The Effect of Subsampling Rate on S^3Bagging Performance

Masahiro Terabe TERABE@MRI.CO.JP
Mitsubishi Research Institute, Inc., 2-3-6 Ohtemachi, Chiyoda, Tokyo 100-8141, Japan

Takashi Washio WASHIO@SANKEN.OSAKA-U.AC.JP
Hiroshi Motoda MOTODA@SANKEN.OSAKA-U.AC.JP
I.S.I.R., Osaka University, 8-7 Mihogaoka, Ibaraki, Osaka 567-0047, Japan

Abstract

Fast classifier induction from large data set is one of the main issues in the field of data mining. One of the approaches on this issue is to reduce the training data size by subsampling. In many cases, the accuracy of the induced classifiers becomes worse when the training data is subsampled. We proposed S^3Bagging (**S**mall **S**ub**S**ampled **Bagging**) that adopts both subsampling and Bagging. The main parameter of S^3Bagging is subsampling rate. The subsampling rate determines the performance of S^3Bagging, i.e., the prediction accuracy of induced classifier and the processing time. In this paper, the effect of subsampling rate on the performance of S^3Bagging is investigated by carefully designed experiments.

1. Introduction

In the data mining process, it is often necessary to induce classifiers iteratively until the human analysts complete extracting valuable knowledge from data. Therefore, the data mining tools need to extract valid knowledge from a large amount of data quickly enough in response to the human demand. Many researchers investigated on the methods for fast induction of classifiers from large data sets. The main approaches are classified into the following types[8].

- Design a fast algorithm: This approach includes a wide variety of algorithm design techniques for reducing the asymptotic complexity, for optimizing the search and representation, for finding approximate solutions, and so on.
- Partitioning the data: This approach involves breaking the data set up into subsets, learning from one or more of the subset, and possibly combining the results.

The former approach includes developing the heuristics which are effective to restrict the search space of classifier and parallel processing method. The latter approaches include the instance selection and subsampling.

In many cases, the purpose of proposed instance selection methods is to select instances which are informative as training data for inducing classifier. Adopting the instance selection method to the training data often improves the prediction accuracy of classifier. However, most of the instance selection methods need much processing time for selecting instances.

Subsampling selects the instances from the training data set to reduce the number of instances of training data set. Simple sampling technique such as random sampling is often adopted so that the process is not time consuming. By reducing the training data size, the classifier can be induced in shorter time. However, the prediction accuracy becomes worse when the training data is subsampled.

In the field of machine learning, many researchers investigated committee learning method for improving the prediction accuracy of a classifier[17]. In the framework of committee learning, multiple training data sets where the composition of instances is varied from that of the original training data set are prepared by sampling. Thus, a committee classifier is constituted by the multiple committee member classifiers induced from each training data set. The committee classifier makes the final prediction through the committee voting.

Boosting[13] and Bagging[1] are the major methods of the committee learning. On the prediction accuracy, Boosting is often superior to Bagging. However, Boosting needs to induce the member classifiers sequentially so that parallel processing is not available. Additionally, it is well known that Boosting tends to be influenced by the noise in training data comparing with Bagging. On the other hands, Bagging can improve the prediction accuracy generally speaking even if the training data includes noise. Furthermore, the committee member classifiers can be induced concurrently so that the committee classifier can be induced in almost the same processing time with the time inducing one committee member classifier if the parallel processing would be available.

We proposed S^3Bagging (**S**mall **S**ub**S**ampled **Bagging**) that adopts both subsampling and a method of committee learning, i.e., Bagging[15]. S^3Bagging is a method which can induce classifier efficiently by reducing the training data size by subsampling and parallel processing. Additionally, the accuracy of the induced classifier is maintained by aggregating the classification results of each classifier through the Bagging process. The performance of S^3Bagging is evaluated by the prediction accuracy of induced classifier and the processing time. The performance of S^3Bagging is determined by the parameter, subsampling rate. In this paper, we investigated the effect of subsampling rate on the performance of S^3Bagging by carefully designed experiments.

