EVALUATING LEARNING ALGORITHMS TO SUPPORT HUMAN RULE EVALUATION BASED ON OBJECTIVE RULE EVALUATION INDICES

H Abe[1 *], S Tsumoto[1], M Ohsaki[2], and T Yamaguchi[3]

*[*1]Dept of Medical Informatics, Shimane University, School of Medicine, 89-1 Enya-cho Izumo Shimane, 6938501, Japan*
Email: abe@med.shimane-u.ac.jp*, tsumoto@computer.org
[2] Faculty of Engineering, Doshisha University, 1-3 Tataramiyakodani Kyo-Tanabe Kyoto, 6100321, Japan
Email: mohsaki@mail.doshisha.ac.jp
[3]Faculty of Science and Technology, Keio University, 3-14-1 Hiyoshi Kohoku-ku Yokohama Kanagawa, 2238522, Japan
Email: yamaguti@ae.keio.ac.jp

ABSTRACT

In this paper, we present an evaluation of learning algorithms of a novel rule evaluation support method for post-processing of mined results with rule evaluation models based on objective indices. Post-processing of mined results is one of the key processes in a data mining process. However, it is difficult for human experts to completely evaluate several thousands of rules from a large dataset with noise. To reduce the costs in such rule evaluation task, we have developed a rule evaluation support method with rule evaluation models that learn from a dataset. This dataset comprises objective indices for mined classification rules and evaluation by a human expert for each rule. To evaluate performances of learning algorithms for constructing the rule evaluation models, we have done a case study on the meningitis data mining as an actual problem. Furthermore, we have also evaluated our method with ten rule sets obtained from ten UCI datasets. With regard to these results, we show the availability of our rule evaluation support method for human experts.

Keywords: Data mining, Post-processing, Rule evaluation support, Objective rule evaluation index

1 INTRODUCTION

In recent years, enormous amounts of data are stored on information systems in natural science, social science, and business domains. People have been able to obtain valuable knowledge because of the development of information technology. Data mining techniques combine different kinds of technologies such as database technologies, statistical methods, and machine learning methods, then, utilize data stored on database systems. In particular, if-then rules, which are produced by rule induction algorithms, are considered as one of the highly usable and readable outputs of data mining. For large datasets with hundreds of attributes including noise, the process often produces many thousands of rules. From such a large rule set, it is difficult for human experts to find out valuable knowledge which is rarely included in the rule set.

To support such rule selection, many efforts use objective rule evaluation indices such as recall, precision, and other interestingness measurements (Hilderman, 2001; Tan, 2002; Yao, 1999) (Hereafter, we refer to these indices as "objective indices"). Further, it is difficult to estimate the subjective criterion of a human expert actually using a single objective rule evaluation index because his/her subjective criterion, such as “interestingness” or “importance,” for the purpose is influenced by the amount of prior knowledge and the passage of time.

In this paper, we present an adaptive rule evaluation support method for human experts with rule evaluation models.

This method predicts the experts' criteria based on objective indices by re-using the results of the evaluations by human experts. Section 2 summarizes previous work; while in Section 3, we describe the rule evaluation model construction method based on objective indices. We present a performance comparison of learning algorithms for constructing rule evaluation models in Section 4.

2 RELATED WORK

Many research efforts have been performed to select valuable rules from mined large rule sets based on objective rule evaluation indices. Some of these works suggest indices to discover interesting rules from a large number of rules.

Focusing on interesting rule selection with objective indices, researchers have developed more than forty objective indices based on number of instances, probability, statistical values, information quantities, distance of rules or their attributes, and rule complexity (Hilderman, 2001; Tan, 2002; Yao, 1999). Most of these indices are used to remove meaningless rules rather than to discover ones of real interest to a human expert because they cannot include domain knowledge. In contrast, a dozen of subjective indices estimate how a rule fits with a belief, a bias, or a rule template formulated beforehand by a human expert. Although these subjective indices are useful to some extent in discovering really interesting rules because of their built-in domain knowledge, they depend on the precondition that a human expert is able to clearly formulate his/her interest. Although interestingness indices were verified as to their availabilities on each suggested domain, nobody has validated their applicability on other domains or their characteristics as related to the background of a given dataset.

Ohsaki et al. (Ohsaki, 2004) investigated the relation between objective indexes and real human interests, taking actual data mining results and their evaluations by human experts. In this work, the comparison shows that it is difficult to predict real human interest with a single objective index. Based on this result, we find indications of the possibility of logical combination of the objective indices to predict actual human interest to experts more exactly.

3 RULE EVALUATION SUPPORT WITH RULE EVALUATION MODEL BASED ON OBJECTIVE INDICES

In practical data mining situations, costly rule evaluation procedures are repeatedly done by a human expert. In these situations, useful results of each evaluation such as focused attributes, interesting combinations, and valuable facts are not explicitly used by any rule selection system, but tacitly stored in the human expert. To address this problem, we suggest a method to construct rule evaluation models based on objective rule evaluation indices as a way to describe criteria used explicitly by a human expert, re-using previous human evaluations. Combining this method with the rule visualization interface, we have designed a rule evaluation support tool, which can carry out more exact rule evaluation with explicit rule evaluation models.

