Establishment and Improvement of Financial Decision Support System Using Artificial Intelligence and Big Data

Cong Si1, Wenquan Shi2*

1Suzhou institute of trade and commerce, Suzhou, Jiangsu Province, China, 215009
2Faculty of Economics and Management, Suzhou Polytechnic Institute of Agriculture, Suzhou, Jiangsu Province, China, 215008

Corresponding author email: swq02000@163.com

Abstract. The traditional financial decision support is not suitable for the intelligent development of enterprise. In order to improve the intelligence and real-time performance of enterprise financial decision-making, and facilitate managers and decision-makers to master comprehensive information and make better judgments, based on artificial intelligence (AI) and big data technology, ID3 algorithm and association rule mining algorithm are adopted to analyze and optimize the financial decision support system (FDSS). The basic knowledge of ID3 algorithm and association rule mining algorithm are introduced in turn. The two algorithms are analyzed and summarized to find their shortcomings. In view of the shortcomings of the algorithm, the improved algorithm is proposed, and its application is analyzed. The results show that the improved algorithm improves the efficiency of FDSS, and has advantages for the establishment and improvement of FDSS. Therefore, the proposed AI and big data will improve the accuracy, automation and timeliness of financial decision-making.

Keywords: Artificial Intelligence, Big Data, Financial Decision Support System, Id3 Algorithm and Association Rule Mining Algorithm

1. Introduction

The rapid development of modern society and market economy make the competition between enterprises fiercer and fiercer. To survive and develop in the market, enterprises must innovate financial management and use the technology and theory of artificial intelligence (AI) and big data, so as to build a suitable system mechanism, which can promote its intelligent development and facilitate managers to master more valuable information to make effective decisions, and achieve the purpose of reducing costs and improving benefits [1]. In foreign countries, the research of financial decision support system (FDSS) is relatively early. In 1970, Michael Scott Merton and Thomas Genity put forward the concept of decision support system [2]. After a long period of theoretical and application research, great development has been made. In China, the development and research of decision support system began in the 1980s. At present, the financial management systems of UFIDA and Kingdee are mainly used, with a small number. Moreover, the current domestic intelligent financial
software is limited to the description of data [3]. The main problem is to integrate the theoretical experience of scholars into the system for use [4]. ID3 algorithm and association rule mining algorithm are analyzed and optimized to realize and develop the design of system module in intelligent decision-making, and make the analysis and evaluation of investment risk more comprehensive, so as to improve the establishment and improvement of FDSS.

2. Method
In a large number of financial data, the hidden information and valuable information are found. Based on AI and big data technology, ID3 algorithm and association rule mining algorithm are used to explain and reason complex problems. These new technologies provide strong technical support for financial analysis, information collection system and financial functionalization. They can be fully used to establish an intelligent analysis system scheme of operation and financial management based on knowledge data, realize the development of financial accounting into decision-making in an information system, and provide a helpful intelligent human-computer interaction system for managers and decision-makers, which is conducive to the analysis and management of business mode, cost control and financial status of enterprises.

2.1. ID3 algorithm
(1) Definition of ID3 algorithm
ID3 algorithm was put forward by Quinlan in 1986. It is a classic decision tree classification algorithm constructed by induction and classification model from data [5]. The inner node of decision tree represents the value range of this attribute. Each internal node corresponds to a non-class attribute of the sample; the attribute basis of the test is the information gain; the internal node is associated with the largest non-class attribute of information. The basic idea is to take the information in the database as the node, split the tree root according to the similar value of each attribute, and establish the branch node; then, each established branch node is taken as the root node, so as to split the similar values of each attribute repeatedly, and divide them uniformly according to this method until the final direct point only contains the positive examples or the counter examples. According to this method, a complete decision tree is constructed, which can realize the classification of new samples [6].

The method of calculating attribute information is to calculate the attribute information gain of each attribute, and then compare their size to obtain the attribute with the maximum information gain [7].

