Master's Thesis

Construction of Mathematical Models and Development of Efficient Algorithms

Guidance

Professor Taro JOHO
Assistant Professor Jiro KOGAKU

Saburo SURI

Department of Applied Mathematics and Physics

Graduate School of Informatics

Kyoto University

February 2019

Abstract

This article proposes how to find fraudulent activities from data of healthcare claim in an unsupervised way. Firstly, we split data into homogenous groups by providers' department, number of speciality, and their geographical location, otherwise we may find providers doing rare operations or providers in rural areas because of its lower submitted money than in urban areas. Next, we apply outlier detection method, One-SVM and LOF to that groups. We find variation due to department and geographical location. Finally, we evaluate the effectiveness by labels of providers who are excluded because of their having committed fraud. Proposed method is above random and may be used as a first line of further investigation.

Contents

1 Health care fraud and Existing works

1.1 Health care fraud

According to 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 provider 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 healthcare costs on behalf of their subscribers, including governmental health departments and private insurance companies [1].

NHCAA highlights below type of fraud.

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 type 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, data-cleaning method and statistical model which detect fraudulent activity heavily depend on who commit fraud and what type of fraud we want to detect.

1.2 Existing Works

To combat fraud, there are three categories of intervention : Prevent, Detect and Respond as shown 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 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 report system 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 frauds”. For example, changing and improving the system or law enforcement initiatives so that the chances of future frauds are reduced are in this area. Most of the existing works focus on the effectiveness of detection [11]. Our research are also categorized as detection. We try data cleansing method before detail analysis of the experts.

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 propriately and take corrective action when integrity breakdown arise.

Supervised method

Ortega and He used neural network [12,13] and Bonchi used Decision trees [14]. These methods need the target data, which is the record of activity that were detected as fraudulent by the domain experts. Richard [15] used random forest with respect to various undersampling ratio data of providers because medicare dataset is highly unbalanced, meaning that most of the providers are labeled normal and tiny number of providers are labeled fraud. Richard [16] combined three data sources of medicare procedure and applied logistic regression, gradient boosting and random forest.

Unsupervised method

Rasim [17] used demographic information and used the quantile as a fraudulent score. Richard [18] used Isolation Forest, Local Outlier Factor, Unsupervised Random Forest, Autoencoder and $k-$ Nearest Neighbor and evaluated their methods using label of provider who committed fraud provided by NHCAA. Van [3] used more story-based method. For example, they used the deviations from simple linear model between total dollar amount reimbursed and the number of reimbursed claims of a provider. They also used the percentage of dental claims claimed for specific tooth code. This idea is based on the report that some dentists claimed over and over for the same set of procedures by only changing patient IDs in order to reimburse as much as possible with the least effort involved.

Others

Pande [19] outline the characteristics of physicians who have been convicted of fraud and the consequence of their conviction and the proposed problems of policies for fraud.

2 Data

In order to detect fraud in the health care receipts submitted by the physicians registered in the US, we used three types of data such as Medicare Provider Utilization and Payment Data in 2014, Physician and Other Supplier (Physician and Other Supplier PUF) [20], Full Replacement Monthly NPI file [21], and All US zip Codes with their corresponding latitude and longitude coordinates processed by MABLE/Geocorr2K ver1.3.3 [22]. The first data : Physician and Other Supplier PUF records a 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 National Provider Identifier(NPI) Healthcare Common Procedure Coding System(HCPCS) code and place of service. Physician and Other Supplier PUF is made public from 2012 to 2016. For the sake of simplicity, we use only data in 2014. This data consists of 9,497,892 records and used to make physicians' feature vector. Each row records how much money each provider submitted for a operation in a year [23]. The second data : the full replacement monthly NPI file includes 5,400,369 records and provides specialties of the providers. Each row records what kind of specialties each provider has. For example, internal medicine are more categorized as specialities like addiction medicine and bariatric medicine. The third data : US zip Code have 43,723 records. Each row records longitude and latitude of US zip code. We use this data to calculate which grid square code [24] each provider belongs to. Tabs. 2, 3 and 4 shows the descriptions of the columns of the three data. They have more columns, but the columns which we do not focus on were omitted. Tabs. 5, 6 and 7 are the example of three raw datas. For example first row of Tab. 5 indicates that provider whose npi code is 101111 submitted 10 dollars on average for the service of which hcpcs code is 91312. The first row of Tab. 6 indicates that provider whose npi codes is 101111 has only one speciality of which code is 207X00000X. The first row of Tab. 7 indicates that the longitude and latitude of zipcode : 01008 are -72.402004 and 42.277543.

