ارائه یادگیری ماشین

Detecting Fraudulent Insurance Claims Using Random Forests and Synthetic Minority Oversampling Technique

ارائه دهنده : سید مهدی مقدسی
دانشجوی دکتری مهندسی پزشکی– بیوالکتریک

دی ۱۴۰۰

چکیده

- در چند سال اخیر فریب عمدی با حذف حقایق و پنهان کردن جزئیات، مقدار قابل توجهی رشد داشته و برای بیمه منجر به ضرر قابل توجهی شده است.

- تقلب در بیمه زمانی رخ می‌دهد که افراد تلاش می‌کنند با عدم تحقق شرایط قرارداد بیمه از آن سود ببرند.

- برای کنترل این خطرات، یک چارچوب مناسب برای نظارت منطقی در مورد تقلب در بیمه مورد نیاز است. در مقاله مورد بررسی، یک رویکرد جدید برای ردیابی خودکار تقلب در بیمه ارائه شده است.

- در این ارائه به شرح مسئله ، توضیح ،پیاده سازی روش های به کار برده شده در مقاله " تشخیص تقلب در بیمه خودرو با استفاده از جنگل های تصادفی و تکنیک بیش نمونه برداری اقلیت مصنوعی "و مقایسه آن ها با پیاده سازی خود خواهیم پرداخت .

مقدمه

- صنعت بیمه از همان ابتدا با تعداد زیادی از چالش های ناشی از کلاهبرداری روبرو بوده است که منجر به افزایش هزینه های حق بیمه ، روند ناکارآمدی و از دست دادن اعتماد شده است .
- گرچه همه بیمه ها و شرکت ها سیستم های تشخیص کلاهبرداری خود را دارند، اما بیشتر این فرایندها بسیار ناکارآمد و زمان بر هستند و به شدت به انسان متکی هستند.
- داده کاوی و تجزیه و تحلیل تقلب، سناریوی تشخیص را تغییر داده است، داده ها را می توان از منابع مختلف جمع آوری و در یک مخزن ترکیبی ذخیره نمود؛ داده کاوی می تواند بسیاری از هزینه شرکت ها را با صرفه جویی و کاهش کلی هزینه کشف تقلب بهبود دهد.
- در تشخیص تقلب، چالشی که مطرح است مشکل عدم توازن داده هاست. به این معنا که تعداد موارد یک کلاس (نمونه های تقلب نکرده) بسیار بیشتر از تعداد موارد کلاس دیگر (نمونه های تقلب کرده) است. به همین دلیل، طبقه اقلیت تمایل دارد که در طول روند دسته بندی نادیده گرفته شود. بنابراین عدم توازن داده ها باید رفع شود و یک راه ساده برای رفع این مشکل ایجاد تعادل بین مجموعه داده هاست که با نمونه گیری بیش از حد انجام می شود .

مقدمه

❖ تقلب ها را می توان به دو دسته ی سخت و نرم دسته‌بندی کرد .

- اگر یک مالک عمدا یک تصادف را برنامه‌ریزی کند یا فقط برای به دست آوردن سود از شرکت بیمه، خسارت وارد کند، در این صورت گفته می‌شود که این یک تقلب سخت است.
- هنگامی که یک جراحت یا سرقت واقعی رخ می‌دهد، و بیمه‌شده در ادعای به دست آوردن پول بیشتر از شرکت اغراق می‌کند، آن را تقلب نرم می‌نامند.

شرح مسئله

- هدف این مقاله **توسعه ی مدلی برای کمک به بیمه ها در تشخیص هر چه بهتر تقلب** است به طوری که بیمه می تواند پس از استفاده از الگوریتم فقط برای اطمینان بیشتر، نمونه های مشکوک را توسط ناظر بررسی و در هزینه ها صرفه جویی کند.
- ارزیابی موفقیت یک مدل ، برای تعیین میزان مناسب بودن آن برای حل یک مساله خاص بسیار مهم است. حتی پیشرفت‌های اندک در عملکرد، می‌تواند منجر به **مزایای اقتصادی بزرگی** شود.