If there are X samples and the number of the values with different attributes is n, n of the same kind is defined as:

$$I(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} p_i \log_2 p_i$$  \hspace{1cm} (1)

$P_i = x_i/x$ is the probability that any sample belongs to $C_i$. If the attribute $W$ has $m$ different values, according to the divided attributes ($w_1, w_2, w_3 \ldots w_m$), $x$ is divided into $m$ subsets ($x_1, x_2, x_3 \ldots x_m$). $w_j (j = 1, 2, 3 \ldots v)$ is the same value of sample on attribute $w$ in $x_j$. If $x_j$ belongs to the sample number similar to $C_i$ in $x_j$, the entropy or expected value of $x_j$, which is divided according to $x$, is as follows.

$$E(W) = -\sum_{i=1}^{v} P_{ij} \log_2 P_{ij}$$  \hspace{1cm} (2)

In the above equation, the lower the entropy is, the higher the purity of subset partition is. $P_{ij} = x_{ij}/x_j$ is the probability of $C_i$ in the sample $x_j$. Through the above equation, it can be obtained that the information gain of attribute $W$ is $\text{Gain}(W) = I(x_1, x_2, x_3, \ldots x_n) - E(W)$.

(2) Improvement of ID3 algorithm
The traditional ID3 algorithm uses the information gain method, which is usually more inclined to
select the attribute with more values. However, in the actual application, many attributes with more values are not always very important attributes, so it is necessary to make a choice and optimize the ID3 algorithm.

If the training real number set is $S$, $S$ is divided into class $V$, $|S|$ is the total number of instances, and the number of instances of class $i$ is $|S_i|$, the probability that an instance belongs to class $i$ is expressed as follows.

$$P(S_i) = \frac{|S_i|}{|S|}$$  \hspace{1cm} (3)

If the attribute is $A$ and the number of instances belonging to class $i$ is $|S_{ij}|$ when $A=aj$, the probability of belonging to class $i$ is as follows.

$$P\left( \frac{S_i}{A=a} \right) = \frac{|S_{ij}|}{|S_j|}$$  \hspace{1cm} (4)

$|S_j|$ denotes the number of instances of the numerator set when $A=aj$. If the instance obtained by $A=aj$ is represented as $Y_j$, the information quantity of attribute $A$, that is, the information gain, is expressed as follows.

$$Gain(A,S) = H(S) - H(S/A)$$  \hspace{1cm} (5)

In the above equation, the larger the value of $Gain(A, S)$ is, the greater the amount of classification information that attribute $A$ can provide is. It shows that when attribute $A$ is used as the sub-node of the tree, the certainty probability of classification increases.

If there are $P$ positive examples and $N$ counterexamples, the values of attribute $A$ is taken as $a_1$, $a_2$, ..., $a_i$. The instance set generated by $A=aj$ is expressed as follows.

$$ \sum_{j=1}^{i} -p_j \left( -\frac{n_j}{n_j + p_j} \right) - n_j \left( -\frac{p_j}{n_j + p_j} \right) = 2 \sum_{j=1}^{i} \frac{p_j n_j}{n_j + p_j}$$  \hspace{1cm} (6)

In the above equation, $n_j$ is the number of counter examples, $p_j$ is the number of positive examples, and $n_j+p_j$ is the total real number examples of this subset.

2 in the above equation is a constant. After simplification, the following equation can be obtained.

$$Gain'(A,S) = \sum_{j=1}^{i} \frac{p_j n_j}{n_j + p_j}$$  \hspace{1cm} (7)

$Gain'(A,S)$ is regarded as the new information gain of attribute $S$, which can simplify the equation for calculating information gain. The information gain of simplified attributes is calculated by the above equation, and the attribute with the smallest value after calculation is the split attribute.

The experimental training sample set is used to test the algorithm. In order to ensure the accuracy and universality of experimental data, the number of experiments with the same data is added in this experiment. The samples are divided into several groups, and each group is given 10 data experiments. By calculating the average time, the experiment and experimental data on the computer with the same data, the same configuration and the same operating system are analyzed [8].

