Selection of Probabilistic Measure Estimation Method based on Recursive Iteration of Resampling Methods

Shusaku Tsumoto and Hiroshi Tanaka
Department of Informational Medicine
Medical Research Institute, Tokyo Medical and Dental University
1-5-45 Yushima, Bunkyo-ku Tokyo 113 Japan
TEL: +81-3-3813-6111 (6159), FAX: +81-3-5684-3618
email: {tsumoto, tanaka}@tmd.ac.jp

Abstract

One of the most important problems in rule induction methods is how to estimate the reliability of the induced rules, which is a semantic part of knowledge to be estimated from finite training samples. In order to estimate errors of induced results, resampling methods, such as cross-validation, the bootstrap method, have been introduced. However, while cross-validation method obtains better results in some domains, the bootstrap method calculates better estimation in other domains, and it is very difficult how to choose one of the two methods. In order to reduce these disadvantages further, we introduce recursive iteration of resampling methods(RECITE). RECITE consists of the following four procedures: First, it randomly splits training samples(S_0) into two equal parts, one for new training samples(S_1) and the other for new test samples(T_1). Second, rules are induced from S_1, and several estimation methods, given by users, are executed by using S_1. Third, the rules are tested by T_1, and test error estimators are compared with each estimator. The second and the third procedure are repeated for certain times given by users. Then the estimation method which gives the best estimator is selected as the most suitable estimation method. Finally, we use this estimation method for S_0 and derive the estimators of statistical measures from the original training samples. We apply this RECITE method to three original medical databases, and seven UCI databases. The results show that this method gives the best selection of estimation methods in almost all the cases.

1 Introduction

One of the most important problems in rule induction methods is how to estimate the reliability of the induced results, which is a semantic part of knowledge to be induced from finite training samples. In order to estimate errors of induced results, resampling methods, such as cross-validation, the bootstrap method, are introduced. However, while cross-validation method obtains better results in some domains, the bootstrap method calculates better estimation in other domains, and it is very difficult how to choose one of the two methods.

In order to reduce these disadvantages further, we introduce recursive iteration of resampling methods(RECITE). RECITE consists of the following four procedures: First, it randomly splits training samples(S_0) into two equal parts, one for new training samples(S_1) and the other for new test samples(T_1). Second, rules are induced from S_1, and several estimation methods, given by users, are executed by using S_1. Third, the rules are tested by T_1, and test error is compared with each estimator. The second and the third procedure are repeated for certain times given by users. And the estimation method which gives the best estimator is selected as the most suitable estimation method. Finally, we use this estimation method for S_0.

We apply this RECITE method to three original medical databases, and seven UCI databases. The results show that this method gives the best selection of estimation methods in almost the all cases.

The paper is organized as follows: in section 2, we introduce our rule induction method based on rough sets, called PRIMEROSE. Section 3 shows the characteristics of resampling methods. In section 4, we discuss the strategy of RECITE and illustrate how it works. Section 5 gives experimental results. Finally, in Section 6 we discuss about the problems of our work.

In this paper, it is notable that we apply resampling methods not to gain predictive accuracy of induced rules, but to estimate more accurate statistical measures of induced rules. So our methodology is quite different from ordinary usage of resampling methods in the community of machine learning. However, in the field of statistics, our methodology is more popular than the above usage.

2 PRIMEROSE

In this paper, we use a rule induction system based on rough sets for our experiments. However, our RECITE method is not dependent on rough set model, and we can also apply it to other rule induction systems, such as AQ[12], and ID3[16]. In this section, we briefly introduce PRIMEROSE method,

2.1 Probabilistic Extension of Rough Sets

Rough set theory is developed and rigorously formulated by Pawlak[15]. This theory can be used to acquire certain sets of attributes which would contribute to class classification and can also evaluate how precisely these attributes are able to classify data.

We are developing an extension of the original rough set model to probabilistic domain, which we call PRIMEROSE(Probabilistic Rule Induction Method based on ROugh Sets)[22, 23]. This extension is very similar to the concepts of Ziarko's VPRS model[26, 27, 28].

PRIMEROSE algorithm is executed as follows: first, we calculate primitive clusters which consists of the samples which are supported by the same equivalence relations. Then we remove irredundant attribute pairs from total equivalence relations if they do not affect increasing classification rate, which we call *Cluster-Based Reduction.* Repeating these procedures, we finally get minimal equivalence relations, called *minimal reducts.* These equivalence relations can be regarded as premises of rules, so we derive the above minimal rules by using the above reduction technique. Next, we estimate two statistical measures of the induced rules, by cross-validation method and bootstrap method. Combined these measures with the induced results, we obtain probabilistic rules, whose form is defined in subsection 2.2. For further information on the extension of rough set model, readers could refer to [23, 27, 28].

