Mining Anomalies in Medicare Big Data using Patient Rule Induction Method

Saad Sadiq*, Yudong Tao*, Yilin Yan†, Mei-Ling Shyu*
Department of Electrical and Computer Engineering
University of Miami, Coral Gables, Florida
Email: *{Saadsadiq, yxt128, shyu}@miami.edu, †y.yan4@umiami.edu

***Abstract*—The public health infrastructure delivers proper health care services as part of the basic needs of the general population. The health care system in the United States is rapidly changing in order to provide a better and convenient healthcare system to the public. Unfortunately, this comprehensive expand has also given rise to healthcare frauds in recent years where losses surge up to $1.8 billion in the country. Organizations such as the Center for Medicare Services (CMS) have started providing accesses to comprehensive medical big data to promote the identification of healthcare frauds as an important research topic. In this paper, we will use the Patient Rule Induction Method (PRIM) based bump hunting method to identify the spaces of higher modes and masses to indicate the peak anomalies in the CMS 2014 dataset. By applying our framework, we can find a way to observe anomalies, which can be attributed to frauds in legal medical practices or other interesting insights in the CMS dataset. This will enable us to characterize the attribute space and explain the events incurring losses to the medicare/medicaid program. The proposed framework is compared with several methods to illustrate the efficiency and effectiveness of the proposed framework for fraud detection.**

I. Introduction

As defined by WHO, the highest attainable standard of health is a fundamental right of every human being [1]. Therefore, the public health systems are related to all the population and can be defined as "all public, private, and voluntary entities that contribute to the delivery of essential public health services within a jurisdiction" [2]. The goal of healthcare systems is to cure as many patients as possible in a reasonable and affordable way. However, the treatment process requires not only medicines and medical facilities but also the services of physicians and other medical staffs. All these necessities are highly valuable and thus very likely unaffordable for individual patients to pay the cost by themselves. Therefore, medical insurance plans are involved to spread the financial burden among all attendees in the network so that the medical system can work more functionally.

On the other hand, misuse or fraudulent activities exist in any insurance system. Estimated by the FBI, 3% to 10% of all insurance billings are frauds [3]. For example, from the $604 billion of total healthcare insurance costs in 2013 [4], the fraudsters can steal $18 to $61 billion in one year. To avoid such a huge waste of public resources, it is urgent and necessary for the government to find an effective way to detect fraudulent activities and then prevent them from happening. However, due to the complexity of the human body, specialized physicians are trained to acquaint only parts of it so that they can give diagnosis and treatment to the patients accordingly and properly. As a result, the treatment plans and procedures can vary dramatically for whom specializes in different fields, which brings more difficulty on how to identify fraudulent claims.

In order to assist in fraudulent healthcare insurance claim detection, The Center for Medicare and Medicaid Services (CMS) has recently began to release the dataset of claims, called Medicaid Dataset [5], which records various medical procedures performed by each medical service provider in the U.S. and the corresponding average amount paid for these drugs, facilities, and treatments. Although there are several existing approaches that focus on information retrieval [6]–[18] and data mining [19]–[31], they do not target medical data, especially for big data. Thus, it is helpful as a basis for developing a fraudulent claim detection system to identify fraudulent information from such medical big data.

Any medical financial system becomes inefficient when it is maliciously and wastefully used. For example, if the use of Urology can be regulated, over $125 million are estimated to be saved [32]. This could potentially endanger the patients who require the same medical resources which were wasted. To standardize the process, when the regulations might be broken is required to be determined. Therefore, the anomaly detection methods can be applied to the Medicaid dataset to detect abnormal behaviors or bad manners of certain physicians, compared to his or her peers [33].

As a specific type of machine learning method, an anomaly detection method [34] can generate an outlier subset from the general set so that those physicians who behave differently from the average can be identified. Please note that it can be more practical when the detected subset is used as a reference since there is no guarantee that all the physicians in the detected outlier subset have actually practiced maliciously. Hence, the further investigation is necessary. In general, one or several statistics of physician behaviors are calculated and then the machine learning method is applied to determine which physicians are the outliers. If a physician is identified as an outlier, he could have provided medical services maliciously or wastefully or have belonged to a special case. For the former situation, the physician behaved aberrantly with possible fraudulent and wasteful use of healthcare insurance, which can be verified by further scrutiny in the practicing habits.

Given a multi-dimensional target function, bump hunting can be applied in order to detect its specific input regions with the most possible related smaller target value. The bump hunting process is performed based on the identification of the intrinsic structure of the target function in an unsupervised way, instead of calculating all the target values in the input space. This identification process can be regarded as a density estimation problem with joint probability densities, as a clustering problem, as a pattern recognition problem, and/or as an anomaly detection process. Since the outliers of the physicians can be modeled as the ones with extreme values among several statistics, bump hunting is applied to discover the possible fraud in the datasets. The fraud is regarded as a kind of bump so it can be detected by the corresponding algorithms. Therefore, the source of insurance loss can be found and then prevent them from happening in the future.

In this paper, we develop a novel fraudulent medical insurance claim detection framework based on Patient Rule Induction Method (PRIM), where abnormal behaviors of the physicians are detected. Our main contributions are shown as follows.

1) We present a novel approach of using PRIM in detecting medicare fraud and also fraud in general. To the best of our knowledge, PRIM has only been used in the medical domains, specifically in cancer data bump hunting, and this is the first time this method has been applied to for fraud detection;
2) The proposed PRIM framework is part of our ongoing research to develop a cross-validated PRIM regression model that extends the existing method from handling only survival cases to also perform general cross-validated regression;
3) We find a new way to characterize the medicare dataset to protrude fraud anomalies. Normalized predictor functions such as ANOVA and Conditional Probability of the charged prices are used in a 1-Rule setting to characterize the input space;
4) The experimental results show our fraud detection framework can effectively shrink the target subset and the detected physicians have a much higher probability than average performing frauds.