2. S^3 Bagging

In this section, we explain S^3Bagging that adopts both subsampling and a method of committee learning, i.e., Bagging. The process of S^3Bagging is depicted in Figure 1.

First, T training data sets are prepared by subsampling from the original training data set D. The size of each subsampled training data sets D_t $(1 \leq t \leq T)$ is r% of that of the original training data set. The sampling by replacement is adopted as subsampling technique. The subsampling rate r is less than 100% so that the size of subsampled training data set $|D_t|$ is smaller than that of the original training data set $|D|$. Next, a committee member classifier (abbreviated here as member classifier) C_t is induced from each of the training data set D_t. The committee classifier C is constituted by the member classifiers C_t. The committee classifier C is the classifier which is induced by S^3Bagging. The committee classifier predicts the class of a test instance based on the class prediction by the member classifiers. Specifically, the committee classifier predicts the most frequent class predicted by member classifiers. Here, it must be noted that each member classifier can be induced concurrently because each

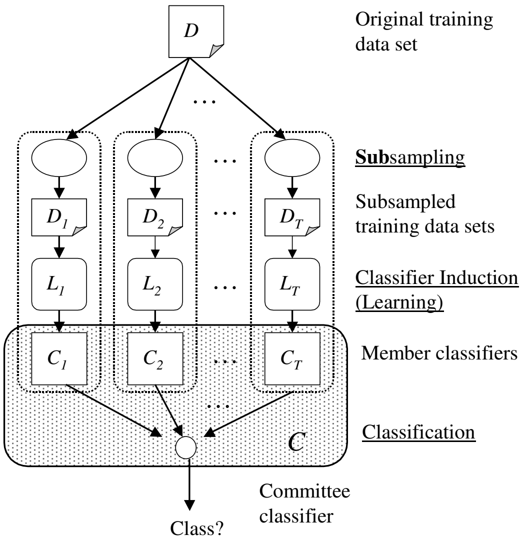

Figure 1. The process of S^3Bagging.

process of subsampling and inducing member classifier is performed independently. Therefore, the processing time to induce committee classifier C, $proc\_time(C)$ is defined as $proc\_time(C) = \max_t proc\_time(C_t)$, where the processing time to induce each member classifier C_t is $proc\_time(C_t)$. The whole processing time to induce committee classifier is almost equivalent to that of inducing each member classifier. Since subsampling is performed, the processing time for inducing committee classifier can also be shortened compared with the case where all original training data are used. S^3Bagging has the features of subsampling and Bagging. In S^3Bagging, the processing time can be shorten by subsampling and parallel processing in member classifier induction. Additionally, the prediction accuracy is improved by Bagging process in the committee classifier.

The main parameter by which S^3Bagging is characterized is subsampling rate r. When the subsampling rate is set large, the prediction accuracy becomes better although the processing time for classifier induction will increase. On the other hand, if the subsampling rate is set small, the prediction accuracy will become worse although the processing time of the classifier induction will be shortened.

In the next section, we show the experimental results. The experiments are conducted to confirm the effect of subsampling rate on the processing time for inducing classifier and the prediction accuracy of the induced classifier.

Table 1. The specification of data sets for experiment 1.

Data set	# of instances	Attribute		Class
		Cont.	Disc.	
BREAST-W	699	9	–	2
CREDIT-A	690	6	9	2
CREDIT-G	1,000	7	13	2
HEPATITIS	155	6	13	2
PIMA	768	8	–	2
SICK	3,772	7	22	2
TIC-TAC-TOE	958	–	9	2
VOTE	435	10	16	2

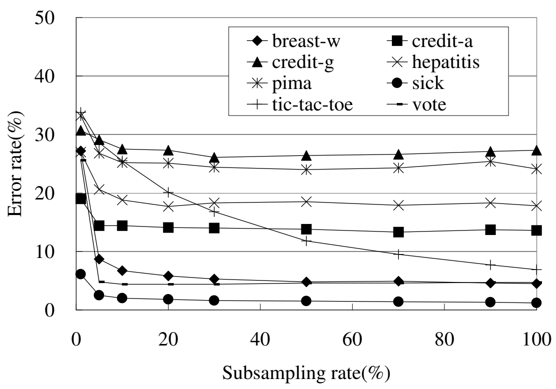

Figure 2. The experimental result: error rate of the committee classifier (%).

