Basket Analysis on Meningitis Data

Takayuki IKEDA, Takashi WASHIO and Hiroshi MOTODA
Institute of Scientific and Industrial Research
Osaka University
{ikeda,washio,motoda}@sanken.osaka-u.ac.jp

Abstract

Basket Analysis is the most representative approach in recent study of data mining. However, it cannot be directly applied to the data including numeric data. In this paper, we claim the importance of the selection and the discretization of numeric attributes in the data preprocessing stage for the wider application of Basket Analysis, and propose the algorithm and the performance measures for the selection and the discretization. Through the application of the developed method and Basket Analysis to the meningitis data, their performance is evaluated, and the desirable feature of the performance measure is clarified.

1 Introduction

Basket Analysis is the most representative approach in recent study of data mining[1]. It can extract the knowledge on the co-occurrence of events, i.e., associations among events, embedded in given data, and its highly efficient algorithm called "apriori-algorithm" is applicable to mine the knowledge from a large database. Because of these ability and practicality, Basket Analysis has become to be widely used in the real world applications of data mining in recent years. Based on this background situation, we decided to apply Basket Analysis to the meningitis data given in this discovery challenge[2].

Though Basket Analysis provides a powerful measure to mine the associations of events, it has a drawback to handle data involving numeric information such as the meningitis data . Basket Analysis cannot be directly applied to the numeric data, because it is to mine the associations among discrete events in principle. Thus, the task to select numeric attributes having some associations with the other attributes and to discretize the values of the selected numeric attributes in the data must be introduced to the mining process. An approach is to embed the task into the mining algorithm. This approach is taken in the decision tree based mining such as C4.5[3]. The mining algorithm directly accepts the numeric data, selects attributes relevant to the class, and discretizes the values of the selected numeric attributes while developing the decision tree. This approach can optimize or sub-optimize the selection and discretization based on the intermediate estimations of the numeric value distributions before the final decision tree is obtained if any appropriate criteria for the selection and the discretization is available. However, this approach is not suitable for Basket Analysis since the apriori-algorithm does not include any process to enable the intermediate estimations of the value distributions for massive data.

Accordingly, we developed another approach, which applies the selection and the discretization in the data preprocessing stage. The principle of the approach and the structure of its algorithm can be designed independent of Basket Analysis because the selection and the discretization are conducted in the data preprocessing stage which is not involved in the Basket Analysis process. An issue in the development of the approach is that the selection of the numeric attributes must be performed while taking into account the dependency among the attributes. This is because the association is a representation of the strong dependency among the events characterized by the attributes. For example, when a numeric attribute a does not have any strong dependency with the other numeric attributes b and c, any discretized value ranges of a do not have associations with any discretized value ranges of b and c. Hence, the selection of a does not result associations in Basket Analysis. The second issue is that the points of the discretizations in the value ranges of the numeric attributes must be chosen to appropriately reflect the dependency of the data distribution among multiple attributes. For example, when the numeric attributes b and c have a strong dependency in the value regions $[b_1, b_2]$ of b and $[c_1, c_2]$ of c respectively, the discretization of b to split its region between b_1 and b_2 fragments the dependent region. Then, the number of data representing the association of the values of b and c in each split region becomes less than the number in the original region, and this will affect the result of Basket Analysis. The

Figure 1: Greedy algorithm for selection and discretization.

third issue is that the discretization must have appropriate granularity. If the granularity is too small, again the excessive fragmentation of the dependent region reduces the number of data representing the association of the values in each fragmented region. If the granularity is too large, any clear associations may not be observed. The forth issue is to establish an efficient algorithm for the selection and the discretization under massive data, though this issue is not crucial for the meningitis data since the number of the data is very limited.

The objectives of this paper are as follows.

(1) The development of an approach for the selection and the discretization of numeric attributes addressing the aforementioned three issues.

(2) The application of the approach and Basket Analysis to the meningitis data, the evaluation of their performance and the discussion on the discovered knowledge.

2 Method for Selection and Discretization

2.1 Algorithm

First, the algorithm to select and discretize the numeric attributes we developed is described[4]. The entire flow chart of the algorithm is depicted in Fig.1. Given a performance measure of selection and discretization, this algorithm takes the greedy strategy to conduct the selection and discretization for large database in efficient manner, and thus does not ensure to achieve the optimum selection and discretization. The detail of the performance measure will be described in the latter subsection.

