Compact Prediction Tree (CPT) Application Research on the Causes of Faults in the Use of Multimedia Classrooms in Colleges and Universities

Haicheng Bai 1, Shuai Yang 2

1 Information Management Department, Yunnan Normal University, Kunming, Yunnan, China
2 Information Management Department, Chuxiong Normal University, Chuxiong, Yunnan, China

*e-mail: 281967246@qq.com

Abstract. In addition to maintaining equipment, the management of multimedia classrooms in colleges and universities is mainly to provide timely assistance to teachers. When a failure occurs, if the cause can be automatically predicted and online real-time assistance is provided in time, this will greatly enhance the teacher teaching experience. Series prediction is one of the hotspots of deep learning research, for it has been widely used in many fields from recommendation systems to speech recognition to natural language processing. The Compact Prediction Tree (CPT) is faster and more accurate than other sequence prediction models. Based on this, this paper studies how to generate prediction trees for multimedia fault text events in colleges and universities, and uses CPT to predict the causes of multimedia classroom failures to improve management efficiency and user experience.

1. Introduction

Many colleges and universities have numerous multimedia classrooms. In addition to the maintenance of hardware and software, most of the daily management of these classrooms is to solve the problems encountered by teachers in time. Taking one university as an example, the general workflow of a teacher after encountering problems is: (1) submit a fault description through the network platform; (2) the administrator provides online or on-site help according to the fault description; (3) the administrator records the cause of the failure and forms a fault event. Since that, a large number of multimedia failure events from this university are collected for analysis. It is found that there are few failures caused by hardware damage. On the contrary, the main problems are caused by improper or unfamiliar operation, and people of the same category often made the same type of mistakes. The cause of the failure can be inferred by analyzing the characteristics of the user and past events. Therefore, if the cause of the failure is predicted in advance based on past events, and the solution is fed back to teachers based on the predicted results, this will greatly enhance their teaching experience. At the same time, according to the prediction results, the relevant knowledge is pushed through the network, which can also help the teacher to reduce the probability of failure.

In recent years, deep learning has made breakthroughs in both theory and practice. In the computer image recognition, natural language processing, automatic driving, speech recognition has achieved notable achievements [1].
Series prediction is one of the hotspots in deep learning research. It has extensive and extremely important applications in many fields from recommendation systems to speech recognition to natural language processing [2]. Series prediction uses a set of training sequences to train the prediction model. The model is used to perform sequence prediction after training. The prediction includes the next item of the prediction sequence. For example, based on the products that the user browses before, it will predict the products that the user may browse, and thereby recommend or pre-load the relevant product data to the user. It has a lot of applications in web prefetching, product recommendation, weather forecasting and even financial forecasting [2, 3].

Research on sequence prediction has been done a lot. There are many different implementation models, the most popular of which is Prediction by Partial Matching (PPM). It is based on the Markov property and at the same time inspires many other models, such as Dependency Graph (DG), All-K-Order-Markov (AKOM), Transfer Directed Acyclic Graph (TDAG), Probability Suffix Tree (PST) and Context Tree Weighting (CTW) [4]. Over the years, these models have improved greatly in terms of time or memory efficiency, but their performance remains almost unchanged in terms of prediction accuracy. At the same time, some compression algorithms are used for sequence prediction, such as lz78 and active lezi. Besides, machine learning algorithms such as neural networks and sequential rule mining have been applied to perform sequence prediction [8].

However, these models have some outstanding limitations. On the one hand, most of them make Markov’s assumption that each event depends only on previous events. If this assumption is not true, the prediction accuracy of using these models will be seriously reduced. In addition, these models are built with partial information contained in the training sequence. Therefore, these models do not use all the information contained in the training sequence to perform predictions, which also seriously reduces the prediction accuracy of the model [5].

Consequently, Ted Gueniche, Philippe Fournier-Viger et al. propose a model called Compact Prediction Tree (CPT) [4]. It exploits the similarity between sub-sequences to compress the training sequence without losing information. Additionally, it is more accurate than the most advanced PPM, DG, and AKOM models on a variety of real data sets. At the same time, based on this, it proposes to use compressed sub-sequences and simple branches as well as noise improving to reduce the space-time complexity of the compressed prediction tree, thereby improving the accuracy of prediction [5].

This paper proposes to use CPT for sequence learning to predict the most possible cause of failure when teachers encounter difficulties in use so as to provide a suitable solution.

2. Compact Prediction Tree (CPT)

CPT is a predictive model. The difference between it and other predictive models is that CPT stores the compressed representation of the training sequence with barely loss or no loss. The CPT facilitates the similarity of the measurement sequence to the training sequence to perform the prediction. Second, CPT utilizes the similarity of the measurement sequence and the training sequence to perform the prediction. The similarity measurement is noise tolerant, thus allowing the CPT to predict the next item of the subsequence that is not previously seen in the training sequence. Other proposed models such as PPM and All-K-Order-Markov [5, 6] fail to perform predictions in this case.

