



Evaluating Learning Algorithms to Support Human Rule Evaluation Based on Objective Rule Evaluation Indices

Hidenao Abe¹⁾, S. Tsumoto¹⁾, M. Ohsaki²⁾, T. Yamaguchi³⁾

Dept. of Medical Informatics, Shimane University, School of Medicine,
Japan¹⁾

Faculty of Engineering, Doshisha University, Japan²⁾

Faculty of Science and Technology, Keio University, Japan³⁾




Outline

- Background and Research Issues
- Rule Evaluation Support Method based on Objective Rule Evaluation indices
- Comparisons of Learning Algorithms for Rule Evaluation Model Construction
- Conclusion



Research Background & Related Work

- Many efforts have done to select rules with single objective index such as recall, precision, and so forth.
- At least 40 objective interestingness measures are developed and investigated to express a human evaluation criterion.

- 
- Ohsaki et al. investigated the relationship between each index and criterion of an expert. However, no single objective index can express the human criterion exactly. [Ohsaki04].
 - Applicable domain of these interestingness measures have been never generalized.



Research Issues

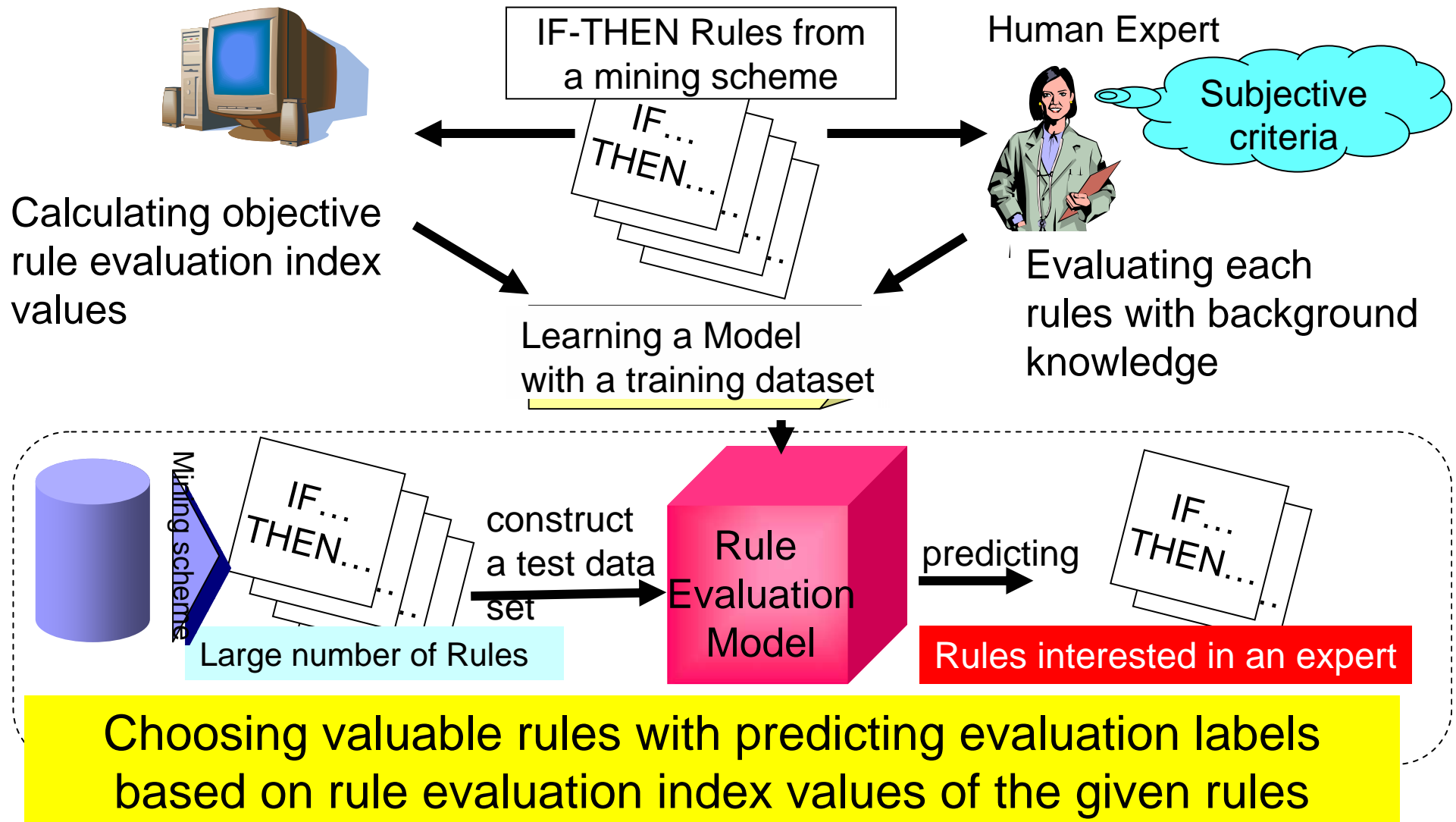
- A novel rule evaluation support method with rule evaluation models (REMs).
 - The system obtains a dataset to combine multiple objective indices and evaluations from a human expert.
- Detailed issues of our rule evaluation support method
 - To construct more accurate REMs to support human experts more exactly
 - To construct a valid REM with smaller training dataset
 - To construct a reasonable REMs to given human evaluation