In this paper, we use some notations used in rough sets, which would make our discussion clearer. For example, we denote a set which supports an equivalence relation R_i by $[x]_{R_i}$ and we call it an *indiscernible set.* For example, if an equivalence relation R is supported by a set {1,2,3}, then $[x]_R$ is equal to {1,2,3} ($[x]_R = \{1, 2, 3\}$). Here we use {1,2,3} as a set of training samples, and each number, say "1", denotes the record number of samples. For example, "3" in {1,2,3} is equal to the samples whose record number is three.

In the context of rule induction, R_i represents the combination of attribute-value pairs, which corresponds to the complexes of the selectors in terms of AQ method[12]. Furthermore, $[x]_{R_i}$ means the set which satisfies such attribute-value relations. This set corresponds to a partial star of AQ methods which supports the complexes of the selectors. For more information on rough sets and on rule induction based on rough sets, readers might refer to [15, 26].

2.2 Definition of Probabilistic Rules

We use the definition of probabilistic measures of diagnostic rules which Matsumura et. al [9] introduce for the development of a medical expert system, RHINOS(Rule-based Headache and facial pain INformation Organizing System). This diagnostic rules, called "inclusive rules" is formulated in terms of rough set theory as follows:

Definition 1 (Definition of Probabilistic Rules) *Let R_i be an equivalence relation and D denotes a set whose elements belong to one class and which is a subset of U. A probabilistic rule of D is defined as a tuple, $< D, R_i, SI(R_i, D), CI(R_i, D) >$ where R_i, SI, and CI are defined as follows.*

R_i is a conditional part of a class D and defined as:

$$R_i \quad s.t. \quad [x]_{R_i} \bigcap D \neq \phi$$

SI and CI are defined as:

$$SI(R_i,D) \;=\; \frac{card\,\{([x]_{R_i} \bigcap D) \bigcup ([x]^c_{R_i} \bigcap D^c)\}}{card\,\{[x]_{R_i} \bigcup [x]^c_{R_i}\}}$$

$$CI(R_i,D) = \frac{card\ \{([x]_{R_i} \bigcap D) \bigcup ([x]^c_{R_i} \bigcap D^c)\}}{card\ \{D \bigcup D^c\}}$$

where D^c or $[x]^c_{R_i}$ consists of unobserved future cases of a class D or those which satisfies R_i,respectively. □

In the above definition, *unobserved future cases* means all possible future cases. So we consider an infinite size of cases, which is called *total population* in the community of statistics.

And SI(Satisfactory Index) denotes the probability that a patient has the disease with this set of manifestations, and CI(Covering Index) denotes the ratio of the number the patients who satisfy the set of manifestations to that of all the patients having this disease. Note that SI(R_i,D) is equivalent to the accuracy of R_i.

For example, let us consider an example of inclusive rules. Let us show an example of an inclusive rule of common migraine(CI=0.75) as follows:

If
history:paroxysmal,
jolt headache:yes,
nature: throbbing or persistent,
prodrome:no,
intermittent symptom:no,
persistent time: more than 6 hours, and
location: not eye,

Then we suspect common migraine (SI=0.9, CI=0.75).

Then SI=0.9 denotes that we can diagnose common migraine with the probability 0.9 when a patient satisfies the premise of this rule. And CI=0.75 suggests that this rule only covers 75 % of total samples which belong to a class of common migraine.

A total rule of D is given by $R = \bigvee_i R_i$, and then total CI(tCI) and total SI(tCI) is defined as: tCI(R,D) = CI($\bigvee_i R_i$,D), and tCI(R,D) = SI($\bigvee_i R_i$,D) respectively.

Since the above formulae include unobserved cases, we are forced to estimate these measures from the training samples. For this purpose, we introduce cross-validation and the Bootstrap method to generate "pseudo-unobserved" cases from these samples as shown in the next subsection.

3 Resampling Estimation Methods

The above equation is rewritten as:

$$\begin{aligned} SI(R_i,D) &= \frac{card\ [x]_{R_i}}{card\ [x]_{R_i} \cup [x]^c_{R_i}} \frac{card\ [x]_{R_i} \cap D}{card\ [x]_{R_i}} \\ &+ \frac{card\ [x]^c_{R_i}}{card\ [x]_{R_i} \cup [x]^c_{R_i}} \frac{card\ [x]^c_{R_i} \cap D^c}{card\ [x]^c_{R_i}} \\ &= \varepsilon_{R_i} \alpha_{R_i} + (1 - \varepsilon_{R_i}) \alpha^c_{R_i} \end{aligned}$$

where ε_{R_i} denotes the ratio of training samples to total population, which consists of both training samples and future cases. α_{R_i} denotes an apparent accuracy, and $\alpha^c_{R_i}$ denotes the accuracy of classification for unobserved cases. This is a fundamental formula of accuracy(SI). Resampling methods focus on how to estimate ε_{R_i} and $\alpha^c_{R_i}$, and makes some assumption about these parameters.

