Master's Thesis

Outlier detection technique to extract candidates of fraud from medical insurance claims

Guidance

Associate Professor Aki-Hiro Sato

Ryosuke Kawamori

Department of Applied Mathematics and Physics

Graduate School of Informatics

Kyoto University

February 2019

Abstract

This thesis proposes how to find fraudulent activities from data of medical insurance claims in an unsupervised way. Firstly, we split data into homogenous groups by the providers' department, a number of specialties, and their geographical location. If this is not accomplished, we may find providers performing rare operations, or providers in rural areas because of its variation in submitted charges. Next, we apply the outlier detection method, One-SVM and LOF to these groups. Finally, we evaluate the effectiveness by using labels of providers who are excluded because they committed fraud. The proposed method is above random and may be used as a first line of further investigation by specialties.

Contents

1 Introduction

The National Health Care Anti-Fraud Association estimated that at least 3%(60 billion USD) of US annual health care expenditures were lost owing to outright fraud [1]. Because of the great attention being paid to the patient safety, expectations, satisfaction, and advances in technology, there have been increases in the costs, which leaves greater room for fraud to occur [2]. To provide quality and safe care to legitimate patients, fraud control systems are needed. As noted above, because of upcoming health care technologies, many kinds of health care methods are in use. This makes it difficult to detect outright fraud manually in the vast amount of claims. Thus, fraud detection models are needed.

To detect outliers from many kinds of operations, we firstly construct vectors which denote features of the healthcare providers. For example, each element of the vector are the gender of the healthcare providers, the number of each operation he or she operated or the amount of money paid to the provider during some period. Next, we apply some fraud detection model to this vector. As noted above, many kinds of healthcare methods are in use, this vector tend to be high dimension. To extract fraudulent activities from the high dimensional feature vectorss, Van [3] used the quantile of each dimension. However, in the case of upcoding, providers submitted a lower amount for one operation and a higher amount for another operation. To cope with this situation, we need to take several dimensions into consideration simultaneously. Shin [4] used a weighted quantile as an outlier score. In the national health care program, providers can execute approximately 7000 types of operations. Thus, it is difficult to mix these operations to calculate one score by the experts. If we have data that has information on fraud labeling, we can build a supervised model to predict fraudulent activities. However, this method needs time-consuming labeling by domain experts and may be unable to find new types of fraudulent activities that are not in the training data.

In our study, we aim to support finding providers that commit fraud by extracting possibly fraudulent cases and use this model as a first line of further investigation by specialists. Our method is applicable to unlabeled data and can be used to extract candidates of outright fraud from a large amount of data. First, we split the data into homogeneous groups because dealing with all data simultaneously is computationally intensive, and if we try to find outliers from all providers, we may only find rare providers and times of computation is very long. Next, we try to find abnormal patterns in these groups. We use a one-class support vector machine (One-SVM) [5] and Local Outlier Factor (LOF) [6] to find abnormal points. The One-SVM estimates a function that is +1 in the region where most data points live and -1 in the other regions and LOF which calculates the ratio of the density of the object and its surroundings. To evaluate our methodology we use the List of Excluded Individuals and Entities(LEIE) [7]. The LEIE contains information of the individuals and the entities that are currently excluded from the participation in Medicare, Medicaid and all other Federal health care programs. Although all physicians in the LEIE are cannot be detected in practical application levels, the evaluation metrics are above random. Thus our methodology can be used as a first line of further investigation. Source codes of our works are available online. [1]

In Section 2, we note what is defined as health care fraud and existing works regarding

[1] https://github.com/Ryosuke-Kawamori/medicare

health care fraud detection. In Section 3, we describe datasets we use in this study and show statistics about the healthcare payments. In Section 4, we outline outlier detection algorithms and evaluation metrics. In Section 5, we show the effectiveness of our methods. In Section 6, we summarize our research and give possible future works.

2 Health care fraud and Existing works

2.1 Health care fraud

According to the National Health Care Anti-Fraud Association (NHCAA) [8], health care fraud is "an intentional deception or misrepresentation that the individual or entity makes knowing that the misrepresentation could result in some unauthorized benefit to the individual, or the entity or to some other party"

There are three parties that may commit fraud: (a) service providers including doctors, hospitals, ambulance companies, and laboratories; (b) insurance subscribers, including patients and patients ' employers; and (c) insurance carriers, who receive regular premiums from their subscribers and pay health care costs on behalf of their subscribers, including government health departments and private insurance companies [1].

NHCAA highlights the type of frauds listed below:

1. Billing for services, procedures and/or supplies that were never provided or performed
2. Intentionally misrepresenting any of the following, for purposes of obtaining a payment-or a greater payment-to which one is not entitled:
    - The nature of services, procedures and/or supplies provided or performed;
    - The dates on which services and/or treatments were rendered;
    - The medical record of service and/or treatment provided;
    - The condition treated or the diagnosis made;
    - The charges for services, procedures and/or supplies provided or performed;
    - The identity of the provider or the recipient of services, procedures and/or supplies.
3. The deliberate performance of medically unnecessary services for the purpose of financial gain.

Furthermore, it is noted that there are two different types of fraud: "hit-and-run " and "steal a little, all the time " [1, 9]."Hit-and-run " perpetrators simply submit many fraudulent claims, receive payment, and disappear."Steal a little, all the time " perpetrators work to ensure fraud goes unnoticed and bill fraudulently over a long period of time. Therefore, data, a data-preprocessing method, and a statistical model that detect fraudulent activity heavily depend on who commits fraud and what types of fraud we want to detect.

2.2 Existing Works

To combat fraud, there are three categories of intervention: Prevent, Detect, and Respond. These are listed in Tab. 1 [10]. Prevention interventions refer to "the interventions that deter potential fraudsters from attempting fraud, and stopping a fraud attempt before the fraud is actually committed." For example, creating an anti-fraud culture and developing compliance systems are in this area. Fraud detection involves "identifying past and new cases of fraud as quickly as possible after a fraud has been committed." For example, detecting fraud by data-mining methods or insider reporting systems are in this area. Response means "administrative and legal actions based on the detection and investigation of the fraudulent cases in order to redress the lost money, fine the fraudsters, and sanction legal punishments to prevent future fraud." For example, changing and improving a system or law enforcement initiatives so that the chance of future fraud is reduced are in this area. Most of the existing works focus on the effectiveness of detection [11]. Our research is also categorized as detection. We try a data extraction method before conducting a detailed analysis of the experts' research. Here we see existing works regarding fraud detection.

Table 1: Three categories of Interventions to combat fraud.

Category	
Prevention	Prevent instances of fraud and misconduct from occurring in the first place.
Detection	Detect instances of fraud and misconduct when they do occur.
Response	Respond appropriately and take corrective action when integrity breakdown arise.

Supervised method

Supervised algorithms are used to categorize datasets into fraudulent cases and normal cases. Ortega and He used a neural network [12, 13] and Bonchi used a decision tree [14]. These methods need target data, which is records of activities that were detected as fraudulent by the specialists. Richard [15] used Random forest with respect to various undersampling ratio data of providers because the Medicare dataset is highly unbalanced, meaning that most of the providers are labeled normal and a tiny number of providers are labeled fraudulent. Richard [16] combined three data sources of Medicare procedures and applied Logistic regression, Gradient boosting and Random forest.

Unsupervised method

Rasim [17] used demographic information and a quantile as a fraudulent score. Richard [18] used an Isolation Forest, Local Outlier Factor, Unsupervised Random Forest, Autoencoder, and $k-$Nearest Neighbor, and evaluated their methods using a label provided by NHCAA of a provider who committed fraud. Van [3] used a more story-based method. For example, they used the deviations from a simple linear model between the total dollar amount reimbursed and the number of reimbursed claims of a provider. They also used the percentage of dental claims for a

specific tooth code. This idea was based on a report that some dentists repeatedly made claims for the same set of procedures by changing patient IDs in order to reimburse as much as possible with the least effort involved. Seo [19] constructed a graph where nodes represent provider and edges represent their similarities and apply a PageRank algorithm to the network. They defined the outliers as nodes that have different specialties from seed nodes of PageRank and have a high PageRank score. Ekina [20] used a Bayesian co-clustering method to simulated data, where the providers and the beneficiaries are represented as network nodes.

Others

Pande [21] outlined the characteristics of physicians who were convicted of fraud, the consequences of their convictions, and the proposed problems of policies against fraud. Li [1] gave a survey, which includes a background of the health care fraud, data management, ,models to detect fraud, and their evaluation.

3 Data

3.1 Description of datasets

Here we describe datasets we use. In order to detect fraud in health care receipts submitted by physicians registered in the US, we used three types of data: Medicare Provider Utilization and Payment Data from 2014, Physician and Other Supplier (Physician and Other Supplier PUF) [22], Full Replacement Monthly NPI file [23], and all US zip codes with their corresponding latitude and longitude coordinates processed by MABLE/Geocorr2K ver1.3.3 [24].

The first set of data, Physician and Other Supplier PUF records, contains information on services and procedures provided to Medicare beneficiaries by physicians and other healthcare professionals. Each row records information on utilization, payment and submitted charges organized by their National Provider Identifier (NPI) and Healthcare Common Procedure Coding System (HCPCS) code. The Physician and Other Supplier PUF information has been made public for the period 2012 to 2016. For the sake of simplicity, we use only data from 2015. This data consists of 9,497,892 records and is used to make physicians' feature vectors. Each row records how much money each provider submitted for each operation in a year [25]. Tab. 2 describes the columns of this data and Tab. 5 shows an example of this data. For example first row of Tab. 5 indicates that a provider whose NPI code is 101111 submitted 10 dollars on average for a service with an HCPCS code of 91312.

The second data set is the full replacement monthly NPI file, includes 5,400,369 records and provides the specialties of the providers. Each row records the specialties of each provider. For example, internal medicine is categorized under specialties such as addiction medicine and bariatric medicine. Tab. 3 describe the columns of this data and Tab. 6 shows the example of this data. For example, The first row of Tab. 6 indicates that a provider whose NPI code is 101111 has only one specialty with code 207X00000X.

The third data set, US zip codes, has 43,723 records. Each row records the longitude and latitude of a US zip code. We use this data to calculate which grid square code [26] each provider belongs to. Tab. 7 shows an example of this data. For example, the first row of Tab. 7 indicates

that the longitude and latitude of zip code: 01008 are -72.402004 and 42.277543 respectively. These three datasets have more columns, but the columns which we do not focus on are omitted.

To evaluate our methodology, we use the List of Excluded Individuals and Entities (LEIE). LEIE contains information on individuals and entities that were excluded from participation in Medicare, Medicaid, and all other federal health care programs in the period from 1977 to 2018 [7]. This data contains 70,491 records. Each row records the excluded physicians, excluded date, and the reason for their exclusion. Because providers who commit fraud are considered to be excluded from registration after their illegal activity, we use the providers' label of which their exclusion dates are after 2015.