To evaluate our methodology we use the List of Excluded Individuals and Entities (LEIE). The LEIE contains information of the individuals and the entities that are currently excluded from participation in Medicare, Medicaid and all other Federal health care programs in 1977-2018 [7]. This data includes 70,491 records. Each row records excluded physicians, excluded date and reason of their exclusion. Because providers who commit fraud are considered to be excluded from registration after their illegal activity, we use providers' label of which their exclusion date are after Physician and Other Supplier PUF records : after 2014. Tab. 8 shows the example of the LEIE data. For example, the first row of Tab. 8 indicates that the provider of which npi code is 101111 was excluded from participation in medicare on 19/04/2018 and the reason of exclusion is code:1128b7. The reasons of exclusion have 22 types. 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 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 [23].

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.
line_srvc_cnt	Number of services provided; note that the metrics used to count the number provided can vary from service to service.
average_medicare_submitted_amount	Average of the Medicare allowed amount for the service.
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.

Table 3: Part of column description 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 speciality of the provider indicated by the npi code.
Healthcare Provider Taxonomy Code_2	Second speciality of provider indicated by the npi code.

Table 4: Part of column description of all US zipcode with their corresponding latitude and longitude coordinates.

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

Table 5: Examples of Medicare Provider Utilization and Payment Data. The raw data has 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 has 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 zipcode to its logitude and latitude file. The raw data has 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 he List of Excluded Individuals and Entities. The raw data has 61,168 rows and 5 columns.

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

2.1 Data Combination

We combined the three types of datasets such as Medicare Provider Utilization and Payment Data, Full Replacement Monthly NPI and zipcode as shown in Fig. 1 We split data by providers ’ department, their number of specialties and their location. We have two reasons to use this data-split method. Firstly, handling all providers’ feature vectors simultaneously is very time consuming, because though LOF has $O(n \log n)$ time complexity on the low dimensional data, One-SVM algorithm has $O(n^3)$ computational complexity, where n is a number of the data points. Even though we split the data, the complexity, decreases in a order of devision of constant, However, the number of provider n is limited, the reduce of the time is effective.

Secondly, variance of providers' feature vector comes from providers' department, specialities, geographical location and so on. With regard to geographical variance, we can easily assume that the providers in the urban area claim much because they have many patients and it is also pointed out that spending per medicare beneficiary also have high variance [25]. Thus, if we try to find fraudulent claim from all providers, we may find providers who have rare speciality or who are in the rural areas. These are the reasons to split the data. To split the data by their location, we calculated first grid square code [24] from their geographical locations (latitude and longitude) by converting their zipcode.

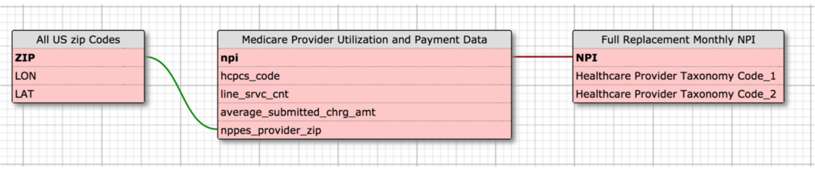

All US zip Codes
ZIP
LON
LAT
Medicare Provider Utilization and Payment Data
npi
hcpcs_code
line_srvc_cnt
average_submitted_chrg_amt
nppes_provider_zip
Full Replacement Monthly NPI
NPI
Healthcare Provider Taxonomy Code_1
Healthcare Provider Taxonomy Code_2

Figure 1: Conceptual illustration of data frame combination.

2.2 Data understanding

Variation by provider speciality

Fig. 2 shows part of histograms of total submitted amount of money by the provider type. We can easily see that different types of provider have very different histogram. For example, Peripheral Vascular Disease have high amount. The mean amount before log transformation is 719014 USD. On the other hand, Certified Nurse Midwife has low amount. The mean amount before log transformation is 9503 USD.