Outline

- Background
- Rule Evaluation Support Method based on Objective Rule Evaluation indices
- Comparisons of Learning Algorithms for Rule Evaluation Model Construction
- Conclusion

Overview of the rule evaluation support with REMs





Outline

- Background
- Rule Evaluation Support Method based on Objective Rule Evaluation indices
- Comparisons of Learning Algorithms for Rule Evaluation Model Construction
- Conclusion



Comparisons of learning algorithms

- Comparison on an actual datamining result
 - To evaluate the availability on solid evaluations from a domain expert
- Comparison on rule sets of benchmark datasets with artificial class distributions
 - To evaluate the availability on evaluations without any particular human criterion
- Evaluation viewpoints for these comparisons:
 - Accuracies to the whole dataset and Leave-One-Out validation, and their recalls and precisions of each class label
 - Estimating minimum size of training subset to construct valid REMs with learning curves
 - Looking at elements of REMs from an actual data mining result



Objective Rule Evaluation indices

calculated on a validation dataset for each classification rule

The 39 objective indices [Ohsaki 04]

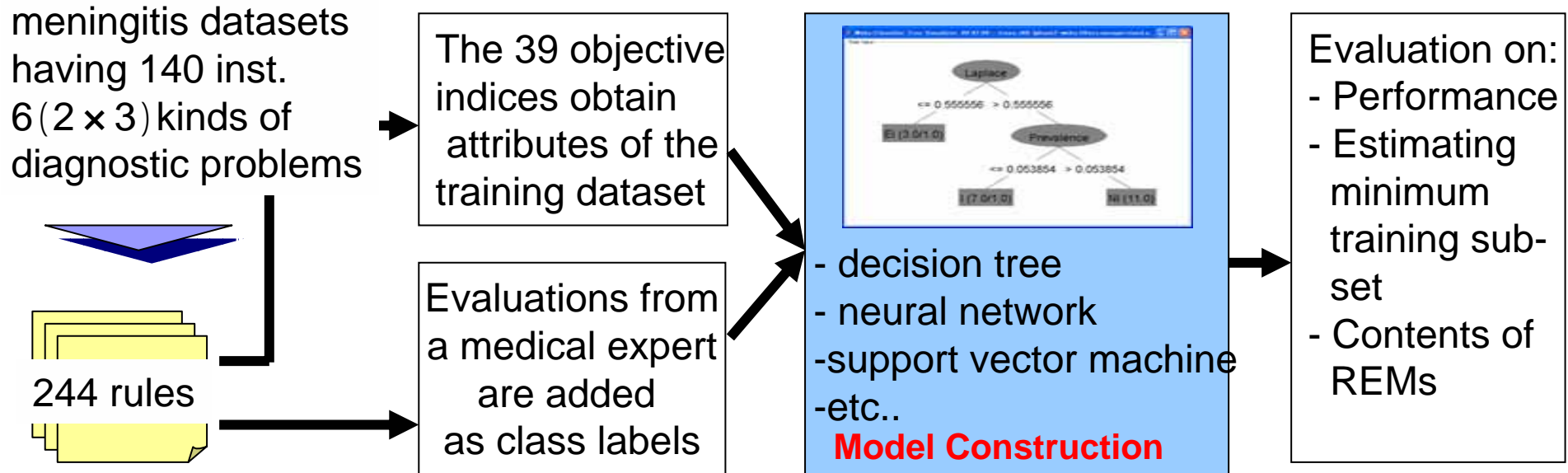
- Based on probability (26 indices)
 - Coverage, Prevalence, Precision, Recall, Support, Accuracy, Specificity, Lift, Leverage, Added Value, Relative Risk, Jaccard, Certainty Factor, Odds ratio, Yule's Q, Yule's Y, Kappa, Koelesgen's Interestingness, Brin's Interestingness, Brin's Conviction, GOI, Credibility, KSI, Laplace Correction, Collective Strength
- Based on test statistics (3 indices)
 - Chi-Square(with only True/Positive, with a whole confusion matrix), Gini Gain
- Based on information theory (6 indices)
 - Mutual Information, J-Measure, YLI 1, YLI 2, YZI, K-Measure
- Based on number of instances (3 indices)
 - coefficient, PSI, Cosine Similarity
- Based on similarity between rules on a validation dataset (2 indices)
 - GBI, Peculiarity



