Research on Two-Stage Model of Project Risk Assessment Based on Knowledge Discovery

Liwei Zhang1,2 Dandan Zhao1
1 Information College, Capital University of Economics and Business, Beijing, China
2 Institute of Scientific and Technical Information of China, Beijing, China
zhangliwei19810@126.com

Abstract

Risk assessment is the most difficult and most time-consuming process of project risk management. In the complex and volatile market environment, the project risk is more difficult to predict. Although numerous projects which have been completed provide much valuable knowledge for reference, it has not been put into good use. Thus, it is urgent to study how to apply the completed projects to help new projects avoid risks effectively. Therefore, this paper proposes the two-stage model of project risk assessment, in which decision tree is applied for discovering the risk rules at the first stage, and case-based reasoning for utilizing the tacit knowledge at the second stage. The model consists of the following four parts: (1) to explore risks and association rules from the project risk cases by using decision tree technology; (2) to use and evaluate rules; (3) to obtain assessment results by using case-based reasoning method; (4) to extend case base and rule base through case study and rule learning. The preliminary tests show that the model can assess project risks more quickly and accurately and the assessment results are accurate and viable.

Keywords: project risk assessment; data mining; case-based reasoning

1 Introduction

With the global economic integration and the deepening reform of investment system, enterprises face more and more risks in the course of project development, such as technical risk, economic risk and so on. These risks not only increase the difficulty of project management, but also increase the probability of projects' potential loss. Only with the successful operation of projects, can enterprises survive and develop in the fierce competition. Therefore, in order to operate the project more effectively, the project risk assessment is particularly important[1].

2 Project Risk Assessment

Project risk assessment is a process of determining the probability of risk occurrence and the severity of its consequence, quantifying the probability and influencing scope of risk, as well as estimating and evaluating its social and economic impact.

Currently, there are lots of project risk assessment methods, mainly divided into two types of qualitative and quantitative methods. For example, the qualitative methods include fault tree analysis, extrapolation, and matrix analysis; the quantitative methods include breakeven analysis, program evaluation and review technique (PERT), and sensitivity analysis. While in recent years, the applications of a number of machine learning technologies such as neural networks[2], fuzzy sets[3], inductive logic programming[4], etc., in the field of project risk assessment have received extensive attention and recognition. But in most applications, only one type of machine language technologies is simply used, which lack system integration. Thus the application performance of machine learning language is weakened.

Therefore, this paper proposes the two-stage model of project risk assessment based on decision tree and case-based reasoning. The model consists of the following four parts: (1) to explore risk association rules from the project risk cases by using decision tree technology; (2) to use and evaluate risk association rules; (3) to obtain assessment results by using case-based reasoning method to match similar cases; (4) to extend case database and rule base through case study and rule learning. The model can assess the project risk quickly and effectively to obtain more accurate and practical assessment results.

3 Construction of Project Risk Assessment Model

In this paper, the decision tree method and the case-based reasoning technology are combined to be applied in the project risk assessment. The premise of establishing the system is to assume that all knowledge can be expressed by rules. Because almost all the knowledge can be acquired by knowledge acquisition tools, and generated by machine language, the knowledge can be expressed as rules easily. At the initial stage of system operation, the risk rule set are input for project risk assessment. The risk assessment system consists of two stages: the knowledge creation stage and the knowledge reasoning stage. System details are as follows:

At the knowledge creation stage, the project data are collected and mined. Firstly, the data set are preprocessed and the data quality is improved. Secondly, the core data set are selected as mining objects. Next, the relationship pattern of data is identified by using the decision tree induction algorithm to find project risk rules and case associated algorithm to mine association rules among project data.

At the knowledge reasoning stage, the cases to be evaluated are input, which can trigger project risk assessment system. Then, the rules discovered at the knowledge creation stage are used to predict the project risk level. Next, the most similar case to the input case is searched by using case-based reasoning algorithm according to the project risk level. Finally, to analyze the most similar

case and obtain the assessment result.

4 Knowledge Creation Stage

At this stage, because the raw data are not only huge but also heterogeneous, they need to be preprocessed firstly. Then, to analyze the case base with decision tree induction algorithm and case-based reasoning algorithm to recognize the relationship patterns among data.

4.1 Decision Tree Induction Algorithm

The decision tree induction algorithm is used to find the project risk rules, which consists of classification tree algorithm and regression tree algorithm.