تشریح الگوریتم

- مقاله مدلی را پیشنهاد داده است که هدف آن تسهیل تصمیم گیری بیمه گران در مورد صحت ادعا، هنگام درخواست خسارت می باشد و روش پیشنهادی با هرگونه داده در زمان واقعی کار می کند.
- در این روش مجموعه داده ی اصلی قبل از اجرا، به یک مجموعه داده متعادل تبدیل می شود و پس از آن می توان هر نوع الگوریتم دسته بندی را روی داده ها به کار گرفت .
- در این مقاله ، یک روش شناسایی خودکار تقلب با استفاده از تکنیک **بیش نمونه گیری اقلیت مصنوعی** (SMOTE) برای تعادل داده ها، اعمال می شود و برای دسته بندی از روش **جنگل های تصادفی** استفاده می شود .

Synthetic Minority Oversampling Technique
Random Forests

مراحل مدل پیشنهادی

✓ مرحله 1: پیش پردازش داده ها

الف) پاکسازی داده ها :

پس از بارگذاری مجموعه داده، داده ها از لحاظ مقادیر از دست رفته، زائد، تکراری و نویز بررسی می شوند.

در مجموعه داده اصلی (carclaims.txt) هیچ مقادیر از دست رفته ای وجود نداشت.

ب) تبدیل داده ها:

تبدیل داده های کیفی به داده های عددی.

برگه پرونده ادعا شامل 4 ستون می باشد. (سال ، ماه ، هفته در ماه و روز)

که آن را برای محاسبه ی راحت تر در قالب تاریخ-زمان عادی به صورت YYYY-MM-DD در آوردند.

ج) مصور سازی داده ها:

داده ها با رسم نمودار، بطور کامل مورد تجزیه و تحلیل قرار می گیرند و این کار برای کشف وابستگی ها انجام می شود.

د) نمونه برداری مجدد داده یا تعادل داده (با استفاده از روش SMOTE)

مراحل مدل پیشنهادی (ادامه)

دانشگاه سمنان

✓ **مرحله 2: طبقه بندی داده ها (با استفاده از جنگل های تصادفی)**

استفاده از طبقه بندی جنگل های تصادفی با اندازه دسته 100 و seedValue = 1

اجرای روشِ 10-fold cross validation

✓ **مرحله 3: آموزش و آزمایش مدل**

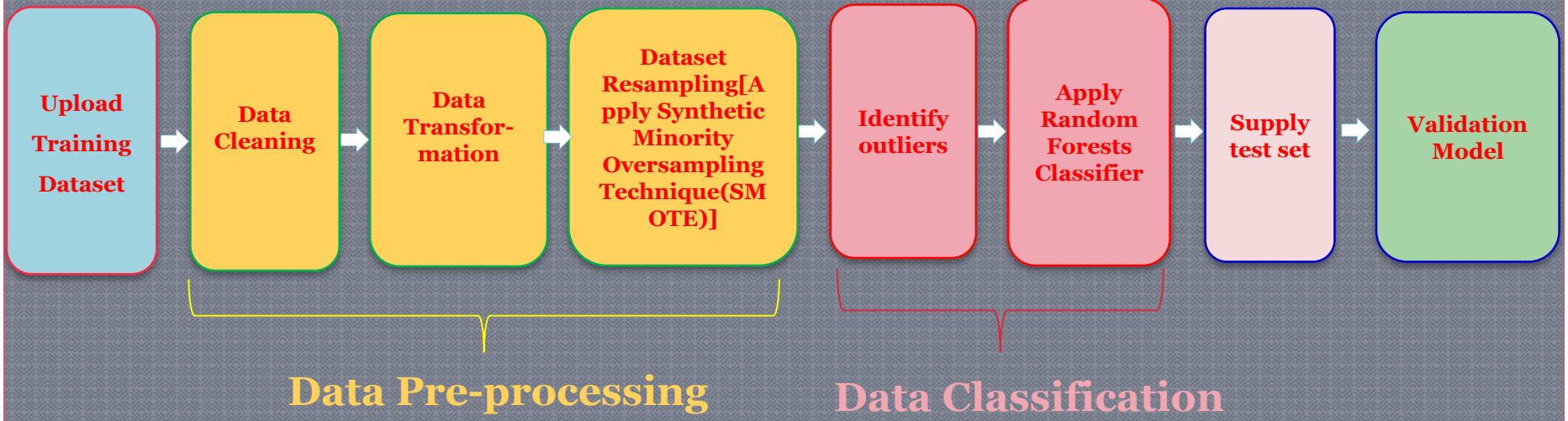
اجرای مدل روی مجموعه تست و آموزش با نسبت 20 به 80