Tab. 8 shows an example of the LEIE data. For example, the first row of Tab. 8 indicates that the provider whose NPI code is 101111 was excluded from participation in Medicare on 19/04/2018, and the reason for the exclusion is code 1128b7. The reasons for exclusion have 22 types. All exclusion reasons are in Appendix A.3. We do not use excluded physicians' records that are not related to fraud such as physicians who are excluded by the failure to disclose required information. We use only fraud-related and excluded physicians whose exclusion reasons are in Tab. 8 as in [17].

Table 2: Part of column description of Medicare Provider Utilization and Payment Data [25]

ID	Explanation
npi	National Provider Identifier (NPI) for the performing provider on the claim.
nppes_provider_zip	The provider's zip code, as reported in NPPES.
hcpcs_code	HCPCS code used to identify the specific medical service furnished by the provider.
nppes_entity_code	Type of entity reported in NPPES. An entity code of 'I' identifies providers registered as individuals and an entity type code of 'O' identifies providers registered as organizations.
nppes_provider_gender	When the provider is registered in NPPES as an individual (entity typecode='I'), this is the provider's gender.
nppes_provider_state	The state where the provider is located, as reported in NPPES.
provider_type	Derived from the provider specialty code reported on the claim.
line_srvc_cnt	Number of services provided; note that the metrics used to count the number provided can vary from service to service.
bene_unique_cnt	Number of distinct Medicare beneficiaries receiving the service.
bene_day_srvc_cnt	Number of distinct Medicare beneficiary/per day services.
average_submitted_chrg_amt	Average of the charges that the provider submitted for the service.
average_medicare_payment_amt	Average amount that Medicare paid after deductible and coinsurance amounts have been deducted for the line item service.

Table 3: Part of column descriptions of Full Replacement Monthly NPI file.

ID	Explanation
Npi	National Provider Identifier (NPI) for the performing provider on the claim.
Healthcare Provider Taxonomy Code_1	first specialty of the provider indicated by the NPI code.
Healthcare Provider Taxonomy Code_2	second specialty of provider indicated by the NPI code.

Table 4: Part of column descriptions of all US zip code with their corresponding latitude and longitude coordinates.

ID	Explanation
zip	Zip code indicated by 5 digits.
lon	Longitude of corresponding zip code.
lat	Latitude of corresponding zip code.

Table 5: Examples of Medicare Provider Utilization and Payment Data. The raw data contains 9,497,882 rows and 19 columns.

Npi	hcpcs_code	···	average_medicare_submitted_amount	line_srvc_cnt
101111	91312	···	10	3
101111	91313	···	10	4
101111	91314	···	25	2
101112	91312	···	50	2

Table 6: Examples of Full Replacement Monthly NPI file. The raw data contains 5,400,369 rows and 15 columns.

npi	Healthcare Provider Taxonomy Code_1	Healthcare Provider Taxonomy Code_2	···
101111	207X00000X	NA	···
101112	207RC0000X	NA	···
101113	174400000X	207RH0003X	···

Table 7: Examples of zip code to its longitude and latitude file. The raw data contains 43,723 rows and 7 columns.

zip	lon	lat
01008	-72.402003	42.277543
01008	-72.954983	42.184969
01010	-72.204899	42.12831

Table 8: Examples of the List of Excluded Individuals and Entities. The raw data contains 70,491 rows and 18 columns. 15 of 5 columns are omitted.

NPI	EXCLDATE	EXCLTYPE
101111	20180419	1128$b7$
000000	19940524	1128$b5$
184783	20110818	1128$a1$

Table 9: Fraud-related reasons of exclusion.

EXCLTYPE	Amendment
1128a1	Conviction of program-related crimes. Minimum Period: 5 years
1128a2	Conviction relating to patient abuse or neglect. Minimum Period: 5 years
1128a3	Felony conviction relating to health care fraud. Minimum Period: 5 years
1128b4	Felony conviction relating to controlled substance. Minimum Period: 5 years

3.2 Data Combination

We combine the three types of datasets (Medicare Provider Utilization and Payment Data, Full Replacement Monthly NPI, and zip code) as shown in Fig. 1 We split data from the provider ' s department, number of specialties, and location. We have two reasons for using this data-split method. First, handling all providers simultaneously is very time consuming because LOF has $O(n \log n)$ time complexity on the low-dimensional data and One-SVM algorithm has $O(n^3)$ computational complexity, where n is the number of the data points. Even though we split the data, the complexity does no decrease in the order of n. They decrease to $O(n \log \frac{n}{M})$ and $O(\frac{n^3}{M^2})$, where M is the number of data split. However the number of provider n is limited, the reduce of the time is effective and this is effective when we use parallel computing. Second, the variance in the provider' s feature vectors come from the provider's types, specialties, geographical location, and so on. With regard to geographical variance, we can easily assume that the providers in urban areas file more claims because they have many patients. It is also pointed out that spending per Medicare beneficiary also has a high variance [27]. Thus, if we try to find fraudulent claims from all providers, we may find providers who have rare specialties or who are in rural areas. These are the reasons to split the data. To split the data by location, we calculated the first grid square code [26] from their geographical locations (latitude and longitude) by converting their zip codes.

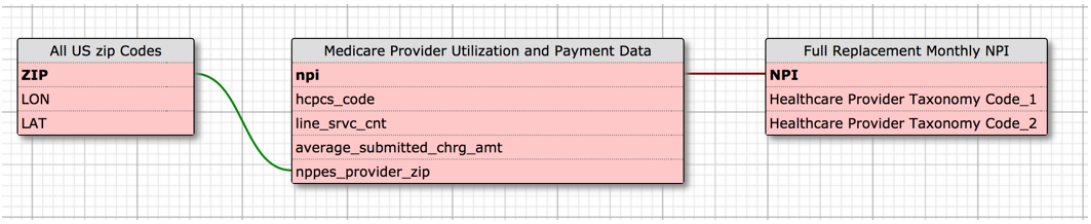

Figure 1: Conceptual illustration of data frame combination.

3.3 Data understanding

Variation by provider specialty

To see variation by provider type, we see histograms of the submitted amount of money. Fig. 2 shows part of the histograms of the total submitted amount of money by provider type. The

provider type has 141 kinds[2] [28]. Here we see part of the histograms of all provider types. We can easily see that different types of providers have very different histograms. For example, Peripheral Vascular Disease has a high dollar amount. The mean amount is 719,014 USD. On the other hand, a Certified Nurse Midwife has a low dollar amount. The mean amount is 9,503 USD.

[2]Physician/General Practice, Physician/General Surgery, Physician/Allergy/ Immunology, Physician/Otolaryngology, Physician/Anesthesiology, Physician/Cardiovascular Disease (Cardiology), Physician/Dermatology, Physician/Family Practice, Physician/Interventional Pain Management, Physician/Gastroenterology, Physician/Internal Medicine, Physician/Osteopathic Manipulative Medicine, Physician/Neurology, Physician/Neurosurgery, Speech Language Pathologist, Physician/Obstetrics & Gynecology, Physician/Hospice and Palliative Care, Physician/Ophthalmology, Oral Surgery (Dentist only), Physician/Orthopedic Surgery, Clinical Cardiac Electrophysiology, Physician/Pathology, Physician/Sports Medicine, Physician/Plastic and Reconstructive Surgery, Physician/Physical Medicine and Rehabilitation, Physician/Psychiatry, Physician/Geriatric Psychiatry, Physician/Colorectal Surgery (Proctology), Physician/Pulmonary Disease, Physician/Diagnostic Radiology, Anesthesiology Assistant, Physician/Thoracic Surgery, Physician/Urology, Chiropractic, Physician/Nuclear Medicine, Physician/Pediatric Medicine, Physician/Geriatric Medicine, Physician/Nephrology, Physician/Hand Surgery, Optometry, Certified Nurse Midwife, Certified Registered Nurse Anesthetist (CRNA), Physician/Infectious Disease, Mammography Center, Physician/Endocrinology, Hospital-Psychiatric Unit, Independent Diagnostic Testing Facility (IDTF), Podiatry, Ambulatory Surgical Center, Nurse Practitioner, Medical Supply Company with Orthotist, Medical Supply Company with Prosthetist, Medical Supply Company with Orthotist-Prosthetist, Other Medical Supply Company, Individual Certified Orthotist, Individual Certified Prosthetist, Individual Certified Prosthetist-Orthotist, Medical Supply Company with Pharmacist, Ambulance Service Provider, Public Health or Welfare Agency, Voluntary Health or Charitable Agency, Psychologist, Clinical, Portable X-Ray Supplier, Audiologist, Physical Therapist in Private Practice, Physician/Rheumatology, Occupational Therapist in Private Practice, Psychologist, Clinical, Clinical Laboratory, Clinic or Group Practice, Registered Dietitian or Nutrition Professional, Physician/Pain Management, Mass Immunizer Roster Biller, Radiation Therapy Center, Slide Preparation Facility, Physician/Peripheral Vascular Disease, Physician/Vascular Surgery, Physician/Cardiac Surgery, Physician/Addiction Medicine, Licensed Clinical Social Worker, Physician/Critical Care (Intensivists), Physician/Hematology, Physician/Hematology-Oncology, Physician/Preventive Medicine, Physician/Maxillofacial Surgery, Physician/Neuropsychiatry, All Other Suppliers, Unknown Supplier/Provider Specialty, Certified Clinical Nurse Specialist, Physician/Medical Oncology, Physician/Surgical Oncology, Physician/Radiation Oncology, Physician/Emergency Medicine, Physician/Interventional Radiology, Advance Diagnostic Imaging, Optician, Physician Assistant, Physician/Gynecological Oncology, Physician/Undefined Physician type, Hospital-General, Hospital-Acute Care, Hospital-Children ' s (PPS excluded), Hospital-Long-Term (PPS excluded), Hospital-Psychiatric (PPS excluded), Hospital-Rehabilitation (PPS excluded), Hospital-Short-Term (General and Specialty), Hospital-Swing Bed Approved, Hospital-Rehabilitation Unit, Hospital-Specialty Hospital (cardiac, orthopedic, surgical), Critical Access Hospital, Skilled Nursing Facility, Intermediate Care Nursing Facility, Other Nursing Facility, Home Health Agency, Home Health Agency (Subunit), Pharmacy, Medical Supply Company with Respiratory Therapist, Department Store, Grocery Store, Indian Health Service facility, Oxygen supplier, Pedorthic personnel, Medical supply company with pedorthic personnel, Rehabilitation Agency, Organ Procurement Organization, Community Mental Health Center, Comprehensive Outpatient Rehabilitation Facility, End-Stage Renal Disease Facility, Federally Qualified Health Center, Hospice, Histocompatibility Laboratory, Outpatient Physical Therapy/Occupational Therapy/Speech Pathology Services, Religious Non-Medical Health Care Institution, Rural Health Clinic, Physician/Sleep Medicine, Physician/Interventional Cardiology, Dentist, Physician/Hospitalist, Physician/Advanced Heart Failure and Transplant Cardiology, Physician/Medical Toxicology, Hematopoietic Cell Transplantation and Cellular Therapy