3. Experiment

3.1 Conditions of Experiments

To confirm the effect of subsampling rate on the performance of classifier induced by S^3Bagging, the following experiments are conducted. In the experiments, we use C4.5 Release 8 for decision tree induction[10]. All of the functional options of C4.5 are set to defaults. The pruned decision trees are used for evaluation. In Bagging and S^3 Bagging, the number of member classifiers T is set 10 as a general setting[11].

In this experiments, the following five kinds of classifiers are induced for comparison.

- **S^3 Bagging (committee):** The committee classifier induced by S^3Bagging,
- **S^3 Bagging (member):** The member classifiers which constitute the committee classifier induced by S^3Bagging,
- **Bagging (committee):** The committee classifier induced by Bagging,
- **Bagging (member) :** The member classifiers which constitute the committee classifier induced by Bagging,
- **All :** The classifier induced from all of the original training data. This is a usual induction method so that the performance of the classifier and the processing time for inducing the classifier are the baseline of the comparison.

Each experimental result of member classifier is the average of that of all member classifiers which constitute the committee classifier.

3.2 Experiment 1: The Effect of Subsampling Rate on Prediction Accuracy

To confirm the effect of subsampling rate on the prediction accuracy of classifier induced by S^3Bagging, the following experiments with test data sets for machine learning algorithms are conducted. The sample data sets are selected from UCI Machine Learning Repository[4]. They are selected from the data sets used in the papers of Quinlan[11] and Reinartz[12]. The specifications of the data sets are summarized in Table 1. In the following experimental results, each value given is the average of a ten fold cross-validation. In order to take into account the influence of the sampling in S^3Bagging and Bagging, the ten fold cross-validation was repeated 10 times and the error rate was evaluated by taing the average over the traials.

The divsrsity of the member classifiers effects the performance of the committee learning. As an index for measuring the diversity of classification result among the member classifiers, entropy $E = -\sum_{i=1}^{|C|} p(c_i) \log_2 p(c_i)$, is calculated, where $|C|$ is the number of classes, $p(c_i)$ is the ratio of member classifier which predicts the class $C = c_i$. The number of classes of all data sets used in this experiments is 2 so that the entropy E takes $0 \leq E \leq 1$. The entropy is 0 when all member classifiers predict the identical class. On the other hands, the entropy is 1 when the member classifiers which classified with each class is the same number.

The error rate of committee classifier is depicted in Figure 2. The error rate is improved when the subsampling rate is set large. In many case, the improvement of error rate becomes less when the subsampling rate is set larger than 30%.

The ratio (B/A) of error rate of committee classifier

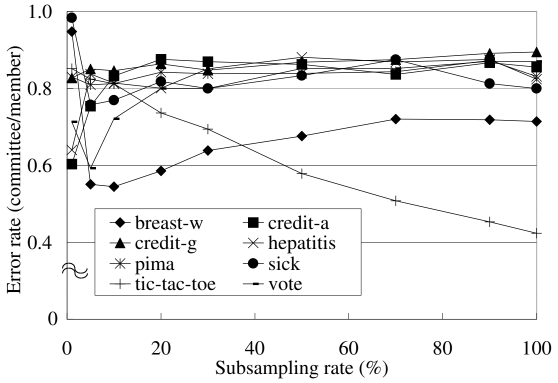

Figure 3. The ratio (B/A) of the error rate of committee classifier (B) to that of member classifier (A).

