Evaluating Learning Algorithms Composed by a Constructive Meta-Learning Scheme for a Rule Evaluation Support Method Based on Objective Indices

Hidenao Abe[1], Shusaku Tsumoto[1], Miho Ohsaki[2], Hideto Yokoi[3] and Takahira Yamaguchi[4]

[1] Department of Medical Informatics, Shimane University, School of Medicine
89-1 Enya-cho, Izumo, Shimane 693-8501, Japan
abe@med.shimane-u.ac.jp, tsumoto@computer.org
[2] Faculty of Engineering, Doshisha University
mohsaki@mail.doshisha.ac.jp
[3] Department of Medical Informatics, Kagawa University Hospital
yokoi@med.kagawa-u.ac.jp
[4] Faculty of Science and Technology, Keio University
yamaguti@ae.keio.ac.jp

Abstract. In this paper, we present evaluations of learning algorithms for a novel rule evaluation support method in data mining post-processing, which is one of the key processes in a data mining process. It is difficult for human experts to evaluate many thousands of rules from a large dataset with noises completely. To reduce the costs of rule evaluation task, we have developed the rule evaluation support method with rule evaluation models, which are learned from a dataset consisted of objective indices and evaluations of a human expert for each rule. To enhance adaptability of rule evaluation models, we introduced a constructive meta-learning system to choose proper learning algorithms for constructing them. Then, we have done a case study on the meningitis data mining result, the hepatitis data mining results and rule sets from the eight UCI datasets.

1 Introduction

In recent years, with huge data stored on information systems in natural science, social science and business domains, developing information technologies, people hope to find out valuable knowledge suited for their purposes. Besides, data mining techniques have been widely known as a process for utilizing stored data on database systems, combining different kinds of technologies such as database technologies, statistical methods and machine learning methods. In particular, if-then rules are discussed as one of highly usable and readable output of data mining. However, to large dataset with hundreds attributes including noise, the process often obtains many thousands of rules, which rarely include valuable rules for a human expert.

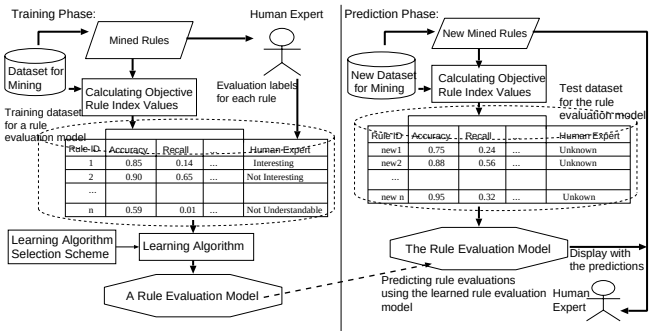


Fig. 1. Overview of the construction method of rule evaluation models.

To support such a rule selection, many efforts have done using objective rule evaluation indices such as recall, precision, and other interestingness measurements [16, 30, 33] (we call them "objective indices" later). However, it is also difficult to estimate a criterion of a human expert with single objective rule evaluation index, because his/her subjective criterion such as interestingness and importance for his/her purpose is influenced by the amount of his/her knowledge and/or a passage of time.

To above issues, we have been developed an adaptive rule evaluation support method for human experts with rule evaluation models, which predict experts' criteria based on objective indices, re-using results of evaluations of human experts. In Section 2, we describe the rule evaluation model construction method based on objective indices. Since our method needs more accurate rule evaluation model to support a human expert more exactly, we present a performance comparison of learning algorithms for constructing rule evaluation models in Section 3. With the results of the comparison, we present the availability of learning algorithms from constructive meta-learning system[1] for our rule evaluation model construction approach.

2 Rule Evaluation Support with Rule Evaluation Model based on Objective Indices

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

At the training phase, attributes of a meta-level training data set is obtained by objective indices such as recall, precision and other rule evaluation values. The human evaluations for each rule are joined as class of each instance. To obtain this data set, a human expert has to evaluate the whole or part of input rules at least once. After obtaining the training data set, its rule evaluation

model is constructed by a learning algorithm. At the prediction phase, a human expert receives predictions for new rules based on their values of the objective indices. Since the task of rule evaluation models is a prediction, we need to choose a learning algorithm with higher accuracy as same as current classification problems.

3 Performance Comparisons of Learning Algorithms for Rule Model Construction

To predict human evaluation labels of a new rule based on objective indices more exactly, we have to construct a rule evaluation model, which has higher predictive accuracy.

In this section, we firstly present the results of an empirical evaluation with the dataset from the result of a meningitis data mining [14], hepatitis data mining [22, 2] and that of the eight rule sets from eight UCI benchmark datasets [15]. With the experimental results, we discuss about the following three view points: performances of rule evaluation models, minimum training subset to construct a valid rule evaluation model, and contents of learned rule evaluation models.

As evaluations of performances of rule evaluation models, we have compared predictive accuracies on the whole dataset and Leave-One-Out. The accuracy of a validation dataset D is calculated with correctly predicted instances $Correct(D)$ as $Acc(D) = (Correct(D)/|D|) \times 100$, where $|D|$ means the size of the dataset. Recalls of class i on a validation dataset is calculated with correctly predicted instances about the class $Correct(D_i)$ as $Recall(D_i) = (Correct(D_i)/|D_i|) \times 100$, where $|D_i|$ means the size of instances with class i. Also the precision of class i is calculated with the size of instances predicted i as $Precision(D_i) = (Correct(D_i)/Predicted(D_i)) \times 100$.