The CPT accepts a set of training sequences corresponding to a set of test sequences. A training sequence is used as input during the training process, and then three different structures are generated, including a Prediction Tree (pt), a lookup table (lt), and an inverted index (ii). In the training process, the three structures are gradually established one by one. For example, it is supposed to have a series of $S=\{s1,s2,s3,s4,s5\}, s1 = \langle a,b,c \rangle, s2 = \langle a,b,c,d \rangle, s3 = \langle b,c \rangle, s4 = \langle a,b,d \rangle$ and $s5 = \langle e,a,b \rangle$, and the letter set is $Z=\{a,b,c,d,e\}$. The process produced by the three structures is the process of generating a prediction tree (pt). The principle is that the first element starts, adds it to the root node, and adds the successor nodes in turn. For each new sequence, it will start again from the root node. It skips if an element is added to the structure, such as $s1=\{a,b,c\}$. Therefore, the first thing to do is to get a root node and a current node that is initially set as the root node, and then check if the first element exists. If not, add a to the sub list of the root node and to the inverted index with a value of 1. Then the above
steps are repeated, with b added after a and c added after b [6]. The process of generating prediction trees is shown in Figure 1.

Figure 1. Example of prediction tree generation process.

The entire process is repeated for all sequences in the data. After that, it is enabled to obtain a full prediction tree and a complete inverted index represented by the hash value. At the same time, it is easy to get the last element of each series and a complete lookup table. After getting the complete pt, lt, and ii, it is possible to predict the test data set.

The data prediction of CPT is divided into three steps. (1) Using the inverted index obtained earlier to find all unique sequences similar to the target sequence and taking the intersection; (2) Finding subsequent sequences similar to the target sequence; (3) Calculating the score of each item of the subsequent sequence, which is calculated as follows:

If the item already exists, \( \text{score} = (1 + (1/\text{number of similar sequences}) +(1/\text{number of items currently in the countable dictionary}+1)*0.001) \times \text{oldscore}. \)

Otherwise, \( \text{score} = 1 + (1/\text{number of similar sequences}) +(1/\text{number of items currently in the countable dictionary}+1)*0.001. \) [6]

Return the maximum value as the predicted value.

3. Using CPT to predict classroom equipment failure

3.1. Processing of raw data

3.1.1. It is required to perform subject extraction of fault events and sequence marking of fault descriptions. In order to use CPT for prediction, it is first necessary to sort out the fault categories in the original data and convert them into CPT alphabet data. A lot of raw data is very vague in describing faults, and there are too many useless phrases. After the invalid data of the original data is deleted, the remaining 14369 fault data is extracted with the Latent Dirichlet allocation, and the manual proofreading is performed. As a result, it is found that the equipment failure includes a total of 10 classifications, which are sequentially labeled as A-J. The specific contents are as shown in Table 1.

| Fault description | tag | Fault description | tag |
|-------------------|-----|-------------------|-----|
| Projector problem | A   | The system cannot be started. | B   |
| There is no sound | C   | Mobile storage device failure | D   |
| Software failures | E   | Computer start-up failure | F   |
| Non-electronic malfunction | G | Computer peripheral malfunction | H |
| Network fault | I | Projection display problem | J |

3
3.1.2. The fault processing sequence is extracted and transformed into the corresponding letter sequence data. Since the extraction of the fault processing sequence is of great importance, the accuracy of the sequence extracted by this step will affect the formation of the prediction tree, thereby affecting the accuracy of the final prediction. First, each fault handling event is divided into five parts, which are user gender (user_gender), subject background (user_subject), age (age), fault description (trouble_des), and fault reason (trouble_rea). The first three items are added because these three items have a lot of impact on the results. In addition, according to the compression strategy of the CPT, the branch that finally generates the root node of the prediction tree can be reduced, and the frequent subsequences are compressed. No doubt, on the other hand, it is possible to increase the branch of the prediction tree, but there is a small increase in the accuracy of the prediction. This paper defines user_gender={M,W} (M:male, W:female), user_subject={L,S} (L:liberal, S:science), age={Y,O} (Y:youth, O:old), trouble_des is shown in Table 1, and trouble_rea is shown in Table 2. It should be noted that the trouble_rea table is obtained by subdividing the subjects from all the original data sets. Some subjects have different descriptions, such as the computer is not powered, the projection cannot be turned on, and the amplifier is not turned on without sound. However, the reason is the same, that is, the device power is not turned on.

Table 2. trouble_des fault reason symbol mark table.

| The cause of the problem (tag) | The cause of the problem (tag) | The cause of the problem (tag) | The cause of the problem (tag) |
|-------------------------------|-------------------------------|-------------------------------|-------------------------------|
| The main power supply is on at the end | P1 | The power supply of the equipment is not turned on | P2 |
| Information switching problem | N1 | Hardware damage | K2 |
| Software damage | Q2 | Software operation problem | Z1 |
| Display damage | K4 | Power amplifier damage | T |
| Poor microphone | V | Central control hardware is bad | X |
| Inability to operate equipment | Z2 | | |

Below we select four typical fault events as an example to illustrate the process of generating sequences, as shown in Table 3.

