Full Paper Bioscience of Microbiota, Food and Health Vol. 33 (3), 129–138, 2014

Technical Aspects of Nominal Partitions on Accuracy of Data Mining Classification of Intestinal Microbiota — Comparison between 7 Restriction Enzymes

Toshio KOBAYASHI[1, 2*] and Kenji FUJIWARA[2, 3]

[1]*Miyagi University, 2–2–1 Hatadate, Taihaku-ku, Sendai City, Miyagi 982-0215, Japan*
[2]*Riken, 2–1 Hirosawa, Wako, Saitama 351-0198, Japan*
[3]*Yokohama Rosai Hospital, JLHWO, Kozukue-cho, Kohoku-ku, Yokohama 222-0036, Japan*

Received July 9, 2013; Accepted February 8, 2014; Published online in J-STAGE May 16, 2014

The application of data mining analyses (DM) is effective for the quantitative classification of human intestinal microbiota (HIM). However, there remain various technical problems that must be overcome. This paper deals with the number of nominal partitions (NP) of the target dataset, which is a major technical problem. We used here terminal restriction fragment length polymorphism data, which was obtained from the feces of 92 Japanese men. Data comprised operational taxonomic units (OTUs) and subject smoking and drinking habits, which were effectively classified by two NP (2-NP; Yes or No). Using the same OTU data, 3-NP and 5-NP were examined here and results were obtained, focusing on the accuracies of prediction, and the reliability of the selected OTUs by DM were compared to the former 2-NP. Restriction enzymes for PCR were further affected by the accuracy and were compared with 7 enzymes. There were subjects who possess HIM at the border zones of partitions, and the greater the number of partitions, the lower the obtained DM accuracy. The application of balance nodes boosted and duplicated the data, and was able to improve accuracy. More accurate and reliable DM operations are applicable to the classification of unknown subjects for identifying various characteristics, including disease.

Key words: human intestinal microbiota, operational taxonomic unit, data mining analysis, decision tree, nominal partitions of data, accuracy of classification, balance node

INTRODUCTION

Human intestinal microbiota (HIM) is related to our health, and practical research on the relationship with the human immune systems and diseases is now being widely performed. Our previous papers [1–3] have assessed HIM data obtained by data mining analysis (DM) for quantitative classification of the relationship between subject characteristics. The results were fruitful, but due to the unique application of DM to HIM, some accumulation of case studies is required for further DM operations. The selection of primer-restriction enzymes and the number of nominal partitions (NP) of assigned characteristics are important factors for reliable applications. This paper aims to compare the effects of both factors for obtaining accurate and dependable DM results, which are the major technical problems of practical applications.

The number of NP, which is a partition of assigned characteristics and depends on the purpose of the analysis, directly affects the accuracy of the DM results. In other words, proper NP application to the data is necessary. Our previous paper [3] already dealt with a simple 2 nominal partition (2-NP), i.e., Yes or No, and examined the accuracy between the 7 restriction enzymes. Here, we aim to further examine another 2 types of NP, 3-NP and 5-NP, and to compare the 3 types of NP, including 2-NP, as shown in Table 1, with 2 characteristics, the latter of which was reported previously [3], but is included here for comparison. The original operational taxonomic unit (OTU) data applied in this paper were the same as reported in our previous papers [1–4], but the detailed NPs are different. As with the previous paper [3], dietary factors for healthy male subjects were controlled, which is an important starting point for the quantitative analysis of HIM.

HIM are represented here as OTUs by terminal

*Corresponding author. Mailing address: Miyagi University, 2–2–1 Hatadata, Taihaku-ku, Sendai City, Miyagi 982-0215, Japan. Fax: +81-3-717-7398. E-mail: toskoba@attglobal.net

Table 1. 2 to 5 nominal partition (NP) of the 92 subjects

characteristics	NP #	mark	area / meaning	N. of subjects
Smoking	2-NP	SA	No, non-smoker + non-respondent	76
		SB	Yes, smoking now	16
	3-NP	SAA	non-smoker + non-respondent	57
		SAP	all previous smokers, not now	19
		SBB	all present smokers	16
	5-NP	SAA	non-smoker + non-respondent	57
		SPA	previous smoker, cess. P. ≧5Y	14
		SP 5	previous smoker, cess. P. <5Y	5
		SBG	smoker, 15cigarettes/d. or less	12
		SBH	heavy smoker, 16cigs./d. or more	4
Drinking	2-NP	DA	No, non-habitual drinker	47
		DB	Yes, habitually drinking now	45
	3-NP	DA	non-habitual drinker	47
		DBS	habitual drinker, 1-3 days/w.	21
		DBF	habitual drinker, 4-7 days/w.	24
	5-NP	DA	non-habitual drinker	47
		DBL	drinker, average <20 ml・AlOH/d.	11
		DB 2	drinker, average 20-40 ml・A/d.	18
		DB 3	drinker, average 41-60 ml・A/d.	6
		DBH	drinker, average 60 ml・A/d.<	10

N.: number; cess. P.: smoking cessation period; 5Y: 5 years; w.: week; d.: day; AlOH, A: alcohol; Shadows at '2-NP' indicated that the results have been reported previously [3], but are shown here for comparison to 3-NP and 5-NP.

restriction fragment length polymorphism (T-RFLP) analysis. The relationship between OTUs and subject characteristics was assessed by cluster analysis, using the methods of Jin [4] and Andoh [5, 6], or by Pearson correlation coefficients and principal component analysis. To date, DM has been applied to the relationships between genes, single nucleotide polymorphisms (Merelli [7]) and inflammatory bowel disease (Merelli [8]), as well as to age-dependent genes (Kirschner [9]) and hormone levels (Modlin [10]), but has not been applied to general HIM. i.e., OTUs.