2.2. Association rule mining algorithm

(1) Definition of association rule algorithm

Association rule algorithm is a kind of important algorithm in data mining, which reflects the knowledge of dependency or association between an event and other events. It is mainly used to find
valuable association between item sets in big data. This algorithm, also known as Apriori algorithm [9], can be applied to sales decision-making, investment decision-making, statement judgment and the judgment on whether there is fraud in enterprise finance. This algorithm is to conduct inventory analysis on the sales situation of enterprises through association rules, and analyze the commodities of enterprises by using AI and big data technology, so as to find out a certain regularity, facilitate the arrangement of purchasing commodities or the decision of managers on commodities, find a better mode suitable for customers' purchase, achieve intelligent decision-making, promote the recycling of enterprise funds, and maximize profits.

Association rule algorithm is an iterative method based on the pruning techniques, which searches one by one at each level. First, it is necessary to scan the given database and calculate the number of times each item appears by AI. According to the calculated minimum support, the most frequent dataset 1 is found out. Then, it is necessary to continue to generate frequent dataset 2 based on frequent dataset 1, and so on.

Here, the support degree is used to determine the frequency of the dataset, the confidence level is used to determine the frequency of a thing in the inclusion of another thing, the expected confidence is used to express the probability of a thing appearing in all things, and the lift is to describe the influence of the appearance of a thing on another.

If \( W = (n_1, n_2, n_3, \ldots, n_m) \) is a set, the set of datasets is \( D \), the set of items is \( C \), and \( W \) contains \( C \), there is only one representation of each object. For example, the number of something is \( CWD \) [10].

**Definition 1: support degree**

In database \( D \), if \( I \) is the thing in \( D \), and it contains the percentage of \( A \) and \( B \) at the same time, according to the rule \( A \rightarrow B \), the probability of support rate is \( P(AB) \), which is expressed as follows.

\[
I(A \rightarrow B) = P(A \rightarrow B) = \frac{|AB|}{|D|} 
\]  

(8)

In the above equation, \(|D|\) represents the quantity recorded in database \( D \), and \(|AB|\) represents the number of simultaneous occurrences of \( AB \) in a database.

**Definition 2: confidence level**

If \( E \) is the confidence level of rule \( A \rightarrow B \), it also contains \( A \) and \( B \) while representing \( E \). Compared with the percentage containing \( A \) itemset, the conditional probability of its confidence level is expressed as follows.

\[
E(A \rightarrow B) = P(AB | A) = \frac{|AB|}{|A|} 
\]  

(9)

In the equation, \(|A|\) represents the quantity recorded in database \( A \).

**Definition 3: threshold**

After two thresholds of minimum confidence level and minimum support degree are set, the rules that satisfy the conditions of interest are found from the huge database.

**Definition 4: itemset and frequent itemset**

The collection of each item is called an itemset. 1 item set is called 1 itemset, \( k \) item set is called \( k \) itemset, and item set satisfying minimum support degree is called frequent \( k \) itemset.

**Definition 5: lift degree**

\[
X(A \rightarrow B) = \frac{P(AB)}{P(A)P(B)} 
\]  

(10)

According to the equation, the lift degree also reflects the influence of \( A \) on \( B \). The greater the lift degree is, the greater the impact of itemset \( A \) on itemset \( B \) is. If \( X(A \rightarrow B) = 1 \), \( P(AB) = P(A)P(B) \), which shows that the appearance of \( A \) and \( B \) has no obvious influence on each other. If \( X(A \rightarrow B) < 1 \),
it means that with the appearance of A, B will not appear. If \( X (A \rightarrow B) > 1 \), it means that the appearance of A and B have influence on each other, and the appearance of A often leads to the appearance of B [11].

(2) Improvement of association rule mining algorithm

There are two defects in Apriori algorithm. First, a large number of key-value pairs will be generated, which will occupy a large amount of memory and consume a lot of network resources in the process of transmission. Second, the algorithm needs to scan the original dataset every time when it counts the local support count, which will generate many operations, and leads to low efficiency. Therefore, the algorithm is improved.

