Journal of Electronics and Information Science (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/jeis.2024.090307
ISSN 2371-9524 Vol. 9 Num. 3

Research on medical insurance fraud identification method based on multi-source datasets

Jinhong Zhang[1]

[1]*School of Economics and Management, Jinzhong College of Information, Jinzhong, Shanxi, 030800, China*
810042948@qq.com

Keywords: Insurance fraud; Active learning; Logistic Regression

Abstract: Due to the fact that datasets pertaining to health insurance are typically stored in different departments and involve multiple databases, conventional data analysis and traditional fraud research methods often fall short in accurately identifying fraudulent activities. To address this challenge, this paper, on one hand, starts by recognizing the unique characteristics of various collected datasets. It employs a multi-source data fusion approach, initially combining their features. Subsequently, an exploratory data analysis is conducted on the fused dataset. Compared to previous single datasets, the merged dataset contains more features, significantly enhancing the model's fitting performance. This approach maximizes the utilization of information within the data, allowing for better exploitation of the data's potential.On the other hand, this paper integrates the strategy of active learning with traditional logistic regression methods, constructing a novel model. The model is initially trained on labeled datasets, and after multiple experiments, it was observed that the fitting accuracy of the active learning model, constructed using the BT strategy (a type of active learning sample extraction strategy), surpassed that of a standalone logistic regression model. This innovative approach provides a new avenue for improving the accuracy of health insurance fraud detection.

1. Introduction

The social medical insurance system has been playing a positive role in safeguarding the health rights and alleviating poverty caused by illness for the general public. It is also an important channel for urban and rural residents to access quality services in the event of accidents or major illnesses.[1] However, in recent years, cases of "health insurance fraud" have been on the rise, and instances of insurance fraud have become increasingly covert. Traditional insurance fraud detection techniques mostly rely on statistical identification methods, where regression models are established based on given claim cases to define key factors for fraud identification. The corresponding weights for these factors are determined through statistical analysis to identify fraud. Currently, machine learning methods have gradually been applied in the field of insurance fraud. When faced with unlabeled data, unsupervised machine learning correlation analysis methods are typically used to discover connections between fraudulent behaviors and explore new types of fraudulent activities within other groups.[2] However, when dealing with high-dimensional datasets, training costs for

unsupervised learning methods are often prohibitively high. Additionally, fraud data is usually imbalanced, making it challenging for unsupervised learning methods to effectively separate fraud and non-fraud samples. While supervised learning methods in machine learning can effectively improve the accuracy of fraud detection, they often require a large number of annotated samples, leading to expensive labeling costs in practical operations. The data we can obtain is typically severely lacking in labels. Based on the above analysis, solely relying on machine learning models for fraud detection is not feasible.

2. Literature Review

Considering the differences in the development history of medical insurance both domestically and internationally, as well as the existing variations in the medical insurance systems, this paper will conduct a literature review from two aspects: domestic research and international research.

In China, scholars began studying the issue of social medical insurance fraud in the 1990s, with one of the prominent figures being Li Lianyou et al. Their research analyzed the inducement of demand and moral damage from the supply side (i.e., medical institutions) and emphasized the need to establish effective constraint mechanisms to counteract medical fraud by the supply side. While this represented an early discussion in the academic community, it was limited to the study of fraud by suppliers. Fraud by healthcare service providers manifests in various forms such as issuing false bills and medical records, excessive services, and splitting service procedures. As China's social medical insurance system undergoes deep reforms, it has exposed several loopholes that have yet to be promptly addressed at the institutional level. Recognizing the serious ethical implications of fraudulent activities, the academic community has given considerable attention to this issue.Currently, theoretical research on medical insurance claims mainly involves three main entities: the insured individuals (insured persons or their agents, employers), healthcare service providers (doctors, hospitals, etc.), and insurance bearers (insurance companies or government agencies).[3] Fraud by insured individuals may involve requesting hospitals to issue unnecessary diagnostic and treatment items or claiming reimbursement for medications on behalf of others. Other instances include uninsured individuals pretending to be insured persons to claim expenses, separate billing, and collusion between healthcare providers and insured individuals.[4] The latter often involves fraudulent practices like "admission for bed occupancy", issuing prescriptions for personal favors and nutritional supplements, among others. Traditional fraud identification methods relying on expert investigations can be prohibitively expensive. Additionally, with the explosive growth of medical insurance data, traditional fraud detection methods are no longer able to meet the needs of current fraud case screening. With the advent of the big data era, machine learning methods, both supervised and unsupervised, have seen rapid development and widespread application in fraud detection.[5] Common classification methods include decision trees, Bayesian networkssupport vector machines, logistic regression XGBoost neural networks, and common unsupervised methods include anomaly detection and clustering. Research on anti-fraud efforts mainly focuses on the causes of fraud and corresponding recommendations, the soundness of systems and legal regulations, among other aspects.

