Research on Group Decision Making of Large-scale Engineering Based on Uncertain Decision Tree Classification Algorithms

Wei Chen

University of Shanghai for Science & Technology, Business School, Shanghai

Keywords: Engineering risk; Decision-making; Classification tree; Case-based reasoning; Pruning

Abstract: A decision-making model of drilling engineering risk based on decision tree classification algorithm is proposed and constructed according to the characteristics of drilling expert decision-making process and the shortcomings of traditional CBR technology to realize intelligent decision-making of drilling engineering risk. Decision tree method is an important classification method in data mining. Decision tree is a tree structure that similar to flow chart. Among them, each internal node of tree represents the test of an attribute, its branches represent the test results, and each leaf node of tree represents a category. The decision tree model is used to classify a record, which is to find a path from root to leaf according to the attribute test results in the model. The attribute value of the last leaf node is the classification result of this record, and thus constructed a risk case retrieval model. The field test results show that the model effectively improves the accuracy and recall of case retrieval. A prototype system of drilling engineering risk decision-making is developed based on the above, which provides efficient decision support for drilling experts and technicians.

1 Introduction

Petroleum drilling engineering is a concealed underground engineering with high investment, high risk and high technology level and exists a lot of complex and uncertain factors. There are risks of varying degrees and forms in all stages of drilling, therefore, drilling risk decision-making control has become a common concern in the drilling industry. The accuracy and efficiency of risk decision-making (including risk identification, treatment and prevention) directly affect drilling costs and economic benefits. In recent years, domestic drilling industry has developed some software systems with the development of drilling engineering information and intelligence technology to assist drilling risk decision-making. But most of the existing systems are developed independently according to the different needs and characteristics of each department, the expert system based on rule-based reasoning can only solve the single factor risk decision-making problem(For example, "Drilling Accident Diagnosis and Processing Expert System" developed by Zhongyuan Oilfield Drilling Institute only aims at accident risk), the software and hardware platforms, terminology describing risk, and the expression of risk decision knowledge are all different；The isolated knowledge organization (that is, knowledge is organized for a specific application and is exclusively shared by the application system) constitutes a "knowledge island" which cannot be shared and reused.

2 Decision tree construction and pruning

2．1 Decision tree construction

The input of the decision tree construction algorithm is a set of examples with class labels, the result of the construction is a binary tree or a multi-branch tree. The internal node (non-leaf node) of a binary tree is generally expressed as a logical judgment, such as a logical judgment in the form of (ai = vi), where ai is an attribute and vi is a value of the attribute. The edge of a tree is the result of a branch of logical judgment. The inner node of a multi-fork tree is the attribute, the edge is all values

DOI: 10.25236/iiicec.2019.075

of the attribute, there are several attribute values, and there are several edges, the leaf nodes of trees are all class markers.

The method of constructing a decision tree is to use a top-down recursive construction. Its construction idea is to start with a single node representing the training sample taking the multi-fork tree as an example, they are regarded as leaf nodes if the samples are all in the same class, and the content of nodes is the class marker. Otherwise, an attribute is selected according to a certain strategy, and the set of examples is divided into several subsets according to the attribute and each value, it enables all instances on each subset to have the same attribute value on that attribute, then recursively process each subset. This kind of thinking is actually the principle of "dividing and governing". The binary tree is the same, but the difference lies only in how to choose a good logical judgment.

The key to construct a good decision tree is how to select good logical judgment or attributes. There can be many decision trees that match this set of examples for the same set of examples. The research results show that the smaller the tree, the stronger the ability to predict of the tree in general. The key of construct a decision tree as small as possible is to select the appropriate attributes to generate branches. Attribute selection depends on the Impurity measure for various subsets of examples. The impurity measurement methods include information Gain, Gain Ra-tio, Gini-index, distance measurement, x2 statistics, evidence weight, minimum description length, etc. Different measures have different effects, it is very important to choose appropriate measurement methods for the results especially for multi-valued attributes. ID3, C4.5 and C5.0 algorithms use the concept of information gain to construct decision trees, the CART algorithm x uses Gini-in-dex, the decision of each classification is related to the target classification selected earlier.

2．2 The decision tree pruning

1)Two basic pruning strategies.

①Forward-Pruning is pruning before the tree's growth process is completed. It is decided whether to continue to partition the impure training subset or to stop in the process of tree growth.

