Developing a Rule Evaluation Support Method Based on Objective Indices

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

[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] Faculty of Science and Technology, Keio University
yamaguti@ae.keio.ac.jp

Abstract. In this paper, we present an evaluation of a rule evaluation support method for post-processing of mined results with rule evaluation models based on objective indices. To reduce the costs of rule evaluation task, which is one of the key procedures in data mining post-processing, we have developed the rule evaluation support method with rule evaluation models, which are obtained with objective indices of mined classification rules and evaluations of a human expert for each rule. Then we have evaluated performances of learning algorithms for constructing rule evaluation models on the meningitis data mining as an actual problem and five rulesets from the five kinds of UCI datasets. With these results, we show the availability of our rule evaluation support method.

Keywords: Data Mining, Post-processing, Rule Evaluation Support, Objective Indices.

1 Introduction

In recent years, it is required by people to utilize huge data, which are easily stored on information systems, developing information technologies. Besides, data mining techniques have been widely known as a process for utilizing stored data, combining database technologies, statistical methods, and machine learning methods. Although, IF-THEN rules are discussed as one of highly usable and readable output of data mining, to large dataset with hundreds attributes including noises, a rule mining process often obtains many thousands of rules. From such huge rule set, it is difficult for human experts to find out valuable knowledge which are rarely included in the rule set.

To support a rule selection, many efforts have done using objective rule evaluation indices[1–3] such as recall, precision and interestingness measurements (called 'objective indices' later), which are calculated by the mathematical analysis and do not include any human evaluation criteria. However, it is also difficult to estimate a criterion of a human expert with single objective rule evaluation

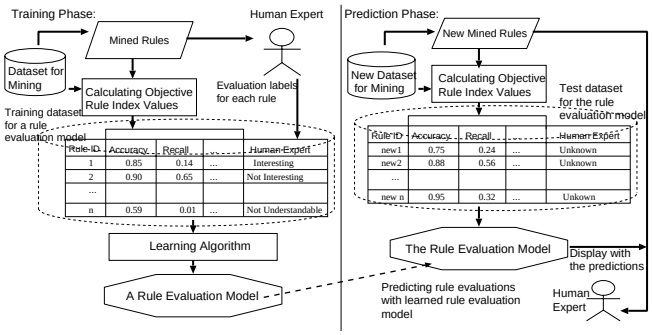

Fig. 1. Overview of the construction method of rule evaluation models.

index[4], because his/her subjective criterion such as interestingness is influenced by the amount of his/her knowledge.

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 this paper, we present a performance comparison of learning algorithms for constructing rule evaluation models. Then we discuss about the availability of 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. Fig.1 shows the process of rule evaluation model construction based on re-use of human evaluations and objective indices.

At the training phase, attributes of a meta-level training data set is obtained by objective indices values. At the same time, a human expert evaluates the whole or part of input rules at least once to join as class of each instance. 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

In this section, we firstly present the result of an empirical evaluation with the dataset from the result of a meningitis data mining[5]. Then to confirm the performance of our approach, we present the result on five kinds of UCI

benchmark datasets [6]. In these case studies , we have evaluated the following three view points: performances of learning algorithms, estimations of minimum training subsets to construct valid rule evaluation models, and contents of learned rule evaluation models.

To construct a dataset to learn a rule evaluation model, 39 objective indices [4] have been calculated for each rule. To these dataset, we applied the following five learning algorithms from Weka[7]: C4.5 decision tree learner[8] called J4.8, neural network learner with back propagation (BPNN)[9], support vector machines (SVM) [10], classification via linear regressions (CLR) [11], and OneR[12].

3.1 Constructing Rule Evaluation Models on an Actual Datamining Result

In this case study, we have taken 244 rules, which are mined from six dataset about six kinds of diagnostic problems as shown in Table1. 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 CAMLET[5]. For each rule, we labeled three evaluations (I:Interesting, NI:Not-Interesting, NU:Not-Understandable), according to evaluation comments from a medical expert.

Table 1. 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

Comparison on Performances In this section, we show the result of the comparisons of performances on the whole dataset, recall and precisions of each class label. Since Leave-One-Out holds just one test instance and remains 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 whole training dataset and the results of Leave-One-Out are also shown in Table2.

Table 2. Accuracies(%), Recalls(%) and Precisions(%) of the five learning algorithms.