OTUs are thought to contain numerous types of bacteria, and their composition directly affects the accuracy of DM classifications. We therefore applied 7 restriction enzymes for better comparisons of subject classification. DM will be applied to classify all OTU data, of which characteristic have various NPs, e.g., types or symptoms of diseases; thus, for effective DM operation, systematic comparisons are required and are examined here.

MATERIALS AND METHODS

As reported previously [4], to avoid the influence of dietary factors, we designed identical meals (1,879 kcal/day), which were fed for 3 days to 92 healthy male volunteers living in Japan. Age and body mass index (BMI) of the subjects were 21–59 years (average: 36.8 years) and 17.3–30.2 kg/m^2 (average: 22.6 kg/m^2), respectively. Fecal samples were analyzed by T-RFLP using 7 restriction enzymes [2, 4]. T-RFLP was applied due to its reproducibility, comparatively low cost and convenience with regard to DM operation. Studies were performed in accordance with the protocols approved by the Riken Research Ethics Committee (Wakou 2009-3rd 21-13), and the OTU data were accumulated by the Benno Laboratory, Riken, Japan.

Bacterial DNA was isolated from feces using a modification of the method described by Matsuki [11]. Amplification of fecal 16S rRNA, restriction enzyme digestion, size fractionation of T-RFs and T-RFLP analysis were carried out as described previously [12–14]. Details of amplification and T-RFLP analysis with the 7 restriction enzymes, i.e., 516f-*Bsl*I, 516f-*Hae*III, 27f-*Msp*I, 27f-*Alu*I, 35f-*Hha*I, 35f-*Msp*I and 35f-*Alu*I, were as described in our previous papers [2, 4].

The amounts for each OTU represent the fluorescence intensity and concentration. The obtained OTU data are abbreviated here as B--- (---: base pair number) for 516f-*Bsl*I, HA--- for 516f-*Hae*III, M--- for 27f*Msp*I, A— for 27f-*Alu*I, QHh—for 35f-*Hha*I, QM—for 35f-*Msp*I and QA—for 35f-*Alu*I. We had 2 groups of OTUs: 516f- + 27f- (4 restriction enzymes), and 35f- (3 restriction enzymes). The component numbers of these 7 enzyme groups were 27·B, 33·HA, 20·M, 40·A, 31·QHh, 34·QM and 48·QA; thus, if we combined all the enzyme components of the 2 groups, the former had a maximum of 120 OTUs, and the latter had a maximum of 113 OTUs. On account of the balance between the number of subjects (92) and OTU components, we did not mix the data from the 2 groups to avoid the problem of field alignment sequences described in previous reports [2, 3]. Various sets of restriction enzymes were combined, and the data were arranged with the answers of the 92 subjects. The resulting 2-dimensional Excel data were analyzed using DM software (IBM-SPSS, Clementine14).

A DM algorithm (Classification and Regression Tree (C&RT) modeling system), which is the most typical method of DM, provides a Decision tree[1] (Dt). The Dt explicitly classifies the various groups of subjects according to the assigned characteristics, as shown in Table 1. C&RT divides subjects into two subsets by comparing the Gini coefficient[2] according to the OTU data, such that the subjects within each subset are more homogeneous than in the previous subset. The C&RT system is flexible, and allows unequal misclassification costs to be considered when comparing to the other modeling systems of DM. A major specialty of DM and the constructed Dt is that a single selected OTU is used