Geographical variance

We computed mean of submitted amount of money in each grid square. Mean of observed value x_{ij}, where i is a index of observed value and j is a data type, in grid k th is defined as follows

$$m_{jk} = \frac{1}{|D_k|} \sum_{i \in D_k} x_{ij}, \tag{1}$$

where D_k is a set of index i included in grid square k and $| \cdot |$ denotes the number of elements included in the set $\cdot$. Fig. 3 shows mean grid square of the total submitted amount of money on each grid k and Fig. 4 shows histogram of mean grid square of the total submitted amount of money. These figures indicate that there is strong geographical dependences with regard to total submitted amount of money.

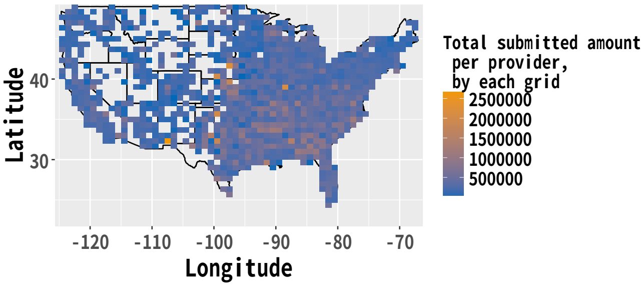


Figure 3: Mean of total submitted amount of money in each grid. The red color corresponds to the high submitted amount of money and blue color corresponds to the low submitted amount of money. We can see total submitted amount of money is high in urban areas and low in rural areas.

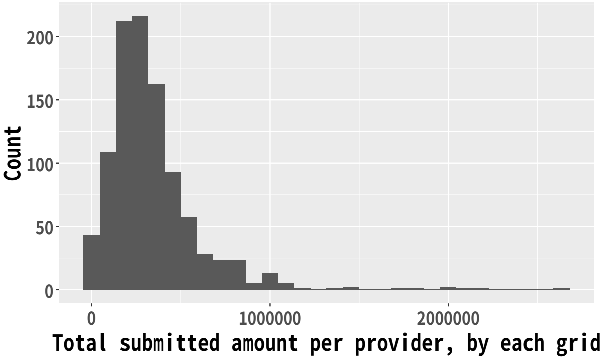


Figure 4: Histograms of total submitted amount of money. x axis shows the mean of total submitted amount in each grid. We can see that the histogram has fat tail and the submitted amount of money is dependent on the area in the US.

Variation by providers ’ number of speciality

Fig. 5 shows histogram of total submitted amount of money by providers ’ number of specialities. We can see that variation by provider ’ s number is as high as by geographical location and type of providers. So we may not need to split data by providers ’ number of speciality.

3 Method

3.1 Subset of all data for evaluation

We focus on physicians in New York, California, and Florida because fraud detection enforcement is strong in these states and we can have fraud label data to evaluate efficiency of our methods. Our main purpose is to cleanse data before investigations by specialists, so in practical application we should apply our data cleansing methods to the data of the state where fraud enforcement is weak too. However, we need metric that evaluate our unsupervised strategy.

3.2 Feature engineering

Feature engineering of existing works

Richard [18]used the mean, median, max, min, sum and standard deviation of line_srvc_cnt, average_medicare_submitted_amt, average_medicare_payment_amt, bene_uniuqe_cnt and used one-hot encoding of providers'gender and specialities. We call this npi-level feature engineering strategy Stat(statistics of all procedures.).

Feature consisting of 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. 10 shows these seven levels of focus. However, because of data limitation, we can look data only above or equal level 3. Above level 3, for example, we pay attention to medical group whose providers belong to 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 above level3.

We create feature vector for each provider of which the j th row is log transformed and normalized total submitted amount of money on each services,

$$\begin{aligned} x'_j &= \\ &\ln\big((\text{average\_medicare\_submitted\_amountof the } j \text{ th hcpcs\_code})\times \\ &(\text{line\_srvc\_cnt of the } j \text{ th hcpcs\_code})\big), \qquad (2) \\ x_j &= \frac{x'_j - m_{jk}}{\sigma_k}, \qquad (3) \end{aligned}$$

where m_k is a mean of x'_j in k th grid and σ_k is a standard deviation of x'_j in k th grid. Tab. 11 shows the providers ' feature based data calculated from Tab. 5. We call this npi-level feature engineering strategy Hcpcs (Each amount of hcpcs procedure).