	On the whole training dataset							Leave-One-Out						
		Recall of			Precision of				Recall of			Precision of		
	Acc.	I	NI	NU	I	NI	NU	Acc.	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

These learning algorithms excepting OneR achieve equal or higher performance with combination of multiple objective indices than sorting with sin-

gle objective index. The accuracies of Leave-One-Out shows robustness of each learningalgorithm. These learning algorithms have achieved from 75.8% to 81.9%.

Estimating minimum training subset to construct a valid rule evaluation model Since the rule evaluation model construction method needs evaluations of mined rules by a human expert, we have estimated minimum training subset to construct a valid rule evaluation model. Table3 shows accuracies to the whole training dataset with each subset of training dataset. As shown in these results, SVM and CLR, which learn hype-planes, achieves grater than 95% with only less than 10% of training subset. Although decision tree learner and BPNN could learn better classifier to the whole dataset than these hyper-plane learners, they need more training instances to learn accurate classifiers.

Table 3. Accuracies(%) on the whole training dataset of the learning algorithms trained by sub-sampled training datasets.

%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

Rule Evaluation Models on the Actual Datamining Result Dataset In this section, we present rule evaluation models to the whole dataset learned with 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.

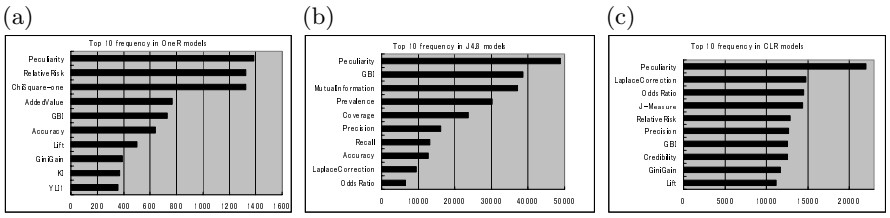

Fig. 2. Top 10 frequencies of indices of learned rule evaluation models by OneR(a), J4.8(b), and CLR(c). Statistics are collected by 10,000 times bootstrap iterations.

As shown in Fig.2, indices used in learned rule evaluation models, they 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, and Coverage are the former type of indices on the models. The later indices are GBI[13] and Peculiality[14], which sums up difference of antecedents between one rule and the other rules in the same ruleset. This corresponds to the comment from the human expert.

3.2 Constructing Rule Evaluation Models on Artificial Evaluation Labels

To confirm the performances without any human criteria, we have also evaluated our method with rule sets from the following five datasets of UCI Machine Learning Repository: Mushroom, Heart, Internet Advertisement Identification (called InternetAd later), Waveform-5000, and Letter. From these datasets, we obtained rule sets with bagged PART, which repeatedly executes PART[15] 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. The left table of Table4 shows us the datasets with three different class distributions.

Table 4. Datasets of the rule sets learned from the UCI benchmark datasets(the left table), accuracies(%) on whole training datasets(the center table), and number of minimum training sub-samples to outperform %Def. class(rhe right table).

	#Mined Rules	#Class labels L1	L2	L3	%Def. class
Distribution I		(0.30)	(0.35)	(0.35)	
Mushroom	30	8	14	8	46.7
InternetAd	107	26	39	42	39.3
Heart	318	97	128	93	40.3
Waveform	518	146	192	180	37.1
Letter	6340	1908	2163	2269	35.8
Distribution II		(0.30)	(0.50)	(0.20)	
Mushroom	30	11	16	3	53.3
InternetAd	107	30	53	24	49.5
Heart	318	99	140	79	44.0
Waveform	824	240	436	148	52.9
Letter	6340	1890	3198	1252	50.4
Distribution III		(0.30)	(0.65)	(0.05)	
Mushroom	30	7	21	2	70.0
InternetAd	107	24	79	9	73.8
Heart	318	98	205	15	64.5
Waveform	824	246	529	49	64.2
Letter	6340	1947	4062	331	64.1