3.1 Constructing a Rule Evaluation Model

We considered the process of modeling rule evaluation of human experts as the process of clarifying the relationships between human evaluation and features of inputted if-then rules. Based on this consideration, we decided that the rule evaluation model construction process can be implemented as a learning task. Figure 1 shows the rule evaluation model construction process based on the re-use of human evaluations and objective indices for each mined rule.

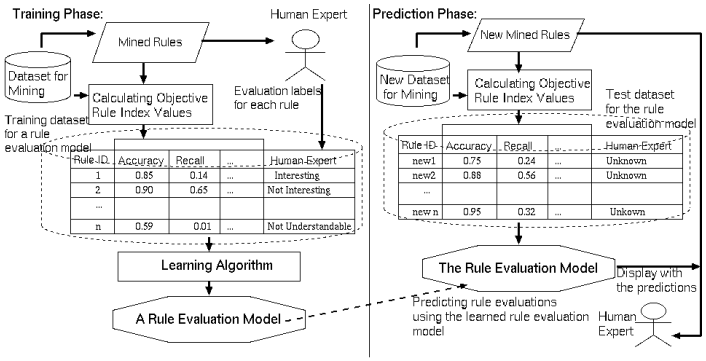


Figure 1. Overview of the construction method of rule evaluation models.

In the training phase, the attributes of a meta-level training data set are obtained by objective indices such as recall, precision, and other rule evaluation values. The human evaluation for each rule is combined as classes of each instance. To obtain this data set, a human expert has to evaluate the whole or a part of the input rules at least once. After obtaining the training data set, its rule evaluation model is constructed by using a learning algorithm.
In the prediction phase, a human expert receives predictions for new rules based on their objective index values. Because rule evaluation models are used for predictions, we need to choose a learning algorithm with high accuracy similar to the current classification problems.

3.2 A Tool to Support Rule Evaluation with Rule Evaluation Models

Our rule evaluation support tool implements interactive support during the time a human expert evaluates rule sets from mining procedure. The first time analyzing a rule set with a totally new task, a human expert sorts them based on some objective indices. Then he/she evaluates the whole or part. On the other hand, if there are previous evaluation results by human experts for the same or similar problem of input rules, possible predictions of the rules can be displayed to a human expert. To obtain the rule set predictions, this tool uses the procedure of the construction of rule evaluation models. Then a human expert corrects the displayed predictions during his/her evaluation. With the corrected evaluations by a human expert, the system rebuilds a rule evaluation model.
With the above procedures, our rule evaluation support tool provides rule evaluation support for a human expert as shown in Figure 2.

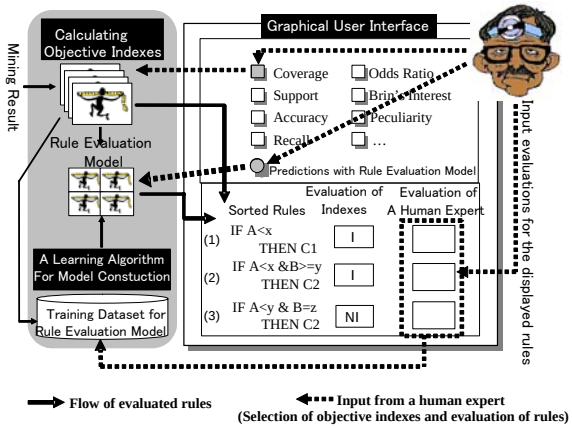


Figure 2. Overview of the visual rule evaluation support tool based on objective indices and rule evaluation models.
A human expert can use this rule evaluation support tool both as both a passive support tool with sorting functions based on objective indices and an active support tool with predictions of rule evaluation models learned from a dataset based on objective indices.

4 PERFORMANCE COMPARISONS OF LEARNING ALGORITHMS FOR RULE MODEL CONSTRUCTION

To predict human evaluation labels of a new rule based on objective indices more accurately, we have to construct a rule evaluation model with a higher predictive accuracy.
In this section, we first present the result of an empirical evaluation with the dataset obtained from the result of a meningitis data mining (Hatazawa, 2000). Then, to confirm the performance of our approach on the other datasets, we evaluated five algorithms on ten rule sets obtained from ten UCI benchmark datasets (Hettich, 1998). Based on the experimental results, we discuss the following: accuracy of rule evaluation models, analysis of learning curves of the learning algorithms, and contents of the learned rule evaluation models.
For evaluating the accuracy of the rule evaluation models, we have compared predictive accuracies on the entire dataset and Leave-One-Out validation. The accuracy of a validation dataset D is calculated with correctly predicted instances:

$Acc(D)=(Correct(D)/|D|)*100$, where $Correct(D)$ is the number of correctly predicted instances, and $|D|$ is the size of the dataset.

