

Application of ID3 Algorithm in the Analysis of Satellite Telemetry Data

Fan Yu^{1,a}, Qin Yan^{2,b}

¹School of Computer Science and Engineering, Xi'an Technological University, Xi'an 710021, China

²School of Computer Science and Engineering, Xi'an Technological University, Xi'an 710021, China

^ayffshun@163.com, ^byanqinxingfu.123@163.com

Keywords: Telemetry data, Correlation analysis, ID3 algorithm, Decision tree

Abstract. In order to deeply exploit the knowledge hidden in the telemetry data and find out the correlation between the data, ID3 algorithm was used in the article to analyze the telemetry data of solar panel substrate. The information gain value of parameter was calculated and the maximum gain value is the split property, the decision tree was constructed based on the split property. The knowledge and the correlation between the property data were discovered; the correlation affected the state of solar panel substrate. These results are necessary for analyzing the satellite on-orbit performance and diagnosing fault. The analysis result shows that ID3 algorithm is suitable for the analysis of satellite telemetry data, and the proposed algorithm improves the quality of telemetry data processing.

Introduction

Telemetry data processing is the core technology of ground-based monitoring software; it is an important part of the remote analysis system on the ground[1]. During the process of satellites on-orbit and in-orbit monitoring, large amounts of information will be produced, including information of fault, status mutation information, monitoring and computing results, record of the satellite's working status, daily management information, and environmental parameters of the operational space. These telemetry data is more realistic than over ground simulation information and testing data, directly reflecting the state of the satellite payloads and satellite operation. The analysis and application of telemetry data have significant meaning for analyzing the in-orbit performance, forecasting and early warning the failure and grasping the state of satellite, even feed backing to the design department of satellite.

Currently, the available literature about the correlation analysis of data primarily in two ways: The first method is to rely on SPSS, which is statistical analysis software. In the references [2], the relevancy between engine speed and other factors is found by calculating the Pearson correlation coefficient of the combat aircraft engine speed, oil pressure, exhaust temperature, pressure and other parameters. In the references [3], with the help of SPSS system, the factors that affect the trajectory consistency are excavated by analyzing the date, which is selected from the database of a shell A and shell B 2 kilometers away as a target distance, within time interval T. The result will provide the basis for improving shells. But the analysis result of SPSS exports in the form of data tables, only the size of the correlation coefficient related to the relevancy that fail to show knowledge between factors, and the output is can not directly opened by the word processing software, only through copying and pasting to interact. Another common method is using the correlation function to calculate the correlation coefficient. In the references [4], the author researches on the measured data from the measuring device on a space tracking ship, the measurement element of selected are the slope distance R, the pitch angle E, Azimuth A, rate of change of slant range V. The correlation during these parameters is expressed by calculating the partial correlation function value; the result

will provide direction for selecting the step when counting the statistical observation random error. However, this method is large computation and high time complexity, so it is not suitable for the analysis of large amounts of data.

ID3 Algorithm

The core idea of ID3 algorithm is information entropy [5], which selects the property of the smallest information entropy as the category property, recursively expand branch of the tree to construct the decision tree.

The basic principles of ID3 algorithm as follows[6]: assuming $E = F_1 \times F_2 \times \dots \times F_n$ is a finite n -dimensional vector space, $F_i (i = 1, 2, \dots, n)$ is a finite set of discrete symbols, the elements $e = (V_1, V_2, \dots, V_n)$ in E is called the sample, in which $V_i \in F_i, i = 1, 2, \dots, n$.

Supposing the number of collection of positive examples and negative examples in the collection of vector space E is the size of p and n , ID3 algorithm is based on the following two assumptions:

- 1) A correct decision tree in the vector space E , the probability for any classification of the sample with the E line in the probability of positive and negative cases.
- 2) The amount of information required for a decision tree can make the right type of judgment to determine:

$$I(p, n) = -\frac{p}{p+n} \log_2 \frac{p}{p+n} - \frac{n}{p+n} \log_2 \frac{n}{p+n} \quad (1)$$

Supposing the property A as the root, it has m property values, as follows: $\{U_1, U_2, \dots, U_m\}$, the space E is divided to m subset, as follows: $\{E_1, E_2, \dots, E_m\}$, assuming the subset E_i has p_i positive examples and n_i negative examples, the entropy of E_i is $I(p_i, n_i)$, the entropy of property A is

$$E(A) = -\sum_{i=1}^m \frac{p_i + n_i}{p+n} I(p_i, n_i) \quad (2)$$

So the information gain of property A is

$$Gain(A) = I(p, n) - E(A) \quad (3)$$

ID3 algorithm selects the property of A^* having the largest information gain as the extended attribute, for the different values of A^* corresponding to E , the recursive call m sub-sets of the above process to produce the child node of A^* .