Initially the minimum value in the value range of data for a numeric attribute is set to be a candidate threshold value. Applying this threshold for the discretization, its performance is evalu-

Figure 2: Regions and class distribution.

ated, and it is compared with the performance of the former candidate threshold if it exists. When the performance of the newest candidate threshold is better, the threshold and the performance are recorded. After increasing the threshold value in some small amount, this search process is repeated until all candidate thresholds for every attribute have been evaluated. Once this repetition is finished, the attribute and its threshold value having the optimum performance is selected and used to discretize the data at the threshold of the attribute. After determining a threshold value of a numeric attribute, the search of another attribute and its threshold is repeated until the number of the threshold becomes to a given upper limit. The process of the selection and the discretization is applied only to the numeric attributes in the data, and the discretized attribute is merged with the original categorical attribute. As easily known by the loop structure of the algorithm, this algorithm needs only the computational time in the order of $O(ND)$ where N and D are the number of data and the number of numeric attributes. Because of the linear order of the computational time in terms of the data size, this algorithm can process a large amount of data efficiently, and hence the issue of the efficiency described in the first section is addressed by this algorithm.

2.2 Performance Measure

To obtain the good selection and discretization of numeric attributes, an appropriate performance measure must be applied to the aforementioned algorithm. The most representative performance measure for the selection and the discretization of the numeric attributes is the information entropy evaluated from the class distribution of data on each numeric attribute axis. This measure is used in the selection and the discretization scheme of C4.5[3]. However, this measure cannot take into account the dependency of the class distribution among multiple numeric attributes since the selection and the discretization is applied to each attribute individually. As pointed out in the first section, the measure in our work must address the issue to reflect the dependency among multiple numeric attributes, and thus the approach of C4.5 is not applicable. Accordingly, the performance measure based on the class distribution in each region space generated by the discretization is used in our work. The distribution $S_{i,j}$ of the data having class $j(= 1, 2)$ in a region i is depicted in Fig.2. In this figure, $|S_{i,1}| = 3$ and $|S_{i,2}| = 2$. Based on all distribution $|S_{i,j}|$s, the performance measure such as the information entropy of the entire discretization can be calculated.

However, the information entropy based on the discretized region space does not suggest the appropriate size of the selection and the discretization, i.e., the appropriate number of the attributes and their thresholds. Therefore, it does not address the issue of the granularity described in the first section. This difficulty is solved by introducing the well-known measure named AIC (Akaike's

Information Criterion) which represents Kullback information entropy[5]. Given a discretization pattern HT, AIC under HT is evaluated by following.

$$AIC(HT) = -2\sum_{i=1}^{M} |S_i| \sum_{j=1}^{K_i} \frac{|S_{i,j}|}{|S_i|} \log \frac{|S_{i,j}|}{|S_i|} + 2\alpha, \tag{1}$$

where M is the total number of the discretized regions containing some data, $|S_i|$ the number of the data in the i-th region, K_i the number of the classes appearing in the i-th region, $|S_{i,j}|$ the number of the data having the j-th class in the i-th region, α the number of thresholds in HT. This measure can estimate the discretization having an appropriate granularity, which does not fragment the dependent regions among attributes under the assumption that rectangular parallelepipeds indicated in Fig.2 can asymptotically subsume each dependent region. However, this assumption does not always hold, and hence we sought the other measure to address the granularity issue.

The principle of the performance measure proposed in this work is the Minimum Description Length (MDL) principle[6]. The description length used in this work, $Length(HT)$, is the sum of the description length of a discretization pattern HT, i.e., code book length, and the description length of the class information of the given data under HT, i.e., coding length. The formula of $Length(HT)$ depends on the coding method of the class information. When the combination of the classes appearing in each discretized region is coded, the formula becomes as follows.

$$\begin{aligned} Length(HT) &= -\sum_{i=1(|S_i|\neq 0)}^{M} \sum_{j=1}^{K_i} (|S_{i,j}|+1) \cdot log_2 \frac{|S_{i,j}|+1}{|S_i|+K_i} \\ &+ \sum_{n=1}^{D} \left\{ -(\alpha+1) \cdot log_2 \frac{\alpha+1}{|T_n|+2} - (|T_n|-\alpha+1) \cdot log_2 \frac{|T_n|-\alpha+1}{|T_n|+2} \right\} \\ &+ M \cdot log_2(2^K - 1), \end{aligned} \tag{2}$$

where $|T_n|$ the number of candidate thresholds for the n-th attribute. When the codes are assigned to all classes even if some classes do not appear in a discretized region, the combination of the classes does not have to be coded. Thus, we obtain the following formula.