For example, the nodes do not continue to split, and the internal nodes become a leaf node when some valid statistics reach a preset threshold. Leaf nodes take the classes with the highest frequency in the subset as their identification, or it may store only the probability distribution functions of these instances. Early pruning can cause trees to stop working before is not fully mature, the tree may be stopped extension should not stop, or called the horizon effect. Moreover, it is difficult to select an appropriate threshold. Higher thresholds may lead to over-simplification of trees, while lower thresholds may lead to too little simplification of trees. Even so, the large-scale practical application of pre-pruning is worth studying because it is quite efficient. Hoizon effect is expected to be solved in future algorithms.

②Post-Pruning is the pruning after the growth process of the decision tree is completed. It is a two-stage method of Fitting-and-simplifying. Firstly, a decision tree is generated which fits the training data perfectly. Then, the tree is pruned from the leaves to the roots from the bottom to the top. A test data set is used for pruning, If there is an accuracy on the test set after a leaf has been cut or other measures are not reduced (not getting worse), the leaf is cut off, otherwise it stops.

Principles to be followed in tree pruning optimization.

Minimum Description Length Principle(MDL). The idea is the simplest explanation is expected, practice are encoded to binary decision tree. Coding needed least binary tree is the "best pruning trees".

Minimum Expected Error Rate Principle. Its idea is to select the subtree with the lowest expected error rate to prune. That is, the expected error rate of the pruning/non-pricing may be calculated for the internal nodes in the tree, and then compared to select.

Occam razor principle. If it is not necessary, do not add entities. That is to say, “The simplest one should be chosen in the theory compatible with observation ". The smaller the decision tree, the easier it will be understood, and the lower the cost of storage and transmission.

3 Overall Structure of Drilling Risk Decision Model

The overall structure of drilling engineering risk decision-making model based on ontology and CBR is divided into three layers: Application layer, decision reasoning layer and knowledge layer.

(1) Application layer: Decision-makers can be provided with decision-making assistance and knowledge sharing application services through a unified human-computer interface, includes auxiliary decision-making of drilling risk, case inquiry, knowledge base maintenance and other modules. The "assistant decision-making" module is used for risk decision-making analysis. First, the decision maker inputs the characteristic information of the current risk and submits it to the system. Then the system automatically starts the case-based reasoning machine for case-based reasoning. Finally, the types of current risks and their solutions are identified and output.

(2) Decision reasoning layer: This layer is the core of risk intelligent decision-making. The main task is to use the knowledge in the knowledge base to realize ontology-based case reasoning, identify risks and form a risk control scheme.

The workflow of case-based reasoning [1] is:①Problem description: Ontology-based case representation is used to describe the current decision-making problem as a new problem case; ② Case retrieval: It retrieves the most similar historical cases from the case base. If the retrieved case matches the new problem perfectly, step ③ will be executed, otherwise step ④ will be executed; ③Case reuse: Reuse the decision-making scheme of similar cases as the solution of new problems；④Case correction: The decision-making schemes of similar cases are amended to obtain solutions suitable for new problems according to the domain knowledge in rule base and the characteristics of new problems.⑤ Case learning: New problems and their final solutions are formed into new cases and evaluated. If they have reserved value, they are added to the case base to achieve self-learning of knowledge.

(3) Knowledge layer: It consists of case base and rule base. Case inventory puts various risk decision cases for existing drilling projects；Rule stocks are used in the relevant rules for case correction and risk classification, For example, a classification rule is described as a production rule:

IF the drill string static before sticking AND (The stuck point is on the drill string OR Normal pump pressure before card) THEN The type of risk is a sticky card drill.

4 Case representation and organization of drilling risk decision based on ontology

If case C has n characteristic attributes, then: $C(c_1,c_2,\cdots,c_n)=D(d_1,d_2,\cdots,d_m)+S(s_1,s_2,\cdots,s_k), m+k=n$, among them, D denotes the numerical feature attribute part of a case and S denotes the conceptual feature attribute part of a case. The calculation method of case similarity is as follows:Different similarity calculation models are used for two types of feature attributes, it first calculates the local similarity between the single attribute of the problem case N and the historical case Hi, and then calculates the overall similarity between the two cases. The concrete calculation model is as follows:

(1)Similarity calculation of numerical feature attributes

The values of numerical characteristic attributes are continuous values(For example, the density of drilling fluid is 1.15g/cm3), the similarity between its individual attributes is defined as:

$$Sim(d_{nj}, d_{ij}) = 1 - \left|a_{nj} - a_{ij}\right| / (\beta - \alpha) \quad (1)$$

In the formula, dnj and dij represent the jth characteristic attribute of problem case N and historical case Hi respectively, anj and aijare the corresponding attribute values, and anj, aij∈[α,β], [α,β] are the range of characteristic attributes.

