Decision Tree Learning Algorithm Dealing with Distorted Learning Data

Kazuyoshi Matsuoka, Hiroaki Kikuchi, and Shohachiro Nakanishi
Dept. of Electric Engineering, Tokai University
1117 Kitakaname, Hiratsuka, Kanagawa 259-1292, Japan
{matsu, kikn, nakanishi}@ep.u-tokai.ac.jp

Abstract— **This research deals with a problem of extracting the characteristics of lectures form a given course evaluation data. In this study, we have a learning data with multiple attributes of questionnaires and multiple classes of lectures, and apply a well-known decision tree learning algorithms. The existing learning algorithm, however, does not deal with a distortion of learning data, in which the number of classes varies by lectures. Because of the distortion issue, some of lectures may be lost during pruning tree. Therefore, based on the ID3 algorithm we propose a new algorithm that is able to deal with an uneven data appropriately. In this paper, we report the experimental result on our proposed algorithm using the actual course evaluation data.**

I. INTRODUCTION

Data mining, a technology that discovers a useful knowledge from a large amount of data, is going to be required with the advance of the information technology. Questionnaires of a course evaluation is one of the prime example of data mining. An algorithm of the data mining such as decision tree learning allows us to extract the characteristics of lectures form a given course evaluation data. In this example, a learning data consists of multiple attributes of questionnaires and multiple classes of lectures.

A number of students per class is not always even. The uneven learning data causes distortion in the learning results. For example, Table I shows the actual statistics of course evaluation, where the most-students class has about 5 times many students of the least. Unfortunately, a pruning step in decision tree learning accelerates the distortion. During a pruning process, classes with smaller number of students may be ignored or even lost. From the viewpoint of extracting the characteristics of every lectures, the class with smaller number of students should be appeared in the resulting tree.

In order to address the distortion issue, based on the ID3 [1] algorithm we propose a new algorithm that is able to deal with an uneven data appropriately. In the proposed algorithm called "inflating method" , even if there is a difference in number of students among the classes, it manipulates the uneven learning data so that all classes have same number of students.

In this paper, after reviewing decision tree algorithm and course evaluation, we evaluate the proposed algorithm using artificially generated learning data and actual course evaluation data. We show the comparison in terms of size of tree, accuracy, and entropy of classes.

II. RELATED WORKS

There are various studies in data mining using decision tree learning algorithms. In [2], Kikuchi applied decision tree learning algorithm C4.5 [3] to the course evaluation data and showed it can be used to extract some meaningful logical propositions, such as "logic of a good lecture". In [4], Takasaki et.al applied decision tree learning algorithm to web pages classification, in which given sample categories from a directory search engine (Yahoo), new web pages are classified into an appropriate category automatically. In [5], Yoneyama et.al studied the issue how to treat the missing value in the decision tree learning algorithms. They proposed a new algorithms of using a missing value and compared it with existing algorithms in terms of the error ratio of a decision tree.

Both researches of [2] and this deal with the analysis of a course evaluation data. The originality of this work is to extract the characteristics of lectures for each lecturer, and to deal with the distortion in learning data.

III. FUNDAMENTAL DEFINITION

A. Course Evaluation Data

In many universities, a course evaluation is performed for an improvement of teaching skills. The results of course evaluation are published from in web pages.

In this research, we investigate a course evaluation data shown in Table II. In the evaluation, a student provides a state of attendance, an achievement through the course, an overall evaluation, and answer to the questions listed in Table III with choice out of three, "good" , "poor" , and "don't care".

B. Survey of Course Evaluation in Japan

The style of evaluation varies in universities. In order to clarify an average style in evaluation, we survey over 100 web pages randomly sampled by search engine. Table IV and V show the results of survey on course evaluation.

According to the survey, most universities perform an anonymous style of questionnaire with similar questions.

TABLE IV
List of Institutes in The Survey

university	90
graduate school	4
junior college	4
junior high school	2
sum	100

C. Decision Tree Learning Algorithm ID3 and C4.5

The ID3 and C4.5 are decision tree algorithms, to get knowledge from a database, which is called a learning data. The learning data D consists of m cases. A case consists of n attributes $(a_1, \ldots, a_n)$ for which boolean values are assigned.

