IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 21, 2020, accepted July 3, 2020, date of publication July 13, 2020, date of current version July 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3009006

Big Data-Driven Abnormal Behavior Detection in Healthcare Based on Association Rules

SHENGYAO ZHOU[1], JIE HE[2,3], HUI YANG[2,3], DONGHUA CHEN[1], (Graduate Student Member, IEEE), AND RUNTONG ZHANG[1,4], (Senior Member, IEEE)

[1]School of Economics and Management, Beijing Jiaotong University, Beijing 100000, China
[2]CETC Big Data Research Institute Company Ltd., Guiyang 550022, China
[3]Big Data Application on Improving Government Governance Capabilities National Engineering Laboratory, Guiyang 550022, China
[4]Beijing Logistics Informatics Research Base, Beijing 100000, China

Corresponding author: Runtong Zhang (rtzhang@bjtu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China (Key Program) under Grant 71532002, in part by the National Social Science Foundation of China (Major Project) under Grant 18ZDA086, in part by the Fundamental Research Funds for the Central Universities under Grant B19JB500230, in part by the National Engineering Laboratory Open Fund Project, and in part by the Beijing Logistics Informatics Research Base.

ABSTRACT Healthcare insurance frauds are causing millions of dollars of public healthcare fund losses around the world in various ways, which makes it very important to strengthen the management of medical insurance in order to guarantee the steady operation of medical insurance funds. Healthcare fraud detection methods can reduce the losses of healthcare insurance funds and improve medical quality. Existing fraud detection studies mostly focus on finding normal behavior patterns and treat those violating normal behavior patterns as fraudsters. However, fraudsters can often disguise themselves with some normal behaviors, such as some consistent behaviors when they seek medical treatments. To address these issues, we combined a MapReduce distributed computing model and association rule mining to propose a medical cluster behavior detection algorithm based on frequent pattern mining. It can detect certain consistent behaviors of patients in medical treatment activities. By analyzing 1.5 million medical claim records, we have verified the effectiveness of the method. Experiments show that this method has better performance than several benchmark methods.

INDEX TERMS Big data, abnormal behavior, healthcare insurance, association rules.

I. INTRODUCTION

Medical insurance is a social insurance system established to compensate workers for economic losses caused by disease risks. The medical insurance funds are established via payments from insured employers and individuals, and their medical expenses for medical treatment will be partly compensated by medical insurance institutions. The establishment and implementation of the medical insurance system can enable patients to obtain the necessary help, reduce the burden of medical expenses, and prevent the diseased members of the society from becoming ''poor after illness'' [1].

In recent years, China's social medical insurance has developed rapidly. Increasing the coverage of social medical insurance has become the most important task for China's social security system. By the end of 2018, 1.345 billion people had registered in the basic medical insurance, covering more than 95 percent of the total population. As shown in Table 1, the total income of the basic medical insurance funds for the whole year of 2018 were 2,109,011 billion yuan, and the total expenditure was 1,760.765 billion yuan [2]. It can be seen from Figure 1 that the amount of China's medical insurance funds keeps increasing every year, while the balance rate keeps decreasing. From 23.0% in 2012 to 10.0% in 2018, there has been a continuous decline. Therefore, how to ensure the normal operation of social medical insurance funds, improve the level of medical insurance management, and reasonably and effectively avoid potential business risks has become an extremely important issue.

In August 2016, the National Audit Office of China authorized local audit institutions to conduct special audits on medical insurance funds, such as basic medical insurance and urban and rural residents' critical illness insurance [3]. This was the most comprehensive audit ever since China's health care reform. The audit randomly selected funds from 28 provinces, 166 cities, and 569 counties (cities and districts) to check their performances in 2015 and the first

The associate editor coordinating the review of this manuscript and approving it for publication was Xin Luo.

TABLE 1. Main indicators of China's medical insurance in 2018.

		Number of insured people (million)	Revenue (billion yuan)	Expenditure (billion yuan)	Fund current balance (billion yuan)
Employee health insurance		317	1325.928	1050.492	275.436
Resident health insurance	Urban and rural areas	897	697.394	628.451	68.943
	The new rural cooperative medical insurance	13	85.689	81.822	3.8670
	Subtotal	1027	783.083	710.273	72.810
Total		1344	2109.011	1760.766	348.246

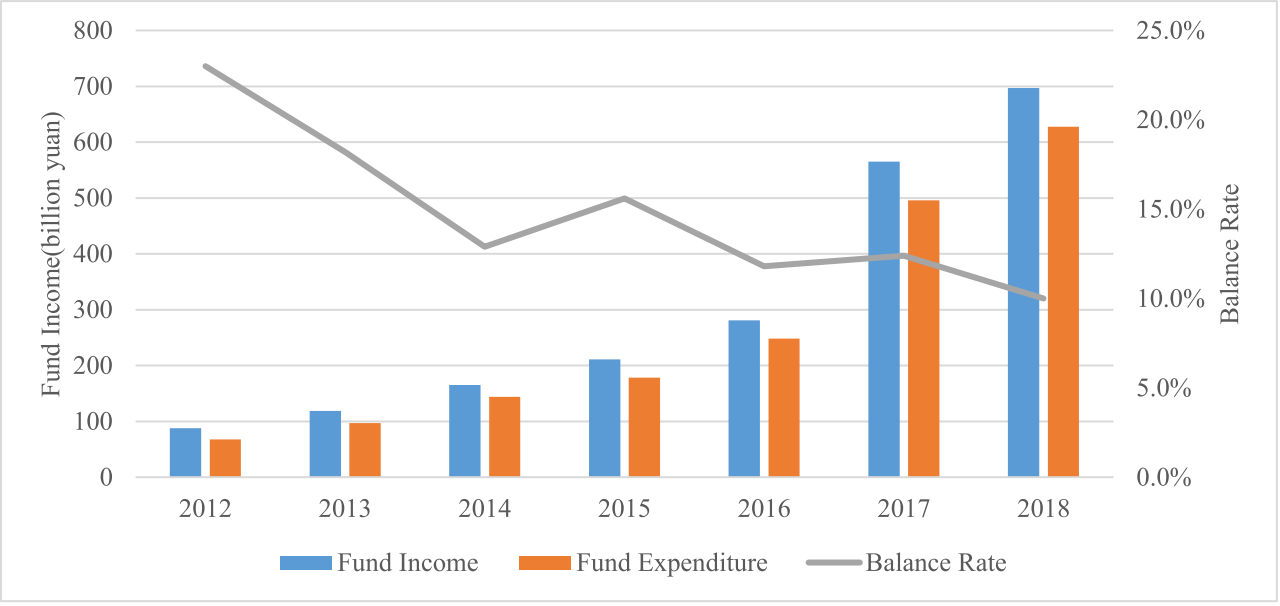

FIGURE 1. Revenue and expenditure of residents' medical insurance from 2012 to 2018.

half of 2016. The total funds reached 343.313 billion yuan, and 1.578 billion yuan of them were illegal, revealing many irregularities including repeated reimbursement of medical expenses, fraudulent medical treatment by some designated agencies and individuals, and fraudulent medical insurance funds through decommissioning and hospitalization.