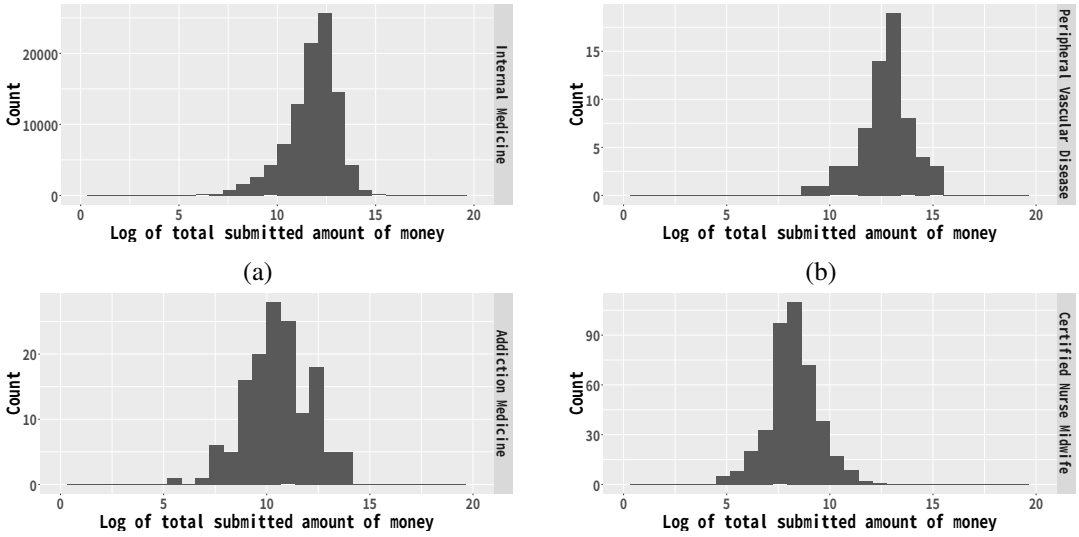

Figure 2: (a) depicts the histograms of the log of the total submitted amount of money of internal medicine. (b) depicts that of Peripheral Vascular Disease. (c) depicts that of Addiction Medicine. (d) depicts that of Certified Nurse Midwife.

Geographical variance

To see variation by geographical location, we computed the mean of the submitted amount of money in each grid square. The mean grid square statistics of the observed values x_{ij}^k, where i is an index of the observed value, j is a data feature, and k is a grid square number, is defined as follows :

$$m_j^k = \frac{1}{N_k} \sum_{i=1}^{N_k} x_{ij}^k, \tag{1}$$

where N_k is a number of observed points in the grid squre k. Fig. 3 shows the mean grid square of the total submitted amount of money on each grid k, and Fig. 4 shows a histogram of the mean grid square of the total submitted amount of money. These figures indicate that there is a strong geographical dependence with regard to the total submitted amount of money.

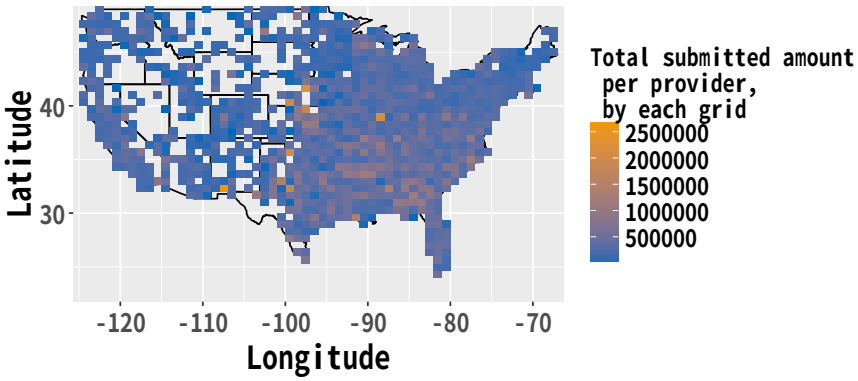

Figure 3: Mean of the total submitted amount of money in each grid. Red corresponds to the high submitted amount of money, and blue corresponds to the low submitted amount of money. We can see that the total submitted amount of money is high in urban areas and low in rural areas.

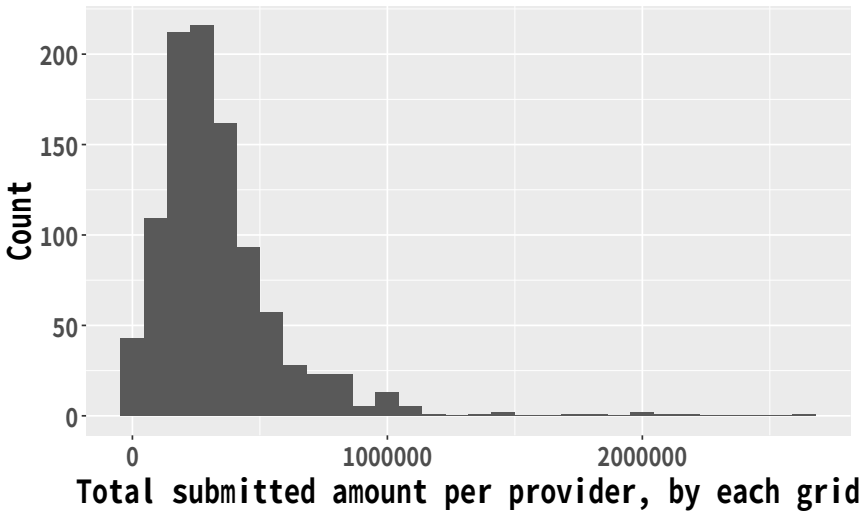

Figure 4: Histograms of the total submitted amount of money. The x-axis shows the mean of the total submitted amount in each grid. We can see that histogram has a fat tail, and the submitted amount of money is dependent on the area in the US.

Variation by providers ' number of specialties

Fig. 5 shows a histogram of the total submitted amount of money by the provider ' s specialty. We can see that variations by the provider ' s number of specialty are not as high as by the

geographical location and the type of providers. Thus, we may not need to split data by the provider's number of specialties.

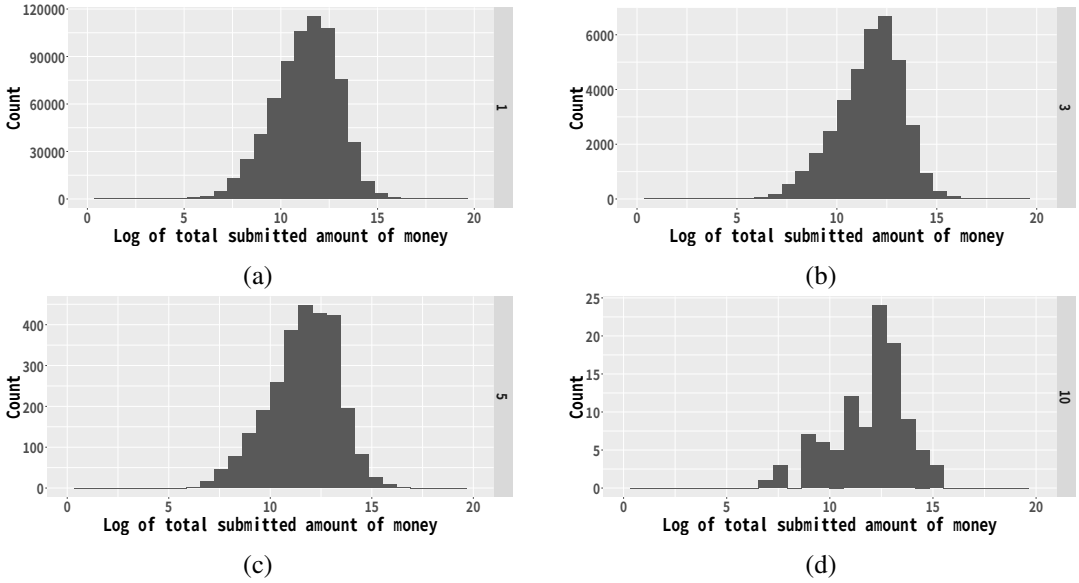

Figure 5: (a) depicts the histogram of the log of the total submitted amount of money by providers who have one specialty. (b) depicts that by providers who have 3 specialties. (c) depicts that by providers who have 5 specialties. (d) depicts that by providers who have 10 specialties.

4 Method

4.1 Subset of all data for evaluation

We focus on providers in New York and California because we can obtain fraud label data to evaluate the efficiency of our methods due to strong fraud detection activities in these states. Our main purpose is to extract candidates of fraud before specialists' investigations. Thus, in practical applications we should apply our extraction methods to the data from states where fraud enforcement is weak as well. However, we need metrics that evaluate our unsupervised strategy. We target individual providers in the states where fraud detection enforcement is strong. Tab. 10 shows the descriptive statistics of our subset of all data used throughout our investigation.

Table 10: Descriptive statistics of our subset of all data used throughout our investigation.

Number of records	1111686
Number of providers	123299
Number of HCPCS codes performed	4104
Number of providers who committed fraud	46

4.2 Feature engineering

Feature engineering of existing works

Richard [18] calculated summary statistics of providers' procedure and construct provider-level feature. They used the mean, median, max, min, sum, and standard deviation of the number of services provided, the number of distinct Medicare beneficiaries, the number of distinct Medicare beneficiary/per day services, the average of the Medicare allowed amount for the service, the average amount that Medicare paid after deductible and coinsurance amounts have been deducted for the line item service, the one-hot encoding of the providers' gender, and specialities. We call this feature engineering strategy STAT(statistics of all procedures).

Feature consisting of the submitted amount of each procedure

From level 1: Single Claim, or Transaction to level 7: Multiparty, Criminal Conspiracies, there are seven levels of healthcare fraud control [1, 9]. Tab. 11 shows these seven levels of focus. However, one row of our data describe what kind of procedure one provider operated in a year and we cannot obtain single claim, we can look at datasets only above or equal to level 3. Above level 3, for example, we can pay attention to the medical group whose providers belong to the same hospital. Here, for the sake of simplicity, we use level3b: One provider and all of its claims and related patients and do not use data above level3.

Table 11: Levels of healthcare fraud control proposed by [1]. This table was constructed by [9].

		Level Focus
Level 1	Single Claim, or Transaction	The claim itself and the related provider and the patient.
Level 2	Patient/Provider	One patient, one provider, and all of their claims.
Level 3	a. Patient	One patient and all of its claims and related providers.
	b. Provider	One provider and all of its claims and related patients.
Level 4	a. Insurer Policy / Provider	Patients that are covered by the same insurance policy and are targeted by one provider.
	b. Patient / Provider Group	One patient being targeted by multiple providers within a practice.
Level 5	Insurer Policy / Provider Group	Patients with the same policy being targeted by multiple providers within a practice.
Level 6	a. Defined Patient Group	Groups of patients being targeted by providers. (e.g. patients living in the same location)
	b. Provider Group	Groups of providers targeting their patients. Groups can be providers within the same practice, clinics, hospitals, or other arrangements.
Level 7	Multiparty, Criminal Conspiracies	Multiparty conspiracies that could involve many relationships.