Under some assumptions, we obtain the formulae of several estimation methods. In the following subsections, due to the limitation of the space, we restrict discussion to the main three methods: cross-validation method, the Bootstrap method and 0.632 estimator. Other methods, such as the Jackknife method[2], generalized cross-validation[4], can be also discussed in the same framework.

3.1 Cross-Validation Method

Cross-validation method for error estimation is performed as following: first, the whole training samples $\mathcal{L}$ are split into V blocks: $\{\mathcal{L}_1, \mathcal{L}_2, \cdots, \mathcal{L}_V\}$. Second, repeat for V times the procedure in which we

induce rules from the training samples $\mathcal{L} - \mathcal{L}_i (i = 1, \cdots, V)$ and examine the accuracy α_i of the rules using $\mathcal{L}_i$ as test samples. Finally, we derive the whole accuracy α by averaging α_i over i, that is, $\alpha = \sum_{i=1}^{V} \alpha_i / V$ (this method is called V-fold cross-validation). Therefore we can use this method for estimation of CI and SI by replacing the calculation of α by that of CI and SI, and by regarding test samples as unobserved cases. This method does not use training samples to estimate measures, so in this case, we can regard the following approximation as an assumption in applying this method: If unobserved cases are expected to be much larger than training samples, then the above formulae can be approximated as follows:

$$\mathrm{SI}(R_i, D) \approx \frac{card\ [x]^c_{R_i} \cap D^c}{card\ [x]^c_{R_i}}$$

The main problems of cross-validation are how to choose the value of V and high variability of estimates, or large mean squared errors of the cross-validation estimates. The first problem suggests that, as the value of V increases, estimates get closer to apparent ones and the variance grows to be smaller. We discuss this phenomenon in [22, 23], in which we conclude that the choice of V depends on our strategy. If it is desirable to avoid the over estimation of statistical measures, we can safely choose 2-fold cross-validation, whose estimators are asymptotically equal to predictive estimators for completely new pattern of data as shown in [3, 4]. In order to solve the second problem, recently, repeated cross-validation method is introduced[24]. In this method, cross-validation methods are executed repeatedly(safely, 100 times), and estimates are averaged over all the trials. This iteration makes the variances to be lower as shown in [23, 24]. For detailed information about these problems and methods, readers might refer to [23, 24]. Since our strategy is to avoid the high variabilities, we choose repeated 2-fold cross-validation method and repeated 10-fold cross-validation method in this paper.

3.2 the Bootstrap Method

On the other hand, the Bootstrap method is executed as follows: first, we create empirical probabilistic distribution(F_n) from the original training samples. Second, we use the Monte-Carlo methods and randomly generate the training samples by using F_n. Third, rules are induced by using these newly generated training samples. Finally, these results are tested by the original training samples and statistical measures, such as error rates are calculated. We iterate these four steps for finite times. Empirically, it is shown that about 200 times' repetition is sufficient for estimation.

This method uses training samples to estimate measures, so in this case, we use the above equation in the subsection 3.1. For example, let {1,2,3,4,5} be original training samples. From this population, we make training samples, say {1,1,2,3,3}. The induced result is equivalent to that of {1,2,3}. Since original training samples are used as test samples, {1,2,3} makes an apparent accuracy, and {4,5} makes a predictive estimator for completely new samples, which can be regarded as test samples generated by cross-validation. In this case, ϵ is estimated as 3/5. We repeat this procedure and take the average over the whole results, which makes this estimation more accurate. That is, ϵ_{R_i} and α_{R_i} are estimated by iteration of Monte-Carlo simulations.

Interestingly, Efron[3, 4] shows that estimators by cross-validation are asymptotically equal to predictive estimators for completely new pattern of data, and that the Bootstrap estimators are asymptotically equal to maximum likelihood estimators[3, 4] and are a little overfitting to training samples. This fact can be explained by the above equation, since in the case of the bootstrap method the term of an apparent accuracy, $\epsilon_{R_i} \alpha_{R_i}$ is included. Furthermore, Efron approximates ϵ_{R_i} to be 0.368 by insightful discussion. For further information, readers could refer to [3].

3.3 0.632 estimator

According to the Efron's explanation, the above estimator of accuracy is asymptotically equal to :

$$SI(R_i, D) \approx 0.368 \alpha_{R_i} + 0.632 \alpha^c_{R_i}.$$

Furthermore, Efron proved that the first term is approximately equal to the apparent accuracy of the bootstrap method and that the second term is approximately equal to the estimator derived by 2-fold cross-validation.

4 RECITE

4.1 Strategy of RECITE

There are many reports on these estimation methods and on their extensions by the researchers in the community of statistics, which are reviewed in [5, 11]. However, since each performance is different in each paper, it is very difficult to determine which method should be selected.