The rest of the paper is organized as follows. Section II introduces the CMS dataset, previous work of fraud detection, and the bump hunting method. Section III discusses the methodology of our approach to fraud detection based on the CMS dataset and Section IV illustrates the experimental results of our method. In the end, Section V concludes our work and contributions.

II. Related Work

A. CMS Dataset

CMS has released its Medicaid datasets for three years, from 2012 to 2014. For each year, there are eight datasets published, recording different aspects of the data in the medical treatment process. Since our fraud detection is designed to distinguish the outliers of physicians, those datasets related to other medical service providers are irrelevant, which leaves us three most interested datasets: Physician and Other Supplier Data, Prescriber Data, and DMEPOS Data, recording the statistics of treatment, drug use, and medical equipment use from each physician. For each piece of data in those big datasets of interest, the identity of the physician, his or her gender, whereabouts, specialized field and type, service number, the average submitted charge amount, the average allowed amount in Medicare, and the average standard amount in Medicare are recorded.

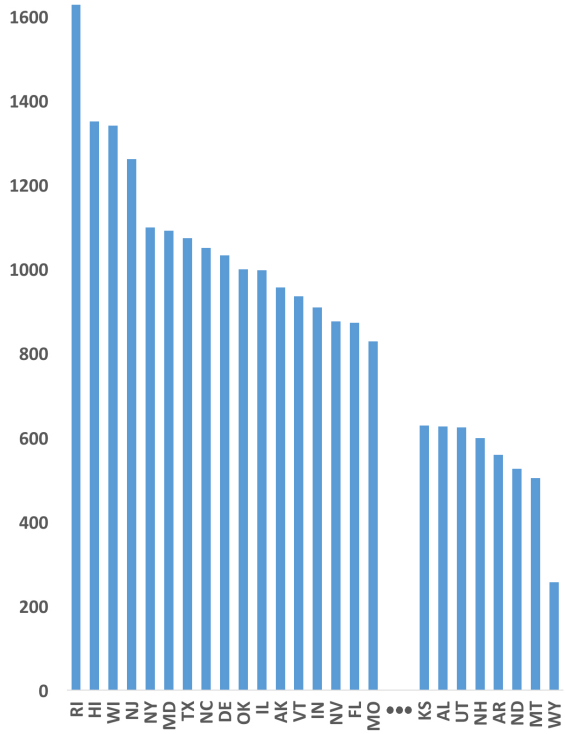


Fig. 1. The average submitted charge amounts of anesthesia for procedures on eyelid in various states

Furthermore, if we summarize the average Medicare payment in different states, we can find an obvious difference between the states. For example, as shown in Fig. 1, the average submitted charge amount of anesthesia around the United States is $2034.40; while the average amount of Florida is $852. Therefore to gain more granularities, fraud detection should be considered state by state to avoid the influence of medical expense differences among states. Furthermore, since Florida has recently suffered severely from several high profile fraudulent malpractices, the focus of this paper is to deal with the state of Florida only. The proposed framework and the experimental results shown in Section IV will reflect the results respectively.

B. Fraud Detection based on the CMS Dataset

Since fraudulent and wasteful use of medical insurance leads to the serious waste of public resources, many researches

on medical insurance fraud detection have been proposed based on these CMS datasets. In paper [35], the physician education background is analyzed to determine how he or she should practice with the 2012 CMS dataset. Their study takes into consideration the medical school tuition, education procedures, possible anomalies, geographical analysis combined with the nationwide school procedure charges, and payment distributions. Therefore, the correlation of the education backgrounds and the physician behaviors can be deducted and used as the evidence to detect who fraudulently and wastefully used the insurance system.

Furthermore, the variability of data in the CMS dataset (big data) can be analyzed alone to detect the fraud activities. [32] proposes a possible approach to perform the detection at the Urology field. The number of patient visits is shown to have a strong correlation with the Medicare reimbursement, which is helpful in fraud detection.

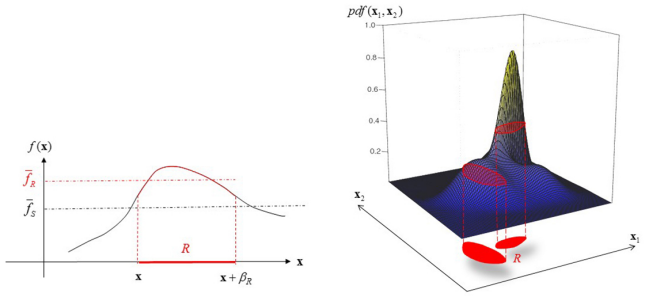
Since the CMS dataset does not provide the label of the fraudulent healthcare providers, the anomaly detection is processed in an unsupervised way. However, Chandola et al. use the fraudulent provider label primarily from the Texas Office of Inspector Generals exclusion database and use a semi-supervised method to detect the frauds [36]. Specifically for the frauds in treatments, typical treatment profiles are used to be compared to what the physicians have provided.