And from a global perspective, the problem of medical insurance anomalies has also attracted much attention [4]. In 2016, a large German public medical insurance company was forced to pay a fine of €7 million to the German Federal Insurance Agency over medical insurance anomalies. The same year, the U.S. Department of Justice cracked down on the biggest medical anomaly case in its history, involving up to $ 900 million and more than 300 people, such as doctors, nurses, and pharmacists, who were accused of participating in medical anomalies. Abnormal medical insurance seriously endangers the entitlements of the insured, and these abnormal behaviors of medical insurance must be put under control.

Common abnormal medical insurance behaviors can be divided into three categories according to different subjects of such behaviors [5]. Category one includes abnormal behaviors of medical insurance individuals, including frequent medical consultations, fraudulent use of other people's health insurance cards, etc. Category two refers to abnormal behaviors of medical institutions, including over-diagnosis and unreasonable medication, and Category three is a kind of joint fraud conducted by individuals and institutions, including fake invoices and admission checklists. Due to the unequal flow of information, category two and three often remain so elusive that even non-medical staff of the General Medical Insurance Bureau can hardly find hidden abnormal behaviors in verification [6].

Although it is difficult to find medical insurance anomalies, medical units have kept a large number of medical visit records and data with the widespread use of medical information systems. Similarly, all medical reimbursement behaviors have been recorded in the medical reimbursement data set. Through researches and analyses of the medical reimbursement data set, abnormal behaviors hidden in it can be discovered. Traditional analyses of medical insurance anomalies mostly take on medical practitioners' experience to make artificial rules and simple statistical analyses. It is difficult to accurately sort out complete abnormal behavior information from complex medical insurance data. As far as the current level of informationization in the medical industry, medical data has developed four basic characteristics of big data [7], [8]. In the context of big data, we can establish a distributed medical insurance abnormal behavior

detection model based on Hadoop, which makes it easier for medical practitioners to find medical insurance anomalies more quickly, dig out abnormal points in massive data, and supervise abnormal behavior for medical practitioners. This has great practical significance and value.

There is a special medical phenomenon in the analysis of medical insurance fraud. This special phenomenon is usually manifested in the fact that multiple medical insurance cards are consumed too frequently at the same time, which is called the medical agglomeration behaviors [9]. This kind of behaviors may be conducted by certain special illness groups, such as chronic patients, and may also be a kind of fraud. It has great meaning on finding who have medical-treatment-related behaviors. On the one hand, efforts could be made to provide targeted management and services for people with special diseases. On the other, fraud shall lead to effective improvement of regulation.

Therefore, we propose a distributed anomaly detection method for medical aggregation behaviors. Our main contributions in this paper are listed as follows:

1) constructing a medical aggregation behavior model that includes a formal description of medical aggregation behaviors;

2) designing a distributed anomaly detection algorithm and corresponding interpretation of the detection results;

3) compared with several benchmark methods for frequent itemset mining, the performance advantage of this method becomes more significant as the amount of data continues to increase, which can significantly improve the accuracy of fraud detection. More specifically, our DCMMAB is better than the comparison method by more than 20% in precision.

At present, this method has been integrated into the medical big data analysis platform to provide decision support for auditors in the medical insurance claims system to assess the possibility of fraud.

The rest of the paper is organized as follows: Section 2 reviews related works on fraud detection issues; Section 3 briefly introduces the framework and related concepts of big data; Section 4 gives the problem definition of medical aggregation behavior detection, and introduces the method of mining medical aggregation behavior based on distributed computing(DCMMAB); Section 5 analyzes the real medical insurance data through our method and interprets the experimental results. And compared with several other benchmark methods, it proves the superiority of our method; Section 6 summarizes our work and discusses several future research directions.

II. RELATED WORK

Medical insurance fraud is not a problem unique to a country, and countries around the world that implement medical insurance systems are facing corresponding problems. At present, the research on medical insurance fraud is mainly divided into three aspects: the causes and characteristics of fraud, how to combat fraud, and the identification of fraud.

In terms of the causes and characteristics of fraud, reference [10] explains the causes of medical fraud based on the perspective of information asymmetry. Reference [11] refers to the profound experience of anti-fraud behaviors, and applies phenomenology of qualitative explanation to explain the causes of fraud behaviors. Reference [12] constructs a patient-centric analysis model by analyzing various frauds. Reference [13] details the classification and causes of fraud in American medical insurance funds.

In terms of how to combat fraud, reference [14] analyzes fraud behavior from the perspective of the costs and benefits of the fraudster, and proposes an impact factor model of fraud behaviors. Reference [15] analyzes the causes of fraud and its harm, and gives corresponding suggestions on how to combat it. Reference [16] analyzed fraud in the process of collection, payment and funds management of medical insurance funds, and proposes a series of measures to combat fraud.

In terms of the identification of fraudulent behaviors, traditional medical fraud detection methods are mainly based on rules established by experts [17]. Once the given rules are violated in the medical records, it will be judged as fraudulent behaviors. The effectiveness of these methods is constrained by the correctness of the rules. With the widespread application of big data technology in the medical field, data mining technology has been applied to the detection of medical fraud. As early as 1999, studies have pointed out that potential data patterns in data set can be discovered through data mining technology, thus providing a basis for scientific decision-making [18]. Reference [19] introduces the successful cases of data mining technology in medical fraud detection. Reference [20] applies data mining and machine learning techniques to the construction of model library and method library from the perspective of risk prevention and control of medical insurance funds.