Each of these methods represents some factors which are important for estimation, part of which we discuss in the above subsection. And it is pointed out that these performances may depend on applied domains. For example, when a domain is noisy and completely new test samples are expected to obtain, a predictive estimator for completely new pairs, such as cross-validation estimator, is preferable.

However, in general, one may want to estimate statistical measures without domain knowledge, or domain-specific knowledge may not be applicable.. One way to solve this problem is to use all methods, and to obtain the range of estimators. For example, if the value of an estimator is given as 0.97 by the Bootstrap, 0.82 by 10-fold cross-validation, then a true value is in the interval between 0.8 and 0.97. If we have knowledge about the distribution of probability density, for example, the target variable's distribution should be normal, then we can also use a formula of asymptotic error rate.

However, if one may want to select the best estimator, instead of the interval, we introduce the other way to select one method from considerable resampling methods. Therefore, this way to solve this problem is to select the best estimation method by using subsets of training samples. For example, if we have training samples, say {1,2,3,4,5,6,7,8,9,10}, then first, we split it into new training samples, say {1,3,5,7,9}, and new test samples, {2,4,6,8,10}. Using new training samples, estimators are calculated, and compared with the result by the new test samples. And then we selected a method whose estimator is close to the test estimator. For example, if test estimator is 0.95, with the bootstrap estimator 0.97, and with the cross validation estimator 0.82, then the bootstrap method is selected as the best estimation method. It may depend on splitting, so we should iterate these procedures for certain times, say 100 times. Then we calculate several statistics over these 100 trials, such as average, variance, and t-statistics.

In this procedure, we assume that the results by subsets can be used as estimation of original samples. This assumption is closely related with cross-validation method, because cross-validation method calculates estimators by using subsets of original samples, and regard them as estimators of original training samples. In other words, estimators by training samples are assumed to be not less than ones by subsets of these samples.

This assumption is concerned with the problem of sampling bias in the field of statistics[11]. The main point of sampling bias is that if original training samples are suitably sampled from population, then the results of these samples are asymptotically equal to those by using total population. Therefore sampling from these training samples, if not biased, gives the same result. And the performance of cross-validation empirically suggests that this assumption be true[4, 8, 11]. We will discuss later about this assumption.

In this paper, we assume that this assumption is true and we introduce RECITE method based on the latter strategy.

4.2 Algorithm for RECITE

Algorithms for RECITE can be derived by embedding a rule induction method and estimation methods into the algorithms shown as follows. An algorithm for RECITE is described as follows:

INPUTS: S_0: Training Samples
B_i: Repeated Times
L_i: List and Subprocedures of Estimation Methods
OUTPUTS: M_i: the Best Estimation Method

1) Counter is set as 0 ($i := 0$).
2) Randomly split training samples(S_0) into two equal parts, one for new training samples(S_1) and the other for new test samples(T_1).
3) Rules are induced from S_1, and several estimation methods of statistical measures (SI and CI), given by users, are executed by using S_1 (S_1 **estimation**).

Table 1: Information of Monk's Databases

Domain	Test Sample	Training Samples	Classes	Attributes
Monk-1	432	124	2	6
Monk-2	432	169	2	6
Monk-3	432	122	2	6

4) The induced rules are tested by T_1 (T_1 **estimation**), and test estimator of statistical measures is compared with each estimator.
5) Increment the counter ($i := i + 1$). If the counter is less than the upper bound($i < B_i$), goto 2). If not, goto 6).
6) Select the best estimation method whose estimators are the nearest to test estimators by using the statistics of Student t-test as the distance between test estimators and the estimators derived by resampling methods.
7) Output this estimation method (M_1), and quit this procedure.

In the step of estimation, we calculate several fundamental statistics, such as average, mode, variances, and t-statistics. And t-statistics are obtained by these fundamental statistics.

4.3 Examples

Here, we illustrate how the RECITE algorithm shown in the above subsection works by applying to Monk's three problems[21] in UCI databases [14]. The Monk's three problems are introduced in order to compare the existing machine-learning methods. Monks-1,2 and 3 consist of training samples, whose sizes are 124, 169 and 122, respectively, and test samples whose sizes are all 432(Table 1). The reason why we choose these problems is that each problem focuses on different problems of machine learning methods and that test samples are clearly given. RECITE procedures are executed as follows. First, we split training samples(S_0) into S_1 and T_1, both of which are composed of 62, 85, 61 samples. Second, and rules are induced from S_1 and the following resampling methods are executed after the induction: repeated 2-fold cross-validation, repeated 10-fold cross-validation, the Bootstrap method, 0.632 estimator. Third, the induced rules are tested by T_1. We repeat these procedures for 100 times, and the statistics of these estimators are calculated as shown in Table 2.