As for estimating minimum training subset to construct a valid rule evaluation model, we obtained learning curves about accuracies to the whole training dataset to evaluate whether each learning algorithm can perform in early stage of a process of rule evaluations.

On the result of the actual data mining, we have investigated elements of the rule evaluation models. Then, we consider the characteristics of objective indices, which are used in these rule evaluation models.

To construct a dataset to learn a rule evaluation model, values of objective indices have been calculated for each rule, taking 39 objective indices as shown in Table1. Thus, each dataset for each rule set has the same number of instances as the rule set. Each instance consists of 40 attributes including the class attribute.

To these dataset, we applied nine learning algorithms to compare their performance as a rule evaluation model construction method. We have taken the following learning algorithms from Weka [31]: C4.5 decision tree learner [27] called J4.8, neural network learner with back propagation (BPNN) [17], support

Table 1. The 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.

Theory	Index Name (**Abbreviation**) [Reference Number of Literature]
P	Coverage(**Coverage**), Prevalence(**Prevalence**) Precision(**Precision**), Recall(**Recall**) Support(**Support**), Specificity(**Specificity**) Accuracy(**Accuracy**), Lift(**Lift**) Leverage(**Leverage**), Added Value(**Added Value**)[30] Klösgen's Interestingness(**KI**)[19], Relative Risk(**RR**)[3] Brin's Interest(**BI**)[6], Brin's Conviction(**BC**)[6] Certainty Factor(**CF**)[30], Jaccard Coefficient(**Jaccard**)[30] F-Measure(**F-M**)[28], Odds Ratio(**OR**)[30] Yule's Q(**YuleQ**)[30], Yule's Y(**YuleY**)[30] Kappa(**Kappa**)[30], Collective Strength(**CST**)[30] Gray andOrlowska's Interestingness weighting Dependency(**GOI**)[12] Gini Gain(**Gini**)[30], Credibility(**Credibility**)[13]
S	χ^2 Measure for One Quadrant(χ^2**-M1**)[11] χ^2 Measure for Four Quadrant(χ^2**-M4**)[11]
I	J-Measure(**J-M**)[29], K-Measure(**K-M**)[23] Mutual Information(**MI**)[30] Yao and Liu's Interestingness 1 based on one-way support(**YLI1**)[33] Yao and Liu's Interestingness 2 based on two-way support(**YLI2**)[33] Yao and Zhong's Interestingness(**YZI**)[33]
N	Cosine Similarity(**CSI**)[30], Laplace Correction(**LC**)[30] ϕ Coefficient(ϕ)[30], Piatetsky-Shapiro's Interestingness(**PSI**)[24]
D	Gago and Bento's Interestingness(**GBI**)[10] Peculiarity(**Peculiarity**)[34]

vector machines (SVM)[5][25], classification via linear regressions (CLR)[6][7], and OneR [18]. In addition, we have also taken the following selective meta-learning algorithms: Bagging [5], Boosting [9] and Stacking[7] [32].

3.1 A Case Study on the Meningitis Datamining Result

In this case study, we have taken 244 rules, which are mined from six datasets about six kinds of diagnostic problems as shown in Table 2. These datasets are consisted of appearances of meningitis patients as attributes and diagnoses for each patient as class. Each rule set was mined with each proper rule induction algorithm composed by a constructive meta-learning system called CAMLET [14]. For each rule, we labeled three evaluations (I: Interesting, NI: Not-Interesting, NU: Not-Understandable), according to evaluation comments from a medical expert.

Constructing a proper learning algorithm to construct the meningitis rule evaluation model We have developed a constructive meta-learning sys-

[5] The kernel function was set up polynomial kernel.

[6] We set up the elimination of collinear attributes and the model selection with greedy search based on Akaike Information Metric.

[7] This stacking has taken the other seven learning algorithms as base-level learner and J4.8 as meta-level learner.

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

Dataset	#Attributes	#Class	#Mined rules	#'I' rules	#'NI' rules	#'NU' rules
Diag	29	6	53	15	38	0
C_Cource	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	29	2	24	3	18	3
TOTAL	—	—	244	48	187	9

tem called CAMLET [1] to choose a proper learning algorithm to a given dataset with machine learning method repository. To implement the method repository, firstly, we identified each functional part called method from the following eight learning algorithms: Version Space [21], AQ15 [20], Classifier Systems [4], Neural Network, ID3 [26], C4.5, Bagging and Boosting. With the method repository CAMLET constructs a proper learning algorithm to a given dataset, searching possible learning algorithm specification space which is obtained by the method repository.

Since we have set up the number of refinement $N = 100$, CAMLET searched up to 400 learning algorithms from 6000 possible learning algorithms for the best one. Fig. 2 shows the constructed algorithm by CAMLET to the dataset of meningitis datamining result.

This algorithm iterates boosting of C4.5 decision tree for randomly split training datasets. Each classifier set generated by C4.5 decision tree learner is reinforced with the method from Classifier Systems. Then, the learned committee aggregates with weighted voting from boosting.

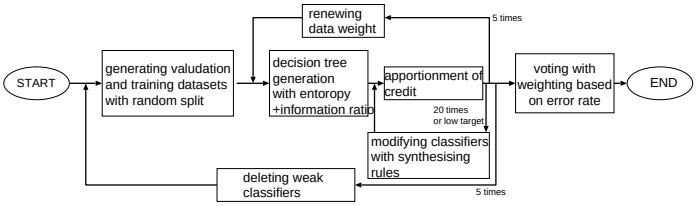


Fig. 2. The learning algorithm constructed by CAMLET for the dataset of the meningitis datamining result.