TABLE I

Uneven Learning Data(course evaluation questionnaire)

lecture	the total number of students	the number of subjects
O	51	1
U	109	2
Y	62	1
M	91	2
N	91	2
I	49	1
A	96	2
M2	161	2
M3	82	2
O2	56	1
K	35	2
O3	184	3
M4	192	3

TABLE II

A Course Evaluation Data

target	Dept. of Electric Engineering Tokai University, 1st-4th years students
duration	2000 Spring term
target number of lecture	13 lectures(24 courses)
total number of students	1259

TABLE III

List of Questions

No	questions
Q_1	Are the contents of the teachings understandable?
Q_2	Did the lecturer hold your attention and make you think?
Q_3	Did the lecturer speak loudly enough to be heard by all participants?
Q_4	Did the lecturer write legibly on blackboard?
Q_5	Were the class well controlled so that no local communication happens?
Q_6	Was the syllabus useful?
Q_7	Did the lecturer guide you for self-learning?
Q_8	Did the lecturer use audio, video and transparency effectively?
Q_9	Were the textbook and the references adequate to the class?
Q_{10}	Was grading and evaluations fair and clear?
Q_{11}	Did the lecturer encourage you to ask question?
Q_{12}	Did the contents cover up-to-date technologies?

TABLE VI

Top 10 Questions Asked in Many Universities

rank	questions	case
1	Are the contents of the teachings understandable?	87
2	Did the lecturer write legibly on blackboard?	79
3	Did the lecturer speak loudly enough to be heard by all participants?	78
4	How was the attendance rate?	71
5	Did the lecturer use audio, video and transparency effectively?	69
6	Was the lecturer enthusiasm?	67
7	Did the lecturer hold your attention and make you think?	61
8	Was the syllabus useful?	57
9	Were the textbook and the references adequate to the class?	54
10	Can you understand the explanations?	48

The learning data D is divided into k subsets, labeled by $C_1, \ldots, C_k$. The algorithms generate the decision tree, which classifies any given data into one of k classes according to the characteristics of the D. In a decision tree, a node, a branch and a leaf correspond to attribute, attribute value and classes, respectively.

The decision tree learning algorithm such as ID3 and C4.5 tries to generate the simplest tree by testing attributes based on expected reduction in entropy and selecting nodes, until a tree is constructed. The C4.5 has improved in the followings.

- continuous-valued attributes
- missing value

TABLE V
STYLES OF QUESTIONNAIRES

anonymous	96
signed	3
either	1
sum	100

When an attribute value is missing, the C4.5 can deal the missing value with a new symbol, which means multiple attribute values with probability distributed according to the relative frequency of known results. In other words, when there are x missing values and y known values out of m, missing value is classified into y with probability of $x \cdot y/(m-x)$. In the questionnaire of our study, we deal answers labeled as "don't care" with missing values.

- tree pruning
  When a learning data is too large, the resulting tree is often very complex with many meaningless contents. For the issue, the C4.5 simplifies the tree according to the following parameters.
  - M(weight)
    This means the minimum number of cases to be classified. The bigger weight, the smaller tree.
  - CF(Confidence level)
    This means an error rate of classification. The smaller value resulting in simpler tree, gives many missclassifications.

D. Proposed Algorithm "inflating method"

A learning data is not always evenly distributed. A distortion in classes may spoil the consistency of resulting decision tree. In order to address the issue of uneven learning data, we propose an algorithm, called "inflating method" , in which informally, the cases of smaller subsets are duplicated until all subsets become the same size.

Consider an uneven learning data D in Table VII. Class A is the largest subset and C is the least.

Inflating Method

1. For the greatest subset (class A), leave it as it was.
2. For single-case subset (class C), duplicate the case as many times as the greatest class has. For instance, case d_6 is copied two times.
3. For other smaller subset, duplicate cases such that the ratio of cases does not change. For example, class B must be inflated up to 3. But, there are 2 possible choices, d_4 and d_5. In this case, we choose one of two cases randomly, say, d_4 to be d_7.

Table VIII shows the result of inflating. Note that the inflating method balances the original data in the cost of accuracy. For example, in class B, the original data has d_4 and d_5, while inflated data has two d_4 and one d_5. So, the rate is different to the original data.

We denote the inflated data by D^*, to which the ID3 is applied.