1) As to the classification tree, the selection of splitting attribute at each stage depends on how much information each attribute contains, namely information entropy. On each branch point, the attribute of minimum entropy is chosen as the splitting attribute. For the discrete attributes, the attribute of minimum entropy is selected for split; for the continuous attributes, the threshold value which makes the entropy of attribute minimum is used for split. The entropy is calculated with the following expression [5]:

$$Entropy(A) = \sum_{i=1}^{2} \frac{\sum_{j=1}^{c} f_{ij}}{R} \sum_{j=1}^{c} P(c_{ij}) Ln(P(c_{ij})) \qquad (1)$$

The c is the number of output values; fij is the appearance frequency of output j in branch i; R is the record number in the two branches. P(cij) is expressed as following:

$$P(c_{ij}) = f_{ij} \Big/ \sum_{k=1}^{c} f_{kj} \qquad (2)$$

2) Regression Tree

For the regression tree, the attribute of minimum branch bias is chosen as the splitting attribute. The normalized standard deviation (NSD) is introduced to choose the splitting attribute. On each branch point, the attribute of minimum NSD is used as the splitting attribute. For the discrete attributes, the attribute of minimum NSD is chosen; for the continuous attributes, the threshold value to make the NSD of attribute minimum is chosen. NSD is calculated with the following formula [5]:

$$NSD(A) = \sum_{k=1}^{2} \frac{R_k}{R} \sqrt{\frac{\sum_{i=1}^{R_k} v_{ik}^2 - \left(\sum_{i=1}^{R_k} v_{ik}\right)^2}{R_k}} \qquad (3)$$

Where vik is the number of output value of record i in branch k; Rk is the number of records in branch k; R is the number of records in all branches.

4.2 Cases Association Algorithm

Case association algorithm is applied to mine association rules from the project risk data. The algorithm is divided into the two steps:

Step 1.Mine frequent item sets: to obtain frequent item sets whose support is

greater than the initial setting threshold value. The cost of the step is much higher than that of Step 2, and the overall performance of mining associated rules is determined by the step.

Step 2.Generate association rules: to generate the association rules by analyzing the frequent item-sets obtained in step 1.

Suppose X is a frequent item-set, and its support is Sx; Y is a item-set and its support is Sy; and $Y \Rightarrow X$. From the above, the possible association rule is $Y \Rightarrow (X-Y)$, in which (X-Y) represents a item-set including X except Y. If users set a minimum confidence, the rule which can meet $Sx/Sy \geq$the minimum confidence is selected from the above rules as the association rule.

Firstly, the database is retrieved to obtain the raw data, which need to be preprocessed and the core part of dataset is used as mining objects. Then, the selected dataset are analyzed by the decision tree induction algorithm and the case association algorithm analysis to identify the relation pattern among data. All the above rules are applied to analyze the test data in order to confirm these rules' accuracy. Finally, these rules are stored in the rule base and used to help new project to predict and avoid project risks.

5 Knowledge Reasoning Stage

At the knowledge reasoning stage, the specific case-based reasoning process is as follows:

5.1 Case-Based Reasoning Process

The process of the case-based reasoning is as follows:

Step 1.Input the risk cases. The risk cases which have completed are input into the case base;

Step 2.Retrieve the case base and match the similar cases. Through the relevant retrieve algorithm, similar cases are found;

Step 3.Confirm results. The result is divided into two types: one can obtain the corresponding assessment results as the final solution, bringing the end of reasoning; another has no match case, then turn Step 4;

Step 4. Adjust or amend risk cases. Combining systematic learning and artificial computing, the new case is assessed, which is treated as the final solution to this reasoning process.

Step 5. Learn case. To determine whether the new case is valuable and whether should be added to the case base.

The process is shown in Figure 1.