We introduced grid separation Eq. (1) to calculate grid square statistics. This is an instance of splitting data in terms of geographical categories. Here let us introduce general data-split k, where k is running from 1 to M. Let us define the i-th data points: $\boldsymbol{x}_i^k = (x_{i1}^k, x_{i2}^k, \ldots, x_{il_k}^k)$ as feature vector in l_k dimensional space, where k is a data-split index. The j-th element of $\boldsymbol{x}_i^k$ is the log-transformed and normalized total submitted amount of money for each service,

$$
\begin{aligned}
x_{ij}^{k\,\prime} = \ln \big(& (\text{Average of the charges that the provider submitted for the service of the } j\text{-th hcpcs\_code} \\
& \text{of } i\text{-th provider with data-split } k) \\
& \times (\text{Number of services provided of the } j\text{-th hcpcs\_code of } i\text{-th provider with data-split } k) \big), \\
x_{ij}^k = & \frac{x_{ij}^{k\,\prime} - m_{jk}}{\sigma_{jk}},
\end{aligned}
\tag{2}
$$

where m_j^k is a mean of j-th data feature with data-split k and σ_j^k is the that of standard deviation. Tab. 12 shows the providers ' features calculated from Tab. 5. We call this feature engineering strategy HCPCS (Each amount of HCPCS procedure).

Table 12: Examples of feature-based data before log transformation and normalization.

NPI	99312	99313	$\cdots$	99314
10111111	30	40	$\cdots$	50
10111112	0	0	$\cdots$	100

4.3 Data-split method

As noted in the previous section, the submitted amount of money have variances with regard to the physician's geographical location, type and number of specialities. We focus on physicians who have only one specialty, and we split the data by their location and specialty. We create a physician feature vector from this split-data frame and apply an outlier detection method. There are 6599 codes in HCPCS code set [29] and the length feature vector is assumed to be 6599. However, some procedures are not performed when we focus on some split-data, then we omit these procedures from columns of the feature vector.

For example, the feature vector of the internal medicine physician with grid square code 306173 has 907 elements, but the feature vector of the same physicians with grid square code 304182 has 420 elements. We apply an outlier detection method to each split data item and combine the results into one data frame in order to calculate the evaluation metrics. The conceptual image of this method is shown in Fig. 6. To check the efficiency of the data-split method based on geographical location, we compare three versions: PROVIDER (split only by physician ' s specialty), GRID (split only by physician's grid square code), and PROVIDER-GRID (split by physician ' s specialty and grid square code). We removed split data whose number of data points is smaller than 20. Tab. 13 shows the number of data-split M of each data-split method of our subset.

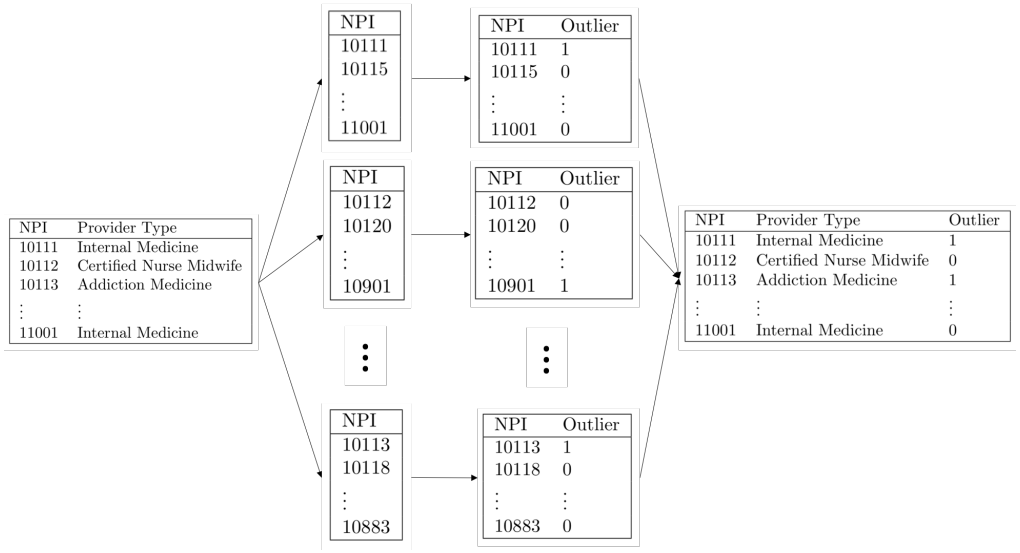

Figure 6: The conceptual illustration of the data-split method. We split data by their type and location, make physicians' feature vector from this split data frame and apply an outlier detection method.

Table 13: Number of data-split M of each data-split method.

Data-split method	M
Provider	66
Mesh	56
Provider-Mesh	967

4.4 One Support Vector Machine

Here we remove the index of data-split k for clarity. Algorithms are applied in each data-separation k. To detect fraudulent provider activity, we use a one-class support vector machine [5] as an outlier extraction method. One support vector machine separates data from the origin to find the decision function that takes +1 in a region capturing most of the data points and -1 elsewhere. Mapping data into the feature space by function $\boldsymbol{\Phi}(\boldsymbol{x}) : X \to F$, where X is an input space and F is a feature space that has an inner product and applying the one support vector machine in this space enables us to solve a nonlinear problem. One support vector machine is expressed as below,

$$\begin{aligned} \min_{\boldsymbol{w}\in F, \boldsymbol{\xi}\in\mathbb{R}^N, \rho\in\mathbb{R}} \quad & \frac{1}{2}||\boldsymbol{w}||^2 + \frac{1}{\nu N}\sum_{i=1}^{N}\xi_i - \rho \\ \text{subject to} \quad & (\boldsymbol{w}, \boldsymbol{\Phi}(\boldsymbol{x}_i)) \geq \rho - \xi_i, \xi_i \geq 0. \end{aligned} \tag{3}$$

Here, $||\cdot||^2$ denotes the l^2-norm of a vector, $(\boldsymbol{w}, \boldsymbol{\Phi}(\boldsymbol{x}_i))$ denotes the dot product of two vectors, $\boldsymbol{w}$ and $\boldsymbol{\Phi}(\boldsymbol{x}_i)$, in inner product space F, N is a number of data points, and ν is a hyperparameter. The decision function is

$$f(\boldsymbol{x}) = \mathrm{sgn}((\boldsymbol{w}, \boldsymbol{\Phi}(\boldsymbol{x})) - \rho), \tag{4}$$

where

$$\mathrm{sgn}(x) = \begin{cases} +1 & (x > 0) \\ 0 & (x = 0) \\ -1 & (x < 0) \end{cases}. \tag{5}$$

To solve main problem Eq. (3) we derive dual problem. The Lagrange function is given as follows :

$$L(\boldsymbol{w}, \boldsymbol{\xi}, \rho, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}||\boldsymbol{w}||^2 + \frac{1}{\nu N}\sum_{i=1}^{N}\xi_i - \rho - \sum_{i=1}^{N}\alpha_i((\boldsymbol{w}, \boldsymbol{\Phi}(\boldsymbol{x}_i)) - \rho + \xi_i) - \sum_{i=1}^{N}\beta_i\xi_i. \tag{6}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_N)$, where N is the number of data points, are dual variables. By taking derivative of the main variables $\boldsymbol{w}, \xi$, and ρ, we get,

$$\boldsymbol{w} - \sum_{i=1}^{N}\alpha_i\boldsymbol{\Phi}(\boldsymbol{x}_i) = 0, \tag{7}$$

$$\frac{1}{\nu l} - \alpha_i - \beta_i = 0, \tag{8}$$

$$-1 + \sum_{i=1}^{N}\alpha_i = 0. \tag{9}$$

By using these equations, we get the dual of this program, given as follows :

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \min_{\boldsymbol{w},\boldsymbol{\xi},\rho} L(\boldsymbol{w}, \boldsymbol{\xi}, \rho, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \min_{\boldsymbol{\alpha}} \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j(\boldsymbol{\Phi}(\boldsymbol{x}_i), \boldsymbol{\Phi}(\boldsymbol{x}_j))$$

$$\text{subject to} \qquad 0 \le \alpha_i \le \frac{1}{\nu}, \quad \sum_{i=1}^{N}\alpha_i = 1. \tag{10}$$

By substituting Eq. (7) into Eq. (4), we get

$$f(\boldsymbol{x}) = \mathrm{sgn}(\sum_{i=1}^{N}\alpha_i(\boldsymbol{\Phi}(\boldsymbol{x}_i), \boldsymbol{\Phi}(\boldsymbol{x})) - \rho). \tag{11}$$

Here we use Gaussian kernel, defined as

$$k(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{\Phi}(\boldsymbol{x}), \boldsymbol{\Phi}(\boldsymbol{y})) = e^{-\frac{||\boldsymbol{x}-\boldsymbol{y}||^2}{c}}. \tag{12}$$

As we can see from the dual problem, we do not need to calculate $\boldsymbol{\Phi}(\boldsymbol{x})$ directly, and we only need the values of kernel function $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ when solving this problem. This technique is called the kernel method. We solve the dual problem expressed by Eq. (10) using R package e1071 [30]. We set hyperparameters $c = \frac{1}{N}$ and $\nu = 0.1$(package default). One support vector machine has two favorable properties. First, ν gives an upper bound of the ratio of the outliers, that is, the training points outside the estimated region. Second, there is a probabilistic guarantee that new points fall inside the estimated region.(For more details and proofs, see [5].)

4.5 Local Outlier Factor

LOF calculates the ratio of the density of the object and its surroundings. The $k-$distance of an object p is defined as

$$k - distance(p) = d(p, o), \tag{13}$$

where $d(p, o)$ is some distance measure between an object p and an object o such that:

1. for at least k objects $o' \in D \setminus \{p\}$ it holds that $d(p, o') \leq d(p, o)$ and

2. for at most $k - 1$ objects $o' \in D \setminus \{p\}$ it holds that $d(p, o') < d(p, o)$,

where D represents the set of all data points. The $k-$distance neighborhood of an object p is defined as

$$N_k(p) = \{q \in D \setminus p \mid d(p, q) \leq k - distance(p)\}. \tag{14}$$

The local eachability distance of an object p with respect to object r is defined as

$$reach - dist_k(p, r) = \max\{k - distance(p), d(p, r)\}. \tag{15}$$

Local reachability density of an object p is defined as

$$lrd_{MinPts}(p) = \left\{ \frac{\sum_{q \in N_{MinPts}(p)} reach - dist_{MinPts}(p, q)}{\mid N_{MinPts}(p) \mid} \right\}^{-1}. \tag{16}$$

This is the inverse of the mean of $reach - dist_{MinPts}$ from p to its surroundings. So, this can be regarded as the density around point p. Local outlier factor of an object p is defined as the mean of the ratio of local density around p and its surroundings o,

$$LOF_{MinPts}(p) = \frac{\sum_{q \in N_{MinPts}(p)} \frac{lrd_{MinPts}(q)}{lrd_{MinPts}(p)}}{\mid N_{MinPts}(p) \mid} \tag{17}$$

In our application, the data points are represented as vectora in $\mathbb{R}^l$ space and distance is Euclid distance, and set hyperparameter $MinPts = 20$. We use R package dbscan [31] to calculate LOF.