✓ **مرحله 4: اعتبارسنجی مدل**

تایید نتایج به دست آمده با استفاده از ماتریس در هم ریختگی

شکل معماری مدل پیشنهادی

دانشگاه سمنان



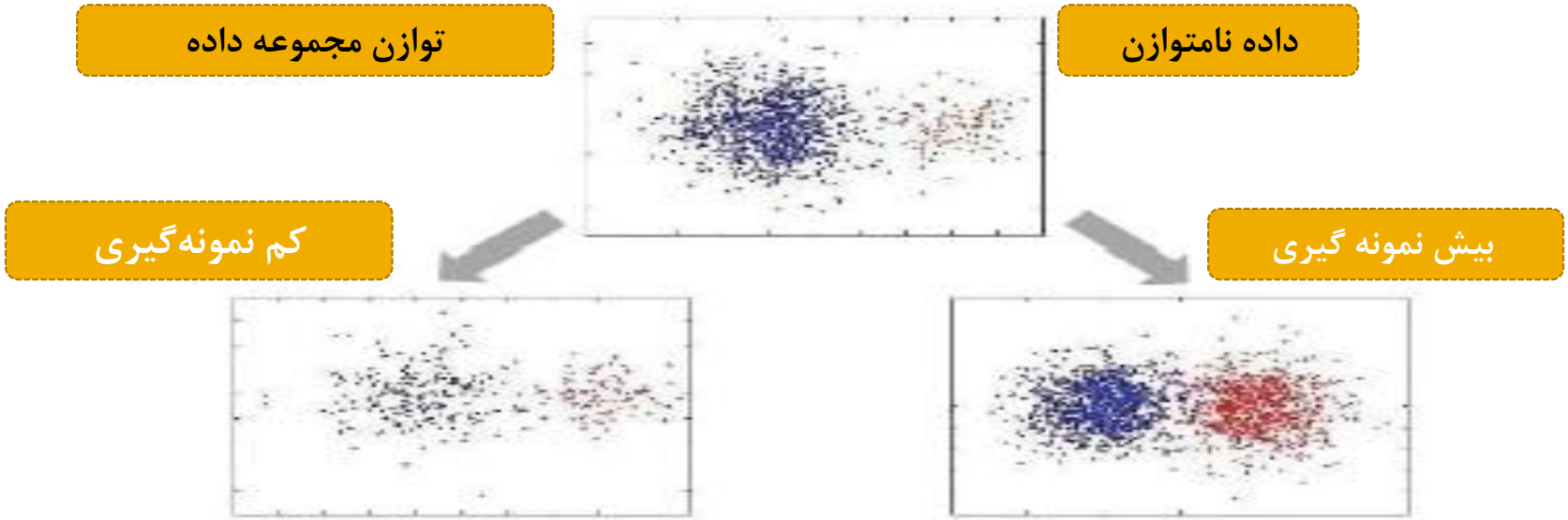
Proposed Architecture for Auto-Insurance Fraud Detection System

رفع عدم تعادل مجموعه دادگان

تکنیک های مختلفی برای متعادل سازی داده ها وجود دارد که به طور گسترده به **بیش نمونه گیری** و **کم نمونه گیری** تقسیم می شود.