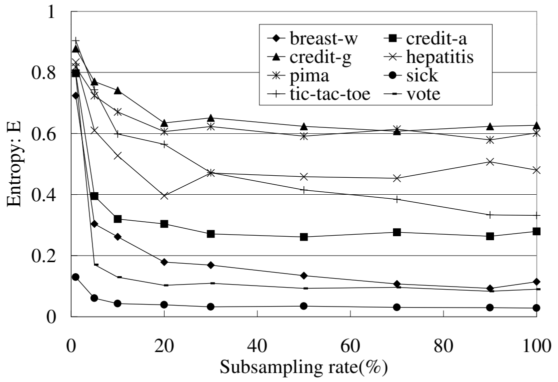

Figure 4. The entropy of classification by member classifiers which constitute the committee classifier.

(B) to that of member classifier (A) is depicted in Figure 3. This ratio indicates the effect of committee learning on prediction accuracy improvement. The effect is larger when the ratio is smaller. All ratios are smaller than 1. This result shows that the committee learning can improve the prediction accuracy at every subsampling rate. Except for the case of TIC-TAC-TOE, the smaller the subsampling rate is, the larger the improvement of prediction accuracy by committee learning is when the subsampling rate is larger than 5%. Thus, committee learning has an effect to improve the prediction accuracy not only in Bagging but also in S^3Bagging which adopts subsampling.

We consider the reason why the committee learning improves the predication accuracy only in the case of TIC-TAC-TOE even when the subsampling rate is set larger than 20%. In order to improve the prediction accuracy of the classifier more by committee learning, it is required to satisfy the following two conditions:

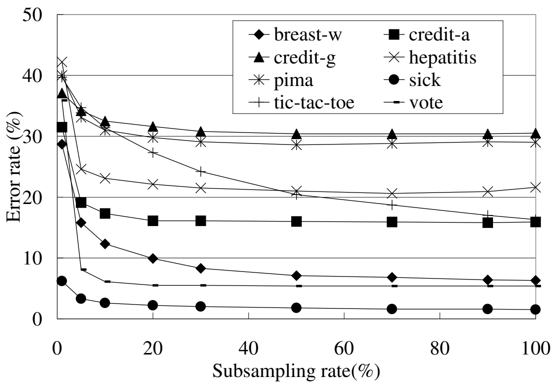

Figure 5. The experimental result: error rate of the member classifier (%).

(a) the characteristics of the member classifiers of the committee classifier are diverse enough, and (b) the prediction accuracy of each member classifier is high enough.

As an index for measuring the diversity of the classification results among member classifiers, entropy is calculated. The result is depicted in Figure 4. The entropy is large[1] in the case of CREDIT-G, HEPATITIS, PIMA, and TIC-TAC-TOE when the subsampling rate is larger than 20%. The large entropy shows that the member classifiers are diverse. Thus, these cases satisfy the condition (a). The error rates of member classifier are depicted in Figure 5. In the case of TIC-TAC-TOE, the error rate is 27% when the subsampling rate is set 20%. The error rate is still improved as the subsampling rate becomes larger. However, in the case of CREDIT-G, HEPATITIS and PIMA, the prediction accuracy of member classifier is about 30% and the prediction accuracy is not improved even when the subsampling rate is set larger. Consequently, the condition (b) is not satisfied in the case of CREDIT-G, HEPATITIS and PIMA but in the case of TIC-TAC-TOE. By the difference of prediction accuracy of member classifier among the cases, the prediction accuracy of committee classifier is much improved only in the case of TIC-TAC-TOE, although the prediction accuracy is not so improved on CREDIT-G, HEPATITIS and PIMA by committee learning.

In the case of BREAST-W, the improvement of the prediction accuracy by committee learning is small when the subsampling rate is set large as shown in Figure 3. This is considered to be due to the small diversity among member classifiers being small as shown in Figure 4, although the prediction accuracy of each

[1]Here, it is based on whether entropy is 0.3 or more.