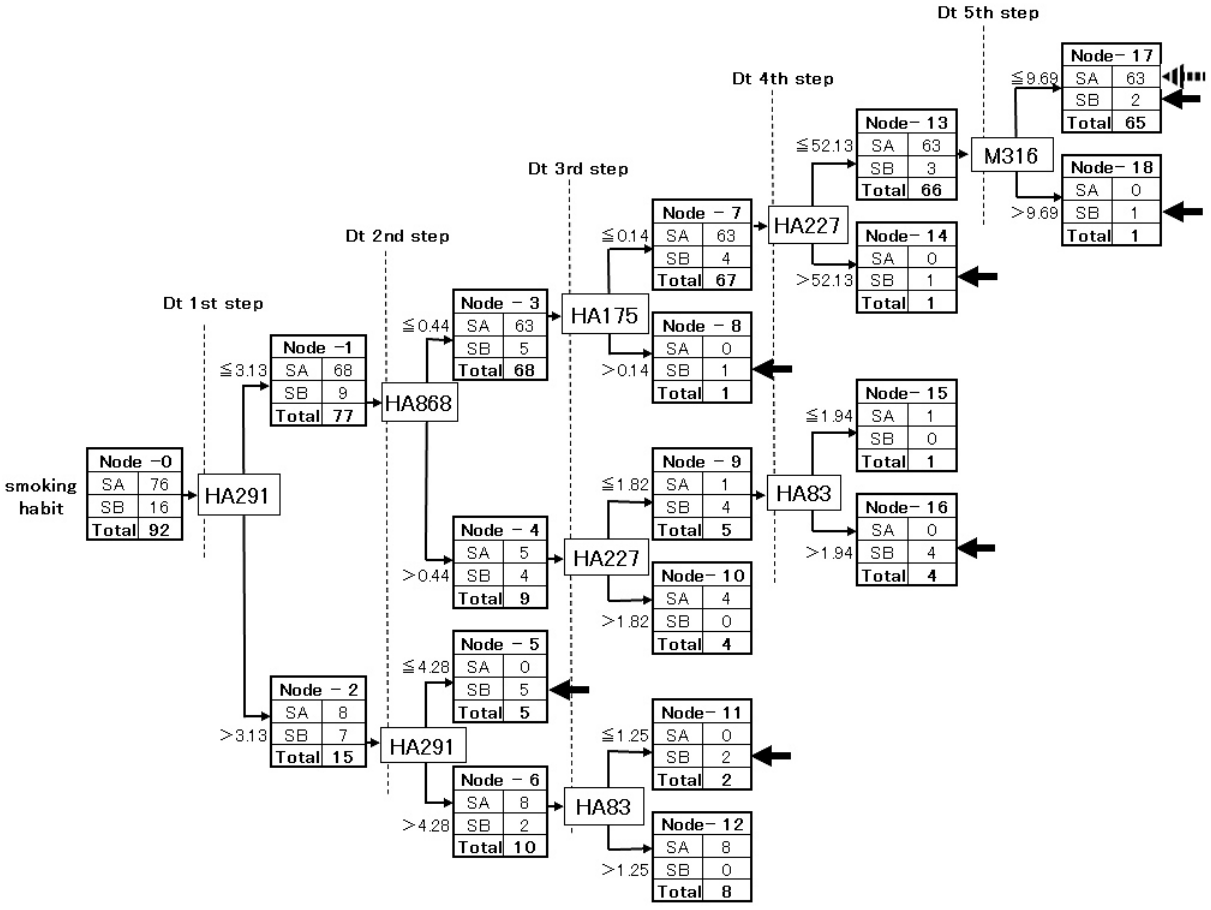


Fig. 1. Decision-tree (Dt) by 2-NP for smoking habit with 53 OTUs.
OTUs: 33HA+20M; marked as * in Table 2; large solid arrows: 7 nodes containing all 16 smokers, 'SB'; large dotted arrow: node of 63 nonsmokers, 'SA'.

for each step of Dt. The default setting of the C&RT system grows a Dt to 5 steps. The balance nodes applied here are for correcting the imbalances in the dataset, which readily develop with higher NPs, and we conform to the specified test criteria and are able to obtain more accurate results. If necessary, balancing is carried out by boosting the occurrence of infrequent values at the time of Dt construction.

RESULTS AND DISCUSSION

Comparison of NPs and restriction enzymes

The Dt produced with 2-NP, as a simple example for understanding and saving space, is shown in Fig. 1, where smoking habit of subjects was explicitly classified into several nodes with certain OTUs. Applying 3-NP and 5-NP as shown in Table 1, the subjects were divided according to the various purposes of DM analysis. Here, the number of partitions was limited to 5 because this was lowest number of subject group (4 as SBH, 5 as SP5 and 6 as DB3) in Table 1. Appendix Fig. A1 shows the results of an actual Dt with 5-NPs for smoking habit, because most of the results with higher NPs required more space to show.

The details of Dt and the pathways to reach the terminal node[3] in these figures clearly show the species of related OTUs, which played a role in dividing the various subject nodes. The Dt also provides quantitative cut off values, namely the 92 men were divided at the 1st step by HA291 for 2 subsets at the left end of Fig. 1 and were subsequently divided. The 1st step was divided at 3.13 by HA291, and the lower 2nd step was recognized as 4.28. The specialty of this Dt was that only 7 OTUs were active out of 53, considering 2 OTUs, i.e., HA291 and HA83 being applied twice, which indicated that the remaining 46 OTUs were neglected in constructing this Dt. In other

words, the 7 OTUs were closely related to subject smoking characteristics, and the other 46 OTUs were recognized as unrelated to smoking. When comparing Fig. 1 with similar results in previous reports [3] (Fig. 1), which had applied 80 OTUs (27·*Bsl*I+33·*Hae*III+20·*Msp*I), HA291 had the same cut-off value at the 1st step, but OTUs later than the 2nd step were different. In addition, 2 wrongly classified subjects were observed in Fig. 1. These were the effects of applied OTU combinations, but we focused only on the accuracies of Dt until the 5th step. With regard to smoking habit, Biedermann et al. [15] recently examined the effects of smoking cessation in 5 subjects, as compared with 10 controls, by T-RFLP and PCA. Their results showed an increase in *Firmicutes* and *Actinobacteria* and a lower proportion of *Bacteroidetes* and *Proteobacteria* at the phylum level.

Similarly in Appendix Fig. A1, 12 OTUs out of 80 were selected to construct the Dt, including HA83. The 92 men were divided at the 1st step by B369 for subset 2 at the left end of Appendix Fig. A1. The 1st step was divided at 1.17 by B369, and the upper 2nd node (Node-1) included 86 subjects, with the lower 2nd node (Node-2) having only 6. These results were the main differences from the former classification methods for HIM, such as clustering, PCA and Pearson correlation coefficient, which considered all OTU data without any selections, and the results inevitably became obscure. Table 2 shows a comparison of the results for 2-NP, 3-NP and 5-NP, with some combination of 7 restriction enzymes for smoking habit. Similarly, with regard to drinking habit, the results shown in Table 3 also show the OTUs for the 1st step and the number of wrongly classified subjects among the 92. The latter indicates the accuracy of evaluation for each set of NP and restriction enzymes, the best value of which is 0.