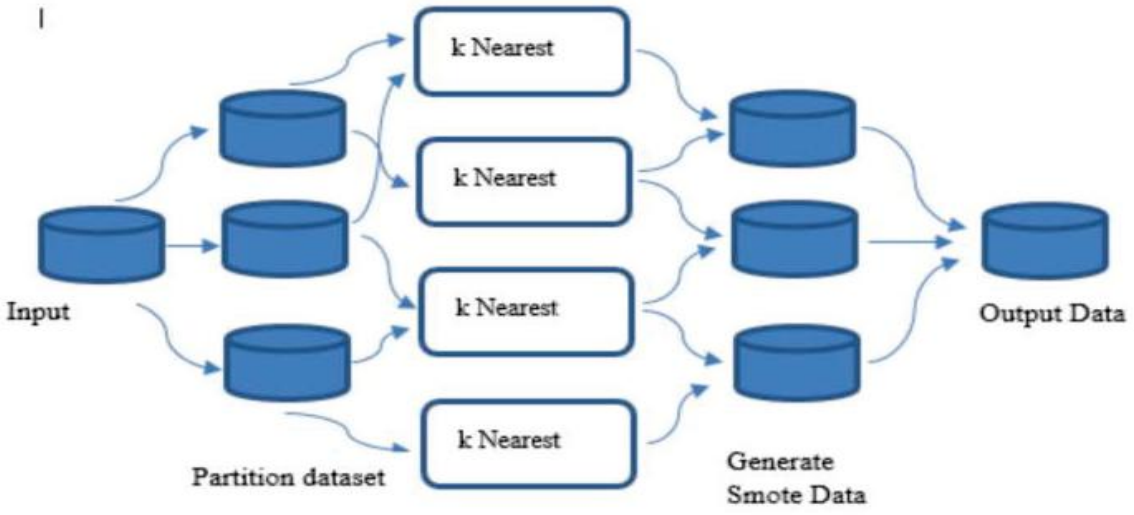
نمونه‌ها از کلاس غالب حذف ⟸ کم نمونه‌گیری