(2)Similarity calculation of conceptual feature attributes

Distance-based semantic similarity computing model is usually used to calculate the similarity of conceptual feature attributes (i.e. concepts) [13] , the basic idea of this model is the concept of semantic distance between the two concepts in concept hierarchy in the network to quantify geometric distance [14, 15]. Existing computational models usually regard the distance of all edges in the network as equally important when calculating the semantic similarity between concepts[13], but the weights of each side may be different in the conceptual hierarchical network composed of case ontology. That is, the semantic similarity between father and child nodes located on different directed edges is different. In this paper, an improved distance-based conceptual semantic similarity computing model is proposed by introducing edge weights based on the existing computing models. The calculation method is as follows:

1)Calculating weights of directed edges

The weights of directed edges are usually related to the following factors in Ontology Conceptual hierarchical networks [14]:The parent node and the child node have the type, depth, strength of the edge, the properties of the concept nodes at both ends, and the density of the parent and child nodes in the hierarchical network. This paper mainly considers the type and depth of directed edges according to the characteristics of semantic information in risk decision-making cases, conceptual node c

The weight of the directed edge with its parent node P is defined as:

$$w(c, p) \propto \text{edge type} \times \text{edge depth} \tag{2}$$

Among them, the types of directed edges are determined by the relationship between concepts. There are mainly four kinds of relationship between concepts in risk case ontology:Synonymous relations, inheritance, composition and associated relations (Through n associated attributes). The relationship between weights of directed edges and edge types is defined as:

$$w(c, p) \propto \begin{cases} 1, & type(c, p) = \text{synonymy} \\ 1/2, & type(c, p) = \text{Inheritance relationship} \\ 2/3, & type(c, p) = \text{Composition relationship} \\ 1/(n+1), & type(c, p) = \text{Correlation} \end{cases} \tag{3}$$

In Ontology Conceptual Hierarchy network, the lower the level of the concept, the more specific it's meaning because each layer is a refinement of the concept of the previous layer. Therefore, the weight of the directed edge is related to its depth, which is defined as:

$$w(c, p) \propto \left(\frac{1}{2^{depth(p)}} + \frac{1}{2^{depth(p)-1}} + \cdots + \frac{1}{2} \right) = \sum_{n=1}^{depth(p)} \frac{1}{2^n} \tag{4}$$

Where depth (p) represents the depth of node p, formula (3) and formula (4) are substituted into formula (2), and the directed edge weight is obtained as follows:

$$w(c, p) = k \times type(c, p) \times \sum_{n=1}^{depth(p)} \frac{1}{2^n}, \ k \text{is a regulatory factor.} \tag{5}$$

2) Calculating the length of directed edges

The directed edge length between concept node c and its parent node p is defined as:

$$Le(c, p) = \frac{\eta}{w(c, p)} - \eta, \ \eta \text{is a regulatory factor.} \tag{6}$$

5 Conclusions

The combination of case-based reasoning and ontology technology is the trend of decision tree algorithm development [4]. This paper presents an engineering risk decision model based on decision tree algorithm, the model standardizes and unifies the concepts and terms involved in the field of drilling engineering risk decision-making by introducing ontology, it realizes the unified representation and retrieval of cases at the semantic level, it improves the accuracy and recall of case retrieval, it provides good scalability and sharing for case knowledge at the same time. The construction of Ontology-based case base makes the decision-making knowledge in the system include not only specific risk case knowledge, but also general decision-making domain knowledge. Thus, it overcomes the limitation of knowledge in traditional decision tree algorithm system, and the effectiveness of system decision support is improved.

References

[1] Liu X, Wu J, Gu F, et al. Discriminative pattern mining and its applications in bioinformatics[J]. Briefings in Bioinformatics, 2015, 16(5):884.

[2] Quesada F J, Palomares I, Martínez L. Using Computing with Words for Managing Non-cooperative Behaviors in Large Scale Group Decision Making[M]// Granular Computing and Decision-Making. Springer International Publishing, 2015:97-121.

[3] Xuan-Hua X U, Hui-Di W U, Business S O, et al. Approach for multi-attribute large group decision-making with linguistic preference information based on improved cloud model[J]. Journal of Industrial Engineering & Engineering Management, 2018.

[4] Zeng Z, Nasri E, Chini A, et al. A multiple objective decision making model for energy generation portfolio under fuzzy uncertainty: Case study of large scale investor-owned utilities in Florida[J]. Renewable Energy, 2015, 75(75):224-242.

[5] Liu B, Huo T, Liao P, et al. A Group Decision-Making Aggregation Model for Contractor Selection in Large Scale Construction Projects Based on Two-Stage Partial Least Squares (PLS) Path Modeling[J]. Group Decision & Negotiation, 2015, 24(5):855-883.