Tables 2 and 3 showed that accuracy is closely related to the combination of restriction enzymes, not only horizontally in the tables, but also vertically with the same restriction enzymes and different NPs. The best accuracies were recognized as having the same OTUs at the 1st step. Higher NPs gave worse accuracy, with the exception of smoking at 3-NP and 3 combinations of restriction enzymes, i.e., QHh+QM, QHh+QM+QA and QM+QA+QHh (marked as &2, &4 and &5 in the lower middle of Table 2), where only 1 subject was misclassified. Comparing the 2 restriction enzymes group, i.e., between 516f-+27f- and 35f-, the former generally seemed to have slightly better accuracy than the latter. Typical OTUs such as HA291 for heavy smokers [1–3] were only observed at 2-NP for smoking in Table 2, but A47 for drinking was widely obtained at the 1st step in Table 3. Comparing the 2 characteristics, i.e., smoking and drinking, the former was rather easier for classification than the latter, which was previously reported [3] only with 2-NP.

Detailed aspects of better accuracy

Tracing the details of the referred exceptional and better cases marked as &1 to &5 in the lower half of Tables 2 and 4 shows the detailed components of the Dt from the 1st step to the 3rd step. For all 5 cases, the 1st step was the same as with QM134, which indicates exceptional accuracy. The reason why these cases had such results is the structure of the Dt configurations. The 3 cases that the best values, i.e., &2, &4 and &5, revealed that the structure of OTUs was the same until the 2nd step, and that the 3rd step was slightly different. Furthermore, the restriction enzymes in these 3 cases, i.e., QHh+QM, QHh+QM+QA and QM+QA+QHh, had a similar Dt configuration until the 5th step. Even though the selected OTUs were different, the locations of missing nodes were similar at the 4th and 5th steps, which are not shown in Table 4. This suggests that OTUs constructed from individual Dt after the 4th step were replaceable with certain OTUs, and that QA was less workable for this classification than QHh and QM. Finally, OTUs for QM134 played the best role in subject classification for smoking with 3-NP and 3 restriction enzymes (35f-), while QHh178 and QHh574 at the 2nd step played secondary important roles.

Subject features for good classification

Although the values for accuracy were simply compared in Tables 2 and 3, each subject had their own individual OTU features, which were classified with varying levels of ease. In other words, some subjects might have cloudy or boundary features for being classified. Thus, for single utilization of the 4 restriction enzymes, i.e., B, HA, M and A, with 3-NP and 5-NP, the misclassified subjects were individually traced and examined. The number of misclassified subjects, redundantly observed subjects, the rate of wrongly observed subjects among the 92 and the rate of always properly classified subjects were examined and are listed in Table 5. Interestingly, values were recognized in the latter 2 rates, namely, that these 2 rates themselves were easily understood due to the features of OTUs. Furthermore, the intermediary values between these 2 rates, i.e., 100 - 26.1 - 55.4=18.5% for smoking and 29.4% for drinking, were the middle features of the 92 subjects, which were classified properly at either 3-NP or 5-NP. These features were closely combined with smoking or drinking characteristics. Differences and specificities were observed clearly with the values in

Table 2. Comparison of nominal partitions for accuracy of DM and smoking habit

Species of R.Enz.		B	HA	M	A	B+HA	HA+M	HA+A	B+HA+M	HA+M+B	A+M+HA	B+HA+M+A	M+A+B+HA
OTU of Dt-1st step	2-NP	B919	HA291	M133	A87	HA291	HA291	HA291	HA291	HA291	HA291	HA291	HA291
N. of wrongly classified subjects among 92	2-NP	1	3	7	4	0	2*	1	0#1	0	1	1	1
OTU of Dt-1st step	3-NP	B494	HA995	M208	A80	HA995	HA995	HA995	HA995	HA995	HA995	HA995	HA995
N. of wrongly classified subjects among 92	3-NP	10	13	15	11	16	14	13	12#2	12	13	12	12
OTU of Dt-1st step	5-NP	B494	HA995	M208	A238	HA995	HA995	HA995	HA995	B369	HA995	B369	B369
N. of wrongly classified subjects among 92	5-NP	17	20	14	19	21	9	16	21#3	11$	16	17	17

Species of R.Enz.		QHh	QM	QA	QHh+QM	QM+QA	QA+QHh	QHh+QM+QA	QM+QA+QHh
OTU of Dt-1st step	2-NP	QHh601	QM124	QA829	QM124	QM124	QA829	QM124	QM124
N. of wrongly classified subjects among 92	2-NP	7	3	7	2	4	4	4	4
OTU of Dt-1st step	3-NP	QHh601	QM134	QA131	QM134	QM134	QA131	QM134	QM134
N. of wrongly classified subjects among 92	3-NP	20	9&1	15	1&2	5&3	16	1&4	1&5
OTU of Dt-1st step	5-NP	QHh728	QM134	QA131	QM134	QM134	QA131	QM134	QM134
N. of wrongly classified subjects among 92	5-NP	25	9	22	9	8	21	8	8

R.Enz.: primer restriction enzymes; N.: number; NP: nominal partition; N. of wrongly classified subjects: number of misclassified subjects up to 5th step=**accuracy**; Combination of R.Enz. indicated sequences in DM processing; *: detailed Dt is shown in Fig. 1; $: detailed Dt is shown in Appendix Fig. A1; #1 – #3: compared with balance nodes in Table 6; &1 – &5: OTUs obtained up to 3rd step are shown in Table 4; Shadow at '2-NP' indicates that the results have been reported previously [3], but are shown here for comparison to 3-NP and 5-NP.