C. Patient Rule Induction Method (PRIM)

Patient Rule Induction Method (PRIM), initially introduced by Fisher and Friedman in 1999 [37] and extended by J. Dazard et. al [38], implements a unified treatment of the "Bump Hunting" task in Survival, Regression and Classification (SRC) settings in a high-dimensional space. This framework is part of our ongoing research to develop a regression model for the PRIMsrc survival framework. PRIM generates the decision rules to delineate a region by recursively peeling non-bump regions from the result and thus the remaining subsets are regarded as bumps. The detected region, called target region R, might be disjointed but with a locally larger or smaller target function value than its average over the whole space, in the input space with a high dimension. The detail implementation of the extended PRIM is given in Algorithm 1.

PRIM uses one or more low-dimensional hyper-rectangles to indicate the approximation of the target region R, whose edges should be parallel to the axes of the input space. Each hyper-rectangle is called a "box" and is defined by the conjunction of several restrictions on the inputs. Therefore, potential fraudulent physicians can be bounded in the identified bump regions.

As shown in Fig. 2a, the target function $f(x)$ is a regression function in $p = 1$. The solution region R (shown in red) corresponds to an interval where $f(x)$ assumes larger average values than over the entire support. Meanwhile, Fig. 2b illustrates an example where the target function is a joint density probability function $pdf(x1, x2)$ simulated from a mixture of bivariate normal distributions in $p = 2$. In higher-dimensional spaces, it is not uncommon to find a solution region R that can be complex and possibly disjointed.



(a) (b)

Fig. 2. Examples of target region(s) of PRIM with different dimensions [38]

Algorithm 1 PRIM

Require: the training dataset $\mathscr{L}_1$ and a maximal box $\{\hat{B}_1\}$ containing it

Ensure: a sequence of boxes $\{\hat{B}_m\}$ identifying all outliers

1: **for** $m \in \{1, ..., M\}$ **do**
2: Generate a box $\hat{B}_m$ containing all training data in $\mathscr{L}_m$
3: **for** $l \in \{1, ..., L\}$ **do**
4: **top-down peeling:** Generate a box $\hat{B}_m^l$ by conducting a stepwise attribute selection
5: **bottom-up pasting:** Expand the box $\hat{B}_m^l$ along any face (pasting) as long as the resulting box maintains a higher box mass than the initial $\hat{B}_1$
6: Stop the peeling loop when a minimal box support $\hat{\beta}_m^L$ of $\hat{B}_m^L$ reaches a minimum box support $0 \leq \beta_0 \leq 1$, expressed by the user as a fraction of the data, i.e., $\hat{\beta}_m^L \leq \beta_0$
7: **end for**
8: Given a sequence of nested boxes $\{\hat{B}_m^L\}$, where L is the estimated number of peeling/pasting steps with different numbers of observations in each box. Call the next box $\hat{B}_{m+1}$.
9: Remove the data in box $\hat{B}_m$ from the training dataset: $\mathscr{L}_{m+1} = \mathscr{L}_m \backslash \hat{B}_m$
10: Stop the covering loop when running out of data or when a minimal number of observations remains within the last box $\hat{B}_M$, where the final box support $\hat{\beta}_M \leq \beta_0$
11: **end for**

III. Methodology

A. Framework

Empirical data indicates that the state of Florida is the hotbed of healthcare frauds. This was ensured by the recent high profile cases caught in South Florida. Therefore, the data will most definitely hold fraudulent outliers and medical payment scammers. In our proposed framework, we begin by formulating a conditional probability of the drug, equipment, or service appearing in the prescription given by a physician belonging to a specific area of medicine.

$$\Pr(\text{Area}_x|\text{Prescription}_y) = \frac{\Pr(\text{Area}_x \cap \text{Prescription}_y)}{\Pr(\text{Prescription}_y)} \quad (1)$$

This derived attribute calculates, for example, if a physician prescribes eye drops to a patient, how likely the physician is an ophthalmologist. A lower value of the conditional probability indicates a higher likelihood that the medical treatment is a fraudulent flag or an improper prescription.

It is challenging to identify fraud in certain cases when common prescriptions are overused and there exists homogeneity in the prescriptions. Therefore, we calculate the mean of that specialist's treatment areas and evaluate ANOVA to prove that the mean value is consistently biased. The ANOVA test is performed to prove the statistical significance against the hypothesis that if a medical specialist is consistently charging similar prices against all his/her patients, then this may be a plausible indication towards a fraud. The F-score in the ANOVA test is calculated using Equation (2a) and Equation (2b).

$$F_1 = \frac{\text{variance between treatments}}{\text{variance within treatments}} \quad (2a)$$

$$F_2 = \frac{MS_{\text{Treatments}}}{MS_{\text{Error}}} = \frac{SS_{\text{Treatments}}/(I-1)}{SS_{\text{Error}}/(n_T - I)} \quad (2b)$$

where MS is the mean square, SS is the sum of square, I is the number of treatments, and n_T is the total number of cases.

We start plotting the individual PDF of the derived attributes as shown in Fig. 3. It was observed in Fig. 3(a) that the population density in the lower conditional probability region was highly unstable and composed of multivariate distribution; whereas the upper probability region indicated a near monotonic region of low anomalies. Therefore, we marked two sections of the conditional probability, i.e., higher and lower than the average. Next, the data was split to 66.6% and 33.4%, and the Patient Rule Induction Method (PRIM) is performed to generate the boxes to characterize the potential fraudulent cases. Once the boxes are generated where the fraudulent medical insurance claims are located, a much smaller pool of potential fraudsters is identified.

B. Transformation and Normalization

Given the derived attributes, Fig. 3(b) and Fig. 3(c) indicate very high kurtosis density plots, where kurtosis is defined as the 4th central moment, indicating a small part of the population consistently requesting incongruent charges. This causes an outlier in the input feed of a linear regression fit and we were having difficulties predicting the fit.