Comparison on classification performances In this section, we show the result of the comparisons of accuracies on the whole dataset, recall of each class label, and precisions of each class label.

The results of the performances of the nine learning algorithms to the whole training dataset and the results of Leave-One-Out are also shown in Table 3. All of the accuracies, Recalls of I and NI, and Precisions of I and NI on the whole training dataset are higher than just predicting each label at random. The accuracies of Leave-One-Out show robustness of each learning algorithm by which have been achieved from 75.8% to 81.9%.

The learning algorithm constructed by CAMLET shows the second accuracy to the whole training dataset, comparing with other learning algorithms. Thus,

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

Learning Algorithms	Evaluation on the training dataset						
	Acc.	Recall			Precision		
		I	NI	NU	I	NI	NU
CAMLET	89.4	70.8	97.9	11.1	85.0	90.2	100.0
Stacking	81.1	37.5	96.3	0.0	72.0	87.0	0.0
Boosted J4.8	99.2	97.9	99.5	100.0	97.9	99.5	100.0
Bagged J4.8	87.3	62.5	97.9	0.0	81.1	88.4	0.0
J4.8	85.7	41.7	97.9	66.7	80.0	86.3	85.7
BPNN	86.9	81.3	89.8	55.6	65.0	94.9	71.4
SVM	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
OneR	82.0	56.3	92.5	0.0	57.4	87.8	0.0

Learning Algorithms	Leave−One−Out(LOO)						
	Acc.	Recall			Precision		
		I	NI	NU	I	NI	NU
CAMLET	80.3	7.4	73.0	0.0	7.4	73.0	0.0
Stacking	81.1	37.5	96.3	0.0	72.0	87.0	0.0
Boosted J4.8	74.2	37.5	87.2	0.0	39.1	84.0	0.0
Bagged J4.8	77.9	31.3	93.6	0.0	50.0	81.8	0.0
J4.8	79.1	29.2	95.7	0.0	63.6	82.5	0.0
BPNN	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
CLR	80.3	35.4	95.7	0.0	60.7	82.9	0.0
OneR	75.8	27.1	92.0	0.0	37.1	82.3	0.0

CAMLET shows higher adaptability than the other selective meta-learning algorithms.

Estimating minimum training subsets for each learning algorithms The left table in Fig.3 shows accuracies to the whole training dataset with each subset of training dataset. Each data point is averaged accuracies from 10 times trials of randomly sub-sampled training datasets. The percentages of achievements for each learning algorithm, comparing with the accuracy with the whole dataset, are shown in the right chart of Fig.3.

%training sample	10	20	30	40	50	60	70	80	90	100
CAMLET	76.7	78.4	80.8	81.6	81.7	82.6	82.8	84.8	84.6	89.3
Stacking	69.6	77.8	75.3	77.9	72.2	82.2	75.4	83.4	86.5	81.1
Boosted J4.8	74.8	77.8	79.6	82.8	83.6	85.5	86.8	88.0	89.7	99.2
Bagged J4.8	77.5	79.5	80.5	81.4	81.8	82.1	83.2	83.2	84.1	87.3
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

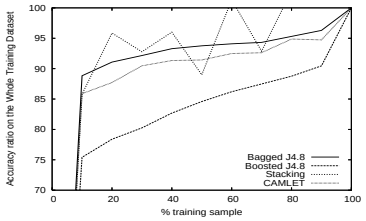


Fig. 3. Accuracies (%) with training sub-samples to the whole training dataset on the left table. And the chart of achieve rates(%) to the accuracies with the whole training dataset on the meta-learning algorithms.

As shown in these results, SVM, CLR and bagged J4.8 achieves higher than 95% with only less than 10% of training subset. Looking at the result of learning algorithm constructed by CAMLET, this algorithm achieves almost as same performance as bagged J4.8 with smaller training subset. However, it can outperform bagged J4.8 with larger training subsets. Although the constructed algorithm based on boosting, the combination of reinforcement method from Classifier Systems and the outer loop has been able to overcome a disadvantage of boosting for smaller training subset.

Rule evaluation models on the meningitis datamining result dataset In this section, we present the statistics of rule evaluation models to the 10000

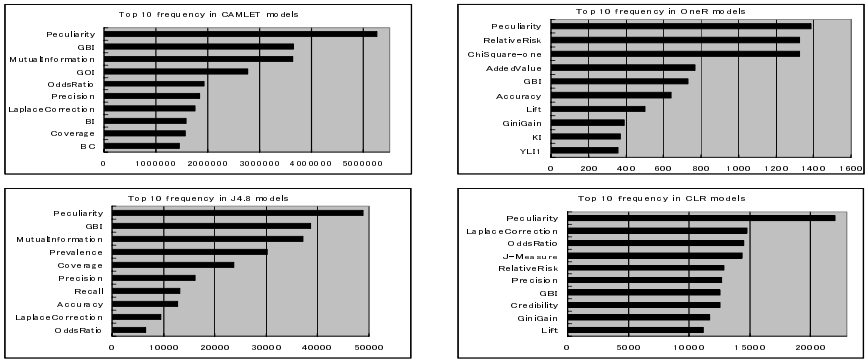


Fig. 4. Top 10 of frequencies of indices used in models of each learning algorithm with 10000 bootstrap samples of the meningitis datamining result dataset and executions.

bootstrap re-sampled dataset learned with the algorithm constructed by CAMLET, OneR, J4.8 and CLR, because they are represented as explicit models such as a rule set, a decision tree, and a set of linear models.