ID3 algorithm is a typical decision tree learning system [7]. In the course of constructing decision tree, selecting properties at all levels is the core idea, using information gain as the attribute selection criteria, using this top-down, divide and conquer strategy that can not be returned [8], ensuring the tree is the most simple, the data need to be tested every time is the smallest amount number, the number of test is lest, so the classification speed is faster.

Correlation analysis of telemetry data based on ID3 algorithm

The type telemetry data of satellite is different, the digital features, distribution and variation of different types are different, and this article selects the current, voltage and temperature parameters of satellite solar panels to analyze.

Selecting the original telemetry data .The original telemetry data is the solar panel telemetry data, which coming from the comprehensive database of the satellite in the history; it is the source data for building the classification tree. The principles when selecting data as follows: First, the data should be targeted; choosing the parameters associated with the state and performance closely, to ensure the decision tree is more complete. According to a large number of test results and monitoring information satellites in orbit, where we select the current, voltage, temperature of solar panels; Second, the amount of data-selected should be moderate, each telemetry data must be chose, so as to ensure the rules generated from the decision tree is generalization.

Remote sensing data preprocessing. Telemetry data processing is an important step in remote sensing data analysis process, aimed at improving the quality of the object, ultimately improves the knowledge quality that gained from the data analysis purposes [9]. The telemetry data is continuous data, however, such data can not be handled by ID3 algorithm, we need adopt some method to discrete them, making their data discrete. In the article, the original data is separated by calculating the rate of change [10]. For the current and voltage class of data, calculating rate of change of the difference between adjacent sampling points, while the temperature class data, calculating the rate of change of the average that time interval between adjacent. The rate of change of sampling points is marked into several intervals, the solar panel telemetry data received training set is got by counted at last, as shown in Tab 1:

Tab 1 The training set of current、 voltage and temperature parameter

No.	rate of change of Current sampling points	rate of change of voltage sampling points	rate of change of average temperature	State of solar panels whether changes
1	>5%	>5.5%	>3.5%	Yes
2	>5%	3%—5.5%	>3.5%	Yes
3	<2.5%	<3%	<2%	No
4	2.5%—5%	>5.5%	2%—3.5%	Yes
5	2.5%—5%	3%—5.5%	<2%	No
6	2.5%—5%	<3%	<2%	No
7	<2.5%	3%—5.5%	>3.5%	Yes
8	>5%	3%—5.5%	2%—3.5%	Yes
9	<2.5%	<3%	2%—3.5%	No
10	2.5%—5%	>5.5%	>3.5%	Yes
11	<2.5%	3%—5.5%	<2%	No
12	2.5%—5%	3%—5.5%	2%—3.5%	Yes
13	2.5%—5%	3%—5.5%	>3.5%	Yes
14	>5%	<3%	2%—3.5%	No
15	<2.5%	>5.5%	2%—3.5%	Yes

Classification decision tree construction. In the training set, classification of property “State of solar panels whether changes” has two different values-“Yes” and” No”. Supposing S_1 corresponds

to "Yes", S_2 corresponds to "No". After adding up, $S_1=9$, $S_2=6$. Based on ID3 algorithm, follow these steps for constructing decision tree:

- 1) According to the formula (1), the expectations of information required is calculated for the given sample classification.

$$I(S_1, S_2) = I(9, 6) = -\frac{9}{15} \log_2 \frac{9}{15} - \frac{6}{15} \log_2 \frac{6}{15} = 0.971.$$

- 2) According to the formula (1) and formula (2), the information entropy is calculated.

Calculating information entropy of rate of change of Current sampling points marks it as A.