Since the regression basis is the squared distances in the attribute space, the outlier distances become large when they are squared. For example, the slope of a regression fit line, as described in Equation (3), is inversely proportional to the variance of X. Thus the outliers change the variance of X to be much higher, causing the fit to rotate down and taking it away from the truth, where X ($x_i \in X$) and Y ($y_i \in Y$) are the attribute space and outcome respectively and i=1 to n. We can apply the log rank transformation (LRT) or the square root transformation to make the slope line more straight. These transformations will pull in the curve too to make the distribution become more Gaussian. Then it is more likely to work with the linear regression model of the modified PRIM.

TABLE I
CONVERSION RANGES AFTER Z-SCORE NORMALIZATION

	Min	1^{st}	Median	Mean	3^{rd} Qu.	Max
U	-99.580	-21.880	-4.463	0.000	12.680	4517
U'	-1.887	-0.045145	-0.08457	0.0	0.24	85.6
S	-97.94	79.31	147.7	241.8	265.4	71650
S'	-0.583	-0.28	-0.16	0.0	0.04	122.6
A	0.0	4800	2753	3.9e27	1.4e4	2.3e32
A'	-0.005	-0.005	-0.005	0.0	0.0057	314.1

$$m(b) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\text{Cov}(X,Y)}{\text{Var}(X)} \quad (3)$$

Moreover, as shown in Fig. 4, we observe that several of the 4th to 8th degree variables have linear relationships; while the others have very large outlier ranges. To rectify this problem, we apply z-score normalization to our dataset $\mathscr{L}_1$ in order to bring the data to unified ranges. The formula for Z-score normalization is given in Equation (4).

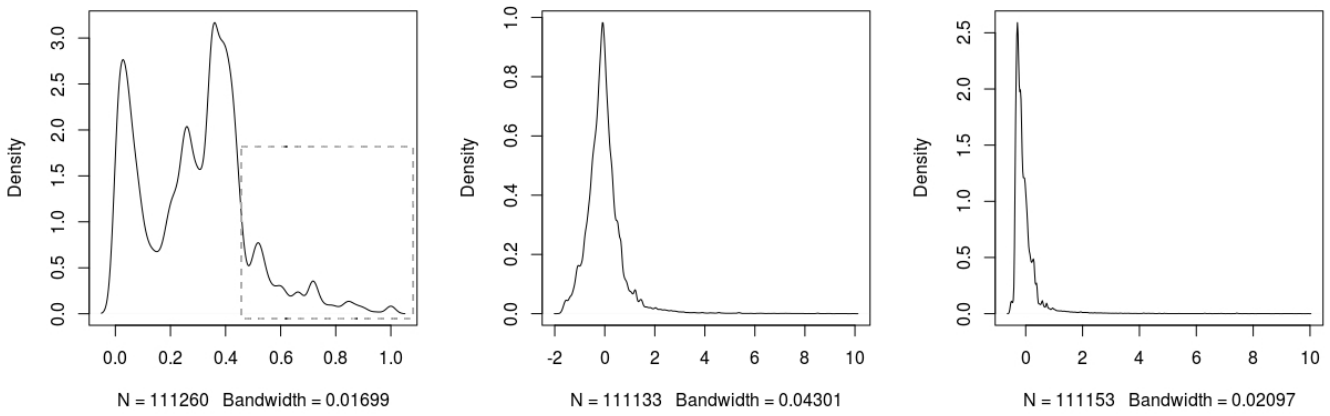
$$Z = \frac{x_j - \mu_{x_j}}{\sigma_{x_j}} \quad (4)$$

where Z represents the normalized matrix and attributes x_j for $j \in \{1, ..., p\}$. The resultant changes in the ranges are also shown in Table I, where U refers to the case when the charged amount is higher than other specialists, S refers to the case when the charged amount is higher than the Medicare standard, and A refers to ANOVA in charged prices. The prime symbols of each of the above mentioned variables represent the normalized ranges.

IV. EXPERIMENT SETUP & RESULTS

After the transformation, we start executing the Patient Rule Induction Method (PRIM) on the derived dataset until we reach a plateau in the peeling steps of the method. The procedure is previously described in Algorithm 1. In the step of top-down peeling, the box $\hat{B}_m$ is shrunk by compressing one face (peeling), so as to peel off an arbitrary fraction α_0 from an attribute x_j for $j \in \{1, ..., p\}$. The direction of peeling j that yields the largest box increase rate is chosen.

Given a sequence of (not necessarily nested) boxes $\{\hat{B}_M\}$, where M is the estimated total number of boxes covering $\mathscr{L}_1$. The decision rules of all boxes $\{\hat{B}_M\}$ can be collected into a combined final decision rule set $\hat{\mathscr{R}}$ applicable to a solution region $\hat{R}$ of the following form:



(a) (b) (c)

Fig. 3. Density plots of (a) Conditional Probability; (b) Charge higher than other physicians; and (c) Charge higher than the Medicare standard