The recall of class i on a validation dataset is calculated using correctly predicted instances about the class $Correct(D_i)$ as:

$Recall(D_i)=(Correct(D_i)/|D_i|)*100$, where $|D_i|$ is the size of instances of class i.

Further, the precision of class i is calculated using the size of instances, which are predicted i as:

$Precision(D_i)=(Correct(D_i)/Predicted(D_i))*100$.

With regard to the learning curves, we obtained curves of accuracies of learning algorithms on the entire training dataset to evaluate whether each learning algorithm can perform in the early stage of rule evaluation process. Accuracies of randomly sub-sampled training datasets are averaged with 10 trials on each percentage of the subset.
By observing the elements of the rule evaluation models on the meningitis data mining result, we consider the characteristics of the objective indices that are used in these rule evaluation models.
In order to construct a dataset to learn a rule evaluation model, the values of the objective indices have been calculated for each rule by considering 39 objective indices as shown in Table 1. Thus, each dataset for each rule set has the same number of instances as the rule set. Each instance has 40 attributes including those of the class.

Theory	Index Nabe (**Abbreviation**) [Reference of Literature]
P	Coverage (**Coverage**), Prevalence (**Prevalence**), Precision (**Precision**), Recall (**Recall**), Support (**Support**), Specificity (**Specificity**), Accuracy (**Accuracy**),Lift (**Lift**),Levarage (**Leverage**), Added Value (**Added Value**) [Tan (2002)], Kloesgen's Interestingness (**KI**) [Kloesgen (1996)], Relative Risk (**RR**) [Ali (1997)], Brin's Interest (**BI**) [Brin (1997)], Brin's Conviction (**BC**) [Brin (1997)], Certainty Factor (**CF**) [Tan (2002)], Jaccard Coefficient (**Jaccard**) [Tan (2002)], F-Measure (**F-M**) [Rijsbergen (1979)], Odds Ratio (**OR**) [Tan (2002), Yule's Q (**YuleQ**) [Tan (2002)]
S	Chi-Square Measure for One Quadrant (**Chi-Square M1**)[Goodman (1979)], Chi-Square Measure for Four Quadrant (**Chi-Square M4**)[Goodman (1979)]
I	J-Measure (**J-M**) [Smyth (1991)], K-Measure (**K-M**) [Ohsaki (2004)], Mutual Information (**MI**) [Tan (2002)], Yao and Liu's Interestingness 1 based on one-way support (**YLI1**) [Yao (1999)], Yao and Liu's Interestingness 2 based on two-way support (**YLI2**) [Yao (1999)],
N	Cosine Similarity (**CSI**) [Tan (2002)], Laplace Correction (**LC**) [Tan (2002)], Phi Corfficient (**Phi**) [Tan (2002)], Piatetsky-Shapiro's Interestingness (**PSI**) [Piatetsky-Shapiro (1991)]
D	Gago and Bento's Interestingness (**GBI**) [Gago (1998)], Peculiarity (**Peculiarity**) [Zhong (2003)]

Table 1. Objective rule evaluation indices for classification rules used in this research. **P:** Probability of the antecedent and/or consequent of a rule. **S:** Statistical variable based on P. **I:** Information of the antecedent and/or consequent of a rule. **N:** Number of instances included in the antecedent and/or consequent of a rule. **D:** Distance of a rule from the others based on rule attributes.

We applied five learning algorithms to these datasets to compare their performances as a rule evaluation model construction method. We used the following learning algorithms from Weka (Witten, 2000): C4.5 decision tree learner (Quinlan, 1993) called J4.8, neural network learner with back propagation (BPNN) (Hinton, 1986), support vector machines (SVM) (Platt, 1999), classification via linear regressions (CLR) (Frank, 1998), and OneR (Holte, 1993).

4.1 Constructing Rule Evaluation Models for an Actual Datamining Result

In this case study, we have considered 244 rules, which are mined from six datasets about six types of diagnostic problems as shown in Table 2. In these datasets, appearances of meningitis patients were considered as attributes and the diagnosis of each patient as a class. Each rule set was mined with its proper rule induction algorithm composed by a constructive meta-learning system called CAMLET (Hatazawa, 2000). For each rule, we labeled three evaluations (I: Interesting, NI: Not-Interesting, NU: Not-Understandable) according to evaluation comments provided by a medical expert.

