Hindawi
International Journal of Intelligent Systems
Volume 2023, Article ID 3660151, 25 pages
https://doi.org/10.1155/2023/3660151

Research Article

An Optimized Association Rules Mining Framework for Chinese Social Insurance Fund Data Auditing

Wu Xiuguo and Du Shengyong

School of Management Science and Engineering, Shandong University of Finance and Economics, Jinan 250014, China

Correspondence should be addressed to Wu Xiuguo; xiuguow@sdufe.edu.cn

Received 27 April 2023; Revised 7 September 2023; Accepted 6 October 2023; Published 20 October 2023

Academic Editor: Frederick E. Petry

Association rules mining with the Chinese social insurance fund dataset can effectively discover different kinds of errors, irregularities, and illegal acts by providing auditors with relationships among the items and therefore improve auditing quality and efficiency. However, traditional positive and negative association rules (PNARs) mining algorithms inevitably produce too many meaningless or contradictory rules when these two types of rules are mined simultaneously, which brings a huge challenge to auditors retrieving decision information. Aimed to reduce the quantity of low-reliability PNARs without missing interesting rules, this paper first proposes an improved PNARs mining algorithm with minimum correlation and triple confidence threshold to control the mined rules number by narrowing the range of confidence settings. Then, a novel pruning algorithm based on the inclusion relation of the rule's antecedent and consequent is given to remove those redundant rules. After that, the proposed optimized PNARs mining approach is applied to the Chinese social insurance fund dataset starting with audit features influence factors mining using the Hash table. The experimental results with different datasets show that the proposed framework not only can ensure effective and interesting rules extraction but also has better performance than traditional approaches in both accuracy and efficiency, reducing the number of redundant PNARs by over 70.1% with experimental datasets and average 78.5% with auditing data mining, respectively.

1. Introduction

As has been shown that auditing plays a paramount role in the process of realizing effective national governance, ensuring healthy and scientific social and economic development around the world [1, 2]. In the era of big data, most governments and companies have accumulated a large amount of management and transaction data with the development of information technology. Similarly, heterogeneous and various types of big data are creating serious troubles in the auditing field since it is extremely difficult and cumbersome to manually find accounting irregularities and financial fraud information from the data itself at the surface level [3, 4]. As an example, in 2022, nearly twenty-three million declarations, payments, treatments, and other records were submitted to the Social Insurance Fund Information Centre of a province in China every month, containing abundant information along with errors, irregularities, and illegal acts [5]. For auditors, it is a challenging task to identify those abnormal economic behaviors from the huge amount of data currently. Furthermore, traditional audit methods only analyze data accuracy and integrity without discovering the hidden relationship between variables, resulting in low quality and efficiency in auditing [6, 7]. With the change of objectives, tasks, emphases, and modes of national governance, the strategies and methods for auditing should also change accordingly, therefore facilitating fraud clues finding [8, 9]. Many studies [6, 8, 9] have proposed several association rule mining methods aimed at exploring the relationships among fields in audit databases, helping the auditors' decision-making with accurate information, including not only positive association rules (PARs) but also negative association rules (NARs), simultaneously. In particular, those NARs provide richer and more valuable information for the auditors compared to PARs for the reason that the auditors' main task is to find fraud clues from large-scale datasets. Nevertheless, it may inevitably produce a large number of irrelevant rules while omitting some interesting ones during the mining process. Furthermore, some researchers attempted to propose solutions that can reduce the number of PARs [10–13]. Still, these approaches are difficult to limit the number of low-reliability rules and easy to miss some interesting rules because of the support-confidence framework [14].

Particularly, the existing association rules mining methods do not work well when mining PARs and NARs simultaneously with Chinese social insurance fund data, and the reasons can be summarized as follows: (1) almost similar data frequently appear in audit databases; for example, the province's social insurance fund data of last month is almost the same as those of current month; (2) the attribute types in the auditing dataset have a variety of sorts, including not only numeric data with decimal but also textual data; (3) part of values in the datasets are described using abbreviations or typos, making the data lack standardization. Additionally, only a few studies have involved the association rules mining in the field of big data auditing until now [15, 16]. How to control the number of mined positive and negative association rules (PNARs) according to the features of the audit data has become an increasingly imminent problem to be solved.

Based on the above analysis, this paper intends to develop an optimized positive and negative association rules mining framework for Chinese social insurance fund data to promote the process of standardizing audit. The main idea is first to propose the triple confidence thresholds setting method according to the confidence values changes with items support degree, then present a PNARs mining algorithm with the minimum correlation and triple confidence, and is followed by an association rules pruning algorithm to remove those worthless rules using the inclusion relation of rules' antecedent and consequent. After that, we investigate and give audit features influence factors mining based on the Hash table and present its application in Chinese social insurance fund data auditing in the end. The main contributions of this paper can be summarized as follows:

(i) An optimized positive and negative association rules mining algorithm based on the minimum correlation and triple confidence (PNARs_M) is proposed, which can not only mine strong PNARs but also control the number of rules with various types flexibly.

(ii) A novel redundant association rules pruning algorithm based on the inclusion relation (PNARs_P) is given to further remove those meaningless rules, which has not been involved in previous studies to the best of our knowledge.

(iii) An audit features influence factors mining algorithm based on a Hash table (AFIFM_H) is proposed, which reduces the potential collisions and improves the mining efficiency.

(iv) These algorithms are applied to the Chinese social insurance fund auditing dataset, and the results indicate the better feasibility and effectiveness of the proposed approach by reducing the number of rules without missing interesting ones.

The rest of the paper is arranged as follows. The next section is devoted to a literature review of related PNARs mining techniques and their application in the audit field and is followed by some basic concepts and existing research results about confidence among four types of association rules in Section 3. Section 4 proposes the main framework and methodology for the optimized PNARs mining strategy. Then, Section 5 details the proposed framework application in Chinese audit dataset analysis. The final section presents conclusions and discusses emerging directions for future research.

2. Literature Review

In this section, we will discuss the related PNARs mining techniques and then present their applications in the audit field.

2.1. PNARs Mining Techniques.

As an important technique in data mining, association rule analysis can extract implicit relations among data items that occur frequently together in many fields [13, 15–18]. Agrawal and Imieliński [19] first proposed the Apriori algorithm by iteratively generating candidate itemsets in a database in 1993. Han et al. [20] developed an efficient FP-tree algorithm for mining the complete set of frequent patterns by pattern fragment growth in 2004. However, sometimes decision-makers pay more attention to items that occur infrequently but are strongly correlated. Therefore, negative association rule mining is getting more and more popular among researchers, whose concept was first mentioned by Brin and Motwani [21]. The general forms of negative rules are of $A \longrightarrow \neg B$, $\neg A \longrightarrow B$, or $\neg A \longrightarrow \neg B$, where the entire antecedent or consequent is either a conjunction of negated attributes or a conjunction of nonnegated attributes [21–23]. The work in reference [21] incorporated frequent itemsets with domain knowledge in the form of a taxonomy to mine negative association rules. Shaheen and Abdullah developed a series of algorithms for different fields, such as exploring positive and negative context-based association rules for conventional/characteristic data [24, 25], and mining context-based association rules on microbial databases to extract interesting and useful associations of microbial attributes with existence of hydrocarbon reserve [26–29]. It should be noted that some contradictory rules may be mined when positive and negative rules are mined simultaneously, such as $A \Rightarrow B$ and $A \Rightarrow \neg B$ are both strong rules [30–32]. In reference [10], an improved PNARs mining algorithm is proposed by removing those contradictory association rules from the candidate rules using a correlation test and dual confidence. However, it is hard to generalize as it is domain-dependent and needs a predefined taxonomy. To solve this problem, the itemsets' correlation was introduced to exclude those meaningless rules and a positive rule is discovered if the correlation is positive; a negative rule is discovered when the correlation is negative. Therefore, the traditional association rules mining paradigm is extended to a correlation-support-confidence framework. Furthermore, in the mining process, it is easy to calculate the support and confidence with positive rules, while it is difficult to obtain these values with negative rules because of its complex calculations. The works in references [33, 34] employed chi-square to measure the correlation, whose tests rely on the normal approximation to the binomial distribution (more precisely, to the hypergeometric distribution). For all that, this approximation breaks down when the expected values are small. In addition, seven correlation measuring methods have been compared and analyzed through providing their connections and differences by Wu et al. [35].

Nevertheless, the reduction of invalid association rules quantity has still not been well solved for the reason that most methods' searching space of negative association rules is all infrequent itemsets, and the number of candidate frequent itemsets is so great if the support degree is used as only one constraint. For massive data, it is meaningless and impossible to count all infrequent itemsets using traditional approaches. In this way, more and more researchers have realized the importance of improving the performance and efficiency of PNARs mining these days. They have begun to concentrate on the invalid rule number reduction with multiple confidence or support thresholds. The work by Cardoni et al. [4] attempted to use multiple confidences for mining positive and negative association rules. Although the test is effective, it still has some limitations in that, it cannot provide enough information about the strength of the relationship. Multiple support degrees are also applied to PNARs mining to improve the interesting rules mining efficiency in reference [36]. Based on the previous works, Bemarisika and Totohasina [37] proposed a two-level confidence threshold-setting method for positive and negative association rules mining to limit the number of frequent and infrequent items. Also, four confidences are introduced to solve the problem that sole confidence usually results in plenty of useless rules in reference [38]. However, these studies neither presented threshold settings ways nor considered the internal constraints among multiple confidences. To solve these problems, researchers in reference [39] gave a double confidence approach for $A \Rightarrow B$, $\neg A \Rightarrow \neg B$, $\neg A \Rightarrow B$, and $A \Rightarrow \neg B$, respectively. Still, it is difficult to effectively control the number of lower confidence rules, and some interesting association rules are missed.

2.2. Association Rules Mining in the Field of Audit. With the rapid development of data mining, association rule mining more and more frequently appears in the audit field in recent years, helping auditors quickly find anything illegal from the audit database. In detail, there are two main types of association rules applications. The first one is network security information system auditing, which is used to protect the security of the audit systems by preventing illegal data intrusions, as can be seen in references [40, 41]. Parkinson et al. proposed a novel method of modelling file system permissions by association rule mining techniques to identify irregular permissions [40]. The second type is data auditing by association rules technology, which is used to verify the integrity and correctness itself. For example, Estrada proposed a new approach toward electronic data processing (EDP) auditing, called electronic auditing (EA), and constructed an infrastructure with the support of emerging technologies so that some of the audit work can be performed electronically and automatically [42]. The work by Sahu and Gmz [39] used data mining for credit card audits in conjunction with the evaluation of the design and effectiveness of internal controls intended to prevent fraud, waste, and abuse in these programs. An innovative fraud detection method was proposed by Singh et al. [43], and they built upon existing fraud detection research and minority reports to deal with the problem of skewed data distributions in data mining. It should be pointed out that auditing business data has high similarity and values characteristics caused by regulations. Aimed at the features of audit data, Dan [44] presented an audit data mining algorithm. Shang et al. [8] gave an audit data mining quality optimization strategy to minimize the possibility of noncoherent data. Reference [45] introduced an audit model framework based on data mining for finding suspicious data from audit data. Djenouri et al. [46] suggested that the audit process be structured in several steps and different data mining algorithms used in each subprocess. In reference [47], a conceptual framework of an improved association algorithm (CFiAA) and its application in audit data mining was proposed. Seong and Lee calculated the importance value of the vocabulary used in the audit report based on machine learning rather than the qualitative research method to improve the audit quality [48]. Zhang et al. presented a correlation analysis algorithm to reveal original characteristics and internal connections in auditing data, and the results demonstrated the validity and effectiveness [49].

2.3. Review. According to the analysis above, although abundant achievements have been obtained in the theory and application of association rules mining in recent years, there are still unsolved problems with the current association rules mining method, particularly in the field of audit. They can be summarized as follows:

(1) Despite the mining efficiency improvement, the rules number of PNARs still have not been well controlled in most mining approaches; some valuable and important rules are still difficult to identify.

(2) Although there has been an increased focus on the application of PNARs in the field of audit, most current approaches do not work well enough owing to the unique features of social auditing data.

(3) Rarely studies have involved audit features influence factors mining, which can effectively identify valuable rules, thereby improving the fraud clues finding efficiency.

In fact, the four confidences of the PNARs are intrinsically correlative, and it is not necessary to use the four confidences threshold according to the actual execution cost of tasks. On the other side, measure standard is another factor that needs to further study for negative association rules, while many achievements have been made in positive rules mining, such as support, confidence, and correlation. As a result, this paper aims to propose an optimized PNARs mining framework from two points according to the intrinsic attributes of four types rules' confidences: a novel PNARs minimum confidences threshold setting method and an efficient PNARs pruning strategy. The proposed approach makes full use of the four types of rules' confidences changing regularity and items inclusion relation, offering greater clarity and feasibility for the confidences thresholds settings. In addition, this study proposes a minimal hashing and pruning algorithm to reduce the potential collisions and improve the mining efficiency.

3. Preliminaries

Suppose $I = \{i_1, i_2, \ldots, i_n\}$ be a set of items and $\mathrm{TD} = \{t_1, t_2, \ldots, t_m\}$ be a set of m transactions over I, where each transaction t_i is a subset of items such that $t_i \subseteq I$. Each transaction is associated with a unique identifier TID. Formally, a positive association rule is an implication of the form "$A \Rightarrow B$", where $A \subseteq I$; $B \subseteq I$, and $A \cap B = \emptyset$.

The rule $A \Rightarrow B$ has support s in the transaction set TD if $(100 * s)\%$ of the transactions in TD contain $A \cup B$, written as $sup(A \Rightarrow B) = s$. In other words, the support of the rule is the probability that A and B hold together among all the possible presented cases. It is said that the rule $A \Rightarrow B$ holds in the transaction set TD with confidence c if $(100 * \mathrm{c})\%$ of transactions in TD that contain A also contain B, written as $conf(A \Rightarrow B) = c$. In other words, the confidence of the rule is the conditional probability that the consequent B is true under the condition of the antecedent A. The problem of discovering all association rules from a set of transactions TD consists of generating the rules that have a support degree and a confidence degree greater than users' given thresholds. These rules are called strong rules, and the framework is known as the support-confidence framework for association rule mining.

The rules of the other three forms $A \Rightarrow \neg B$, $\neg A \Rightarrow B$, and $\neg A \Rightarrow \neg B$ are referred to as negative associations between itemsets. In contrast to positive rules, a negative rule encapsulates the relationship between the occurrences of one set of items with the absence of the other set of items. The rule $A \Rightarrow \neg B$ has support s in the transaction set if $(100 * s)\%$ of transactions in TD contain A but do not contain B. The support of a negative association rule, $sup(\neg A \Rightarrow B)$, is the frequency of transactions occurrence with B in the absence of A. $sup(\neg A \Rightarrow \neg B)$ is the number of transactions in TD neither containing A nor B to the number of all transactions. The rule $A \Rightarrow \neg B$ holds in the given dataset (database) with confidence $c\%$ representing that $(100 * c)\%$ of transactions contain A but do not contain B, written as $conf(A \Rightarrow \neg B) = P(A \cup \neg B)/P(A)$, where $P(X)$ is the probability function. In the same way, the rule $\neg A \Rightarrow B$ with confidence $c\%$ describes that $(100 * c)\%$ of transactions not containing A but B, written as $conf(\neg A \Rightarrow B) = P(\neg A \cup B)/P(\neg A)$, and the rule $\neg A \Rightarrow \neg B$ with confidence $c\%$ describes that $(100 * c)\%$ of transactions neither contain A nor contain B, written as $conf(\neg A \Rightarrow \neg B) = P(\neg A \cup \neg B)/P(\neg A)$. The support and confidence of itemsets are calculated during iterations. However, it is difficult to count the support and confidence of nonexisting items in transactions. The relations among these rules can be described as follows [33]:

Theorem 1. *Suppose an itemset $A \subset I$, an itemset $B \subset I$, and $A \cap B = \emptyset$, then*

(1) $sup(\neg A) = 1 - sup(A)$;

(2) $sup(\neg A \cup B) = sup(B) - sup(A \cup B)$;

(3) $sup(A \cup \neg B) = sup(A) - sup(A \cup B)$;

(4) $sup(A \cup \neg B) = 1 - sup(A) - sup(B) + sup(A \cup B)$.