TABLE VII
UNEVEN LEARNING DATA D (BEFORE INFLATING)

case	a_1	a_2	a_3	class
d_1	1	0	0	A
d_2	1	1	1	A
d_3	1	0	1	A
d_4	1	1	0	B
d_5	0	1	1	B
d_6	0	1	0	C

TABLE VIII
LEARNING DATA D^* APPLIED THE INFLATING METHOD

	a_1	a_2	a_3	class
d_1	1	0	0	A
d_2	1	1	1	A
d_3	1	0	1	A
d_4	1	1	0	B
d_5	0	1	1	B
d_6	0	1	0	C
d_7	1	1	0	B
d_8	0	1	0	C
d_9	0	1	0	C

IV. EVALUATION

A. Evaluation with The Randomly Generated Data

For the purpose of evaluating the proposed algorithm, we set up a learning data in the following way. Note that the generated data is "ideal" , i.e., it doesn't contain noise and contradicted data.

1. Generate completely balanced decision tree T with leaves chosen at random. Fig. 1 show the instance of randomly generated tree with 5 classes ($k = 5$), 8 attributes ($n = 8$), and depth of 5.
2. For each class, generate 120 cases with attributes assigned random valve resulting in the total of 600 cases (D_0). Sampling 20 cases for each class at random gives a 100-cases data (D_1).
3. Make the sampling data distorted by adding and removing randomly. By D_2, we denote the distorted data which is generated from D_1 by randomly removing class C, D and E, and adding some cases in class A and B randomly chosen from D_0. Applying the inflating method to D_2, we have D_2^*.
4. We set up a test data D_3, which is a 100-cases data randomly chosen from D_0, independent of the learning data D_1.

In comparison between the ID3 and the inflating method, for the distorted data D_2, we show the results of the ID3 and the inflating method in Fig. 2 and 3, respectively. The parameters of pruning are $M = 2$ and $CF = 25$.

An error rate means the ratio of wrongly classified cases to all cases. We define a recall R_i and a precision P_i for i-th class as follows.

$$R_i = \frac{\text{the \# of correctly classified cases}}{\text{the \# of all cases in } i\text{-th class}} \quad (1)$$

$$P_i = \frac{\text{the \# of correctly classified cases}}{\text{cases were classified for } i\text{-th class}} \quad (2)$$

Table XI shows the confusion matrix of the inflating method evaluated over the test data D_3. The numbers of correctly classified cases are indicated at diagonal elements. Table XII shows the recall and precision of the inflating method in the test data.

TABLE IX
THE NUMBER OF CASES FOR FIVE CLASSES IN FIVE LEARNING DATA

		class				
		A	B	C	D	E
all data	D_0	120	120	120	120	120
sampling data	D_1	20	20	20	20	20
distorted data	D_2	33	24	19	12	5
inflated data	D_2^*	33	33	33	33	33
test data	D_3	20	20	20	20	20

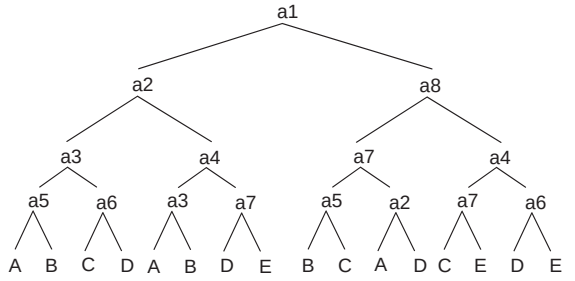

Fig. 1. random tree T for randomly generating data

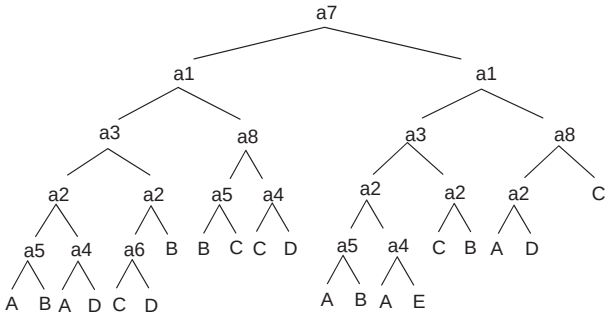

Fig. 2. the result of applying ID3 to the distorted data D_2

B. Evaluation with The Actual Data

We evaluate the proposed algorithm using the actual questionnaire data in Table II in the same way.