3. Problem Description

For medical insurance fraud data, the cost of investigation is prohibitively high.[6] Faced with vast medical data, it is impractical to manually label each piece of data. This not only increases the expenses associated with medical insurance fraud but also places significant pressure on experts in the field. Effectively selecting the most valuable data for investigation is the most direct way to reduce manual costs. This paper introduces an active learning strategy that combines unlabeled and

labeled data for learning, automatically annotating unlabeled data to reduce the cost of manual labeling. The goal of this strategy is to intelligently select data for annotation, enhancing model performance and reducing labeling costs while working with a limited set of labeled samples.

4. Active Learning and Logistic Regression Model

The Concept of Active Learning

The concept of Active Learning involves the following idea: if the cost of acquiring labels is extremely high, one should seek samples for labeling that would most effectively improve the current algorithm's classification ability, achieving significant results with less effort.[7] This method assumes multiple rounds of active interaction between an active learner and an expert.

The active learning labeling strategy employed in this paper is based on the maximum entropy method for data annotation. Information entropy is used to assess the stable state of an event by measuring uncertainty.[8] The larger the entropy value of information entropy, the more uncertain (or unstable) the event (or state); conversely, the smaller the entropy value, the more certain (or stable) the event (or state). The formula for entropy calculation is:

$$H(Z) = -\sum_{i=1}^{n} p_i \log p_i \tag{1}$$

With the entropy calculation for a single variable Z in place, it becomes straightforward to derive the joint entropy of multiple variables. For two random variables Z and Y, the formula for joint entropy is:

$$H(Z,Y) = -\sum_{i=1}^{n} p(z_i, y_i)\, log\, p\,(z_i, y_i) \tag{2}$$

Maximum Entropy Selection Strategy (MEs): In this paper, we use conditional entropy to describe the confidence level of a sample belonging to a particular class. If the conditional entropy value is larger, it indicates that the classification of a certain sample point is more uncertain (confidence in classification is lower). [9]Conversely, if the conditional entropy value is smaller, it indicates that the classification of a certain sample point is more certain (confidence in classification is higher). The conditional entropy is calculated using the following formula:

$$H(Y|Z) = H(Z,Y) - H(Z) \tag{3}$$

The conditional entropy H(Y | Z) can be calculated using the following formula:

$$\begin{aligned} & H(Z,Y) - H(Z) \\ & = -\sum_{z,y} p(z,y) log p(z,y) + \sum_{z,y} p(z) log v(z) \\ & = -\sum_{z,y} p(z,y) log p(z,y) + \sum_{z} (\sum_{y} p(z,y)) log z(z) \\ & = -\sum_{z,y} p(z,y) log p(z,y) + \sum_{z} p(z,y) log g(z) \\ & = -\sum_{z,y} p(z,y) \log \frac{p(z,y)}{p(z)} \\ & = -\sum_{z,y} p(z,y) log p(y|z) \\ & = -\sum_{z,y} p(z) p(y \mid z) \log p(y \mid z) \end{aligned} \tag{4}$$

Based on the above formula, the uncertainty of each sample's classification can be computed. Use $H(C) = \{H(y_i|z_i)|i = 0,1,2,3 \ldots n\}$o represent the classification uncertainty of all samples. By sorting the set H(C) and identifying the sample with the highest numerical value, it is selected for manual labeling.