At present, some anomaly-detection methods [21]–[24] have also been applied to the detection of medical fraud. Reference [25] uses IBM Bluemix platform and open cloud platform to build a medical reimbursement data analysis and display platform. He describes a diagnosis and treatment process between patients and doctors with diagrams, where patients and doctors are different nodes, and a diagnosis and treatment process is regarded as an edge so that the connection between the patient and the doctor can be analyzed. Reference [26] comprehensively applies the semi-supervised IsoMap method and LOF method to detect the abnormal expenses, and time constructed a medical insurance claim data anomaly detection system. In addition, many studies [27]–[29] have focused on dividing patients into different groups by using certain rules, and using different models for each group to fit medical expenses to determine whether there are abnormalities.

In summary, we can see that the method of detecting medical insurance fraud has gone through three stages. The first stage is the prevention and treatment of medical insurance fraud from the perspective of system management

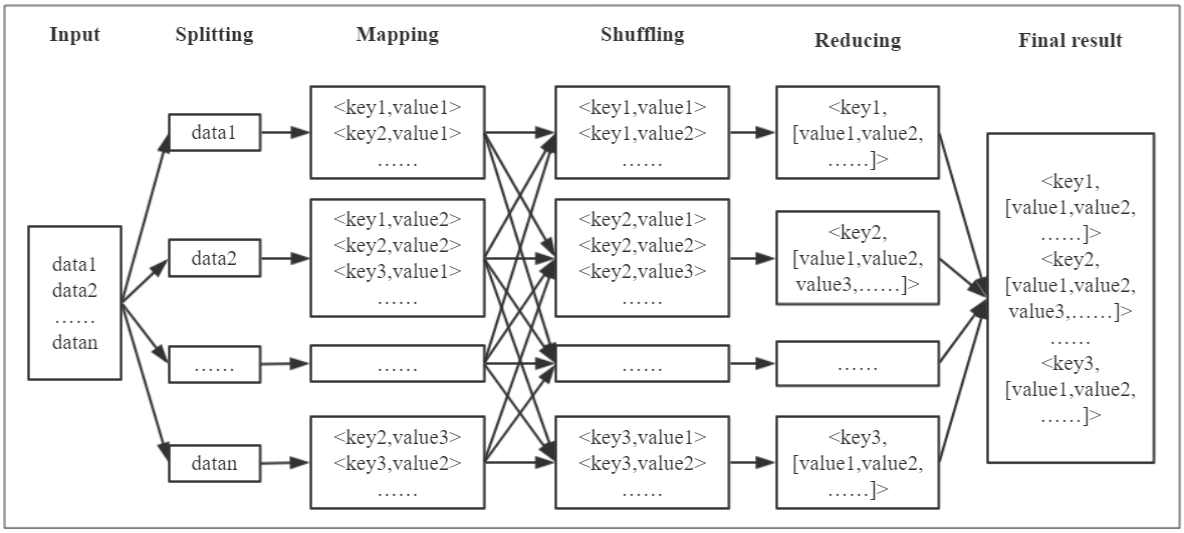

FIGURE 2. MapReduce processing flow.

and funds payment model, where the application is relatively simple and insufficient. The second stage is the introduction of data mining technology to control the risk of medical insurance fraud. Compared with the first stage, the efficiency of medical insurance fraud detection has been greatly improved. With the development of medical information, medical insurance data have grown rapidly, and medical insurance fraud has become more diverse. So it has entered into the technology stage of machine learning. Therefore, on the basis of previous researches and big data related technology [30], this paper proposes a distributed detection algorithm for mining medical aggregation behavior fraud, and applies it to the actual data analysis.

III. BACKGROUNDS AND PRELIMINARIES

In this section, we discuss some preliminary terms and concepts that need to be understood before understanding the general concept of MapReduce [31] and HDFS (Hadoop Distributed File System) [32].

Hadoop is a distributed computing platform developed based on the Java language by Apache. Because of its high reliability, high scalability, high efficiency, high fault tolerance, low cost, and complete open source, it is widely used in many industries and scientific research fields. Hadoop provides users with a distributed infrastructure with transparent underlying details of the system. Its distributed file system HDFS and distributed computing model MapReduce have been proven to be able to successfully analyze and process big data in parallel on a large number of computer clusters. In MapReduce-based development, developers only need to pay attention to the segmentation of the data set, the division of Map and Reduce tasks, and the implementation of Map and Reduce functions. All other complex parallel computing programming problems, such as distributed storage, task scheduling, load balancing, fault tolerance and network communication are all completed by the MapReduce framework, which greatly reduces the difficulty of development. Users can use Hadoop to easily organize computer resources, build distributed platforms and complete parallel programming with the help of MapReduce computing models, providing a feasible and efficient solution for the storage and processing of massive data.

The whole processing process of MapReduce is shown in Figure 2. The principle is to use a set of input key-value pairs <key, value> to generate a set of output key-value pairs. The user expresses this calculation process by customizing Map and Reduce functions: The Map function accepts an input key-value pair and generates a set of intermediate key-value pairs. MapReduce then aggregates all intermediate value values with the same intermediate key value and passes them to the Reduce function. The Reduce function accepts an intermediate key value and a corresponding value set, and merges with the intermediate value set to form a smaller value set. Therefore, when encountering a situation in which the amount of data is too large to fit into memory, a large number of intermediate value sets can be provided to the Reduce function for processing through an iterator.

MapReduce's calculation process is specifically described as follows:

***Input*:** The input data set is divided into M splits of the same size, the split information and configuration information are stored on HDFS, and the task is submitted to JobTracker. JobTracker assigns M Map sub-tasks and R Reduce sub-tasks to idle TaskTracker, and puts all tasks in a queue.

Map subtask: Obtain data from HDFS, generate <key, value> after processing, call Map function to receive all input key-value pairs, generate an intermediate set as the output of Map function, and divide it into R parts by the same Hash function. The result is written into the file and the location information is sent to the JobTracker. JobTracker sends the location information to the node that assumes the Reduce subtask.

Reduce subtask: The node obtains the output subset (1 / R) of the Map task according to the received location

information, sorts them based on the key value, and then combines all <key, value> with the same key value to form a smaller set as input to the Reduce function. After the Reduce function finishes running, it outputs the results to a file.

Output: After all Map and Reduce sub-tasks are completed, JobTracker returns the output results of the Reduce sub-tasks to the client program, which is merged by the client program to obtain the final result.

IV. MATERIALS AND METHODS

In this section, we propose a distributed fraud detection method based on the definition of medical insurance gathering behavior in reference [9], which can mine data records that may participate in aggregation fraud from medical insurance data. This problem is a novel and practical fraud detection problem. Obviously, it overlaps with the work on frequent pattern mining, but it is significantly different from them.