It is necessary to reduce the number of visits and change the generation process of frequent itemset. The candidate K itemset is formed by combining frequent K-1 itemset and frequent 1 itemset. The support count of the candidate set obtained needs to scan the original dataset many times. Therefore, it is necessary to vertically divide the experimental data, convert the rows and columns of the experimental dataset, and then make statistics on the transformed dataset. If it satisfies the minimum support degree, the item is a frequent itemset. If not, it means that the item is an infrequent itemset. In this case, it is necessary to delete the itemset that cannot meet the minimum support degree, and get the frequent 1 itemset. Then, the frequent 1 itemset and the frequent 1 itemset are connected, and the corresponding intersection is taken to generate the candidate 2 itemset. The candidate 2 itemset is filtered and pruned to obtain the frequent 2 itemset. Then, the frequent 2 itemset and the frequent 1 itemset are connected to generate the candidate 3 itemset. By analogy, frequent K-1 itemset and frequent 1 itemset are combined to generate frequent K itemset [12].

The algorithm is further improved and divided into two parts: generating frequent 1 item and generating frequent K item. The whole process needs K running processes. In the first process, the frequent 1 itemset is generated, which reads raw data. In the second process, the process frequent 2 itemset is generated, and candidate 2 itemset is generated by using the partitioned partial frequent 1 itemset and frequent 1 itemset. The generated 2 itemset is summarized to generate frequent 2 itemset. According to this process, the K-1 itemset after partitioning are connected with frequent 1 itemset to generate candidate K itemset. At the end of each running process, the generated frequent itemset will be divided before the next run, so that each partitioned frequent itemset is connected with frequent 1 itemset to generate a higher frequent itemset.

In order to reduce the amount of network transmission, a large number of key-value pairs will be generated in the process of running. If it is directly transmitted through the network, it will definitely occupy more network resources. Therefore, the optimization method is to add a Count process after the execution. In this process, the same key-value pairs in the running key-value pairs can be statistically merged. For example, if \((X1,Y1)\) and \((Y1,W1)\) are merged, the result of merging is \((Y1,X1,W1)\), so as to reduce the number and burden of transmission and improve the efficiency of processing.

3. Results and discussion

3.1. Performance comparison of ID3 algorithm before and after improvement

Figure 1 shows the time cost and accuracy of ID3 algorithm on the decision tree before and after improvement.
Figure 1. Comparison of ID3 improved algorithm and original algorithm.

Figure 1 shows that with the increase of the number of records, the average time of the ID3 algorithm before and after improvement is on the rise as a whole. However, the average time of the improved ID3 algorithm is obviously less than that of the original ID3 algorithm, which fully proves that the improved ID3 algorithm is feasible.

3.2. Application of ID3 algorithm in investment decision

Investment decision is an important part of intelligent financial support system, and the determination of risk plays an important role in the return of investment. Data mining is used to analyze and judge the investment project or scheme, select the most appropriate and optimal scheme, and then get the maximum return. ID3 algorithm is used to calculate the index, determine the branch node, and get the decision content, so as to effectively predict the probability of risk occurrence, results and development trend, and evaluate the project risk. Investment risk is taken as an example. From the bank's long-term loan capital, retained earnings capital and capital flow ratio, the risk of investment and the factors that have the greatest impact on the investment risk are analyzed and judged. The simplified ID3 algorithm is illustrated.

ID3 algorithm is used to simplify the equation. First, the information gain value of each attribute is calculated.

\[
Gain'(cfr) = \sum_{j=1}^{i} \frac{p_j n_j}{n_j + p_j} = \frac{2 \times 3}{2+3} + \frac{4 \times 0}{4+0} + \frac{3 \times 2}{3+2} = 2.4
\]  

(11)

\[
Gain'(bltlc) = \sum_{j=1}^{i} \frac{p_j n_j}{n_j + p_j} = \frac{2 \times 2}{2+2} + \frac{4 \times 2}{4+2} + \frac{3 \times 1}{3+1} = 3.083
\]  

(12)

\[
Gain'(rec) = \sum_{j=1}^{i} \frac{p_j n_j}{n_j + p_j} = \frac{3 \times 4}{3+4} + \frac{6 \times 2}{6+1} = 2.571
\]  

(13)

The value of the information gain calculation of attribute shows that Gain’ (capital flow ratio) is the smallest, which indicates that the information about capital flow ratio is the most helpful for classification. Among the three factors tested, the proportion of capital flow ratio has the greatest
impact on investment risk, and the most important investment object to be considered is the capital flow ratio.