TABLE II
CROSS-VALIDATED BOXES GENERATED AND THEIR ATTRIBUTES

Cross Validation	Learning Set		Test Set	
	$Obj_{\hat{B_m}}$	$\hat{\beta_m}$	$Obj_{\hat{B_m}}$	$\hat{\beta_m}$
CV_1	B1 0.024	0.304	B1 0.024	0.217
	B2 0.031	0.182	B2 0.030	0.179
	B3 0.050	0.261	B3*0.062	0.604
	B4*0.083	0.253		
CV_2	B1 0.024	0.264	B1 0.024	0.290
	B2 0.025	0.186	B2 0.034	0.220
	B3 0.048	0.319	B3 0.047	0.233
	B4*0.087	0.229	B4*0.087	0.256
CV_3	B1 0.022	0.246	B1 0.023	0.176
	B2 0.030	0.215	B2 0.031	0.227
	B3*0.065	0.539	B3*0.061	0.597

$$\hat{\mathscr{R}} = \bigcup_{m=1}^{M} \hat{\mathscr{R}}_m \tag{5a}$$

$$\hat{R} = \bigcap_{j \in J} (x_j \in [t_{-j,m}, t_{+j,m}]) \tag{5b}$$

where $\hat{\mathscr{R}}$ refers to the combined final decision rule set, $\hat{R}$ refers to the common solution region from M sub-spaces having $[t_{-j,m}, t_{+j,m}]$ boundaries, and J is the total number of attributes in the input space. The final decision rule set $\hat{\mathscr{R}}$ can then be applied to identify whether a claim is more likely to be fraudulent or wasteful based on the given dataset.

Meanwhile, Fig. 4 illustrates the relationships between the different attributes. Here we can observe that several of the variables display strong relationships, which validates the hypothesis of co-variance among the attributes. Table II shows the cross-validated boxes generated by PRIM from the learning set and test set respectively, where $Obj_{\hat{B_m}}$ and $\hat{\beta_m}$ represent the box-function and box-mass of individual boxes

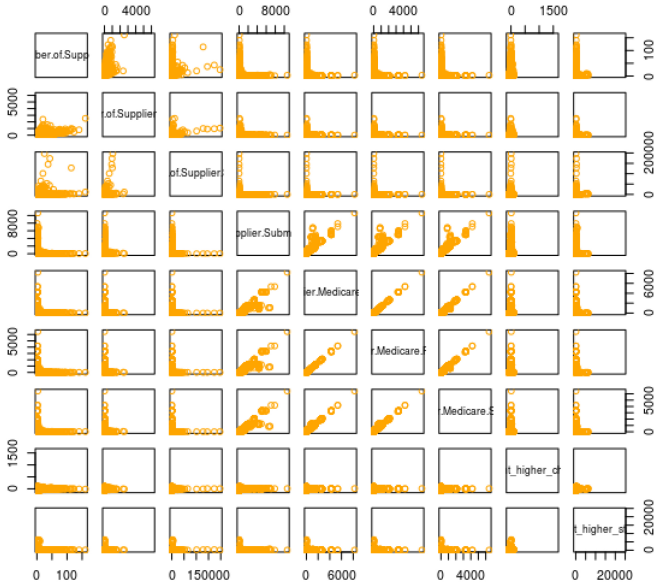


Fig. 4. Relationships between the attributes

respectively. The PRIM algorithm starts with 12-dimensional data with the first Box $\hat{B_1}$ being 0.024 (i.e., only 2.4% of the learning data linked to the low conditional probability). Then further boxes were generated encompassing more areas of the learning set and testing set respectively. A combination of the generated boxes was to characterize the input space and also to predict the test set.

Let $\hat{B_M}$ be the final box derived using PRIM and having region $\hat{R}$. To evaluate if the final box $\hat{B_M}$ having box-mass $\hat{\beta_M}$ characterizes the low conditional probability in region $\hat{R}$, we compare the number of observations contained in $\hat{R}$ and also those residing outside of $\hat{R}$. We adopt the confusion matrix illustrated in Table III. After calculating the numbers of a, b, c, and d in Table III, we calculate the sensitivity, recall, and F1 measure using the following equations.

TABLE III
CONFUSION MATRIX OF SIMULATED OBSERVATIONS

		Classified	
		Obs. inside $\hat{R}$	Obs. outside $\hat{R}$
Actual	Obs. inside $\hat{R}$	**a:** the number of observations inside $\hat{R}$, classified correctly	**b:** the number of observations inside $\hat{R}$, classified incorrectly
	Obs. outside $\hat{R}$	**c:** the number of observations outside $\hat{R}$, classified incorrectly	**d:** the number of observations outside $\hat{R}$, classified correctly

$$\text{Sensitivity} = a/(a+b) \tag{6a}$$

$$\text{Recall} = a/(a+c) \tag{6b}$$

$$F1 = 2 \times \frac{\text{Sensitivity} \times \text{Recall}}{\text{Sensitivity} + \text{Recall}} \tag{6c}$$

The F1 measure is defined as the F-score and it has the values between zero and one. The value close to one implies that most of the observations are classified correctly. Another important measure is to evaluate the accuracy defined in Equation (7). However, this measure is susceptible to imbalanced observations in $\hat{R}$.

$$\text{Accuracy} = (a+d)/(a+b+c+d) \tag{7}$$

A. Visualization of the Boxes

There is a wide variety of health systems around the world to meet the needs of populations in different countries and regions. In the United States, the prominent public healthcare system is known as the Medicare program. In the experimental part, we feed data from the Medicare database into our model and try to identify healthcare frauds. The visualizations drawn are shown in Fig. 5 with 2-dimensional vectors of several derived attributes. The solid boxes in Fig. 5(a) and Fig. 5(b) represent the shrunk region that we identify as a bump, in space of the two corresponding variables along the axis. Fig. 5(a) shows the small region between the $Percent\_higher\_charged$ and $Anova\_F$ variables that cause the bump in our outcome variable. Similarly, Fig. 5(b) shows that the bump is along the entire range of $Percent\_higher\_charged$ when compared to $Avg\_Supplier\_Medicare\_Std\_Amount$. The box in Fig. 5(c) was made transparent to accommodate the scatter plot of the points between $Percent\_higher\_charged$ and $Percent\_higher\_standard$.