A. PROBLEM DEFINITION

According to the definition of the medical insurance gathering behavior in the reference [9], we can know that this behavior usually manifests as multiple medical insurance cards being consumed in the same place too frequently at the same time when the patient is in the hospital. This phenomenon of medical agglomeration may be a tendency to violate the rules: one person with multiple medical insurance cards may consume several cards for one treatment. Therefore, the manifestation of medical agglomeration behavior can be simplified as a kind of consistency: multiple medical insurance cards being used in one exact hospital at the same time. This kind of consumption behavior can be regarded as an anomaly if it is too frequent, and we will supervise it.

Definition 1: The two core data in the medical record are the visit time and the visit place. In our model, we use one day as the unit of consultation time, so let d be the set of visit time, and l be the set of visit place. The two together form a medical visit matrix. According to definition, the medical records of each medical insurance card MC_i can be expressed as

$$MC_i = \begin{bmatrix} l_1d_1 & \cdots & l_1d_n \\ \vdots & \ddots & \vdots \\ l_md_1 & \cdots & l_md_n \end{bmatrix},$$

where $l_md_n = \langle 0 \mid 1 \rangle$. If MC_i sees a doctor at the hospital l_m at d_n time, then $l_md_n = 1$, otherwise $l_md_n = 0$, so the main form of MC_i is

$$MC_i = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix},$$

and the medical database MDB composed of MC can be defined as

$$MDB = \begin{bmatrix} MC_1 \\ \vdots \\ MC_m \end{bmatrix} \quad (i \le m).$$

According to definition 1, the medical gathering behavior can be expressed as in the MDB, $\{MC_1, MC_2, \cdots, MC_i\}$ has the same value on l_md_n, so the following definition is introduced.

Definition 2: Let S be the medical behavior matrix composed of $\{MC_1, MC_2, \cdots, MC_i\}$, then S is a subset of MDB, and the same row in S represents the medical records of MC every day, and the same column represents the medical records of different MC on the same day. According to definition 1, if $\{MC_1, MC_2, \cdots, MC_i\}$ there is a medical gathering behavior, each row of S has the same value.

Let the parameter *min_row* be the shortest number of rows and the parameter *min_column* be the shortest number of columns. If rows number *S.row* of the matrix S is not less than *min_row*, S is considered to be a frequent pattern, and if the number of columns *S.column* of S is not less than *min_column*, S is considered to be abnormal. The mining of medical gathering behavior needs to find all the abnormal matrix S.

Definition 3: Mining the medical gathering behavior for a given *min_row* and *min_column*, find all matrices S that simultaneously meet the following conditions:

1) $S.row \ge min.row$;

2) $S.column \ge min.column$.

Definition 3 transforms the aggregation behavior mining problem into mining frequent patterns. The distributed aggregation behavior mining algorithm in the big data environment is introduced in detail below to solve the above problems.

B. MEDICAL AGGREGATION BEHAVIOR MINING METHOD BASED DISTRIBUTED COMPUTING

The vast majority of databases have a horizontal data format {ID,MC,l,d, $\cdots$}. The distributed medical aggregation behavior mining method first scans the original medical database, deletes data that does not meet the requirements, and then transpose the format to generate vertical data format. The first order matrix is used to generate the second order matrix, and so on. Each new matrix intersects with the first-order matrix to generate a higher-order matrix until no new matrix is generated.

In the original medical database, if S.column is less than min_column, it cannot exist in a pattern greater than or equal to *min_column*. Therefore, when scanning the original medical database, the algorithm needs to delete this part of data first. Then the *MCs* with the same l and d are merged and transposed, so that the medical database format is converted to $\{l_id_j, [MC_1, MC_2, \cdots, MC_i]\}$.

During cross calculation, the *S.row* can be calculated at the same time, so it is no longer necessary to repeatedly scan the entire medical database.

In addition, since most of l_id_j values are 0, the efficiency will be greatly improved compared to other frequent pattern mining algorithms. After the transposition is completed, the algorithm can generate the matrix S_1 simultaneously.

After generating S_1, higher-order matrices can be continuously generated through crossover operations.

Algorithm 1 MRTranspose

Input: Transaction database MDB
Output: Key: $R_i\{l, d\}$, Value: $R'_i\{MC\}$
Map
1) **let** $R_i := MDB.MC_i$
2) **write** $< R_i\{l, d\}, R_i\{MC\} >$
Reduce
3) **for** each j in $R_i\{MC\}$ **do**
4) **if** $R_i\{l, d\} = R_j\{l, d\}$ **then**
5) **let** $R'_i\{MC\} := R_i\{MC\} \cup R_j\{MC\}$
6) **end if**
7) **end for**
8) **write** $< R_i\{l, d\}, R'_i\{MC\} >$

Algorithm 2 MRFrequencyItemSet

Input: $R_i\{l, d\}, R'_i\{MC\}, min\_row$
Output: Key: S, Value: $S.row$
Map
1) **for** each i in $R'_i\{MC\}$ **do**
2) **let** $S_i := R'_i\{MC\}$
3) **for** each S in S_i **do**
4) **write** $< S, 1 >$
5) **end for**
6) **end for**
Reduce
7) **for** each *item* **do**
8) **let** $S.row := S.row + 1$
9) **end for**
10) **if** $S.row \geq min\_row$ **then**
11) **write** $< S, S.row >$
12) **end if**

Algorithm 3 MRJoin

Input: $S, S.row, R_i\{l, d\}, min\_column, min\_row$
Output: Key: S'_{len+1}, Value: $R'_i\{l, d\}$
Map
1) **for** each $R_i\{l, d\}$ **do**
2) **if** $S_i.column < min\_column$ **then**
3) delete $R_i\{l, d\}$
4) **end if**
5) **end for**
6) **if** $S_1 \notin S_{len}$ **then**
7) **add** $S'_1 \cup S'_{len}$ to S_{len+1}
8) **write** $< S_{len+1}, R'_i\{l, d\} >$
9) **end if**
Reduce
10) **if** S_{len+1} not all Subset Frequent
11) **if** $\| S'_1 \cap S'_{len} \| . row \geq min\_row$ **then**
12) **add** $S'_1 \cap S'_{len}$ to S'_{len+1}
13) **write** $< S'_{len+1}, R'_i\{l, d\} >$
14) **end if**
15) **end if**

For S_i, S_{i+1} can be generated by the self-linking of S_i. But for $\forall S \subseteq S_{i+1}$, if S can be generated by S_i self-connection, then S can also be generated by connecting S_i with S_1. Therefore, connecting S_i and S_1 to generate S_{i+1} can replace the self-connection operation of S_i. Since S_i contains a large amount of data, this method can greatly reduce the number of cross operations.

