# An Anomaly Detection Method for Medicare Fraud Detection

Weijia Zhang, Xiaofeng He*
Shanghai Key Laboratory of Trustworthy Computing,
School of Computer Science and Software Engineering, East China Normal University,
Shanghai, China
Email: wjzhang081@163.com, xfhe@sei.ecnu.edu.cn

***Abstract*—With the improvement of medical insurance system, the coverage of medicare increases a lot. However, while the expenditure of this system is continuously rising, medicare fraud is causing huge losses for this system. Traditional medicare fraud detection greatly depends on the experience of domain experts, which is not accurate enough and costs much time and labor.**

**In this study, we propose a medicare fraud detection framework based on the technology of anomaly detection. Our method consists of two parts. First part is a spatial density based algorithm, called improved local outlier factor (imLOF), which is more applicable than simple local outlier factor in medical insurance data. Second part is robust regression to depict the linear dependence between variables. Some experiments are conducted on real world data to measure the efficiency of our method.**

## I. Introduction

With the increasement of the participants of medical insurance system of China, medicare fraud becomes a prominant problem. Medicare fraud refers to the behaviors that medical service providers or individuals misuse the medical insurance fund in various ways.

Typically, there are three kinds of medicare fraud that we summarize as follows:

- *Excessive Medical Treatment*: Some doctors and nurses intentionally induce the patients to consume some unnecessary and expensive medical treatment or overdose some medicine.
- *Decomposing Hospitalization*: In the medical insurance system, there is a ceiling to a medicare claim. Beyond this ceiling, medical insurance fund will decrease the amount of payment and even refuse to pay for it. To avoid reaching the ceiling, some medical service providers decompose one hospitalization by discharging the patients away and then admitting them soon.
- *Illegal Claim*: This refers to the behaviors of cheating the medical insurance fund to pay for the expense which should be paid by themselves.

The Green Book of Health Reform and Development of China, published in 2014, predicted the expenditure would exceed the income in 2017 and the deficit would reach 735 billion RMB in 2024. Medicare fraud worsens this financial strain greatly. However, facing the situation of wasting medical resources caused by medicare fraud, the medical insurance institutes take limited countermeasures. The biggest problem is to detect the fraud activities. Traditional medicare fraud detection usually uses rule-based method proposed by domain experts: a hospitalization will be suspected as medicare fraud if any of these rules have been violated [1]. The suspicious instances must be examined manually by domain experts. However, the number of instances that violate the rules is still large and the percentage of normal instances is high. As a result, both the efficiency and accuracy are badly influenced. In addition, these methods greatly depend on the experience of domain experts and rule-based systems are not flexible enough to deal with medicare fraud behaviors with constant change.

This study proposes a data-driven method based on anomaly detection technology. Anomaly detection, which is an important domain of machine learning, refers to find data that do not conform to expected behavior. In the domain of medical insurance, the medicare fraud behaviors can be considered as anomalies because they are vastly different from the normal ones. Because of the variety of medicare fraud, we outline a framework with a two-tier detection method. Targeting excessive medical treatment and decomposing hospitalization, we propose the improved local outlier factor (imLOF), which enhances the simple local outlier factor (simLOF) [2] by using a clustering algorithm. The rest one medicare fraud, illegal claim, is detected using robust regression because it is insensitive to outliers.

Our main contributions are summarised as follows:

- By analysing the medical data and the medicare fraud behaviors, we design a framework to detect various medicare fraud. Applying anomaly detection techniques, this framework can greatly improve the efficiency of medicare fraud detection.
- Based on simple local outlier factor which is a very widely-used anomaly detection method, we analyze its limitation in medicare domain and propose an improved local outlier factor (imLOF) to decrease false negative rate. The robust regression model is used for modeling the relationship between hospitalization expenditure and medicare reimbursement. The residual of each instance is regarded as its anomaly score.
- Two groups of experiments are conducted with the real-world data to demonstrate our method is effective.

*Corresponding author.

978-1-5386-3120-1/17 $31.00 © 2017 IEEE
DOI 10.1109/ICBK.2017.47

The rest of the paper is organized as follows. In section II, we introduce some methods of anomaly detection and works of others about medicare fraud detection. In section III, the datasets we use are introduced. Section IV details the method we propose in this study, including the framework design and the models. The result of experiments is presented in section V. The last section covers our conclusion and future work.

