Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 154 (2019) 232–237

Procedia Computer Science

www.elsevier.com/locate/procedia

8th International Congress of Information and Communication Technology, ICICT 2019

Study on the Test Data Fault Mining Technology Based on Decision Tree

Hui Lian Han[**a,b], Hong Ying Ma[c], Ye Yang[a,b]

[a]National Key Laboratory of Science and Technology on Aerospace Intelligence Control, Beijing 100854, China
[b]Beijing Aerospace Automatic Control Institute,hanhuilian08@163.com, Beijing 100854, China
[c]Hebei Jiaotong Vocational and technical College, mahongy@126.com, Shijiazhuang 050091, China

Abstract

Aiming at the shortage of greatness of test data and the traditional decision-making arithmetic of decision tree, the paper put forward a way to diagnose faults based on vehicle test data mining. Attribute list was built by the way of training data set in tests, fault decision tree fabricated based on better-deciphered arithmetic –C4.5. To overcome the time-consuming shortage of C4.5 in the discretization process of continuous valves, the paper studied Fayyad theorem, prospered improved arithmetic by calculating via dividing line, experiments proved the fault detection and diagnose strategy effectively and exactitude

Keywords: Big Data; Decision list; Decision tree; Information entropy; Test optimize

1. Introduction

Tests go with the process of vehicle development, production and flight experimentation process at all times. Vehicles test data is largeness, multi-dimensionality, complex relation, relevant and specialized. It is useful to analyze data availability and compute capacity for the develop management personnel to judge the performance of aircraft. And the test data provided many kinds of hereunder in effect to run and manage vehicles. As a new knowledge acquisition technology, data mining has a wide application in fault detection and diagnosis, production

* * Corresponding author. Tel.:+ 18513157797
Emial: hanhuilian08@163.com

10.1016/j.procs.2019.06.035

optimization, repository enrichment and decision-making[1-2]. It is important for test data assistant to improve flight fault detection diagnosis and decision-making, and it is momentous to improve the quality of fault detection and diagnosis and ensure the safety in flight.

As an important investigate content, fault detection and diagnosis arithmetic belong to supervisory learning method, meaning that data swatch, training sorter, marked with sort label. Accordingly, those sort labels of unknown test swatches can be obtained from the trained sorters. Decision tree sort arithmetic[3-4] is right sort way for decision-makers to find out valuable information from a great deal of data. For its simple pick-up rules, easily comprehended and low computation complexity, decision tree sort arithmetic is widely used in data mining in recent years. On the other hands, decision tree sort arithmetic can process discrete data and continuous data contemporary, the arithmetic can be implemented in distributed software on parallel hardware platform.

Various data give birth to constantly while vehicle in run time, ordinarily or manpower diagnose method cannot carry through, it is necessary to implore a simple and practicality way to construe test data. The paper brought forward an intelligent fault detection and diagnosis method based on decision tree, by the way of training data set in the test attribute list was built, fault decision tree fabricated based on better-deciphered arithmetic – C4.5[5-7]. By deeply mining the vehicle test data, fault diagnoses information were attained exactly, the method not only improved fault diagnoses efficiency but also acquired high accuracy and low false positives.

2. Introduction of Decision Tree Arithmetic

Decision tree diagnosis arithmetic is an important technology of sort arithmetic; the procreant configuration is similar to the tree model of program flow chart. The method is a progress of establishing different rule nodes via seeking characteristic attributes with the greatest information in data set, reinstitute different tree deprecating by different eigenvalues, the wholly tree was established by circulation. Currently, there are 2 kinds of universal fault detection and diagnosis arithmetic: ID3[8-9] and C4.5. Compared with the other methods, decision tree has widely applications because of its computation rapid and accurate.

The essential of decision tree's from top downwards recursion is that the sample data arrayed from the root node, and the samples are sorted via the recursion of Non-leafage sub node, and were denoted the sort results by the bottom-leafage sub node. Non-leafage sub node on the decision tree represents the test of an attribute value in sample data, hereinto 'YES' and 'NO' represent right examples and reversed examples in example sets, this kind of sort mode has the advantage of two-classification problems.