Dataset	#Att.	#Class	#Mined rules	#'I' rules	#'NI' rules	#'NU' rules
Diag	29	6	53	15	38	0
C_Course	40	12	22	3	18	1
Culture+diag	31	12	57	7	48	2
Diag2	29	2	35	8	27	0
Course	40	2	53	12	38	3
Cult_find	31	2	24	3	18	3
TOTAL	-	-	244	48	187	9

Table 2. Description of the meningitis datasets and the results of the datamining

4.1.1 Comparison of Classification Performances

In this section, we present the result of the accuracy comparison over the entire dataset, recall of each class label, and their precision. Because Leave-One-Out holds just one test instance and the remaining as the training dataset repeatedly for each instance of a given dataset, we can evaluate the performance of a learning algorithm to a new dataset without any ambiguity.
The results of the performances of the five learning algorithms to the entire training dataset and the results of Leave-One-Out are also shown in Table 3. All the Accuracies, Recalls of I and NI, and Precisions of I and NI are higher than those of the predicting default labels.

	Over the entire training dataset							Leave-One-Out						
	Acc.	Recall			Precision			Acc.	Recall			Precision		
		I	NI	NU	I	NI	NU		I	NI	NU	I	NI	NU
J4.8	85.7	41.7	97.9	66.7	80.0	86.3	85.7	79.1	29.2	95.7	0.0	63.6	82.5	0.0
BPNN	86.9	81.3	89.8	55.6	65.0	94.9	71.4	77.5	39.6	90.9	0.0	50.0	85.9	0.0
SVM	81.6	35.4	97.3	0.0	68.0	83.5	0.0	81.6	35.4	97.3	0.0	68.0	83.5	0.0
CLR	82.8	41.7	97.3	0.0	71.4	84.3	0.0	80.3	35.4	95.7	0.0	60.7	82.9	0.0
OneR	82.0	56.3	92.5	0.0	57.4	87.8	0.0	75.8	27.1	92.0	0.0	37.1	82.3	0.0

Table 3. Accuracies (%), Recalls (%), and Precisions (%) of the five learning algorithms.

As compared to the accuracy of OneR, the other learning algorithms achieve equal or higher performances using combinations of multiple objective indices than by sorting with a single objective index. With regard to the Recall values over class I, BPNN has achieved the highest performance. The other algorithms exhibit lower performance than that of OneR because they tend to be learned classification patterns for the major class NI.
The accuracy of Leave-One-Out demonstrates the robustness of each learning algorithm. The Accuracy (%) of these learning algorithms ranges from 75.8% to 81.9%. However, these learning algorithms have not been able to classify the instances of class NU because it is difficult to predict a minor class label in this dataset.

4.1.2 Learning Curves of the Learning Algorithms

Since the rule evaluation model construction method requires the mined rules to be evaluated by a human expert, we have investigated learning curves of each learning algorithm to estimate a minimum training subset to construct a valid rule evaluation model. The table in the upper portion of Figure 3 shows the accuracies to the entire training dataset with each subset of training dataset. The percentage of achievements of each learning algorithm compared with their accuracy over the whole dataset is shown in the lower section of Figure 3.
As observed in these results, SVM and CLR, which use hyper-planes, obtained an achievement ratio greater than 95% using less than 10% of training subset. Although a decision tree learner and BPNN could determine better classifiers to the entire dataset than the hyper-plane learners, they need more training instances to determine accurate classifiers.

%training sample	10	20	30	40	50	60	70	80	90	100
J4.8	73.4	74.7	79.8	78.6	72.8	83.2	83.7	84.5	85.7	85.7
BPNN	74.8	78.1	80.6	81.1	82.7	83.7	85.3	86.1	87.2	86.9
SMO	78.1	78.6	79.8	79.8	79.8	80.0	79.9	80.2	80.4	81.6
CLR	76.6	78.5	80.3	80.2	80.3	80.7	80.9	81.4	81.0	82.8
OneR	75.2	73.4	77.5	78.0	77.7	77.5	79.0	77.8	78.9	82.4

%training sample	10	20	30	40	50	60	70	80	90	100
J4.8	82.7	85.3	92.8	88.7	93.2	92.7	93.2	92.9	94.0	97.9
BPNN	84.6	86.6	90.4	90.2	92.2	91.9	92.7	93.9	94.2	89.8
SMO	93.3	92.7	96.8	96.1	95.9	95.8	96.3	96.0	96.3	97.3
CLR	88.3	89.6	94.4	94.0	94.3	94.1	94.1	94.2	94.3	97.3
OneR	88.4	84.0	92.4	91.4	92.0	92.3	93.4	92.7	92.1	96.3