## II. Related Work

Medical insurance systems all over the world are beset with the problem of medicare fraud. There are researches about using anomaly detection methods to detect this fraud activities both in China and abroad.

Bauder et al. [3] use the Physician and other Supplier PUF datasets that the Centers for Medicare and Medicaid Services (CMS) released in 2012 and 2013 to train several models. Then They compare the values generated by models with the corresponding actual values to calculate the anomaly score. Tsai and Ko [4] employ a knowledge engineering methodology to analyze the medical insurance fraud, which reducing the large labor cost and time consuming of the existing domain systems. Tang et al. [5] cluster the consumer data and label them on certain criteria. Then the Hidden Markov Model are used to find hidden temporal patterns from the clusters of consumers. Finally, an anomaly score is assigned to each consumer. A medical insurance company of Chile designs a system based on neural networks to discover medical abuse and fraud [6], which is able to process the claims online.

Application of anomaly detection in the medical insurance system of China is not that much. Shi et al. [7] propose a fraud resilient medical insurance claim system. This system can detect behaviors of both abnormal categories and abnormal frequencies using semi-supervised isomap cluster method and simple local outlier factor. Xie et al. [8] improve simple Local Outlier Factor by introducing information entropy when calculating the distance between data instances. In addition, to accelerate training, the algorithm is parallelized on Hadoop. Shao et al. [9] cluster the medical insurance data by k-means and gaussian mixture model first. Then, with the help of domain experts, the clusters are labeled as normal or anomalous. At last, C4.5 decision tree is used to classify the testing data. Li [10] proposes a sampling method based on the kmeans clustering algorithm and smote sampling method to deal with the problem of unbalanced medicare data. The data after sampled is fed to an improved random forest classifier to detect anomalies.

## III. Data

We first describe all the three datasets used in this study. These datasets are provided by the medical insurance bureau of a city in Sichuan Province, including medical insurance claims, hospitalization details and information of insured residents from 2005 to 2015.

**Medical Insurance Claims**: This dataset contains 722,883 claims (corresponding to the same amount of hospitalizations), which involve 3,228 hospitals and 221,101 patients. A piece of data in this dataset consists of code of patient, code of hospital, level of hospital, code of disease, various fees, reimbursement, date of being in and out hospital and so on. The various fees include:

- class-A fee: fee of the most needed and relatively cheap medical treatment
- class-B fee: fee of the alternative and relatively expensive medical treatment
- class-C fee: fee of medical treatment not in the list of class-A and class-B

**Hospitalization Details**: There are total 50,637,729 records in this dataset. The relationship between a hospitalization and the records in this dataset is one-vs-many. Each record contains the following fields: code of medical treatment or medicine, price, amount, category and so on. The categories are also class-A, class-B and class-C, whose meanings are the same as what we described before.

**Information of Insured Residents**: In this City, 692,336 residents are covered by the medical insurance system. This dataset contains information of all these residents, including personal code, status (working or retired), sex, age, wage and so on.

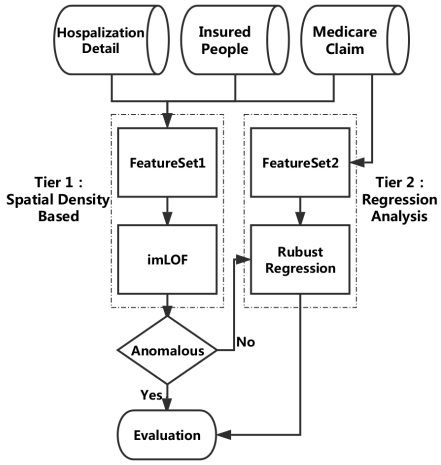



Fig. 1. Framework

## IV. Methodology

This section details the method we use in this study. Figure 1 is the simplified flowchart of the framework. Our method consists of two tiers of detection. One uses the spatial density information to detect anomalies in the sparse region. The other uses regression analysis to detect anomalies which disturb the linear dependence between variables.

### A. Framework

The proposed framework consists of several components as follows:

- **Feature Extraction**: It harvests the medical insurance datasets and transform the raw data into features that the following models need.