In addition, the confidence of these negative rules can be calculated using the following theorem.

Theorem 2. *Suppose itemset $A \subset I$, itemset $B \subset I$, and $A \cap B = \emptyset$, then*

(1) $conf(A \Rightarrow B) = (sup(\mathrm{A}) - sup(A \cup B))/(sup(A))$;

(2) $conf(A \Rightarrow B) = (sup(B) - sup(A \cup B))/(1 - sup(A))$;

(3) $conf(A \Rightarrow B) = (1 - sup(A) - sup(B) + sup\ (A \cup B))/(1 - sup(A))$.

There also exists a confidence constraint relation among these four types of association rules, as can be seen in Theorem 3.

Theorem 3. *Let itemset $A \subseteq I$; itemset $B \subseteq I$, and $A \cap B = \emptyset$, conf(X) is the confidence degree of rule X, then*

(1) $conf(A \Rightarrow B) + conf(A \Rightarrow \neg B) = 1$;

(2) $conf(\neg A \Rightarrow B) + conf(\neg A \Rightarrow \neg B) = 1$.

For Theorem 3, it is easy to prove using the above formulae in Theorems 1 and Theorem 2. Theorem 3 shows that there is a complementary relationship between the confidences of these four types of rules with the same antecedent.

Based on these constraint relations, the study about rules' confidence can help us set the confidence threshold rationally and therefore improve mining efficiency and quality. In this way, the values of confidence range for four types rules based on items support degree can be seen in Theorem 4.

Theorem 4. *Let itemset $A \subseteq I$; itemset $B \subseteq I$, and $A \cap B = \emptyset$, sup(X) and conf(X) are the support and confidence degrees of rule X, respectively. Max(x, y) and min(x, y) are maximum and minimum value functions, respectively. Then, the confidence value range of four types of association rules can be described as follows:*

(1) $max(0,\ (sup(A) + sup\ (B) - 1)/(sup(A)) \leq conf\ (A \Rightarrow B) \leq min(1,\ (sup(B)/sup(A)))$;

(2) $1 - min(1,\ (sup(B)/sup(A)) \leq conf(A \Rightarrow \neg B) \leq 1 - max(0,\ (sup(A) + sup(B) - 1)/(sup(A)))$;

(3) *max(0, $(sup(B) - sup(A))/(1 - sup(A))) \leq$ conf(¬A ⇒ B) min(1, $(sup(B))/(1 - sup(A)))$;*

(4) *1 − min(1, $(sup(B) - sup\ (A))/(1 - sup(A))) \leq$ conf (¬A ⇒ ¬B) ≤ 1 − max(0, $(sup(B) - sup(A))/(1-sup\ (A)))$.*

From Theorem 4, it is clear that the rule's confidence range is closely related to high or low itemset support degree. Obviously, only one-level confidence threshold does not work well without taking into account the confidence constraint relation of these four types of PNARs. In this way, we try to further narrow the range of confidence in order to set a reasonable and appropriate confidence threshold, not only mining the interesting rules but also excluding those meaningless rules in social insurance auditing data application.

4. An Optimized PNARs Mining Framework with Correlation and Triple Confidence Thresholds

In this section, we will present an optimized PNARs mining framework using minimum correction coefficient and triple confidence thresholds, which can mine strong PNARs with reasonable rules' quantity and flexibility.

4.1. Correction Coefficient. The support-confidence framework is the most popular approach used to mine positive rules up to now. However, it usually results in some self-contradictory rules when mining negative rules directly. For example, suppose there are 10,000 goods, A is used to describe transactions of buying X; B is used to describe transactions of buying Y; $A \cup B$ is transactions for buying both X and Y. Transactions details can be seen in Table 1.

Table 1: Transactions of buying A and B.

Items	A	$\neg A$	$\sum_{\text{row}}$
B	2500	2500	5000
$\neg B$	3500	1500	5000
$\sum$col	6000	4000	10000

Suppose the user-specified minimum support and minimum confidence thresholds are set as $min\_sup = 0.2$, $min\_conf = 0.3$, then

$sup(A \Rightarrow B) = 2{,}500/10{,}000 = 0.25 > min\_sup$;
$conf(A \Rightarrow B) = 2{,}500/6{,}000 = 0.42 > min\_conf.$

So, $A \Rightarrow B$ is a strong positive association rule. On the other hand, $A \Rightarrow \neg B$ is also a strong negative association rule for the reason that

$sup(A \Rightarrow \neg B) = 3{,}500/10{,}000 = 0.35 > min\_sup$, and
$conf(A \Rightarrow \neg B) = 3{,}500/6{,}000 = 0.58 > min\_conf.$

However, both of these two strong association rules are contradictory and meaningless. In addition, $A \Rightarrow \neg B$ is more reliable because the value of $conf(A \Rightarrow \neg B)$ is larger than the value of $conf(A \Rightarrow B)$, meaning that A and B are negatively correlated. It is obvious that the minimum support and confidence thresholds usually lead to contradictory rule extraction.

Generally, the correlation coefficient measures the strength and direction of the linear relationship between a pair of random variables. It is also known as the covariance between two variables divided by their standard deviation (σ), defined as follows:

$$\operatorname{corr}(A, B) = \frac{\operatorname{cov}(A, B)}{\sigma_A \sigma_B}, \tag{1}$$

where cov(A, B) is the covariance of the two variables and σ represents the standard deviation [50, 51]. This definition is available for positive attributes but does not specify those negative attributes. In this study, we use the correlation coefficient to measure the strength of the correlation of itemsets A and B in the four types of rules $A \Rightarrow B$, $A \Rightarrow \neg B$, $\neg A \Rightarrow B$, and $\neg A \Rightarrow \neg B$.

Definition 5 (Correlation degree). The correlation degree of itemset A and itemset B is defined as

$corr(A,\ B) = sup(A \cup B) - sup(A) \times sup(B) = P(A \cup B) - P(A) \times P(B)$, where

(1) $corr(A,\ B)$ is the itemset correlation degree

(2) $sup(A \cup B)$, $sup(A)$, and $sup(B)$ are support degrees of itemsets $A \cup B$, A and B, respectively

(3) $P(A \cup B)$, $P(A)$, and $P(B)$ are occurrence frequencies of itemsets $A \cup B$, A, and B, respectively

The value of $corr(A,\ B)$ represents the correlation strength of itemset A and itemset B. Although the value of the correlation degree of itemset A and itemset B is uncertain, there are three types of possible values' ranges:

(1) $corr(A, B) > 0$, means that itemset A and itemset B are positively correlated

(2) $corr(A, B) < 0$, means that itemset A and itemset B are negatively correlated

(3) $corr(A, B) = 0$, means that itemset A and itemset B are independent

The correlation degree provides the positive and negative correlation relation between itemsets A and B, indicating that if the occurrence of itemset A increases (decreases), then itemset B will increase (decrease) correspondingly. If the correlation degree is equal to 0, it means that they are independent and can be ignored in the future. Therefore, it is important to set the value of the minimum correlation threshold.

According to references in references [38, 52], if a rule $A \Rightarrow B$ satisfies $sup(A \cup B) - sup(A) \times sup(B) < \varepsilon$ $(\varepsilon = 10^{-6})$, then itemsets A and B would be regarded as independent. Accordingly, we introduce a user-specified minimum correlation threshold min_corr, and only those itemsets $sup(A \cup B) - sup(A) \times sup(B) \geq min\_corr$ will be mined and regarded as interesting rules. On the other hand, the value of $sup(A \cup B) - sup(A) \times sup(B)$ is lower than zero sometimes; hence, we use their absolute value |

$sup(A \cup B) - sup(A) \times sup(B)| \geq min\_corr$ when rule $A \Rightarrow B$ is interesting.

Theorem 6. *Suppose itemset $A \subset I$, itemset $B \subset I$, $A \cap B = \emptyset$, min_corr is the minimum correlation threshold. For a positive rule $A \Rightarrow B$, if $corr(A, B) \geq min\_corr$, then itemsets' correlation in its corresponding negative rules satisfies $corr(\neg A, B) \geq min\_corr$, $corr(A, \neg B) \geq min\_corr$, and $corr(\neg A, \neg B) \geq min\_corr$. Formally, as can be described as if $|sup(A \Rightarrow B) - sup(A) \times sup(B)| \geq min\_corr$, then*

(1) $|sup(\neg A \Rightarrow B) - sup(\neg A) \times sup(B)| \geq min\_corr$;

(2) $|sup(A \Rightarrow \neg B) - sup(A) \times sup(\neg B)| \geq min\_corr$;

(3) $|sup(\neg A \Rightarrow \neg B) - sup(\neg A) \times sup(\neg B)| \geq min\_corr$.

Proof. (1) It is known that $|sup(A \Rightarrow B) - sup(A) \times sup(B)| \geq min\_corr$ since $corr(A, B) \geq min\_corr$ for positive rule $A \Rightarrow B$. Then,

$|sup(\neg A \Rightarrow B) - sup(\neg A) \times sup(B)| = |sup(B) - sup(A \cup B) - (1 - sup(A)) \times sup(B))| = |-sup(A \cup B) + sup(A) \times sup(B)| = |sup(A \Rightarrow B) - sup(A) \times sup(B)| \geq min\_corr$.

The proof is complete.

(2) and (3) can be proved in the same way as (1).

Theorem 6 shows that the itemset correlation in negative rules is not necessary to calculate specifically and can be obtained by its positive itemset. It also tells us a fact that these four types of association rules can be filtered only with one appropriate minimum correlation degree at the same time. □

4.2. Triple Confidence Threshold. Infrequent itemsets are also significant for decision-makers, especially for auditors, because they can imply relevant suspicious behaviors in social insurance fund data auditing. According to the traditional support-confidence framework, $sup(A \Rightarrow B)$ must be a small value if $sup(A)$ and $sup(B)$ are small, while $conf(A \Rightarrow B)$ is uncertain for positive rule $A \Rightarrow B$. However, for the negative rule $\neg A \Rightarrow \neg B$, it is obvious that $conf(\neg A \Rightarrow \neg B)$ must be a large value for the reason that $conf(\neg A \Rightarrow \neg B) = 1 - (sup(B) - sup(A \cup B))/(1 - sup(A))$, while $sup(B) - sup(A \cup B)$ is a small value closing to zero. Such cases may usually happen when only one confidence threshold is used: if the predefined confidence threshold value is small, too many worthless rules would be generated; on the other hand, many important rules would be missed if the predefined confidence threshold value is too large. In practice, the approaches by Yu et al. [9] and Shen et al. [53] have not effectively controlled the number of association rules with double confidence thresholds under the correlation-support-confidence framework. Although Dong et al. [33] have presented multiple confidences, they just took into account such two cases: $sup(A) = sup(B) = 0.1$ and $sup(A) = sup(B) = 0.9$. Based on Theorem 4, we will discuss the initial confidence values by analyzing the relationship between support and confidence for those positive and negative association rules.

According to confidence value ranges in previous studies [52–54], we transform and simplify them for convenience of calculations by uniformly setting the left bound with a maximum function and the right bound with a minimum function, as can be seen in the following:

(1) $\max(0,\ 1 - sup(B)/sup(A)) \leq conf(A \Rightarrow \neg B) \leq \min(1, (1 - sup(B))/sup(A))$;

(2) $\max(0, (1 - sup(A) - sup(B))/(1 - sup(A))) \leq conf(\neg A \Rightarrow \neg B) \leq \min(1, 1 - sup(B))/(1 - sup(A)))$,

Here, functions $\max(x, y)$ and $\min(x, y)$ return the bigger and smaller values of x and y, respectively.

However, the range of confidence value is still too large, and it is necessary to further narrow the range of rules' confidence according to the relation of support and confidence. In this way, we divide into the following four cases and analyze them for rules $A \Rightarrow B$, $A \Rightarrow \neg B$, $\neg A \Rightarrow B$, and $\neg A \Rightarrow \neg B$.

Case 1: $sup(A) \leq sup(B)$ and $sup(A) + sup(\mathrm{B}) \leq 1$;

Case 2: $sup(A) > sup(B)$ and $sup(A) + sup(\mathrm{B}) \leq 1$;

Case 3: $sup(A) \leq sup(B)$ and $sup(A) + sup(B) > 1$;

Case 4: $sup(A) > sup(B)$ and $sup(A) + sup(B) > 1$.

For Case 1, the left bound of $conf(A \Rightarrow B)$ is $\max(0, (sup(A) + sup(B) - 1)/sup(A))) = 0$, for the reason that $sup(A) + sup(B) \leq 1$; and the right bound is $\min(1, sup(B)/sup(A)) = 1$ since $sup(A) \leq sup(B)$. So, the confidence value of $A \Rightarrow B$ ranges from 0 to 1, written as $conf(A \Rightarrow B) \in [0, 1]$.

Similarly, those four types of rules' confidence can be calculated with the other cases, and the results can be seen in Table 2.

Based on the conclusions in Table 2, we will further discuss the confidence values with different support degrees in detail based on the proposed triple confidence threshold settings. In order to show the whole process more clearly, Table 3 provides the examples of confidence range of four types rule with different values of $sup(A)$ and $sup(B)$.

(1) Itemset A is positively correlated with itemset B. The rules $A \Rightarrow B$ and $\neg A \Rightarrow \neg B$ are valid in such a situation. Also, the values of $sup(A)$ and $sup(B)$ have little difference for the reason of correlation constraint. We use a positive value closing to zero $\varepsilon \longrightarrow 0$ representing their difference, that is, $\varepsilon = |sup(A) - sup(B)|$. It is mainly divided into the following two categories:

(i) Both of $sup(A)$ and $sup(B)$ are smaller values, satisfying $sup(A) + sup(B) \leq 1$. At this point, for case 1, $conf(A \Rightarrow B) \in [0, 1]$; $sup(B)/(1 - sup(A))$ is a very small value, then $conf(\neg A \Rightarrow \neg B) \in [1 - \varepsilon, 1]$, showing that its left bound is relatively high. For case 2, $conf(A \Rightarrow B) \in [0,\ 1 - \varepsilon]$; $conf(\neg A \Rightarrow \neg B) \in [1 - \varepsilon, 1]$. In this way, the left bound of $conf(\neg A \Rightarrow \neg B)$ is a high value.

(ii) Both of $sup(A)$ and $sup(B)$ are bigger values, satisfying $sup(A) + sup(B) > 1$. In this situation, the left bound of $conf(A \Rightarrow B)$ is the same for case 3 and case 4, their left bounds values are $(1 + (sup(B) - 1)/sup(A)) = 1 + sup(B)/sup(A) - (1/sup(A)) \longrightarrow 2 - (1/sup(A))$. Since $sup(A)$ is a bigger value and $sup(A) \leq 1$,

Table 2: Confidence ranges of rule with different support degrees.