$$\begin{aligned} &Length(HT) \\ &= -\sum_{i=1(|S_i|\neq 0)}^{M} \sum_{j=1}^{K} (|S_{i,j}|+1) \cdot log_2 \frac{|S_{i,j}|+1}{|S_i|+K} \\ &\quad + \sum_{n=1}^{N} \left\{ -(\alpha+1) \cdot log_2 \frac{\alpha+1}{|T_n|+2} - (|T_n|-\alpha+1) \cdot log_2 \frac{|T_n|-\alpha+1}{|T_n|+2} \right\}. \end{aligned} \tag{3}$$

MDL principle suggests that the selection and the discretization pattern which gives the minimum $Length(HT)$ is the best in terms of the parsimonious description of the data, which is considered to be a general description.

Eq.(1), Eq.(2) and Eq.(3) are applied to the algorithm of Fig.1 as the performance measures respectively. When the value of the measure becomes minimum before the number of the threshold reaches at a given upper limit, the selection and discretization process is stopped, and the attributes and their threshold values are considered to represent the resultant selection and discretization pattern.

3 Application

The meningitis data provided in this JKDD01 Challenge contain 140 case consisting of 21 numeric attributes and 13 categorical attributes. They represent the contents and the results of medical examination, inspection and treatment. The data also include two class attributes, DIAG and DIAG2. DIAG takes 6 values and DIAG2 2 values. We applied our approach to the class DIAG which takes the values of ABSCESS, BACTERIA, BACTERIA(E), BT(E), VIRUS and VIRUS(E). The objectives of the mining analysis are as follows.

(a) To obtain association rules which indicate the conditions of the class of DIAG.

(b) To obtain association rules describing the relations among multiple attributes.

(c) To compare the empirical characteristics of the three performance measures.

(d) To obtain the review on the mining results by a medical expert and to reflect the review comments to the further analyses and discussions.

The procedure of the analysis consists of the following five stages.

1. The selection and discretization method described in the former section is applied to the numeric attributes involved in the meningitis data except the attribute CSF_CELL3 containing some missing values. Thus, the selection and the discretization are applied to the 20 numeric attributes.

2. The categorical attributes except THERAPY2 representing the treatment method, i.e., totally 12 attributes, are combined with the discretized attributes. THERAPY2 was removed because the treatment is a consequence of the diagnosis but not a condition.

3. Each attribute and its value are combined together, and they are transformed into the form of an item. This is because the original data have a table format, but Basket Analysis in the latter stage accepts only the data in an item transaction format.

4. Basket Analysis is applied to the data preprocessed in the former stages.

5. The association rules containing the class attribute DIAG in the head part are collected for the aforementioned object (a). The other rules are separately collected for the object (b).

Every performance measure of Eq.(1), Eq.(2) and Eq.(3) is applied to this mining process.

4 Result and Expert's Evaluation

Table 1 shows the attributes and their threshold values derived through the selection and discretization process by each performance measure. AIC Eq.(1) selects and discretizes only a small number of attributes, but the selection of the attributes has some variety. MDL Eq.(2) selects and discretizes many attributes, but does not show much variety in the attribute selection. MDL Eq.(3) selects and discretizes attributes as many as MDL Eq.(2). This may be because of the similarity of the criterion formulae. However, the discretized attributes show much variety in MDL Eq.(3). The value of Eq.(2) has a high sensitivity to the number of the classes included in each region, because the combination of the classes appearing in each region is coded in the measure, and the number of the combination is exponential to the number of the classes. The sensitivity makes the class distribution in each region to be dominated by a class, and reduces the chance to discretize a region including various classes where the regions dominated by a class are hardly obtained by the discretization. Accordingly, the region already dominated by a class has a tendency to be further selected and discretized on the attributes, which have been already used for the discretization.

By applying the subsequent stages of the mining process to the cases discretized by the three performance measures, we obtained dozens of association rules, and collected the rules containing the class attribute DIAG in the head part and the other rules separately. They are presented to the medical expert who provided this data. The expert suggested that the attribute RISK(Grouped) should be removed from the data, since it is generated by grouping the values of another attribute RISK, and shares some redundant information with RISK. After the removal of RISK(Grouped), the mining process is repeated. The followings are the examples of the sets of association rules derived by the second mining process under some minimum support and minimum confidence levels.