The second column of this table shows test estimators by T_1. From the third to sixth column, we give estimators derived by repeated 2-fold cross-validation (2-fold CV), repeated 10-fold cross-validation (10-fold CV), the Bootstrap method (BS), and 0.632 estimator (0.632), respectively. In this table, they are shown as S_1 estimators.

From these statistics, we select the best estimation method by using the Student t-test statistics as the distance between test estimators and the estimators derived by resampling methods. In Table 2, we show T_1 estimators and these best estimators in the **bold** characters.

As shown in this table, 10-fold cross-validation, 0.632 estimator, and 2-fold and 10-fold cross-validation are the best for Monks-1, 2, and 3 problems, respectively. Hence we use 10-fold cross-validation for Monk-1, 0.632 estimator for Monk-2, and 2-fold cross-validation for Monk-3 as an estimation method for S_0. The comparisons of the estimators with test estimators are shown in Table 3. The first column shows estimators derived by test samples, whose size are equal to 10, and which are given independently in these databases, as mentioned above. In the second and third columns, estimators by original training samples(S_0) and the estimation methods are given respectively. Finally, the fourth and the fifth column show the best estimators and the best estimation methods for S_0. For example, in the case of Monk-1, while the estimation method derived by RECITE is repeated 10-fold cross-validation, the best estimation method is the Bootstrap method, although this difference is not significant in terms of t-test. On the other hand, as to the other two methods, the selected methods coincide with the best estimation methods.

Table 2: Estimation of Accuracy(SI) by using subsets(S_1 and T_1)

		T_1 Test	S_1 Estimators			
Domain	Samples	Estimator	2-fold CV	10-fold CV	BS	0.632
Monk-1	124	**85.0±1.5**	72.8±1.3	**84.2±1.0**	91.2±0.5	82.9±1.2
Monk-2	169	**64.6±2.0**	56.0±2.2	56.6±1.9	84.8±0.7	**72.1±2.0**
Monk-3	122	**91.8±0.6**	**91.9±0.7**	91.7±0.8	92.5±0.5	94.8±0.6

Table 3: Experimental Results of Monk's problems

	Test	S_0	S_0	the best	the best
Domain	estimator	estimator	method	estimator	method
Monk-1	100.0	95.7	10-fold CV	96.3	BS
Monk-2	80.1	77.8	0.632	77.8	0.632
Monk-3	92.1	91.8	2-fold CV	91.8	2-fold CV

Test estimator is derived by using 432 test samples, which are given independently.

5 Experimental Results

We apply this RECITE method to three original medical databases, which were collected by us, and four UCI databases, which consists of lymphography, primary cancer, breast cancer, and breast cancer from Wisconsin.

For estimation methods, we use repeated 2-fold, repeated 10-fold cross-validation, the Bootstrap method, and 0.632 estimator.

Unfortunately, in these databases, test samples are not given independently. So we first have to generate test samples from the original training samples. to evaluate our RECITE methods in the same way as evaluation shown in subsection 4.3. We first randomly split given samples into training samples(S_0) and test samples(T_0). This T_0 corresponds to test samples of Monks problems, and S_0 corresponds to training samples of Monks problems.

Then we apply RECITE method to new training samples. We repeat this splitting procedure for 100 times in order for the effect of random sampling to be small.

The precise information of databases is given in Table 4. In these databases, primary tumor domain, meningitis, and CVD have many missing values in some of the cases. And in the case of primary tumor domain, breast cancer, meningitis, and CVD, the sets of attributes are incomplete as they are not sufficient to get complete classification of each data. That is, although some samples have the same combination of attribute–value pairs, classes assigned to those samples are different. Moreover,

These data are incomplete, and include many inconsistencies.

The results of S_1 estimation are given in Table 5. The notations of this table are the same as those of Table 2. Using this S_1 estimation, the best estimation methods are selected as shown in Table 6. The first column gives estimators derived by test samples T_0, generated by splitting of the original training samples. In the second and third columns, estimators by training samples(S_0), which is generated by splitting of the original training samples, and the estimation methods are given respectively. Finally, the fourth and the fifth column show the best estimators and the best estimation methods for S_0.

In all the cases, the selected methods coincide with the best estimation methods, and furthermore, the derived estimators are very close to test estimators.

