Medical and Health Data Classification Method Based on Machine Learning

Yu Zeng and Fuchao Cheng

College of Computer Science, Chengdu University, Chengdu 610106, China

Correspondence should be addressed to Yu Zeng; zengyu@cdu.edu.cn

Received 16 October 2021; Revised 2 November 2021; Accepted 5 November 2021; Published 16 November 2021

Academic Editor: Kalidoss Rajakani

Copyright © 2021 Yu Zeng and Fuchao Cheng. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The information defined in medical health data is researched based on machine learning-related algorithms. Also, this paper used random forest and other related algorithms to perform health data training and fitting. Research shows that the algorithm proposed in the paper can improve the progress of health data classification. The algorithm can provide technical support for the improvement of medical data classification.

1. Introduction

With the development of China’s medical industry, the medical market has become more and more complicated. Establishing a sound medical credit system is one of the important means to regulate the medical market. The lack of standard measures for participants in China’s medical market has led to frequent breaches of trust, such as registration breaches [1]. This seriously wastes limited medical resources. This article studies the dishonesty behaviour in medical treatment. The purpose of the research is to increase the medical industry’s management of market participants and improve the market access threshold and management level. In recent years, the development of computer technology has made great progress in data mining and machine learning technology. Supporting massive amounts of data and using machine learning algorithms to utilize the data effectively can enhance the value of data. When the scenario of the algorithm is in the medical field, the relevant historical behaviour data of medical market participants can be used to predict whether there is a risk of dishonesty. This assists medical market managers in making decisions.

This paper studies the decision tree and random forest algorithm. Since the random forest algorithm is based on the ensemble learning idea, it effectively avoids noise in the training data set, so there will be no overfitting phenomenon. The simulation results also show that this method performs better than logistic regression and K-nearest algorithm in identifying dishonest behaviours [2]. This has important reference significance for the establishment of the medical credit system.

2. Theoretical Basis

2.1. Decision Tree Algorithm. In recent years, decision trees have been one of the most widely used algorithms in machine learning. Compared with the neural network algorithm decision tree, it has the characteristics of flexibility and strong interpretability. Flexibility embodied in the decision tree can prune the tree structure according to the wishes of the algorithm designer and improve the algorithm’s performance. Interpretability is embodied in the decision-making standard with extremely high confidence when each root node of the decision tree makes a decision. The decision tree is composed of three parts: directed edges of internal nodes and leaf nodes. The basic process is shown in Figure 1.

It can be seen from Figure 1 that a classification tree can divide a data set into different classifications $C_i$ using different feature dimensions $A$. When the classification tree is classified, different classification tree algorithms are based on different node classification standards [3]. Figure 1 is
based on information gain. The method of defining information gain is as follows.

First, define information entropy. The definition of information entropy of data set $D$ is as follows:

$$\text{Info}(D) = -\sum_{i=1}^{T} \frac{|C_i|}{|D|} \log_2 \left( \frac{|C_i|}{|D|} \right).$$

(1)

Different features have conditional information entropy $\text{Info}_A(D)$ for $D$:

$$\text{Info}_A(\{|D|\}) = -\sum_{k=1}^{K} \frac{|C_k|}{|D|} \times \text{Info}(D_k)$$

(2)

$$= -\sum_{k=1}^{K} \frac{|C_k|}{|D|} \sum_{i=1}^{T} \frac{|C_{ki}|}{|D|} \log_2 \left( \frac{|C_{ki}|}{|D|} \right).$$

The information gain of the feature at this time is

$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(\{|D|\}).$$

(3)

Academically, different types of decision trees use different node decision criteria. This article also uses the Gini coefficient as the decision criterion [4]. The decision tree at this time is called CART. For multiclassification problems, when there are $K$ different categories, mark $p_k$ as the probability that the current sample is category $k$. At this time, the Gini coefficient can be defined according to the probability distribution:

$$\text{Gini}(p) = \sum_{k=1}^{K} p_k (1 - p_k).$$

(4)

When there is a sample set $D$, the Gini coefficient of the sample can be written as

$$\text{Gini}(D) = 1 - \sum_{k=1}^{K} \left( \frac{|C_k|}{|D|} \right)^2.$$

(5)

The sample set $D$ is divided into different subsets $D_1$ and $D_2$ according to feature $A$:

$$D_1 = \{(x, y) \in D | A(x) = a\},$$

$$D_2 = D - D_1.$$  

(6)

