The Thirty-Third International
FLAIRS Conference (FLAIRS-33)

MedFroDetect: Medicare Fraud Detection with Extremely Imbalanced Class Distributions

Yuping Su,[1,2,3] **Xingquan Zhu,**[3] **Bei Dong,** [1,2] **Yumei Zhang,**[1,2] **Xiaojun Wu**[1,2]
[1] Key Laboratory of Modern Teaching Technology, Ministry of Education Shaanxi Normal University, Xian, China
[2] School of Computer Science, Shaanxi Normal University, Xi'an, China
[3] Department of Computer and Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, USA
ypsu@snnu.edu.cn, xzhu3@fau.edu, {dongbei, zym0910, xjwu}@snnu.edu.cn

Abstract

In this paper, we propose to use machine learning to automate Medicare fraud detection. By cross checking Medicare payment database and provider exclusion database, we build datasets with millions of service providers, including a handful of convicted fraudulent service providers. One essential challenge is that the dataset created is extremely imbalanced, making it extremely difficult to learn accurate classifiers for fraud detection. To tackle the challenge, we first use feature engineering to design effective features, by taking the difference between each service provider and its group cohort into consideration. At the instance level, we also use a synthetic instance generation approach to generate positive samples to alleviate the data imbalance challenge. By combining feature engineering, synthetic instance generation, and under sampling based ensemble learning, our method outperforms baseline approaches for Medicare fraud detection.

1 Introduction

The United States Medicare services cover 18% of the US population. A report from the Centers for Medicare & Medicaid Services (CMS) shows that the Medicare improper payment rate in 2018 was 8.12%, contributing substantially to rising healthcare costs. Therefore, fraud detection is of great significance. The current Medicare fraud detection mainly depends on auditors' manual reviewing which is time-consuming and requires a significant amount of efforts and costs. Fortunately, the CMS has released a series of publicly available Medicare provider utilization and payment data (CMS 2019). The availability of the Medicare data, combining the recent advancement of machine learning research, makes automatic fraud detection possible.

In this paper, we use CMS *Medicare Provider Utilization and Payment Data: Physician and Other Supplier* (CMS 2019), also known as Medicare Part B data, to design fraud detection algorithm. The CMS payment data provide information about services and procedures, allowing us to observe behaviors of service providers to find fraudulent providers. To find label for each provider, we use the List of Excluded Individuals/Entities (LEIE) database (OIG 2019). By cross checking these two databases, we can build a dataset including millions of service providers and a handful of convicted fraud service providers. While the learning objective is clear, one essential challenge is that the dataset created is extremely imbalanced, making the learning extremely difficult.

To tackle data imbalance, a couple of existing research on Medicare fraud detection propose to construct simple features and use basic re-sampling techniques (Bauder and Khoshgoftaar 2018), including random under-sampling (RUS), random over-sampling (ROS), Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al. 2002) and Adaptive Synthetic (ADASYN) (Haibo He et al. 2008) sampling.

While the existing research mainly relies on simple features and re-sampling techniques, our research is motivated by designing new/effective features to characterize the similarity/distance between each providers and their cohort groups. The assumption is that a provider dissimilar to the cohort groups is more likely to be a fraudulent provider.

2 MedFroDetect: Medicare Fraud Detection System

Medicare Payment and Exclusion Datasets

We use two publicly available data sources: the Medicare Part B dataset (CMS 2019) and the LEIE (OIG 2019) dataset. The former is used to construct features and the latter is used to find labels for each service provider in Medicare Part B dataset.

The Medicare Part B dataset provides information on services and procedures provided to Medicare beneficiaries by providers. Each row in this dataset is a claim from a service provider, which is differentiated by the National Provider Identifier (NIP). Each claim provides information on utilization, payment and submitted charges of services specified by NPI, HCPCS code, and place of service. So far the data covers calendar years 2012 through 2017 and we choose 2013 and 2014 as our test bed.

For the Medicare Part B dataset, each provider may have multiple claims, depending on the HCPCS codes involved and places the services were provided. Therefore, we first aggregate the dataset and use one row to denote a service provider (denoted by NPI) and construct features for the providers.

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Table 1: A summary of CMS Medicare payment datasets

Year	# of Providers (NPI-level)	# of Fraud (NPI-level)	Percentage of fraud (%)
2013	908,833	1,013	0.1115
2014	937,311	802	0.0856

In order to obtain label for each provider, we use LEIE dataset to find fraudulent providers. It is assumed that providers appear in LEIE dataset are considered fraudulent and otherwise as non-fraudulent.

Although cross checking NPIs between the Medicare dataset and the LEIE dataset can help us find label for provider, one challenge is that only about 7% of providers in the LEIE dataset have an NPI number. In order to maintain the most accurate fraud label mapping, we only use provider NPI and exclude providers without an NPI number. In addition, when finding fraudulent providers in Medicare dataset, we only consider providers in LEIE dataset whose exclusion date is not before the year of Medicare data. We use the September 2019 version of LEIE dataset in this paper.