AIC Eq.(1)
Rules, which include the class, attribute DIAG in the head part
Minimum Support=45% and Minimum Confidence=60%

```
{ [LOC_DAT]:- }==>{ [CRP]:under3.1, [Cell_Poly]:under221, [DIAG]:VIRUS }
{ [LOC_DAT]:- }==>{ [Cell_Poly]:under221, [CT_FIND]:normal, [DIAG]:VIRUS }
{ [LOC_DAT]:- }==>{ [Cell_Poly]:under221, [RISK]:n, [DIAG]:VIRUS }
{ [Cell_Poly]:under221, [RISK]:n }==>{ [CRP]:under3.1, [DIAG]:VIRUS }
{ [Cell_Poly]:under221, [RISK]:n }==>{ [DIAG]:VIRUS, [LOC_DAT]:- }
{ [Cell_Poly]:under221, [RISK]:n }==>{ [DIAG]:VIRUS, [FOCAL]:- }
{ [Cell_Poly]:under221, [RISK]:n }==>{ [DIAG]:VIRUS, [C_COURSE]:negative }
{ [Cell_Poly]:under221 }==>{ [CRP]:under3.1, [DIAG]:VIRUS }
{ [Cell_Poly]:under221 }==>{ [DIAG]:VIRUS, [LOC_DAT]:- }
```

Table 1: Discretized attributes and threshold values

-	AIC Eq.(1)		MDL Eq.(2)		MDL Eq.(3)	
m	attributes	thresholds	attributes	thresholds	attributes	thresholds
1	Cell_Poly	2210	Cell_Poly	307.9	Cell_Poly	307.9
2	Cell_Mono	12.0	CSF_CELL	33.8	Cell_Mono	83.3
3	CRP	3.1	Cell_Poly	131.6	FEVER	7.0
4	Cell_Mono	320.0	SEIZURE	1.86	AGE	46.2
5	HEADACHE	3.0	Cell_Poly	6935.2	NAUSEA	4.0
6	CSF_GLU	55.0	CSF_CELL	8468.4	CSF_PRO	44.2
7	BT	37.0	Cell_Poly	657.8	Cell_Poly	55.2
8	-	-	Cell_Poly	2115.7	LOC	0.0013
9	-	-	CSF_CELL	2668.3	HEADACHE	5.0
10	-	-	Cell_Poly	55.2	WBC	6846.9
11	-	-	WBC	19680.8	BT	37.8
12	-	-	CSF_CELL	84.7	CSF_GLU	49.0
13	-	-	CSF_GLU	108.8	Cell_Poly	2.6
14	-	-	Cell_Poly	28.9	Cell_Poly	7.6
15	-	-	Cell_Mono	117.8	Cell_Poly	15.9

```
{ [Cell_Poly]:under221 }==>{ [DIAG]:VIRUS, [FOCAL]:- }
{ [Cell_Poly]:under221 }==>{ [DIAG]:VIRUS, [C_COURSE]:negative }
{ [Cell_Poly]:under221, [C_COURSE]:negative }==>{ [DIAG]:VIRUS, [COURSE(Grouped)]:n }
{ [Cell_Poly]:under221, [COURSE(Grouped)]:n }==>{ [DIAG]:VIRUS, [C_COURSE]:negative }
```

Rules describing the relations among multiple attributes
Minimum Support=60% and Minimum Confidence=95%

```
{ [Cell_Poly]:under221, [ONSET]:ACUTE }==>{ [RISK]:n }
{ [Cell_Poly]:under221, [RISK]:n }==>{ [ONSET]:ACUTE }
{ [Cell_Poly]:under221, [C_COURSE]:negative }==>{ [ONSET]:ACUTE, [COURSE(Grouped)]:n }
{ [Cell_Poly]:under221, [C_COURSE]:negative }==>{ [RISK]:n }
{ [Cell_Poly]:under221, [COURSE(Grouped)]:n }==>{ [ONSET]:ACUTE, [C_COURSE]:negative }
{ [Cell_Poly]:under221, [COURSE(Grouped)]:n }==>{ [RISK]:n }
{ [CRP]:under221, [RISK]:n, [C_COURSE]:negative }==>{ [ONSET]:ACUTE }
{ [CRP]:under221, [RISK]:n, [COURSE(Grouped)]:n }==>{ [ONSET]:ACUTE }
{ [CRP]:under221, [C_COURSE]:negative }==>{ [ONSET]:ACUTE, [RISK]:n }
{ [CRP]:under221, [COURSE(Grouped)]:n }==>{ [ONSET]:ACUTE, [RISK]:n }
```