The Gini coefficient divided according to feature $A$ is

$$\text{Gini}(D, A) = \frac{|D_1|}{|D|} \text{Gini}(D_1) + \frac{|D_2|}{|D|} \text{Gini}(D_2).$$

(7)

### 2.2 Integrated Learning and Random Forest

In actual engineering applications, the ability of a single decision tree is limited. The noise and outliers in the training data will cause overfitting of the decision tree, which will seriously affect the accuracy of the decision tree to classify unknown data. Therefore, pruning operations are required after the decision tree is generated. To avoid overfitting, random forest (RF) can also be used [5]. The establishment of the random forest depends on the guidance of integrated learning thought (bagging). Bagging is characterized by random sampling of samples and classifiers. It includes two steps: selecting samples for model training and classifying based on classifiers. The process is shown in Figure 2.

![Decision Tree Algorithm Flow](image)

**Figure 1: Decision tree algorithm flow.**
Ensembling the following formula: bedenotedasvar(mr),anditsupperboundcanbegivenby
todefinetheintervalfunctionoftherandomforest:
classificationperformanceoftherandomforest,weneed
differentclassifiersgetastrongclassifier[7].Tomeasure
classifierscanbeobtained.\[P\]hen,assignthesameweightto
tobe pruned. After the training is completed,
Duringthetrainingprocess,thedecisiontreedoesnotneed
to be pruned. After the training is completed, $k$ training setscantrain $k$ decision trees.

$k$ tree is generated. $K$ training setscantrain
increased, and the correlation between each tree is reduced.

randomforesthasanupperlimit. $fP$hisupperlimitcanbe
to be pruned. After the training is completed, $k$ training sets can train $k$ decision trees.

Duringthetrainingprocess,thedecisiontreedoesnotneed
prediction, the decision tree does not need to be pruned. After the training is completed, $k$ weak
classifiers can be obtained. Then, assign the same weight to
different classifiers to get a strong classifier [7]. To measure
classification performance of the random forest, we need
to define the interval function of the random forest:
For the classifier \{h_1 (x), h_2 (x), \ldots, h_N (x)\}, there isa sample set \{X, Y\}. Its distribution is $X, Y$:
\[mg (x, y) = av_k I (h_k (x) = y) − \max_{j \neq k} av_j I (h_j (x) = j).\]  
(9)

The interval function of the random forest algorithm can be
denoted as var(mr), and its upper bound can be given by
the following formula:
\[\text{var(mr)} \leq \frac{P (1 - s^2)}{s^2}.\]  
(10)

$s$ represents the classification strength of a single decision
tree, and $P$ represents the correlation between decision trees.
The above formula shows that the interval function of the
random forest has an upper limit. This upper limit can be
lowered when the strength of a single classification tree
is increased, and the correlation between each tree is reduced.

3. Method Implementation

3.1. Data Input. This article investigates the three parties of
patients, hospitals, and medical companies. Then, from the
patients’ perspective, random forest algorithms identify
dishonest behaviours to build a healthy medical platform.
First of all, this article obtained data sets related to citizen
credit from public data sets [8]. In the random forest
training, because the privacy of medical patients is involved,
the residents’ credit data can only be obtained from the
foreign platform Lending Club. Then, add relevant medical
record information to each data to ensure that the data are
suitable for the application scenarios required in this article.

After the data set is collected, the data are preprocessed
according to the data preprocessing process shown in Figure 3 [9].

When dealing with outliers, mainly eliminate data items
that are seriously inconsistent with logic. The method used
in the article is Turkey’s algorithm. This method can define
the data of 1.5 times the 4th quartile range as outliers
according to the distribution characteristics of the data and
eliminate them. The data normalization process uses the
following formula:
\[
\frac{(X – \text{min}(X))}{\text{max}(X) – \text{min}(X)}
\]  
(11)

After the data are normalized, the value of the data itself
will not cause an offset to the evaluation result, and the
evaluation result only depends on the influence of the data
attribute. The final step of preprocessing is to perform
correlation analysis on the data. We eliminate the more
relevant attributes in the data to reduce the dimensionality
of the input data under the premise of ensuring the classifi-
cation accuracy, thereby improving the efficiency of model
training and classification.

3.2. Algorithm Training and Testing. In the random forest, it
is necessary to reasonably set the relevant parameters of
the random forest algorithm according to the feature
vector dimension and the data dimension. This paper
analyzes the error of random forest under different pa-
rameters [10]. The error analysis results are shown in
Figures 4 and 5.