Table 10: Levels of healthcare fraud control proposed by [1]. This table is 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.

Table 11: Examples of feature-based data before log transforming and normalizing.

NPI	99312	99313	···	99314
10111111	30	40	···	50
10111112	0	0	···	100

Data-split method

As noted in previous section, submitted amount of money have variance with regard to physicians‘ geographical location, speciality and number of specialities. We focus on the physicians whose speciality is only one and we split data by their location and speciality and make physicians' feature vector from this split data frame. So, if medical procedure in specific area is limited, the length of feature vector of physicians' in that area became short. For example, feature vector of the internal medicine physician in meshcode : 306173 has 907 but, the feature vector of same physicians in meshcode : 304182 has 420. We applied outlier detection method on each split data and combine the results to one data frame to calculate evaluation metrics. To check the efficiency of data-split method based on geographical location, we compare two versions. Provider : split only by physicians' speciality and Provider-mesh : split by physicians' speciality and their location.

3.3 One Support Vector Machine

To detect fraudulent provider activity, we use one-class support vector machine [5] as an outlier based method. One support vector machine separate data from 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 $\mathbf{\Phi}(\boldsymbol{x}) : X \longrightarrow F$,where X is a input space and F is a feature space that has inner product, and applying support vector machine in this space enable us to solve a nonlinear problem. One support vector machine is expressed as below,

$$\begin{aligned} &\min_{\boldsymbol{w}\in F,\boldsymbol{\xi}\in\mathbb{R}^l,\rho\in\mathbb{R}} \frac{1}{2}||\boldsymbol{w}||^2 + \frac{1}{\nu l}\sum_i \xi_i - \rho \\ &\text{subject to} \qquad (\boldsymbol{w}, \mathbf{\Phi}(\boldsymbol{x}_i)) \geq \rho - \xi_i, \xi_i \geq 0. \end{aligned} \tag{4}$$

Here, $||\cdot||^2$ denotes l^2-norm of a vector, $(\cdot,\cdot)$ denotes dot product of two vectors in inner product space F and ν is a hyperparameter. The decision function is

$$f(\boldsymbol{x}) = \mathrm{sgn}((\boldsymbol{w}, \mathbf{\Phi}(\boldsymbol{x})) - \rho), \tag{5}$$

where

$$\mathrm{sgn}(x) = \begin{cases} +1 & (x > 0) \\ 0 & (x = 0) \\ -1 & (x < 0) \end{cases} . \tag{6}$$

The Lagrange function is given by follows,

$$\begin{aligned} L(\boldsymbol{w}, \boldsymbol{\xi}, \boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = &\frac{1}{2}||\boldsymbol{w}||^2 + \frac{1}{\nu l}\sum_i \xi_i - \rho \\ &- \sum_i \alpha_i((\boldsymbol{w}, \mathbf{\Phi}(\boldsymbol{x}_i)) - \rho + \xi_i) - \sum_i \beta_i \xi_i. \end{aligned} \tag{7}$$

By taking derivative of variables, we get,

$$\boldsymbol{w} - \sum_i \alpha_i \mathbf{\Phi}(\boldsymbol{x}_i) = 0, \tag{8}$$

$$\frac{1}{\nu l} - \alpha_i - \beta_i = 0, \tag{9}$$

$$-1 + \sum_i \alpha_i = 0. \tag{10}$$

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

$$\begin{aligned} &\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \min_{\boldsymbol{w},\boldsymbol{\xi},\rho} L(\boldsymbol{w}, \boldsymbol{\xi}, \rho, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \min_{\boldsymbol{\alpha}} \sum_{ij} \alpha_i \alpha_j (\mathbf{\Phi}(\boldsymbol{x}_i), \mathbf{\Phi}(\boldsymbol{x}_j)) \\ &\text{subject to} \qquad 0 \leq \alpha_i \leq \frac{1}{\nu}, \quad \sum_i \alpha_i = 1. \end{aligned} \tag{11}$$