4.6 Evaluation Metric

We evaluate the effectiveness of data our data-extraction method using the true positive rate, the false negative rate, Receiver Operating Characteristic (ROC) curve, and Area Under Curve (AUC). The true positive rate(TPR) and the false negative rate (FNR) are defined as follows :

$$TPR = \frac{TP}{TP + FN}, \tag{18}$$

$$FPR = \frac{FN}{FN + TP}, \tag{19}$$

where TP is the number of physicians that are truly classified as fraudulent and FN is the number of physicians that are falsely classified as normal. The ROC curve created by plotting TPR against the FPR and the AUC is the area of the ROC curve. If physicians are randomly selected, the ROC curve is a straight line through two points $(0, 0)$ and $(1, 1)$ and the AUC is 0.5. If we can correctly classify physicians, the ROC curve is above that line and the AUC is above 0.5. When applying One-SVM, we cannot get the outlier score but only labels and cannot draw the ROC curve in a standard way. However, hyperparameter ν works as an upper bound of the ratio of the outliers, thus, TPR and FPR assumed to increase as ν increases. We plot TPR against the FPR with regard to 20 hyperparameter values ν running from 0.001 to 0.98. We regard this line as a ROC curve.

5 Results

5.1 One-SVM

Fig. 7 shows the ROC curve of the data-split method for PROVIDER and PROVIDER-GRID of STAT and HCPCS. All ROC curves are slightly above the straight line through points $(0, 0)$ and $(1, 1)$. However, they do not change much with respect to the data-split and the feature. Outliers found by One-SVM do not match with the fraudulent activities.

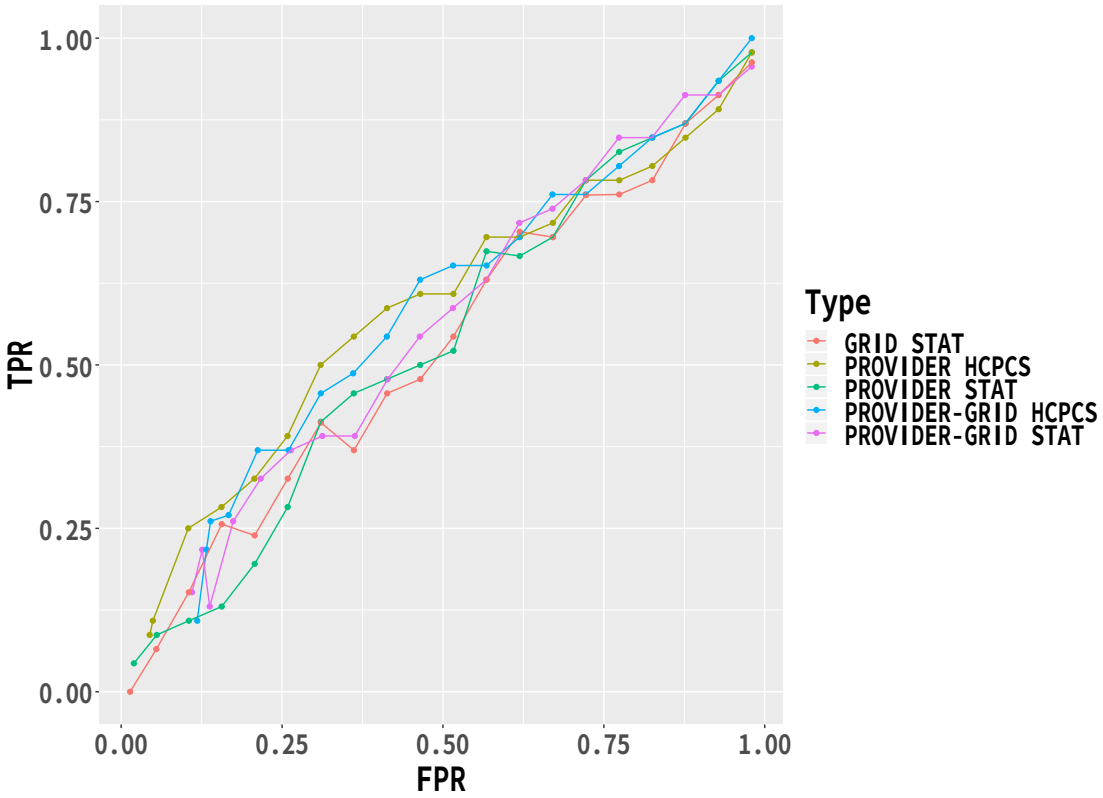

Figure 7: ROC curve of One-SVM with respect to the data-split: PROVIDER and PROVIDER-GRID of the feature: STAT and HCPCS. Type shows the data-split method and how features are created. The x-axis shows true positive rate, and the y-axis shows false positive rate. All ROC curves are slightly above straight line through points $(0, 0)$ and $(1, 1)$. However, they do not change much with respect to data-splits and features.

5.2 Local Outlier Factor

Fig. 8 shows the ROC curve of the data split: PROVIDER, GRID, and PROVIDER-GRID of the feature: STAT and HCPCS. All ROC curves are above the random extraction. All ROC curves are slightly above the straight line through points $(0, 0)$ and $(1, 1)$. In particular, data-split with data-split: PROVIDER and feature: HCPCS is the best of all.

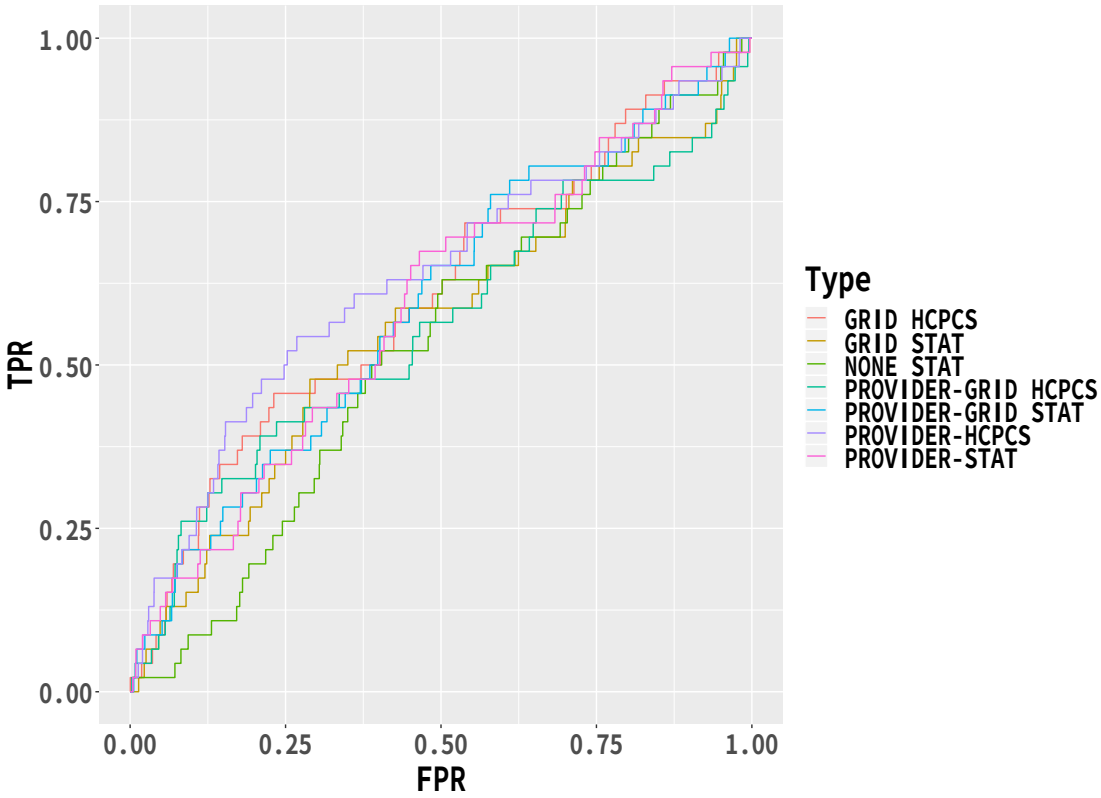

Figure 8: ROC curve of LOF with respect to data split: PROVIDER, GRID and PROVIDER-GRID of the feature: STAT and HCPCS. Type shows their data-split method and how their features are created. The x-axis shows true positive rate, and the y-axis shows false positive rate. None means there was no data split. All ROC curves are slightly above straight line through points $(0, 0)$ and $(1, 1)$. In particular, data split for PROVIDER of the feature with HCPCS had the best result.

5.3 AUC Score

Tab. 14 shows the AUC of the outlier detection models of each data-split and feature. The best AUC was achieved with LOF of data split: PROVIDER and feature : HCPCS. We cannot calculate the data-split: GRID of feature: HCPCS and NONE: without any data-split with the One-SVM because of the time of the calculation. First, with respect to outlier detection algorithms, 3 out of 5 AUC scores of LOF are above the AUC score of One-SVM and 1 out of 5 AUC score of LOF are almost same with the AUC score of One-SVM. Next, with respect to feature, 4 out of 5 AUC scores of the feature: HCPCS are above STAT. Third, with respect to data-split, data-split: PROVIDER-GRID of the feature: STAT and the data-split: PROVIDER of the feature: HCPCS are good combination. There are two reasons for the false negatives. First, a provider who commits fraud steals a little and his or her fraudulent activity is not apparent in the data. However, this provider was excluded owing to the insider report. Second, the fraudulent

submission occurred before 2015, although we used the submission data for after 2015. The first case cannot be captured easily by the data-mining method, but the second case can be captured by a combination of available datasets.

Fig. 8 shows that the ROC curve of the model of the best AUC score is through $(0.25, 0.50)$. This means that if we have 100 fraud cases out of 10000 providers, we need to search 2000 cases to find 50 fraud cases. We have not had any opportunities to interview specialists yet. However it seems that the detection costs are expensive because specialists need to research many non-fraudulent cases to find truly fraudulent cases.

Table 14: AUC of outlier detection models of each data-split and feature.