B. Performance Comparison

The goal is to compare bump hunting with other algorithms by measuring the reproducibility of finding the fraudulent region of the conditional probability. This is done by treating the prediction as a classification problem and identifying if the same bump is found repeatedly in the low conditional probability region. The second step is to characterize the attribute space using variable importance and/or variable correlation metrics. The comparison is to see if some other leading frameworks are able to persistently identify and characterize the attribute space that causes the low conditional probability. To conduct the comparison, our PRIM algorithm is evaluated against several popular classifiers including Support Vector Machine (SVM) [39], [40], Naive Bayes (NB) [41], Random Forest (RF) [42], [43], discriminant analysis classifier (DAC) [44]–[46] and Logistic Regression (LR).

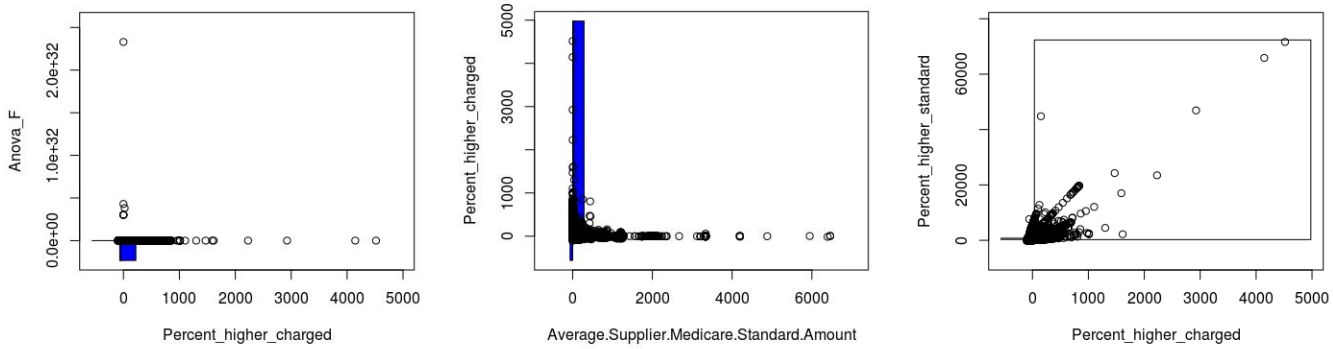
TABLE IV
F-SCORES AND ACCURACY COMPARIASION

Classifier	F-score	Accuracy
SVM	0.654	0.564
NB	0.639	0.567
RF	0.586	0.578
DAC	0.689	0.535
LR	0.688	0.539
PRIM	**0.785**	**0.699**

The general idea of the Support Vector Machine (SVM) classifier is to build a separating hyperplane to classify the dataset so that the geometric margin is maximized. The "Naive Bayes" (NB) approach is based on applying the Bayes theorem with strong independence assumptions between the attributes in the CMS dataset. Random forest (RF) is an ensemble learning method for classification by constructing a multitude of decision trees at the training time and outputting the class that is the mean prediction of the individual trees. The discriminant analysis classifier (DAC) assumes that the data from different classes are generated based on different Gaussian distributions. In the training phase, the fitting function calculates the parameters of a Gaussian distribution for each class; while in the testing stage, the trained classifier finds the class with the smallest misclassification cost. Logistic regression (LR) is another predictive analysis and a kind of generalized linear model. It can be used when the dependent variable is binary, which is the case in this study. Instead of just predicting binary-valued labels in linear regression, logistic regression uses a different hypothesis class to predict the probability that a given example belongs to the positive (fraud) class versus the probability that it belongs to the negative (non-fraud) class by a logistic function. As shown in Table IV, it is clear that other comparative classifiers struggle to classify the uniqueness of the health care data. Thus potential frauds having a low conditional probability (i.e., indicating malpractice) may go unidentified. PRIM, however, achieves better F-score and accuracy results because it characterizes the input space and then classifies new instances.

C. Further Discussion

The rationale behind this study is that we believe any fraud is an individual act and organizations will not commit frauds because the risk to benefit ratio is too high. We develop our ground truth from the known fraud cases in Florida



(a) (b) (c)

Fig. 5. Box Relationship plots of (a) ANOVA vs. % Higher charged; (b)% higher charged vs. Standard medicare amount; (c)% higher charged vs % higher than medicare standard

[47]–[50]. By observing these real-world cases, we observed that the majority of these fraudulent activities are built on false claims. False claims are attributed to overcharging the medicare program by billing for unnecessary items, billing an item for a higher price, or both. To identify unnecessary items, we find the conditional probability of a prescribed item falling in a category. If the conditional probability is too low, then it is a matter of concern. The method to design fraud detecting frameworks usually starts with scenarios where you know what is supposed to be in the box and compare it with several other methods like SVM, Random Forests, or clustering. Bump hunting identifies what cases are at the extremes and characterizes their X-space while trying to identify the partitions in that attribute space with respect to the response. We believe that a certain level of depth into the data is necessary to build any substantial hypothesis for predicting the frauds. Therefore, we work on the data of individual claims prescribed by the physicians so that we are able to trace and compare individual prescriptions.