نمونه‌های بیشتری به کلاس اقلیت اضافه ⟸ بیش نمونه گیری

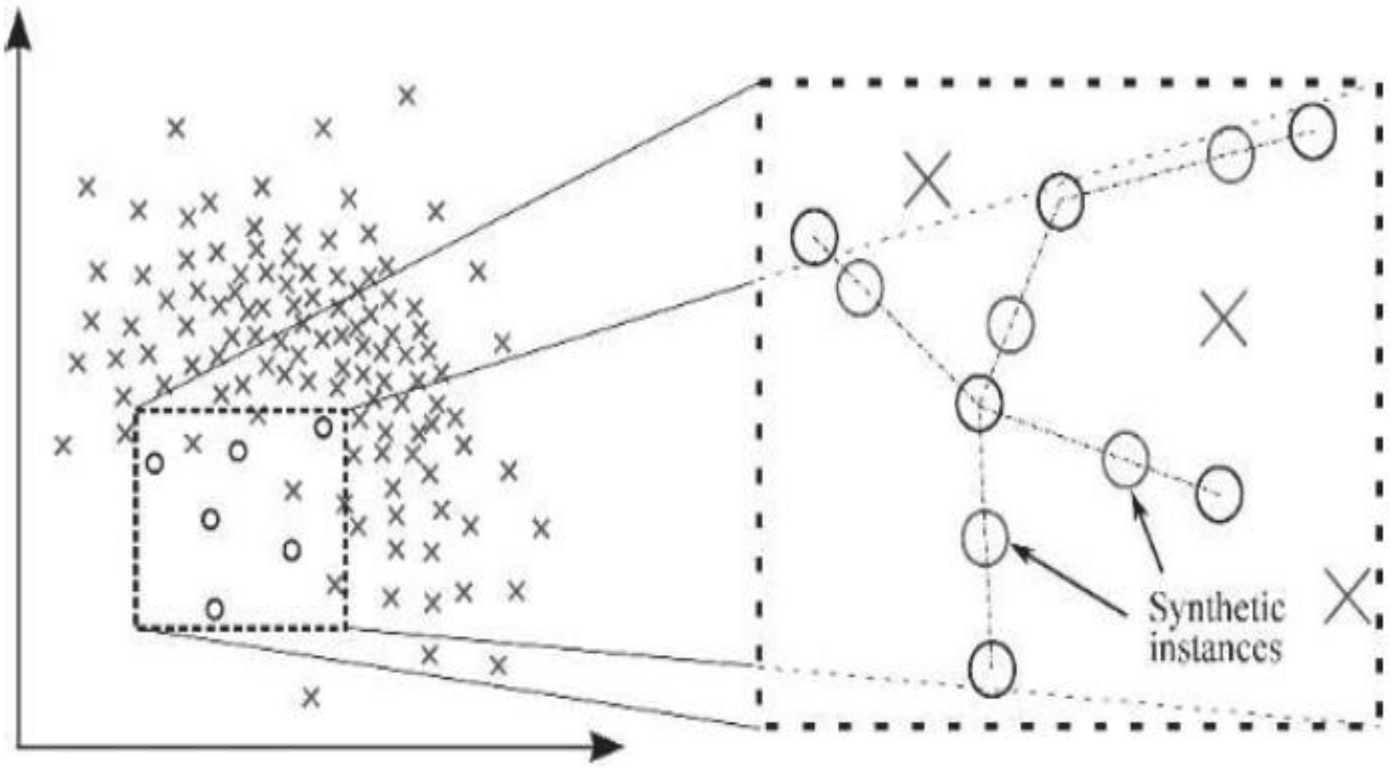


- روش **بیش نمونه گیری اقلیت مصنوعی** (SMOTE) برای نمونه گیری از مجموعه داده در رویکرد پیشنهادی مقاله استفاده شده است.
- این روش یک روش شناخته شده برای بیش نمونه گیری است که یک مجموعه داده نامتعادل را تغییر می دهد تا یک مجموعه متعادل ایجاد کند و نمونه های کلاس اکثریت و کلاس اقلیت را توزیع می کند.
- این روش موارد مصنوعی یا موارد مشابه از کلاس اقلیت ایجاد می کند و هدف آن کاهش تعصب پیش بینی دسته بند نسبت به کلاس اکثریت است .

- نمونه ها با توجه به الگوریتم k نزدیکترین همسایه تولید می شوند. به طوری که هر همسایه از k نزدیکترین همسایه به طور تصادفی انتخاب و در کنار هم قرار می گیرند.
- نمونه های جدید در امتداد پاره خط هایی که بین k نزدیکترین همسایگان کلاس اقلیت قرار دارند، ایجاد می شود.
- مقدار نمونه جدید به مقدار نیاز به داده جدید بستگی دارد .