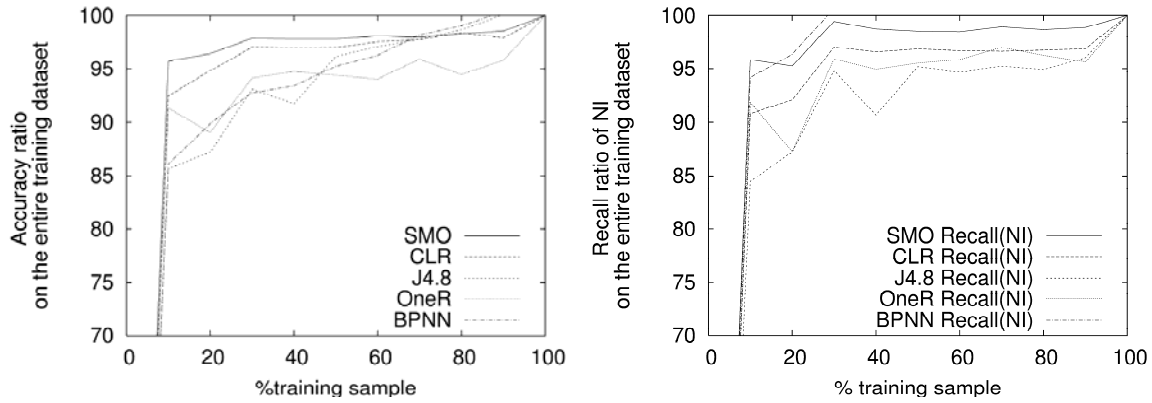


Figure 3. Learning curves of Accuracies (%) on the learning algorithms over subsampled training dataset: The left table shows accuracies (%) of each training dataset to the entire dataset. The left graph shows their achievement ratios (%). The right table shows recalls (%) and the graph shows their achievement ratios (%).

In order to eliminate known ordinary knowledge from a large rule set, the non-interesting rules need to be classified correctly. The right upper table in Figure 3 shows percentage of Recalls on NI. The right lower chart in Figure 3 also shows the percentage of achievements of Recall of NI and compares it with the Recall of NI of the entire training dataset. From this result, we can eliminate the NI rules with rule evaluation models from SVM and BPNN although only 10% of rule evaluations are conducted by a human expert. This fact is guaranteed with no less than 80% precision for all learning algorithms.

4.1.3 Rule Evaluation Models on the Actual Datamining Result Dataset

In this section, we present rule evaluation models for the entire dataset learned using OneR, J4.8 and CLR. This is because they are represented as explicit models such as a rule set, a decision tree, and linear model set.

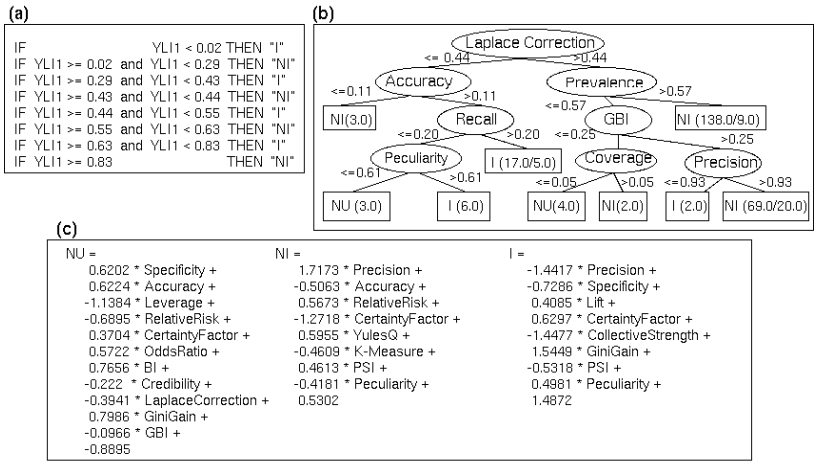


Figure 4. Learned models for the meningitis data mining result dataset.

Figure 4 shows rule evaluation models for the actual data mining results: The rule set of OneR is shown in Figure 4 (a), Figure 4 (b) shows the decision tree learned with J4.8, and Figure 4 (c) shows linear models used to classify each

class.

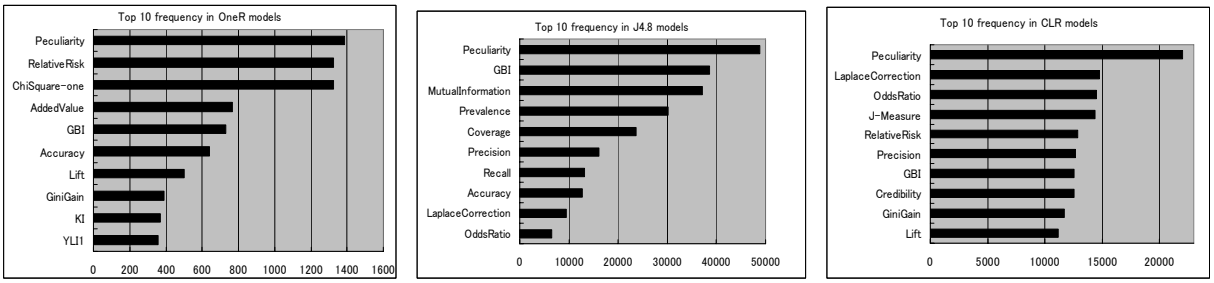
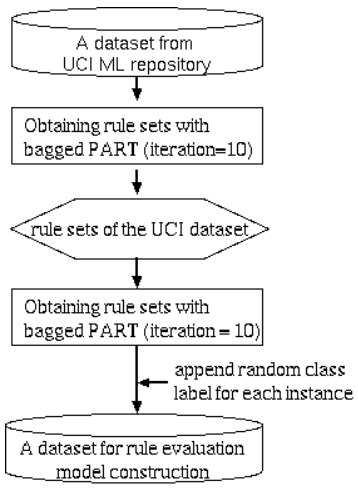


Figure 5. Top 10 frequencies of the indices used by the models of each learning algorithm with 10000 bootstrap samples of the meningitis datamining result dataset and executions.