V. Conclusion

Public health policies and regulations have been continuously striving to improve the quality of public health systems and better serve the people of the United States. One service in the system is the Medicare/Medicaid program that enables health benefits to low-income families and individuals. However, the program is filled with fraudsters causing annual losses of almost $1.8 billion. Our proposed framework analyzed the CMS big data to identity a subset of physicians who are potentially involved in fraudulent and wasteful use of Medicare insurance. The state of Florida was isolated for this study because of the high and persistent evidence of health care fraud in Florida. The experimental results show that our fraud detection framework can effectively shrink the target dataset and deduce a potential suspect subset of physicians who involve several anomalous claims and probably are qualified as fraudsters. The attribute sub-space and their correlations are used in PRIM to characterize the low conditional probability region. The attribute space was characterized by PRIM that provides a deeper understanding of how certain attributes are the key predictors in identifying frauds. The identified bumps were validated and compared with other classification methods to demonstrate the efficiency and effectiveness of the fraud characterization.

References

[1] WHO. (2015) Health and human rights. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs323/en/

[2] CDC. (2014) The public health system and the 10 essential public health services. [Online]. Available: https://www.cdc.gov/nphpsp/essentialservices.html

[3] L. Morris, "Combating fraud in health care: An essential component of any cost containment strategy," *Health Affairs*, vol. 28, no. 5, pp. 1351–1356, 2009.

[4] K. M. King. (2014) Medicare fraud: Progress made, but more action needed to address medicare fraud, waste and abuse. United State Government Accountability Office. [Online]. Available: http://www.gao.gov/products/GAO-14-560T

[5] CMS. (2016) Medicare provider utilization and payment data. [Online]. Available: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/index.html

[6] M. L. Shyu, Z. Xie, M. Chen, and S. C. Chen, "Video semantic event/concept detection using a subspace-based multimedia data mining framework," *IEEE Transactions on Multimedia*, vol. 10, no. 2, pp. 252–259, Feb 2008.

[7] L. Lin, G. Ravitz, M.-L. Shyu, and S.-C. Chen, "Video semantic concept discovery using multimodal-based association classification," in *Proceedings of the IEEE International Conference on Multimedia & Expo*, July 2007, pp. 859–862.

[8] L. Lin, M.-L. Shyu, G. Ravitz, and S.-C. Chen, "Video semantic concept detection via associative classification," in *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*. IEEE, 2009, pp. 418–421.

[9] X. Li, S.-C. Chen, M.-L. Shyu, and B. Furht, "Image retrieval by color, texture, and spatial information," in *Proceedings of the 8th International Conference on Distributed Multimedia Systems*, September 2002, pp. 152–159.

[10] S.-C. Chen, M.-L. Shyu, and C. Zhang, "An intelligent framework for spatio-temporal vehicle tracking," in *Proceedings of the 4th IEEE International Conference on Intelligent Transportation Systems*, August 2001, pp. 213–218.

[11] X. Huang, S.-C. Chen, M.-L. Shyu, and C. Zhang, "User concept pattern discovery using relevance feedback and multiple instance learning for content-based image retrieval," in *Proceedings of the Third International Workshop on Multimedia Data Mining, in conjunction with the 8th ACM International Conference on Knowledge Discovery & Data Mining*, July 2002, pp. 100–108.

[12] S.-C. Chen, S. Sista, M.-L. Shyu, and R. Kashyap, "Augmented transition networks as video browsing models for multimedia databases and multimedia information systems," in *Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence*, 1999, pp. 175–182.

[13] S.-C. Chen, A. Ghafoor, and R. L. Kashyap, *Semantic Models for Multimedia Database Searching and Browsing*. Norwell, MA, USA: Kluwer Academic Publishers, 2000.

[14] S.-C. Chen, M.-L. Shyu, and R. Kashyap, "Augmented transition network as a semantic model for video data," *International Journal of Networking and Information Systems*, vol. 3, no. 1, pp. 9–25, 2000.

[15] M.-L. Shyu, S.-C. Chen, and R. Kashyap, "Generalized affinity-based association rule mining for multimedia database queries," *Knowledge and Information Systems (KAIS): An International Journal*, vol. 3, no. 3, pp. 319–337, August 2001.

[16] C. Chen, Q. Zhu, L. Lin, and M.-L. Shyu, "Web media semantic concept retrieval via tag removal and model fusion," *ACM Transactions on Intelligent Systems and Technology*, vol. 4, no. 4, pp. 61:1–61:22, October 2013.

[17] C. Haruechaiyasak, M.-L. Shyu, and S.-C. Chen, "Web document classification based on fuzzy association," in *Proceedings 26th Annual International Computer Software and Applications*, 2002, pp. 487–492.

[18] Q. Zhu, L. Lin, M.-L. Shyu, and S.-C. Chen, "Effective supervised discretization for classification based on correlation maximization," in *Proceedings of the IEEE International Conference on Information Reuse and Integration*, 2011, pp. 390–395.

[19] M.-L. Shyu, C. Haruechaiyasak, and S.-C. Chen, "Category cluster discovery from distributed www directories," *Information Sciences*, vol. 155, no. 3, pp. 181–197, 2003.

[20] M.-L. Shyu, S.-C. Chen, and R. L. Kashyap, "Generalized affinity-based association rule mining for multimedia database queries," *Knowledge and Information Systems*, vol. 3, no. 3, pp. 319–337, 2001.

[21] S.-C. Chen, M.-L. Shyu, C. Zhang, and M. Chen, "A multimodal data mining framework for soccer goal detection based on decision tree logic," *International Journal of Computer Applications in Technology*, vol. 27, pp. 312–323, 2006.

[22] M.-L. Shyu, C. Haruechaiyasak, S.-C. Chen, and N. Zhao, "Collaborative filtering by mining association rules from user access sequences," in *Proceedings of the International Workshop on Challenges in Web Information Retrieval and Integration*, April 2005, pp. 128–135.