As shown in Fig. 4, indices used in learned rule evaluation models are not only the group of indices increasing with a correctness of a rule, but also they are used some different groups of indices on different models. Almost indices such as YLI1, Laplace Correction, Accuracy, Precision, Recall, Coverage, PSI and Gini Gain are the former type of indices on the models. The later indices are GBI and Peculiality, which sums up difference of antecedents between one rule and the other rules in the same rule set.

3.2 A Case Study on the Chronic Hepatitis Datamining Results

In this case study, we have taken four datamining results about chronic hepatitis as shown in the left table of Table 4. These datasets are consisted of patterns for each laboratory test value about blood and urine of chronic hepatitis patients as attributes. Firstly, we have done datamining processes to find out relationships between patterns of attributes and patterns of GPT as class, which is one of the important test items to grasp conditions of each patient, two times. Second, we have also done other datamining processes to find out relationships between patterns of attributes and results of interferon (IFN) therapy two times. For each rule, we labeled three evaluations (EI: Especially Interesting, I: Interesting, NI: Not-Interesting, NU: Not-Understandable), evaluated by another medical expert.

Constructing proper learning algorithms for chronic hepatitis datamining results As same as the construction of the proper learning algorithm for the meningitis data mining result, we constructed proper learning algorithms for the datasets of the four chronic hepatitis datamining results. The right table of Table 4 shows an overview of constructed learning algorithms for each dataset.

Table 4. Description of datasets of the chronic hepatitis datamining results (left table). And Overview of constructed learning algorithms by CAMLET to the datasets of the chronic hepatitis datamining results (right table).

	#Rules	Class Distribution				%Def class
		EI	I	NI	NU	
GPT						
Phase1(GPT1)	30	3	8	16	3	53.33
Phase2(GPT2)	21	2	6	12	1	57.14
IFN						
First Time(IFN1)	26	4	7	11	7	42.31
Second Time(IFN2)	32	15	5	11	1	46.88

	original classifier set	overall control structure	final eval. method
GPT1	C4.5 tree	Bagging	Best selection
GPT2	C4.5 tree	CS+Boost+Iteration	Weighted Voting
IFN1	C4.5 tree	CS+Boost+Iteration	Weighted Voting
IFN2	C4.5 tree	CS+Boost+Iteration	Weighted Voting

CS means including reinfoecement of classifier set from Classifiser Systems
Boost means including methods and control structure from Boosting

Comparison on classification performances The results of the performances of the nine learning algorithms to the whole training dataset and the results of Leave-One-Out are shown in Table5. Almost of the accuracies on the whole training dataset are higher than just predicting each default class. The accuracies of Leave-One-Out show robustness of each learning algorithm. To GPT1 and IFN1, they are lower than just predicting default classes, because the medical expert evaluated these datamining results without certain criterion in his mind.

Table 5. Accuracies(%), Recalls(%) and Precisions(%) of the nine learning algorithms on training dataset(the left table) and Leave-One-Out(the center table). Minimum training instances to construct valid rule evaluation models with each learning algorithm (the right table).