As shown in Figure 4 and Figure 5, the indices used in the learned rule evaluation models are not only taken from a group of indices that increase with correctness of a rule but also from different groups of indices. YLI1, Laplace Correction, Accuracy, Precision, Recall, Coverage, PSI and, Gini Gain are indices which are formally used on the models. The latter indices are GBI and Peculiarity, which sum up the difference in antecedents between one rule and the other rules in the same rule set. This corresponds to the comments provided by the human expert who said that he evaluated these rules not only according to their correctness but also to their interestingness based on his expertise

4.2 Constructing Rule Evaluation Models on Artificial Evaluation Labels

We have also evaluated our rule evaluation model construction method using rule sets obtained from five datasets of the UCI machine learning repository to confirm the lower limit performances on probabilistic class distributions.
We selected the following ten datasets: Anneal, Audiology, Autos, Balance-scale, Breast-cancer, Breast-w, Colic, Credit-a, Waveform, and Letter. With these datasets, we obtained rule sets with bagged PART, which repeatedly executes PART (Frank, 1998) to the bootstrapped training subsample datasets. For these rule sets, we calculated 39 objective indices as attributes of each rule. With regard to the classes of these datasets, we used three class distributions with multi-nomial distribution. Table 4 shows the process flow diagram for obtaining these datasets and their description with three different class distributions. The class distribution for "Distribution I" is $P=(0.3,0.35,0.35)$ where p_i is the probability of class i. Thus, the number of class i instances in each dataset D_j become p_iD_j. Similarly, the probability vector of "Distribution II" is $P=(0.3,0.5,0.2)$ and that of "Distribution III" is $P=(0.3,0.65,0.05)$.



	#Mined Rules	#Class labels			%Def. class
		L1	L2	L3	
Distribution I		(0.30)	(0.35)	(0.35)	
Anneal	95	33	39	23	41.1
Audiology	149	44	58	47	38.9
Autos	141	30	48	63	44.7
Balance-scale	281	76	102	103	36.7
Breast-cancer	122	41	34	47	38.5
Breast-w	79	29	26	24	36.7
Colic	61	19	18	24	39.3
Credit-a	230	78	73	79	34.3
Waveform	518	146	192	180	37.1
Letter	6340	1908	2163	2269	35.8
Distribution II		(0.30)	(0.50)	(0.20)	
Anneal	95	26	47	22	49.5
Audiology	149	44	69	36	46.3
Autos	141	40	72	29	51.1
Balance-scale	281	76	140	65	49.8
Breast-cancer	122	40	62	20	50.8
Breast-w	79	29	36	14	45.6
Colic	61	19	35	7	57.4
Credit-a	230	78	110	42	47.8
Waveform	824	240	436	148	52.9
Letter	6340	1890	3198	1252	50.4
Distribution III		(0.30)	(0.65)	(0.05)	
Anneal	95	26	63	6	66.3
Audiology	149	49	91	9	61.1
Autos	141	41	95	5	67.4
Balance-scale	281	90	178	13	63.3
Breast-cancer	122	42	78	2	63.9
Breast-w	79	22	55	2	69.6
Colic	61	22	36	3	59.0
Credit-a	230	69	150	11	65.2
Waveform	824	246	529	49	64.2
Letter	6340	1947	4062	331	64.1

Table 4. Flow diagram to obtain datasets and the datasets of the rule sets learned from the UCI benchmark datasets.

4.2.1 Accuracy Comparison on Classification Performances

In the above mentioned datasets, we have used the five learning algorithms to estimate if their classification results reach or exceed the accuracies of that of just predicting each default class. The left table of Table 5 shows the accuracies of the five learning algorithms applied to each class distribution of the three datasets. As shown in Table 5, J4.8 and BPNN always perform better for just predicting a default class. However, their performances suffer from probabilistic class distributions for larger datasets such as Waveform and Letter.