By cross checking NPIs of 2013 and 2014 Medicare datasets with the September 2019 LEIE dataset, respectively, the number of fraudulent providers and the corresponding percentage for each Medicare dataset is provided in Table 1.

Feature Engineering

Based on the aggregated Medicare dataset, we design 4 categorical features and 12 numerical features as shown in Table 2. For the state type, providers located within the fifty U.S. States and the District of Columbia belong to one state type, and providers located in other places (*e.g.*, Guam) belong to another state type. We use one-hot encoding to convert categorical features as numerical values.

Among all 12 numerical features in Table 2, the first 8 numerical features provide statistical summary of the provider with respect to the services, beneficiaries, involved HCPCS codes etc. Because fraudulent providers are assumed to carry out services/procedures different from normal providers, we create another four numerical features by comparing each provider's HCPCS code based service distribution with respect to the same distribution of all normal providers. In the following, we elaborate detailed explanation of the motivation and the design method of these four features. For convenience, we always use "service distribution" to represent "service *vs.* HCPCS code distribution" in the analysis.

In our research, we consider providers of each specialty as a cohort group. Each specialty may contain from a few to tens of thousands providers. We first create average service distribution of each cohort by aggregating number of services at the HCPCS code level (the average service of an HCPCS code is calculated by the total services of the HCPCS code, divided by the number of providers in the cohort group). For each provider, we also create the service distribution, by listing the services *w.r.t.* each HCPCS code.

Fig. 1 (left panel) shows the service distributions of Clinical Laboratory specialty (1st row) and Internal Medicine

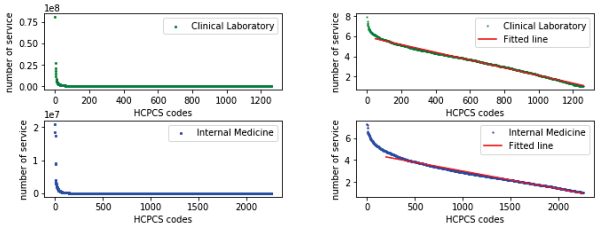

Figure 1: Service distributions of the Clinical Laboratory specialty/cohort and the Internal Medicine cohort.

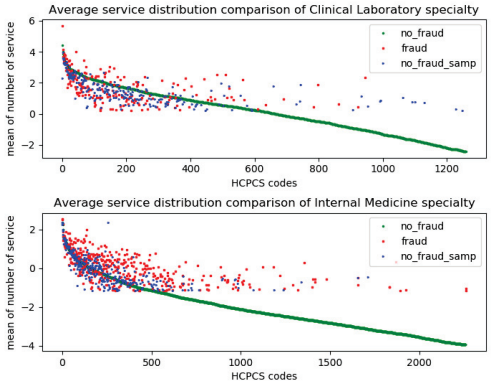

Figure 2: Service distribution comparison between fraudulent and non-fraudulent providers of a certain specialty.

specialty (2nd row) of 2014 Medicare dataset, where the number of services for all HCPCS codes are sorted in a descending order and the $x-$axis shows the rank order of the corresponding HCPCS code. The results show that service distributions largely follow the 80/20 rule, where most services are provided to a few number of HCPCS codes. The right panel of Fig. 1 reports the corresponding service distribution in log10 scale and the fitted line. The results show that the service distribution curve of each speciality largely follows a negative exponential function. Because log scale represents a fair importance for all HCPCS codes, we will only consider log10 scale service distribution in the following analysis.

To observe the service distribution difference between fraudulent and non-fraudulent providers, we report the average service distribution comparison of fraudulent and non-fraudulent providers for Clinical Laboratory and Internal Medicine cohorts (2014 Medicare dataset) in Fig. 2. We can find that fraudulent and non-fraudulent providers have different service distribution patterns for both specialties.

To further check whether fraudulent and non-fraudulent providers have different service distributions without taking the specialty into consideration, Fig. 3 compares average service distributions of all 802 fraudulent providers 936,509 non-fraudulent providers, and a random sample of 802 non-fraudulent providers.