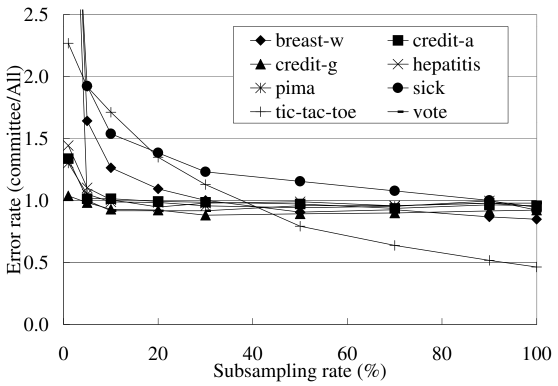

Figure 6. The ratio (B/C) of the error rate of committee classifier (B) to the error rate of classifier (C) induced from all original training data.

member classifier is high enough.

Next, the ratio (B/C) of the error rate of committee classifier induced by S^3Bagging (B) to that of the classifier induced from all original training data (C) is depicted in Figure 6. In all cases, the prediction accuracy is improved by S^3Bagging when the subsampling rate is set large enough. Furthermore, in many cases, the deterioration of the prediction accuracy of classifier induced by S^3Bagging is only about 20% even when the subsampling rate is set up to 30%. When the training data size is reduced especially as the subsampling rate is set up to 1%, the prediction accuracy of committee classifier becomes worse. This is because the number of instances of subsampled training data is too small to induce the correct member classifier.

3.3 Experiment 2: The Effect of Subsampling Rate on Processing Time

To confirm the shortening effect of processing time for classifier induction by S^3Bagging, the experiment with large size data is conducted. The processing time (CPU Time) is defined as the sum of processing time for subsampling and classifier induction as mentioned in section 2. In S^3Bagging and Bagging, we assume that each member classifier can be induced in parallel. Therefore, the processing time for inducing committee classifier is the longest processing time among each member classifier induction. A personal computer having the specification of OS: Linux OS, CPU: PentiumIII 700 MHz, and main memory: 256 M bytes is used in this experiment.

For the large size data sets, CENSUS-INCOME (abbreviated here as CENSUS), LED(10%) and WAVEFORM are selected. CENSUS is selected from UCI KDD Archive[3]. LED(10%) and WAVEFORM are selected from UCI Machine Learning Repository[4]. These data sets are used in the paper of Provost et al.[9]. "LED(10%)" means that 10% of instances are includes noise in one of the attribute values. The specifications of the data sets are summarized in Table 2.

Table 2. The specification of data sets for experiment 2.

Data set	# of instances	Attribute		Class
		Cont.	Disc.	
CENSUS	299,285	7	33	2
LED(10%)	500,000	–	7	10
WAVEFORM	300,000	21	–	3

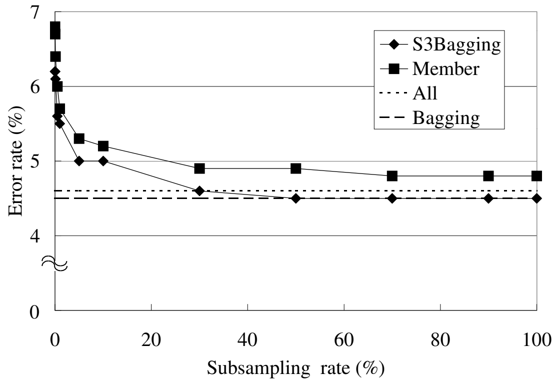

Figure 7. The experimental result on CENSUS : the error rate (%).

The experimental results on prediction accuracy are depicted in Figure 7, 8 and 9. The prediction accuracy of committee classifier is always higher than that of member classifier. When the subsampling rate is set 30% in the case of CENSUS, 0.1% in the case of LED(10%) and WAVEFORM, the error rate of committee classifier induced by S^3Bagging is almost equivalent to that of the classifier (All) induced from the all of the training data.