3.3. Application of association rule mining algorithm in investment decision

There are commodities 1, 2, 3, 4, 5, 6 selling things X1, X2, X3,…,X9, and the minimum support count is 2. Table 1 shows the data.

Table 1. Information of commodity sales.

| Sales serial number | Commodity A | Commodity B | Commodity C | Commodity D |
|---------------------|-------------|-------------|-------------|-------------|
| X1                  | I           | II          | III         |             |
| X2                  | IV          | II          |             | V           |
| X3                  | VI          | I           | II          | III         |
| X4                  | VI          | II          |             |             |
| X5                  | IV          | II          | III         |             |
| X6                  | IV          | I           | III         |             |
| X7                  | VI          | I           | II          | III         |
| X8                  | IV          | III         |             |             |
| X9                  | IV          | I           | V           | III         |

Based on the sales of commodities, the database is scanned and the Hash table is used to record. The sales serial number carried by the records of each line in each item is the support count. The support count is obtained by the statistics of the included sales data. Therefore, the Hash table [13] is obtained as shown in Table 2.

Table 2. Hash table.

| Item | Sales serial number | Support count |
|------|---------------------|---------------|
| I    | X1, X3, X6, X7, X9  | 5             |
| II   | X1, X2, X3, X4, X5, X7 | 6         |
| III  | X1, X3, X5, X6, X7, X8, X9 | 7       |
| IV   | X2, X5, X6, X8, X9  | 5             |
| V    | X2, X9              | 2             |
| VI   | X3, X7              | 3             |

Table 2 shows that in the frequent 1 itemset {I: 5, II: 6, III: 7, IV: 5, VI: 3}, when the 2 itemset is judged, it is only necessary to scan the sales serial number of the same row of the item in the Hash table. In this way, the support can be obtained. The final frequent 2 itemset is obtained as {VI, II}: 3, {III, IV}: 4, {I, II}: 3, {I, III}: 5, {II, III}:4}. By using this method, the frequent 3 itemset can be obtained, and the support count is 3. According to the frequent itemset, the association rules with a confidence level of 60% or above are obtained [14]. Table 3 shows the filtered association data.

Table 3. Filtered association rules.

| Association rules | Confidence level | Expected confidence | Lift |
|-------------------|------------------|---------------------|------|
| {I, II} = {III}   | 100%             | 7/9                 | 1.29 |
| {I, III} = {II}   | 60%              | 6/9                 | 0.9  |
| {II, III} = {I}   | 75%              | 5/9                 | 1.35 |
| {I} = {II, III}   | 60%              | 4/9                 | 1.35 |

According to the information in Table 3, the law of commodity sales can be found. For example, the rule {I, II} = {III} shows that consumers are more inclined to buy commodity III when they buy commodity I and commodity II. Therefore, it can be convenient for managers to place commodities,
purchase commodities and use funds effectively.

4. Conclusion
In recent years, with the continuous progress of software development technology, a variety of new ideas, design patterns and technical methods have emerged in the field of object-oriented. The continuous progress of AI technology and big data provides technical support and guarantee for the development of intelligent decision support system, and further enhances the maintainability of software. Based on the background of AI and big data, the FDSS has been established and improved. ID3 algorithm and association rule mining algorithm are mainly used. The basic concepts and knowledge of these two algorithms are introduced and their defects and deficiencies in the application are analyzed. In view of the existing problems, the two algorithms are optimized and improved, so that they can be better applied in FDSS. Then, the two improved algorithms are analyzed by examples to verify the effectiveness and optimization of the improved algorithm.

