Design and Development of Financial Fraud Audit System Based on Big Data Technology

Binglan Meng
Dalian University of Finance and Economics, Dalian City, Liaoning Province, 116600, China

Keywords: Big Data Technology, Data Mining, Audit of Financial Fraud, Selection of Features, Classifier Model.

Abstract: Based on the combination of big data technology and financial fraud audit, Hadoop framework, Relief algorithm under data mining technology, Logistic, SVM and Random Forest classifier are combined to complete the sample data feature acquisition and financial fraud identification model construction, and the financial fraud audit system is packaged and published in Python language environment. The system is presented in the form of Web, which is convenient for auditors to query all kinds of financial data or non-financial data, identify financial fraud and assess the risk of financial fraud through simple and convenient operation. It provides comprehensive application solutions for the problems of complexity, concealment, difficulty and risk in the audit of financial fraud in the data age.

1 INTRODUCTION

At present, under the background of digital economy era, the business operation mode of enterprises is becoming more and more complex with the empowerment of the new generation of digital information technology. As a result, the means and forms of financial fraud have also changed, showing the characteristics of diversification, complexity and concealment. (Jiao, 2021) At the same time, the traditional audit procedures and means have gradually fallen behind, and it is more and more difficult to complete the audit of many types of financial data by relying solely on auditors' personal ability and work experience. In addition, the change of network and digital environment has prompted a qualitative leap in the quantity and dimension of enterprise data information, resulting in increasing risk of financial fraud audit failure. For this reason, this paper believes that taking big data technology as the core, Hadoop framework as the foundation, using HDFS, HBase and other distributed storage frameworks to capture, clean and store all kinds of data information in enterprises, combining with data mining technology, Relief algorithm, Logistic, SVM and RandomForest classifier to complete the selection of sample data characteristics and the construction of financial fraud identification model, and to complete the design and development of financial fraud audit system in Python environment. The system is convenient for internal auditors of enterprises to complete the whole process of financial fraud audit through simple and efficient Web application operation, which is not only conducive to the innovation of the working mode and method of financial fraud audit, but also greatly improves the working efficiency of auditors.

2 OVERVIEW OF KEY TECHNOLOGIES

2.1 Big Data Technology

The big data (mega data) can be called huge data, which is a kind of data collection whose scale is so large that its acquisition, storage, management and analysis greatly exceed the capability of traditional database software tools. (Zhao, 2022) The embodiment of the value of big data depends on big data processing technology, that is, big data technology.

The Hadoop is an open source framework written by Java language, which stores massive data on distributed server clusters and runs distributed analysis applications. (Shi, 2021) Hadoop has quickly become the most popular and powerful big data tool with its application advantages of high reliability, high scalability, high fault tolerance and high effi-

Meng, B.
Design and Development of Financial Fraud Audit System Based on Big Data Technology.
DOI: 10.5220/0011737600003607
In *Proceedings of the 1st International Conference on Public Management, Digital Economy and Internet Technology (ICPDI 2022)*, pages 378-381
ISBN: 978-989-758-620-0

ciency. The core of Hadoop architecture is distributed file system (HDFS), distributed computing programming framework (MapReduce) and resource distributed scheduling framework (YARN).

HBase is a distributed database with column storage, but HBase itself is not directly involved in file storage, and its actual functions are still realized by HDFS under Hadoop framework. The design core of HBase is to realize random and real-time read/write access of HDFS system.