- **imLOF**: It detects excessive medical treatment and decomposing hospitalization using spatial density information.
- **Robust Regression**: It detects illegal claims through checking the linear dependence between the expenditure and the reimbursement.
- **Evaluation**: The result will be evaluated by the domain experts.

### B. Feature Extraction

Following is the key feature of each kind of medicare fraud:

- *Excessive Medical Treatment*: higher expenditure than normal case
- *Excessive Medical Treatment*: a patient being in hospital again shortly after discharged from hospital
- *Decomposing Hospitalization*: the claim not proportionating to the expanse

Because imLOF targets detecting excessive medical treatment, which differs from normal hospitalization in its relatively higher expenditure, and decomposing hospitalization, which can be inferred from the frequency and time interval of hospitalizations, the features that imLOF needs contain both all kinds of expenditure and the information of insured residents:

- *Expenditure during a hospitalization*: This part contains three fields of the fees of medical treatments (class-A, B and C) and nine fields of the fees of drugs.
- *Information of Insured Residents*: This part contains the sex of insured people, being working or retired, hospitalization frequency in a year and days since last hospitalization.

The training of the robust regression model only needs all the expenditures as the input and the reimbursement as the response value.

### C. Simple Local Outlier Factor

Simple Local Outlier Factor (simLOF), proposed by Breunig et al. [2], is a well-known anomaly detection algorithm that has been widely used in many fields such as the detection of credit card fraud [11], network of wireless sensor [12] and detection of network intrusion [13]. The method does not label a data instance with normal or anomalous, but gives an anomaly score to each instance. It detects anomaly according to the following assumption: *The density of a normal instance will be similar to that of its neighbors while the density of an anomalous instance will be lower than that of its neighbors.* If the density of a data instance is lower than that of its neighbors, it will get a large simLOF score which indicates a higher possibility of being an anomaly. Steps to calculate the simple local outlier factor are as follows:

1) Calculate the $k-distance$ of each instance $p$. Let $D$ be the dataset. For any positive integer $k$, the $k-distance$ of instance $p$, denoted as $dist_k(p)$, is defined as the distance $d(p,o)$ between $p$ and an instance $o \in D$ such that:
   - At least $k$ instances $o' \in D\backslash\{p\}$ satisfy $d(p,o') \leq d(p,o)$;
   - At most $k$-1 instances $o' \in D\backslash\{p\}$ satisfy that $d(p,o') < d(p,o)$.

   where $d(p,o)$ is the euclidean distance between instance $p$ and instance $o$.
2) Calculate $k-distance\ neighborhood$ of each instance $p$. The $k-distance\ neighborhood$ of $p$ is an instance set that contains every instance whose distance from $p$ is not greater than the $dist_k(p)$, denoted by $N_{k-dist}(p)$.
3) Calculate $reachability\ distance$ of an instance $p$ with regard to instance $o$, which is defined as:

$$reach-dist_k(p,o) = max\{dist_k(o), d(p,o)\} \quad (1)$$

4) Calculate the $local\ reachability\ density$ of each instance $p$, denoted as $ldr_k(p)$. It is defined as

$$lrd_k(p) = \frac{1}{\sum_{o\in N_k(p)} reach-dist_k(p,o)/|N_k(p)|} \quad (2)$$

5) Calculate the $local\ outlier\ factor$, denoted as $LOF_k(p)$

$$LOF_k(p) = \frac{\sum_{o\in N_k(p)} lrd_k(o)/lrd_k(p)}{|N_k(p)|} \quad (3)$$

### D. The Improved Local Outlier Factor

It is the local relative density that decides simLOF, but if the anomalous instances are neighbors of each other, the simLOF score of these instances will not be large. So, we propose an improved local outlier factor (imLOF) which can solve the problem of simLOF.

*1) Limitation Of simLOF:* In some cases, anomalies are neighbors of each other, which contributes to lower simLOF. Figure 2 is an example. Evaluated by simLOF, instance $o_1$, instance $o_2$ and a few instances in cluster $c_2$ are anomalous. The instances in $c_1$ will not get a very large anomaly score, because, as a whole, the density of this local area is low. However, $c_1$ is a cluster containing very few objects, which should also be regarded anomalous and the instances in these clusters should be regarded as anomalies.