Table 2: Features constructed from Medicare data

Feature Type	Feature	Description
Categorical Features	gender	Provider's gender
	state	The state where the provider is located
	provider_type	provider type(or specialty) of the provider
	state_type	Identify whether the provider is located in the fifty U.S. states and the District of Columbia or not
Numerical Features	tot_num_ser	# of services provided by the provider
	tot_num_benefi	# of distinct Medicare beneficiaries receiving the services
	tot_num_disti_benefi_ser	# of distinct Medicare beneficiary/per day services
	num_benefi_and_num_ser_ratio	Ratio between *tot_num_benefi* and *tot_num_disti_benefi_ser*
	num_disti_ser_and_num_ser_ratio	Ratio between *tot_num_disti_benefi_ser* and *tot_num_ser*
	tot_num_HCPCS_code	# of unique HCPCS codes involved by services
	num_ser_O	# of services when the place of service is non-facility (value of 'O')
	num_ser_F	# of services when the place of service is a facility (value of 'F')
	cos_simi_group	Cosine similarity between provider's service distribution and the non-fraudulent group average distribution
	cos_simi_global	Cosine similarity between provider's service distribution and the non-fraudulent global average distribution
	percent_vs_aver_group	Percentage of unique HCPCS codes with # of service above the non-fraudulent group average of service
	percent_vs_aver_global	Percentage of unique HCPCS codes with # of service above the non-fraudulent global average of service

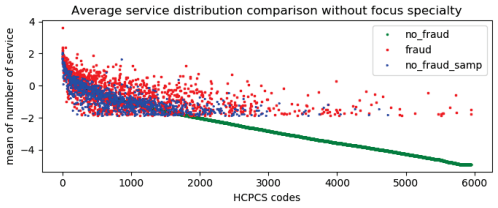

Figure 3: Average service distribution comparison between fraudulent *vs.* non-fraudulent providers without specialty.

Synthetic Samples & Imbalanced Learning

In order to tackle the extremely imbalanced class distribution, we propose to create synthetic samples to increase the population and diversity of positive samples for learning. More specifically, we first use Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al. 2002) to generate a small percentage of synthetic positive samples.

SMOTE generate new positive samples by interpolating synthetic instances between nearest neighbours in the set of minority class instances. It should be noted that SMOTE can only handle all numerical features. For this paper's case that both mixed dataset of numerical and categorical features, a simple generalization of SMOTE called SMOTE-NC (Chawla et al. 2002) can be used.

After the above process, we combine all positive samples and generated synthetic instances as positive instances. Even so, negative instances are far more than positive instances. To further alleviate the class imbalance, we use RUS to randomly sample a percentage of negative class instances, and combine them with positive instances to learn a classifier.

Because synthetic instance generation and under sampling both result in bias in the training data. To reduce bias, we repeat the instance generation and sampling for a number of times, and combine all trained classifiers as an ensemble.

3 Experiments

We implemented MedFroDetect and baseline methods using imbalanced-learn package (Lemaitre, Nogueira, and Aridas 2017) and compare algorithm performance on 2013 and 2014 Medicare datasets. All results are based on 5-fold cross validation with AUC being used as the performance metrics.

Baseline Methods: we use ROS, SMOTE-NC, and RUS sampling techniques as baseline to compare the performance of MedFroDetect for Medicare fraud detection. For comparison, the performance of original dataset without using any sampling technique is also reported.

Experimental Settings: In our experiments, we generate following class distributions (β=Positive:Negative): 1:99, 5:95, 10:90, 25:75, 35:65, 50:50 for MedFroDetect and baseline methods. We use Random Forest (RF) and decision tree to make classification and use scikit-learn (Pedregosa and et al 2011) to implement these classifiers. Unless otherwise specified, we use default parameters for both RF and decision tree classifiers. For each of the re-sampling technique, we sample the training data for 10 times and get the final prediction on the test set through vote of 10 trained models.

Performance Comparison: In Table 3, we report detailed AUC values between MedFroDetect and baseline techniques using Random Forest classifier. For MedFroDetect, α is the synthetic sample generation ratio and denotes that the number of synthetic samples is α times the original positive samples. Results in Table 3 show that MedFroDetect can achieve the best performance across most sampling ratios. Among all baseline methods, RUS achieves the best performance. This indicates that it's helpful to generate synthetic positive samples before performing RUS.

Table 3: Performance Comparison between MedFroDetect and baseline using Random Forest classifier.