Outlier Detection Model	Data-Split	Feature	AUC
LOF	NONE	STAT	0.529
LOF	PROVIDER	STAT	0.591
LOF	PROVIDER	HCPCS	0.634*
LOF	PROVIDER-GRID	STAT	0.593
LOF	PROVIDER-GRID	HCPCS	0.559
LOF	GRID	STAT	0.560
LOF	GRID	HCPCS	0.605
One-SVM	NONE	STAT	NA
One-SVM	PROVIDER	STAT	0.513
One-SVM	PROVIDER	HCPCS	0.562
One-SVM	PROVIDER-GRID	STAT	0.535
One-SVM	PROVIDER-GRID	HCPCS	0.561
One-SVM	GRID	STAT	0.509
One-SVM	GRID	HCPCS	NA

LOF with the data-split: PROVIDER and feature: HCPCS has the best AUC score. Here we see more detail result of the model and matches between outliers and frauds. Tab. 15 shows the AUC scores by the provider types. We see several providers who commit fraud. The AUC score of Neurosurgery is the highest. The NPI code of the provider who committed fraud is 1003904830. His score gets top 3 score out of 441 Neurosurgery providers. Fig. 9 shows the histograms of the submitted amount of money on each HCPCS procedure in Neurosurgery and the submitted amount of money with NPI code, 1003904830. He submitted most for these four HCPCS codes. Fig. 9 shows that he submitted more than normal providers with these HCPCS codes and this can be a sign of fraud. Next, in the case of Diagnostic Radiology, the NPI code of provider who committed fraud is 1619117538. His outlier score was top 294 out of the 4298 diagnostic radiology providers. News reports that he wrote more than 250 illegal prescriptions over a four-year period, starting in 2010, for Oxycodone, Percocet and, Hydrocodone [32]. Fig. 10 shows the histograms of the submitted amount of money on each HCPCS procedure in Diagnostic Radiology and the submitted amount of money with NPI code, 1619117538. His records of submission do not record high amounts for these kinds of operations, though he gets higher outlier score than other diagnostic radiology providers. We think that there are two possible reasons for the illegal activities not being recorded in the datasets. First, the illegal submission

procedure is not included in the Medicare program. Second, this illegal operation was stopped before 2015 and were not recorded in our dataset. We cannot explain why he gets high outlier score from the histograms of his submission amounts.

Table 15: AUC Score of each data-split.

Provider type (data-split k)	AUC Score	Number of fraud providers	Number of providers
Neurosurgery	0.995	1	441
General Practice	0.980	1	764
Neurology	0.962	1	2188
Diagnostic Radiology	0.925	1	4928
Cardiology	0.893	1	3604
Obstetrics/Gynecology	0.866	3	3127
Physical Therapist	0.803	1	7719
Optometry	0.732	1	3000
Internal Medicine	0.686	4	17738
Emergency Medicine	0.686	2	6597
General Surgery	0.680	1	2547
Podiatry	0.653	2	2768
Psychiatry	0.626	4	3884
Family Practice	0.602	10	9546
Chiropractic	0.516	6	4321
Physical Medicine and Rehabilitation	0.485	1	1132
Otolaryngology	0.462	2	1173
Anesthesiology	0.384	2	6214
Physician Assistant	0.266	1	6583
Pain Management	0.117	1	112

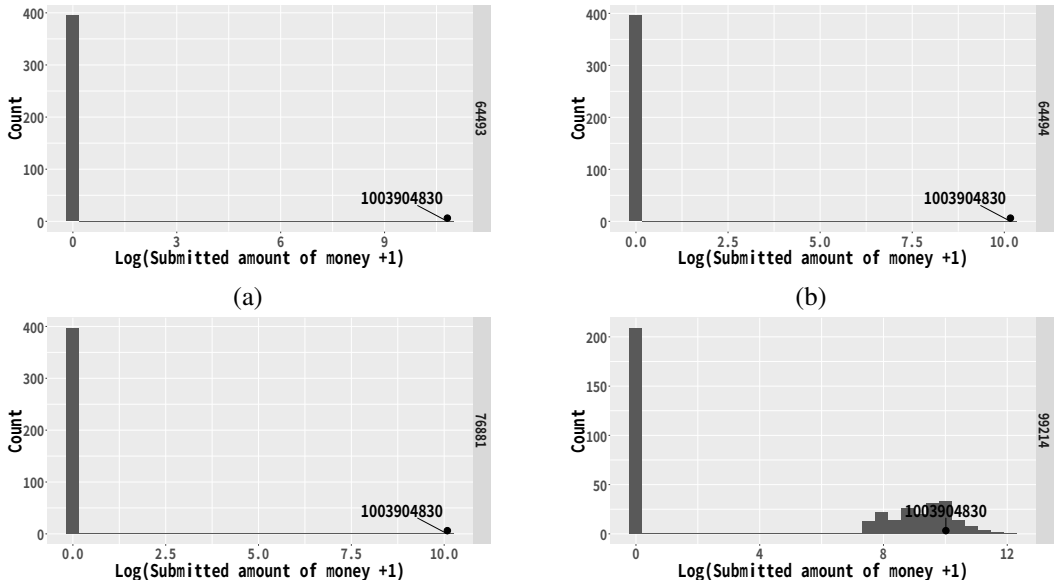

Figure 9: (a) depicts histogram of log(total submitted amount of money + 1) of HCPCS code, 64493(Injections of lower or sacral spine facet joint using imaging guidance) of Neurosurgery providers and the value of fraudulent providers. (b) depicts that of HCPCS code 64494 (Injections of lower or sacral spine facet joint using imaging guidance). (c) depicts that of HCPCS code 76881 (Ultrasound of leg or arm). (d) depicts that of HCPCS code 99214 (Established patient office or other outpatient, visit typically 25 minutes). He submitted more than normal providers with these HCPCS codes and this can be a sign of fraud.

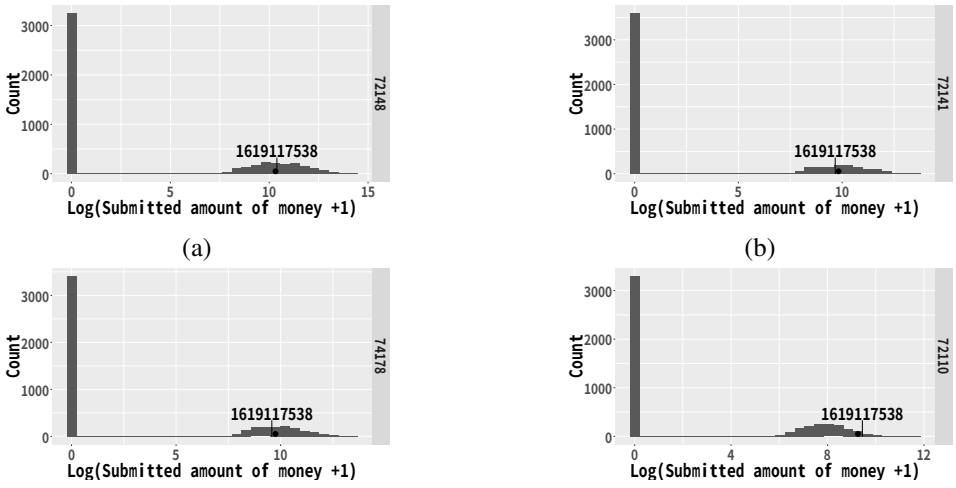

Figure 10: (a) depicts histogram of log(total submitted amount of money + 1) of HCPCS code, 72148(MRI scan of lower spinal canal) of Neurosurgery providers and the value of fraudulent providers. (b) depicts that of HCPCS code 72141 (MRI scan of upper spinal canal). (c) depicts that of HCPCS code 74178 (CT scan of abdomen and pelvis before and after contrast). (d) depicts that of HCPCS code 72110 (X-ray of lower and sacral spine, minimum of 4 views.) We cannot explain why he gets high outlier score from the histograms of his submission amounts.

We evaluated whether outliers and frauds are matched in the healthcare insurance claims. However, the score in other domain can be different from Medicare cases. We applied One-SVM and LOF to credit card fraud detection. Actually, the performance of credit card fraud detection is different from the medicare cases. One-SVM outperformed LOF and the score is significantly high. The result of is not different from random extraction. This implies that the detection performance, matches between fraud and outlier, is determined not only by detection algorithm but also data features.

5.4 Ratio of the outliers in One-SVM

Next, we changed hyperparameter ν of One-SVM at 20 points from 0.001 to 0.98. Fig. 11 shows the ratio of outliers on various ν. As we expected, we can easily see as ν increases, the ratio of outliers also increases. Thus, we can regulate the ratio of the outliers by changing ν.

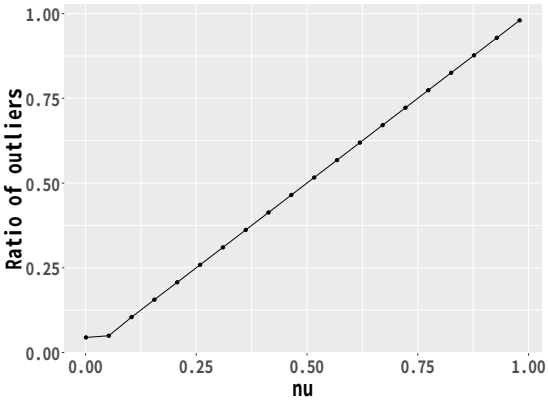

Figure 11: Ratio of the outliers in terms of ν. The x-axis shows the hyperparameter ν which is an upper bound of the ratio of the outliers and the y-axis shows the ratio of outlier detected by One-SVM of which $\nu = x$.

6 Conclusions

We applied data linkage methods to the NPI, Healthcare Provider Taxonomy Core, and zip code. We used world grid square codes to segment regions regarding physicians' geospatial heterogeneities and investigated world grid square statistics about the submitted amount of money and the provider type variance. We applied One-SVM and LOF to the feature vectors, HCPCS: submitted amount of money for each HCPCS code and STAT: summary statistics of the submitted amount for money and one-hot encoding of physicians' characteristics. By dividing large records based on physicians' locations using world grid square codes and types into many segments, we obtained many small datasets about the healthcare receipts. Finally, we evaluated the method using LEIE labels. LOF performs better than One-SVM. LOF with data-split: Provider and feature: HCPCS is the best of all methods. Data-split is also beneficial in terms of the time of calculation.

As future work, first, as we see in the design of feature vectors and their performance, the feature engineering is crucial when we define the fraud claims in the health care receipts. We may need deep knowledge of the healthcare domain in order to select effective features. Second, we should interview healthcare fraud subject matter experts whether providers with higher outlier scores are fraudulent. Some providers who are not in LEIE datasets, but have high outlier score may be the real fraudulent providers. Third, the tradeoff between the true positive rate and the false positive rate should be adjusted. If we allow only low true positive rates, then there are more rooms for benefits by increasing the cost of the experts who find the providers actually committing fraud from fraudulent cases in order to prevent them from receiving money. On the other hand, if we allow a high true positive rate, the cost of the experts increase, and the costs are higher than the money given to the providers who commit fraud. Fourth, more sophisticated datasets are needed. We used Medicare Provider Utilization and Payment Data, Full Replacement Monthly NPI in 2015 and List of Excluded Individuals after 2015. However, if their exclusion resulted from illegal activities before 2015, then their illegal activities are not recorded in the

Provider Utilization and Payment Data, Full Replacement Monthly NPI and List of Excluded Individuals includes providers. In addition List of Excluded Individuals include providers who were excluded not only for their fraud-related activities but for their other illegal activities such as physical contact with their patients. Thus, more sophisticated datasets that record provider activities and frauds related to these activities are needed.