Case	$conf(A \Rightarrow B)$	$conf(A \Rightarrow \neg B)$	$conf(\neg \mathrm{A} \Rightarrow \mathrm{B})$	$conf(\neg \mathrm{A} \Rightarrow \neg \mathrm{B})$
Case 1	$[0, 1]$	$[0, 1]$	$[(\sup(B) - \sup(A))/(1 - \sup(A)), (\sup(B))/(1 - \sup(A))]$	$[1 - (\sup(B))/(1 - \sup(A)), (1 - \sup(B))/(1 - \sup(A))]$
Case 2	$[0, (\sup(B)/\sup(A))]$	$[1 - (\sup(B)/\sup(A)), 1]$	$[0, (\sup(B))/(1 - \sup(A))]$	$[1 - (\sup(B))/(1 - \sup(A)), 1]$
Case 3	$[1 + (\sup(B) - 1)/(\sup(A)), 1]$	$[0, (\sup(B) - \sup(A))/(\sup(A))]$	$[(\sup(B) - \sup(A))/(1 - \sup(A)), 1]$	$[0, (1 - \sup(B))/(1 - \sup(A))]$
Case 4	$[1 + (\sup(B) - 1)/(\sup(A)), (\sup(B)/\sup(A))]$	$[1 - (\sup(B)/\sup(A)), (1 - \sup(B))/(\sup(A))]$	$[0, 1]$	$[0, 1]$

then $1/sup(A) \longrightarrow 1^+$, $2 - (1/sup(A)) \longrightarrow 1^-$. In this way, $conf(A \Rightarrow B)$ is a high value while $conf(\neg A \Rightarrow \neg B) \in [0, 1]$ in this case.

Consequently, when itemset A is positively correlated with itemset B, $conf(\neg A \Rightarrow \neg B)$ is a bigger value if both $sup(A)$ and $sup(B)$ are smaller values, and the fourth column and seventh column in rows 1 and 2 in Table 3 describe such a situation. Similarly, $conf(A \Rightarrow B)$ is bigger if both $sup(A)$ and $sup(B)$ are bigger values, and the fourth column and seventh column in rows 3 and 4 in Table 3 describe such a situation.

(2) Itemset A is negatively correlated with itemset B. The rules $A \Rightarrow \neg B$ and $\neg A \Rightarrow B$ are valid in such a situation. When $sup(A) + sup(B) \longrightarrow 1$, it is obvious that the following conclusions can be drawn:

(i) $\max(0, (sup(A) + sup(B) - 1)/sup(A)) \longrightarrow 0$;
(ii) $\min(1, sup(B)/sup(A)) \longrightarrow 1$;
(iii) $\max(0, (sup(B) - sup(A))/(1 - sup(A))) \longrightarrow 0$;
(iv) $\min(1, sup(B)/(1 - sup(A))) \longrightarrow 1$.

As can be seen from Table 2, the values of $conf(A \Rightarrow \neg B)$ and $conf(\neg A \Rightarrow B)$ are both between 0 and 1. Furthermore, considering $conf(A \Rightarrow \neg B)$ and $conf(\neg A \Rightarrow B)$, respectively:

(i) For case 1, $conf(A \Rightarrow \neg B) \in [0, 1]$, which is uncertain with different $sup(A)$ and $sup(B)$. For $conf(\neg A \Rightarrow B)$, when $sup(A) + sup(B) \longrightarrow 1$, the right bound of $conf(\neg A \Rightarrow B)$ is $sup(B)/(1 - sup(A)) \longrightarrow 1^-$-; the left bound of $conf(\neg A \Rightarrow B)$ is $\varepsilon/(1 - sup(A))$, which is a monotone increasing function of ε. The $conf(\neg A \Rightarrow B)$ is uncertain when ε is getting smaller; the $conf(\neg A \Rightarrow B)$ is high when ε is getting higher. While $sup(A) + sup(B) \nrightarrow 1$, the range $conf(A \Rightarrow \neg B)$ changes with the difference of $sup(A)$ and $sup(B)$, as the same increases (or decrease) with ε. As can be seen in the sixth column in rows 5 and 6 in Table 3, where such a situation is given with different $sup(A)$ and $sup(B)$.

(ii) For case 2, when the difference of $sup(A)$ and $sup(B)$ is getting smaller, both of the left bound of $conf(A \Rightarrow \neg B)$=$(1 - (sup(B)/sup(A))$ and the right bound of $conf(\neg A \Rightarrow B) = sup(B)/(1 - sup(A))$ are getting smaller, and the value of $conf(A \Rightarrow \neg B)$ is uncertain, while $conf(\neg A \Rightarrow B)$ is lower at this time. In the same way, the left bound of $conf(A \Rightarrow \neg B)$ and the right bound of $conf(\neg A \Rightarrow B)$ get higher as the value of ε increases. In this way, $conf(A \Rightarrow \neg B)$ is higher and $conf(\neg A \Rightarrow B)$ is uncertain. As can be seen in the sixth column in rows 7, 8, and 9 in Table 3, where such a situation is given with different $sup(A)$ and $sup(B)$.

For Case 3 and Case 4, the values changes are the same as those in Case 1 and Case 2, respectively.

Thus, the values of $conf(A \Rightarrow \neg B)$ and $conf(\neg A \Rightarrow B)$ change from 0 to 1 when itemset A and itemset B are negatively correlated and $\mathrm{sup}(A) + sup(B) \longrightarrow 1$, and either $conf(A \Rightarrow \neg B)$ or $conf(\neg A \Rightarrow B)$ synchronously changes with the difference $sup(A)$ and $sup(B)$.

In order to make the best of these relations to control the number of mined association rules, we introduce triple minimum confidence thresholds for PNARs, named *min_conf_P*, *min_conf_NH*, and *min_conf_NL*, respectively.

Definition 7 (Triple confidence thresholds). Let I be a set of items, itemset $A \subseteq I$, itemset $B \subseteq I$, and $A \cap B = \emptyset$, min_corr is the minimum correlation threshold, $\varepsilon_{\min}$ is used to measure support degree difference of itemset A and B, then

(1) *min_conf_P* is the minimum confidence threshold for PARs when itemset A and B are positively correlated. Rule $A \Rightarrow B$ and rule $\neg A \Rightarrow \neg B$ are both regarded as strong association rules if and only if (i) $corr(A, B) \geq min\_corr$, and (ii) $conf(A \Rightarrow B) \geq min\text{-}conf\_P$, $conf(\neg A \Rightarrow \neg B) \geq minconf\_P$, respectively;

(2) *min_conf_NH* is the high minimum confidence threshold for NARs when (i) itemsets A and B are negatively correlated and (ii) $|sup(A) - sup(B)| \geq \varepsilon\mathrm{min}$. Rule $A \Rightarrow \neg B$ and rule $\neg A \Rightarrow B$ are both regarded as strong association rules if and only if (i) $corr(A, B) \leq -min\_corr$ and (ii) $conf(A \Rightarrow \neg B) \geq min\_conf\_NH$; $conf(\neg A \Rightarrow B) \geq min\_conf\_NH$, respectively;

(3) *min_conf_NL* is the low minimum confidence threshold for NARs when (i) itemsets A and B are negatively correlated and (ii) $|sup(A) - sup(B)| < \varepsilon\mathrm{min}$. Rule $A \Rightarrow \neg B$ and rule $\neg A \Rightarrow \neg B$ are both regarded as strong association rules if and only if (i) $corr(A, B) \leq -min\_corr$ and (ii) $conf(A \Rightarrow \neg B) \geq min\_conf\_NL$; $conf(\neg A \Rightarrow B) \geq min\_conf\_NL$, respectively.

As can be seen from the above definition, the constraints $corr(A, B) \geq min\_corr$ and $corr(A, B) \leq -min\_corr$ are mainly used to ensure that the mined results are strongly correlated association rules.

4.3. PNARs Mining Algorithm with Minimum Correlation and Triple Confidence Model.

As discussed in the previous subsection, such conclusions can be drawn as follows:

(i) $conf(\neg A \Rightarrow \neg B)$ is a higher value when itemsets A and B are positively correlated and $sup(A)$ and $sup(B)$ are actually small. In this way, the confidence threshold *min_conf_P* should be set with a higher value in order to reduce the number of such type rules;

(ii) $conf(A \Rightarrow B)$ is a higher value when itemsets A and B are positively correlated, and $sup(A)$ and $sup(B)$ are actually higher. Thus, the confidence threshold *min_conf_P* should be set to be larger to ensure high confidence in such type rules;

(iii) Either the value of $conf(A \Rightarrow \neg B)$ or $conf(\neg A \Rightarrow B)$ is determined by the difference of $sup(A)$ and $sup(B)$ when $sup(A) + sup(B) \longrightarrow 1$. Meanwhile, the smaller of $\varepsilon = |sup(A) - sup(B)|$, the values of $conf(A \Rightarrow \neg B)$ or $conf(\neg A \Rightarrow B)$ may be uncertain; and the confidence threshold *min_conf_NL* should be initialized with a smaller value so that no interesting rules be excluded. On the other hand, the values of

Table 3: An example of confidence range setting with difference *sup*(A) and *sup*(B).

Row	*sup*(A)	*sup*(B)	*conf*($A \Rightarrow B$)	*conf*($A \Rightarrow \neg$B)	*conf*($\neg A \Rightarrow B$)	*conf*($\neg A \Rightarrow \neg B$)
1	0.05	0.05	[0, 1]	[0, 1]	[0, 0.05]	[0.95, 1]
2	0.05	0.10	[0, 1]	[0, 1]	[0, 0.11]	[0.89, 1]
3	0.95	0.85	[0.84, 0.90]	[0.11, 0.16]	[0, 1]	[0, 1]
4	0.95	0.95	[0.95, 1]	[0, 1]	[0, 0.05]	[0, 1]
5	0.30	0.55	[0, 1]	[0, 1]	[0.36, 0.76]	[0.24, 0.64]
6	0.80	0.15	[0, 0.19]	[0.81, 1]	[0, 0.25]	[0.25, 1]
7	0.30	0.70	[0, 1]	[0, 1]	[0.58, 1]	[0, 0.43]
8	0.95	0.05	[0, 0.05]	[0.95, 1]	[0, 1]	[0, 1]
9	0.05	0.95	[0, 1]	[0, 1]	[0.95, 1]	[0, 0.05]

conf($A \Rightarrow \neg B$) or *conf*($\neg A \Rightarrow B$) may be larger with the value of ε increasing. In this way, the confidence threshold *min_conf_NH* should be initialized with a bigger value to exclude those meaningless rules.

In conclusion, the confidence ranges are narrowed down combined with triple confidence thresholds according to the above analysis, as can be seen in Table 4.

In this study, we propose an improved PNARs mining algorithm with triple confidence and minimum correlation coefficient, called PNARs_M, as shown in Algorithm 1. The main idea can be described in more detail as follows:

(i) Setting triple confidence thresholds *min_conf_P*, *min_conf_NH*, and *min_conf_NL*, representing the low confidence thresholds when positively related, the high confidence thresholds when negatively related, and the low confidence thresholds when negatively related, respectively.

(ii) Selecting those itemsets A and B satisfying a minimum correlation degree. If A and B are positively correlated, then $A \Rightarrow B$ and $\neg A \Rightarrow \neg B$ are valid and can be extracted by the *min_conf_P*. On the other hand, if A and B are negatively correlated, then $A \Rightarrow \neg B$ and $\neg A \Rightarrow B$ are valid and can be extracted by the *min_conf_NH* and *min_conf_NL*, respectively. According to the difference between *sup*(A) and *sup*(B), *min_conf_NH* is the high minimum confidence threshold for NARs; *min_conf_NL* is the low minimum confidence threshold for NARs.

The proposed approach combines the user's requirement of high confidence and interesting rules in practice instead of random selection. The PNARs mining algorithm mainly involves two steps: (i) searching all the frequent items that meet the user's requirements from transaction database D and (ii) generating strong positive and negative association rules from frequent items set. Suppose the frequent itemsets have been mined and saved in itemsets L by any existing frequent itemsets mining algorithm (i.e., Apriori or AFIFM_H in 5.2 of this paper). Algorithm 1 is used to extract strong association rules from frequent itemsets L.

In Algorithm 1, line (1) initializes both rule sets PARs and NARs to be empty, and lines (2) to (22) are used to extract all PARs and NARs from L. Correlation of itemsets A and B *corr*(A, B) can be calculated using Definition 5. If *corr*(A, B) $\geq$ *min_corr* and the rule's confidence is greater than its own minimum confidence threshold, the algorithm generates rules like $A \Rightarrow B$ and $\neg A \Rightarrow \neg B$ (line (9) to line (10)). Otherwise, if *corr*(A, B) $\leq -$*min_corr*, the rules like $\neg A \Rightarrow B$ and $A \Rightarrow \neg B$ are generated according to their support difference (line (12) to line (18)). In the end, line (21) returns the results and ends the whole algorithm.

Next, we will analyze the time complexity of the proposed algorithm. Considering a set L with n frequent itemsets and X is the set of 2-frequent itemsets satisfying $A \cup B = X$ and $A \cap B = \emptyset$, then the size of X is C_n^2. Confidence and minimum correlation analyses are necessary to compute with two times, then the total computation is $2 * c_n^2 = n * (n-1)$. In addition, the time complexity of steps (6) to (18) is $O(1)$, then the total time complexity is $O(n^2)$, indicating that the improved approach does not increase extra time consumption.

4.4. Rules Pruning Algorithm for PNARs.

In the above subsection, PNARs mining algorithm has been proposed with a triple confidence and minimum correlation model. Nevertheless, it does not work well occasionally because there are still many redundant rules, and it is necessary to find and prune them using an effective approach. In this subsection, we will present an association rules pruning algorithm based on the inclusion relation of rule's antecedent and consequent.

Theorem 8. *Let* I *be a set of items, itemset* $A \subseteq I$*, itemset* $B \subseteq I$*, and* $A \cap B = \emptyset$*, min_corr is the minimum correlation threshold, rule* $A \Rightarrow B$ *is a strong positive association rule mined using PNARs_M algorithm. If there also exists a valid positive association rule* $A \Rightarrow B'$ *in PARs satisfying* $B' \subseteq B$*, then* $A \Rightarrow B'$ *is a redundant rule of* $A \Rightarrow B$*.*

Proof. According to Definition 7, we need to prove that the following two inequalities are simultaneously satisfied: (i) *conf*($A \Rightarrow B'$) $\geq$ *min_conf_P* and (ii) *corr*(A, B') $\geq$ *min_corr* under the condition that $A \Rightarrow B$ is a strong association rule.

(1) It is obvious that *conf*($A \Rightarrow B$) $\geq$ *min_conf_P* according to Definition 7. We also know that $B' \subseteq B$, then $t(B') \supseteq t(B)$, and $t(A \cup B') \supseteq t(A \cup B)$, where $t(x)$ is the transactions set including itemset x. In this way, $|t(A \cup B')| \geq |t(A \cup B)|$, $P(A \cup B') \geq P(A \cup B)$ and $P(A \cup B')/P(A) \geq P(A \cup B)/P(A)$, where $P(x)$

Table 4: The values of confidence threshold settings under different support degrees.