In addition, according to the nature of association rule mining, any sub-pattern of frequent pattern is frequent, so S_i can be pre-expanded to S_{i+1} before cross-operation. If S_{i+1} has infrequent S_i, S_{i+1} can be deleted.

We assign the scanning, construction, and generation results of each stage to the map subtask and reduce subtask. Algorithm 1 to Algorithm 3 are the pseudo-codes of each stage functions.

V. EXPERIMENT RESULTS

In this section, we present the experimental results using the proposed method. First, we describe the experimental environments and provide implementation details. Then, we demonstrate the effectiveness of this method by comparing with several benchmark methods. Finally, we show the practical usage of this method in real systems.

A. EXPERIMENTAL SETTINGS

Experimental platform is a Hadoop analysis platform built by 14 servers. The detailed description of the nodes is shown in Table 2. The medical insurance data in the experiment are selected from the medical insurance claim system of the medical insurance administration department of a county in China, so there is almost no sparse input data in the data set. The data set covers the county's outpatient records for the past three years (January 2017 to October 2019). Table 3 shows an example of the original medical insurance record. After cleaning the data set, we obtained 1,574,775 outpatient reimbursement information from 151,679 patients.

B. RESULTS AND ANALYSIS

After detecting and analyzing this data set through our algorithm, we obtained 872,042 pieces of correlation data in which the number of correlation data distribution are shown in Figure 3 to Figure 5. Figure 3 is "the quantitative distribution of the associated data over different *S.row* values", which can be regarded as a two-dimensional display of Figure 5 to some extent. From Figure 3, we can see that as *S.row* increases, the number of associated data keeps falling and reaches a stable state after $S.row = 12$. Obviously, a higher *S.row* means more simultaneous patient visits and less correlated data, and the higher possibility of fraud, which is in line with what we know. Figure 4 is "the quantitative distribution of the associated data over different *S.column* values". Similarly, it can also be regarded as a two-dimensional display of Figure 5 to some extent. From Figure 4, we can see that with the increase of *S.column*, the number of associated data increases first and then decreases, presents a symmetrical relationship before and after. Likewise, a higher *S.column* means that the more patients go to the clinic at the same time, and the higher possibility of fraud, which also fit

TABLE 2. Hadoop node configuration description.

Node	Last Contact	Admin State	Capacity	Used	Non DFS Used	Eemaining	Blocks	Block pool used	Failed Volumes
slave1:50010	0	In Service	1.78 TB	235.92 GB	98.21 GB	1.46 TB	4927	235.92 GB	0
slave2:50010	1	In Service	1.78 TB	243.82 GB	97.98 GB	1.45 TB	5061	243.82 GB	0
slave3:50010	0	In Service	1.78 TB	237.56 GB	98.17 GB	1.46 TB	4930	237.56 GB	0
slave4:50010	2	In Service	1.78 TB	310.58 GB	97.94 GB	1.38 TB	5521	310.58 GB	0
slave5:50010	0	In Service	1.78 TB	247.39 GB	96.67 GB	1.45 TB	5097	247.39 GB	0
slave6:50010	1	In Service	1.78 TB	334.63 GB	96.99 GB	1.36 TB	5738	334.63 GB	0
slave7:50010	2	In Service	1.78 TB	242.24 GB	96.86 GB	1.45 TB	5077	242.24 GB	0
slave8:50010	0	In Service	1.78 TB	244.53 GB	96.76 GB	1.45 TB	4991	244.53 GB	0
slave9:50010	2	In Service	1.78 TB	228.3 GB	96.83 GB	1.47 TB	4866	228.3 GB	0
slave10:50010	1	In Service	1.78 TB	252.11 GB	97.36 GB	1.44 TB	5060	252.11 GB	0
slave11:50010	0	In Service	1.78 TB	238.53 GB	97.4 GB	1.45 TB	4977	238.53 GB	0
slave12:50010	0	In Service	1.78 TB	234.01 GB	97.88 GB	1.46 TB	4925	234.01 GB	0
slave13:50010	1	In Service	1.78 TB	248.43 GB	97.35 GB	1.45 TB	5108	248.43 GB	0
slave14:50010	1	In Service	1.78 TB	237.3 GB	97.06 GB	1.46 TB	5010	237.3 GB	0

TABLE 3. Example of medical insurance records.

Outpatient Compensation Number	Medical Card Number	Sex	Age	Disease Code	Consultation Date	Compensation Date	Total Cost	Actual Compensation Amount	Out-of-pocket Amount
198459880	5226281103050008	Female	35	H10.201	2017/1/7	2017-01-07 12:16	53.67	26.84	26.83
198459820	5226280806030014	Male	2	R05.01	2017/1/7	2017-01-07 12:15	23	11.5	11.5
198459629	5226280901030004	Male	1	R10.4	2017/1/7	2017-01-07 12:12	44.67	22.34	22.33
198459605	5226280901030004	Male	1	R10.4	2017/1/7	2017-01-07 12:12	16.5	8.25	8.25
198459584	5226280901030004	Male	1	R10.4	2017/1/7	2017-01-07 12:11	16.42	8.21	8.21
198459441	5226281005170023	Male	74	J40.03	2017/1/7	2017-01-07 12:09	29.14	14.57	14.57
198459426	5226281005170023	Male	74	J40.03	2017/1/7	2017-01-07 12:09	2.56	1.28	1.28
198459408	5226281005170023	Male	74	J40.03	2017/1/7	2017-01-07 12:09	13	6.5	6.5
198459316	5226281209040058	Male	87	I00.03	2017/1/7	2017-01-07 12:09	49.08	34.36	14.72
198459383	5226281005170023	Male	74	J40.03	2017/1/7	2017-01-07 12:09	2.82	1.41	1.41
198459223	5226280906010036	Female	1	K52.905	2017/1/7	2017-01-07 12:06	3.91	1.96	1.95

our perception. Of course, it is not comprehensive enough to analyze the associated data by *S.row* or *S.column*, we need to combine them. Figure 5 is the distribution of the number of associated data under different values of *S.row* and *S.column*. Compared with Figure 3 and Figure 4, the distribution of the associated data in Figure 5 has some changes, this is because *S.row* and *S.column* have a mutual influence on each other. Although the increase of *S.row* or *S.column* means the higher possibility of fraud, but Figure 5 shows that they will not increase at the same time, so we need to find a balance point as *min_row* and *min_column*.