Table 5; smoking was comparatively easy to classify, and drinking was more ambiguous than smoking, which were recognized with the physiological stresses to the subjects.

Balance node, application of boosted apparent subjects

As shown in Table 1, with large NP, i.e., 5-NP, numbers of component data became small and imbalanced, e.g., SB5, SBH and DB3. If the minimum number vs. the maximum data was less than 15%, the obtained Dt was considered to be less stable, and was easily shifted using a slight change in the minimum component data. To overcome these problems, the DM software provides special methods for applying balance nodes, which boosts and duplicates subjects during Dt construction. Boosting refers to the multiple utilization of minor data components, which allows the total apparent data to be balanced. However, the total number of subjects increases naturally depending on the applied multiple rates for each component. After Dt was constructed, the original data for the 92 subjects without any boosting was applied to the obtained Dt, and the accuracy was normally examined. The detailed mechanisms of boosting and preparing the subjects are shown in Fig. 2 and Table 6 for the cases of smoking marked as #2 and #3 shown in the upper middle part of Table 2. In the left half of Table 6, the original dataset without boosting is shown, and in the middle of the table, multiple rates for boosting and number of apparent subjects are indicated. On the right side, the results examined normally with the original dataset, i.e., 92, are shown. Comparing the results in Table 6, with and without the balance nodes, the advancement improved, particularly at the case of imbalanced datasets, i.e., 5-NP.

Table 3. Comparison of nominal partitions for accuracy of DM and drinking habit

Species of R.Enz		B	HA	M	A	B+HA	B+A	A+B	HA+M	M+A+B	M+A+HA	B+HA+M+A	M+A+B+HA
OTU of Dt-1 st step	2-NP	B657	HA130	M45	A47	B657	A47	A47	M45	A47	A47	A47	A47
N. of wrongly classified subjects among 92	2-NP	5	2	5	3	5	0	0	2	0	0	3	0
OTU of Dt-1 st step (1)	3-NP	B657	HA130	M45	A47	B657	A47	A47	M45	A47	A47	A47	A47
N. of wrongly classified subjects among 92	3-NP	10	7	7	12	8	6	6	7	6	7	7	7
OTU of Dt-1 st step	5-NP	B657	HA194	M45	A47	HA194	A47	A47	HA194	A47	HA194	HA194	HA194
N. of wrongly classified subjects among 92	5-NP	21	21	15	20	21	21	21	21	16	17	17	16

Species of R.Enz		QHh	QM	QA	QHh+QA	QA+QHh	QM+QA	QHh+QM+QA	QM+QA+QHh
OTU of Dt-1 st step	2-NP	QHh601	QM194	QA422	QA422	QA422	QA422	QA422	QA422
N. of wrongly classified subjects among 92	2-NP	12	10	11	5	5	9	10	10
OTU of Dt-1 st step (1)	3-NP	QHh584	QM194	QA422	QA422	QA422	QA422	QA422	QA422
N. of wrongly classified subjects among 92	3-NP	19	19	26	19	19	14	14	14
OTU of Dt-1 st step	5-NP	QHh601	QM194	QA422	QA422	QA422	QA422	QA422	QA422
N. of wrongly classified subjects among 92	5-NP	17	19	27	18	18	14	16	16

All notations are the same as in Table 2.

Table 4. Comparison of detailed Dts with better 3-NP, marked as &1 – &5 in Table 2

Species of R.Enz.		QM	QHh+QM	QM+QA	QHh+QM+QA	QM+QA+QHh
OTU of Dt-1st step	3-NP	QM134	QM134	QM134	QM134	QM134
OTU of Dt-2nd step	3-NP	QM134, QM124	QHh178, QHh574	QA422, QA58	QHh178, QHh574	QHh178, QHh574
OTU of Dt-3rd step	3-NP	QM124, QM194, QM544, QM83	QM194, QM171, QHh361, QHh555	QM134, QM200, QM124, QM134	QA237, QM171, QHh361, QA422	QA237, QM171, QM124, QA422
N. of wrongly classified subjects among 92 up to Dt 5th step	3-NP	9&1	1&2	5&3	1&4	1&5

Smoking habit; all of their Dt 1st step: QM134; &2, &4 and &5 are as described in the text; Other notations are the same as for Tables 1 and 2.

Table 5. Comparison of wrongly classified subjects with NPs and characteristics

characteristics	Nominal Partition	Total of wrongly classified subjects among 92 with single use of R.Enz. in Table 2-3	N. of subjects redundantly observed in left column	N. of subjects wrongly observed with both 3-NP and 5-NP in Table 2-3	N. of subjects who were always properly classified with both 3-NP and 5-NP
Smoking	3-NP	49	12	24 (26.1%)	51 (55.4%)
	5-NP	70	21		
Drinking	3-NP	36	3	27 (29.3%)	38 (41.3%)
	5-NP	77	21		

Single use of 4 R.Enz.: B, HA, M and A; (…): rates among 92 subjects (%).

Namely, the accuracy improved from 77.2% to 89.1%, which shows the advantage of balance nodes. However, in the case of 3-NP, components that are less imbalanced (i.e., 28.1%), the progress was a slight, from 87.0 to 88.0, but the obtained Dt configuration was different.