Learning algorithms for comparisons

- **Decision Tree**: J4.8 (an Java implementation of C4.5)
- **Neural Network**: BPNN (with back-propagation)
 - Parameters of BP: learning rate=0.3 , momentum= 0.2
 - Each unit corresponds to each class label in output layer
- **Classification Via Linear Regression**: CLR
 - Linear regression expressions: “1-the other” for each class label
 - explanatory variable selection: greedy search with AIC
- **SVM**: Sequential Minimal Optimization [Platt98]
 - SVM for multiple class: learning “1-the other” expressions for each class label
 - Kernel function setting: polynomial kernel
- **OneR**
 1. sorting with single objective index
 2. setting thresholds based on class labels
 3. constructs a rule set with the objective index

The Flow of the comparison with the meningitis datamining result [Hatazawa 00]



Sample of the data set

ruleID	Accuracy	Added_Value	...	YulesQ	YulesY	HumanExpert
Rule1	0.81	0.41	...	0.73	0.44	NI
Rule10	0.81	0.43	...	0.75	0.45	NI
Rule11	0.85	0.46	...	0.79	0.49	I
Rule12	0.84	0.56	...	0.87	0.58	I
Rule13	0.94	0.44	...	0.88	0.59	I
Rule14	0.81	0.43	...	0.75	0.45	NI