And by substituting (8) to (5), we get

$$f(\boldsymbol{x}) = \mathrm{sgn}(\sum_i \alpha_i (\mathbf{\Phi}(\boldsymbol{x}_i), \mathbf{\Phi}(\boldsymbol{x})) - \rho). \tag{12}$$

Here we use a gaussian kernel defined as

$$k(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{\Phi}(\boldsymbol{x}), \boldsymbol{\Phi}(\boldsymbol{y})) = e^{-\frac{||\boldsymbol{x} - \boldsymbol{y}||^2}{c}}. \tag{13}$$

As we can see from dual problem, we do not need to calculate $\boldsymbol{\Phi}(\boldsymbol{x})$ directly and only needs value of kernel function $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ when solving this problem. This technique is called kernel method. We solve the dual problem expressed by (11) using R package e1071 [26]. We set hyperparameter $c = \frac{1}{l}$ and $\nu = 0.1$(package default). One support vector machine have two favorable properties. First, ν gives an upper bound of ratio of outliers, that is, training points outside the estimated region. Second, there are probabilistic guarantee that new points fall inside the estimated region. For more details and proofs, see [5].

3.4 Local Outlier Factor

LOF calculates the degree of how the object is isolated from its surroundings. The $k-$distance of an object p is defined as

$$k - distance = d(p, o), \tag{14}$$

where $d(p, o) = \sqrt{\sum_i (p_i - o_i)^2}$ is Euclidean distance 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

$$N_k(p) = \{q \in D \setminus p \mid d(p, q) \leq k - distance(p)\}. \tag{15}$$

Reachability distance of an object p with respect to object r is defined as

$$reach - dist_k(p, r) = \max\{k - distance(p), d(p, r)\}. \tag{16}$$

Local reachability density of an object p is defined as

$$lrd_{MinPts}(p) = \left\{ \frac{\sum_{q \in N_{MinPts}(p)} reach - dist_{MinPts}(p, q)}{\mid N_{MinPts}(p) \mid} \right\}^{-1}. \tag{17}$$

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

$$LOF_{MinPts}(p) = \frac{\sum_{q \in N_{MinPts}(p)} \frac{lrd_{MinPts}(q)}{lrd_{MinPts}(p)}}{\mid N_{MinPts}(p) \mid} \tag{18}$$

We set hyperparameter $MinPts = 20$.

3.5 Evaluation Metric

Our methods were evaluated using true positive rate, false negative rate, Receiver Operating Characteristic (ROC) curve, and Area Under Curve (AUC). True positive rate(TPR) and false negative rate (FNR) is defined as

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

$$FNR = \frac{FN}{FN + TP}, \tag{20}$$

where TP is a number of physicians that are truely classified as positive and FN is a number of physicians that are falsely classified as negative. ROC curve created by plotting TPR against the FPR and AUC is the are of ROC curve. If physicians are randomly selected, ROC curve is a straight line through two points $(0, 0)$ and $(1, 1)$ and AUC is 0.5. If we can correctly classify physicians, ROC curve is above that line and AUS is above 0.5. When applying One-SVM, we cannot get outlier score and cannot draw ROC curve in standard way. We plot TPR against the FPR with regard to hyperparameter values : ν.

4 Results

Evaluation

One-SVM

Fig. 6 shows the ROC curve of the data split method Provider and Provider-mesh of Stat and Hcpcs. All ROC curves are slightly above the straight line through points $(0, 0)$ and $(1, 1)$. However, they didn't change so much with respect to the data-split and the feature.