The experimental results on the processing time are depicted in Figure 10, 11 and 12. Inducing the classifier from all of the training data requires much processing time. The processing time is 300, 11 and 970 seconds in the case of CENSUS, LED(10%) and WAVEFORM respectively. On S^3Bagging, the experimental result shows that processing time for S^3Bagging is almost proportional to the subsampling rate. Therefore, S^3Bagging can shorten the processing time sharply by setting the subsampling rate small. For example, in the case of CENSUS, the processing time for S^3Bagging

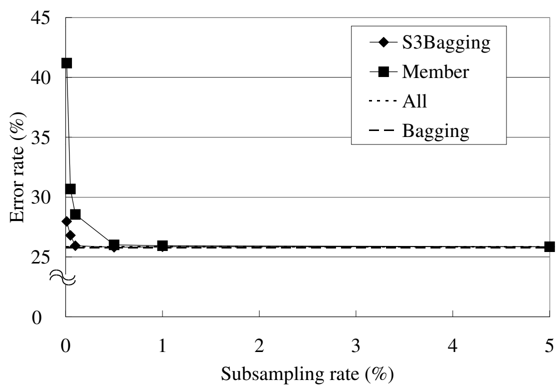

Figure 8. The experimental result on LED(10%) : the error rate (%).

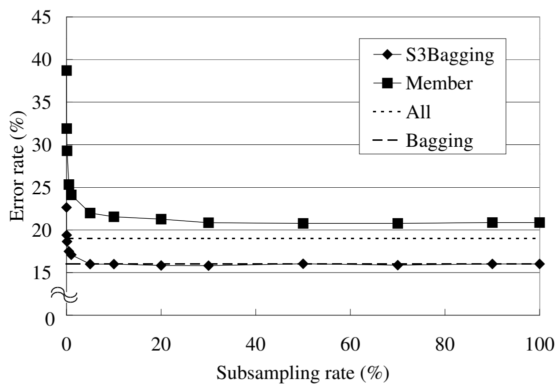

Figure 9. The experimental result on WAVEFORM : the error rate (%).

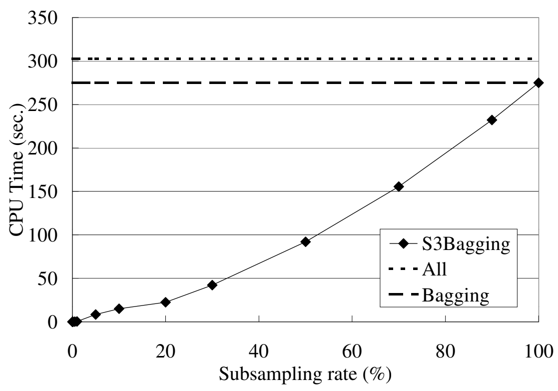

Figure 10. The experimental result on CENSUS : the processing time (sec.).

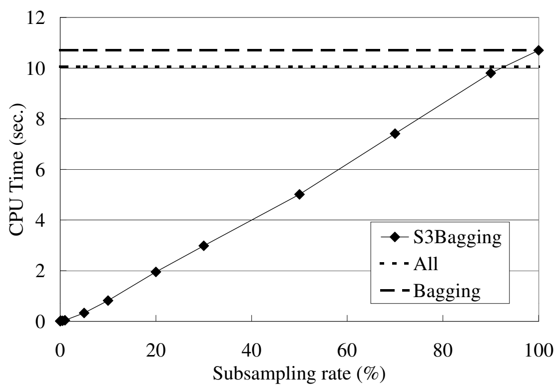

Figure 11. The experimental result on LED(10%) : the processing time (sec.).

is only 22 seconds when the subsampling rate is set up to 30%. The processing time is less than 8% in the case inducing the classifier from all original training data (All).