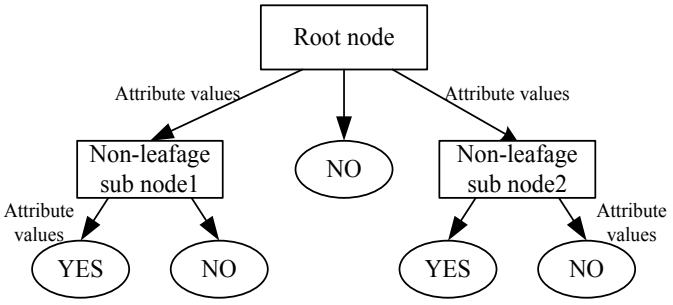

Fig.1 Sample of framework of decision tree

The knowledge denote method based on decision tree classify rules is an effective way to building repository. Thereunto ID3 algorithmic is the basis of all the algorithmic, it select the types of examples on the basis of the values of attribute sets, use information entropy as the objective evaluate function, adopting the manner from top downwards recursion to search a part of the whole space. The decision tree constructed by ID3 which has low deepness, rapid classifies, and suit to non-increment learning task. For ID can not solve the increment training example, it must give off the intrinsic decision tree as gaining an example, the algorithmic expense increased and the efficiency play down. ID3 can only process discrete data, as C4.5 not only can process discrete data but also can

process data with continuous attributes.

3. C4.5 Algorithmic

3.1 summarize of C4.5 Algorithmic

Behaving ID3 core of integrated decision tree building system, algorithmic inheriting all the advantages of ID3 algorithmic and improve on it.C4.5 algorithmic impose information gain ratio as the criterion of splitting attributes At all level-nodes, enhancing the accuracy of classification and simplifying decision tree at one time.

C4.5 algorithmic was improved on the basis of ID3 algorithmic, it not only has classify rapid speed, high accuracy, easily comprehended and is capable of dealing with the lack of continuous attributes or attribute data. Replacing information gain with information gain ratio, the algorithmic avoid high-branch attributes not be chosen, for too many branches cause decision tree depend on a certain attribute excessively, the method overcome the objection of ID3 using information gain select branch attributes leaning to excessive data.

3.2 C4.5 Algorithmic basic process

Suppose sample data-set X classifies into n types. Suppose there are Ci sample data in the *i*th class, total samples of X data-set is |X|, so the probability that a sample data attribute to the *i*th class is $P(C_i) \approx \frac{C_i}{|X|}$. If select attribute *a* (suppose *a* has m different values) to test, its uncertainty (condition entropy) is:

$$H(X \mid a) = -\sum_{j=1}^{m} p(a = a_j) \sum_{i=1}^{n} p(C_i \mid a = a_j) \log_2 p(C_i \mid a = a_j)$$

Information gain ratio is defined as the ratio of average mutual information and the cost to obtain a information，namely：

$$E(X \mid a) = \frac{I(X \mid a)}{H(X \mid a)}$$

A decision tree T which is created by giving training data-set using information gain ratio as the criterion to select attribute, select E(X,C)the maximum attribute as the test attribute. The steps of build-up decision tree are as follows:

(1) Preprocess data source, discredit the continuous data, form training dataset of decision tree (if data-set has not continuous data ignore it);

(2) Calculate every data-set information gain and information gain ratio;

(3) Every attribute to the root nodes corresponding to a sunset, recursive of sample carry out former step(2) course until get the division every subset and supervise data get same classified data, then generate decision tree;

(4) Classify new data-sets according to the decision tree’s rules of pick-up classify.

The flow chart of decision tree based on C4.5 is shown as figure 2.

3.3 The Shortages of C4.5 Algorithmic

Despite C4.5 has a good many advantage, the Algorithmic presences several shortages, incorporate on two aspects:

(1) Low computation efficiency

C4.5 has low computation efficiency， specially facing training samples continuous training data[11]. For example: Suppose training sample set S has N samples, sample A is continuous, it is necessary for the continuous data to be discredited before building decision tree. Suppose A have M continuous values, while discrediting, every attribute data is the dividing line, every dividing line should traverse a time, accordingly figure out the information gain of every diving line, then select the maximal threshold, the traversal takes a lot of time, It is evidence especially

sample sets have many samples or M is very big.