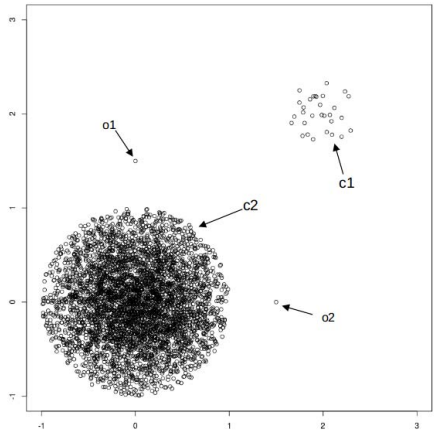



Fig. 2. Limitation of simLOF: It can not detect anomalies which group themselves into small and sparse clusters, such as cluster $c_1$.

This problem happens in medicare datasets. We will take a dataset of hypertension as an example, which is mixed with 200 medicare fraud instances. We first cluster the data using DBSCAN algorithm. Then the simple local outlier factor of each instance is calculated to detect fraud and we take the top 2% as anomalous hospitalizations. The result is shown in the Table I.

TABLE I
AN EXAMPLE OF MEDICARE DATA THAT REVEALS THE LIMITATION OF SIMLOF

| cluster | # of instances | # of fraud | % of fraud | # of fraud by LOF |
|---|---|---|---|---|
| 1 | 4,214 | 22 | 0.52% | 40 |
| 2 | 2,995 | 36 | 1.20% | 68 |
| 3 | 1,798 | 43 | 2.39% | 44 |
| 4 | 1,092 | 32 | 2.93% | 35 |
| 5 | 583 | 24 | 4.11% | 14 |
| 6 | 99 | 19 | 19.19% | 3 |
| 7 | 31 | 13 | 41.94% | 3 |
| 8 | 18 | 7 | 38.9% | 0 |
| 9 | 6 | 3 | 50% | 1 |

As we can see, the percentage of fraud in a small cluster is relatively higher than that in a large cluster, which means that the anomalous hospitalizations group themselves in one cluster. However, the simLOF algorithm tends to ignore this because, as shown in the table, the fraud it detects is rare in small clusters. As we mentioned before, it is the relative density deciding the anomaly score of an instance. But if the neighbors of an anomaly are all anomalies whose density are all low, the relative density of this anomaly will be low. As a result, the simLOF scores of these instances tends to be small, which leads to medicare fraud undetected.

*2) Improved Local Outlier Factor:* Considering the limitation of simLOF, we propose the improved local outlier factor (imLOF), which improves simLOF by using DBSCAN algorithm.

DBSCAN [14] is a density-based clustering algorithm. Density-based means that objects in a dense area will be grouped together. One advantage of DBSCAN is noise immunity. Another is that DBSCAN can find clusters of arbitrary shapesTo cover the shortage of simLOF in detecting self-grouped anomalies, we introduce the size of the cluster an object belonging to into its simLOF score. By taking the size of cluster into consideration, instances in small cluster tend to be assigned with higher anomaly score. Algorithm 1 illustrates this procedure.

### *E. Robust Regression*

The imLOF can detect anomaly by considering the spatial density information, but the linear dependence between the expenditure and the reimbursement is ignored. So we use regression analysis as a complement.

Linear regression is a statistical tool for estimating the relationship between independent variable and dependent variable. Ordinary least square (OLS) is an algorithm to estimate the unknown parameters in a linear regression.

**Algorithm 1** Improved Local Outlier Factor

**Input:** datasets $D$; parameter of simLOF: K-nearest neighbors, $K$; parameters of DBSCAN: An object is a core if within distance $eps$ it has at least $minPts$ neighbors, $minPts$ and $eps$

**Output:** anomaly score of each data instance;

1: Apply DBSCAN to divide $D$ into several clusters $C = \{c_1, c_2, c_3...\}$;
2: **for** each data instance $p$ in $D$ **do**
3: Calculate *reachability distance* of $p$ with regard to its k-nearest neighbors;
4: Calculate *local reachability density* of $p$;
5: Calculate *imLOF* of $p$

$$imLOF = \frac{\sum_{o \in N_k(p)} lrd_k(o)/lrd_k(p)}{|N_k(p)| * log_a|c_n + 1|}, p \in c_n$$

6: **end for**
7: **return** imLOF;

However, data from real world usually contains outliers. Typically in the domain of medicare, when illegal claims happen, the linear relation between expenditure and reimbursement is broken, which produces outliers.