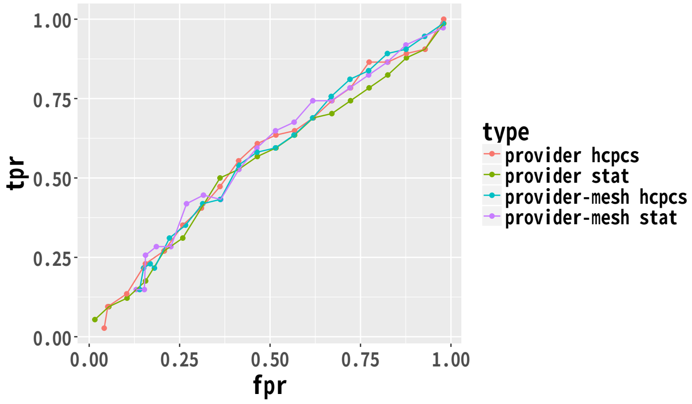


Figure 6: ROC curve of One-SVM with respect to the data-split : Provider and Provider-mesh of the feature : Stat and Hcpcs. Type shows their data-split method and how their feature are created. X-axis shows the true positive rate and Y-axis shows the false positive rate. All ROC curves are slightly above the straight line through points $(0, 0)$ and $(1, 1)$. However, they didn't change so much with respect to the data-split and the feature.

Local Outlier Factor

Fig. 7 shows the ROC curve of the data split method Provider and Provider-mesh of Stat and Hcpcs. All ROC curve is above the random. All ROC curves are slightly above the straight line through points $(0, 0)$ and $(1, 1)$. Especially data-split : Provider of feature : Hcpcs is best of all.

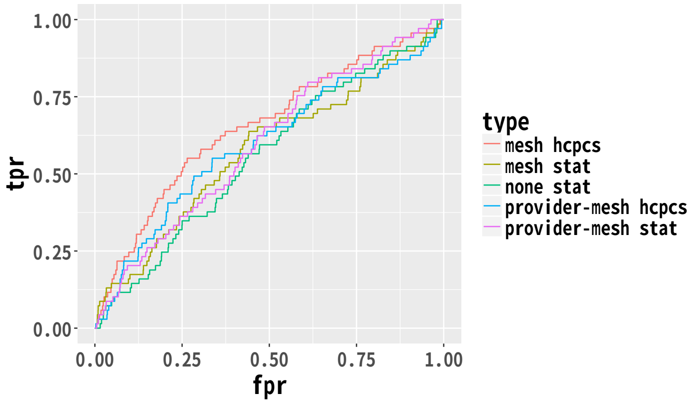


Figure 7: ROC curve of LOF with respect to the data-split : Provider and Provider-mesh of the feature : Stat and Hcpcs. Type shows their data-split method and how their feature are created. X-axis shows the true positive rate and Y-axis shows the false positive rate. None means there were no data-split. All ROC curves are slightly above the straight line through points $(0, 0)$ and $(1, 1)$. Especially data-split : Provider of feature : Hcpcs is best of all.

AUC Score

Tab. 12 shows the AUC of outlier detection models of each data-split and feature. The best AUC was achieved with LOF of data-split : Provider and feature : Hcpcs. Though because of time of calculation, we cannot calculate the data-split : None of SVM, Of all data-split and feature, AUC of LOF are above One-SVM,. The AUC of the One-SVM changed slightly with respect to data-split and features, but that of LOF changed a lot. By comparing the data-split : Provider and None of LOF, data-split : Provider is effective, though data-split : Provider-Mesh is not effective or even worse when using LOF.

There are two reason of false negative. Firstly, provider who commit fraud steal a little and their fraudulent activity was not apparant in the data but they were excluded due to the insider report. Secondly, their fraudulent submission was before 2015 though we used the submission data after 2014. First case cannot be captured easily by the data-mining method, but, second case can be captured by the combination of available datas.

Table 12: AUC of outlier detection models of each data-split and feature.

Outlier Detection Model	Data-Split	Feature	AUC
LOF	None	HCPCS-Stat	0.558
LOF	Provider-Mesh	HCPCS-Amount	0.594
LOF	Provider-Mesh	Provider-Stat	0.597
LOF	Provider	HCPCS-Amount	0.654*
LOF	Provider	Provider-Stat	0.577
One-SVM	Provider-Mesh	HCPCS-Amount	0.541
One-SVM	Provider-Mesh	Provider-Stat	0.550
One-SVM	Provider	HCPCS-Amount	0.548
One-SVM	Provider	Provider-Stat	0.527

Ratio of outlier

Next, we changed hyperparameter ν of One-SVM at 20 points from 0.001 to 0.98. Fig. 8 shows the ratio of outlier on various ν. As we expected, we can easily see as ν increases, ratio of outlier also increases. Thus, we can regulate ratio of outlier by changing ν.