Acknowledgments

The author would like to express his sincere gratitude to Associate Professor Aki-Hiro Sato for his helpful advices and appreciate the feedback offered by the members of my laboratory.

References

[1] Jing Li, Kuei-Ying Huang, Jionghua Jin, and Jianjun Shi. A survey on statistical methods for health care fraud detection. *Health care management science*, 11(3):275–287, 2008.

[2] Passard C. Dean, Josseibel Josseibel Vazquez-Gonzalez, and Lucy Fricker. Causes and challenges of healthcare fraud in the us. *International Journal of Business and Social Science*, 4(14), 2013.

[3] Guido van Capelleveen, Mannes Poel, Roland M. Mueller, Dallas Thornton, and Jos van Hillegersberg. Outlier detection in healthcare fraud: A case study in the medicaid dental domain. *International journal of accounting information systems*, 21:18–31, 2016.

[4] Hyunjung Shin, Hayoung Park, Junwoo Lee, and Won Chul Jhee. A scoring model to detect abusive billing patterns in health insurance claims. *Expert Systems with Applications*, 39(8):7441–7450, 2012.

[5] Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

[6] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. Lof: Identifying density-based local outliers. *SIGMOD Rec.*, 29(2):93–104, 2000.

[7] National Hospice and Palliative Care Organization. OIG's List of Excluded Individuals/Entities (LEIE) [Online]. `https://www.nhpco.org/sites/default/files/public/regulatory/LEIE_Exsclusion_List.pdf`. [Accessed on 11 Feb 2019].

[8] National Health Care Anti-Fraud Association. Consumer Info & Action [Online]. `https://www.nhcaa.org/resources/health-care-anti-fraud-resources/consumer-info-action.aspx`. [Accessed on 11 Feb 2019].

[9] Dallas Thornton, Roland M. Mueller, Paulus Schoutsen, and Jos van Hillegersberg. Predicting healthcare fraud in medicaid: a multidimensional data model and analysis techniques for fraud detection. *Procedia technology*, 9:1252–1264, 2013.

[10] KPMG Fraud risk management. Developing a strategy for prevention, detection, and response [Online]. https://assets.kpmg.com/content/dam/kpmg/pdf/2014/05/fraud-risk-management-strategy-prevention-detection-response-O-201405.pdf. [Accessed 11 Feb 2019].

[11] Arash Rashidian, Hossein Joudaki, and Taryn Vian. No evidence of the effect of the interventions to combat health care fraud and abuse: a systematic review of literature. *PLoS ONE*, 7(8):41988, 2012.

[12] Pedro A Ortega, Cristián J. Figueroa, and Gonzalo A. Ruz. A medical claim fraud/abuse detection system based on data mining: A case study in chile. *DMIN*, 6:26–29, 2006.

[13] Hongxing He, Jincheng Wang, Warwick Graco, and Simon Hawkins. Application of neural networks to detection of medical fraud. *Expert Systems with Applications*, 13(4):329–336, 1997.

[14] Francesco Bonchi, Fosca Giannotti, Gianni Mainetto, and Dino Pedreschi. A classification-based methodology for planning audit strategies in fraud detection. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 175–184. ACM, 1999.

[15] Richard Bauder and Taghi Khoshgoftaar. Medicare fraud detection using random forest with class imbalanced big data. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 80–87, 2018.

[16] Matthew Herland, Taghi Khoshgoftaar, and Richard A. Bauder. Big data fraud detection using multiple medicare data sources. *Journal of Big Data*, 5(1):29, 2018.

[17] Rasim M. Musal. Two models to investigate medicare fraud within unsupervised databases. *Expert Systems with Applications*, 37(12):8628 – 8633, 2010.

[18] Richard Bauder, Raquel da Rosa, and Taghi Khoshgoftaar. Identifying medicare provider fraud with unsupervised machine learning. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 285–292, 2018.

[19] Jiwon Seo and Ofer Mendelevitch. Identifying frauds and anomalies in medicare-b dataset. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3664–3667, 2017.

[20] Tahir Ekina, Francesca Leva, Fabrizio Ruggeri, and Refik Soyer. Application of bayesian methods in detection of healthcare fraud. *chemical engineering Transaction*, 33, 2013.

[21] Vivek Pande and Will Maas. Physician medicare fraud: characteristics and consequences. *International Journal of Pharmaceutical and Healthcare Marketing*, 7(1):8–33, 2013.

[22] Center for Medicare & Medicaid Services. Medicare Provider Utilization and Payment Data: Physician and Other Supplier [Online]. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/

Medicare-Provider-Charge-Data / Physician-and-Other-Supplier . html. f[Accessed on 11 Feb 2019].

[23] Center for Medicare & Medicaid Services. NPI FIles [Online]. http://download.cms.gov/nppes/NPI_Files.html. [Accessed on 11 Feb 2019].

[24] MCDC Data Applications. Geocorr 2000: Geographic Correspondence Engine [Online]. http://mcdc.missouri.edu/applications/geocorr2000.html. [Accessed on 11 Feb 2019].

[25] CENTER FOR MEDICARE & MEDICAID SERVICES. Medicare Fee-For-Service Provider Utilization & Payment Data Physician and Other Supplier Public Use File: A Methodological Overview [Online]. http : / / www . cms . gov / Research-Statistics-Data-and-Systems / Statistics-Trends-and-Reports / Medicare-Provider-Charge-Data / Downloads/Medicare-Physician-and-Other-Supplier-PUF-Methodology.pdf. [Accessed on 11 Feb 2019].

[26] Aki-Hiro Sato, Shoki Nishimura, and Hiroe Tsubaki. World grid square codes: Definition and an example of world grid square data. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 4238–4247, 2017.

[27] CONGRESS OF THE UNITED STATES CONGRESSIONAL BUDGET OFFICE. Geographic Variation in Health Care Spending [Online]. http://www.cbo.gov/sites/default/files/cbofiles/ftpdocs/89xx/doc8972/02-15-geoghealth.pdf. [Accessed on 11 Feb 2019].

[28] CENTERS FOR MEDICARE & MEDICAID SERVICES. CROSSWALK MEDICARE PROVIDER/SUPPLIER to HEALTHCARE PROVIDER TAXONOMY [Online]. https : / / data . cms . gov / Medicare-Enrollment / CROSSWALK-MEDICARE-PROVIDER-SUPPLIER-to-HEALTHCARE / j75i-rw8y. [Accessed 11 Feb 2019].

[29] Centers for Medicare & Medicaid Services. Details for title: 2018 [Online]. https : / / www . cms . gov / medicare / coding / hcpcsreleasecodesets / alpha-numeric-hcpcs-items / 2018-alpha-numeric-hcpcs-file- . html, [Accessed on 11 Feb 2019].

[30] Evgenia Dimitriadou, Kurt Hornik, Friedrich Leisch, David Meyer, and Andreas Weingessel. Misc functions of the department of statistics (e1071), tu wien. *R package*, 1:5–24, 2008.

[31] Michael Hahsler, Matthew Piekenbrock, Sunil Arya and David Mount. 2018. Package 'dbscan'. https://CRAN.R-project.org/package=dbscan. [Accessed 11 Feb 2019].

[32] BUFFALO LAW JOURNAL. Cowie, physician and opioid addict, sentenced to 24 months in prison [Online]. https://www.bizjournals.com/buffalo/news/2017/02/02/

cowie-physician-and-opioid-addict-sentenced-to-24.html. [Accessed 11 Feb 2019].

[33] Statistics Bureau, Ministry of Internal Affairs and Communications. Grid Square Statistics[Online]. https://www.stat.go.jp/english/data/mesh/index.html. [Accessed 11 Feb 2019].

[34] Andrea D. Pozzolo, Olivier Caelen, Reid A. Johnson, and Gianluca Bontempi. Calibrating probability with undersampling for unbalanced classification. In *2015 IEEE Symposium Series on Computational Intelligence*, pages 159–166, 2015.

A Appendix

A.1 World Grid Square Code

World grid square code [26] is an extension of the Japanese Industrial Standard (JIS) for gird square codes (JIS X0410 [33]) established in 1976. World grid square code is defined from 1-st grid square code to 6 -th grid square code. However we only used 1- st grid square code. Here, we give the definition of 1-st world grid square code. See [26] for more details. Firstly, we define three binary variable x, y, z to separate the earth into eight areas based on latitude and longitude and construct 0-th level gird square code. The three variables are given as follows:

$$x = \begin{cases} 0 & \text{(if latitude is positive)} \\ 1 & \text{(otherwise)} \end{cases}, \tag{20}$$

$$y = \begin{cases} 0 & \text{(if longitude is positive)} \\ 1 & \text{(otherwise)} \end{cases}, \tag{21}$$

$$z = \begin{cases} 0 & \text{(if } \mid \text{longitude } \mid < 100^\circ) \\ 1 & \text{(otherwise)} \end{cases}. \tag{22}$$

We define the 0-th level grid square code o as

$$o = 2^2 x + 2y + z + 1. \tag{23}$$

The 1- st grid square code can be calculated by

$$\text{1- st grid square code} = \begin{cases} o00p0u & (p < 10, u < 10) \\ o0p0u & (10 \leq p < 100, u < 10) \\ op0u & (100 \leq p, u < 10) \\ o00pu & (p < 10, 10 \leq u) \\ o0pu & (10 \leq p < 100, 10 \leq u) \\ opu & (100 \leq p, 10 \leq u) \end{cases}, \tag{24}$$

where p and u are defined as

$$p := \lfloor (1 - 2x)\text{latitude} \times 60 \div 40 \rfloor, \tag{25}$$

$$u := \lfloor (1 - 2y)\text{longitude} - 100z \rfloor \qquad (u \text{ is one or two digits}). \tag{26}$$

For example, latitude is 35.010348 and longitude is 135.768738, then 1 -st grid square code is 105235.

A.2 Outlier detection to extract credit card fraud

In this section, we apply outlier detection techniques to another domain, Credit card fraud detection [34]. This dataset is composed of credit card transactions from September 2013. It contains a subset of online transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. It consists of 31 columns, where Time column denotes the second elapsed between each transaction and the first transaction in the dataset, Amount column denotes the transaction, Class column takes 1 in case of fraud and 0 otherwise and the meanings of remaining columns are not revealed for confidentiality reason and the features have been transformed by means of principal components. Excluding Time and Class columns, the feature vectors are made. Listing 1 shows the program of outlier detections and their evaluations. Fig. 12 shows the roc curve of the two models. Tab. 16 shows the AUC score of the two models. Contrary to the fraud detection of medicare transaction, One-SVM performs well and LOF performs worse and is below random sampling.