Table XIII, Fig. 4 and 5 show the result of learning, the decision trees generated by the ID3 and the inflating method, respectively. The pruning parameters of $M = 10$, $CF = 30$ is used for all data. Where, the right and left branches indicate values of "good" and "bad". Following nodes from root means the conjunction of the attributes in the path.

C. Remarks on Experiments

- The number of inflating cases increases depending on the distortion. In the both examples, it increases about twice.
- The size of decision tree grows in inflating. The randomly generated data grows 1.2 times, while the actual data grows 3 times. The entropy of class increases with inflated data.
- Error rate doesn't change by inflating. An actual data has higher error rate than the randomly generated data. The contradictions and noises in the actual data are the main sources of the error.
- Minority classes, which have been removed in Fig. 4, are appropriately treated in the decision tree in Fig. 5. Therefore, we claim that the proposed inflating method can deal with any distorted data.
- The proposed algorithm deals with any distorted data. But, the extremely distorted data may not be deal with correctly since an enormous data are necessary to be inflated, which costs too much.
- From the observation of Table XII, class D and E cause many wrong classification and both of precision and recall decrease as many cases are inflated. Hence, the accuracy of classification strongly depends on the quantity of inflating.

TABLE XII
ACCURACY OF THE INFLATING METHOD

	recall R_i	precision P_i	# of inflated cases
A	1	1	0
B	1	0.9	7
C	1	0.625	14
D	0.75	0.625	21
E	0.05	0.5	28

TABLE XIII
THE EXPERIMENTAL RESULT OF THE INFLATING METHOD FOR THE ACTUAL DATA ($M = 10, CF = 30$)

	learning data	
	ID3	inflating method
# of cases m	1259	2496
size	29	86
error rate[%]	69.9	69
entropy[bit]	2.96	3.5
# of lecturers (class)	9	13

V. CONCLUSION

We propose an inflating algorithm for distorted learning data, and evaluate size of decision tree, error rate, and en-

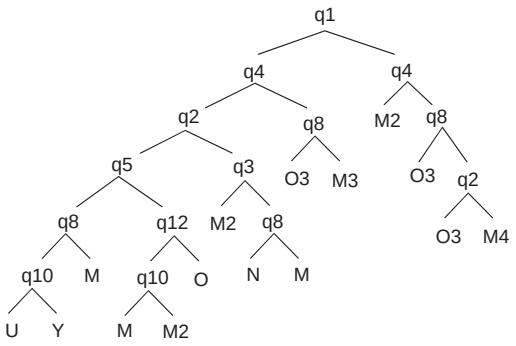

Fig. 4. the result of ID3 for actual data

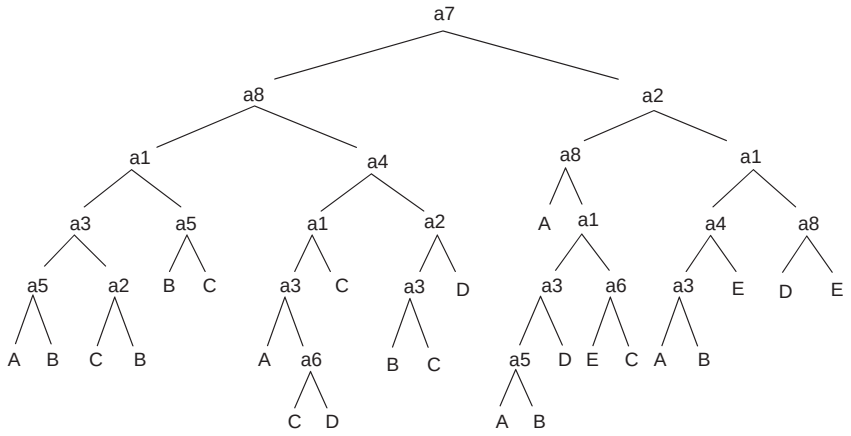

Fig. 3. the result of the ID3 to the inflated data D_2^*

TABLE X
THE EXPERIMENTAL RESULT OF THE INFLATING METHOD FOR RANDOMLY GENERATED DATA ($M = 2, CF = 25$)