The obtained OTUs up to the 3rd step were also shown in Table 6 for comparison. The configuration of the Dt indicates the effects of balance nodes. First, the obtained OTUs with the balance node became different from OTUs with normal DM. Second, HA291, which was recognized as the most related OTU to heavy smokers [2, 3], appeared after the application of balance nodes with 5-NP at the 1st step and the lower 2nd step, which are underlined in Table 6. This indicates that the Dt structure after applying balance nodes is similar to the Dt with 2-NP, which is shown in the upper middle part of Table 6. If the OTU dataset has imbalanced components, a more stable Dt configuration and OTUs are obtained with the application of balance nodes. Wide imbalances in a dataset, such as having uneven components, take place occasionally with HIM analyses of large NPs (5 or more).

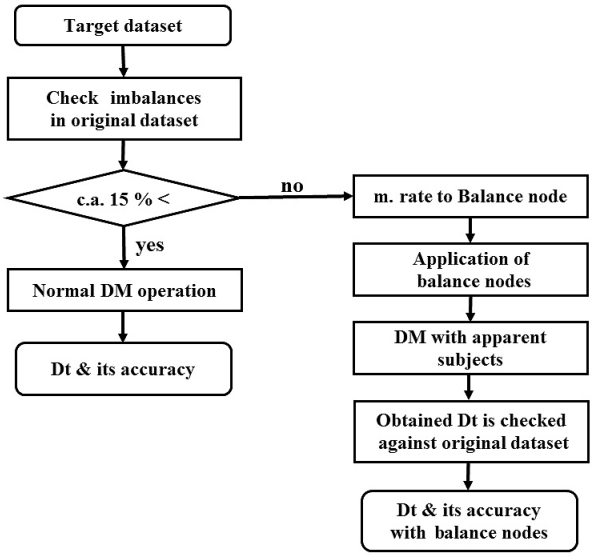


Fig. 2. Flow-chart at utilization of balance nodes.
m.rate: multiple rate for boosting data; Balance nodes are used to correct imbalances in a target dataset. Practical criteria for application were unclear between 10% and 20%. The applied results are shown in Table 6.

Effects of nominal partitions

With regard to the selection of effective restriction enzymes to obtain the accurate DM results, Tables 2 and 3 gave us a good example for smoking and drinking habits. The applications of 2 to 3 combined restriction enzymes revealed better results. Furthermore, in these limited cases, the 516f- + 27f- group exhibited better results than the 35f- group.

Table 6. Application of balance nodes, accuracy and Dt configuration

NP	mark	area / meaning	real subjects, without balance node					with balance node only at Dt constr.				
			N. of real subjects	% to largest data	N. of wrongly classified subjects among 92	accuracy %	Dt 1st step ～ Dt 3rd step	m.rate to balance node	N. of aparent subjects at Dt constr.	N. of wrongly classified subjects among 92	accuracy %	Dt 1st step ～ Dt 3rd step
2-NP	SA	No, non-smoker + non-respondent	76	100	0 #1	100	HA291					
	SB	Yes, smoking now	16	21.1			B469, HA291					
		N. of subjects at Dt constr.	92				B749, B124, –, B919					
3-NP	SAA	non-smoker + non-respondent	57	100	12 #2	87.0	HA995	1.000	57	11	88.0	B919
	SAP	all previous smokers, not now	19	33.3			B494, B494	3.000	57			B369, B940
	SBB	all present smokers	16	28.1			–, B105, B919, –	3.562	57			HA291, –, HA336, B168
		N. of subjects at Dt constr.	92						171			
5-NP	SAA	non-smoker + non-respondent	57	100	21 #3	77.2	HA995	1.0	57	10	89.1	HA291
	SPA	previous smoker, cess. P.* ≧5Y	14	24.6			B494, B494	4.0	56			B919, HA291
	SP 5	previous smoker, cess. P.* <5Y	5	8.8			B110, B106, HA778, –	11.0	55			B469, B754, B124, HA336
	SBG	smoker, 15cigarettes/d. or less	12	21.1				5.0	60			
	SBH	heavy smoker, 16cigs./d. or more	4	7.0				14.0	56			
		N. of subjects at Dt constr.	92						284			

Smoking habit; R.Enz.: 27B+33HA+20M; NP: nominal partition; N.: number; constr.: construction; N. of real subjects: original dataset: 92; m.rate: multiple rate at boosting the data; N. of apparent subjects: boosted subjects for Dt construction; #1 – #3: compared with balance nodes in Table 2; 1st step – 3rd step were connected only vertically, not related to the left end column, i.e., SAA – SP5; "-" at Dt configuration: missing OTU; Shadow at '2-NP' indicates that the results have been reported previously [3], but are shown here for comparison to 3-NP and 5-NP; Basic application schemes for balance nodes are shown in Fig. 2 as a flow-chart.

Focusing on the effects of NPs, which were also observed in Tables 2 and 3, the more NPs were applied, the less accuracy was generally obtained. This provided valuable information about both the selection of related OTUs, and confirmed an effective and stable method for DM processing. Moreover, we obtained in parallel lists of classified subjects, who were situated in the terminal Dt nodes. This means that one is able to classify or discriminate individual subjects, which were visually understood in Appendix Fig. A1 with 5-NP.