NPI	Provider Type
10111	Internal Medicine
10112	Certified Nurse Midwife
10113	Addiction Medicine
⋮	⋮
11001	Internal Medicine

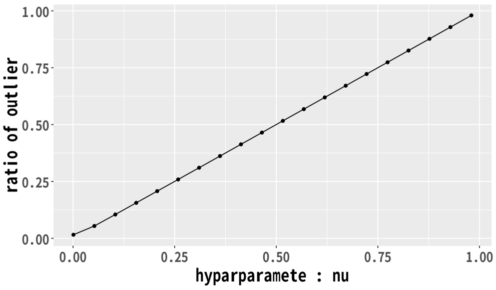


Figure 8: Ratio of outlier in terms of ν. x axis shows the hyperparameter ν which is a upper bound of ratio of outlier and y axis shows the ratio of outlier detected by One-SVM of which $\nu = x$.

5 Conclusions

We applied data linkage methods regarding about NPI, Healthcare Provider Taxonomy Core, and zipcode. We used world grid square codes to segment regions regarding physicians' geospatial heterogeneity and investigates world grid square statistics about the submitted amount of money and the provider type variance. We applied One-SVM and LOF for feature vectors, Hcpcs : submitted amount of money of each hcpcs code and Stat : summary statistics of submitted amount of money and one-hot encoding of physicians' characteristics. By dividing large records based on physicians' location using world grid square codes and speciality into many segments, we get many small datasets about the healthcare receipts. Finally, evaluated the method using LEIE labels. LOF performs better than One-SVM. LOF with date-split : Provider and feature : Hcpcs is the best of all methods. Data-split is also beneficial in terms of time of calculation. As future work, Firstly, as we see in the design of feature vectors and its performance, the feature engineering is crucial when we define the fraud claims in the health care receipts. We may need more deep knowledge of healthcare domain to build feature. Secondly, the trade off between true positive rate and false positive rate should be adjusted. If we allow only low true positive rate. there are more rooms for benefit by increasing the cost of experts that find the providers actually committing fraud from fraudulent cases to prevent giving them money. On the other hand, if we allow high true positive rate, the cost of experts increases and its costs is higher than the money given to the providers committing fraud. Those two work seems to need the expert corporation

NPI	Provider Type
10111	Internal Medicine
10115	Internal Medicine
⋮	⋮
11001	Internal Medicine

NPI	Provider Type
10113	Addiction Medicine
10118	Addiction Medicine
⋮	⋮
10883	Addiction Medicine

Acknowledgments

The author would like to express his sincere gratitude to Professor Taro Joho and Assistant Professor Jiro Kogaku for their helpful advices.

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 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.

NPI	Provider Type
10112	Certified Nurse Midwife
10120	Certified Nurse Midwife
⋮	⋮
10901	Certified Nurse Midwife

[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, May 2000.

[7] `https://www.nhpco.org/sites/default/files/public/regulatory/LEIE_Exsclusion_List.pdf` Accessed on 2018 Nov 25.

[8] `https://www.nhcaa.org/resources/health-care-anti-fraud-resources/consumer-info-action.aspx` Accessed on 2018 Dec 19.

[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. Available : `https://assets.kpmg.com/content/dam/kpmg/pdf/2014/05/fraud-risk-management-strategy-prevention-detection-response-O-201405.pdf` Accessed 2018 Nov 25.

[11] Rashidian A, Joudaki H, and Vian T. 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] R. Bauder and T. 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, July 2018.

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

[17] Rasim Muzaffer 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. IEEE, 2018.

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

[20] https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier.html.

[21] http://download.cms.gov/nppes/NPI_Files.html.

[22] http://mcdc.missouri.edu/applications/geocorr2000.html.

