شامل ادعا های خسارت بیمه خودرو است که در آن هر سطر، یک ادعای خسارت انجام گرفته است، و ویژگی های معمول این مسأله یعنی ستون ها، شامل زمان تصادف، مشخصات راننده، مشخصات خودرو و ... می باشد.

**خروجی:** خروجی اول یک مدل است که با استفاده از آن می توان ورودی های جدید را بررسی و ساختگی یا عادی بودن تصادف را دسته بندی کرد .

خروجی دوم ماتریس است که حاوی الگوهای استخراج شده از هستند و هر سطر این ماتریس ها شامل یک الگو می باشد که می تواند در وضع قوانین و ... به سازمان ها و شرکت های بیمه ای کمک کند .

1. **کارهای گذشتگان**

الگوریتم های متفاوتی برای تشخیص تقلب و استخراج الگو وجود دارند همچنین برای متوازن سازی داده ها دو رویکرد کم نمونه گیری و بیش نمونه گیری وجود دارد. در جدول 1، خلاصه ای از کارهای گذشتگان برای حل مسأله آمده است:

**جدول 1-خلاصه ای از مقالات در حوزه ی کشف و استخراج الگو تقلب بیمه خودرو**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **مقاله** | **سال** | **عنوان** | **شباهت** | | **الگوریتم** |
| [4] | 2018 | Effect of Class Imbalanceness in Detecting  Automobile Insurance Fraud | کم نمونه گیری |  | Support-vector machine,  Multilayer perceptron,  Decision tree |
| بیش نمونه گیری |  |
| استخراج الگو |  |
| رانش مفهوم |  |
| [5] | 2018 | Credit Card Fraud Detection: A Realistic Modeling  and a Novel Learning Strategy | کم نمونه گیری |  | Online learnig,  DDM,  Ensemble learning |
| بیش نمونه گیری |  |
| استخراج الگو |  |
| رانش مفهوم |  |
| [6] | 2019 | Detecting Fraudulent Insurance Claims Using  Random Forests and Synthetic Minority  Oversampling Technique | کم نمونه گیری |  | Random Forests |
| بیش نمونه گیری |  |
| استخراج الگو |  |
| رانش مفهوم |  |
| [7] | 2020 | Empirical Oversampling Threshold Strategy for Machine Learning Performance Optimisation in Insurance Fraud Detection | کم نمونه گیری |  | K-nearest neighbors |
| بیش نمونه گیری |  |
| استخراج الگو |  |
| رانش مفهوم |  |
| [8] | 2020 | Detecting Fraudulent Motor Insurance Claims Using Support Vector Machines with Adaptive Synthetic Sampling Method | کم نمونه گیری |  | Support-vector machine,  Random Forests |
| بیش نمونه گیری |  |
| استخراج الگو |  |
| رانش مفهوم |  |

جدول بالا نشان می دهد که الگوریتم های متفاوتی برای تشخیص تقلب به کار گرفته شده است و اکثرا برای متوازن سازی داده ها روی بیش نمونه گیری تمرکز شده است و توجه خاصی به کم نمونه گیری برای متوازن سازی داده ها بهره نشده است در حالیکه در مقالات موازی از هر دوی این روش ها یا ترکیب این دو روش بهره برده می شود . همچنین در حوزه استخراج الگو کار خاصی نشده و به رانش مفهوم نیز توجه خاصی نشده است .

از گپ های مقالات، می توان به عدم بررسی تاثیر کم نمونه گیری در متوازن سازی داده ها اشاره کرد. اکثرا فقط سعی کرده اند تعداد داده کلاس اقلیت را با بیش نمونه گیری افزایش دهند و به کم نمونه گیری از کلاس غالب توجه ای نکرده اند. از دیگر گپ های مقالات، عدم وجود مقالات و توجه به مفهوم رانش و استخراج الگوها در این حوزه است .