Case	Description	A and B are positively correlated	A and B are negatively correlated
1	Both $sup(A)$ and $sup(B)$ are smaller values	$conf(\neg A \Rightarrow \neg B)$ is higher: $minconf\_P \longrightarrow$ higher	—
2	Both $sup(A)$ and $sup(B)$ are higher values	$conf(A \Rightarrow B)$ is higher: $minconf\_P \longrightarrow$ higher	—
3	The absolute value of the difference between $sup(A)$ and $sup(B)$ is smaller	—	$conf(A \Rightarrow \neg B)$ or $conf(\neg A \Rightarrow B)$ is smaller: $min\_conf\_NL \longrightarrow$ smaller
4	The absolute value of the difference between $sup(A)$ and $sup(B)$ is higher	—	$conf(A \Rightarrow \neg B)$ or $conf(\neg A \Rightarrow B)$ is higher: $min\_conf\_NH \longrightarrow$ higher

```
Input: frequent itemsets $L$, minimum correction $min\_corr$, triple confidences thresholds $min\_conf\_P$, $min\_conf\_NH$ and $min\_conf\_NL$, support degree difference threshold $\varepsilon$min;
Output: positive association rule set PARs, negative association rule set NARs;
PARs = ∅; NARs = ∅;
For (any frequent itemsets $X$ in $L$)
  Begin
    For (any itemsets A and $B$)
      Begin
        If (($A \cup B = X$) AND ($A \cap B = \varnothing$)) Then
          Begin
            If $corr(A, B) \geq min\_corr$ Then
              If ($conf(A \Rightarrow B) \geq minconf\_P$) Then PARs = PARs ∪ {$A \Rightarrow B$};
              If ($conf(\mathrm{A} \Rightarrow \mathrm{B}) \geq minconf\_P$) Then NARs = NARs ∪ {$\neg A \Rightarrow \neg B$};
            ElseIf ($corr(A, B) \leq -min\_corr$) Then
              If ($|sup(A) - sup(B)| \geq \varepsilon$min Then
                If $conf(A \Rightarrow \neg B) \geq (min\_conf\_NH)$ Then NARs = NARs ∪ {$A \Rightarrow \neg B$};
                If $conf(A \Rightarrow B) \geq (min\_conf\_NH)$ Then NARs = NARs ∪ {$\neg A \Rightarrow B$};
            Else
                If ($conf(A \Rightarrow \neg B) \geq minconf\_NL$ Then NARs = NARs ∪ {$A \Rightarrow \neg B$};
                If ($conf(\neg A \Rightarrow B) \geq minconf\_NL$ Then NARs = NARs ∪ {$\neg A \Rightarrow B$};
            End//If
      End//For
    End//For
  Return PARs, NARs;
End.
```

Algorithm 1: Positive and negative association rules mining with minimum correlation and triple confidence (PNARs_M).

is the probability function. As a result, $conf(A \Rightarrow B') = P(A \cup B')/P(A) \geq P(A \cup B)/P(A) = conf(A \Rightarrow B) \geq min\_conf\_P$.

(2) On the other hand, $corr(A, B) = sup(A \cup B) - sup(A) \times sup(B)$ according to Definition 5; and $corr(A, B) \geq min\_corr$ according to Definition 7. $corr(A, B') = sup(A \cup B') - sup(A) \times sup(B') = P(A \cup B') - P(A) \times P(B')$. If $B' \subseteq B$, then it is evident that $t(B') \supseteq t(B)$, $|t(B')| \geq |t(B)|$. In this way, $P(B') \geq P(B)$, $P(A \cup B') \geq P(A \cup B)$, and $P(A \cup B') - P(A) \times \mathrm{P}(\mathrm{B}') \geq P(A \cup B) - P(A) \times \mathrm{P}(\mathrm{B}) = corr(A, B) \geq min\_corr$.

Consequently, we can conclude that $A \Rightarrow B'$ can be derived from rule $A \Rightarrow B$ once it is a strong association rule; therefore, $A \Rightarrow B'$ is a redundant of rule $A \Rightarrow B$.

The proof is complete. □

Theorem 9. *Let I be a set of items, itemset $A \subseteq I$, itemset $B \subseteq I$, and $A \cap B = \varnothing$, min_corr is the minimum correlation threshold, $A \Rightarrow \neg B'$ is a strong negative association rule mined using algorithm PNARs_M. If there also exists a valid negative association rule $A \Rightarrow \neg B$ in NARs satisfying $B' \subseteq B$, then $A \Rightarrow \neg B$ is a redundant rule of $A \Rightarrow \neg B$'.*

Proof. There are two cases when $A \Rightarrow \neg B$ is a strong negative association rule: $|sup(A) - sup(B)| \geq \varepsilon$min and $|sup(A) - sup(B)| < \varepsilon$min.

When $|sup(A) - sup(B)| \geq \varepsilon$min, it is necessary to explain that $A \Rightarrow \neg B$ can be concluded from a strong association rule $A \Rightarrow \neg B$'. We have to prove the following two inequalities: $conf(A \Rightarrow \neg B) \geq min_conf_$NH and $corr(A, \neg B) \leq -min\_corr$.

(1) It is easy to show that $conf(A, \neg B') \geq \min_conf_$NH according to Definition 7. We know that $B' \subseteq B$, then $t(B') \supseteq t(B)$; $\neg B' \supseteq \neg B$, and $t(\neg B') \subseteq t(\neg B)$, where $t(x)$ is the transactions including itemset x. At the same time, it is evident that $(A \cup \neg B') \supseteq (A \cup \neg B)$; $t(A \cup \neg B') \subseteq t(A \cup \neg B)$. In this way, $conf(A, \neg B) = P(A \cup \neg B)/P(A) \geq P(A \cup \neg B')/P(A) = conf(A, \neg B') \geq min\_conf\_NH$.

(2) Meanwhile, $corr(A, \neg B') \leq -min\_corr$ according to Definition 7. That is to say, $corr(A, \neg B') = P(A \neg B') - P(A) \times P(\neg B')$. Since $B' \subseteq B$, then $\neg B' \supseteq \neg B$, $P(\neg B') \geq P(\neg B)$, and $P(A \neg B') \geq P(A \neg B)$. $corr(A, \neg B) = P(A \neg B) - P(A) \times P(\neg B) \leq P(A \neg B') - P(A) P(\neg B') = corr(A, \neg B) \leq -min\_corr$.

According to (1) and (2), rule $A \Rightarrow \neg B$ can be inferred from the rule $A \Rightarrow \neg B'$ if and only if $B' \subseteq B$ when $|sup(A) - sup(B)| \geq \varepsilon$min, indicating that $A \Rightarrow \neg B$ is a redundant rule of $A \Rightarrow \neg B'$.

Using the similar way, it is easy to show that the rule $A \Rightarrow \neg B$ can be inferred from the rule $A \Rightarrow \neg B'$ if and only if $B' \subseteq B$ when $|sup(A) - sup(B)| < \varepsilon$min.

The proof is complete.

In the same way, the following theorems can be obtained: □

Theorem 10. *Let I be a set of items, itemset $A \subseteq I$, itemset $B \subseteq I$, and $A \cap B = \varnothing$, min_corr is the minimum correlation threshold, $\neg A \Rightarrow B$ is a strong negative association rule mined using algorithm PNARs_M. If there also exists a valid negative association rule $\neg A \Rightarrow B'$ in NARs* *satisfying $B' \subseteq B$, then $\neg A \Rightarrow B'$ is a redundant rule of $\neg A \Rightarrow B$.*

Theorem 11. *Let I be a set of items, both A and B are nonempty itemsets, $A \subseteq I$, $B \subseteq I$, and $A \cap B = \varnothing$, min_corr is the minimum correlation threshold, $\neg A \Rightarrow \neg B'$ is a strong negative association rule mined using algorithm PNARs_M. If there also exists a valid negative association rule $\neg A \Rightarrow \neg B$, satisfying $B' \subseteq B$, then $\neg A \Rightarrow \neg B$ is a redundant rule of $\neg A \Rightarrow \neg B'$.*

We do not take a lot of space proving both of them due to the length limitation of paper. They can be proved with a similar approach as in Theorems 8 and 9 (see Appendix).

Using the above theorems, we can prune those redundant rules from the mining results. The pruning redundant PNARs algorithm based on inclusion relation, called PNARs_P, is described as follows, where PARs and NARs are the outputs of Algorithm 1 with frequent itemsets L as input.

In Algorithm 2, line (1) initializes the PNARs to be an empty set. Lines (2) to (16) scan each rule r in turn and classify them into different categories according to their type. In line (4), if r is a type of $A \Rightarrow B$, then search the rules with the same antecedent satisfying $B' \subseteq B$, and delete them. Similarly, lines (7), (10), and (13) are used to process different types of rules. Line (17) merges the two sets to PNARs, and line (18) returns the results and ends the whole algorithm. As can be seen from the PNARs_P algorithm, most of the time is taken by the scanning of rules set in PARs and NARs. For any rule r, it is necessary to test their inclusion relation with any other rule in PARs or NARs. In this way, the total execution time is n^2, where n is the number of PARs and NARs. That is to say, the total time complexity of the PNARs_P algorithm is $O(n^2)$.

4.5. Performance Evaluations of Optimized PNARs Mining and Pruning Algorithms. To evaluate the performance of the proposed PNARs mining and pruning algorithms, we have implemented simulations with Java programming language on a 64 bit Windows 10 Profession platform, whose hardware configurations are set as Intel(R) Core(TM) i7-12700H CPU @ 3.40 GHz, 256 G Memory. Furthermore, some popular datasets with different numbers of instances and attributes are used, including mushroom, nursery, and chess databases from the UCI database (https://archive-beta.ics.uci.edu/ml/datasets). The other three datasets are synthetic with an IBM data generator. Table 5 describes the details of these datasets.

4.5.1. Simulation and Analysis for Association Rules Mining Algorithm. To test the effectiveness of the PNARs_M algorithm, it is necessary to compare a common single-confidence model with different confidence settings. In the first five simulations, the sole confidence threshold is initially set to be the same values as 0.1, 0.3, 0.5, 0.7, and 0.9, respectively. The number of various types of mined association rules is compared with different datasets. In contrast, the triple confidence approaches are assigned with fixed values separately in the last two simulations. These confidence setting values are set with mc1 and mce2, respectively, as shown in Table 6.

Figure 1 depicts the comparison results, from where we can see that the number of mined rules from different datasets decreases along with the increase of confidence thresholds. However, the higher value of confidence thresholds usually results in a large number of PNARs, while a lower value usually leads to some interesting rules missing. Furthermore, the number of rules is relatively reasonable when multiple confidences are set for each type of rules. Also, these results are in line with those in studies by Kishor and Porika [11] and Dong et al. [13], who showed that multiple minimum confidences can effectively control the number of rules. On the other hand, mined rule numbers with mc1 and mc2 are larger compared to the rule number of 0.9. The reason is that the proposed triple confidences control the rules number according to their rule types.

In the second simulation for the algorithm of positive and negative association rules mining with minimum correlation and triple confidence (PNARs_M), we introduce the approach in reference [54] as the benchmark in this study to test the performance of the proposed algorithm. Mining results are compared with them in the number of mined rules on the dataset of chess in UCI and DS3, whose descriptions are shown in Table 5. As can be seen in Table 7, where *P_mc* is the confidence threshold for rule $A \Rightarrow B$ and rule $\neg A \Rightarrow \neg B$; *N_mc* is the confidence threshold for rule $\neg A \Rightarrow B$ and rule $A \Rightarrow \neg B$; satisfying equation $P\_mc + N\_mc = 1$. NPAR and NNAR are used to denote the number of positive association rules and negative association rules, respectively.

As can be seen from Table 7, the number of mined positive association rules between the methods in previous studies and the proposed algorithm in this study are almost equivalent when the confidence thresholds are set with the same values, indicating that the proposed algorithm has equally excellent performance when extracting the positive rules. However, the number of mined negative association rules has significantly reduced compared to the approach in reference [54]. The reason can be ascribed to the introduction of triple confidence, which can control the negative rules number by *minconf_NH* and *minconf_NL*. The sum values of *P-mc* and *N-mc* are set to a fixed value in the double confidence approach, and if the value of *P-mc* is high, then the value of *N-mc* is low and vice versa, which will inevitably lead to the extraction of many useless and boring rules. By contrast, the proposed approach fully considers and satisfies the internal regularity of confidence and sets them with more reasonable values, which not only extract the interesting rule but also control the number of rules with lower confidence. Based on the analysis, it is obvious that the proposed algorithm is more effective than the double threshold approach in controlling rule number and ensuring the rules' interest. Furthermore, the average reduction of negative mined rules is down by sixty percent, indicating the validity of triple confidence settings.

```
Input: PARs, NARs;
Output: a set of nonredundant PNARs;
PNARs = $\varnothing$;
For (any rules $r$ in PARs or NARs)
  Begin
    If (r is a type of $A \Rightarrow B$) Then
      For (any rule $A \Rightarrow B'$)
        If (B' $\subseteq$ B) Then Delete rule $A \Rightarrow B'$ From PARs;
  ElseIf (r is a type of $A \Rightarrow \neg B'$) Then
    For (any rule $A \Rightarrow \neg B$)
      If (B' $\subseteq$ B) Then Delete rule $A \Rightarrow \neg B$ From NARs;
  ElseIf (r is a type of $\neg A \Rightarrow B$) Then
    For any rule $\neg A \Rightarrow B'$
      If (B' $\subseteq$ B) Then Delete rule $\neg A \Rightarrow B'$ From NARs;
  ElseIf (r is a type of $\neg A \Rightarrow \neg B'$) Then
    For any rule $\neg A \Rightarrow \neg B$
      If (B' $\subseteq$ B) Then Delete rule $\neg A \Rightarrow \neg B$ From NARs;
  End
  PNARs = PARs $\cup$ NARs;
Return PNARs;
End.
```

Algorithm 2: Pruning redundant of PNARs (PNARs_P).

Table 5: The details of simulation datasets.

Dataset	Data size (Kb)	Number of items	Average items per transaction	Number of total transactions
Mushroom	365	23	13	8,124
Nursery	1,035	9	9	12,960
Chess	241	36	36	3,196
DS1	2,168	45	40	8,546
DS2	3,286	62	43	8,412
DS3	5,864	80	62	19,854

Table 6: The values of different minimum confidence.