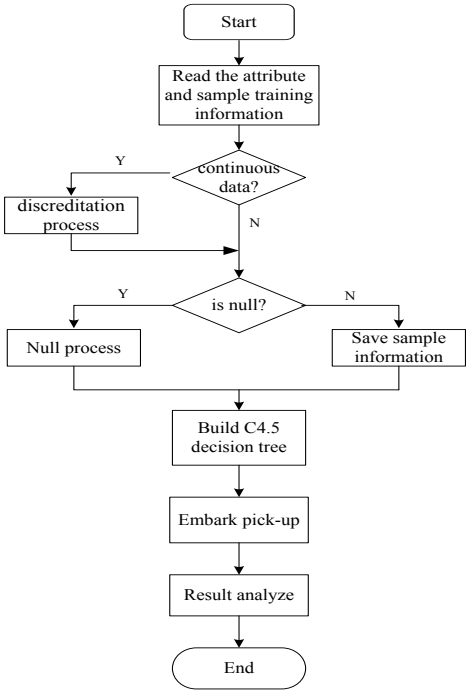

Fig2 The process of C4.5

(2) High complex model of decision tree

C4.5 has great advantage on training time and accuracy of predict, but the model has high complexity, and the number of decision tree model nodes is excessive. For C4.5 select attribute to filiations as discredited, the subtree has the same number with the discredited data.

The first, while building decision tree, the arithmetic must process scanning and taxis one by one repeatedly, causing low classify efficiency.

The second, the formulas of the algorithmic deal with a great deal of logarithm operation, computer transfers functions in high frequency, increase time spending of the algorithm.

The third, C4.5 algorithm does not consider the relativity of condition attributes while choosing split attribute, it only computes the expectation information between every attribute and class attribute, it maybe effect the validity of attribute select.

The forth, despite C4.5 algorithm is improved based on ID3 algorithm, it can deal with the data set of not integrity and continuous data while it cannot process a good many other styles data set, the width of data set modes of C4.5 algorithm need to be enhanced.

4. The Improved C4.5 Algorithmic Based on Fayyad Borderline Principle

4.1 Fayyad borderline principle [10]

C4.5 uses information gain ratio as the criterion scaling classify attribute of training sample. It is necessary for the continuous samples data to be discredited before building decision tree. It is necessary for the continuous samples data to be discredited before building decision tree. While getting a good many of continuous values, using the algorithm need to carry through a mass of computation of information gain ratio.

Fayyad brought forward borderline principle in allusion to the problem: in despite of how many different attribute

valves the training sample data set have, also despite how the sorts distributed, the best borderline(the best dividing threshold) of continuous attribute is always on the borderline of the training sample attribute data.

As compute the best borderline of continuous attribute, only need to compare the least information entropy value of attribute sequence, the most modify information gain ratio can be calculated out, reduce borderline be an alternate, consequently improve the calculate efficiency greatly. Using borderline principle, can reduce the discrete time prodigiously, accordingly improve the efficiency of the entire time.

5. Fault Detection and Diagnosis Experiment Scheme Based on Improved C4.5

5.1 fault detection and diagnosis experiment scheme based on improved C4.5

The fault detection and diagnosis experiment scheme based on improved C4.5 are shown as figure 3.

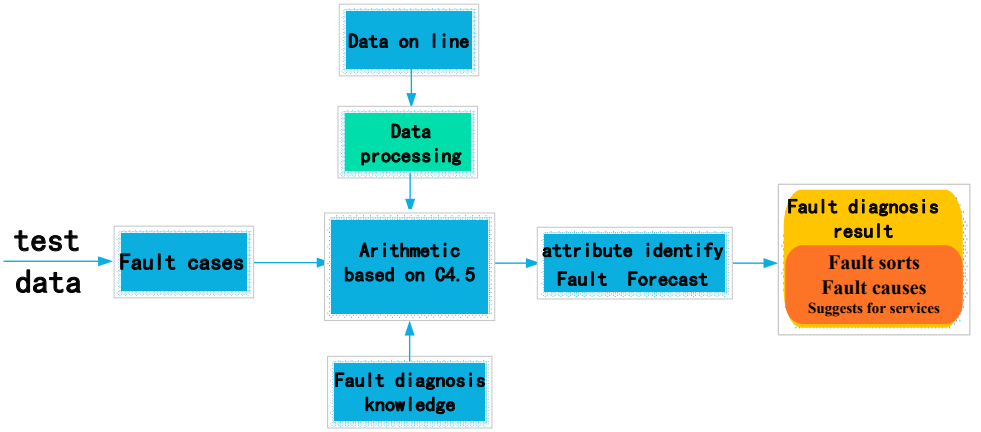

Fig.3 fault detection and diagnosis experiment scheme based on C4.5

5.2 fault detection and diagnosis result based on C4.5

For temperature data, the training quantum is 5000, there were 3 types of fault, the gross of fault sample is 603. The quantum of test data is 5000, fault types is also 3, the usual error took place at time index 2000~2200, the excursion error took place at time index 3000~3200, the impulsion error took place at time index 4000~4200.