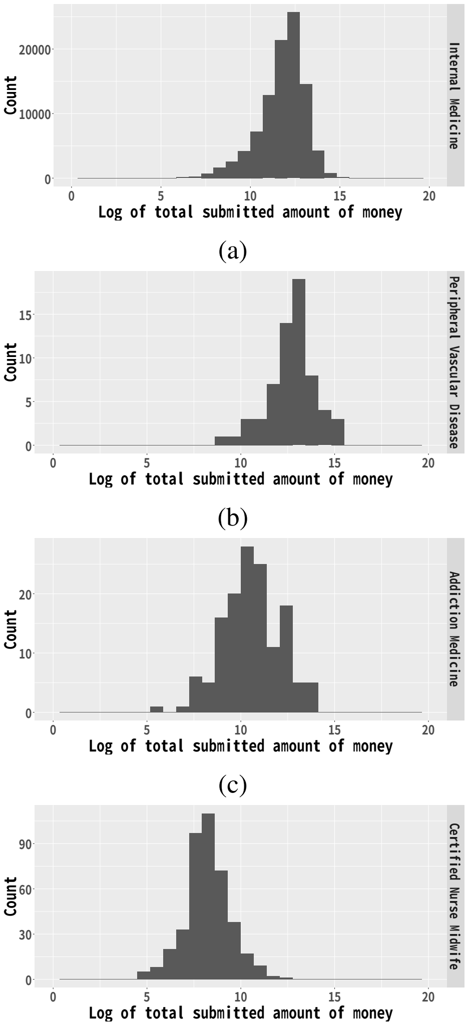
[23] http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Downloads/Medicare-Physician-and-Other-Supplier-PUF-Methodology.pdf.

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

[25] http://www.cbo.gov/sites/default/files/cbofiles/ftpdocs/89xx/doc8972/02-15-geoghealth.pdf.

[26] 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.

A Appendix



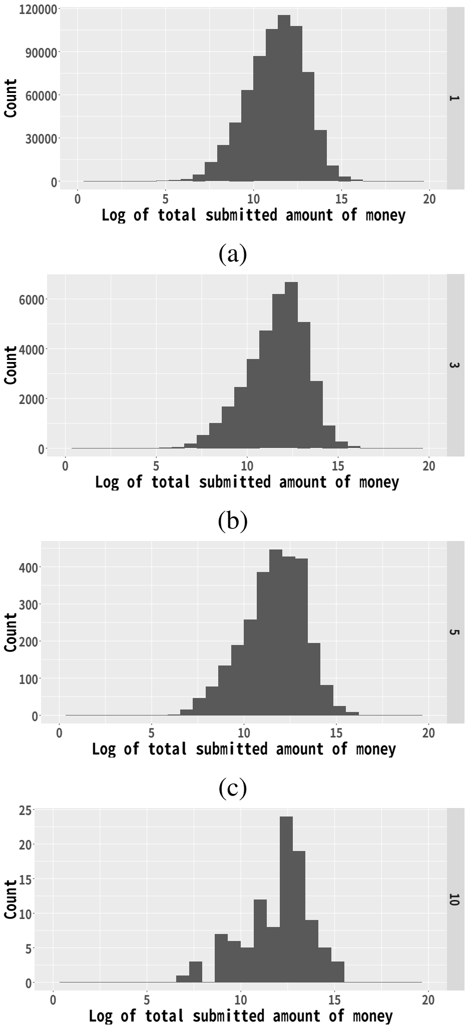
(a)

(b)

(c)

(d)

Figure 2: (a) depicts the histograms of log of 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.



(a)

(b)

(c)

(d)

Figure 5: (a) depicts histogram of log of total submitted amount of money by providers who has 1 speciality. (b) depicts that by providers who has 3 specialities. (c) depicts that by providers who has 5 specialities. (d) depicts that by providers who has 10 specialities.

Master's Thesis

Construction of Mathematical Models and Development of Efficient Algorithms

Guidance

Professor	Taro JOHO
Assistant Professor	Jiro KOGAKU

Saburo SURI

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 Saburo SURI February 2019

Construction of Mathematical Models and Development of Efficient Algorithms

Saburo SURI

Abstract

This article proposes how to find fraudulent activities from data of healthcare claim in an unsupervised way. Firstly, we split data into homogenous groups by providers' department, number of speciality, and their geographical location, otherwise we may find providers doing rare operations or providers in rural areas because of its lower submitted money than in urban areas. Next, we apply outlier detection method, One-SVM and LOF to that groups. We find variation due to department and geographical location. Finally, we evaluate the effectiveness by labels of providers who are excluded because of their having committed fraud. Proposed method is above random and may be used as a first line of further investigation.