The research results have theoretical and practical significance for the research of enterprises in intelligent decision support system. Intelligent FDSS is in the developing stage. On the basis of this topic, the future research will focus on the combination of AI, machine learning technology and FDSS, so as to strengthen the analysis of the decision-making research of intelligent financial system, and make the system more intelligent and the decision-making process more effective, intelligent and accurate.

Acknowledgments
Phased research results of the sub-projects of the Industry-University Cooperation Cooperative Education Project of the Department of Higher Education, Ministry of Education (201902323030816).

References
[1] Ferreira O, Plomaritis T A, Costas S. Effectiveness assessment of risk reduction measures at coastal areas using a decision support system: Findings from Emma storm[J]. The ence of the Total Environment, 2019, 657(MAR.20), pp. 124-135.
[2] Bhornsawan T. Bayesian-Based Decision Support System for Assessing the Needs for Orthodontic Treatment[J]. Healthcare Informatics Research, 2018, 24(1), pp. 22-28.
[3] Wilkinson T A, Dixon B E, Xiao S, et al. Physician clinical decision support system prompts and administration of subsequent doses of HPV vaccine: A randomized clinical trial[J]. Vaccine, 2019, 37(31), pp. 4414-4418.
[4] Kohler M, Vellasco M, Silva E, et al. SimProxy Decision Support System: A Neural Network Proxy Applied to Reservoir and Surface Integrated Optimization[J]. IEEE Systems Journal, 2020, (99), pp. 1-11.
[5] Yu H H, Chang K H, Hsu H W, et al. A Monte Carlo Simulation-based Decision Support System for Reliability Analysis of Taiwan's Power System: Framework and Empirical Study[J]. Energy, 2019, 178(JUL.1), pp. 252-262.
[6] Gudmundsson H T, Hansen K E, Halldorsson B V, et al. Clinical decision support system for the management of osteoporosis compared to NOGG guidelines and an osteology specialist: a validation pilot study[J]. BMC Medical Informatics and Decision Making, 2019, 19(1), pp. 27.
[7] Pratama Y, Saragi H S. Cassava Quality Classification for Tapioca Flour Ingredients by Using ID3 Algorithm[J]. Indonesian Journal of Electrical Engineering and Computer ence, 2018, 9(3), pp. 799-805.
[8] Savas S K, Nasibov E. A fuzzy ID3 induction for linguistic data sets[J]. International Journal of Intelligent Systems, 2018, 33(4), pp. 858-878.
[9] Vasquez J, Comendador B E. Competency Discovery System: Integrating the Enhanced ID3 Decision Tree Algorithm to Predict the Assessment Competency of Senior High School Students[J]. International Journal on Advanced ence Engineering and Information
Technology, 2019, 9(1), pp. 60.

[10] Jorge P A, Marcio D, Mario I P. A multi-objective gene clustering algorithm guided by apriori biological knowledge with intensification and diversification strategies[J]. Biodata Mining, 2018, 11(1), pp. 16.

[11] Zhao S, Huang B, Liu F. Linear Optimal Unbiased Filter for Time-Variant Systems Without Apriori Information on Initial Conditions[J]. IEEE Transactions on Automatic Control, 2017, 62(2), pp. 882-887.

[12] Rohde M, Bonn B R, Zimmermann M, et al. Relevance of ID3-TCF3-CCND3 pathway mutations in pediatric aggressive B-cell lymphoma treated according to the non-Hodgkin Lymphoma Berlin-Frankfurt-Münster protocols[J]. Haematologica, 2017, 102(6), pp. 1091-1098.

[13] Kyei F, Asante D B, Sarpong E, et al. Over-expression of Id2 and Id3 Proteins Regulates Growth and Survival of Human Colon Carcinoma (HCT116) Cells -[J]. Journal of Applied Biology & Biotechnology, 2017, 5(1), pp. 010-017.

[14] Kermarrec G, Schoen S. A priori fully populated covariance matrices in least-squares adjustment—case study: GPS relative positioning[J]. Journal of Geodesy, 2017, 91(5), pp. 465-484.