Robust regression [15] is to solve this problem. By assigning a weight to each data instance according to its residual, robust regression updates its parameters iteratively. The weight of an instance is calculated according to its residual in previous iteration. Frequently-used weight functions, mapping the residual of an instance in previous iteration into its weight in current iteration, are Bisquare, Huber and Hampel, defined as following:

$$w_{Bisquare} = I(|r| - 1)(1 - r^2)^2 \tag{4}$$

$$w_{Huber} = \frac{1}{max(1, |r|)} \tag{5}$$

$$w_{Hampel} = e^{-r^2} \tag{6}$$

The function $I(x)$ is a signal function, defined as:

$$I(x) = \begin{cases} 1 & x \leq 0 \\ 0 & x > 0 \end{cases} \tag{7}$$

The value $w$ in these equations is the weight and the value $r$ is calculated according to residual in previous iteration.

Generally speaking, the weights of data instances with large residual in previous iteration will be reduced in the next iteration. The Huber function reduces the weight only if the residual passes an threshold, while the Bisquare function and the Hampel function reduce the weight as long as the residual grows large. The difference between the Bisquare and the Hampel is that the Bisquare function always returns zero if the residual is larger than a threshold.

Excluding the weight function, one iteration of robust regression is same as OLS. While iterating with changing

TABLE II
RESULT OF FIRST EXPERIMENT

| Diseases | imLOF | | | simLOF | | | KBLOF | | | KNN | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Pre. | Recall | F1 | Pre. | Recall | F1 | Pre. | Recall | F1 | Pre. | Recall | F1 |
| Coronary Heart | 0.301 | 0.525 | **0.383** | 0.229 | 0.4 | 0.291 | 0.243 | 0.425 | 0.310 | 0.198 | 0.286 | 0.235 |
| Hypertension | 0.294 | 0.5 | **0.37** | 0.238 | 0.405 | 0.3 | 0.241 | 0.41 | 0.304 | 0.174 | 0.215 | 0.193 |
| Bronchitis | 0.311 | 0.562 | **0.4** | 0.261 | 0.442 | 0.328 | 0.239 | 0.421 | 0.305 | 0.203 | 0.256 | 0.226 |

TABLE III
RESULT OF SECOND EXPERIMENT

| Diseases | OLS | | | RRwithBisquare | | | RRwithHuber | | | RRwithHampel | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Pre. | Recall | F1 | Pre. | Recall | F1 | Pre. | Recall | F1 | Pre. | Recall | F1 |
| Coronary Heart | 0.479 | 0.67 | 0.558 | 0.557 | 0.78 | **0.65** | 0.529 | 0.74 | 0.617 | 0.507 | 0.71 | 0.591 |
| Hypertension | 0.359 | 0.61 | 0.451 | 0.367 | 0.625 | 0.463 | 0.376 | 0.64 | **0.474** | 0.341 | 0.585 | 0.433 |
| Bronchitis | 0.426 | 0.632 | 0.509 | 0.498 | 0.684 | **0.574** | 0.481 | 0.691 | 0.567 | 0.452 | 0.611 | 0.52 |

weights, the effect of outliers to parameter estimating decreases continuously. After converging, the absolute value of the residual of each data instance is regarded as the anomaly score.

## V. EXPERIMENTS

In this section, two groups of comparision experiments are conducted to prove the efficiency of our method. The first one is comparing imLOF with simLOF, kmeans-based LOF and a k-nearest-neighbor based method [16]. The second one is comparing robust regression with different weight functions with OLS. Experiment data contains history data of three diseases in 2013-2015, which are listed in Table IV.

TABLE IV
EXPERIMENT DATA

| Diseases | Claims | HospitalizationDetails | InsuredResidents |
|---|---|---|---|
| Coronary Heart | 9,289 | 73,183 | 6,839 |
| Hypertension | 8,065 | 54,267 | 3,584 |
| Bronchitis | 13,012 | 90,124 | 7,204 |

### A. Performance metrics

Performance metrics for our method are Precision, Recall and F1 score. In this study, Precision is the fraction of detected instances that are fraudulent, while Recall is the fraction of fraudulent instances that are detected. F1 score is a combination of Precision and Recall. In addition, Precision-Recall curve (PR curve), whose x-axis is Recall and y-axis is Precision, is ploted as complement.