The OTU of the 1st step indicated here in the figures and tables was the most related OTU to the assigned characteristics, and the 4th and 5th steps were thought to show some indirect effects, such as local effects in certain areas of OTUs. Focusing only to the OTU of the 1st step, with increasing NP (3-NP or more), less accuracy for DM was observed, as shown in Tables 2 and 3, which is an essential problem of DM processing. However, 5-NP has 4 borders within the dataset and gave worse accuracy when compared to 2-NP, which has only 1 border. The greater the number of NPs, the more the subjects are situated in the border zones of partitions. Therefore, to obtain a clear and simple Dt structure and steady OTU, it is preferable to utilize small NPs (2-NP or 3-NP) than large NPs (5-NP or more). On the other hand, there is a remedy for large NPs and imbalanced components; application of the balance node, as shown in Fig. 2 and Table 6. However, there is a limit on improving accuracy due to the principal mechanisms of a dataset, such as the existence of subjects situated at the border zone.

[1] Decision tree: decision supporting pathway that makes use of a treelike graph, growing left to right.

[2] Gini coefficient: ***g(t)*** is used for quantitative evaluation of group impurity, and is defined at node ***t*** in C&RT, as

$$g(t) = \sum_{j \neq i} p(j|t)p(i|t)$$

, where ***i*** and ***j*** are categories of the target field.

[3] terminal node: tree nodes that do not split further.

ACKNOWLEDGEMENTS

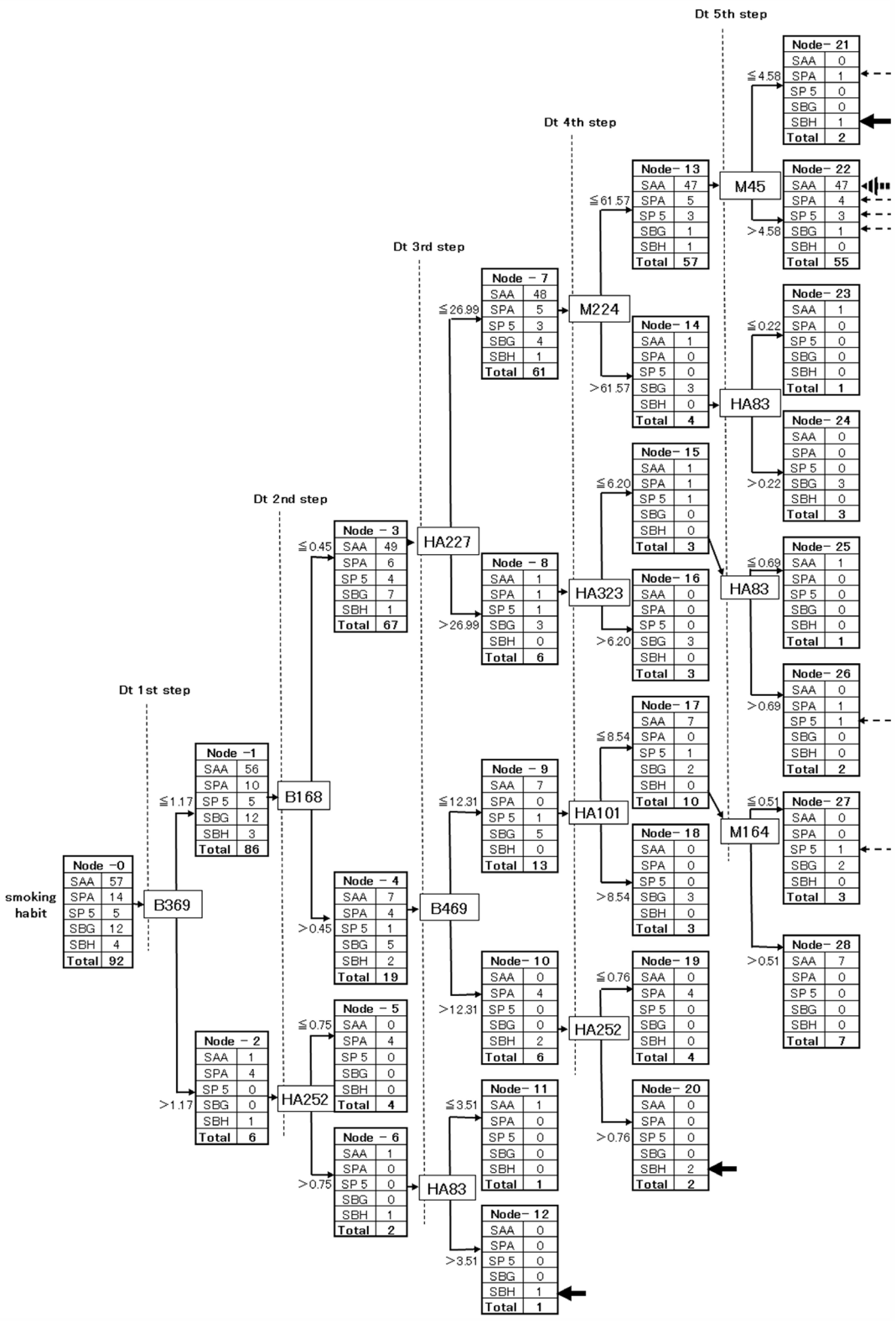
All data processed here were accumulated cooperatively with Jong-Sik Jin, Mutsumi Touyama, Ryoko Kibe, Yoshiki Tanaka, Yoshiko Benno and Yoshimi Benno, Riken, and we appreciate their collaborative efforts. In addition, we are grateful to Dr. Saburou Nakazawa (MD, PhD) the former General Director of the Japanese Society of Gastroenterology, for his insightful encouragement regarding these studies. Finally, we appreciate the steady and long-term financial support provided by Dr. Tatsuo Shimamoto (MD, PhD) and Mr. Toshiaki Teramura, Arteriosclerosis Research Foundation, Japan, for the DM facilities.