4. Discussion

From the experimental results in section 3.2 and 3.3, it is confirmed that the prediction accuracy of the classifier induced by S^3Bagging is not so inferior to that of the classifier induced from all instances of original training data even when the subsampling rate is set up to 30%. In the case of TIC-TAC-TOE, the prediction accuracy becomes better when the subsampling rate set large. TIC-TAC-TOE is an artifical data set which does not include any redundant instances[7]. By this feature of data set, the experimental result of TIC-TAC-TOE is characteristic. The effect of preventing the deterioration of prediction accuracy by subsampling is derived from the committee learning adopted in S^3Bagging. Next, the experimental result in section 3.3 shows that the processing time for classifier induction is reduced significantly by the effect of subsampling adopted in S^3Bagging. The processing time for S^3Bagging is mostly proportional to the subsampling rate.

From the viewpoint of the balance between the cost of processing time and the quality of the classifier represented with prediction accuracy, the experimental results shows that the subsampling rate should be set up to 30%, i.e., the number of sampling instances for the training data should be set up to several thousands. However, this is based on the result at the time of applying C4.5 as the learner. In the case where the other algorithm such as a neural network is adopted as learner, more experiments and discussions are required.

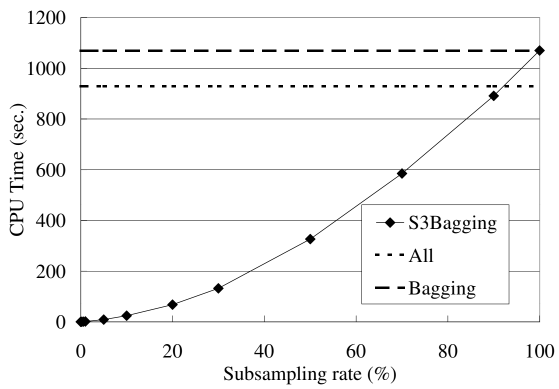

Figure 12. The experimental result on WAVEFORM : the processing time (sec.).

5. Related Work

The effect of the data size reduction by subsampling is investigated by Catlett[5]. In the experiments, the random sampling and the stratified sampling are adopted. They focused on the problem that the C4.5 needs much processing time for inducing decision tree when the data set includes many continuous attributes. Thus, they confirmed that the processing time could be shortened by adopting the subsampling to the training data.

Harris-Jones et al.[6] investigated the relation between training data size and prediction accuracy of classifier induced by some learning algorithms. Their experimental results shows that the prediction accuracy of classifier reaches a plateau at small subsampling rate in many cases.

Weiss et al. investigated the relation between subsampling rate and the prediction accuracy of classifier induced from the subsampled training data with some kinds of induction algorithms[16]. As the experimental result, the subsampling rates at which the best prediction accuracy classifier is induced are different among each algorithm. Additionally, they conducted the experiments which confirm the change of prediction accuracy of committee classifier when the member classifiers are induced from subsampled training data. In the experiment, they used Arcing[2] as the committee learning method. They show that the deterioration of the prediction accuracy of classifier can be prevented by adopting committee learning. However, they did not discuss from the viewpoint of the processing time for inducing classifier.

Reinartz investigated the effect of the sampling techniques both the prediction accuracy and the processing time[12]. The sizes of data sets used in the experiments are small, and they did not investigate the performance of classifier when the sampling techniques are adopted to the large size data.

6. Conclusion

In this paper, we investigated the effect of subsampling rate on the performance of S^3Bagging by carefully designed experiments. The effect was investigated through the experiment using test data sets which include large size data set.

The processing time of S^3Bagging for inducing classifier is mostly proportional to the subsampling rate. When the subsampling rate is set up to about 30%, the processing time becomes 30% compared with the case where all original training data is used. Additionally, by performing committee learning as the way of Bagging, S^3Bagging can prevent the deterioration of the prediction accuracy which is derived from subsampling. From the experimental results, when the subsampling rate is set about 30%, the deterioration of the prediction accuracy could be suppressed to about 30% in many cases.