On Training		Precision				Recall				Leave-One-Out		Precision				Recall				Min.
	Acc	EI	I	NI	NU	EI	I	NI	NU	Acc	EI	I	NI	NU	EI	I	NI	NU	Estim.
GPT1																			
J4.8	96.7	100.0	88.9	100.0	100.0	66.7	100.0	100.0	100.0	50.0	0.0	60.0	60.0	0.0	0.0	75.0	56.3	0.0	14
BPNN	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	30.0	0.0	12.5	50.0	0.0	0.0	12.5	50.0	0.0	14
SVM	56.7	0.0	100.0	68.2	14.3	0.0	12.5	93.8	33.3	46.7	0.0	0.0	65.0	11.1	0.0	0.0	81.3	33.3	20
CLR	63.3	0.0	66.7	62.5	0.0	0.0	50.0	93.8	0.0	40.0	0.0	14.3	50.0	0.0	0.0	12.5	68.8	0.0	16
OneR	60.0	0.0	66.7	59.3	0.0	0.0	25.0	100.0	0.0	43.3	0.0	25.0	55.6	0.0	0.0	37.5	62.5	0.0	14
BagJ4.8	93.3	75.0	87.5	100.0	100.0	100.0	87.5	93.8	100.0	33.3	0.0	12.5	50.0	0.0	0.0	12.5	56.3	0.0	14
BooJ4.8	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	43.3	0.0	42.9	62.5	0.0	0.0	37.5	62.5	0.0	12
Stacking	70.0	0.0	62.5	72.7	0.0	0.0	62.5	100.0	0.0	36.7	0.0	33.3	61.5	0.0	0.0	37.5	50.0	0.0	24
CAMLET	73.3	0.0	50.0	87.5	100.0	0.0	75.0	87.5	66.7	43.3	0.0	6.7	33.3	3.3	0.0	6.7	33.3	3.3	16
GPT2																			
J4.8	90.5	66.7	85.7	100.0	0.0	100.0	100.0	91.7	0.0	76.2	0.0	66.7	90.9	0.0	0.0	100.0	83.3	0.0	6
BPNN	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	66.7	0.0	83.3	81.8	0.0	0.0	83.3	75.0	0.0	5
SVM	95.2	100.0	100.0	92.3	100.0	50.0	100.0	100.0	100.0	81.0	0.0	100.0	91.7	25.0	0.0	83.3	91.7	100.0	5
CLR	85.7	50.0	100.0	85.7	0.0	50.0	83.3	100.0	0.0	76.2	0.0	83.3	84.6	0.0	0.0	83.3	91.7	0.0	16
OneR	85.7	0.0	75.0	92.3	0.0	0.0	100.0	100.0	0.0	81.0	0.0	66.7	91.7	0.0	0.0	100.0	91.7	0.0	11
BagJ4.8	90.5	100.0	75.0	100.0	0.0	100.0	100.0	91.7	0.0	76.2	0.0	66.7	90.9	0.0	0.0	100.0	83.3	0.0	6
BooJ4.8	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	76.2	0.0	66.7	100.0	0.0	0.0	100.0	83.3	0.0	6
Stacking	61.9	66.7	0.0	100.0	0.0	100.0	0.0	91.7	0.0	71.4	0.0	83.3	76.9	0.0	0.0	83.3	83.3	0.0	11
CAMLET	81.0	0.0	75.0	84.6	0.0	0.0	100.0	91.7	0.0	76.2	0.0	28.6	47.6	0.0	0.0	28.6	47.6	0.0	8
INF1																			
J4.8	88.5	80.0	100.0	83.3	100.0	100.0	71.4	90.9	100.0	19.2	37.5	0.0	20.0	0.0	75.0	0.0	18.2	0.0	8
BPNN	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	26.9	40.0	22.2	25.0	25.0	50.0	28.6	18.2	25.0	6
SVM	46.2	26.7	0.0	70.0	100.0	100.0	0.0	63.6	25.0	34.6	21.4	0.0	54.5	0.0	75.0	0.0	54.5	0.0	10
CLR	53.8	100.0	0.0	47.6	66.7	50.0	0.0	90.9	50.0	19.2	33.3	0.0	28.6	0.0	25.0	0.0	36.4	0.0	16
OneR	50.0	0.0	50.0	50.0	0.0	0.0	85.7	63.6	0.0	19.2	0.0	11.1	23.5	0.0	0.0	14.3	36.4	0.0	18
BagJ4.8	96.2	80.0	100.0	100.0	100.0	100.0	100.0	90.9	100.0	26.9	33.3	37.5	22.2	0.0	50.0	42.9	18.2	0.0	10
BooJ4.8	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	23.1	42.9	0.0	27.3	0.0	75.0	0.0	27.3	0.0	8
Stacking	11.5	0.0	12.5	14.3	0.0	0.0	14.3	18.2	0.0	23.1	0.0	33.3	28.6	0.0	0.0	57.1	18.2	0.0	16
CAMLET	76.9	100.0	60.0	80.0	100.0	100.0	85.7	72.7	50.0	30.8	11.5	0.0	19.2	0.0	11.5	0.0	19.2	0.0	14
INF2																			
J4.8	90.6	88.2	100.0	90.9	0.0	100.0	80.0	90.9	0.0	75.0	76.5	66.7	75.0	0.0	86.7	40.0	81.8	0.0	6
BPNN	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	37.5	50.0	28.6	22.2	0.0	53.3	40.0	18.2	0.0	8
SVM	56.3	72.7	0.0	45.0	100.0	53.3	0.0	81.8	100.0	31.3	36.4	0.0	28.6	0.0	26.7	0.0	54.5	0.0	8
CLR	65.6	63.2	100.0	60.0	0.0	80.0	60.0	54.5	0.0	34.4	41.2	20.0	30.0	0.0	46.7	20.0	27.3	0.0	16
OneR	68.8	62.5	0.0	87.5	0.0	100.0	0.0	63.6	0.0	68.8	60.0	0.0	100.0	0.0	100.0	0.0	63.6	0.0	16
BagJ4.8	90.6	88.2	100.0	90.9	0.0	100.0	80.0	90.9	0.0	71.9	70.0	100.0	72.7	0.0	93.3	20.0	72.2	0.0	8
BooJ4.8	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	71.9	76.5	100.0	70.0	0.0	86.7	60.0	63.6	0.0	6
Stacking	40.6	46.2	0.0	33.3	0.0	80.0	0.0	9.1	0.0	53.1	58.8	0.0	58.3	0.0	66.7	0.0	63.6	0.0	12
CAMLET	90.6	83.3	100.0	100.0	100.0	100.0	100.0	72.7	100.0	43.8	18.8	0.0	18.8	0.0	18.8	0.0	18.8	0.0	8

The learning algorithm constructed by CAMLET shows almost the same predictive performance as Boosted J4.8 on LOO, because these algorithms consist of C4.5 decision tree learner with boosting control structure.

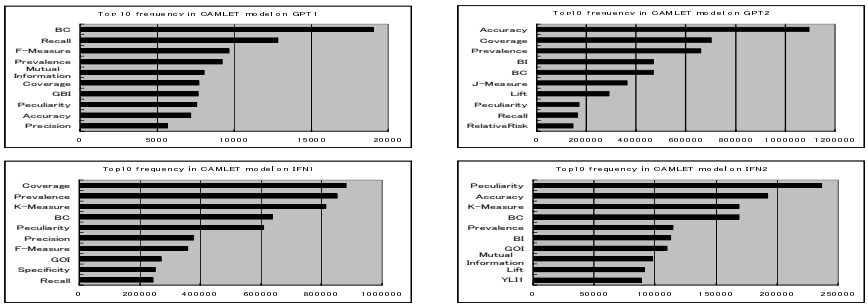


Fig. 5. Top 10 of frequencies of indices used in models of learning algorithms composed by CAMLET with 10000 bootstrap samples of the chronic hepatitis datamining results' datasets.