← 39 objective rule evaluation indices →

Performance Comparison of the five algorithms

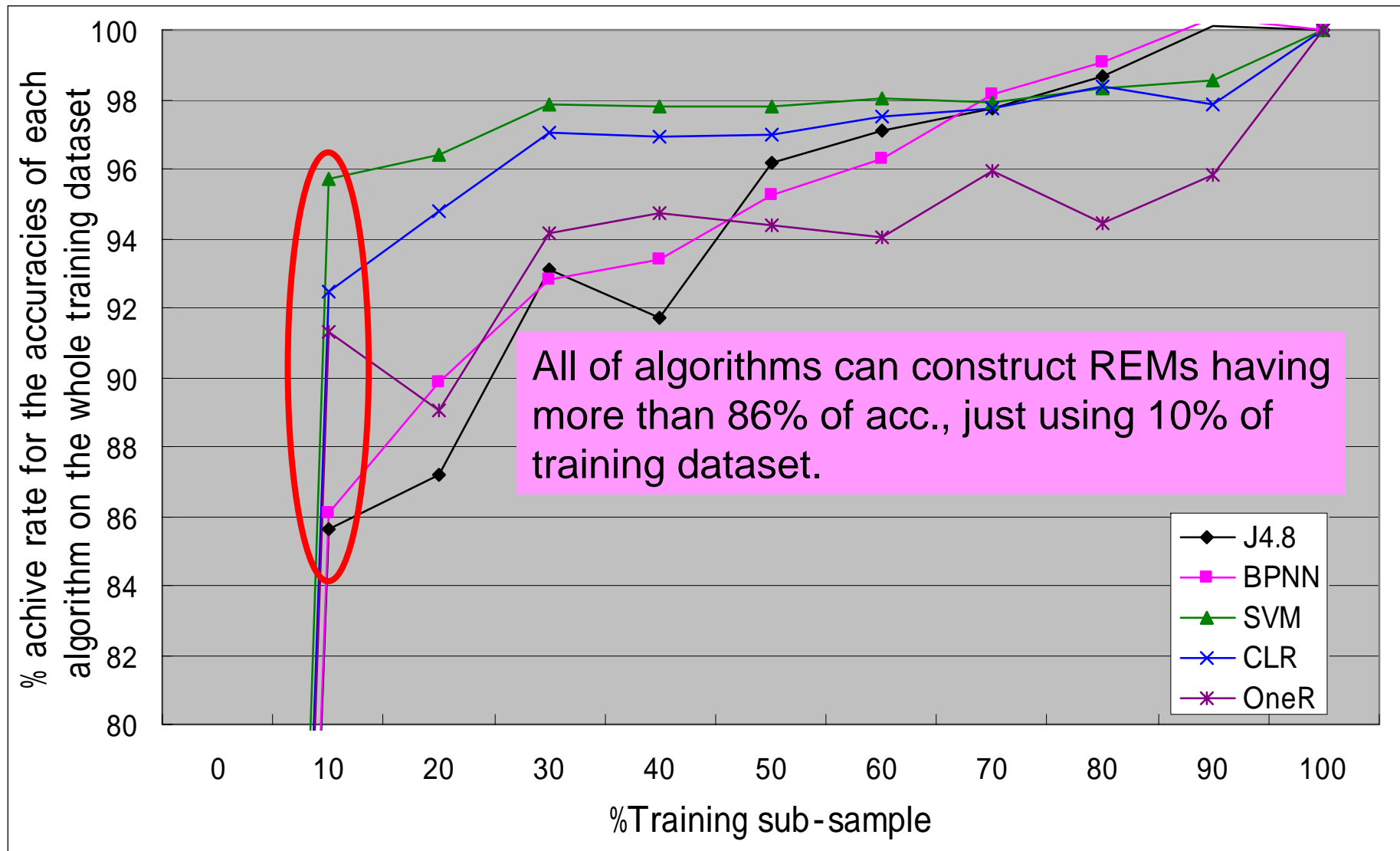
(All of rules =244 , 'I'=48(19.7%), 'NI'=187(76.6%), 'NU'=9(3.7%))

Learning Algorithms	Evaluation on the Whole Training Dataset						
	Acc.	Recall			Precision		
		I	NI	NU	I	NI	NU
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	Evaluation with Leave - One-Out(LOO)						
	Acc.	Recall			Precision		
		I	NI	NU	I	NI	NU
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