The following issues remain for our future work.

- Experiments on S^3Bagging performance in the parallel processing environment: The execution of S^3Bagging in the parallel processing environment, such as PC cluster, is one of the most important future work.

- Determination of adequate subsampling rate: We proposed the description length of sampled data as a reference index at the time of determining the subsampling rate[14]. However, it is not taking the correlation among attributes into consideration. The new index which reflects the correlation among attributes should be investigated. As another direction of future work, adopting progressive sampling method[9] to S^3Bagging is also interesting.

- Investigation of the effective presentation method of the extracted information by S^3Bagging: In the process of S^3Bagging, not only committee classifier but also classification of each member classifier and their distribution are extracted. This information includes useful information for human analysts. The effective method to present this information to the human analysts should be investigated.

References

[1] Breiman, L.: Bagging Predictors, *Machine Learning*, **24** (2), pp.123–140 (1996).

[2] Breiman, L.: Bias, variance, and arcing classifiers, Technical Report 460 UC-Berkeley, Berkeley, CA. (1996).

[3] Bay, S. D.: The UCI KDD Archive, `http://kdd.ics.uci.edu`, Irvine, CA: University of California, Department of Information and Computer Science (1999).

[4] Blake, C., Keogh, E. 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).

[5] Catlett, J.: Megainduction: a Test Flight, In *Proc. of the Eighth International Workshop on Machine Learning*, pp.596–599. (1991).

[6] Harris-Jones, C. and Haines, T.L.: Sample Size and Misclassification: Is More Always Better?, Working Paper AMSCAT-WP-97-118, AMS Center for Advanced Technologies (1997).

[7] Matheus, C.J., Adding Domain Knowledge to SBL thorough Feature Construction, In *Proc.of the Eighth National Conference on Artificial Intelligence*, pp.803–808 (1990).

[8] Provost, F. and Kolluri, V.: A Survey of Methods for Scaling Up Inductive Algorithms, *Knowledge Discovery and Data Mining*, **3** (2), pp.131–169 (1999).

[9] Provost, F., Jensen, D., and Oates, T.: Efficient Progressive Sampling, In *Proc. of the Fifth International Conference on Knowledge Discovery and Data Mining*, pp.23–32 (1999).

[10] Quinlan, R.: *C4.5: Programs for Machine Learning*, Morgan Kaufmann (1993).

[11] Quinlan, R.: Bagging, Boosting, and C4.5., In *Proc. of the Thirteenth National Conference on Artificial Intelligence*, pp.725–730 (1996).

[12] Reinartz, T.: Similarity-Driven Sampling for Data Mining, Zytkow, J.M. and Quafafou, M. (eds.). *Principles of Data Mining and Knowledge Discovery: Second European Symposium PKDD '98*, pp.423–431 (1998).

[13] Shapire, R.E.: A Brief Introduction to Boosting, In *Proc. of the Sixteenth International Joint Conference on Artificial Intelligence*, pp.1401–1406 (1999).

[14] Terabe, M., Washio, T., Motoda, H.: Fast Classifier Generation by S^3Bagging (in Japanese), *IPSJ Transaction on Mathematical Modeling and Its Applications* (submitted).

[15] Terabe, M., Washio, T., Motoda, H.: S^3Bagging: Fast Classifier Generation with Subsampling and Bagging, In *Proc. of the fourth International Symposium on Intelligent Data Analysis* (2001, to appear).

[16] Weiss, S.M. and Indurkhya, N.: *Predictive Data Mining – a practical guide –*, Morgan Kaufmann (1998).

[17] Zheng, Z.: Generating Classifier Committees by Stochastically Selecting both Attributes and Training Examples, In *Proc. of the Third Pacific Rim Conference on Artificial Intelligence*, pp.12–33 (1998).