6 Related Works

In order to select estimation methods, we introduce recursive iteration of resampling methods(RECITE). As shown in the experimental results, RECITE gives the best estimation method for each database without using domain knowledge. This method is based on the assumption that sampling bias of our

Table 4: Information of Databases

Domain	Samples	Classes	Attributes
primary tumor	339	23	17
lymphography	148	4	18
breast cancer	286	2	9
breast cancer (from Wisconsin)	699	2	10
headache	232	10	20
meningitis	198	3	25
CVD	261	6	27

Table 5: Estimation of Accuracy(SI) by using subsets(S_1 and T_1)

	T_1 Test	S_1 Estimators			
Domain	Estimator	2-fold CV	10-fold CV	BS	0.632
primary tumor	**56.8±1.2**	42.4±1.5	51.4±1.4	**53.0±0.4**	63.5±1.0
lymphography	**71.2±1.4**	61.3±1.6	**71.6±1.3**	88.0±0.6	75.5±1.0
breast cancer	**63.2±1.5**	**61.8±2.0**	73.2±2.5	85.4±0.2	75.8±1.2
breast cancer (from Wisconsin)	**94.0±1.0**	89.6±2.0	96.9±2.1	97.0±0.2	**93.4±1.2**
headache	**78.6±2.4**	65.4±3.0	81.1±2.1	95.6±0.2	**78.1±2.1**
meningitis	**77.1±1.9**	62.4±2.9	**77.3±1.6**	86.2±0.1	76.2±1.4
CVD	**74.1±2.5**	59.3±3.4	69.2±2.7	92.1±0.1	**74.2±2.6**

resampling procedures is negligible. Moreover, the experimental results show that this assumption is also effective, at least for medical domain.

Our research is motivated by the following three works. The first one is Breiman's work [1] which introduce cross-validation method to induce tree-formed discriminant rules for CART. The second is Walker's work [24] which introduce repeated cross-validation method to solve the high variance of cross-validation. Finally, the third one is Schaffer's work which applies cross-validation in order to select classification methods.

In the following subsection, we discuss the relation between each work and our approach.

6.1 Breiman's CART

The problem of the number of folds in cross-validation is first discussed by Breiman et al. [1]. They use 10-fold cross-validation to estimate pruned tree size and its error rate in CART. They point out that 10-fold cross-validation gave adequate accuracy, while 2 and 5-fold cross-validation also gave sufficient accuracy in some examples. Furthermore, they stress that they have not come across any situations where taking the number of fold larger than 10 gave a significant improvement. This is why users of cross-validation persist in using 10-fold cross-validation. However, it should be noted that Breiman et al. do not regard 10-fold as the best. They discuss that 10-fold cross-validation is the most stable empirically for estimation of a pruned tree size.

In our experience, in some domain, 10-fold cross-validation gives a estimator close to the bootstrap estimator, and 2-fold cross-validation performs better than 10-fold. In other words, 10-fold cross-validation sometimes generates overfitting estimation. So we should select these two methods by some criterion. We adopt RECITE method to choose 10-fold cross-validation and 2-fold cross-validation as shown in the above section on experimental results.

Table 6: Experimental Results

Domain	T_0 Test estimator	S_0 estimator	S_0 method	the best estimator	the best method
primary tumor	64.1	61.8	BS	61.8	BS
lymphography	74.4	73.2	10-fold CV	74.6	2-fold CV
breast cancer	65.9	63.2	2-fold CV	63.2	2-fold CV
breast cancer (from Wisconsin)	94.6	96.2	0.632	96.2	0.632
headache	89.6	89.4	0.632	89.4	0.632
meningitis	85.0	84.2	10-fold CV	84.2	10-fold CV
CVD	82.5	81.4	0.632	81.4	0.632

T_0 Test estimator is derived by applying test samples T_0 to the results induced by training samples S_0. Both T_0 and S_0 are randomly generated by splitting original training samples into T_0 and S_0. We repeat this splitting for 100 times, and the above test estimators are derived by averaging those estimators over 100 trials.

6.2 Walker's Repeated Cross-Validation

Walker performs Monte Carlo simulation to evaluate the conditional probability estimation methods[24, 25]. For simulations, he uses data from 14 distributions, with one to eight dimensions, continuous and categorical variables, noise variables, sample sizes from 50 to 1000 data points, and Bayes' error rates ranging from 0.01 to 0.4. They adopt eight estimation methods: resubstitution, bootstrap 0.632, repeated cross-validation, Breiman's method and Breiman's method with repeated cross-validation, and Breiman's method with the bootstrap 0.632. Note that his bootstrap 0.632 is different from our 0.632 estimator. As shown in [24], estimators derived by bootstrap test samples is used for estimation of $\alpha^c_{R_i}$, while we use estimators derived by 2-fold cross-validation.

His conclusion is that for sample size of 200 or less, Breiman's method with repeated cross-validation, or with the bootstrap 0.632 is generally accurate than the other estimates and that for sample size of 500 or 1000, the Breiman's method with repeated cross-validation is generally the most accurate.

As shown in his results, it is very difficult to say which method gives the best result for given samples. While repeated cross-validation performs better in some domains, the 0.632 estimator performs better in other domains. So if one may want to select one from these methods, one should introduce a selection method. Our RECITE method will be one approach for selection.

Furthermore, it is notable that our method can be applied to other rule induction methods, such as AQ[12], CART[1], and C4[17]. Even in the same domain, the best estimation method may be different when we use different method. For example, in some cases, it may be true that the bootstrap is the best for CART, while 10-fold cross-validation is the best for AQ. Our RECITE method can also deal with such complicated cases, since our method is only based on recursive iteration of resampling, and is not on specific algorithms.