Estimating minimum training subset to construct a valid rule evaluation model As same as the case study of meningitis dataset, we have estimated the minimum training subsets for a valid model, which works better than just predicting a default class as shown in Table 5.

To GPT1 and IFN1, these algorithms need more instances to learn valid rule evaluation models than that of GPT2 and IFN2. This caused by the difference of human criteria when evaluating each datamining result.

Rule evaluation models on the chronic hepatitis datamining result dataset In this section, we present the statistics of rule evaluation models to the 10000 times bootstrap re-sampled dataset learned with the algorithm constructed by CAMLET to compare the difference among the models.

As shown in Fig. 5, these models consist of not only indices expressing correctness of rules but also other types of indices as shown in meningitis rule evaluation models (Fig. 4). This shows that the medical expert evaluated these rules with both of correctness and interestingness based on his background knowledge.

On each problem, the variance of indices has been reduced in each second time datamining process. This indicates that the medical expert evaluated each second time datamining result with more certain criterion than it of each first time datamining process.

3.3 An Experiment on Artificial Evaluation Labels

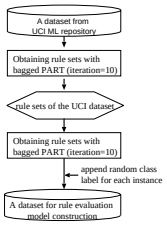
We have also evaluated our rule evaluation model construction method with rule sets from four datasets of UCI Machine Learning Repository [15] to investigate the performances without any human criteria.

We have taken the following eight dataset: anneal, audiology, autos, balance-scale, breast-cancer, breast-w, colic, and credit-a. With these datasets, we obtained rule sets with bagged PART, which repeatedly executes PART [8] to bootstrapped training sub-sample datasets.

To these rule sets, we calculated the 39 objective indices as attributes of each rule. As for the class of these datasets, we set up three class distributions

with multinomial distribution. Table 6 shows us the process flow diagram to obtain the datasets and the description of datasets with three different class distributions. The class distribution for 'Distribution I' is $P = (0.35, 0.3, 0.3)$ where p_i is the probability for class i. Thus, the number of class i in each instance D_j become $p_i D_j$. As the same way, the probability vector of 'Distribution II' is $P = (0.3, 0.65, 0.05)$. We have investigated performances of learning algorithms on these balanced class distribution and unbalanced class distribution.

Table 6. Flow diagram to obtain datasets and the datasets of the rule sets learned from the UCI benchmark datasets



	#Mined Rules	#Class labels L1	L2	L3	%Def. class
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
Distribution II		(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

Constructing proper learning algorithms for rule sets from UCI datasets As same as the construction of the proper learning algorithm for the meningitis data mining result, we constructed proper learning algorithms for the datasets of rule sets from the eight UCI datasets. Table7 shows an overview of constructed learning algorithms for each dataset which has two different class distributions.

Table 7. Overview of constructed learning algorithms by CAMLET to the datasets of the rule sets learned from the UCI benchmark datasets

	Distribution I			Distribution II		
	original classifier set	overall control structure	final eval. method	original classifier set	overall control structure	final eval. method
anneal	C4.5 tree	Win+Boost+CS	Weighted Voting	C4.5 tree	Boost+CS	Weighted Voting
audiology	ID3 tree	Boost	Voting	Random Rule	Simple Iteration	Best Select.
autos	Random Rule	Win+Iteration	Weighted Voting	Random Rule	Boost	Weighted Voting
balance-scale	Random Rule	Boost	Voting	Random Rule	CS+GA	Voting
breast-cancer	Random Rule	GA+Iteration	Voting	Random Rule	Win+Iteration	Weighted Voting
breast-w	ID3 tree	Win	Weighted Voting	ID3 tree	CS+Iteration	Weighted Voting
colic	Random Rule	CS+Win	Voting	ID3 tree	Win+Iteration	Voting
credit-a	C4.5 tree	Win+Iteration	Voting	ID3 tree	CS+Boost+Iteration	Best Select.

CS means including reinfoecement of classifier set from Classifiser Systems
Boost means including methods and control structure from Boosting
Win means including methods and control structure from Window Strategy
GA means including reinforcement of classifier set with Genetic Algorithm

Accuracy Comparison on Classification Performances To above datasets, we have attempted the nine learning algorithms to estimate whether their classification results can go to or beyond the percentages of just predicting each default class. The left table of Table 8 shows the accuracies of the nine learning algorithms to each class distribution of the eight datasets. The learning algorithms constructed by CAMLET, boosted J4.8, bagged J4.8, J4.8 and BPNN always work better than just predicting a default class. However, their performances are suffered from probabilistic class distributions to larger datasets.

Table 8. Accuracies(%) on whole training datasets labeled with three different distributions(The left table). Number of minimum training sub-samples to outperform %Def. class(The right table).