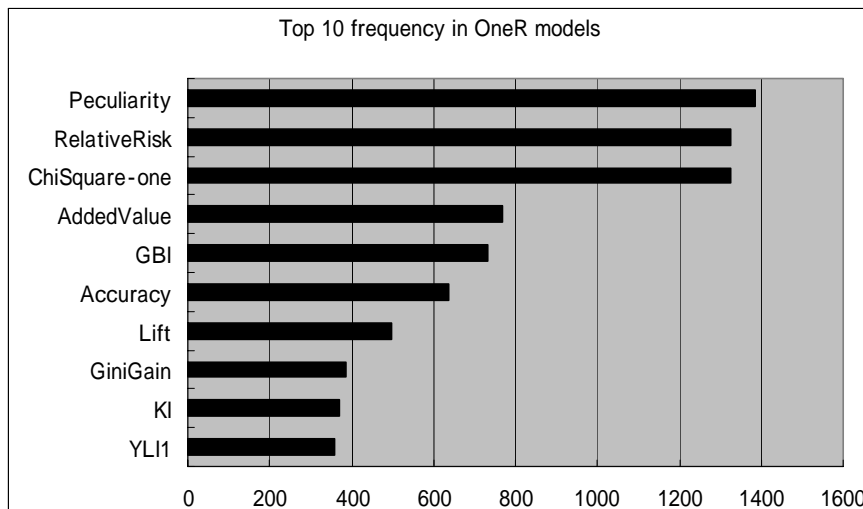
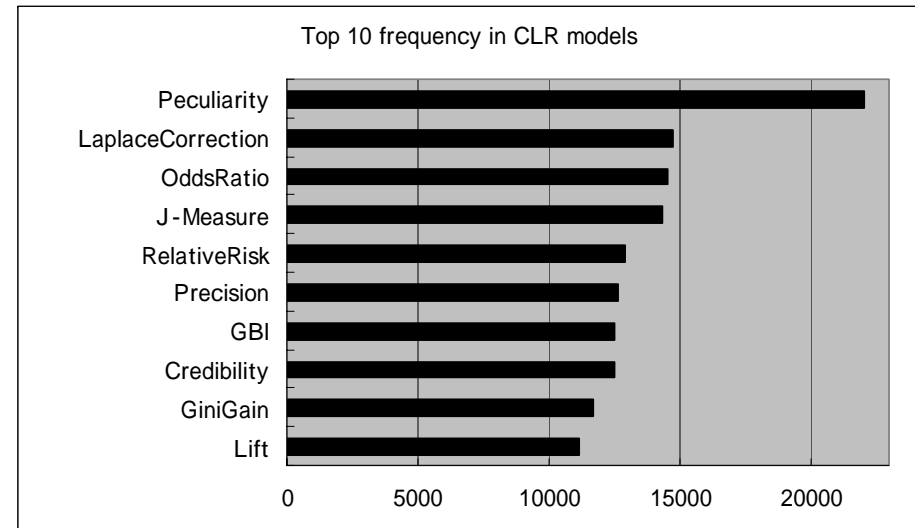
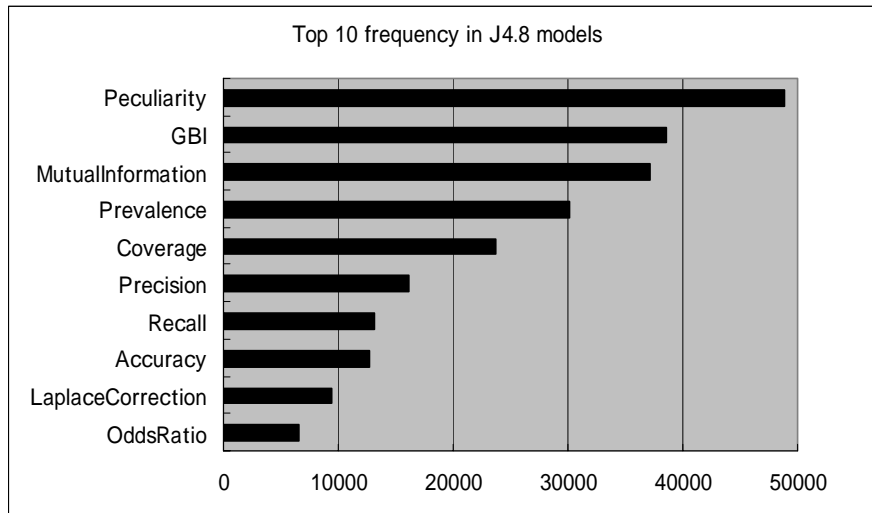
1. J4.8 and BPNN achieve higher than 85.7% of acc. with more than 77.5% reliability. (BPNN tend to be over fitting, looking at it's LOO acc., recalls and precisions)
2. To predict very minor class 'NU', a proper learning algorithm will be needed.

Leaning curves on achieve rates

(achieve rate = (acc. of each sub-sample / acc. of whole sample) * 100)



Contents of Rule Evaluation Models (Statistics of 10,000 bootstrap iterations)



• These models include not only indices which express correctness of rules, but also other kinds of indices such as Peculiarity and GBI.