For further analysis, we need to limit min_row and min_column. When min_row = 5, the data set related records are reduced to 719. Then we remove the record of a single medical insurance card, that is, min_column = 2. At this time the data set contains 291 pieces of data, and the quantity distribution is shown in Figure 6. We performed frequent itemsets S.row analysis and support S.column analysis on 291 pieces of data, and obtained the following findings:

1) Among 291 pieces of data, when min_column = 2, there are 18 data of S.row $\geqslant$ 7, among which the S.rowmax = 11. That is, during the three years from January 2017 to October 2019, two medical insurance cards appeared in the same hospital more than 7 times at the same time, and this phenomenon occurred 18 times. At most, two medical insurance cards appear 11 times in the same hospital

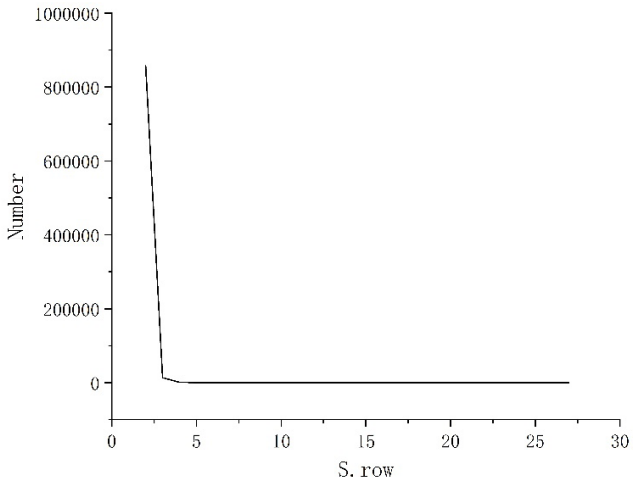

FIGURE 3. The quantitative distribution of the associated data over different *S.row* values.

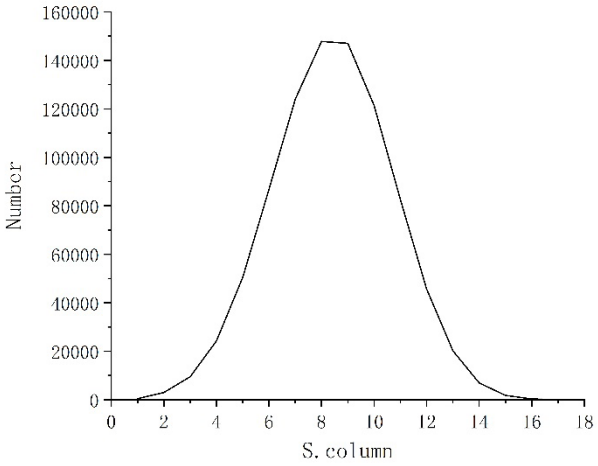

FIGURE 4. The quantitative distribution of the associated data over different *S.column* values.

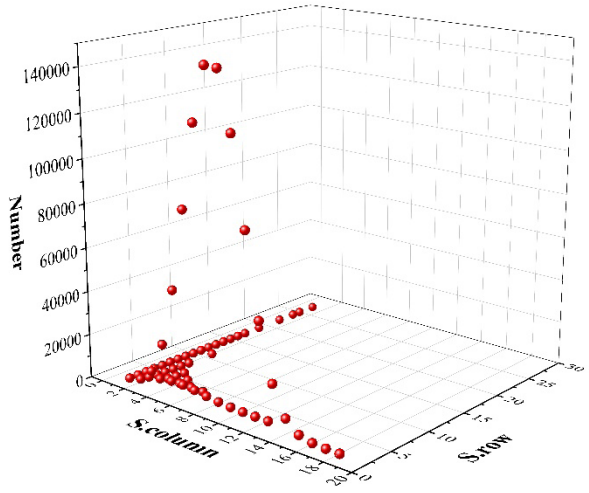

FIGURE 5. Distribution of the number of associated data under different values of *S.row* and *S.column*.

at the same time, so it's reasonable to believe that this is not an accidental phenomenon, but more likely to be a fraud.

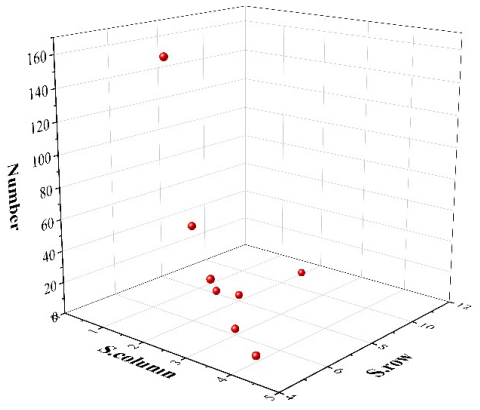

FIGURE 6. Distribution of the number of associated data under different values of *min_row* = 5 and *min_column* = 2.

2) Among 291 pieces of data, $S.column_{max} = 4$. There are four pieces of data at this time, and all the values of *S.row* are 5. That means for the three years from January 2017 to October 2019, four medical insurance cards appeared five times in the same hospital at the same time. This phenomenon occurred four times. Since *S.column* is big enough, we have reasons to believe that it is likely to be a fraud.

In this method, *S.row* and *S.column* in the medical behavior matrix S explain this anomaly from two different angles. Of course, through data analysis, anomalous data can be detected, but it is not certain that this must be a fraud. There is just such a possibility, and the greater the *S.row* or *S.column*, the higher the possibility. Although in many cases it still needs such professionals as doctors or government regulators to determine if it is really a fraud, this paper can provide the anomaly data detected by the algorithm, making it a meaningful job.

As in Reference [9], we apply the classic frequent item set mining algorithms Apriori [33], Eclat [34] and BP-Growth [35] to the frequent pattern mining in this paper and compare with this method. When $min\_row = 2$, the running time of these methods under different data amounts is shown in Figure 7.

The experimental results show that under the same condition of *S.row*, Apriori has the longest running time, which is because it needs to scan the database repeatedly, so it takes a great deal of time. At the same time, many candidate sets are generated in the process of pattern growth, which requires repeated crossover operation. Compared with Apriori, BP-Growth adopts a tree structure, it can directly obtain frequent sets without generating candidate frequent sets, which greatly reduces the number of times to scan the transaction database, so it is more efficient. However, it is not suitable for parallel computing. Although Eclat also adopts longitudinal format data mining, our method preprocesses the data before transposing the format, and deletes some data that does not meet the requirements in advance. At the same time, it uses the connection operation between S_i and S_1 instead of S_i self-connection operation, so it has higher efficiency.