Logistic Regression is a commonly used machine learning method that is employed to address binary classification problems (0 or 1). It is also utilized to estimate the likelihood of an event occurring.[10] Examples of such events include the probability of a mall user purchasing a specific

product, the likelihood of diagnosing a particular disease, and the probability of a user clicking on an advertisement on a website. [11]It is important to note that in this context, "likelihood" refers to the occurrence probability and not the mathematical concept of "probability." The results obtained from logistic regression are not mathematical probabilities and cannot be directly interpreted as such. Instead, these results are often used in weighted sums with other feature values rather than being directly multiplied.

Logistic Regression is a type of Generalized Linear Model (GLM)[12], and linear regression is also a type of Generalized Linear Model. Logistic Regression assumes that the dependent variable y follows a Bernoulli distribution, while linear regression assumes that the dependent variable y follows a Gaussian distribution.

The general form of the logistic regression model is as follows (Allison, 2001):

$$P(Y = 1 \mid X_1, \ldots, X_n) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\cdots+\beta_N X_N)}} \tag{5}$$

5. Data Processing and Exploratory Data Analysis

The data used in this paper comes from the Healthcare Provider Fraud Analysis dataset provided by Kaggle (verified by experts as suitable for insurance fraud detection research). Due to the characteristics of the data, this paper primarily analyzes fraudulent activities from the perspective of different hospitals. The original data consists of eight different CSV files. Four of these files belong to datasets labeled with potential fraudulent providers: training beneficiaries, training outpatient patients, training inpatient patients, and labeled training providers. The remaining four files belong to unlabeled datasets (not labeled as potential fraudulent providers): test beneficiaries, test outpatient patients, test inpatient patients, and test providers. The labeled data includes a total of 558,211 beneficiaries and 5,410 providers, while the unlabeled data includes a total of 135,392 beneficiaries and 1,353 providers.

We first merge the datasets of the four training sets: training beneficiaries' individual information, training inpatient beneficiaries' data, and training outpatient beneficiaries' data. A hierarchical fusion approach is adopted, based on feature concatenation. Initially, the data of inpatient beneficiaries and outpatient beneficiaries are merged. The inpatient beneficiaries' dataset has 30 different features, while the outpatient beneficiaries' dataset has 27 different features. The difference lies in the presence of features such as AdmissionDt (admission date), DeductibleAmtPaid (inpatient deductible amount), and ClmAdmitDiagnosisCode (diagnosis code at admission) in the inpatient beneficiaries' dataset. Since the inpatient dataset has 40,474 observations and the outpatient dataset has 517,737 observations, to maximize the utilization of information in the samples, we merge these two datasets based on the features of the inpatient dataset. For the features such as AdmissionDt in the outpatient dataset, we handle them by assigning zero values. Next, we merge the beneficiaries' personal information with the above data. The merged data has 558,211 observations and 55 features. On this basis, we process the features of this data. Based on whether the "DOD" (date of death) variable in the dataset is empty, we add a variable "IsDead." If the value of DOD is non-empty, the variable "IsDead" is assigned a value of 1; otherwise, it is assigned a value of 0. Then, subtracting the date of visit from the patient's date of birth, we obtain the patient's age variable (age), and convert its numerical unit to years. We create a variable "NumPhysicians," which represents the number of physicians and is the sum of Attending Physician, Operating Physician, and Other Physician. Finally, we create a variable "AdmissionDays," whose value is the difference between DischargeDt (discharge date) and AdmissionDt (admission date). Lastly, we check whether the processed dataset has missing values. For variables with missing values, we assign them a value of zero. We then remove object variables such as AdmissionDt and

DischargeDt, retaining only numerical variables. Subsequently, based on the provider indicator, we sum up or take the average of patient data (i.e., summing or averaging the data of all patients within the same healthcare provider). After these processing steps, the resulting dataset has 5,410 observations and 32 features.