MDL Eq.(2)
Rules, which include the class, attribute DIAG in the head part
Minimum Support=45% and Minimum Confidence=60%

```
{ [LOC_DAT]:- }==>
   { [DIAG]:VIRUS, [SEIZURE]:under1.85, [WBC]:under19680, [CSF_GLU]:under108 }
{ [FOCAL]:- }==>
   { [DIAG]:VIRUS, [SEIZURE]:under1.85, [WBC]:under19680, [CSF_GLU]:under108 }
{ [CT_FIND]:normal }==>{ [DIAG]:VIRUS, [SEIZURE]:under1.85, [CSF_GLU]:under108 }
{ [SEIZURE]:under1.85, [Cell_Mono]:over17.7 }==>{ [DIAG]:VIRUS, [CSF_GLU]:under108 }
{ [Cell_Mono]:over17.7, [ONSET]:ACUTE, [RISK]:n }==>
   { [DIAG]:VIRUS, [SEIZURE]:under1.85, [CSF_GLU]:under108 }
{ [Cell_Mono]:over17.7, [ONSET]:ACUTE }==>
   { [DIAG]:VIRUS, [SEIZURE]:under1.85, [CSF_GLU]:under108, [RISK]:n }
{ [SEIZURE]:under1.85, [Cell_Mono]:over17.7, [CSF_GLU]:under108, [RISK]:n }==>
   { [DIAG]:VIRUS, [WBC]:under19680 }
{ [SEIZURE]:under1.85, [Cell_Mono]:over17.7, [CSF_GLU]:under108, [RISK]:n }==>
   { [DIAG]:VIRUS, [ONSET]:ACUTE }
{ [SEIZURE]:under1.85, [Cell_Mono]:over17.7, [CSF_GLU]:under108 }==>
   { [DIAG]:VIRUS, [WBC]:under19680, [RISK]:n }
{ [SEIZURE]:under1.85, [Cell_Mono]:over17.7, [CSF_GLU]:under108 }==>
   { [DIAG]:VIRUS, [ONSET]:ACUTE }
{ [SEIZURE]:under1.85, [CSF_GLU]:under108, [ONSET]:ACUTE, [RISK]:n }==>{ [DIAG]:VIRUS }
{ [SEIZURE]:under1.85, [CSF_GLU]:under108, [ONSET]:ACUTE }==>{ [DIAG]:VIRUS }
```

Rules describing the relations among multiple attributes
Minimum Support=80% and Minimum Confidence=95%

```
{ null }==>{ [SEIZURE]:under1.85, [WBC]:under19680 }
{ null }==>{ [SEIZURE]:under1.85, [CSF_GLU]:under108 }
{ null }==>{ [WBC]:under19680, [CSF_GLU]:under108 }
{ [C_COURSE]:negative }==>
   { [SEIZURE]:under1.85, [WBC]:under19680, [CSF_GLU]:under108, [COURSE(Grouped)]:n }
{ [COURSE(Grouped)]:n }==>
   { [SEIZURE]:under1.85, [WBC]:under19680, [CSF_GLU]:under108, [C_COURSE]:negative }
{ [RISK]:n }==>{ [SEIZURE]:under1.85, [WBC]:under19680, [CSF_GLU]:under108 }
```

MDL Eq.(3)
Rules, which include the class, attribute DIAG in the head part
Minimum Support=45% and Minimum Confidence=60%

```
{ [LOC]:under0.0013 }==>{ [DIAG]:VIRUS }
{ [LOC_DAT]:- }==>{ [DIAG]:VIRUS, [CT_FIND]:normal }
{ [LOC_DAT]:- }==>{ [DIAG]:VIRUS, [RISK]:n }
{ [FOCAL]:- }==>{ [DIAG]:VIRUS, [RISK]:n }
{ [CT_FIND]:normal }==>{ [DIAG]:VIRUS  [LOC_DAT]:- }
{ [C_COURSE]:negative }==>{ [DIAG]:VIRUS }
```