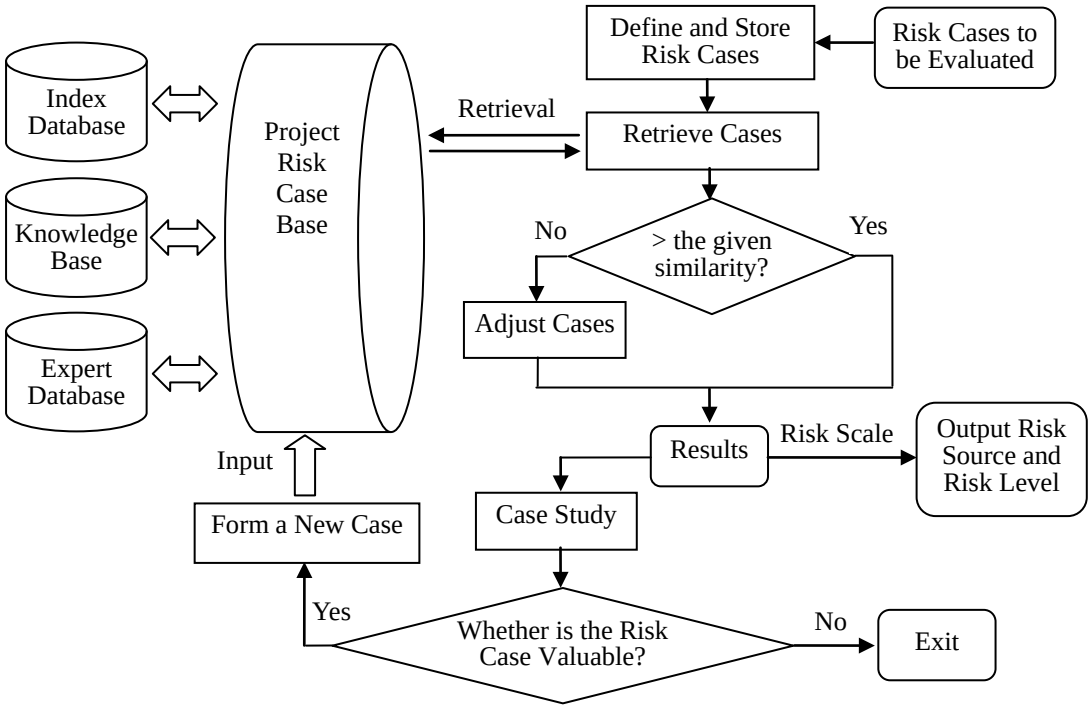

Figure 1 Project Risk Assessment Model Based on CBR

5.2 The Key Algorithm of Case-based Reasoning—Case Similarity Determination

Because the range and the magnitude order of data are different, the data should be normalized before determining the similarity of cases. Then, the similarities of numerical attributes and non-numerical attributes need to be determined separately[6].

(1) The similarity of numerical attribute:

$$d_{jk1} = \sum_{m=1}^{n} W_m d\left(I_{jm}, X_m\right) \qquad (4)$$

Where d(Ijm,Xm) represents the Euclidean distance in attribute m between the new case X and case Ij. Wm denotes the weight value of attribute m.

(2) The similarity of non-numerical attributes:

$$d_{jk2} = \sum_{h=1}^{n} W_h d\left(I_{jh}, X_h\right) \qquad (5)$$

Where d(Ijh,Xh) represents whether the new case X and case Ij have the same attribute h.

$$d\left(I_{jh}, X_h\right) = \begin{cases} 1 & identical \\ \alpha & similar \\ 0 & different \end{cases} \quad (6)$$

Weights:

$$\sum_{m=1}^{n} W_m + \sum_{h=1}^{n} W_h = 1 \quad (7)$$

Finally, the similarity function between the two cases is defined as:

$$Similarity(X, I_j) = d_{jk1} \odot d_{jk2}; \quad (8)$$

Where $\odot$ represents the synthesis between d_{jk1} and d_{jk2}.

6 Conclusions

In order to enable enterprises to effectively utilize the finished projects to avoid the risk, a two-stage model of project risk assessment is proposed, which supplies a new method for enterprises to deal with project risks effectively. In the future research, more machine learning techniques, such as neural network, genetic algorithm, and so on, can be combined with case-based reasoning methods in project risk assessment. At the same time, the application field is not limited to project risk assessment, but also can be extended to more research fields.

Acknowledgment

This study was supported by Beijing Natural Science Foundation(9132004), Beijing Social Sciences Foundation(14JGC113),China Postdoctoral Science Foundation (2012M510522), Reserve Academic Leader Program of Capital University of Economics and Business, and Improving Scientific Research Level Program of Beijing Municipal Commission of Education.

Reference

Zhang Liwei. Technical Risk Assessment of High-Tech Project[M]. Beijing:Scientific and Technical Documentation Press, 2011.

Li Hua,Cao Xiaolong,Cheng Jiangrong. Research on Risk Assessment of Software Project Based on BP Neural Network[J].Computer Simulation,2011, 28(7): 374-377.

Zhang Xuejun, Wei Guiwu. The Group Evaluation Method of Venture Capital Project Risk Based on Intuitionistic Fuzzy Sets[J].Science and Technology Management Research, 2009(7):448～450.

Hung Juan, Feng Yuqiang, Wang Hongwei.The Integrated Intelligent Systems in Credit Risk Assessment Based on Inductive Reasoning[J]. Application Research of Computers, 1999(9):14～16.

Attar Software Limited. Xpert Ruler Miner: Knowledge from Data[M].Attar Software Limited Inc, 2002.

Zhu Weidong, Xu Ning. Pre-warning The Personnel Risk in Listed Company Based on Case-based Reasoning [J].Computer Applications and Software,2007,24(1): 107-109.