	ID3		inflating method	
	learning data D_2	evaluation data D_3	learning data D_2^*	evaluation data D_3
# of cases m	93	100	165	100
size (# of nodes)	39	39	47	47
error rate [%]	3.2	22	7.3	24
entropy [bit]	2.18	1.97	2.28	2.08

TABLE XI
THE CONFUSION MATRIX IN THE INFLATING METHOD

	classified as					
class	A	B	C	D	E	the # of inflated cases
A	20					0
B		20				7
C			20			14
D		1	3	15	1	21
E		1	9	9	1	28

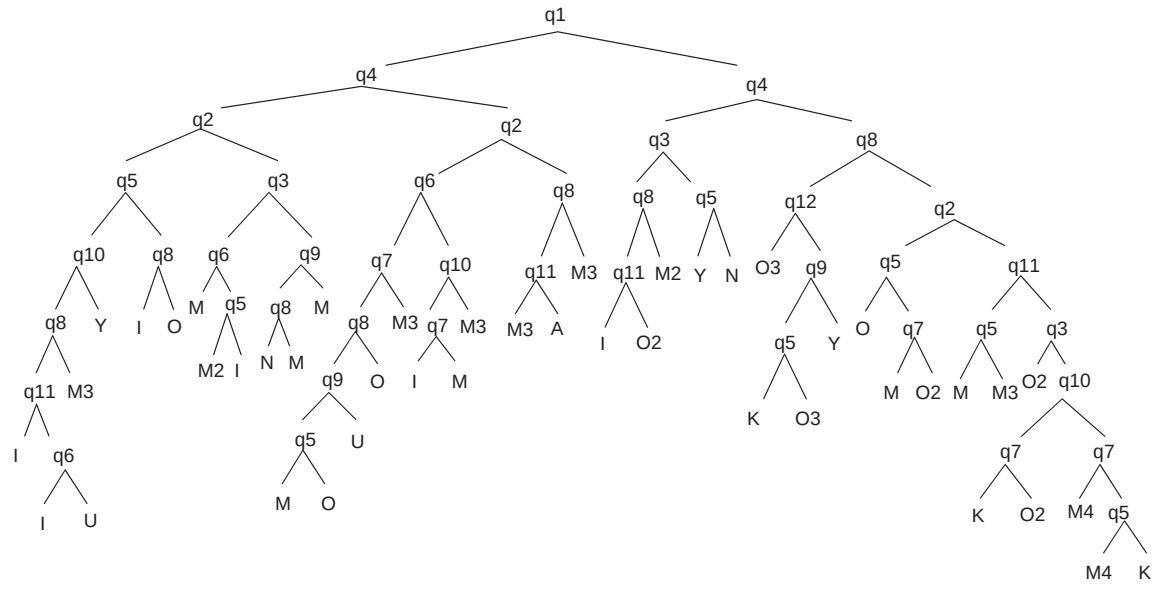

Fig. 5. the result of inflating algorithm to the actual data

tropy. We conclude that the proposed algorithm generates appropriately decision tree which classifies most minority classes that have been ignored so far.

Future studies include an investigation on the number of leaves for each class, a treatment of missing value, and improvement of the efficiency of the proposed algorithm.

Acknowledgement

In carrying out this research, we thank Mr.Isao Takasaki of graduate school of engineering, Tokai university for his contribution.

References

[1] Motohide Umano, "ID3" , Japanese Joumal of Fuzzy Theory and Systems vol. 6, No. 3, pp. 502-504, 1994.

[2] Hiroaki Kikuchi, "Decision Tree Analysis of Course Evaluation Data" , 6th Workshop on Evaluation of Heart and Mind, 2001.

[3] J. Ross Quinlan, "C4.5: Programs for Machine Leaning", Morgan Kaufmann pub., 1993.

[4] Isao Takasaki, Kazunori Yoneyama, Hiroaki Kikuchi, and Shohachiro Nakanishi, "On Extraction of Keywords for Classification of Webpages" , 7th Workshop on Evaluation of Heart and Mind, pp. 9-14, 2002.

[5] Kazunori Yoneyama, Hiroaki Kikuchi, and Shohachiro Nakanishi, "Handing with Missing Attribute in Decision Tree" , 17th Fuzzy System Symposium, pp. 743-746, 2001.