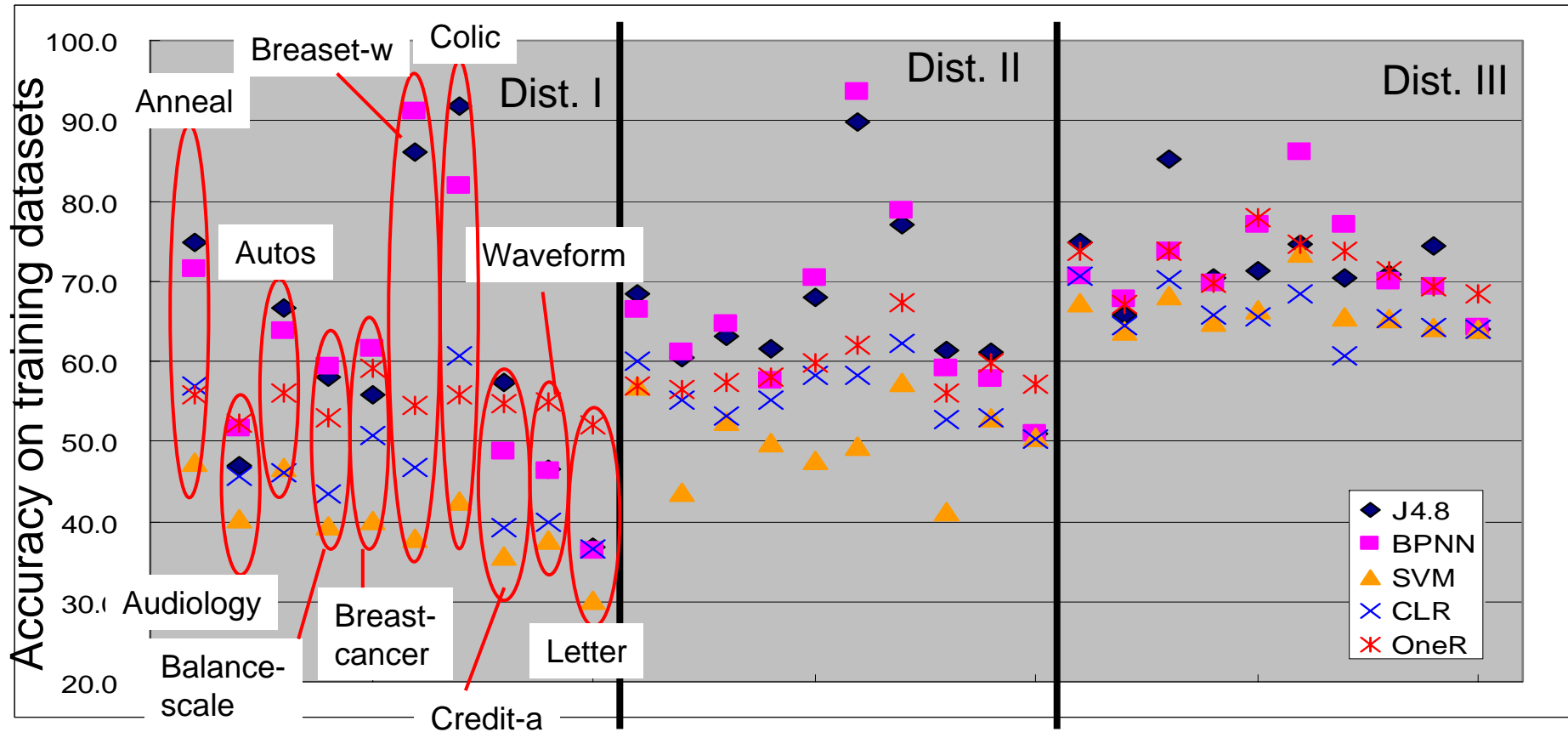
Datasets from rule sets learned with the ten UCI benchmark data

(To make sure the availability of our method without any human criteria)

	#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	824	240	247	310	37.6
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

*All of rule sets are obtained by bagged PART with Weka [Witten 00]

Performances of REMs on the training datasets with three kinds of class distributions



- Performances of algorithms are suffered from probabilistic class distribution especially in larger datasets.
- Hyper-plane type learner (SVM and CLR) could not work well, because of the probabilistic class distributions.

Estimation of minimum training subset to construct valid REMs (from learning curve analysis)

	J48	BPNN	SVM	CLR	OneR
Distribution I					
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					
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					
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

- In Dist. I and II, almost learner succeeded in learning valid REMs with less than 20% of each data set.
 - It is more difficult to construct valid REMs with smaller training subset on 'Distribution III', which has unbalanced class distribution.
- > If we construct REMs without particular human criterion, we should prepare small (<100) dataset with balanced class distribution.



Outline

- Background
- Rule Evaluation Support Method based on Objective Rule Evaluation indices
- Comparisons of Learning Algorithms for Rule Evaluation Model Construction
- Conclusion



Conclusion

■ Summary

- Comparing learning algorithms to construct rule evaluation models for supporting a post-processing of data mining exactly
 - Our method can construct valid rule evaluation models with the 39 objective rule evaluation indices at least the five learning algorithms.
 - The algorithms have been able to construct valid rule evaluation models with 10% of training subset with evaluations based on solid expert's criterion.

■ Future works

- Introducing algorithm selection
 - for attribute construction and attribute selection algorithm
 - for learning algorithm
- Applying this method to other data from other domains