6. Model Training

We use the Logistic Regression analysis model as our baseline model. Its basic definition is as follows: Logistic Regression is a statistical model used to determine whether independent variables have an impact on a binary dependent variable.[13] This means that there are only two potential outcomes under consideration. For the well-processed dataset, we decide to use stratified sampling as the method to divide the test set and training set. Stratified sampling is defined as a method of randomly sampling individuals from different strata (or layers) of a population that can be divided into different categories according to specified proportions. The advantages of this method are better representativeness of the sample and smaller sampling errors.[14] Moreover, our dataset is imbalanced, with non-fraudulent cases accounting for around 90% and fraudulent cases accounting for around 10%. If a random sampling strategy is adopted, the sampling error caused will be larger compared to a balanced dataset.

We stratify the data according to the fraud label in a 70%, 30% ratio for training and testing sets. We train a Logistic model on the training set as our baseline model, achieving an accuracy of 92.4%. We then employ the strategy of the active learning framework for training, with the following specific steps:

1) Determine the sample size (batch size) and the number of sampling rounds (rounds) for each sampling.

2) Randomly select a small portion of samples from the entire training set.

3) Stratify this small portion of samples to create a training set and a test set for training. Obtain an accuracy score on the test set.

4) Use the trained model to fit the data not selected in the entire dataset, obtaining accuracy scores for the remaining data. Then, extract data from the remaining data according to three different sampling strategies.[15]

5) Merge the newly extracted data with the previously extracted data, and repeat steps 3-5 until the set number of sampling rounds is reached.[16]

In this experiment, we set the sample size for each extraction to be 500, and the number of extractions to be 10, then proceed with training. The accuracy results of the training are shown in the following Table 1:

Table 1: Training Accuracy

rounds	1	2	3	4	5	6	7	8	9	10
RS Strategy	93.3%	86.3%	86.2%	82.1%	79.9%	76.3%	75.5%	76.8%	75.7%	73.4%
LC Strategy	93.3%	83.3%	81.3%	78.0%	78.0%	78.6%	74.8%	75.6%	76.7%	75.3%
BT Strategy	89.3%	92.3%	93.3%	93.1%	93.2%	92.7%	92.6%	92.8%	93.0%	94.1%

RS Strategy: Randomly selects samples for training in each round, serving as a control group for the other two strategies.

LC Strategy: Identifies the least confident predictions of the trained classifier, i.e., samples with low probabilities for the most probable class.

BT Strategy: Targets samples where the trained classifier is "on the fence," i.e., samples with closely matched probabilities for the two predicted classes.

Observing the results, the RS strategy exhibits lower final accuracy compared to the other strategies. While initially showing relatively better accuracy at 93.3% (baseline classifier accuracy is 92.7%), its accuracy decreases over the active learning process, indicating lower robustness and generalization capability, resulting in a final accuracy of 73.4%. The LC strategy also performs relatively poorly on this dataset, with a final accuracy of 75.3%, though higher than the random sampling's 73.4%. The BT strategy performs the best on this dataset, starting with an accuracy of 89.3%, reaching around 92%, and ultimately achieving a final accuracy of 94.1%, surpassing the baseline classifier's accuracy of 92.7%. Therefore, we decide to adopt the BT strategy as our sample selection strategy. For the four unlabeled test sets, we perform the same merging operations as with the training set and use our established model to annotate them through the BT strategy. This approach significantly reduces manual and time costs, greatly enhancing efficiency.