6.3 Schaffer's Selection of a Classification Method

Schaffer introduces cross-validation to select the classification method best for some domain without using domain knowledge [19].

He gives three constituent strategies: ID3 [16], C4 [17], and Back Propagation [10], and introduces the fourth strategy, called CV, which conducts a 10-fold cross-validation using training data to compare the three constituent strategies. For each fold, the best strategies are selected. And this selection is repeated for 10 times. For example, in the case of Glass databases, CV chooses ID3 tree method for 6 times, and C4 rule induction method for 4 times. The results shows this CV performs better than a single classification method in average. Finally he concludes that cross-validation may be seen as a way of applying partial information about the applicability of alternative classification strategies.

This method is also based on the assumption mentioned in Section 4. That is, the results induced by subsets reflects the results induced by the original samples. As shown in his paper[19], his results also suggest that this assumption be true. Furthermore, he points out that this assumption is closely

related with the performance of cross-validation, which is precisely discussed in [18]. We will discuss about this assumption in the next subsection.

The main difference between Schaffer's method and ours is that we use recursive iteration of resampling methods: we generate new training samples(S_1) and test samples(T_1) from original training sample(S_0). And we apply estimation methods to S_1, and test the induced results by T_1. On the other hand, Schaffer only uses training samples for selection, and does not test his results by using test samples. Although CV's superiority to each of the constituent strategies is significant at above the .999 level, using a one-sided paired t test, it may not be true when test samples are applied. For example, in the case of Image databases, while CV chooses ID3 tree method for 10 times, C4 may perform better than ID3 tree method for newly derived samples.

Therefore, we can apply our concepts of RECITE in order to strengthen this Schaffer's procedure. First, we generate S_1 and T_1 from original training samples. Then Schaffer's method is applied to S_1. Finally, we induce classification rules by S_1 and compare the estimated accuracy with test estimators derived by T_1. If the selected method is not the best, then this selection method is not suitable for this domain.

6.4 Schaffer's Overfitting Avoidance as Bias

Schaffer stresses that any overfitting avoidance avoidance strategy, such as pruning methods, amounts to a form of bias[18]. Furthermore, he clearly explains this result from the viewpoint of information theory. As one of the pruning methods, he also discusses about cross-validation, and points out that the main idea of this strategy is "A bias is as good as it is appropriate". This is exactly the same idea as ours in the subsection 4.1. In terms of statistical theory, this assumption is closely related with **sampling bias**. As mentioned above, the main point of sampling bias is that if original training samples are suitably sampled from population, then the results of these samples are asymptotically equal to those by using total population. Therefore sampling from these training samples, if not biased, gives the same result.

In the field of statistics, these ideas are applied to studying the effectiveness of the Bootstrap sampling[5], since its sampling procedure is based on Monte Carlo simulation, which is rigorously studied in mathematics. The idea of the Bootstrap method is also captured formulated by the Edgeworth expansion[6], since the idea of this sampling is easy to formulate in terms of Monte Carlo simulation.

On the other hand, cross-validation is not studied in this way. Because procedures of cross-validation method are algebraic, and they seem not to have analytic aspects like the Bootstrap. Therefore, from the viewpoint of sampling bias, cross-validation is different from the Bootstrap method, and, from the results of experimental results in [8, 24], it can be concluded that cross-validation sampling generates higher biases than the Bootstrap sampling. Moreover, in our experiences, we feel that higher biases cause high variances of estimators derived by cross-validation.

Schaffer's points are very important, and the research on cross-validation should be started from his results. Unfortunately, statisticians do not study in this direction, maybe because cross-validation sampling is difficult to deal with in a rigorous way like the Bootstrap. This direction towards the problems of sampling bias of cross-validation is a main future research, which may give the justification of applying this method.

7 Conclusion

One of the most important problems in rule induction methods is how to estimate the reliability of the induced results, which is a semantic part of knowledge to be induced from finite training samples.

In order to estimate errors of induced results, resampling methods, such as cross-validation, the bootstrap method, are introduced. However, while, in some domains, cross-validation method obtains better results, in other domains, the bootstrap method calculates better estimation, and it is very difficult how to choose one of the two methods. In order to solve this problem, we introduce recursive iteration of resampling methods(RECITE). We apply this RECITE method to three original medical databases, and seven UCI databases. The results show that this method gives the best selection of estimation methods in almost the all cases.

Acknowledgements

The authors would like to thank Dr. Patrick M. Murphy and Dr. David W. Aha for giving them UCI databases. The authors would also like to thank Dr. M. Zwitter and Dr. M. Soklic for providing primary tumor, lymphography and breast cancer databases, and to thank Dr. William H. Wolberg for donating the breast cancer databases of the University of Wisconsin. This research is supported by Grants-in-Aid for Scientific Research No.04229105 from the Ministry of Education, Science and Culture, Japan.