Table 3. List of fault events.

| S1 | Zhang *, young female teacher of college of arts: the projector cannot be turned on. |
| S2 | Li *, an elderly male teacher of art college: the projector has no signal. |
| S3 | Li *, male teacher of the school of mathematics: the projector has no signal. |
| S5 | Song *, teacher of college of physics and electrical engineering: no sound. |

According to the previous definitions, after conversion to alphabet sequence data, S1={W,L,Y,A,P2}, S2={M,L,O,A,N1}, S3={M,S, A, P2}, S4={S, C, Z1}. It can be found that some sequences lack the user_gender item or the user_age item, which is missing from the original data set.

After all the data is converted into alphabet sequence data, since the data set is relatively small, the method of K-fold verification is selected. In this case, K=3.

3.2. CPT training and testing

CPT provides libraries written in JAVA. For convenience, the CPT Python library written by Neeraj Singh Sarwan is chosen [7] (https://github.com/analyticsvidhya/CPT) and appropriate modifications are made to train CPT. the training hardware platform uses a MacBook pro2018 with 16G running memory. The first training time is 13 minutes and 35 seconds, and the training accuracy is 0.98431. As
shown in Figure 2, the prediction accuracy on the test set is 0.89357, which is not particularly high.
After several times of debugging, it is found that if all the missing sequence user_gender, user_subject, user_age can be added, the training precision on the test set will increase. Therefore, the data is complemented by a random method. The prediction accuracy of the test set rises to 0.94231 and the training time is 13 minutes and 56 seconds.

On the other hand, it is worth noting that, in order to reduce the prediction tree branching, unlike other items, the complement or non-complement user_age has little change in prediction accuracy. However, if user_age is completely deleted, the training time will decrease rapidly, but it will affect the prediction accuracy.

![Training and validation accuracy](image1)

Figure 2. Comparison of the first training accuracy and test accuracy.

Therefore, the final choice is to complement user_gender with user_subject, leaving user_age unchanged. This increases the prediction accuracy of the test set and effectively reduces the branching of the prediction tree. In this way, the final training accuracy is 0.98531, the test accuracy is 0.95893, and the time is 13 minutes and 49 seconds, as shown in Figure 3.

![Training and validation accuracy](image2)

Figure 3. Comparison of final training accuracy and test accuracy.

4. Conclusion
This paper uses CPT to predict the cause of the user’s use of multimedia classroom equipment failure. The final test prediction accuracy is not very high, but at least it is realistic. Although the fault has certain certainty, the formation of some fault causes is very complicated. In addition, everyone has different familiarity with the system, their descriptions are relatively discrepant, so it is difficult to make accurate descriptions. Therefore, it is unlikely that accurate predictions will be made. In fact, CPT greatly reduces the workload of the administrator and at the same time enhances the user experience, thereby better solving problems in a timely manner.

In what way is the raw data converted into sequence data to best perform CPT. Compared with other CPT applications, this study consumes a longer training time and there is room for further optimization. All of these are the questions that need to be considered in the next step of research work.
References
[1] Huang liwei, jiang bitao, lu shouye, liu yanbo, li dey.(2018)Research review of recommendation system based on deep learning. Acta computerica sinica, 2018,41(07):1619-1647.
[2] Deshpande,M., g. Karypis.(2014) Selective Markov models for predicting Web page accesses.ACM the Transactions on the Internet Technology, vol. 4 no. 2:163-184.
[3] Padmanabhan,V.N., Mogul, J.C. (1998) Using Prefetching to Improve World Wide Web Latency. Computer Communications, vol. 16: 358-368.
[4] Gueniche T, Fournier -Viger P. Tseng V s. (2013) Compact Prediction Tree: Lossless Model for Accurate Sequence Prediction. In: Advanced Data Mining and Applications: 9th International Conference.Hangzhou,China. pp.14-16.
[5] Gueniche T, Fournier - Viger P, Raman R, et al. (2015) CPT + : Decreasing the Time/Space Complexity of the Compact Prediction Tree. In: Pacific - Asia Conference on Knowledge Discovery and Data Mining.Ho Chi Minh,Vietnam.pp.625-635
[6] Neeraj Singh Sarwan. (2018)A Guide to Sequence Prediction using Compact Prediction Tree (with codes in Python).https://www.analyticsvidhya.com/blog/2018/04/guide-sequence-prediction-using-compact-prediction-tree-python/.
[7] Neeraj Singh Sarwan. (2018),CPT (Compact Prediction Tree).https://github.com/analyticsvidhya/CPT.
[8] Fournier-Viger P, Gueniche T, S.(2012) Tseng V .Using Partially - Ordered Sequential Rules to Generate More Accurate Sequence Prediction.In:Advanced Data Mining and Applications: 8th International Conference. Nanjing, China. pp.431-442.