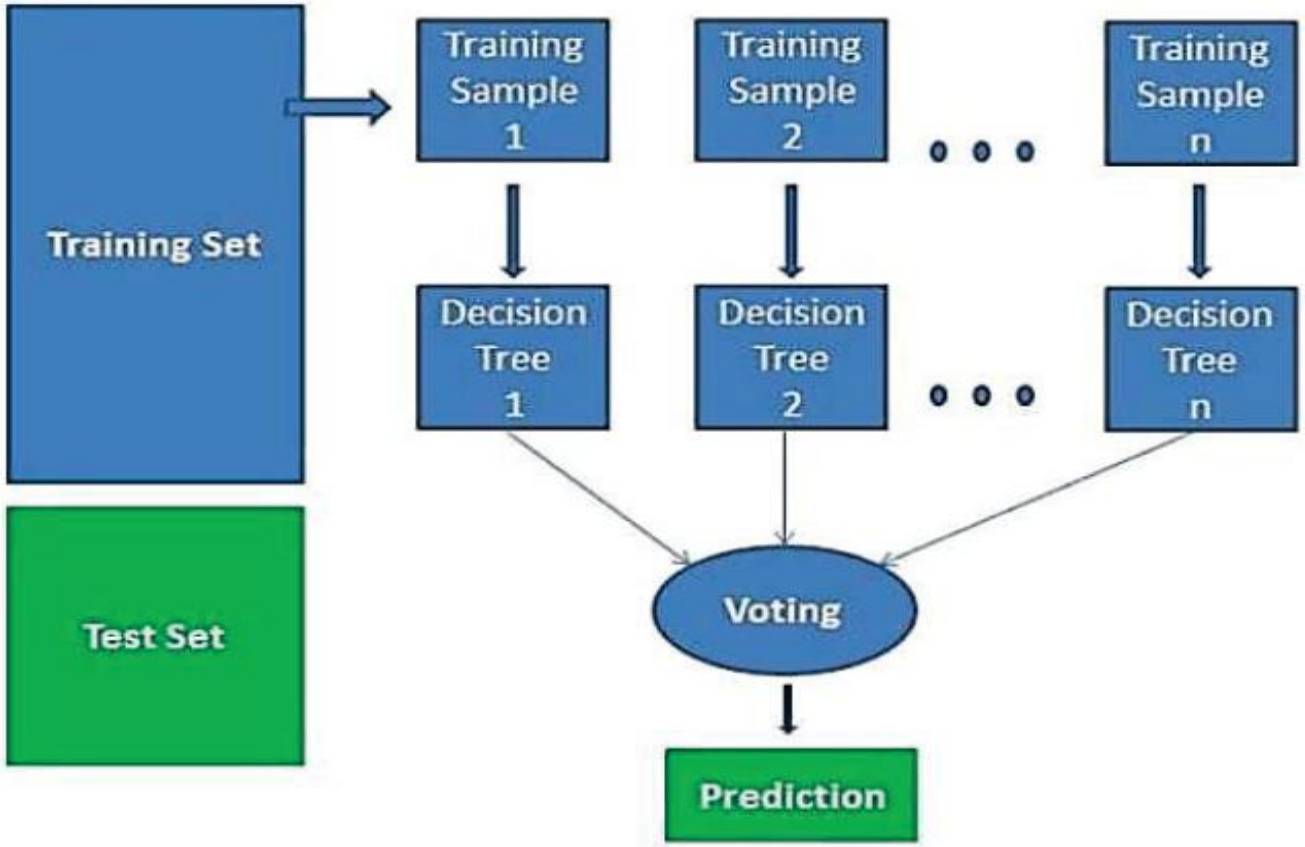
تولید نمونه‌های مصنوعی



دسته بند جنگل های تصادفی

- جنگل های تصادفی (RF)، که به عنوان تصمیم تصادفی نیز نامیده می شوند، یک الگوریتم یادگیری ماشین با ناظر بر پایه درخت است که می تواند برای **دسته بندی** و **رگرسیون** استفاده شود.
- RF درخت‌های تصمیم‌گیری را براساس نمونه‌های داده، تصادفی ایجاد می‌کند، مقدار پیش‌بینی را از هر درخت دریافت و بهترین راه حل را با روند رأی گیری انتخاب می کند.
- این الگوریتم، شاخص خوبی برای بدست آوردن ویژگی های مهم یک مجموع دادگان است.
- یکی از کاربردهای گسترده تکنیک RF در **تشخیص تقلب** و **فعالیت های کلاهبرداری** است.

دسته بند جنگل های تصادفی



پیاده سازی و مقایسه نتایج

دانشگاه سمنان

آشنایی با مجموع دادگان:

- مجموع دادگانی که در مقاله بررسی شده است، یک مجموع دادگان معروف در بین مقالات کشف تقلب در حوزه بیمه است. (carclaims.txt) این مجموع دادگان دارای 15420 رکورد با 32 ویژگی است .
- نمای کوچکی از مجموع دادگان را در زیر می بینیم :

```
df.head()
```

	Month	WeekOfMonth	DayOfWeek	Make	AccidentArea	DayOfWeekClaimed	MonthClaimed	WeekOfMonthClaimed	Sex	MaritalStatus	...	AgeOfVehicle
0	Dec	5	Wednesday	Honda	Urban	Tuesday	Jan	1	Female	Single	...	3 years
1	Jan	3	Wednesday	Honda	Urban	Monday	Jan	4	Male	Single	...	6 years
2	Oct	5	Friday	Honda	Urban	Thursday	Nov	2	Male	Married	...	7 years
3	Jun	2	Saturday	Toyota	Rural	Friday	Jul	1	Male	Married	...	more than 7
4	Jan	5	Monday	Honda	Urban	Tuesday	Feb	2	Female	Single	...	5 years

پیاده سازی مدل پیشنهادی

مرحله 1 :پیش پردازش

✓ الف) پاکسازی داده ها

همانطور که در مقاله مطرح شد، در مجموعه داده اصلی (carclaims.txt) مقادیر از دست رفته ای وجود نداشت.

✓ ب) تبدیل داده ها

داده ها را با استفاده از لیبل انکدر تبدیل به داده های عددی و قابل پردازش کردیم .

	Month	WeekOfMonth	DayOfWeek	Make	AccidentArea	DayOfWeekClaimed	MonthClaimed	WeekOfMonthClaimed	Sex	MaritalStatus	...	AgeOfVehicle
0	2	5	6	6	1	6	1	1	0	2	...	1
1	4	3	6	6	1	2	1	4	1	2	...	4
2	10	5	0	6	1	5	11	2	1	1	...	5
3	6	2	2	17	0	1	7	1	1	1	...	6
4	4	5	1	6	1	6	2	2	0	2	...	3

پیاده سازی مدل پیشنهادی (ادامه)

دانشگاه سمنان

✓ ج) مصور سازی داده ها

نامتوازن بودن مجموع دادگان را به وضوح در شکل می بیند.