1. **کارهای آینده**

راه حل پیشنهادی این پایان نامه جهت حل مسأله به ترتیب شامل پیش پردازش، پاکسازی و آماده سازی داده های بیمه خودرو، سپس استخراج ویژگی های مرتبط و بهینه ، متوازن سازی داده ها ، استخراج الگوها و تشخیص تقلب می باشد.

همانطور که بحث شد یکی از چالش های اساسی در مقوله تشخیص تقلب عدم توازن داده هاست. در کار های گذشتگان برای متوازن سازی داده ها فقط از روش بیش نمونه گیری استفاده شده است و به روش کم نمونه گیری توجه خاصی نشده است . در این پایان نامه با ترکیب دو روش کم نمونه گیری و بیش نمونه گیری برای متوازن سازی داده ها، بطوریکه علاوه بر بیش نمونه گیری با حذف مقداری داده از کلاس غالب سعی در رسیدن به یک توازن مناسب خواهیم داشت .

چالش دیگری در بحث تشخیص تقلب این است که افراد متقلب سعی در یافتن شیوه های جدید و بهبود روش خود برای فریب شرکت های بیمه را دارند ، در این پایان نامه بنا بر این است که مدل ارائه شده تحت تاثیر داده های نامتوازن و رانش مفهوم قرار نگیرد .

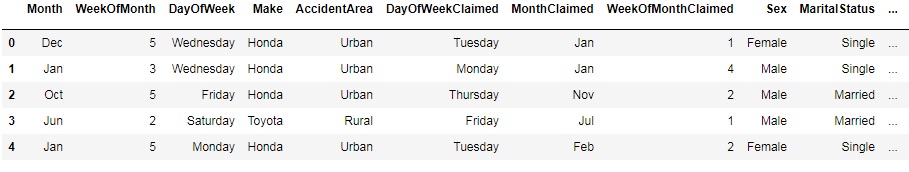
استخراج الگوها در کنار تشخیص تقلب نیز می تواند به سازمان ها و شرکت های بیمه ای کمک کند که در وضع قوانین ، مشتری سنجی و ... از این الگوها استفاده لازم را ببرند.

سامانه ی پیشنهاد شده روی پایگاه داده ی ذکر شده در قسمت 6 پیاده سازی و اجرا شده و با معیارهای ارزیابی اشاره شده در قسمت 8 تست و ارزیابی خواهد شد.

1. **پایگاه داده**

در این پایان نامه از پایگاه داده با داده های واقعی یک شرکت بیمه در مورد مطالبات خودرو که توسط Angoss Knowledge Seeker Software ارائه شده و با نام "carclaims.txt" شناخته شده است استفاده خواهیم کرد. این مجموعه داده شامل 15420 ادعا است كه بین ژانویه 1994 و دسامبر 1996 گزارش شده است که از این مقدار تعداد 14497 نمونه عدم تقلب و 923 نمونه تقلب کرده اند بطوریکه 6٪ موارد تقلب و 94٪ موارد عدم تقلب را دارا است. ]9[

بخش کوچکی از این پایگاه داده در تصویر 1 مشاهده می شود:



**تصویر 1-بخشی از پایگاه داده ی carclaims.txt**

در جدول 2، ویژگی های این پایگاه داده به همراه توضیحات هر ویژگی آورده شده است:

**جدول 2-پایگاه داده ی carclaims.txt** ]10[

|  |  |  |
| --- | --- | --- |
| **ماه رخ دادن حادثه** | **Month** | **1** |
| **دوره هفته در یک ماه که حادثه رخ داده است .** | **WeekOfMonth** | **2** |
| **روزی که حادثه رخ داده است** | **DayOfWeek** | **3** |
| **بیمه مالک مارک خودرو** | **Make** | **4** |
| **محل حادثه : "شهری" یا "روستایی"** | **AccidentArea** | **5** |
| **ماهی که مالک بیمه ادعا می کند.** | **DayOfWeekClaimed** | **6** |
| **دوره هفته در یک ماه که مالک بیمه ادعا می کند** | **MonthClaimed** | **7** |
| **روزی که صاحب بیمه ادعا می کند.** | **WeekOfMonthClaimed** | **8** |
| **جنسیت صاحب بیمه** | **Sex** | **9** |
| **وضعیت تاهل صاحب بیمه** | **MaritalStatus** | **10** |
| **سن صاحب بیمه** | **Age** | **11** |
| **مقصری که باعث حادثه شده است** | **Fault** | **12** |
| **نوع خط مشی اعمال شده در موارد تصادف** | **PolicyType** | **13** |
| **رده خودرو صاحب بیمه** | **VehicleCategory** | **14** |
| **قیمت وسیله نقلیه صاحب بیمه** | **VehiclePrice** | **15** |
| **تقلبی بودن یا نبودن ادعا** | **FraudFound** | **16** |
| **شماره بیمه نامه** | **PolicyNumber** | **17** |
| **شماره از نماینده** | **RepNumber** | **18** |
| **پولی که باید قبل از تأمین بودجه توسط شرکت بیمه طبق ادعای ارائه شده توسط مالک بیمه پرداخت شود** | **Deductible** | **19** |
| **مهارت راننده خودرو (درجه 1 - 5)** | **DriverRating** | **20** |
| **خط مشی در صورت تصادف (در روز)** | **Days:Policy-Accident** | **21** |
| **خط مشی در مورد روند ادعا (در چند روز)** | **Days:Policy-Claim** | **22** |
| **تعداد ادعاهایی که تا به حال مطرح شده است** | **PastNumberOfClaims** | **23** |
| **عمر خودرو صاحب بیمه** | **AgeOfVehicle** | **24** |
| **سن صاحب بیمه (براساس رده سنی)** | **AgeOfPolicyHolder** | **25** |
| **پر کردن پرونده گزارش تصادف به پلیس** | **PoliceReportFiled** | **26** |
| **حضور شاهد** | **WitnessPresent** | **27** |
| **نوع نماینده در روند مطالبات بیمه** | **AgentType** | **28** |
| **تعداد پیوست ها یا اطلاعات تکمیلی** | **NumberOfSuppliments** | **29** |
| **تغییر آدرس صاحب بیمه هنگام مطالبه** | **AddressChange-Claim** | **30** |
| **تعداد وسایل نقلیه متعلق به دارندگان بیمه** | **NumberOfCars** | **31** |
| **سال مطالبات بیمه اتومبیل** | **Year** | **32** |
| **بیمه نامه های اساسی برای مواردی که صاحبان بیمه تجربه می کنند استفاده می شود** | **BasePolicy** | **33** |

1. **روش ارزیابی**

برای ارزیابی الگوریتم ها و روند پیشنهادی برای حل مسأله، رویکرد زیر را در پیش گرفته و نتایج بدست آمده را با هم مقایسه می کنیم. این بخش کاربرد ومقایسه معیارهای ارزیابی مدل ما و انتخاب برای تصمیم گیری در مورد بهترین الگوریتم های دسته بندی ارائه شده توسط یک مطالعه موردی را ارائه می دهد.

با در نظر داشتن ﻧﻤﺎدﻫﺎي TP(ﻣﺜﺒﺖ واﻗﻌﻲ)، TN(ﻣﻨﻔﻲ واﻗﻌﻲ)، FP (ﻣﺜﺒﺖ ﻛﺎذب) و FN(منفی کاذب) در این بخش از سه معیار زیر استفاده خواهیم کرد:

1) معیار دقت (Accuracy): تعداد پیش بینی های صحیح ساخته شده توسط مدل نسبت به انواع پیش بینی های انجام شده است. دقت یک معیار عالی برای سنجش کارکرد مدل است اما وقتی مجموعه داده شده متوازن و متعادل باشد. با این حال، وقتی داده ها مانند پایگاه داده ی مورد مطالعه ما متعادل نباشند، معیار دقت اثربخشی دسته بند را جلب نمی کند، زیرا ممکن است مدل تحت تاثیر داده قرار بگیرد و در تصمیم گیری تنها با پیش بینی "عدم تقلب" به دقت بالایی برسد (پارادوکس دقت) .