	Distribution I				
	J4.8	BPNN	SVM	CLR	OneR
Anneal	74.7	71.6	47.4	56.8	55.8
Audiology	47.0	51.7	40.3	45.6	52.3
Autos	66.7	63.8	46.8	46.1	56.0
Balance-scale	58.0	59.4	39.5	43.4	53.0
Breast-cancer	55.7	61.5	40.2	50.8	59.0
Breast-w	86.1	91.1	38.0	46.8	54.4
Colic	91.8	82.0	42.6	60.7	55.7
Credit-a	57.4	48.7	35.7	39.1	54.8
Waveform	46.5	46.4	37.6	39.8	54.9
Letter	36.8	36.4	30.1	36.6	52.1
	Distribution II				
	J4.8	BPNN	SVM	CLR	OneR
Anneal	68.4	66.3	56.8	60.0	56.8
Audiology	60.4	61.1	43.6	55.0	56.4
Autos	63.1	64.5	52.5	53.2	57.4
Balance-scale	61.6	57.7	49.8	55.2	58.0
Breast-cancer	68.0	70.5	47.5	58.2	59.8
Breast-w	89.9	93.7	49.4	58.2	62.0
Colic	77.0	78.7	57.4	62.3	67.2
Credit-a	61.3	59.1	41.3	52.6	56.1
Waveform	61.2	57.8	52.9	53.0	59.7
Letter	51.0	51.0	50.4	50.4	57.0
	Distribution III				
	J4.8	BPNN	SVM	CLR	OneR
Anneal	74.7	70.5	67.4	70.5	73.7
Audiology	65.8	67.8	63.8	64.4	67.1
Autos	85.1	73.8	68.1	70.2	73.8
Balance-scale	70.5	69.8	64.8	65.8	69.8
Breast-cancer	71.3	77.0	66.4	65.6	77.9
Breast-w	74.7	86.1	73.4	68.4	74.7
Colic	70.5	77.0	65.6	60.7	73.8
Credit-a	70.9	70.0	65.2	65.2	71.3
Waveform	74.4	69.3	64.2	64.2	69.3
Letter	64.1	64.3	64.1	64.1	68.3

	Distribution I				
	J4.8	BPNN	SVM	CLR	OneR
Anneal	20	14	17	29	29
Audiology	21	18	65	64	41
Autos	38	28	76	77	70
Balance-scale	12	14	15	15	32
Breast-cancer	16	17	22	41	22
Breast-w	7	10	10	18	14
Colic	8	8	9	22	14
Credit-a	9	12	16	30	28
Waveform	60	52	46	355	152
Letter	189	217	–	955	305
	Distribution II				
	J4.8	BPNN	SVM	CLR	OneR
Anneal	29	20	16	42	46
Audiology	36	45	–	61	67
Autos	49	39	49	123	88
Balance-scale	81	84	69	221	168
Breast-cancer	31	28	102	40	46
Breast-w	14	11	23	30	26
Colic	24	20	36	42	36
Credit-a	51	74	–	134	109
Waveform	251	355	763	–	533
Letter	897	>1000	451	–	>1000
	Distribution III				
	J4.8	BPNN	SVM	CLR	OneR
Anneal	54	58	64	76	–
Audiology	64	73	45	76	107
Autos	66	102	84	121	98
Balance-scale	118	103	133	162	156
Breast-cancer	50	31	80	92	80
Breast-w	44	36	31	48	71
Colic	28	24	46	30	42
Credit-a	118	159	–	–	173
Waveform	329	425	191	–	601
Letter	>1000	>1000	998	>1000	>1000

Table 5. Accuracies (%) on entire training datasets labeled with three different distributions (left table). Number of minimum training sub samples for outperforming the Accuracy (%) of default class (right table).

4.2.2 Evaluation of Learning Curves

Similar to the evaluations of the learning curves on the meningitis rule set, we have estimated minimum training subsets for a valid model, which works better for just predicting a default class. The right table in Table 5 shows the sizes of the minimum training subsets, which can help construct more accurate rule evaluation models than percentages of a default class formed by each learning algorithm. With smaller datasets, these learning algorithms have been able to construct valid models with less than 25% of the given training datasets. However, for larger datasets such as Waveform and Letter, they need more training subsets to construct valid models because their performance with the entire training dataset fall to the percentages of default class of each dataset as shown in the left table in Table 5.

5 CONCLUSION

In this paper, we have described the evaluation of five learning algorithms for a rule evaluation support method with rule evaluation models to predict evaluations for an if-then rule based on objective indices by re-using evaluations by a human expert. Based on the performance comparison of the five learning algorithms, rule evaluation models have achieved higher accuracies for just predicting each default class. Considering the difference between the actual evaluation labeling and the artificial evaluation labeling, it is shown that the evaluation of the medical expert considered particular relations between an antecedent and a class another antecedent in each rule. By using these

learning algorithms for estimating the robustness of a new rule with Leave-One-Out, we have achieved accuracy greater than 75.8%. By evaluating learning curves, SVM and CLR were observed to have achieved an achievement ratio greater than 95% using less than 10% of the subset of the training dataset, which includes certain human evaluations. These results indicate the availability of this rule evaluation support method for a human expert.
In the future, we will introduce a selection method of learning algorithms to construct a proper rule evaluation model according to each situation. We also apply this rule evaluation support method to estimate other data mining results such as decision tree and rule set and combine them with objective indices, which evaluate all the mining results.