Datasets	Sampling Method	Positive:Negative Sampling Ratio (β in MedFroDetect)						
		No Sampling	1:99	5:95	10:90	25:75	35:65	50:50
2013	ROS	0.6290	0.7594	0.7602	0.7523	0.7573	0.7615	0.7539
	SMOTE-NC	-	0.7925	0.7943	0.7936	0.7938	0.7908	0.7962
	RUS	-	0.8013	0.8175	0.8210	0.8230	0.8219	**0.8211**
	MedFroDetect$_{\alpha=0.2}$	-	**0.8056**	0.8188	**0.8230**	**0.8241**	**0.8223**	0.8208
	MedFroDetect$_{\alpha=0.5}$	-	0.8036	**0.8191**	0.8228	0.8239	**0.8223**	0.8206
	MedFroDetect$_{\alpha=1}$	-	0.8048	0.8187	0.8207	0.8225	0.8211	0.8193
	MedFroDetect$_{\alpha=3}$	-	0.8021	0.8160	0.8183	0.8205	0.8185	0.8176
2014	ROS	0.6053	0.7354	0.7409	0.7399	0.7406	0.7393	0.7385
	SMOTE-NC	-	0.7827	0.7863	0.7905	0.7905	0.7829	0.7850
	RUS	-	0.7925	0.8126	0.8149	0.8176	0.8158	0.8138
	MedFroDetect$_{\alpha=0.2}$	-	0.7996	**0.8158**	**0.8185**	0.8185	0.8166	0.8151
	MedFroDetect$_{\alpha=0.5}$	-	**0.8049**	0.8146	0.8180	**0.8202**	**0.8193**	0.8159
	MedFroDetect$_{\alpha=1}$	-	0.7984	0.8137	0.8182	0.8182	0.8180	**0.8167**
	MedFroDetect$_{\alpha=3}$	-	0.8037	0.8111	0.8123	0.8155	0.8157	0.8137

Table 4: Performance comparison between MedFroDetect and RUS using decision tree classifier

Datasets	Sampling Method	Positive:Negative Sampling Ratio (β in MedFroDetect)						
		No Sampling	1:99	5:95	10:90	25:75	35:65	50:50
2013	RUS	0.5037	0.6153	0.7163	**0.7487**	**0.7770**	0.7794	0.7784
	MedFroDetect$_{\alpha=0.2}$	-	0.6202	**0.7208**	0.7457	0.7766	0.7751	0.7824
	MedFroDetect$_{\alpha=0.5}$	-	**0.6214**	0.7130	0.7410	0.7676	**0.7832**	**0.7845**
	MedFroDetect$_{\alpha=1}$	-	0.6172	0.7025	0.7324	**0.7770**	0.7776	0.7788
	MedFroDetect$_{\alpha=3}$	-	0.6074	0.6881	0.7247	0.7625	0.7646	0.7843
2014	RUS	0.5032	0.6053	0.7014	0.7385	0.7662	0.7720	0.7737
	MedFroDetect$_{\alpha=0.2}$	-	0.6162	0.7140	**0.7464**	**0.7727**	0.7793	0.7835
	MedFroDetect$_{\alpha=0.5}$	-	**0.6210**	**0.7177**	0.7428	0.7708	0.7795	0.7799
	MedFroDetect$_{\alpha=1}$	-	0.6171	0.6969	0.7411	0.7692	**0.7834**	**0.7881**
	MedFroDetect$_{\alpha=3}$	-	0.5961	0.6826	0.7272	0.7651	0.7704	0.7837

To further check the effectiveness of MedFroDetect *w.r.t.* other learning algorithms, Table 4 reports the performance of MedFroDetect and RUS sampling using decision tree classifier. It still shows that MedFroDetect achieves better performance across almost all sampling ratios.

4 Conclusion

In this paper, we proposed a machine learning framework for Medicare fraud detection. We took the difference of service distributions between fraudulent and non-fraudulent service providers into consideration to design features. At the instance level, we combined synthetic instance generation and random under sampling to generate synthetic positive samples and reduce negative samples. By combining feature engineering and an ensemble based combination sampling framework, our method shows better performance than all baseline approaches for Medicare fraud detection.

5 Acknowledgment

This work is partially supported by the U.S. National Science Foundation under grant Nos. IIS-1763452 & CNS-1828181, National Natural Science Foundation of China (Nos. 61701291, 11772178, 61703258 & 11872036), China Postdoctoral Science Foundation (Nos. 2017M613053 & 2017M613054), Natural Science Foundation of Shaanxi Provincial (Nos. 2018JQ6089 & 2019GY-217), and Shaanxi Postdoctoral Science Foundation (No. 2017BSHYDZZ33). This research was conducted while the first author was a visiting scholar at Florida Atlantic University.

References

Bauder, R., and Khoshgoftaar, T. 2018. Medicare fraud detection using random forest with class imbalanced big data. In *IEEE Intl. Conf. on Info. Reuse and Integration (IRI)*, 80–87.

Chawla, N. V.; Bowyer, K. W.; O.Hall, L.; and Kegelmeyer, W. P. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16:321–357.

CMS. 2019. Medicare provider utilization and payment data, [online] https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/medicare-provider-charge-data/index.html.

Haibo He; Yang Bai; Garcia, E. A.; and Shutao Li. 2008. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE IJCNN*, 1322–1328.

Lemaitre, G.; Nogueira, F.; and Aridas, C. 2017. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research* 18(17):1–5.

OIG. 2019. Leie downloadable databases, [online] https://oig.hhs.gov/exclusions/exclusions_list.asp.

Pedregosa, F., and et al. 2011. Scikit-learn: Machine learning in Python. *J. of Machine Learning Research* 12:2825–2830.