For $A_1: > 5\%$: There are four attribute values, $S_{11}=3, S_{21}=1, I(S_{11}, S_{21}) = I(3, 1) = 0.813$;

For $A_2: 2.5\% \sim 5\%$: There are six attribute values, $S_{12}=4, S_{22}=2, I(S_{12}, S_{22}) = I(4, 2) = 0.927$;

For $A_3: < 2.5\%$: There are five attribute values, $S_{13}=2, S_{23}=3, I(S_{13}, S_{23}) = I(2, 3) = 0.971$;

$$E(A) = \frac{4}{15} I(3, 1) + \frac{6}{15} I(4, 2) + \frac{5}{15} I(2, 3) = 0.911.$$

Similarly, $E(B) = 0.403$, $E(C) = 0.371$.

- 3) According to the formula (3), calculating the information gain of each property.

$$Gain(A) = I(S_1, S_2) - E(A) = 0.06;$$

$$Gain(B) = I(S_1, S_2) - E(B) = 0.568;$$

$$Gain(C) = I(S_1, S_2) - E(C) = 0.600.$$

According to the above results, the information gain of rate of change of average temperature attribute is the largest, hosing it as the division property, which is the root of the decision tree. The training set samples is divided into three branches by temperature, the state of solar panels all change in the branch samples that the rate of change greater than 3.5%; the state of solar panels don't change in the branch samples that the rate of change less than 2%, so the two branches stop dividing. But the branch samples that the rate of change between 2% and 3.5% also includes changes and does not change, and there are current and voltage two attributes in the attribute set, so it is necessary to go on calculating information gain of the remaining attributes. At last, the decision tree is established based on the ID3 algorithm, shown in the Fig.1:

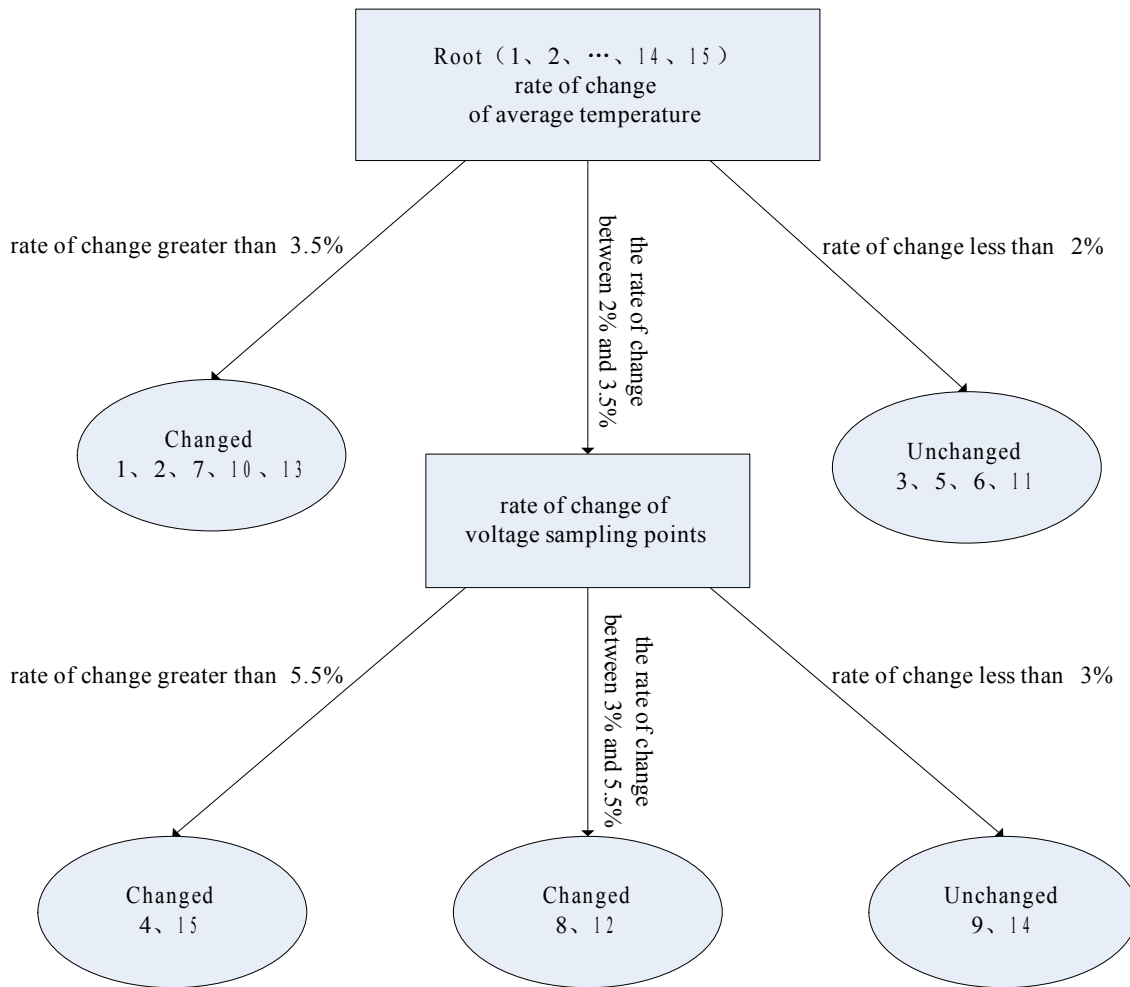


Fig.1 Correlation decision tree of telemetry data

Explanation of the decision tree. In order to more clearly understand the knowledge expressed by the decision tree, the leaf nodes from the root to the only path into the form of rules, as follows:

- (1) If the rate of change average temperature is greater than the 2%, then the state of solar panels change;
- (2) If the rate of change average temperature is less than the 3.5%, then the state of solar panels don't change;
- (3) If the rate of change average temperature between 2% and 3.5% and the rate of change of voltage sampling points greater than the 5.5%, then the state of solar panels change;
- (4) If the rate of change average temperature between 2% and 3.5% and the rate of change of voltage sampling points between 3% and 5.5%, then the state of solar panels change;
- (5) If the rate of change average temperature between 2% and 3.5% and the rate of change of voltage sampling points less than 3%, then the state of solar panels do not change.

Conclusion

With the help of ID3 algorithm, the telemetry data of satellite solar panels is analyzed. The rules and knowledge hidden in the telemetry data is found, the factors that affect the state of solar panels and the potential relationship between factors are excavated. Current, voltage and temperature parameters of the three properties are interrelated; they restrict the change of state solar panels

together. The analysis results also provide decision basis for failure analysis of the solar panels, the targeted analysis of failures can reduce the blindness of analysis, thus improving the working efficiency; At the same time, the results has practical significance for analyzing the performance and status of solar panels during the satellite in-orbit operation process. Therefore, the analysis method for telemetry data discussed in the article is suit to dealing with the satellite telemetry data; it has great potential and provides a new way for the future analysis of telemetry data.

References

- [1] Zhang Chenguang, Lü Ming, Wang Gang. Research on Launch Vehicle Telemetry Data Processing [J].Missiles and Space Vehicles, 2002(2):9-11.
- [2] Lu Bingyan, Chen Youlong ,Li Yingying.The Correlation Analysis of the Pre-Processing Data based on SPSS[J].Science & Technology Information, 2009,15:487-488.
- [3] Zhang Lingke, Wang Zhongyuan, Guo Wenjuan. Application of SPSS in Trajectory Consistency Test and Data Analysis [J]. Journal of Nanjing University of Science and Technology, 2006,30(2):209-212.
- [4] Li Xiaoyong, Zhang Zhonghua. Analysis of Correlation Relationship in Dynamic Data from TT&C Ship-borne Trajectory Measurement Equipment[J]. 2008,30(5):47-52.
- [5] Qu Wei, Zhu Shibing. Elements of Information Theory and Applications[M].Bei Jing: Tsinghua University Press,2005.
- [6] Zhai Junhai,Zang Sufang,Wang Xizhao. Verification on Optimal Algorithm[J].Journal of Hebei University: Natural of Science Edition, 2006,26(5):547-548.
- [7] Hu Jianghong. Research of Classification Algorithms Based on Decision Tree[D].Wuhan:Wuhan University of Technology, 2006.
- [8] Zuo Jifeng, Qiao Junjian. The Proof of Rationality of ID3 Algorithm and Experimental Analysis [J].Journal of Baoding University, 2008,21(4):24-27.
- [9] Zhang Ling. Research on ID3 Algorithm and Application in Data Mining System of Grade Policy-Making [D]. Xiamen:Xiamen University,2007.
- [10]Liu Guanghui. Research on Processing method of Missile Telemetry Data[D].Changsha:National University of Defense Technology,2005.