	Distribution I								
	J4.8	BPNN	SVM	CLR	OneR	Bagged J4.8	Boosted J4.8	Stacking	CAMLET
anneal	74.7	71.6	47.4	56.8	55.8	87.4	100.0	27.4	77.9
audiology	47.0	51.7	40.3	45.6	52.3	87.2	47.0	21.5	63.1
autos	66.7	63.8	46.8	46.1	56.0	89.4	66.7	29.8	53.2
balance-scale	58.0	59.4	39.5	43.4	53.0	83.3	58.0	39.5	39.5
breast-cancer	55.7	61.5	40.2	50.8	59.0	88.5	70.5	23.8	41.0
breast-w	86.1	91.1	38.0	46.8	54.4	96.2	100.0	34.2	77.2
colic	91.8	82.0	42.6	60.7	55.7	88.5	100.0	29.5	67.2
credit-a	57.4	48.7	35.7	39.1	54.8	91.3	57.4	26.5	55.7
	Distribution II								
	J4.8	BPNN	SVM	CLR	OneR	Bagged J4.8	Boosted J4.8	Stacking	CAMLET
anneal	74.7	70.5	67.4	70.5	73.7	84.2	94.7	67.4	66.3
audiology	65.8	67.8	63.8	64.4	67.1	83.2	67.1	59.7	65.1
autos	85.1	73.8	68.1	70.2	73.8	87.9	100.0	66.7	67.4
balance-scale	70.5	69.8	64.8	65.8	69.8	80.1	85.8	62.6	63.0
breast-cancer	71.3	77.0	66.4	65.6	77.9	86.9	79.5	73.0	73.0
breast-w	74.7	86.1	73.4	68.4	74.7	87.3	100.0	63.3	70.9
colic	70.5	77.0	65.6	60.7	73.8	85.2	100.0	49.2	60.7
credit-a	70.9	70.0	65.2	65.2	71.3	85.7	87.8	61.7	65.2

	Distribution I								
	J4.8	BPNN	SVM	CLR	OneR	Bagged J4.8	Boosted J4.8	Stacking	CAMLET
anneal	20	14	17	29	29	16	14	36	20
audiology	21	18	65	64	41	21	14	56	27
autos	38	28	76	77	70	28	28	77	31
balance-scale	12	14	15	15	32	14	9	51	128
breast-cancer	16	17	22	41	22	14	14	41	36
breast-w	7	10	10	18	14	10	6	19	11
colic	8	8	9	22	14	8	8	24	8
credit-a	9	12	16	30	28	9	8	51	19
	Distribution II								
	J4.8	BPNN	SVM	CLR	OneR	Bagged J4.8	Boosted J4.8	Stacking	CAMLET
anneal	54	58	64	76	–	42	38	64	46
audiology	64	73	45	76	107	50	50	103	84
autos	66	102	84	121	98	45	39	76	76
balance-scale	118	103	133	162	156	86	92	132	–
breast-cancer	50	31	80	92	80	38	36	60	41
breast-w	44	36	31	48	71	34	34	52	53
colic	28	24	46	30	42	28	22	48	54
credit-a	118	159	–	–	173	76	76	120	109

Estimating minimum training subset to construct a valid rule evaluation model As same as the case study of meningitis dataset, we have estimated the minimum training subsets for a valid model, which works better than just predicting a default class as shown in the right table in Table8. To datasets with balanced class distribution, almost of learning algorithm can construct valid models with less than 20% of given training datasets. However, to datasets with unbalanced class distribution, they need more training subsets to construct valid models, because their performances with whole training dataset fall to the percentages of default class of each dataset as shown in the left table in Table8.

4 Conclusion

In this paper, we have described the evaluation of the nine learning algorithms for a rule evaluation support method with rule evaluation models to predict evaluations for an IF-THEN rule based on objective indices, re-using evaluations of a human expert.

As the result of the performance comparison with the nine learning algorithms on the dataset of meningitis data mining result, rule evaluation models have achieved higher accuracies than just predicting each default class. To this dataset, the learning algorithm constructed by CAMLET shows higher accuracy with higher reliability than the other eight learning algorithm including three meta-learning algorithm. From the results on the datasets of hepatitis datamining results, we find out that the difference of human evaluation criteria appear as the differences of rule evaluation models on both of performances and their contents. To datasets of rule sets obtained from the eight UCI datasets, although hyper-plane type learners, such as SVM and CLR, and Stacking have failed to go to the percentage of default class of some datasets, the other learning algorithms have been able to go to or beyond each percentage of default class with smaller than 50% of each training dataset. Considering the difference between the actual evaluation labeling and the artificial evaluation labeling, it is shown that the medical expert evaluated with noticing particular relations between an antecedent and a class/another antecedent in each rule. This indicates that our

approach can detect differences of human criteria as differences of performances of rule evaluation models.

As future work, we will improve the method repository of CAMLET to construct more suitable learning algorithms for rule evaluation models. We also apply this rule evaluation support method to other datasets from various domains.