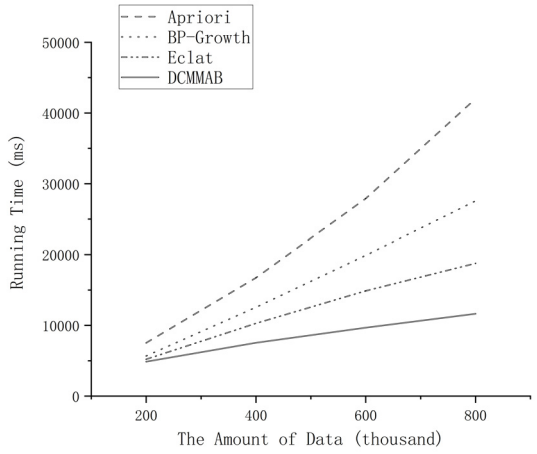

FIGURE 7. Running time comparison when *min_row* is 2.

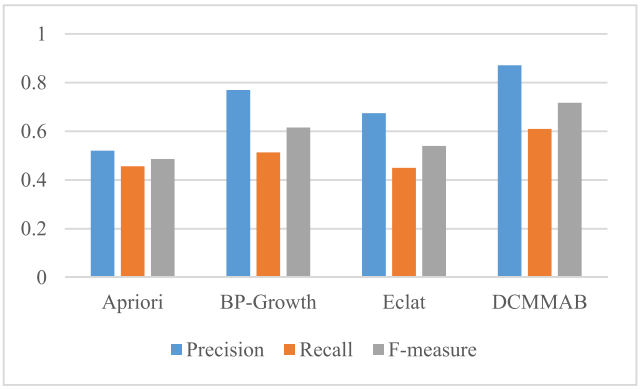

FIGURE 8. Precision, Recall, F-measure Comparison.

Figure 8 shows the performance of DCMMAB against other approaches. We have several interesting observations which confirm our research motivation from Figure 8. Due to the extremely low percentage of positive data, the performance of the Apriori method needs to be improved, as most fraudsters will try their best to bypass routine detection rules. The BP-Growth method has high precision but low recall because there is few behavior pattern in the crowd. And Eclat can hardly find meaningful frequent itemsets from the whole crowd because of the curse of cardinality. In contrast, our DCMMAB method significantly improve the precision by more than 20%. This observation shows that our approach can effectively reduce the false positives. Moreover, our method also performs better in terms of other metrics. For example, the recall rate of our method is 15% higher than the Eclat method. As a result of high precision and high recall, when these two metrics are combined together to form the f-measure shown in Figure 8, DCMMAB consistently beats the comparison approaches in the experiments. On average, DCMMAB outperformed the comparison method by more than 10% on the F-measure.

Experimental results show that our method has better performance for medical insurance fraud of medical aggregation behavior.

VI. CONCLUSION

In this paper, we give a definition of medical aggregation behavior and propose an effective method of fraud identification. The method DCMMAB combines the MapReduce distributed computing model and association rule mining to detect abnormal behaviors in the medical insurance reimbursement process. We use a real dataset from a county's medical insurance system in China, which contains 1.5 million records of medical claims activity from 150,000 users. Experimental results show that as the amount of data increases, the performance advantages of this method become more obvious, which can significantly improve the accuracy of fraud detection. More specifically, our PCDHIFD is better than the comparison method by more than 20% in precision.

At present, this method has been integrated into the medical big data analysis platform to provide decision support for auditors in the medical insurance claims system to assess the possibility of fraud. In subsequent research, we will focus on the differences between different diseases, and explore the potential links between disease types and medical behaviors.

REFERENCES

[1] W. Wang, "Analysis of the causes of medical insurance abnormalities and their supervision," *Manag. Observ.*, vol. 34, no. 8, pp. 164–166, 2014.
[2] National Healthcare Security Administration of The People's Republic of China, China. (Feb. 28, 2019). *The Statistical Report of the Development of Medical Insurance in 2018*. [Online]. Available: http://www.nhsa.gov.cn/art/2019/2/28/art_7_942.html
[3] National Audit Office of The People's Republic of China, China. (Jan. 24, 2017). *Announcement No. 1 of 2017: Audit Results of Medical Insurance Fund*. [Online]. Available: http://www.audit.gov.cn/n5/n25/c92641/content.html
[4] Y. Gao, "Research on key issues of fraud detection in medical insurance big data," Ph.D. dissertation, Dept. Comput. Sci. Tech., Shandong Univ., Shandong, China, 2018.
[5] Y. Li and B. You, "Analysis of the characteristics of medical insurance fraud," *China Soc. Secur.*, vol. 22, no. 2, pp. 76–79, 2015.
[6] Y. Lin, "An analysis on medical insurance fraud researches in the domestic and overseas market," *Insurance Stud.*, vol. 31, no. 12, pp. 115–122, 2010.
[7] S. F. Wamba, S. Akter, A. Edwards, G. Chopin, and D. Gnanzou, "How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study," *Int. J. Prod. Econ.*, vol. 165, pp. 234–246, Jul. 2015.
[8] Z. Zhang, Y. Zhou, and S. Du, "Medical big data and its opportunities and challenges," *J. Med. Informat.*, vol. 36, no. 6, pp. 2–8, 2014.
[9] S. Zhou, R. Zhang, J. Feng, D. Chen, and L. Chen, "A novel method for mining abnormal behaviors in social medical insurance," in *Proc. IEEE 9th Annu. Inf. Technol., Electron. Mobile Commun. Conf. (IEMCON)*, Nov. 2018, pp. 683–687.
[10] G. Dionne and R. Gagne, "Replacement cost endorsement and opportunistic fraud in automobile insurance," *J. Risk Uncertainty*, vol. 24, no. 3, pp. 213–230, May 2002.
[11] J. M. Skiba, "A phenomenological study of the challenges and barriers facing insurance fraud investigators," *J. Insurance Regulation*, vol. 33, pp. 87–114, 2014.
[12] J. H. Krause, "A patient-centered approach to health care fraud recovery," *J. Criminal Law Criminology*, vol. 96, no. 2, pp. 579–619, 2006.
[13] F. A. Lorenz, "Healthcare fraud in the United States: Assessing current policy and its role in fraud prevention," M.S. thesis, Dept. Public Health, California State Univ. Northridge, Los Angeles, CA, USA, 2013.
[14] L. Li, "The study of social medical insurance fraud based on cost-benefit theory," *Theory Pract. Finance Econ.*, vol. 31, no. 1, pp. 32–36, 2010.
[15] M. Wang and S. Tao, "Analysis of the influencing factors of the implementation of major illness medical insurance in China," *Manager J.*, vol. 29, no. 21, pp. 298–305, 2013.