6 REFERENCES

Ali, K., Manganaris,S.,& Srikant, R. (1997) Partial Classification Using Association Rules. *Proc. of Int. Conf. on Knowledge Discovery and Data Mining KDD-1997,* 115—118.

Brin, S., Motwani, R., Ullman, J., & Tsur, S. (1997) Dynamic itemset counting and implication rules for market basket data. *Proc. of ACM SIGMOD Int. Conf. on Management of Data,* 255—264.

Frank, E., Wang, Y., Inglis, S., Holmes, G, & Witten, I. H. (1998) Using model trees for classification. *Machine Learning. 32(1),* 63—76.

Frank, E. &Witten, I. H. (1998) Generating accurate rule sets without global optimization. *Proc. of the Fifteenth International Conference on Machine Learning,* 144—151.

Gago, P., & Bento, C. (1998) A Metric for Selection of the Most Promising Rules. *Proc. of Euro. Conf. on the Principles of Data Mining and Knowledge Discovery PKDD-1998,* 19—27.

Goodman, L. A. & Kruskal, W. H. (1979) Measures of association for cross classifications. *Springer Series in Statistics, 1,* Springer-Verlag.

Gray, B. & Orlowska, M. E. (1998) CCAIIA: Clustering Categorical Attributes into Interesting Association Rules. *Proc. of Pacific-Asia Conf. on Knowledge Discovery and Data Mining PAKDD-1998,* 132—143.

Hamilton, H. J., Shan, N., & Ziarko, W. (1997) Machine Learning of Credible Classifications.
Proc. of Australian Conf. on Artificial Intelligence AI-1997, 330—339.

Hatazawa, H., Negishi, N., Suyama, A., Tsumoto, S., & Yamaguchi, T. (2000) Knowledge Discovery Support from a Meningoencephalitis Database Using an Automatic Composition Tool for Inductive Applications. *Proc. of KDD Challenge 2000 in conjunction with PAKDD2000,* 28—33.

Hettich, S., Blake, C. L., & Merz, C. J. (1998) UCI Repository of machine learning databases, Irvine, CA: University of California, Department of Information and Computer Science.

Hilderman, R. J. & Hamilton, H. J. (2001) *Knowledge Discovery and Measure of Interest*. Kluwe Academic Publishers.

Hinton, G. E. (1986) Learning distributed representations of concepts. *Proc. of 8th Annual Conference of the*

Cognitive Science Society, Amherest, MA. REprinted in R.G.M.Morris (ed.).

Holte, R. C. (1993) Very simple classification rules perform well on most commonly used datasets. *Machine Learning, 11,* 63—91.

Kloesgen, W. (1996) Explora: A Multipattern and Multistrategy Discovery Assistant. in Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy R. (Eds.): *Advances in Knowledge Discovery and Data Mining.* AAAI/MIT Press, California, 249—271.

Ohsaki, M., Kitaguchi, S., Kume, S., Yokoi, H., & Yamaguchi, T. (2004) Evaluation of Rule Interestingness Measures with a Clinical Dataset on Hepatitis. *Proc. of ECML/PKDD 2004, LNAI 3202,* 362—373.

Piatetsky-Shapiro, G. (1991) Discovery, Analysis and Presentation of Strong Rules. in Piatetsky-Shapiro, G., Frawley, W. J. (eds.): *Knowledge Discovery in Databases*. AAAI/MIT Press, 229—248.

Platt, J. (1999) Fast Training of Support Vector Machines using Sequential Minimal Optimization. In: B. Schoelkopf, C. Burges, and A. Smola (eds.): *Advances in Kernel Methods - Support Vector Learning*, MIT Press, 185—208.

Quinlan, R.: C4.5: *Programs for Machine Learning*. Morgan Kaufmann Publishers.

Rijsbergen, C. (1979) *Information Retrieval*, Chapter 7, Butterworths, London, [http://www.dcs.gla.ac.uk/Keith/Chapter.7/Ch.7.html].

Smyth, P., Goodman, R. M. (1991) Rule Induction using Information Theory. in Piatetsky-Shapiro, G., Frawley, W. J. (eds.): *Knowledge Discovery in Databases*. AAAI/MIT Press, 159—176.

Tan, P. N., Kumar V., & Srivastava, J. (2002) Selecting the Right Interestingness Measure for Association Patterns. *Proc. of Int. Conf. on Knowledge Discovery and Data Mining KDD-2002,* 32—41.

Witten, I. H. & Frank, E. (2000) *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufmann.

Yao, Y. Y. & Zhong, N. (1999) An Analysis of Quantitative Measures Associated with Rules. *Proc. of Pacific-Asia Conf. on Knowledge Discovery and Data Mining PAKDD-1999*, 479—488.

Zhong, N., Yao, Y. Y., & Ohshima, M. (2003) Peculiarity Oriented Multi-Database Mining. *IEEE Trans. on Knowledge and Data Engineering, 15(4*), 952—960.