Rules describing the relations among multiple attributes
Minimum Support=60% and Minimum Confidence=95%

```
{ [FEVER]:under7.02 }==>{ [ONSET]:ACUTE }
{ [FEVER]:under7.02, [C_COURSE]:negative }==>{ [ONSET]:ACUTE, [COURSE(Grouped)]:n }
{ [FEVER]:under7.02, [COURSE(Grouped)]:n }==>{ [ONSET]:ACUTE, [C_COURSE]:negative }
{ [LOC]:under0.0013, [CT_FIND]:normal }==>{ [LOC_DAT]:- }
{ [LOC_DAT]:-, [CT_FIND]:normal }==>{ [LOC]:under0.0013 }
{ [Cell_Mono]:over83.3 }==>{ [ONSET]:ACUTE }
```

The contents of the rule sets derived by the mining process shows the strong dependency to the performance measures used in the selection and the discretization stage. In case of MDL Eq.(2), many rules have mutually similar body parts and/or head parts. Especially, the rules, which include the class attribute DIAG in the head part are derived from almost identical frequent itemsets except the first three rules. The first three rules are also derived from mutually similar frequent itemsets. The reason why only similar frequent itemsets are derived is because the discretization of MDL Eq.(2) does not show much variety in the attribute selection as pointed out earlier. This property of MDL Eq.(2) limits the variety of frequent itemsets derived in Basket Analysis. The case of AIC Eq.(1) also shows the tendency that many rules have mutually similar body parts and/or head parts. Because the number of the numeric attributes selected and discretized by this performance measure is small as shown in Table 1, the number of the frequent itemsets found in Basket Analysis becomes small. This effect also reduces the variety of frequent itemsets derived in Basket Analysis. On the other hand, the case of MDL Eq.(3) shows more variety of the combinations of the items appearing within a small number of rules. This is because the variety and the number of the attributes selected and discretized by this performance measure were large as mentioned earlier. In addition, the number of itemsets included in each rule is smaller than the other cases. This is also due to the large variety and the large number of the discretized attributes. This property of the selection and the discretization increases the number of the discretized regions in the numeric attribute space, and the number of the data in a region is reduced. This effect makes the size of the frequent itemsets smaller.

5 Discussion and Conclusion

The medical experts evaluated the rule sets derived by using MDL Eq.(3) contains interesting rules more than the other cases. He ordered the performance measures in terms of the ability to derive interesting rules as follows.

$$MDLEq.(3) > MDLEq.(2) > AICEq.(1) \tag{4}$$

This order almost matches with the order of the variety of itemset combinations in the association rules. The rule set describing the relations among various attributes and their thresholds is considered to suggest many potential mechanisms underlying the data.

In this paper, first, we claimed the importance of the selection and the discretization of numeric attributes in the preprocessing stage of data for the wider application of Basket Analysis. Second, we pointed out three issues in the development of the discretization method. Third, the algorithm and the performance measures of the discretization are developed to address the issues. Finally, through the application of the developed method and Basket Analysis to the meningitis data, their performance is evaluated. In conclusion, the performance measure for the discretization of the

values of the numeric attributes strongly affects the results of Basket Analysis. The performance measure which selects variety of the attributes and many threshold values is considered to have a tendency to catch interesting relations among events for domain experts under the analysis of the meningitis data. This insight should be validated through the analysis of the extensive data in the future, and should be reflected to the development of better performance measures for the discretization.

References

[1] R.Agrawal and R.Srikant: Fast algorithms for mining association rules, In *Proceedings of the 20th VLDB Conference,* pp.487-499,1994.

[2] http://wwwada.ar.sanken.osaka-u.ac.jp/pub/washio/jkdd/jkddcfp.html

[3] R. Quinlan: C4.5:Programs for Machine Learning, Morgan Kaufman, 1993.

[4] M. Tsukada, A. Inokuchi, T. Washio and H. Motoda: Comparison of MDLP and AIC for the discretization of numeric attributes, Working notes of knowledge based systems workshop, JSAI, SIG-FAI-9802, pp.45-52, 1998 (in Japanese).

[5] H. Akaike: A new look at the Bayes procedure, In *Biometrika*, Vol.65, pages 53–59, 1978.

[6] U.M. Fayyad and K.B. Irani: Multi-Interval Discretization of Continuous-Valued Nbutes for Classification Learning, Proc. of IJCAI-93: 13th Int. Joint Conf. on Artificial Intelligence, Vol.2, pp.1022-1027, 1993.