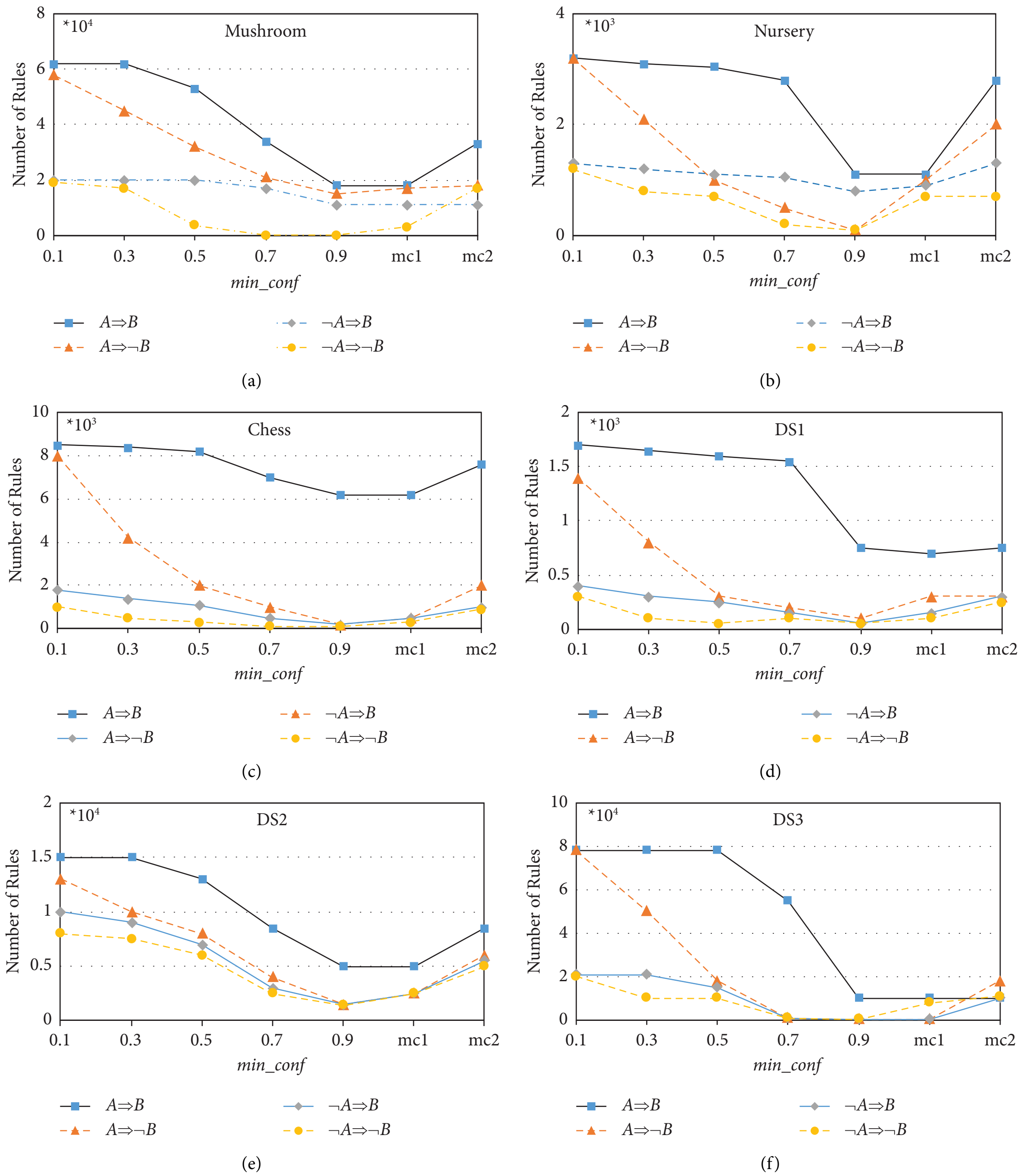
Datasets	mc1			mc2		
	minconf_P	*minconf_NH*	*minconf_NL*	*minconf_P*	*minconf_NH*	*minconf_NL*
Mushroom	0.70	0.57	0.42	0.68	0.56	0.42
Nursery	0.56	0.56	0.44	0.58	0.48	0.40
Chess	0.72	0.68	0.48	0.72	0.58	0.44
DS1	0.62	0.62	0.45	0.60	0.52	0.41
DS2	0.74	0.64	0.46	0.71	0.60	0.52
DS3	0.70	0.56	0.42	0.72	0.52	0.40

4.5.2. Simulation and Analysis for Association Rules Pruning Algorithm. In this subsection, we will verify the efficiency and effectiveness of the proposed association rules pruning algorithm with different minimum correlations. Figure 2 shows the experimental results. For the reason that there was no related research currently, we set the number of mined rules before pruning as the benchmark that the pruned rules can be compared.

In Figures 2(a)–2(e), PARN and NARN are used to describe the number of association rules mined by algorithms PNARs_M and PNARs_P, respectively. As we can see that these two values decrease gradually with the increase of minimum correlations. Specifically, the number of association rules mined by PNARs_P is far less than those of rule by PNARs_M. In particular, when the minimum correlation is set to 0.15, the pruning effect of some datasets is acceptable with a pruning rate more than 70%, indicating the pruning validity of the proposed strategy.

Also, the pruning results are verified in the following experiments based on Algorithm 2. Table 8 presents the pruning rate of PNARs on datasets with different confidence configurations.

As can be seen from Table 8, the pruning rates decline with the increase of the confidence threshold until it reaches a critical value and will gradually increase with a larger confidence threshold. The reason can be attributed that the



(a)

(b)

(c)

(d)

(e)

(f)

FIGURE 1: Number of association rules comparison with different minimum confidences. (a) Mushroom. (b) Nursey. (c) Chess. (d) DS1. (e) DS2. (f) DS3.

TABLE 7: Number comparison of mined rules on different datasets.

Algorithm	Confidence setting			Chess		DS3	
	P_mc/minconf_P	*N_mc/minconf_NH*	*minconf_NL*	PARN	NARN	PARN	NARN
Approach in reference [54]	0.90	0.10	—	1,076	171	1,224	14,457
	0.85	0.15	—	1,076	211	1,316	15,784
	0.80	0.20	—	1,076	241	1,415	16,996
This study	0.90	0.60	0.30	1,085	107	1,108	6,322
	0.85	0.55	0.40	1,088	131	1,239	6,041
	0.80	0.55	0.40	1,190	182	1,301	6,401

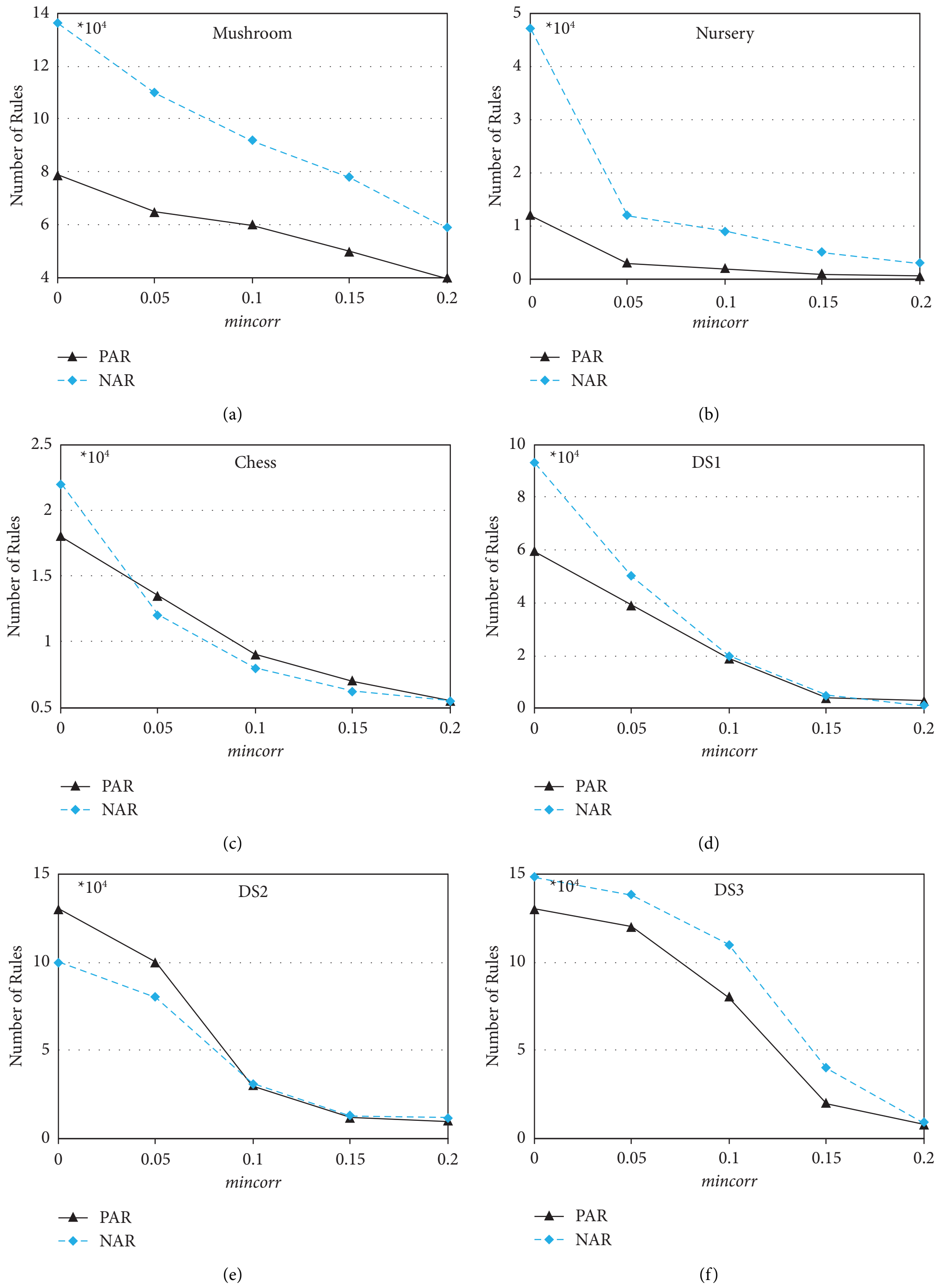


Figure 2: Number of association rules comparison after pruning with different minimum correlations.

PNARs_M filters some rules and limits them joining into PNARs with appropriate confidence threshold setting. In particular, all of the pruning ratios are greater than fifty percent, and the pruning rate on the mushroom is even more than 82.5% when the confidence is equal to 0.15. On the whole, the proposed method can reduce the number of redundant PNARs by over 70.1% with experimental datasets, indicating the better performance of the proposed pruning strategy. In this way, the experimental results show that the proposed pruning algorithm can reduce the number of those meaningless rules.

To further describe the performance of the proposed pruning algorithm, we define the concept of “accuracy” to

Table 8: The pruning rate (%) on different datasets with different confidence thresholds.

Confidence threshold			Mushroom	Nursery	Chess	DS1	DS2	DS3
minconf_P	*minconf_NH*	*minconf_NL*						
0.15	0.50	0.25	82.5	81.6	78.4	82.3	81.6	79.4
0.30	0.55	0.30	78.5	76.3	75.4	76.3	70.5	74.2
0.45	0.60	0.35	76.3	72.5	71.4	70.6	68.5	64.3
0.60	0.65	0.40	72.1	68.5	65.8	65.7	60.9	60.1
0.75	0.70	0.45	74.3	70.2	63.4	60.8	60.4	50.7
0.90	0.75	0.50	75.4	72.6	65.5	64.5	66.6	57.7

describe the ratio of the useful association rules to the total mined rules. In detail, it can be defined as accuracy = 1 − $n_{\text{useful}}/n_{\text{total}}$, where n_{pruning} is the number of useful mining rules; n_{total} is the number of total mining rules. It is obvious that accuracy reflects the efficiency of useless rules pruning. Table 9 lists the accuracy of the proposed mining algorithm.

It is obvious from Table 9 that the accuracy of the useful rules is higher, indicating that many useless rules have been removed from the mined result. Consequently, the proposed strategy has better effectiveness and adaptability with different sizes and characteristics of datasets. In addition, it takes a little extra time when mining association rules, indicating its high efficiency.

Table 9: The accuracy on different datasets with different confidence thresholds.

Confidence threshold			DS1	DS2	DS3
minconf_P	*minconf_NH*	*minconf_NL*			
0.15	0.50	0.25	92.6	89.9	85.6
0.30	0.55	0.30	93.1	92.5	92.5
0.45	0.60	0.35	90.3	90.6	90.2
0.60	0.65	0.40	94.4	88.7	92.1
0.75	0.70	0.45	86.9	92.3	82.4
0.90	0.75	0.50	92.5	94.3	87.5

4.5.3. Simulation and Analysis for Comprehensive Mining Approach. To comprehensively test the performance of the proposed mining approach, it is compared with the existing algorithms in terms of precision, recall, and *F*-measure. Furthermore, the rules extracted from the dataset are divided into true positives, false positives, true negatives, and false negatives according to the measures above. By referring the definitions in reference [24, 25], precision, recall, and F-measure are redefined according to usability of mined rules as follows: precision is expressed as the number of relevant mind rules divided by the number of total retrieved rules; recall can be described as the number of relevant mind rules divided by the number of total relevant rules; and F-measure can be regarded as the twice value of (precision ∗ recall)/(precision + recall). The comparisons of the algorithms based on average values on different datasets are shown in Table 10, where TP rate is the rate of true positives and FP rate is the rate of false positives (instances falsely extracted as a rule).

Table 10: Performance comparison of average precision, recall, and *F*-measure with different datasets.

Algorithm	TP rate	FP rate	Precision	Recall	*F*-measure
Apriori [19]	0.53	0.31	0.56	0.91	0.69
PNARM [54]	0.56	0.30	0.54	0.93	0.68
CBPNARM [26]	0.73	0.02	0.94	0.92	0.93
CARM [24]	0.79	0.01	0.97	0.94	0.95
This study	0.78	0.01	0.99	0.95	0.97

The values given in Table 10 are calculated by comparing the results of the algorithms for extracting association rules with the real rules that are used by the expert of the domain, from where as can be seen that, higher precision, recall, and F-measure for the proposed approach indicate that the algorithm has extracted more useful rules, compared with other approaches.

5. Application with Social Insurance Fund Auditing Data Using Proposed Association Rule Mining Framework

As a complex and systematic task, audit is regarded as one particularly important work in promoting economic and social development. In recent years, big data auditing has gradually become a new auditing paradigm, bringing us not only the changes in auditing methods but also the comprehensive transformation of the audit mode. The application of association rule mining technology in the audit field can provide auditors with abnormal information and therefore improve audit capability and efficiency.

5.1. Insurance Fund Auditing Data Mining Framework with Association Rule. The insurance fund data used in this study involves more than several hundred-thousand people from eight provinces in China, including employee profiles, basic annuities insurance premium payment information, retirees' basic information, and so on. The datasets are rich in information with various types of data formats. In this way, a framework of association rule mining in auditing is proposed in this study, as can be seen in Figure 3. First, the relevant insurance fund data are gathered, and the quality of the data is verified. Generally, the assembled data contain missing or incomplete attributes, noise (containing errors or outlier values that deviate from what is expected), and data inconsistencies. Therefore, the collected data must be cleaned and transformed before it can be utilized in data mining

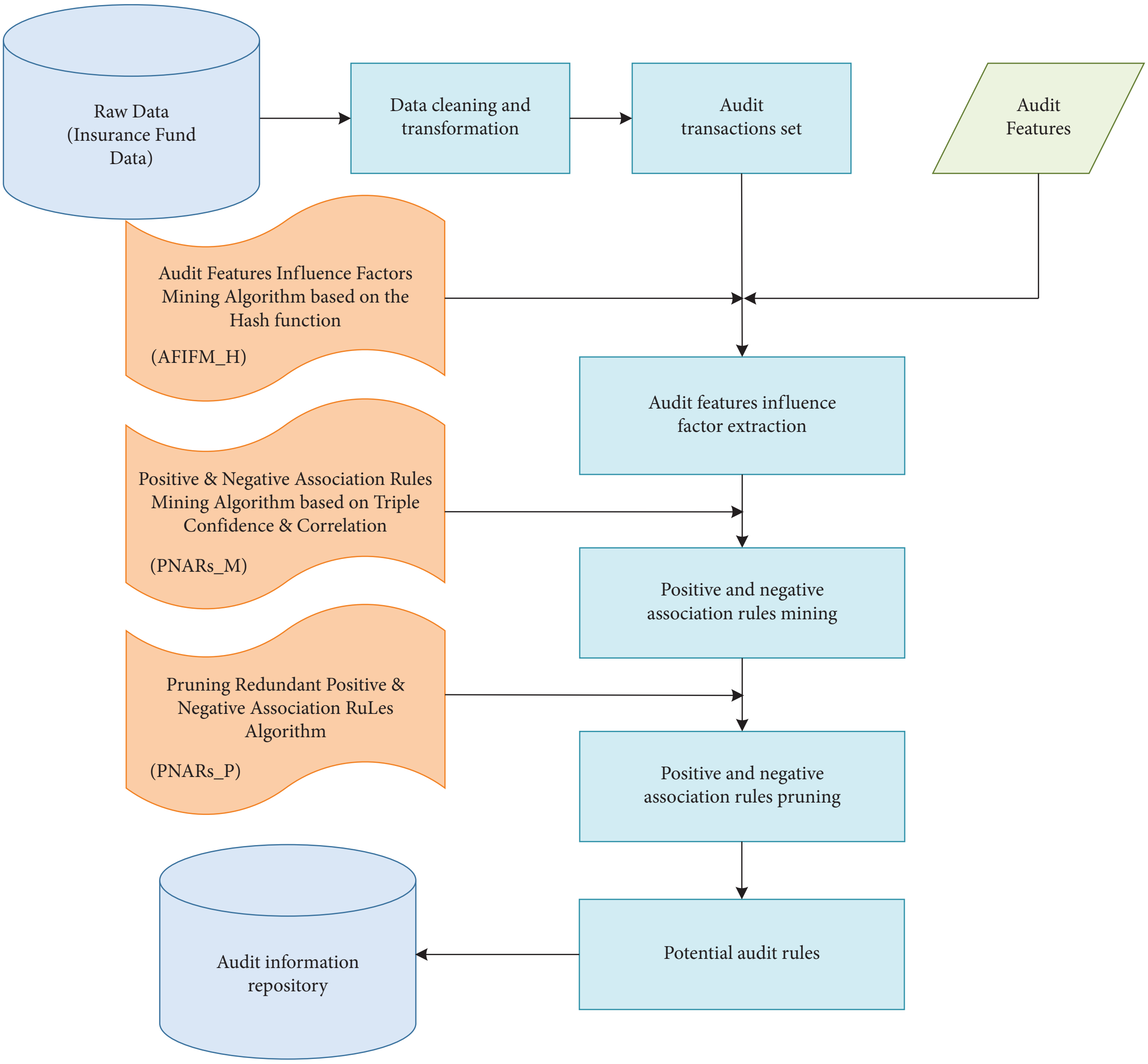


Figure 3: Auditing data mining framework with association rule.

systems in order to produce better-quality results. Insurance fund data cleaning involves several processes, such as filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies. Then, the cleaned datasets are transformed into a form of relation table that is suitable for audit data mining. After data preprocessing is complete, feature influence factors are extracted using a frequent item mining algorithm (seen in the next subsection). After this, we compute the support and confidence between feature factors to mine positive and negative rules (PNARs_M Algorithm). Then, the PNARs pruning strategy (PNARs_P algorithm) is applied to exclude those redundant association rules. The ultimate rules set will be obtained as the final step. Once the rules are accepted by the auditing institutions, they are going to be new potential auditing rules and be added to the audit information repository for finding auditing clues.