این مجموعه داده یک توزیع کلاسی نامتوازن است زیرا تقریبا ۹۴ % از ادعاهای غیرکلاهبرداری و تقریبا ۶ % از ادعاهای کلاهبرداری را دارد.

پیاده سازی مدل پیشنهادی (ادامه)

دانشگاه سمنان

✓ د) نمونه برداری مجدد داده یا ایجاد تعادل داده (با استفاده از روش SMOTE)

با استفاده از روش شرح داده شده، به مقدار داده ی تولید شده در مقاله، بیش نمونه گیری انجام دادیم و تعداد رکورد مجموع دادگان به 16343 افزایش یافت و کلاس اقلیت همانند مقاله از 6 درصد به 11 درصد افزایش یافت .

پیاده سازی مدل پیشنهادی (ادامه)

دانشگاه سمنان

مرحله 2: طبقه بندی داده ها (با استفاده از جنگل های تصادفی)

با استفاده از دسته بندی جنگل های تصادفی با اندازه دسته ۱۰۰ و seedValue = 1

داده ها را دسته بندی کردیم.

مرحله 3: آموزش و آزمایش مدل

پس از اجرای مدل روی مجموعه تست و آموزش با نسبت ۲۰ به ۸۰ نتایج مقاله به صورت زیر است :

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
No	0.999	0.549	0.935	0.999	0.966	0.933
Yes	0.451	0.001	0.986	0.451	0.619	0.933
Weighted Avg.	0.937	0.487	0.940	0.937	0.927	0.933

پیاده سازی مدل پیشنهادی (ادامه)

نتایج به دست آمده **در پیاده سازی ما** نیز به صورت زیر می باشد و می بینیم که پیاده سازی ما نیز توانسته نتایجی مشابهی کسب کند.

	precision	recall	f1-score	support
No	0.92	1.00	0.96	14497
Yes	0.98	0.36	0.52	1846
accuracy			0.93	16343
macro avg	0.95	0.68	0.74	16343
weighted avg	0.93	0.93	0.91	16343

پیاده سازی مدل پیشنهادی (ادامه)

دانشگاه سمنان

❖ برای مقایسه آسانتر نتایج در جدول زیر آورده شده است:

class	Article Precision	Implementation Precision	Article Recall	Implementation Recall	Article f1-score	Implementation f1-score
No	0.935	0.92	0.999	1.00	0.966	0.96
Yes	0.986	0.98	0.451	0.36	0.619	0.52
Weighted avg	0.940	0.93	0.937	0.93	0.927	0.91

مرحله 4: اعتبارسنجی مدل

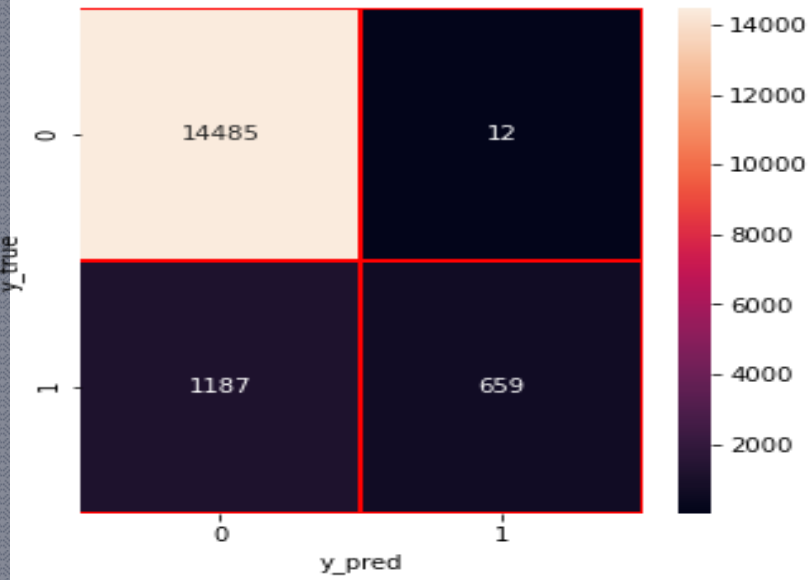
نتایج به دست آمده با استفاده از ماتریس در هم ریختگی اعتبار سنجی می شوند .