از این رو از معیار حساسیت و ویژگی نیز در کنار معیار دقت برای رفع پارادوکس دقت بهره خواهیم برد.

2) معیار حساسیت (Recall): معیاری است که به ما می گوید چه نسبت ادعاهایی که در واقع تقلب بوده اند، توسط الگوریتم به عنوان تقلب پیش بینی شده است.

در بسیاری از حوزه های تشخیص داده های پرت شناسایی این داده ها تا حد امکان (به عبارت دیگر ارزشیابی معیار های حساسیت یا recall در تشخیص داده های پرت) بسیار مهم تر از منحرف نشدن اشیاء نرمال به عنوان داده های پرت است. در نتیجه هنگامی که از یک روش دسته بندی برای تشخیص داده های پرت استفاده می شود ، بسیار مناسب است که بر روی معیار ارزشیابی recall تاکید شود . ]1[

3) معیار ویژگی (Specificity): معیاری است که به ما می گوید چه نسبت ادعاهایی که در واقع تقلب نبوده اند، توسط مدل پیش بینی شده است که تقلبی نیست.

# **منابع**

|  |  |
| --- | --- |
| [1] | J. Han, J. Pei and M. Kamber, Data mining: concepts and techniques, Elsevier, 2011. |
| [2] | https://www.khabaronline.ir/news/1484816/یک-سوم-تصادفات-کشور-ساختگی-و-برای-گرفتن-خسارت-از-بیمه-است |
| [3] | https://bimebazar.com/blog/افزایش-تخلف‌های-ساختگی-برای-گرفتن-خسا/ |
| [4] | S. Subudhi and S. Panigrahi, "Effect of Class Imbalanceness in Detecting Automobile Insurance Fraud," 2018 2nd International Conference on Data Science and Business Analytics (ICDSBA), Changsha, China, 2018, pp. 528-531, doi: 10.1109/ICDSBA.2018.00104. |
| [5] | A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi and G. Bontempi, "Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy," in IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 8, pp. 3784-3797, Aug. 2018, doi: 10.1109/TNNLS.2017.2736643. |
| [6] | S. Harjai, S. K. Khatri and G. Singh, "Detecting Fraudulent Insurance Claims Using Random Forests and Synthetic Minority Oversampling Technique," 2019 4th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2019, pp. 123-128, doi: 10.1109/ISCON47742.2019.9036162. |
| [7] | Bouzgarne Itri, Youssfi Mohamed, Bouattane Omar and Qbadou Mohamed, “Empirical Oversampling Threshold Strategy for Machine Learning Performance Optimisation in Insurance Fraud Detection” International Journal of Advanced Computer Science and Applications(IJACSA), 11(10), 2020. http://dx.doi.org/10.14569/IJACSA.2020.0111054 |
| [8] | C. Muranda, A. Ali and T. Shongwe, "Detecting Fraudulent Motor Insurance Claims Using Support Vector Machines with Adaptive Synthetic Sampling Method," 2020 61st International Scientific Conference on Information Technology and Management Science of Riga Technical University (ITMS), Riga, Latvia, 2020, pp. 1-5, doi: 10.1109/ITMS51158.2020.9259322. |
| [9] | Soufiane, E., Baghdadi, S. E. E., Berrahou, A., Mesbah, A., & Berbia, H. Automobile Insurance Claims Auditing: A Comprehensive survey on Handling Awry Datasets.‏ |
|  |  |
|  |  |
| [10] | T. Badriyah, L. Rahmaniah and I. Syarif, "Nearest Neighbour and Statistics Method based for Detecting Fraud in Auto Insurance," 2018 International Conference on Applied Engineering (ICAE), Batam, Indonesia, 2018, pp. 1-5, doi: 10.1109/INCAE.2018.8579155. |