2.2 Data Mining Technology

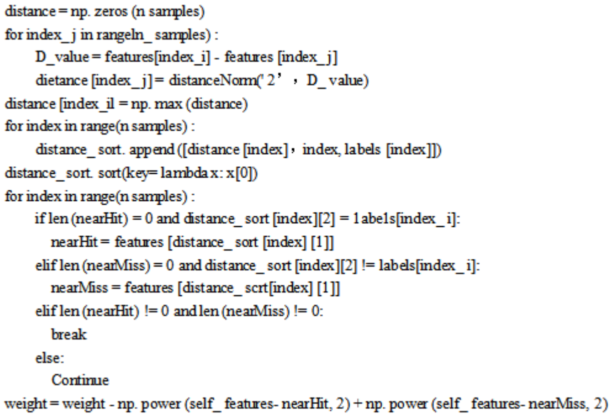
As a kind of computer science and technology, data mining is a processing method for big data, which aims to extract information and knowledge that people don't know in advance but have potential usefulness from a large number of, incomplete, noisy, fuzzy and random actual data. (Wang, 2021) The construction of data mining model is the core of the whole data mining work, which corresponds to the data analysis method. According to the functional requirements of the financial fraud audit system studied in this paper, the construction of data mining model aims to complete the identification and risk assessment of financial fraud, that is, classifying all kinds of sample data into fraud samples and non-fraud samples, which belongs to the standard two-category problem. So, we can choose single classifiers to solve it.

In the past research, we found that there are many indicators that affect financial fraud. According to the application environment and sample number requirements of this system, Relief method is selected as the representative of filtering feature selection method, which can score data features according to correlation, and build the optimal feature data set based on the score, so as to improve the accuracy of subsequent data mining results.

2.3 Development Process

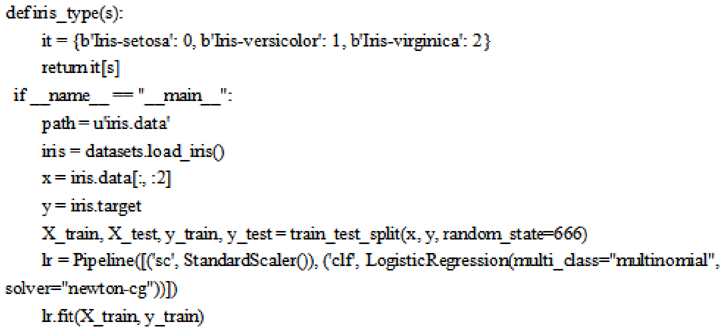
According to the application requirements of the above related technologies, complete the configuration and deployment of the development environment of the financial fraud audit system. the Hadoop cluster architecture is built with Linux as the operating system, the version is CentOS 6.7(x86_64), and the JDK version is jdk-8u291-linux-x64. According to the application requirements of the system, Hadoop cluster will be set up into seven nodes. The version of Hadoop is 2.7.7, which is installed in each node, and components such as Yarn, HDFS, Zookeeper and HBase are also deployed in each node.

Secondly, for the development of Web application server, the operating system is Windows10.0. The Web server is Nginx server, the project development language is Python 3.6.7, the development tool is PyCharm 2018.3.1 x64, and the database is MySQL5.7 to complete the construction and support of the system database system. In the server, Django framework is adopted, and the development and construction of modules, algorithms and models will be completed in the directory of "mysite" according to the requirements of system functions. Figure 1 shows the key code for the implementation of Relief method. In addition, the implementation of each classifier will also depend on the sklearn module of Python. As shown in Figure 2, the key code of Logistic regression model is realized by Pipeline() method. Through the introduction of the above key technical theories, the overall environment of the system development, the configuration of related software and tools, and the technical feasibility of the overall project of the financial fraud audit system are determined.



```
distance = np. zeros (n samples)
for index_j in rangeln_ samples) :
      D_value = features[index_i] - features [index_j]
      dietance [index_j]= distanceNorm('2' ， D_value)
distance [index_il = np. max (distance)
for index in range(n samples) :
      distance_ sort. append ([distance [index]，index, labels [index]])
distance_ sort. sort(key= lambda x: x[0])
for index in range(n samples) :
      if len (nearHit) = 0 and distance_ sort [index][2] = 1abe1s[index_ i]:
         nearHit = features [distance_ sort [index] [1]]
      elif len (nearMiss) = 0 and distance_ sort [index][2] != labels[index_ i]:
         nearMiss = features [distance_ scrt[index] [1]]
      elif len (nearHit) != 0 and len (nearMiss) != 0:
         break
      else:
         Continue
weight = weight - np. power (self_ features- nearHit, 2) + np. power (self_ features- nearMiss, 2)
```