5.2. Audit Features Influence Factors Mining. The audit feature describes the facts and current situation of auditing transactions. Audit feature impact factors are elements that play an important role in audit feature description. For example, “high individual base payment” is a feature in social insurance fund auditing, and its impact factors that influence this feature may include “his/her administrative position” and “his/her age.” In general, the most critical step in extracting audit impact factors is to find those frequent itemsets satisfying the minimum support degree and minimum confidence degree from audit datasets.

The main way to find auditing features from social insurance data in this study is to seek out attributes that satisfy the minimum support threshold using the frequent itemsets mining algorithm. Li et al. [16] proposed a hash-based algorithm, called DHP, for frequent item discovery. It is effective in the generation of candidate itemsets for large 2-itemsets. However, with the size of transaction data increasing, the possibility of hash collisions greatly increases, thus resulting in a lower efficiency of the algorithm. To address this problem, we propose a minimal hashing and pruning algorithm to reduce the potential collisions and thus improve the mining efficiency.

Suppose there are four transactions in a database and the predefined minimum support is 50%, meaning that the number of frequent items should be more than two. The minimal hashing and pruning algorithm first scans the database once, generating $C1$ and $L1$ subsequently, and then generates a hash table H. The hash function is defined as

$H(x, y) = (P(x) - 1) * (P(y) - 1) + P(y) - P(x) * ((P(x) -1))/2 - 2$, where $P(x)$ and $P(y)$ are the sequence numbers of x and y in L_{k-1} separately.

All the 2-items are mapped to the hash table using the hash function value. For example, the hash value of 2-item $\{B, E\}$ can be calculated as follows:

$H(2, 5) = (2 - 1) * (5 - 1) + 5 - 2 * (2 - 1)/2 - 2 = 6.$

In this way, the 2-item $\{B, E\}$ is hashed to the sixth position in the hash table, as can be seen in Figure 4.

In this study, we propose an audit features influence factors mining algorithm based on the hash function (AFIFM_H). It greatly reduces the database scanning time using not only hashing but also deletion. As can be seen in Figure 4, {AD} and {CD} can be generated from transaction 100, while only one of these 2-item exists and will be removed due to infrequency. In this way, {AC} is the only candidate itemset. At the same time, the 1-itemsets $\{A\}$ and $\{C\}$ are lower than the minimum support threshold, meaning that no 3-itemset will be generated from transaction TID 100, and it can be deleted from the database. The audit features influence factors mining algorithm can be described as follows (see Algorithm 3):

In Algorithm 3, line (3) generates 1-item by scanning the transactions dataset and inserts them into a hash tree with hash function $h_2(x)$. In essence, the proposed AFIFM_H algorithm uses a hashing function to filter out unnecessary itemsets for the generation of the next candidate itemsets. The main advantage of the algorithm is that it greatly reduces hash collisions and decreases scanning times, thus improving the efficiency of mining frequent itemsets.

Next, we will evaluate the time performance of the proposed algorithm compared to the other frequent item mining strategies: FP-Growth and DHP [55–57]. All the experiments were performed on a server with an Intel Pentium dual core 3.4 GHz CPU, running on a Windows Server operating system and 256 GB of memory. All programs are implemented using Java version 1.8. We test and verify the usability of our approach on a dataset consisting of 20 attributes out of which 7 are numerical and 13 are categorical with almost 50,000 transactions. Figure 5 shows the average time cost comparisons for these three algorithms with different minimum support degrees.

The experimental results in Figure 5 show that the execution time of FP-Growth, DHP, and AFIFM_H gradually decrease with the increase of the minimum support threshold. The main reason is that more and more itemsets satisfy the minimum support threshold when the value of min_sup is lower for the same dataset. Therefore, the generation of frequent items will take an increasing amount of time as the number of rules increases. In addition, it is worth noting that the AFIFM_H algorithm has a lower time complexity than the other strategies. The reason is that the hash function reduces itemsets collision, and 2-itemsets can be generated by first scanning the database directly, as can be seen that the proposed algorithm is an efficient approach for mining frequent items because of scanning time reduction.

5.3. PNARs Mining and Pruning with Insurance Fund Data. In this subsection, we present PNARs mining and pruning with audit feature influence factors in social insurance fund data. The experimental data used in this paper is from the social insurance departments of eight provinces in China. To mine positive and negative auditing rules, we design an association rule mining application using Java programming language, and its interface is shown in Figure 6.

These datasets mainly involve basic annuities insurance premium payment information, named IC01, and retiree basic information, named IC02. The main fields of these tables are listed in Tables 11 and 12, respectively. These datasets are saved in independent files according to their source, named A_Prvns to H_Prvns, to compare association relationships with different provinces.

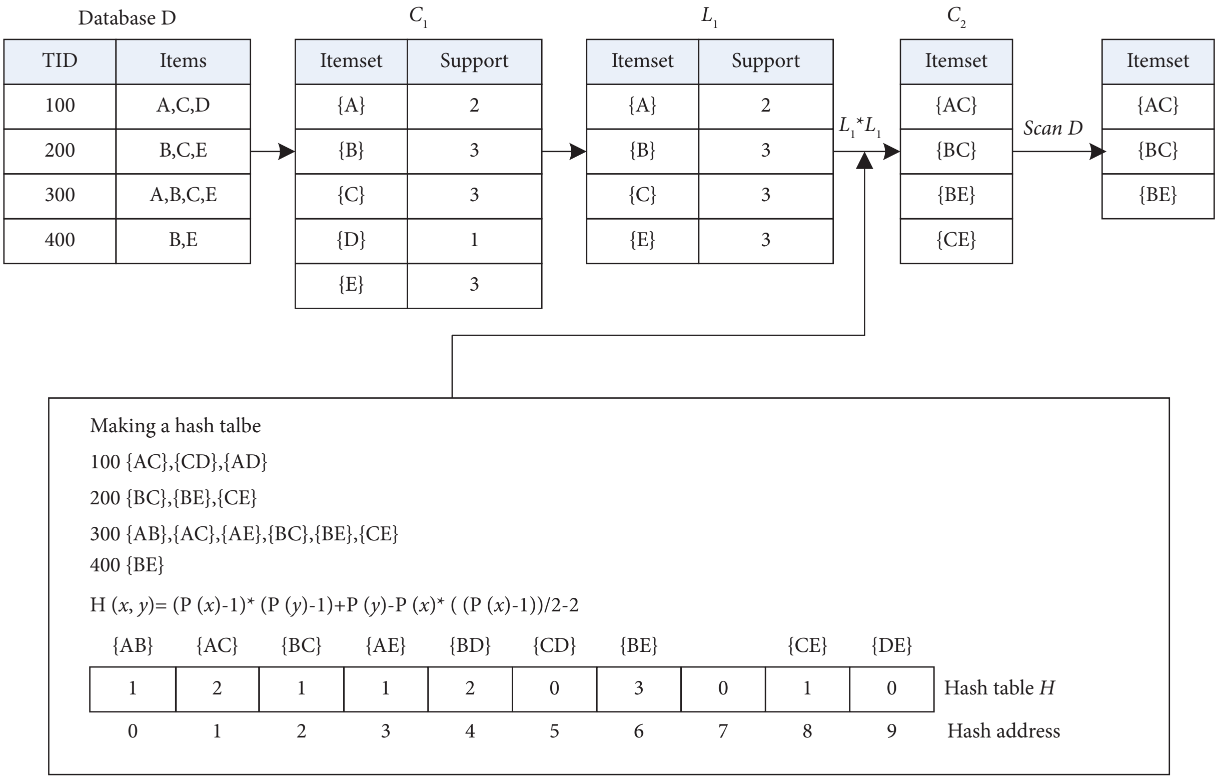
As an example, the mining results from Tables 11 and 12 of A_Prnvs using AFIFM_H are shown in Tables 13 and 14, respectively, where the minimum support threshold is set to 0.2.

From the features influence factors listed in Tables 13 and 14, we can obtain much useful information; for example, “Basic_pension_month” is an important factor that decides the amount of pension. At the same time, the proposed algorithm efficiency is also tested regarding access time in audit features influence factors mining.

5.4. Mining Results and Discussion. In the experiments, the parameters are set as follows: $min\_corr = 0.25$, $min\_conf\_P = 0.65$, $min\_conf\_NH = 0.52$, and $min\_conf\_NH = 0.40$.

In order to find fraud clues from auditing data as soon as possible, not only positive rules but also negative rules play an important role in social insurance fund auditing. Figure 7 presents rules number comparisons with different social insurance fund datasets, including LOGIC [13], PNARC [56], and the strategy proposed in this study.

From Figure 7, we can see that the proposed strategy exhibits some advantage in controlling the number of mined rules with different audit datasets. The number of PNARs mined by PNARs_M and PNARs_P is far less than the number of those mined by PNARC and LOGIC, no matter which dataset is used. The main reason can be attributed to the triple confidence threshold settings, and many meaningless negative association rules with low confidence may be excluded. Also, the proposed strategy considers confidence intrinsic relation and the change rule, therefore effectively controlling the mined rules quantity. It will neither miss those interesting association rules nor produce too many association rules with low reliability. In this way, the approach in this study has great advantages in



Database D

TID	Items
100	A,C,D
200	B,C,E
300	A,B,C,E
400	B,E

C_1

Itemset	Support
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

L_1

Itemset	Support
{A}	2
{B}	3
{C}	3
{E}	3

C_2

Itemset
{AC}
{BC}
{BE}
{CE}

Itemset
{AC}
{BC}
{BE}

Making a hash talbe

100 {AC},{CD},{AD}

200 {BC},{BE},{CE}

300 {AB},{AC},{AE},{BC},{BE},{CE}

400 {BE}

H (x, y)= (P (x)-1)* (P (y)-1)+P (y)-P (x)* ((P (x)-1))/2-2

	{AB}	{AC}	{BC}	{AE}	{BD}	{CD}	{BE}		{CE}	{DE}
Hash table H	1	2	1	1	2	0	3	0	1	0
Hash address	0	1	2	3	4	5	6	7	8	9

Figure 4: An example of searching for frequent itemsets.

```
Input: Insurance fund database $D$; min_sup;
Output: Frequent 1-itemset $L_1$; 2-itemset hash table $H_2$; Frequent 2-itemset $L_2$;
  Begin
  $H_2 = \varnothing$
  For all transactions $t \in$ D
    Count 1-item occurrences and insert them into a hash tree;
    For all 2-subsets $x$ of $t$ do
      $H_2[h_2(x)]$++;
  $L_1 = \{c|c.count \geq min\_sup\}$;
  $C_2 = L_1 * L_1$;
  For all candidates 2-itemsets $c$
    If $c \in C_2$ and $H_2[h_2(C)] < min\_sup$ Then
      Delete $c$ from $C_2$;
Scan the database and count each 2-item;
$L_2 = \{c|c.count \geq min\_sup$;
End.
```

Algorithm 3: Audit features influence factors mining algorithm based on the hash function (AFIFM_H).

controlling the rules number and ensuring their interests compared with methods in references [58–60]. Compared with the other approaches, the proposed method can reduce the number of redundant PNARs by average 78.5% with audit data mining.

At the same time, auditors usually pay much attention to the effectiveness and feasibility of mined rules, which can be measured by the fraud clues number that is found. Those audit fraud clues can help auditors quickly distinguish abnormal instances from insurance fund data, thus further identifying them accurately. Table 15 presents the ratios comparison of the number of fraud clues to mined rules.

From Table 15, it is obvious that ratios of the fraud clues number to mined rules are a little different among the eight datasets with the same model. We can observe that most of the rules mined from the proposed strategy are regarded as fraud clues by the auditors compared to previous studies, such as LOGIC [13] and PNARC [56]. Furthermore, the average ratio of 89% mined rules can help the auditors identify abnormal samples, improving the efficiency of

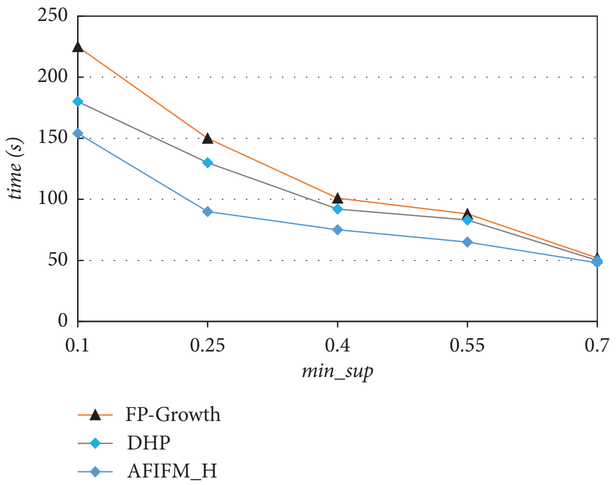
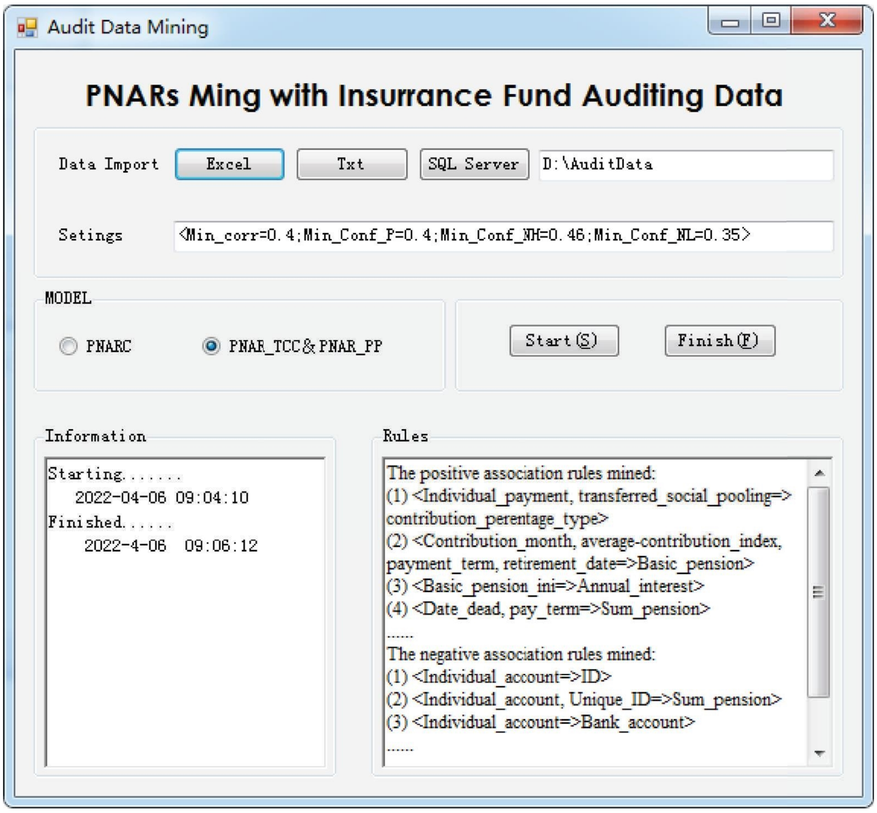


Figure 5: Execution time comparisons with different min_sup.