REFERENCES

1. Kobayashi T, Fujiwara K. 2013. Identification of heavy smokers through their intestinal microbiota by data mining analysis. Biosci Microb Food Health 32: 77–80. [CrossRef]
2. Kobayashi T, Jin J, Kibe R, Toyama M, Tanaka Y, Benno Y, Fujiwara K, Shimakawa M, Maruo T, Toda T, Matsuda I, Tagami H, Matsumoto M, Seo G, Sato N, Chounan O, Benno Y. 2013. Identification of human intestinal microbiota of 92 men by Data Mining for 5 characteristics, *i.e.* ages, BMI, smoking habit, cessation period of previous smokers and drinking habit. Biosci Microb Food Health 32: 129–137. [CrossRef]
3. Kobayashi T, Fujiwara K. 2013. Comparison of the accuracy and mechanism of Data mining identification of the intestinal microbiota with 7 restriction enzymes. Biosci Microb Food Health 32: 139–148. [CrossRef]
4. Jin JS, Toyama M, Kibe R, Tanaka Y, Benno Y, Kobayashi T, Shimakawa M, Maruo T, Toda T, Matsuda I, Tagami H, Matsumoto M, Seo G, Sato N, Chounan O, Benno Y. 2013. Analysis of the human intestinal microbiota from 92 volunteers after ingestion of identical meals. Benef Microbes 4: 187–193. [Medline] [CrossRef]
5. Andoh A, Kuzuoka H, Tsujikawa T, Nakamura S, Hirai F, Suzuki Y, Matsui T, Fujiyama Y, Matsumoto T. 2012. Multicenter analysis of fecal microbiota profiles in Japanese patients with Crohn's disease. J Gastroenterol 47: 1298–1307. [Medline] [CrossRef]
6. Andoh A, Imaeda H, Aomatsu T, Inatomi O, Banba S, Sasaki M, Saito Y, Tsujikawa T, Fujiyama Y. 2011. Comparison of the fecal microbiota profiles between ulcerative colitis and Crohn's disease using terminal restriction fragment length polymorphism analysis. J Gastroenterol 46: 479–486. [Medline] [CrossRef]
7. Merelli I, Calabria A, Cozzi P, Viti F, Mosca E, Milanesi L. 2013. A gene-centric data mining tool for diseases associated SNP prioritization in GWAS. BMC Bioinformatics (Suppl 1): S9. [Medline]
8. Merelli I, Viti F, Milanesi L. 2012. A galaxy-interacting, integrative database for supporting inflammatory bowel disease high throughput data analysis. BMC Bioinformatics 13:(Suppl 14): S5. [Medline] [CrossRef]
9. Kirschner M, Pujol G, Radu A. 2002. Oligonucleotide microarray data mining: search for age-dependent gene expression. Biochem Biophys Res Commun 298: 772–778. [Medline] [CrossRef]
10. Modlin IM, Gustafsson BI, Drozdov I, Nadler B, Pfragner R, Kidd M. 2009. Principal component analysis, hierarchical clustering, and decision tree assessment of plasma mRNA and hormone levels

as an early detection strategy for small intestinal neuroendocrine (carcinoid) tumors. Ann Surg Oncol 16: 487–498. [Medline] [CrossRef]

11. Matsuki T, Watanabe K, Fujimoto J, Kado Y, Takada T, Matsumoto K, Tanaka R. 2004. Quantitative PCR with 16S rRNA-gene-targeted species-specific primers for analysis of human intestinal bifidobacteria. Appl Environ Microbiol 70: 167–173. [Medline] [CrossRef]
12. Nagashima K, Hisada T, Sato M, Mochizuki J. 2003. Application of new primer-enzyme combinations to terminal restriction fragment length polymorphism profiling of bacterial populations in human feces. Appl Environ Microbiol 69: 1251–1262. [Medline] [CrossRef]
13. Nagashima K, Mochizuki J, Hisada T, Suzuki S, Shimomura K. 2006. Phylogenetic analysis of 16S ribosomal RNA gene sequences from human fecal microbiota and improved utility of terminal restriction fragment length polymorphism profiling. Biosci Microflora 25: 99–107.
14. Matsumoto M, Sakamoto M, Benno Y. 2009. Dynamics of fecal microbiota in hospitalized elderly fed probiotic LKM512 yogurt. Microbiol Immunol 53: 421–432. [Medline] [CrossRef]
15. Biedermann L, Zeitz J, Mwinyi J, Sutter-Minder E, Rehman A, Ott S, Steurer-Stey C, Frei A, Frei P, Scharl M, Loessner M, Vavricka S, Fried M, Schreiber S, Schuppler M, Rogler G. 2013. Smoking cessation induces profound changes in the composition of the intestinal microbiota in humans. PLoS ONE 8: e59260. [Medline] [CrossRef]



Appendix Fig. A1. Decision-tree by 5-NP for smoking habit with 80 OTUs.
OTUs: 33HA+20M+27B; marked as $ in Table 2; large solid arrows: 3 nodes for heavy smokers, 'SBH' in Table 1; large dotted arrow: node for 47 nonsmokers, 'SAA'; thin dotted arrows: misclassified subject(s) until 5th step, of which the total number was 11, marked as $ in Table 2.