### B. Baseline Experiments

This subsection introduces four baseline experiments. To illustrate the efficiency of imLOF, there are two widely-used method in the baseline experiments: simLOF and a k-nearest-neighbor (KNN) based method. In addition, an experiment of a variation of imLOF, kmeans-based LOF (KBLOF), is also conducted, which shows the reason why we choose DBSCAN clustering algorithm. The only difference between kmeans-based local outlier factor (KBLOF) and imLOF is that instead of DBSCAN, kmeans [17] algorithm is applied to cluster the data. The baseline of robust regression is OLS.

### C. Result

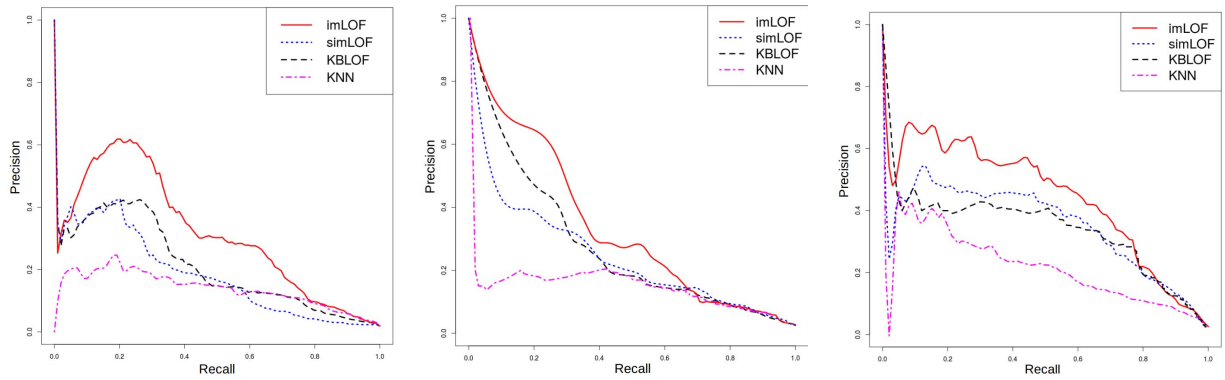
In the first experiment, the datasets are added data of medicare fraud instances. We compare the result of imLOF with that of simLOF, KNN and KBLOF. For comparing Precision, Recall and F1 score, the instances of top 3% large anomaly score are labeled as anomaly. Table II shows these scores. Figure 3 presents the PR curves of the four datasets. The following observations and analysis are drawn from the result:

- Generally speaking, the scores of cluster-based LOF are higher than those of simLOF. This indicates that introducing the size of clusters into simLOF relieves the problem of self-clustered anomaly.
- The scores vary in the different clustering algorithms. The DBSCAN-based LOF performs better than the kmeans-based algorithm in terms of these scores. That is because the kmeans algorithm is sensitive to outliers, while these datasets contain outliers caused by medicare fraud. In addition, the kmeans algorithm has difficulty in dealing with non-convex data.
- The PR curves also indicate that the imLOF algorithm performs better than other methods do.

In the second experiment, the training datasets are the original datasets removing the anomalous instances detected by imLOF. We compare the result of robust regression with different weight functions with that of ordinary least square.

Table III presents Precision, Recall and F1 score of robust regeression and OLS. From this table, we observe that:

- Most robust regression methods can achieve a higher accuracy than OLS does since outliers influence parameter estimating of OLS.
- The scores of robust regression with Bisquare function and with Huber function do not vary that much, which are




(a) PR curve of Coronary Heart dataset (b) PR curve of Hypertension dataset (c) PR curve of Bronchitis dataset

Fig. 3. Comparing imLOF with other methods

both higher than those of robust regression with Hampel function.

## VI. Conclusion

Medicare fraud tends to be a burden of the meidical insurance system of China. In this study, we first summarize the fraud behaviors into three categories. Then a framework which can detect all kinds of these fraud is designed. One important part of this framework is imLOF, an algorithm we propose to detect excessive medical treatment and decomposing hospitalization. Another important part of this framework is robust regression, aiming at detecting illegal claims. Experiments show our method is effective.