7. Conclusion

In view of the high cost of labeling medical insurance fraud data sets and the relatively small number of public insurance fraud data sets, this paper develops an active learning strategy combined with a logistic regression model. Compared with traditional supervised learning and unsupervised learning methods, active learning methods can automatically label data. This paper innovatively integrates the collected data sets. Most previous studies only used a single data set. The features of a single data set are fewer than those of multiple data sets, which may lead to a lower fit of the model. This paper fuses multiple collected data sets and obtains relatively more features. Then calculations are performed on the basis of the original features to obtain new features. Then we manually screened the features, combined with our own knowledge about medical insurance, and then asked relevant professionals, and finally selected the features we needed. We significantly improved the goodness of fit of the model by using the active learning method of the BT strategy. A model built using multi-source data sets can also improve the generalization ability of the model, because the model is equivalent to validating multiple models to a certain extent. Compared with the model trained with a single dataset, the prediction accuracy is relatively high, higher than the traditional logistic regression model, and this strategy can be used to label unlabeled data.

References

[1] Wang S L, Pai H T, Wu M F, et al. The evaluation of trustworthiness to identify health insurance fraud in dentistry[J]. Artificial intelligence in medicine, 2017, 75: 40-50.

[2] Thornton D. Brinkhuis M, Amrit C, et al. Categorizing and Describing the Types of Fraud in Healthcare [J]. Procedia Comput Science, 2015, 64(1): 713-720.

[3] Faseela V S, Thangam D P. A Review on Health Insurance Claim Fraud Detection[J]. International Journal of Engineering Research Science (IJOER), 2015, (4):47-49.

[4] Yip W, Hsiao W C. What Drove the Cycles of Chi- nese Health System Reforms[J]. Health System Re form, 2015, 1(1): 52-61.

[5] Yu, H. Universal Health Insurance Coverage for 1. 3 Billion[J]. Health Poli- People: What Accounts for China's Successcy, 2015, 119(9): 1145-1152.

[6] LIU J, BIER E, WILSON A, et al. Graph analysis for detecting fraud, waste, and abuse in healthcare data[C]// Proceedings of the 27th Conference on Innovative Applications of Artificial Intelli-gence. Palo Alto, CA: AAAI Press, 2015:3912-3919.

[7] Wilhelm W K. The Fraud Management Lifecycle Theory: A Holistic Approach to Fraud Management[J]. Journal of Eco- nomic Crime Management, 2004, 2(2): 1-38.

[8] Faseela V S, Thangam P. A Review on Health Insurance Claim Fraud Detection[J]. International Journal of Engineering Research Science, 2015, 1(1):47-49.

[9] Hubick, K. T.. Artificial neural networks in Australia. Canberra: Common wealth of Australia, 1992.
[10] Stijn Viaene, Richard A. Derrig, Guido Dedene. A Case Study of Applying Boosting Naive Bayes to Claim Fraud Diagnosis. Ieee Transaction on Knowledge and Data Engineering, 2004: Vo1. 18, NO. 5
[11] Fletcher Lu, J. Efrim Boritz. Detecting fraud in health insurance data: learning to model incomplete benford's law distributions. Lecture notes in computer science, 2005(3720):633-640.
[12] Jing Li, Kuei-Ying Huang, Jionghua Jin, Jianjun Shi. A Survey on Statistical 51 Methods for Health Care Fraud Detection. Health Care Manage Science, 2007, (5):1-21.
[13] James H. Bisker, Kate Ehrlich. Health Insurance Fraud Detection Using Socia Network Analytics, United States Patent Application Publication. 2008:US 2008/0172257
[14] Ekina T, Leva F, Ruggeri F, et al. Application of bayesian methods in detection of healthcare frad[J]. chemical engineering Transaction, 2013, 33-42.
[15] Sadiq S, Tao Y, Yan Y, et al. Mining anomalies in medicare big data using patient rule inductionmethod[C]2017 IEEE third international conference on multimedia Big Data (BigMM). IEEE, 2017:185-192.
[16] Chuishi Meng, Xiuwen Yi, Lu Su, Jing Gao, and Yu Zheng. 2017. City-wide Traffic Volume Inference with Loop Detector Data and Taxi Trajectories. In Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '17). Association for Computing Machinery, New York, NY, USA, Article 1, 1–10.