Figure 1: Python implementation of the Relief algorithm key code.



```
def iris_type(s):
     it = {b'Iris-setosa': 0, b'Iris-versicolor': 1, b'Iris-virginica': 2}
     return it[s]
 if __name__ == "__main__":
     path = u'iris.data'
     iris = datasets.load_iris()
     x = iris.data[:, :2]
     y = iris.target
     X_train, X_test, y_train, y_test = train_test_split(x, y, random_state=666)
     lr = Pipeline([('sc', StandardScaler()), ('clf', LogisticRegression(multi_class="multinomial",
solver="newton-cg"))])
     lr.fit(X_train, y_train)
```

Figure 2: Key code of Logistic Logistic regression classifier implemented by SkLearn module.

Table 1: Classification results of three classifiers.

Number	Classifier	Accuracy rate	Precision rate	Recall rate
1	Logistic	0.88	0.78	0.44
2	SVM	0.90	0.74	0.66
3	RandomForest	0.91	0.77	0.74

3 FUNCTION REALIZATION

3.1 Administrator's Side

Under the indicator management module, the administrator can finish the primary selection of indicators, and select as many indicators as possible that have certain influence on financial fraud. After research in this paper, 55 indicators were selected, and they were divided into 8 categories according to different meanings: ratio structure, solvency, profitability, operating ability, development potential, cash flow, risk level and governance ability. (Lu, 2022) The determination of indicators will directly affect the determination of the sample data range of subsequent financial fraud design, and will also have an impact on the final fraud identification.

With the sample management module, the administrator can import all kinds of sample data. The sample determination will be completed in combination with the indicator information, which contains 59 key fields such as indicator information, primary key number, fiscal year, industry code, and fraud judgment.

3.2 Audit Client

In the model management module, audit users can add models, view models and delete models. Among them, Logistic, SVM and RandomForest models supported by the system will be divided into two groups according to the completion of training. If the training is completed, "Completed Training" will be displayed in the "Details" column on the page, and users can click the model name to view the details of the model. When the user chooses to add a model, the system will automatically complete the training of the new model, and the new model will complete the training will automatically enter the model list, convenient for the user's subsequent use.

In the financial fraud identification module, the audit user selects the unmarked sample data existing in the system, that is, the financial data and non-financial data of the enterprise in a certain fiscal year. There are 55 indicators contained in the sample data, which will be selected by the Relief algorithm. On the premise of the threshold of 0.001, 25 indicators will be selected as the results of special diagnosis of the sample data. According to the feature information, the classification results of the three classifier models are shown in Table 1.

According to the classification result, the system will automatically determine whether the sample data is a financial fraud sample. If the prediction result shows fraud, the sample will be identified as a fraud sample, otherwise, it will be a non-fraud sample.

4 CONCLUDING REMARKS

In this paper, based on the challenges in the process of financial fraud audit in the era of digital economy, an online interactive financial fraud audit system based on big data technology, data mining technology as the core and Web application technology as the framework is proposed. By using the Relief algorithm, Logistic, SVM and RandomForest classifier under the system data mining technology, the sample data features are selected and the financial fraud identification model is constructed. Finally, the financial fraud is automatically identified and judged by the accuracy and precision of classification.

REFERENCES

Jiao Haixian. Research on the Problems and Countermeasures of Corporate Financial Fraud Audit under the New Situation [J]. Money China.2021.08

Lu Xin, Li Huiming, et al. Construction of Financial Fraud Identification Framework —— Based on Accounting Information System Theory and Big Data Perspective [J]. Accounting research.2022.03

Shi Fang Xia, Gao Yi. Application analysis of Hadoop big Data Technology [J]. Modern electronic technology.2021.09

Wang Lili. Application of data mining technology in the context of big data [J]. Computer and network.2021.10

Zhao Peng, Zhu Yilan. Overview and development prospect of big data technology [J]. Astronaut systems Engineering Technology.2022.01