	J48	BPNN	SVM	CLR	OneR
Distribution I					
Mushroom	100.0	93.3	56.7	66.7	53.3
InternetAd	84.1	82.2	29.9	53.3	80.7
Heart	76.0	75.8	40.3	42.5	54.7
Waveform	46.5	46.4	37.6	39.8	54.9
Letter	36.8	36.4	30.1	36.8	52.1
Distribution II					
Mushroom	93.3	93.3	80.0	80.0	76.7
InternetAd	73.8	79.4	49.5	59.8	80.7
Heart	72.3	69.2	35.9	47.8	55.7
Waveform	61.2	57.8	52.9	53.0	59.7
Letter	51.0	51.0	50.4	50.4	57.0
Distribution III					
Mushroom	93.3	96.7	70.0	70.0	76.7
InternetAd	86.0	90.7	70.1	69.2	72.0
Heart	76.0	77.7	64.5	65.7	71.4
Waveform	74.4	69.3	64.2	64.2	69.3
Letter	64.1	64.3	64.1	64.1	68.3

	J48	BPNN	SVM	CLR	OneR
Distribution I					
Mushroom	8	8	12	18	14
InternetAd	14	14	–	30	14
Heart	42	31	66	114	98
Waveform	60	52	46	355	152
Letter	189	217	–	955	305
Distribution II					
Mushroom	6	4	4	6	12
InternetAd	24	24	52	42	70
Heart	52	40	–	104	92
Waveform	251	355	763	–	533
Letter	897	>1000	451	–	>1000
Distribution III					
Mushroom	22	14	22	28	22
InternetAd	80	66	–	–	–
Heart	114	94	142	318	182
Waveform	329	425	191	–	601
Letter	>1000	>1000	998	>1000	>1000

Accuracy Comparison on Classification Performances As shown in the center table of Table4, J48 and BPNN always work better than just predicting a default class. However, their performances are suffered from probabilistic class distributions to larger datasets such as Heart and Letter.

Evaluation on Learning Curves As shown in the right table of Table4, to smaller dataset, such as Mushroom and InternetAd, they can construct valid models with less than 20% of given training datasets. However, to larger dataset, they need more training subsets to construct valid models, because their performances with whole training dataset fall to the percentages of default class.

4 Conclusion

In this paper, we have described rule evaluation support method with rule evaluation models to predict evaluations for an IF-THEN rule based on objective indices. As the result of the performance comparison with the five learning algorithms, rule evaluation models have achieved higher accuracies than just predicting each default class. Considering the difference between the actual evaluation labeling and the artificial evaluation labeling, it is shown that the medical expert evaluated with certain subjective criterion. In the estimations of minimum training subset for constructing a valid rule evaluation model on the dataset of

the actual datamining result, SVM and CLR have achieved more than 95% of achievement ratio compared to the accuracy of the whole training dataset with less than 10% of subset of the training dataset with certain human evaluations. These results indicate the availability of our method to support a human expert.

As future work, we will introduce a selection method of learning algorithms to construct a proper rule evaluation model according to each situation.

References

1. Hilderman, R. J. and Hamilton, H. J.: *Knowledge Discovery and Measure of Interest*, Kluwer Academic Publishers (2001)
2. Tan, P. N., Kumar V., Srivastava, J.: Selecting the Right Interestingness Measure for Association Patterns. in Proc. of Int. Conf. on Knowledge Discovery and Data Mining KDD-2002 (2002) 32–41
3. Yao, Y. Y. Zhong, N.: An Analysis of Quantitative Measures Associated with Rules. in Proc. of Pacific-Asia Conf. on Knowledge Discovery and Data Mining PAKDD-1999 (1999) 479–488
4. 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
5. 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
6. 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)
7. Witten, I. H and Frank, E.: DataMining: *Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, (2000)
8. Quinlan, R.: *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, (1993)
9. 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)
10. 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
11. Frank, E., Wang, Y., Inglis, S., Holmes, G., and Witten, I. H.: Using model trees for classification, *Machine Learning*, **32**(1) (1998) 63–76
12. Holte, R. C.: Very simple classification rules perform well on most commonly used datasets, *Machine Learning*, **11** (1993) 63–91
13. Gago, P., Bento, C.: A Metric for Selection of the Most Promising Rules. in Proc. of Euro. Conf. on the Principles of Data Mining and Knowledge Discovery PKDD-1998 (1998) 19–27
14. Zhong, N., Yao, Y. Y., Ohshima, M.: Peculiarity Oriented Multi-Database Mining. *IEEE Trans. on Knowledge and Data Engineering*, **15**(4) (2003) 952–960
15. 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