Figures 4 and 5, respectively, show the impact of dif-
ferent ntry and mtry on the model’s accuracy during
training. For random forests, the value of ntry should be large
enough to ensure that the model can converge during the
training process. Figure 4 shows the model error under
different ntry when the default mtry is $mtry = \sqrt{M}$. It
can be seen that when ntry reaches 1000, the model error
drops to a stable level. Figure 5 shows the effect of different
mtry on model accuracy when ntry = 1000. It can be seen
that when mtry < 6, the model error increases as mtry in-
creases. When mtry > 6, the model error increases as mtry
increases. Therefore, the optimal value of mtry is 6.
In addition to mtry and ntry, the important parameters of random forests also include classes. In the subsequent model training and testing, the value of each parameter is shown in Table 1.

After setting the parameters of the model, we trained and tested the model. To better identify the efficiency of the batch model in the medical dishonesty behavior, this paper uses logistic regression (LR) and $k$-neighbor algorithm ($k$-NN).
for comparison. The recognition accuracy of each algorithm is shown in Table 2.

In Table 2, the actual rate is the rate predicted by the positive model for the sample. The true negative rate is the negative rate but is predicted to be positive by the model. It can be seen from Table 2 that the actual rate of the RF algorithm is 12.9%, which is an improvement over k-NNLR. The accuracy rate is 1.4% and 1.3% higher than that of k-NN and LR, respectively. The true negative rate has dropped to a certain extent. It can be seen that the RF algorithm has better performance when recognizing unpredictable medical behaviours.

4. Conclusion

This article is based on the idea of machine learning and data mining to research medical and health data. The article establishes a prevention and monitoring model based on the random forest algorithm. This article focuses on the input features used in related medical models. In addition to combining the historical medical information of medical participants, we also introduce the patient’s social credit status, which can effectively compensate for the behavior identification of medical record personnel and the prevention of dishonesty. The random forest algorithm used in this article can avoid the overfitting phenomenon in the training process and improve the prediction accuracy. The content of this article has certain practical significance for the behavioral norms of medical market participants.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Z. Yin, L. M. Sulieman, and B. A. Malin, “A systematic literature review of machine learning in online personal health data,” *Journal of the American Medical Informatics Association*, vol. 26, no. 6, pp. 561–576, 2019.

[2] G. Silahtaroğlu and N. Yılmaztürk, “Data analysis in health and big data: a machine learning medical diagnosis model based on patients’ complaints,” *Communications in Statistics - Theory and Methods*, vol. 50, no. 7, pp. 1547–1556, 2021.

[3] G. A. Kaisiss, M. R. Makowski, D. Rückert, and R. F. Braren, “Secure, privacy-preserving and federated machine learning in medical imaging,” *Nature Machine Intelligence*, vol. 2, no. 6, pp. 305–311, 2020.

[4] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, “Comparing different supervised machine learning algorithms for disease prediction,” *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, pp. 281–316, 2019.

[5] M. D. McCradden, S. Joshi, M. Mazwi, and J. A. Anderson, “Ethical limitations of algorithmic fair solutions in health care machine learning,” *The Lancet Digital Health*, vol. 2, no. 5, pp. e221–e223, 2020.

[6] A. L. Beam, A. K. Manrai, and M. Ghassemi, “Challenges to the reproducibility of machine learning models in health care,” *Jama*, vol. 323, no. 4, pp. 305-306, 2020.

[7] A. Vellido, “The importance of interpretability and visualization in machine learning for applications in medicine and health care,” *Neural Computing and Applications*, vol. 32, no. 24, pp. 18069–18083, 2020.

[8] C. K. Fisher, A. M. Smith, and J. R. Walsh, “Machine learning for comprehensive forecasting of Alzheimer’s Disease progression,” *Scientific Reports*, vol. 9, no. 1, pp. 1–14, 2019.

[9] J. G. Richens, C. M. Lee, and S. Johri, “Improving the accuracy of medical diagnosis with causal machine learning,” *Nature Communications*, vol. 11, no. 1, pp. 3923–3929, 2020.

[10] V. Mhasawade, Y. Zhao, and R. Chunara, “Machine learning and algorithmic fairness in public and population health,” *Nature Machine Intelligence*, vol. 3, no. 8, pp. 659–666, 2021.