Data Import D:\AuditData

Setings <Min_corr=0.4;Min_Conf_P=0.4;Min_Conf_NH=0.46;Min_Conf_NL=0.35>

○ PNARC ● PNAR_TCC&PNAR_PP

```
Starting.......
   2022-04-06 09:04:10
Finished......
    2022-4-06   09:06:12
```

The positive association rules mined:
(1) <Individual_payment, transferred_social_pooling=> contribution_perentage_type>
(2) <Contribution_month, average-contribution_index, payment_term, retirement_date=>Basic_pension>
(3) <Basic_pension_ini=>Annual_interest>
(4) <Date_dead, pay_term=>Sum_pension>
......
The negative association rules mined:
(1) <Individual_account=>ID>
(2) <Individual_account, Unique_ID=>Sum_pension>
(3) <Individual_account=>Bank_account>
......

Figure 6: The interface of the proposed association rule mining approach.

insurance fund data auditing. On the one hand, the proposed strategy reduces the association rules number by the reasonable minimum confidence settings. On the other hand, the pruning method further eliminates those meaningless rules while the previous studies have not involved, indicating its higher efficiency and better performance than previous studies. Based on the analysis above, the proposed algorithms not only discover positive association rules but also negative rules among the properties. Nearly all the mined rules are meaningful to auditing, which can improve audit quality.

A part of the mined rules is analyzed as follows:

(1) <Individual_pay_amount, Transferred_social_pooling ⇒ Contribution_percentage_type>

In the mining procedure, three frequent itemsets are found: "Individual_pay_amount," "Transferred_social_pooling," and "Contribution_percentage_type." The implicit association relation is "Individual_pay_amount, Transferred_social_pooling ⇒ Contribution_percentage_type." This can be interpreted as attribute "Transferred_social_

Table 11: Fields of insurance premium payment information (IC01).

Field	Type	Description
Corporate_ID	Char(20)	Corporate ID
Personal_ID	Char(18)	Personal ID
Payment_term	Char(10)	Payment term
Date_org	DateTime	The expiry date of organization deposit
Payer_type	Char(10)	The type of payer
Date_indiv	DateTime	Expiry date of individual deposit
Crate_enterprise	Decimal(8, 2)	Enterprise contribution rate
Crate_individual	Decimal(8, 2)	Individual contribution rate
Amount_corporate	Decimal(8, 2)	Amount of corporate contribution
Amount_individual	Decimal(8, 2)	Amount of individual contribution
Avg_wage	Decimal(8, 2)	Average monthly wage
Transferred_social_pooling	Decimal(8, 2)	Transferred to pooling
Time_con	DateTime	Actual contribution time
...	...	...
Memo_p	Text	Other important personal information

Table 12: Fields of retiree basic information (IC02).

Field	Type	Description
Corporate_ID	Char(20)	Corporate ID
Personal_ID	Char(18)	Personal ID
Personal_name	Char(12)	Personal name
Personal_tel	Char(11)	Personal phone
Address	Char(20)	Address
Bank_account	Char(22)	Bank account
Basic_pension_month	Decimal(8, 2)	Basic pension each month
Retirement_type	Char(12)	Retirement type
Average_contribution_index	Decimal(8, 2)	Average contribution index
Date_Retire	Date	Date of retirement
Accounts_individual	Decimal(8, 2)	Deposit amount of individual accounts
Interest_individual	Decimal(4, 2)	Interest rate for calculation of individual accounts
Contribution_time	DateTime	Actual contribution time
...	...	...
Memo_p	Text	Other important personal information

Table 13: Feature influence factors mining from A_Prvns_IC01.

Feature	Support threshold
Payment_term	0.37
Payer_type	0.31
Individual_pay_base	0.29
Individual_pay_amount	0.28
Transferred_social_pooling	0.27
Contribution_from_corporate	0.27
Contribution_percentage_type	0.26
Social_spooling_from_corporate	0.26
Date_social_spooling_from_corporate	0.25
Inter_annual_interest_social_transferred	0.25
Annual_interest_social_spooling_transferred	0.24
Pament_type	0.24

Table 14: Feature influence factors mining from A_Prvns_IC02.

Feature	Support threshold
Basic_pension_month	0.36
Treatment_Date	0.35
Retirement_Type	0.32
Retirement_Date	0.31
Contribution_month	0.30
Retirement_date_basic_pension	0.29
Average_Contribution_Index	0.26
Total_Amount_When_Retire	0.25
Percentage_Amount_Basic_Pension	0.25

pooling" and "Transferred_social_pooling" in Table 12 mean the social pooling from corporate contributions; and the corresponding field "Contribution_percentage_type" can be acquired; then the association relation between them is the rule for audit analysis.

(2) <Contribution_month, Average_contribution_index, Payment_term, Retirement_date ⇒ Basic_pension_month>

"Contribution_month," "Average_contribution_index," "Payment_term," "Retirement_date," and "Basic_pension_month" are frequent itemsets found satisfying the minimum support threshold. Also, the rule "Contribution_month, Average_contribution_index, Payment_term, Retirement_date ⇒ Basic_pension"

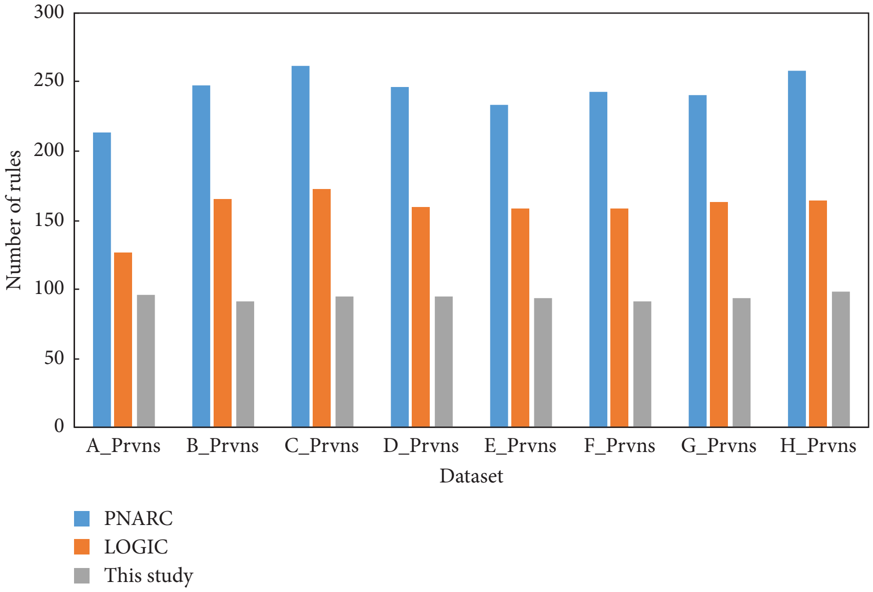


Figure 7: Mined rules number comparisons with different social insurance fund datasets.

Table 15: Ratios comparison of the number of fraud clues to mined rules.

Datasets	PNARC	LOGIC	This study
A_Prvns	40	64	88
B_Prvns	36	51	92
C_Prvns	32	48	91
D_Prvns	34	52	92
E_Prvns	35	52	85
F_Prvns	34	50	93
G_Prvns	37	48	89
H_Prvns	33	50	81
Average	35	52	89

can be interpreted as the “Contribution_month,” “Average_contribution_index,” “Payment_term,” and “Retirement_date” are the most important factors for “Basic_pension_month.” In this way, writing errors and calculation errors may result in the wrong individual basic pension, which can be regarded as an important clue in auditing.

(3) <Individual_account ⇒ Personal_ID>

This is a useful association rule, indicating the relation between the “Individual_account” and “Personal_ID.” In practice, a unique individual account corresponds to a personal_ID. Once a rule like <Individual_account ⇒ Personal_ID> is mined, indicating that an “Individual_account” is mapping to more than one “Personal_ID,” this signifies an audit clue of duplicate Personal_ID.

6. Conclusion and Future Work

Currently, data mining-based auditing is playing a more and more important role in the government supervisory system and has caused increasing attention all around the world. Nevertheless, the lack of means to mine the hidden audit clues behind the data, the difficulty of finding increasingly hidden cheating techniques caused by the electronic and networked environment, and the inability to solve the quality defects of the audited data are very common currently. Therefore, it is a great challenge for auditors to find any relevant faults or mistakes when facing large-scale and complex datasets [61, 62]. Particularly in PNARs mining, the existing confidence threshold values setting methods are more difficult to control the number of low-reliable rules mining and miss the interesting rules under the traditional framework of relevance-support-confidence. If the value of the rule confidence threshold is set too high, many useless and worthless will be mined; while if the value is set too low, some valuable and important rules may be missed. To solve the above problem, this paper proposes an improved PNARs mining framework, including a positive and negative association rule mining algorithm based on triple confidences and minimum correlation, and a PNARs prune algorithm combined with inclusion relation of rules’ antecedent and consequent. The results of the proposed algorithms are compared with other similar algorithms and outperforms the existing algorithms in terms of rule number and other metrics, including precision, recall, and F-measure. In addition, the proposed association rules mining framework is particularly suitable to apply in the file of big data auditing because of its special data features. Using the proposed association rules mining schema, auditors can easily perceive the hidden relation among the social insurance fund auditing data, therefore facilitating fraud clues finding by narrowing the range of confidence settings. On the one hand, the triple confidence threshold effectively reduces the generation of invalid rules mining in audit data, and on the other hand, the pruning algorithm further removes those redundant rules. Furthermore, its application in audit data mining can quickly discover different kinds of errors, irregularities, and illegal acts; therefore, the efficiency of audit work could be improved. The application in Chinese social insurance fund data auditing shows better performance than traditional approaches in both accuracy and efficiency, reducing the number of redundant PNARs by over 70.1% with experimental datasets and average 78.5% with auditing data mining, respectively.

Nonetheless, the article has several possible extensions in the near future. One is to optimize the association rules pruning algorithm by discovering the hidden inclusion relation between itemsets from the perspective of semantic analysis, which can further improve the mined rules quality. The other is to provide a method for determining the rules' specific support value depending on context and utility. These can help enhance the reliability and accuracy of association rules mining results. Also, we wish to conduct experiments on other real datasets and compare the performance of our strategy with other related algorithms, such as data in commercial bank auditing and financial auditing.

Appendix

Theorem A.12. *Let I be a set of items, $A \subseteq I$, $B \subseteq I$, and $A \cap B = \varnothing$, min_corr is the minimum correlation threshold, rule $\neg A \Rightarrow B$ is a strong negative association rule mined using algorithm PNARs_M. If there also exists a valid negative association rule $\neg A \Rightarrow B'$ in NARs satisfying $B' \subseteq B$, then $\neg A \Rightarrow B'$ is a redundant rule of $\neg A \Rightarrow B$.*

Proof. There are two cases when $\neg A \Rightarrow B'$ is a strong negative association rule: $|sup(A) - sup(B)| \geq \varepsilon\text{min}$ and $|sup(A) - sup(B)| \leq \varepsilon\text{min}$.

When $|sup(A) - sup(B)| \geq \varepsilon\text{min}$, it is necessary to explain that $\neg A \Rightarrow B'$ can be concluded from a strong association rule $\neg A \Rightarrow B$. The authors have to prove the following two inequalities: $\text{conf}(\neg A \Rightarrow B') \geq minconf\_NH$ and $corr(\neg \text{A}, \text{B}') \leq -min\_corr$.

(1) It is easy to show that $conf(\neg A \Rightarrow B') \geq min\_conf\_NH$ according to Definition 7. We know that B' ⊆ B, then $t(B') \supseteq t(B)$, where $t(x)$ is the transactions including itemset x. At the same time, it is evident that $\neg A \cup B' \subseteq \neg \text{A} \cup B$; $t(A \cup \neg B') \supseteq t(A \cup \neg B)$. In this way, $conf(\neg A \Rightarrow B') = P(\neg A \cup B')/P(\neg A) \geq P(A \cup \neg B)/P(\neg A) = conf(\neg A \Rightarrow B) \geq \text{minconf\_NH}$.

(2) Meanwhile, $corr(\neg A, B) \leq -\text{min\_}corr$ according to Definition 7. That is to say, $corr(\neg A, B) = P(\neg A \cup B) - P(\neg A)P(B) \leq -min\_corr$. According to Definition 5, $corr(\neg A, B') = P(\neg AB') - P(\neg A)P(B')$, and $P(\neg AB') \geq P(\neg AB)$, $P(B') \geq P(B)$ for the reason that $B' \subseteq B$. In this way, $corr(A, B') = P(\neg AB') - P(\neg A)P(B') \geq P(\neg AB) - P(\neg A)P(B') \geq P(\neg AB) - P(\neg A)P(B) = corr(\neg A, B)$. Consequently, $corr(\neg A, B') \geq corr(\neg A, B)$, and $corr(\neg A, B') \leq -min\_corr$ holds with the inequality $corr(A, B) \leq -min\_corr$.

Using a similar way, it is easy to show that the rule $\neg A \Rightarrow B'$ can be inferred from the rule $\neg A \Rightarrow B$ if and only if B'⊆B when $|sup(A) - sup(B)| \leq \varepsilon\text{min}$, The proof is complete. □

Theorem A.13. *Let I be a set of items, both A and B are nonempty sets, $A \subseteq I$, $B \subseteq I$, and $A \cap B = \varnothing$, min_corr is the minimum correlation threshold, rule $\neg A \Rightarrow \neg B'$ is a strong negative association rule mined using algorithm PNARs_M. If there also exists a valid negative association rule $\neg A \Rightarrow \neg B$ in NARs satisfying $B' \subseteq B$, then $\neg A \Rightarrow \neg B$ is a redundant rule of $\neg A \Rightarrow \neg B'$.*

Proof. It is necessary to prove the following inequalities according to Definition 7: $corr(A, B) \geq min\_corr$ and $conf(\neg A \Rightarrow \neg B) \geq minconf\_P$. Also, the following inequalities hold from the given rule, $corr(A, B') \geq min\_corr$ and $conf(\neg A \Rightarrow \neg' B) \geq minconf\_P$.