[16] H. Xia, K. Wang, and S. Zhang, "Fraud and anti-fraud issues in medical insurance," *Mod. Preventive Med.*, vol. 34, no. 20, pp. 3907–3908, 2007.
[17] W. W. Cohen, "Fast effective rule induction," *Mach. Learn. Proc.*, vol. 16, no. 2, pp. 115–123.
[18] S. Biafore, "Predictive solutions bring more power to decision makers," *Health Manage. Technol.*, vol. 20, no. 10, p. 12, 1999.
[19] D. Delen, C. Fuller, C. McCann, and D. Ray, "Analysis of healthcare coverage: A data mining approach," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 995–1003, Mar. 2009.
[20] J. Zhang and Z. Lin, "A data mining technique on the medical insurance fund risk prevention and control platform," *Comput. Appl. Softw.*, vol. 28, no. 8, pp. 120–122, 2011.
[21] S. J. Roberts, W. Penny, and D. Pillot, "Novelty, confidence & errors in connectionist systems," in *Proc. IEE Colloq. Intell. Sensors*, Sep. 1996, pp. 1–6.
[22] Z. Xie, X. Li, and W. Wu, "An improved outlier detection algorithm to medical insurance," in *Proc. 17th Int. Conf. Intell. Data Eng. Automated Learn.*, Oct. 2016, pp. 436–444.
[23] C. Sun, Y. Shi, and Q. Li, "A hybrid approach for detecting fraudulent medical insurance claims," in *Proc. 15th Int. Conf. Auto. Agents Multiagent Syst.*, May 2016, pp. 1287–1288.
[24] Y. Gao, C. Sun, R. Li, Q. Li, L. Cui, and B. Gong, "An efficient fraud identification method combining manifold learning and outliers detection in mobile healthcare services," *IEEE Access*, vol. 6, pp. 60059–60068, 2018.
[25] L. G. Moyano, "GraPhys: Understanding health care insurance data through graph analytics," in *Proc. Int. World Wide Web Conf. Steering Committee*, 2016, pp. 227–230.
[26] Y. Shi, C. Sun, and Q. Li, "A fraud resilient medical insurance claim system," in *Proc. 30th AAAI Conf. Artific. Intellig.*, Feb. 2018, pp. 4393–4394.
[27] R. A. Bauder and T. M. Khoshgoftaar, "A novel method for fraudulent medicare claims detection from expected payment deviations (application paper)," in *Proc. IEEE 17th Int. Conf. Inf. Reuse Integr. (IRI)*, Jul. 2016, pp. 11–19.
[28] C. Sun, Z. Yan, Q. Li, Y. Zheng, X. Lu, and L. Cui, "Abnormal group-based joint medical fraud detection," *IEEE Access*, vol. 7, pp. 13589–13596, 2019.
[29] C. Sun, Q. Li, H. Li, Y. Shi, S. Zhang, and W. Guo, "Patient cluster divergence based healthcare insurance fraudster detection," *IEEE Access*, vol. 7, pp. 14162–14170, 2019.
[30] H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," *IEEE Access*, vol. 2, pp. 652–687, 2014.
[31] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, 2008.
[32] R. Harpaz, H. S. Chase, and C. Friedman, "Mining multi-item drug adverse effect associations in spontaneous reporting systems," *BMC Bioinf.*, vol. 11, no. S9, pp. 1–8, Oct. 2010.
[33] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proc. 20th Int. Conf. Very Large Data Bases*, Sep. 1994, pp. 487–499.
[34] M. J. Zaki, S. Parthasarathy, M. Ogihara, and W. Li, "New algorithms for fast discovery of association rules," in *Proc. 3rd Int. Conf. Knowl. Discovery Data Mining*, 1997, pp. 283–286.
[35] X. Li, H. Cao, E. Chen, H. Xiong, and J. Tian, "BP-growth: Searching strategies for efficient behavior pattern mining," in *Proc. IEEE 13th Int. Conf. Mobile Data Manage.*, Jul. 2012, pp. 238–247.

SHENGYAO ZHOU received the B.S. degree from Yanshan University, Qinhuangdao, China. He is currently pursuing the Ph.D. degree with the School of Economics and Management, Beijing Jiaotong University, Beijing, China. His current research interests include machine learning, big data analysis, and medical informationization.

JIE HE was born in Longchang, Sichuan, China, in 1976. He is currently pursuing the Ph.D. degree with Tianjin University. He is also a Senior Engineer at CETC Big Data Research Institute Company Ltd. His main research interest includes the application of big data in government affairs.

HUI YANG was born in Sichuan, China, in 1969. He received the M.S. degree from Sichuan University. He is currently a Senior Engineer at CETC Big Data Research Institute Company Ltd. His main research interest includes the application of big data.

DONGHUA CHEN (Graduate Student Member, IEEE) received the B.S. and M.S. degrees from the Department of Industrial Engineering, Beijing Jiaotong University, Beijing, China, in 2013 and 2015, respectively, where he is currently pursuing the Ph.D. degree with the Department of Information Management, School of Economics and Management.

His research interests include medical informatics, consumer health informatics, machine learning techniques, and big data analysis.

RUNTONG ZHANG (Senior Member, IEEE) was born in Chaoyang, Liaoning, China, in November 1963. He received the B.S. degree in computer science and automation from Dalian Maritime University, China, in 1985, and the Ph.D. degree in production engineering and management from the Technical University of Crete, Greece, in 1996.

He was also with the Swedish Institute of Computer Science, as a Senior Researcher, and the Port of Tianjin Authority, as an Engineer. He is currently a Professor and the Head of the Department of Information Management, Beijing Jiaotong University, China. He has published over 300 articles in referenced journals and conferences and 40 books. He has been a PI for more than 100 research projects and holds nine patents. His current research interests include big data, health-care management, operations research, and artificial intelligence. He has been the general chair or co-chair for over ten IEEE sponsored international conferences.