This research demonstrates that medicare fraud behaviors can be detected by using novel model of artificial intelligence. In the future work, we intend to seek the help of medicare experts to get more labeled data and apply some supervised model. In addition, because the DBCAN algorithm can not be applied to new-coming data, the imLOF method is not able to run in the online environment. So, it will be very meaningful to adapt the imLOF into an automated method.

## Acknowledgements

This work is supported by the National Key Research and Development Program of China under Grant No. 2016YFB1000904.

## References

[1] E. W. T. Ngai, Yong Hu, Y. H. Wong, Yijun Chen, and Xin Sun. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. 50(3):559–569, 2011.

[2] Markus M Breunig, Hans Peter Kriegel, Raymond T Ng, and Jrg Sander. Lof: identifying density-based local outliers. In *ACM SIGMOD International Conference on Management of Data, May 16-18, 2000, Dallas, Texas, Usa*, pages 93–104, 2000.

[3] Richard A. Bauder and Taghi M. Khoshgoftaar. A novel method for fraudulent medicare claims detection from expected payment deviations (application paper). In *IEEE International Conference on Information Reuse and Integration*, pages 11–19, 2016.

[4] Yao Hsu Tsai, Chieh Heng Ko, and Kuo Chung Lin. Using commonkads method to build prototype system in medical insurance fraud detection. *Journal of Networks*, 9(7), 2014.

[5] Ming Jian Tang, B. Sumudu. U. Mendis, D. Wayne Murray, Yingsong Hu, and Alison Sutinen. Unsupervised fraud detection in medicare australia. In *Australasian Data Mining Conference*, pages 103–110, 2011.

[6] Pedro A. Ortega, Cristin J. Figueroa, and Gonzalo A. Ruz. A medical claim fraud/abuse detection system based on data mining: A case study in chile. In *International Conference on Data Mining, Dmin 2006, Las Vegas, Nevada, Usa, June*, pages 224–231, 2006.

[7] Yuliang Shi, Chenfei Sun, Qingzhong Li, Lizhen Cui, Han Yu, and Chunyan Miao. A fraud resilient medical insurance claim system. In *Thirtieth AAAI Conference on Artificial Intelligence*, pages 4393–4394, 2016.

[8] Zhiping Xie, Xiaoyu Li, Wenyi Wu, and Xiaoling Zhang. *An Improved Outlier Detection Algorithm to Medical Insurance*. Springer International Publishing, 2016.

[9] Shao X. A k-means clustering algorithm. Master's thesis, University of Electronic Science and Technology of China, Sichuan, 2016.

[10] Li X. Research on the classification ensemble algorithm for medical insurance anomaly detection. Master's thesis, University of Electronic Science and Technology of China, Sichuan, 2016.

[11] Mei Chih Chen, Ren Jay Wang, and An Pin Chen. An empirical study for the detection of corporate financial anomaly using outlier mining techniques. In *International Conference on Convergence Information Technology*, pages 612–617, 2007.

[12] Mahsa Salehi, Christopher Leckie, James C Bezdek, and Tharshan Vaithianathan. Local outlier detection for data streams in sensor networks: Revisiting the utility problem invited paper. In *Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2015 IEEE Tenth International Conference on*, pages 1–6. IEEE, 2015.

[13] Nerijus Paulauskas and kAvzuolas Faustas Bagdonas. Local outlier factor use for the network flow anomaly detection. *Security and Communication Networks*, 8(18):4203–4212, 2015.

[14] Martin Ester, Hans Peter Kriegel, Jrg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *International Conference Knowledge Discovery and Data Mining*, pages 226–231, 1996.

[15] Peter J. Rousseeuw and Annick M. Leroy. Robust regression and outlier detection. *Journal of the American Statistical Association*, 31(2):260–261, 1987.

[16] Simon Byers and Adrian E. Raftery. Nearest-neighbor clutter removal for estimating features in spatial point processes. *Journal of the American Statistical Association*, 93(442):577–584, 1998.

[17] J. A. Hartigan and M. A. Wong. A k-means clustering algorithm. *Applied Statistics*, 28(1):100–108, 1979.