Every group data was infused different types and different amplitude faults, the usual error was marked in fault code 1, the excursion error was marked in fault code 2, the impulsion error was marked in fault code 3, and the errors were applied as the input data of the fault decision tree model C4.5. A part of decision tree face to temperature data via C4.5 was shown as figure 4.

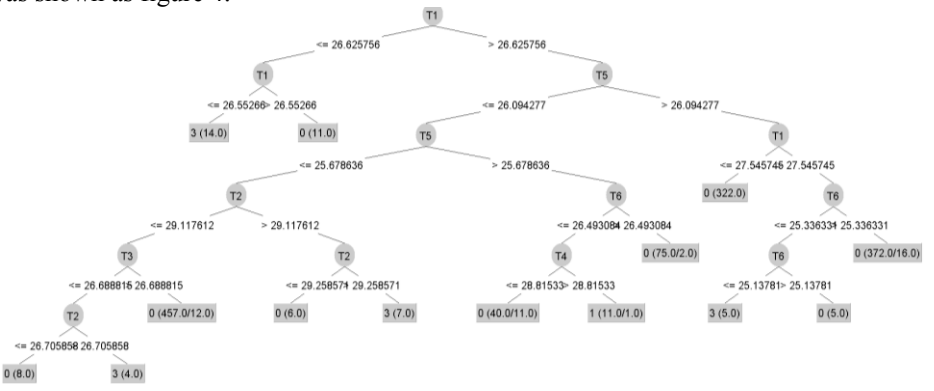

Fig.4 A part of decision tree face to temperature data

Based on C4.5, the paper implored the decision tree facing temperature data, etc. By emulationally infused usual errors, excursion and impulsion fault at originality data different situations, the experiments proved the can diagnosis fault in effect.

6. Conclusion

By analyzing the decision tree sort arithmetic of C4.5, and facing the shortage of C4.5, the paper improved on the computation of optimization dividing line of continuous attribute. The method can pick-up decision list, predigest the search steps, apply the decision tree rules on the fault diagnoses, buildup the vehicle's fault-tolerant, also the method can improve the accuracy of fault tree diagnoses. The experiments proved the method effect in fault detection and diagnosis, and it can be used for reference in vehicle fault detection and diagnosis.

References

1. 1. Wang M X. Survey of data mining. Software Guide, 2013, 12(10):13}-137. (In Chinese)
2. 2 Jeong S, Shimoyama K. Review of data mining for multidisciplinary design optimization[C]//Proceedings of the Institution of Mechanical Engineers. Part G. Journal of Aerospace Engineering, 2011.225(5): 469-479.
3. 3 Quinlan J R.C4.5:Programs for Machine Learning. Morgan Kaufman，1993.
4. 4. TSAI Yingchieh, CHENG Chinghsue. Entropy －based fuzzy rough classification approach for extracting classification rules, Expert Systems with Applications, 2006, 31(2): 436－443.
5. 5. Hashim H. Talab A A. Satty A, et al. Data Mining Methodologies to Study Student's Academic Performance Using the C4.5 Algorithm. 2015,5(2): 59-68.
6. 6. Mantas C J, Abellán J, Castellano J G. Analysis of Credal-C4.5 for classification in noisy domains. Expert Systems with Applications, 2016, 61: 314-326..
7. 7. Elomaa B T. In Defense of C4.5: Notes on Learning One-Level Decision Trees. Proc. of the 11th Int. Conf. on Machine Learning. 2015.
8. 8.Du J, Cui H,Li W,et al. fault detection and diagnosis of vacuum circuit breakers based on ID3 method[C]//Electric Power Equipment-Switching Technology(ICEPE-ST), 2011 1st International Conference on.IEEE, 2011:283-286..
9. 9. Meichun H U, Tian D, School B. Improved ID3 Algorithm Based on Parameter Correction and Simplified Standard[J].Computer & Digital Engineering,2015.
10. 10. Zitzler E, Thiele L. Multiobjective evolutionary algorithms: A comparative case study and the strength Pareto approach. Evolutionary Computation, 1999, 3(4):257-271.
11. 11. Jing Tian, Zihan Song, Fei Gao, Feng networks using the C4.5 algorithm. Grid pattern recognition in road Cartography and Geographic Information Science, 2016,433.