(1) From the known, $B' \subseteq B$, then $P(B') \geq P(B)$, $P(AB') \geq A \Rightarrow B\, P(AB)$. Then, $corr(A, B) = P(AB) - P(A) \times P(B) \geq P(AB') - P(A) \quad P(B) \geq P(AB') - P(A) \quad P(B') = corr(A, \text{B}') \geq min\_corr$.

(2) Because of $B' \subseteq B$, then $\neg B' \supseteq \neg B$, that is to say, $\neg B \subseteq \neg \text{B}'$, and $|\neg B| \geq |\neg B'|$, where $|A|$ represents the number of set A. Then, $P(\neg B) \geq P(\neg B')$, $P(\neg A \neg B) \geq P(\neg A \neg B')$.

According to Definition 7, $\text{conf}(\neg A \Rightarrow \neg B) = P(\neg A \neg B)/P(\neg A) \geq P(\neg A \neg B')/P(\neg A) = \text{conf}(\neg A \Rightarrow \neg B') \geq min\_conf\_P$.

From (1) and (2), as can be concluded that $\neg A \Rightarrow \neg B$ is a redundant rule of $\neg A \Rightarrow \neg B'$ for any itemset B' satisfying $B' \subseteq \text{B}$.

The proof is complete. □

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 that they have no conflicts of interest.

Acknowledgments

The authors are grateful to the Shandong Province Audit Bureau in China for providing data support, which has made this research possible. This work was supported partially by the Natural Science Foundation of Shandong Province of China (ZR2020MG031, ZR2022MG024) and the Ministry of Education, Humanities, and Social Sciences Project of China (22YJAZH113).

References

[1] H. Brown-Liburd, H. Issa, D. Lombardi, and D. R. Lombardi, "Behavioral implications of big data's impact on audit judgment and decision making and future research directions," *Accounting Horizons*, vol. 29, no. 2, pp. 451–468, 2015.

[2] M. Seera, C. P. Lim, A. Kumar, L. Dhamotharan, and K. H. Tan, "An intelligent payment card fraud detection system," *Annals of Operations Research*, vol. 3, pp. 1–23, 2021.

[3] J. Tang and K. E. Karim, "Financial fraud detection and big data analytics-implications on auditors' use of fraud

brainstorming session," *Managerial Auditing Journal*, vol. 34, no. 3, pp. 324–337, 2019.

[4] A. Cardoni, E. Kiseleva, and F. De Luca, "Continuous auditing and data mining for strategic risk control and anticorruption: creating fair value in the digital age," *Business Strategy and the Environment*, vol. 29, no. 8, pp. 3072–3085, 2020.

[5] J. Gao and Gao, "Analysis of the financial internal control strategies of SME based on the background of big data," *Technium Social Sciences Journal*, vol. 32, no. 1, pp. 352–358, 2022.

[6] J. J. O'Leary and O'Leary, "The auditor's responsibility to detect errors, irregularities, and illegal acts by clients," *Journal of Corporate Accounting & Finance*, vol. 1, no. 3, pp. 239–253, 1990.

[7] K. Hummel, C. Schlick, M. Fifka, and Fifka, "The role of sustainability performance and accounting assurors in sustainability assurance engagements," *Journal of Business Ethics*, vol. 154, no. 3, pp. 733–757, 2019.

[8] T. Shang, F. Zhang, X. Chen, J. Liu, X. Lu, and Lu, "Identity-based dynamic data auditing for big data storage," *IEEE Transactions on Big Data*, vol. 7, no. 6, pp. 913–921, 2021.

[9] Y. Yu, S. Cai, & Z. Kong, and Ning, "Efficient and secure identity-based public auditing for dynamic outsourced data with proxy," *Ksii Transactions on Internet & Information Systems*, vol. 11, no. 10, pp. 5039–5061, 2017.

[10] H. Behera and D. Mohapatra, "Positive and negative association rule mining using correlation threshold and dual confidence approach," *Springer India*, vol. 26, pp. 249–260, 2016.

[11] P. Kishor and S. Porika, "An efficient approach for mining positive and negative association rules from large transactional databases," in *Proceedings of the International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, India, August 2016.

[12] M. Antonie and O. Zaïane, "Mining positive and negative association rules: an approach for confined rules," in *Proceedings of the Knowledge Discovery in Databases*, pp. 27–38, PKDD, Pisa, Italy, September 2004.

[13] X. Dong, F. Hao, L. Zhao, and T. Xu, "An efficient method for pruning redundant negative and positive association rules," *Neurocomputing*, vol. 393, pp. 245–258, 2020.

[14] Z. Tan, H. Yu, W. Wei, J. Liu, and Tan, "Top-k interesting preference rules mining based on maxclique," *Expert Systems with Applications*, vol. 143, pp. 113043.1–113043.18, 2020.

[15] S. Bashir, "An efficient pattern growth approach for mining fault tolerant frequent itemsets," *Expert Systems with Applications*, vol. 143, pp. 113046.1–113046.15, 2020.

[16] Y. Li, Z. H. Zhang, W. B. Chen, and F. Min, "TDUP: an approach to incremental mining of frequent itemsets with three-way-decision pattern updating," *International Journal of Machine Learning and Cybernetics*, vol. 8, no. 2, pp. 441–453, 2017.

[17] L. Wang, S. L. Li, H. Sun, K. X. Peng, and Peng, "A classification and regression algorithm based on quantitative association rule tree," *Journal of Intelligent and Fuzzy Systems*, vol. 31, no. 3, pp. 1407–1418, 2016.

[18] S. Chakraborty and S. Biswas, "Prediction and analysis on covid-19 using positive and negative association rule mining," in *Proceedings of Research and Applications in Artificial Intelligence*, pp. 1–11, Springer, Singapore, 2021.

[19] R. Agrawal and T. Imieliński, "Mining association rules between sets of items in large databases," in *Proceeding of the Acm Sigmod International Conference on Management of Data*, pp. 207–216, Philadelphia, PA, USA, June 1993.

[20] J. Han, J. Pei, Y. Yin, and R. Mao, "Mining frequent patterns without candidate generation: a frequent-pattern tree approach," *Data Mining and Knowledge Discovery*, vol. 8, no. 1, pp. 53–87, 2004.

[21] R. Brin and C. Motwani, "Beyond market baskets: generalizing association rules to correlations," in *Proceeding of the 1997 ACM SIGMOD International Conference on Management of Data*, pp. 265–276, Tucson, AZ, USA, May 1997.

[22] T. Azeez and S. Ayemobola, "Network intrusion detection with a Hashing based Apriori algorithm using Hadoop MapReduce," *Computers*, vol. 8, no. 4, pp. 86–100, 2019.

[23] A. Caratas, "A conceptual model for the analysis of internal audit effectiveness designed to improve corporate governance quality," in *Proceedings of Vision 2020: Innovation, Development Sustainability and Economic Growth*, pp. 1–12, USA, 2021.

[24] M. Shaheen and U. Abdullah, "CARM: context based association rule mining for conventional data," *Computers, Materials & Continua*, vol. 68, no. 3, pp. 3307–3322, 2021.

[25] M. Mansoor, Z. Rehman, M. Shaheen, M. A. Khan, and M. Habib, "Deep learning based semantic similarity detection using text data," *Information Technology and Control*, vol. 49, no. 4, pp. 495–510, 2020.

[26] M. Khan, "From data mining to wisdom mining," *Journal of Information Science*, vol. 49, no. 4, pp. 1–24, 2021.

[27] S. Khan, "Wisrule: first cognitive algorithm of wise association rule mining," *Journal of Information Science*, vol. 50, no. 5, pp. 1–20, 2022.

[28] M. Shaheen, M. Shahbaz, and M. Shahbaz, "An algorithm of association rule mining for microbial energy prospection," *Scientific Reports*, vol. 7, no. 1, Article ID 46108, 2017.

[29] S. Shahbaz, M. Ahsan, R. Shaheen, S. Nawab, and A. Masood, "Automatic generation of extended ER diagram using natural language processing," *Journal of American Science*, vol. 7, no. 8, pp. 1–10, 2011.

[30] X. Xiaoying and Z. Yingtao, "A novel sampling learning based approach to algorithm design for effective mining of association rules," *Journal of Zhejiang Normal University (Natural Sciences)*, vol. 41, pp. 44–49, 2018.

[31] L. Zhou, L. Cai, L. Jiang, and L. Chen, "Power grid enterprise intelligent risk identification model considering multi-attribute and low correlation data," *IEEE Access*, vol. 7, pp. 111324–111331, 2019.

[32] W. Ariya, "An enhanced incremental association rule discovery with a lower minimum support," *Artificial Life and Robotics*, vol. 21, no. 4, pp. 1–12, 2016.

[33] F. Dong, X. Sun, and R. Han, "Study of positive and negative association rules based on multi-confidence and chi-squared test," in *Proceeding of the International Conference on Advanced Data Mining & Applications*, pp. 100–109, Xian, China, August 2006.

[34] D. Liang, F. Lin, and S. Wu, "Electronically auditing EDP systems: with the support of emerging information technologies," *International Journal of Accounting Information Systems*, vol. 2, pp. 130–147, 2001.

[35] T. Wu, Y. Chen, and J. Han, "Re-examination of interestingness measures in pattern mining: a unified framework," *Data Mining and Knowledge Discovery*, vol. 21, no. 3, pp. 371–397, 2010.

[36] A. Swesi, "Mining positive and negative association rules from interesting frequent and infrequent itemsets," in *Proceedings of the International Conference on Fuzzy Systems & Knowledge Discovery*, pp. 650–655, Chongqing, China, May 2012.

[37] P. Bemarisika and A. Totohasina, “An efficient method for mining informative association rules in knowledge extraction,” in *Proceedings of the International Federation for Information Processing 2020, Machine Learning and Knowledge Extraction*, pp. 227–247, Dublin, Ireland, August 2020.

[38] X. Yongshun, D. Tiantian, and L. Xiangjun, “e-NSPFI: efficient mining negative sequential pattern from both frequent and infrequent positive sequential patterns,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 31, no. 2, Article ID 175002, 2016.

[39] H. Sahu and K. Gmz, “A dual approach for credit card fraud detection using neural network and data mining techniques,” in *Proceeding of the IEEE 17th India Council International Conference (INDICON)*, pp. 1–7, New Delhi, India, December 2021.

[40] S. Parkinson, V. Somaraki, and R. Ward, “Auditing file system permissions using association rule mining,” *Expert Systems with Applications*, vol. 55, pp. 274–283, 2016.

[41] S. Zhao, H. Tu, and Y. Chen, “Efficient association rule mining algorithm based on user behavior for cloud security auditing,” in *Proceeding of the 2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS)*, pp. 145–149, Chongqing, China, May 2016.

[42] J. Estrada, “Fraud detection using the fraud triangle theory and data mining techniques: a literature review,” *Computers*, vol. 10, pp. 1–22, 2021.

[43] N. Singh, K. Lai, M. Vejvar, and T. C. E. Cheng, “Data-driven auditing: a predictive modeling approach to fraud detection and classification,” *Journal of Corporate Accounting & Finance*, vol. 30, no. 3, pp. 64–82, 2019.

[44] I. T. Dan, “The auditor’s responsibility for finding errors and fraud from financial situations: case study,” *International Journal of Academic Research in Accounting, Finance and Management Sciences*, vol. 7, pp. 342–352, 2017.

[45] D. Moon, H. Im, I. Kim, and J. H. Park, “DTB-IDS: an intrusion detection system based on decision tree using behavior analysis for preventing APT attacks,” *The Journal of Supercomputing*, vol. 73, no. 7, pp. 2881–2895, 2017.

[46] Y. Djenouri, D. Djenouri, J. C. W. Lin, and A. Belhadi, “Frequent itemset mining in big data with effective single scan algorithms,” *IEEE Access*, vol. 6, pp. 68013–68026, 2018.

[47] H. Lu, C. B. Sivaparthipan, and A. Antonidoss, “Improvement of association algorithm and its application in audit data mining,” *Journal of Interconnection Networks*, vol. 22, no. 3, Article ID 2144002, 2021.

[48] D. Seong and U. Lee, “Group-wise keyword extraction of the external audit using text mining and association rules,” *Journal of Korean Society for Quality Management*, vol. 50, no. 1, pp. 77–89, 2022.

[49] Y. Zhang, C. Liao, Y. Shang, J. Feng, W. Du, and X. Zhong, “Application of extended matrix pencil method in multiport frequency-dependent network equivalent and the transient analysis of multiconductor transmission line system,” *IEEE Transactions on Power Delivery*, vol. 38, no. 1, pp. 95–104, 2023.

[50] W. Xiufeng and L. Zhanlong, “Research on mining positive and negative association rules based on dual confidence,” in *Proceeding of the 5th International Conference on Internet Computing for Science and Engineering*, pp. 102–105, Harbin, China, November 2011.

[51] I. Cafaro and M. Epicoco, “Data mining: mining frequent patterns, associations rules, and correlations,” *Encyclopedia of Bioinformatics and Computational Biology*, vol. 1, pp. 358–366, 2019.

[52] A. Ghazikhani, R. Monsefi, H. Sadoghi Yazdi, and S. Yazdi, “Online neural network model for non-stationary and imbalanced data stream classification,” *International Journal of Machine Learning and Cybernetics*, vol. 5, no. 1, pp. 51–62, 2014.

[53] W. Shen, G. Yang, J. Yu, H. Zhang, F. Kong, and R. Hao, “Remote data possession checking with privacy-preserving authenticators for cloud storage,” *Future Generation Computer Systems*, vol. 76, no. 9, pp. 136–145, 2017.

[54] C. Liu, “Two-level confidence threshold setting method for positive and negative association rules,” *Journal of Computer Applications*, vol. 38, no. 5, pp. 1315–1319, 2018.

[55] H. Jin, K. Zhou, H. Jiang, D. Lei, R. Wei, and C. Li, “Full integrity and freshness for cloud data,” *Future Generation Computer Systems*, vol. 80, pp. 640–652, 2018.

[56] L. Zhao, T. Xu, F. Hao, and X. Dong, “Positive and negative association rules mining for mental health analysis of college students,” *Eurasia Journal of Mathematics, Science and Technology Education*, vol. 13, no. 8, pp. 5577–5587, 2017.

[57] Y. Djenouri and M. Comuzzi, “Combining Apriori heuristic and bio-inspired algorithms for solving the frequent itemsets mining problem,” *Information Sciences*, vol. 420, pp. 1–15, 2017.

[58] F. Agustina, N. Nurkholis, and M. Rusydi, “Auditors professional skepticism and fraud detection,” *International Journal of Research in Business and Social Science*, vol. 10, no. 4, pp. 275–287, 2021.

[59] X. G. Wu and Y. Du, “An analysis on financial statement fraud detection for Chinese listed companies using deep learning,” *IEEE Access*, vol. 10, pp. 22516–22532, 2020.

[60] A. Paul, “Positive and negative association rule mining using correlation threshold and dual confidence approach,” in *Proceedings of the 2016 International Conference on Computational Intelligence in Data Mining*, pp. 249–260, Bhubaneswar, Odisha, India, December 2016.

[61] B. Le, D. Phuong Le, and M. T. Tran, “Hiding sensitive association rules using the optimal electromagnetic optimization method and a dynamic bit vector data structure,” *Expert Systems with Applications*, vol. 176, Article ID 114879, 2021.

[62] X. G. Wu and S. Y. Du, “An optimized association rules mining framework and its application in Chinese social insurance fund data auditing,” 2022, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4292653.