[23] X. Chen, C. Zhang, S.-C. Chen, and M. Chen, "A latent semantic indexing based method for solving multiple instance learning problem in region-based image retrieval," in *Multimedia, Seventh IEEE International Symposium on*, Dec 2005, pp. 37–44.

[24] S.-C. Chen and R. Kashyap, "Temporal and spatial semantic models for multimedia presentations," in *Proceedings of the 1997 International Symposium on Multimedia Information Processing*, 1997, pp. 441–446.

[25] S.-C. Chen, M.-L. Shyu, and C. Zhang, "Innovative shot boundary detection for video indexing," in *Video Data Management and Information Retrieval*, S. Deb, Ed. Idea Group Publishing, 2005, pp. 217–236.

[26] S.-C. Chen, M.-L. Shyu, C. Zhang, and R. L. Kashyap, "Identifying overlapped objects for video indexing and modeling in multimedia database systems," *International Journal on Artificial Intelligence Tools*, vol. 10, no. 4, pp. 715–734, 2001.

[27] M.-L. Shyu, T. Quirino, Z. Xie, S.-C. Chen, and L. Chang, "Network intrusion detection through adaptive sub-eigenspace modeling in multiagent systems," *ACM Transactions on Autonomous and Adaptive Systems*, vol. 2, pp. 9:1–9:37, 2007.

[28] M.-L. Shyu, C. Haruechaiyasak, and S.-C. Chen, "Category cluster discovery from distributed www directories," *Journal of Information Sciences*, vol. 155, pp. 181–197, 2003.

[29] X. Chen, C. Zhang, S.-C. Chen, and S. Rubin, "A human-centered multiple instance learning framework for semantic video retrieval," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 39, no. 2, pp. 228–233, 2009.

[30] M.-L. Shyu, Z. Xie, M. Chen, and S.-C. Chen, "Video semantic event/concept detection using a subspace-based multimedia data mining framework," *IEEE Transactions on Multimedia*, vol. 10, pp. 252–259, February 2008.

[31] S.-C. Chen, M.-L. Shyu, S. Peeta, and C. Zhang, "Learning-based spatio-temporal vehicle tracking and indexing for transportation multimedia database systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 4, no. 3, pp. 154–167, Sept 2003.

[32] J. S. Ko, H. Chalfin, B. J. Trock, Z. Feng, E. Humphreys, S.-W. Park, H. B. Carter, K. D. Frick, and M. Han, "Variability in medicare utilization and payment among urologists," *Urology*, vol. 85, no. 5, pp. 1045–1051, 2015.

[33] C. K. Reddy and C. C. Aggarwal, Eds., *Healthcare Data Analytics*. CRC Press, 2015.

[34] M.-L. Shyu, K. Sarinnapakorn, I. Kuruppu-Appuhamilage, S.-C. Chen, L. Chang, and T. Goldring, "Handling nominal features in anomaly intrusion detection problems," in *15th International Workshop on Research Issues in Data Engineering: Stream Data Mining and Applications (RIDE-SDMA 2005)*, 2005, pp. 55–62.

[35] K. Feldman and N. V. Chawla, "Does medical school training relate to practice? Evidence from big data," *Big Data*, vol. 3, no. 2, pp. 103–113, 2015.

[36] V. Chandola, S. R. Sukumar, and J. C. Schryver, "Knowledge discovery from massive healthcare claims data," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, 2013, pp. 1312–1320.

[37] J. H. Friedman and N. I. Fisher, "Bump hunting in high-dimensional data," *Statistics and Computing*, vol. 9, no. 2, pp. 123–143, 1999.

[38] J. E. Dazard and J. S. Rao, "Local sparse bump hunting," *Journal of Computational and Graphical Statistics*, vol. 19, no. 4, pp. 900–929, 2012.

[39] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.

[40] S.-i. Amari and S. Wu, "Improving support vector machine classifiers by modifying kernel functions," *Neural Networks*, vol. 12, no. 6, pp. 783–789, 1999.

[41] K. P. Murphy, "Naive bayes classifiers," *University of British Columbia*, 2006.

[42] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: http://dx.doi.org/10.1023/A:1010933404324

[43] M. R. Segal, "Machine learning benchmarks and random forest regression," *Center for Bioinformatics & Molecular Biostatistics*, 2004.

[44] L. Lin, G. Ravitz, M.-L. Shyu, and S.-C. Chen, "Effective feature space reduction with imbalanced data for semantic concept detection," in *Proceedings of the IEEE International on Sensor Networks, Ubiquitous, and Trustworthy Computing*, June 2008, pp. 262–269.

[45] T. Meng and M.-L. Shyu, "Leveraging concept association network for multimedia rare concept mining and retrieval," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, Melbourne, Australia, July 2012.

[46] L. Lin and M.-L. Shyu, "Weighted association rule mining for video semantic detection," *International Journal of Multimedia Data Engineering and Management*, vol. 1, no. 1, pp. 37–54, 2010.

[47] E. R. A. Batchelor, "12 arrested in miami-based medical fraud scheme," newspaper, WPLG, Aug. 9, 2016.

[48] T. Jones, "Doctor, therapist arrested in miami medical fraud investigation," newspaper, CBS Miami, Dec. 20, 2015.

[49] D. Mangan, "$1 billion medicare bust in florida is biggest 'criminal health-care fraud case ever," newspaper, CNBC, Jul. 22, 2016.

[50] K. Kennedy, "3 charged in $1 billion health care fraud that took advantage of medicare in Miami," newspaper, NBC 6 South Florida, Jul. 22, 2016.