References

1. Abe, H. and Yamaguchi, T.: Constructive Meta-Learning with Machine Learning Method Repositories, in Proc. of the 17th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems IEA/AIE 2004, LNAI 3029, (2004) 502–511
2. Abe, H., Ohsaki, M., Yokoi, H., and Yamaguchi, T.: Implementing an Integrated Time-Series Data Mining Environment based on Temporal Pattern Extraction Methods – A Case Study of an Interferon Therapy Risk Mining for Chronic Hepatitis –, JSAI2005 Workshops, LNAI 4012, 425–435
3. Ali, K., Manganaris,S., Srikant, R.: Partial Classification Using Association Rules. in Proc. of Int. Conf. on Knowledge Discovery and Data Mining KDD-1997 (1997) 115–118
4. Booker, L. B., Holland, J. H., and Goldberg, D. E.: Classifier Systems and Genetic Algorithms, Artificail Inteligence, 40 (1989) 235–282
5. Breiman, L.: Bagging Predictors, Machine Learning, 24(2) (1996) 123–140
6. Brin, S., Motwani, R., Ullman, J., Tsur, S.: Dynamic itemset counting and implication rules for market basket data. Proc. of ACM SIGMOD Int. Conf. on Management of Data (1997) 255–264
7. Frank, E., Wang, Y., Inglis, S., Holmes, G., and Witten, I. H.: Using model trees for classification, Machine Learning, Vol.32, No.1 (1998) 63–76
8. Frank, E, Witten, I. H., Generating accurate rule sets without global optimization, in Proc. of the Fifteenth International Conference on Machine Learning, (1998) 144–151
9. Freund, Y., and Schapire, R. E.: Experiments with a new boosting algorithm, in Proc. of Thirteenth International Conference on Machine Learning (1996) 148–156
10. Gago, P., Bento, C.: A Metric for Selection of the Most Promising Rules. Proc. of Euro. Conf. on the Principles of Data Mining and Knowledge Discovery PKDD-1998 (1998) 19–27
11. Goodman, L. A., Kruskal, W. H.: Measures of association for cross classifications. Springer Series in Statistics, 1, Springer-Verlag (1979)
12. Gray, B., Orlowska, M. E.: CCAIIA: Clustering Categorical Attributes into Interesting Association Rules. Proc. of Pacific-Asia Conf. on Knowledge Discovery and Data Mining PAKDD-1998 (1998) 132–143
13. Hamilton, H. J., Shan, N., Ziarko, W.: Machine Learning of Credible Classifications. in Proc. of Australian Conf. on Artificial Intelligence AI-1997 (1997) 330–339
14. Hatazawa, H., Negishi, N., Suyama, A., Tsumoto, S., and Yamaguchi, T.: Knowledge Discovery Support from a Meningoencephalitis Database Using an Automatic Composition Tool for Inductive Applications, in Proc. of KDD Challenge 2000 in conjunction with PAKDD2000 (2000) 28–33
15. Hettich, S., Blake, C. L., and Merz, C. J.: UCI Repository of machine learning databases [http://www.ics.uci.edu/˜mlearn/MLRepository.html], Irvine, CA: University of California, Department of Information and Computer Science, (1998).

16. Hilderman, R. J. and Hamilton, H. J.: Knowledge Discovery and Measure of Interest, Kluwe Academic Publishers (2001)
17. Hinton, G. E.: "Learning distributed representations of concepts", in Proc. of 8th Annual Conference of the Cognitive Science Society, Amherest, MA. REprinted in R.G.M.Morris (ed.) (1986)
18. Holte, R. C.: Very simple classification rules perform well on most commonly used datasets, Machine Learning, Vol. 11 (1993) 63–91
19. Klösgen, W.: 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 (1996) 249–271
20. Michalski, R., Mozetic, I., Hong, J. and Lavrac, N.: The AQ15 Inductive Learning System: An Over View and Experiments, Reports of Machine Learning and Inference Laboratory, No.MLI-86-6, George Mason University (1986).
21. Mitchell, T. M.: Generalization as Search, Artificial Intelligence, 18(2) (1982) 203–226
22. Ohsaki, M., Sato, Y., Kitaguchi, S., Yokoi, H., and Yamaguchi, T.: Comparison between Objective Interestingness Measures and Real Human Interest in Medical Data Mining, in Proc. of the 17th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems IEA/AIE 2004, LNAI 3029, (2004) 1072–1081
23. Ohsaki, M., Kitaguchi, S., Kume, S., Yokoi, H., and Yamaguchi, T.: Evaluation of Rule Interestingness Measures with a Clinical Dataset on Hepatitis, in Proc. of ECML/PKDD 2004, LNAI3202 (2004) 362–373
24. Piatetsky-Shapiro, G.: Discovery, Analysis and Presentation of Strong Rules. in Piatetsky-Shapiro, G., Frawley, W. J. (eds.): Knowledge Discovery in Databases. AAAI/MIT Press (1991) 229–248
25. Platt, J.: Fast Training of Support Vector Machines using Sequential Minimal Optimization, Advances in Kernel Methods - Support Vector Learning, B. Schölkopf, C. Burges, and A. Smola, eds., MIT Press (1999) 185–208
26. Quinlan, J. R. : Induction of Decision Tree, Machine Learning, 1 (1986) 81–106
27. Quinlan, R.: C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, (1993)
28. Rijsbergen, C.: Information Retrieval, Chapter 7, Butterworths, London, (1979) http://www.dcs.gla.ac.uk/Keith/Chapter.7/Ch.7.html
29. Smyth, P., Goodman, R. M.: Rule Induction using Information Theory. in Piatetsky-Shapiro, G., Frawley, W. J. (eds.): Knowledge Discovery in Databases. AAAI/MIT Press (1991) 159–176
30. Tan, P. N., Kumar V., Srivastava, J.: Selecting the Right Interestingness Measure for Association Patterns. Proc. of Int. Conf. on Knowledge Discovery and Data Mining KDD-2002 (2002) 32–41
31. Witten, I. H and Frank, E.: DataMining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, (2000)
32. Wolpert, D. : Stacked Generalization, Neural Network 5(2) (1992) 241–260
33. Yao, Y. Y. Zhong, N.: An Analysis of Quantitative Measures Associated with Rules. Proc. of Pacific-Asia Conf. on Knowledge Discovery and Data Mining PAKDD-1999 (1999) 479–488
34. Zhong, N., Yao, Y. Y., Ohshima, M.: Peculiarity Oriented Multi-Database Mining. IEEE Trans. on Knowledge and Data Engineering, 15, 4, (2003) 952–960