➢ ماتریس در هم ریختگی مقاله

```
=== Confusion Matrix ===

     a     b   <-- classified as
 14485    12 |     a = No
  1013   833 |     b = Yes
```

➢ ماتریس در هم ریختگی پیاده سازی ما



✓ **نتایج کسب شده در دسته تقلب نشده یکسان و نتایج در دسته تقلب کرده ها مقدار کمی متفاوت با نتایج مقاله است که این مقدار کم برای ما به علت تفاوت داده ها در تقسیم برای تست و آموزش قابل قبول است.**

پیاده سازی مدل پیشنهادی (ادامه)

دانشگاه سمنان

مقایسه روش های دسته بندی:

- پس از ساخت مدل، مقاله به بررسی دسته بند ارائه شده با سایر دسته بندها اعم از **ماشین بردار پشتیبان**، **درخت تصمیم** و **پرسپترون چندلایه** پرداخته است.

ما نیز این روش ها را پیاده کرده و به بررسی نتایج کلی با نتایج ارائه شده توسط مقاله پرداخته ایم.

- **نتایج مقاله برای مقایسه روش پیشنهادی اش با سایر دسته بند ها به شرح زیر است :**

Performance Metrics (in %)	Support Vector Machine (SVM)	Decision Tree	Multi-layer Perceptron (MLP)	Proposed Model
Accuracy	58.41	57.39	74.98	94.33
Sensitivity (or Recall value)	90.53	86.94	47.83	**99.9**
Specificity	36.86	38.14	18.75	45.1

پیاده سازی مدل پیشنهادی (ادامه)

نتایج ما نیز بدین شکل حاصل شده اند :

Performance Metrics (in %)	Support Vector Machine (SVM)	Decision Tree	Multi-layer Perceptron (MLP)	Proposed Model
Accuracy	80	89	73	93
Sensitivity	89	94	79	100
Specificity	11	52	27	36

باتوجه به اینکه مقاله از نرم افزار داده کاوی Weka و ما از ابزار python برای پیاده سازی استفاده کردیم و تقسیم داده ها به آموزش و تست و همچنین random state ها، مقداری تفاوت در نتایج انتظار میرفت، ولی با این حال نتایج به دست آمده در کل، تفاوت زیادی با نتایج مقاله ندارد و با همین ابزار و نتایج نیز می توان نتیجه گرفت که مدل پیشنهادی مقاله، نسبت به سایر روش ها عملکرد بهتری داشته است.

جمع‌بندی و نتیجه‌گیری

- از روش پیشنهادی مقاله می توان نتیجه گرفت که در تشخیص تقلب، الگوریتم جنگل های تصادفی نتایج بسیار خوبی ارائه می دهد، همچنین برای تشخیص صحیح، متوازن کردن داده ها یک عمل ضروری است زیرا در داده های نامتوازن روش های دسته بندی، تعصب بیشتری روی کلاس اکثریت دارند .
- با پیاده سازی روش پیشنهادی مقاله نیز به این نتیجه دست یافتیم با اینکه ابزارهای پیاده سازی، تقسیم داده ها به مجموع آموزش و آزمایش و همچنین random state ها متفاوت بودند ولی نتایج کلی کسب شده مشابه نتایج مقاله به دست آمدند و در مقایسه ی روش پیشنهادی مقاله با سایر روش ها، روش پیشنهادی مقاله توانست دقت خوبی نسبت به سایر دسته بند ها کسب کند.
- همچنین به عنوان کار آتی با استفاده از سایر روش های متوازن کردن داده ها حتی استفاده ترکیبی از روش بیش نمونه گیری و کم نمونه گیری انتظار میرود بتوان نتایج را بهبود بخشید .

منابع

1. Sonakshi Harjai, Detecting Fraudulent Insurance Claims Using Random Forests and Synthetic Minority Oversampling Technique Inc. 2019 [online] Available : https://ieeexplore.ieee.org/abstract/document/9036162

2. Data Mining: Concepts and Techniques, 3rd Edition. Jiawei Han, Micheline Kamber, Jian Pei. Database Modeling and Design: Logical Design, 5th Edition.

3. https://blog.faradars.org/sampling-methods-in-dats-science/

با تشکر از توجه شما