Table 16: AUC score of the two outlier detection models for credit transaction fraud detection.

Outlier Detection Model	AUC
LOF	0.506
One-SVM	0.942

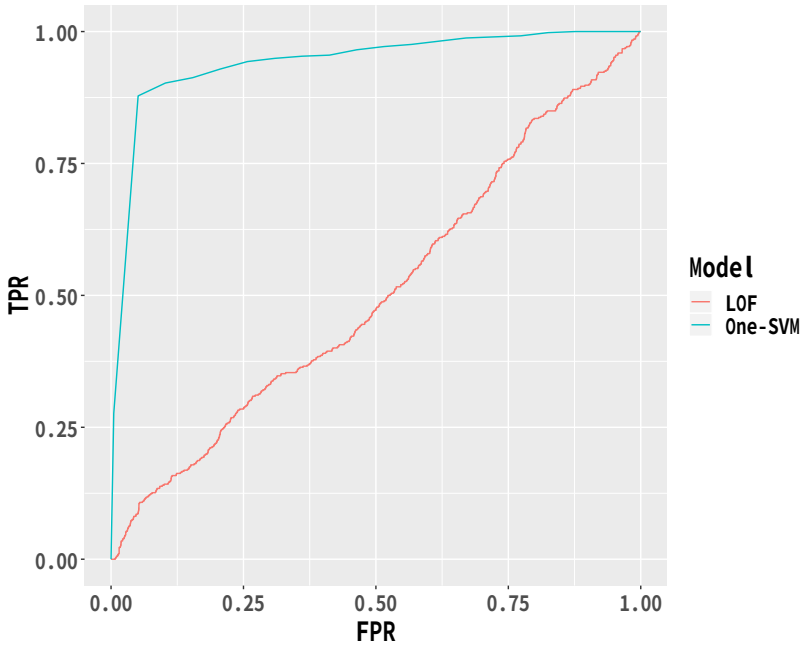

Figure 12: ROC curves of the LOF and One-SVM. The Model corresponds to the outlier detection methods. One-SVM outperforms LOF.

Listing 1: The code for outlier detection.

```
library(tidyverse)
library(e1071)
library(magrittr)
library(reshape2)
library(DMwR)


# define functions
    ------------------------------------------------------------

#scale function
scale_this <- function(x){
return((x-mean(x, na.rm=TRUE))/sd(x, na.rm=TRUE))
}

#calc area under curve
auc <- function(fpr, tpr){
  return((fpr-lag(fpr,default=0))*(tpr+lag(tpr,default=0))*0.5)
}

#one support vector machine
one_svm <- function(mat, nu=0.1){
  svm_model <- svm(mat, y=NULL, nu=nu, type="one-classification")
  outlier_df <- tibble(predict = factor(if_else(predict(svm_model)==TRUE
      , 0, 1)))
  return(outlier_df)
}

```

```
#calc confusion matrix from datafrme which have class column and predict
    column
conf_mat <- function(df){
  df <- df %>%
    complete(class, predict, fill=list(n=0))
  row <- df$class %>% as.character %>% as.numeric
  col <- df$predict %>% as.character %>% as.numeric

  conf_mat <- matrix(0, nrow=2, ncol=2)
  rownames(conf_mat) <- c("false", "true")
  colnames(conf_mat) <- c("false", "true")
  for(i in row){
    for(j in col){
      conf_mat[i+1,j+1] <- df %>% filter(class==i, predict==j) %>% .$n
    }
  }
  return(conf_mat)
}

#false positive rate from confusion matrix
fpr <- function(conf_df){
  return(conf_df["false","true"]/(conf_df["false","false"]+conf_df["
      false","true"]))
}

#true positive rate from confusion matrix
tpr <- function(conf_df){
  return(conf_df["true","true"]/(conf_df["true","true"]+conf_df["true","
      false"]))
}

set.seed(100)
credit_row <- read_csv("data/credit/creditcard.csv")

N <- nrow(credit_row)
MIN_PTS <- 20
nu_range <- seq(0.001, 0.98, length=20)
#N <- 100000

#credit_fraud <- credit_row %>% filter(Class==1)
#credit_normal <- credit_row %>% filter(Class==0) %>% sample_n(N-nrow(
    credit_fraud))
#credit <- bind_rows(list(credit_fraud, credit_normal))
credit <- credit_row

x <- credit %>%
  select(-Time, -Class) %>%
  mutate_if(is.double, scale_this) %>%
  as.matrix()

y <- credit %>%
  select(Class) %>%
  as.matrix() %>%
  .[,1]


# One SVM evaluation
    ------------------------------------------------------

predict_onesvm_df <- tibble(nu = nu_range, x=rep(list(x), length(nu_
    range)), class=rep(list(tibble(class=y)), length(nu_range))) %>%
  dplyr::mutate(predict = map2(x, nu, function(data_x, hyper_nu){
```

```
    print(hyper_nu)
    one_svm(mat=data_x, nu=hyper_nu)}))

predict_onesvm_roc <- predict_onesvm_df %>%
  mutate(df = purrr::map2(class, predict, function(x,y){bind_cols(list(x
      ,y))})) %>%
  mutate(conf_df = purrr::map(df, . %>% count(class,predict))) %>%
  mutate(fpr = purrr::map(conf_df, function(x){fpr(conf_mat(x))})) %>%
  mutate(tpr = purrr::map(conf_df, function(x){tpr(conf_mat(x))})) %>%
  select(nu, fpr, tpr) %>%
  unnest() %>%
  bind_rows(list(., tibble(fpr=c(0,1), tpr=c(0,1)))) %>%
  arrange(tpr) %>%
  mutate(s = auc(fpr=fpr, tpr=tpr))
onesvm_auc <- sum(predict_onesvm_roc$s)


# LOF detection
    ------------------------------------------------------------

predict_lof_df <- lof(x, k=MIN_PTS)

predict_lof_roc <- bind_cols(list(credit, tibble(lof=predict_lof_df))) 
    %>%
  dplyr::arrange(desc(lof)) %>%
  mutate(class=Class) %>%
  dplyr::mutate(not_class = if_else(class == 0, 1, 0)) %>%
  dplyr::mutate(tpr = cumsum(class)/sum(class)) %>%
  dplyr::mutate(fpr = cumsum(not_class)/sum(not_class)) %>%
  mutate(s = auc(fpr=fpr, tpr=tpr))

lof_auc <- sum(predict_lof_roc$s)

predict_onesvm_roc <- read_rds("data/predict_onesvm_roc.rds")
predict_lof_roc <- read_rds("data/predict_lof_roc.rds")


tibble(Model=c("One-SVM","LOF"), list(predict_onesvm_roc, predict_lof_
    roc)) %>%
  unnest() %>%
  ggplot(aes(fpr, tpr, color=Model))+
  geom_line()+
  theme(text=element_text(size=20))+
  labs(x="FPR", y="TPR")

#write_rds(x = predict_lof_roc, "data/predict_lof_roc.rds")
#write_rds(x = predict_onesvm_roc, "data/predict_onesvm_roc.rds")
print(lof_auc)
print(onesvm_auc)
```

A.3 List of exclusions

Here, we show the all exclusion reasons. Tab. 17 shows the exclusion codes and their descriptions.

Table 17: Exclusion code and its reason.

Social Security Act	Amendment
1128	Scope of exclusions imposed by OIG expanded from Medicare and State health care programs to all Federal health care programs, as defined in section 1128B(f)(1).
1128(a)(1)	Conviction of program-related crimes. Minimum Period: 5 years
1128(a)(2)	Conviction relating to patient abuse or neglect. Minimum Period: 5 years
1128(a)(3)	Felony conviction relating to health care fraud. Minimum Period: 5 years
1128(a)(4)	Felony conviction relating to controlled substance. Minimum Period: 5 years
1128(c)(3)(G)(i)	Conviction of second mandatory exclusion offense. Minimum Period: 10 years
1128(c)(3)(G)(ii)	Conviction of third or more mandatory exclusion offenses. Permanent Exclusion
1128(b)(1)(A)	Misdemeanor conviction relating to health care fraud. Baseline Period: 3 years
1128(b)(1)(B)	Conviction relating to fraud in non-health care programs. Baseline Period: 3 years
1128(b)(2)	Conviction relating to obstruction of an investigation or audit. Baseline Period: 3 years
1128(b)(3)	Misdemeanor conviction relating to controlled substance. Baseline Period: 3 years
1128(b)(4)	License revocation, suspension, or surrender. Minimum Period: Period imposed by the state licensing authority.
1128(b)(5)	Exclusion or suspension under federal or state health care program. Minimum Period: No less than the period imposed by federal or state health care program.
1128(b)(6)	Claims for excessive charges, unnecessary services or services which fail to meet professionally recognized standards of health care, or failure of an HMO to furnish medically necessary services. Minimum Period: 1 year
1128(b)(7)	Fraud, kickbacks, and other prohibited activities. Minimum Period: None

Table 17: *(continued)*

Social Security Act	Amendment
1128(b)(8)	Entities controlled by a sanctioned individual. Minimum Period: Same as length of individual's exclusion.
1128(b)(8)(A)	Entities controlled by a family or household member of an excluded individual and where there has been a transfer of ownership/control. Minimum Period: Same as length of individual's exclusion.
1128(b)(9), (10), and (11)	Failure to disclose required information, supply requested information on subcontractors and suppliers; or supply payment information. Minimum Period: None
1128(b)(12)	Failure to grant immediate access. Minimum Period: None
1128(b)(13)	Failure to take corrective action. Minimum Period: None
1128(b)(14)	Default on health education loan or scholarship obligations. Minimum Period: Until default or obligation has been resolved.
1128(b)(15)	Individuals controlling a sanctioned entity. Minimum Period: Same as length of entity's exclusion.
1128(b)(16)	Making false statement or misrepresentations of material fact. Minimum period: None.
1156	Failure to meet statutory obligations of practitioners and providers to provide medically necessary services meeting professionally recognized standards of health care (Quality Improvement Organization (QIO) findings). Minimum Period: 1 year

Master's Thesis

Outlier detection technique to extract candidates of fraud from medical insurance claims

Guidance

Associate Professor Aki-Hiro Sato

Ryosuke Kawamori

Department of Applied Mathematics and Physics

Graduate School of Informatics

Kyoto University

February 2019

Outlier detection technique to extract candidates of fraud from medical insurance claims

Ryosuke Kawamori

Abstract

This thesis proposes how to find fraudulent activities from data of medical insurance claims in an unsupervised way. Firstly, we split data into homogenous groups by the providers' department, a number of specialties, and their geographical location. If this is not accomplished, we may find providers performing rare operations, or providers in rural areas because of its variation in submitted charges. Next, we apply the outlier detection method, One-SVM and LOF to these groups. Finally, we evaluate the effectiveness by using labels of providers who are excluded because they committed fraud. The proposed method is above random and may be used as a first line of further investigation by specialties.