References

[1] Breiman, L., Freidman,J., Olshen,R., and Stone,C. *Classification And Regression Trees.* Belmont,CA:Wadsworth International Group, 1984.

[2] Efron, B. *The Jackknife, the Bootstrap and Other Resampling Plans.* Pennsylvania: CBMS-NSF, 1982.

[3] Efron, B. Estimating the error rate of a prediction rule: improvement on cross validation. *J.Amer.Statist.Assoc.* **78**, 316-331, 1983.

[4] Efron, B. How biased is the apparent error rate of a prediction rule ? *J. Amer. Statist. Assoc.* **82**, 171-200, 1986.

[5] Efron, B. and Tibshirani, R. *An Introduction to the Bootstrap.* Chapman and Hall, London, 1994.

[6] Hall, P. *The Bootstrap and Edgeworth Expansion*, Springer Verlag, New York, 1992.

[7] Katzberg, J.D. and Ziarko, W. Variable Precision Rough Sets with Asymmetric Bounds, *Proceedings of RSKD-93*,in this issue.

[8] Konishi, S. and Honda, M. Comparison of procedures for estimation of error rates in discriminant analysis under nonnormal populations. *J, Statist. Comput. Simul.*, **36**, 105-115, 1990.

[9] Matsumura, Y., et al., Consultation system for diagnoses of headache and facial pain: RHINOS, *Medical Informatics*, **11**, 145-157, 1986.

[10] McClelland, J.L., and Rumelhart, D.E. *Explorations in parallel distributed processing.*, MIT Press, Cambridge, MA, 1988.

[11] McLachlan, G.J., *Discriminant Analysis and Statistical Pattern Recognition,* John Wiley and Sons, New York, 1992.

[12] Michalski, R.S., A Theory and Methodology of Machine Learning. Michalski, R.S., Carbonell, J.G. and Mitchell, T.M., *Machine Learning - An Artificial Intelligence Approach*, Morgan Kaufmann, Palo Alto, CA, 1983.

[13] Michalski, R.S., et al. The Multi-Purpose Incremental Learning System AQ15 and its Testing Application to Three Medical Domains, in: *Proceedings of AAAI-86*, 1041-1045, AAAI Press, Palo Alto, CA, 1986.

[14] Murphy, P.M. and Aha, D.W. *UCI Repository of machine learning databases* [Machine-readable data repository]. Irvine, CA: University of California, Department of Information and Computer Science.

[15] Pawlak, Z., *Rough Sets*, Kluwer Academic Publishers, 1991.

[16] Quinlan, J.R., Induction of decision trees, *Machine Learning*, **1**, 81-106, 1986.

[17] Quinlan, J.R. Simplifying Decision Trees. *International Journal of Man-Machine Studies*, **27**, 221-234, 1987.

[18] Schaffer,C. Overfitting Avoidance as Bias. *Machine Learning*, **10**, 153-178, 1993.

[19] Schaffer,C. Selecting a Classification Method by Cross-Validation. *Machine Learning*, **13**, 135-143, 1993.

[20] Slowinski, K. et al., Rough sets approach to analysis of data from peritoneal lavage in acute pancreatitis, *Medical Informatics*, **13**, 143-159, 1988.

[21] Thrun, S.B. et al. The Monk's Problems- A performance Comparison of Different Learning algorithms. Technical Report CS-CMU-91-197, Carnegie Mellon University, 1991.

[22] Tsumoto,S.and Tanaka,H. PRIMEROSE:Probabilistic Rule Induction based on Rough Sets, *Proc. of RSKD'93*, 1993.

[23] Tsumoto,S.and Tanaka,H. PRIMEROSE:Probabilistic Rule Induction based on Rough Set Theory, *TMD-IM-TR 93-001*, 1993.

[24] Walker, M.G. and Olshen, R.A., Probability Estimation for Biomedical Classification Problems. *Proceedings of SCAMC-92*, McGrawHill, New York, 1992.

[25] Walker, M.G. Probability Estimation for Classification Trees and DNA Sequence Analysis. Technical Report STAN-CS-92-1422, Stanford University, 1992.

[26] Ziarko, W., The Discovery, Analysis, and Representation of Data Dependencies in Databases, in:*Knowledge Discovery in Database*, Morgan Kaufmann, Palo Alto, CA, pp.195-209, 1991.

[27] Ziarko, W., Variable Precision Rough Set Model, *Journal of Computer and System Sciences*, **46**, 39-59, 1993.

[28] Ziarko, W., Analysis of Uncertain Information in the Framework of Variable Precision Rough Sets, *Foundation of Computing and Decision Science*, **18**, 381-396, 1993.