

# Effect of Subsampling Rate on Subbagging and Related Ensembles of Stable Classifiers

Faisal Zaman\* and Hideo Hirose

Kyushu Institute of Technology,  
680-4 Kawazu, Iizuka-shi, Fukuoka, Japan  
`zaman@ume98.ces.kyutech.ac.jp,`  
`hirose@ces.kyutech.ac.jp`

**Abstract.** In ensemble methods to create multiple classifiers mostly bootstrap sampling method is preferred. The use of subsampling in ensemble creation, produce diverse members for the ensemble and induce instability for stable classifiers. In subsampling the only parameter is the subsample rate that is how much observations we will take from the training sample in each subsample. In this paper we have presented our work on the effect of different subsampling rate (SSR) in bagging type ensemble of stable classifiers, Subbagging and Double Subbagging. We have used three stable classifiers, Linear Support Vector Machine (LSVM), Stable Linear Discriminant Analysis (SLDA) and Logistic Linear Classifier (LOGLC). We also experimented on decision tree to check whether the performance of tree classifier is influenced by different SSR. From the experiment we see that for most of the datasets, the subbagging with stable classifiers in low SSR produces better performance than bagging and single stable classifiers, also in some cases better than double subbagging. We also found an opposite relation between the performance of double subbagging and subbagging.

**Keywords:** Subsample rate, Stable Classifiers, Subbagging, Double Subbagging.

## 1 Introduction and Notation

An important theoretical approach to analyzing the performance of combination of learning machines is through studying their stability. Various notions of stability of ensemble of machines (combination of machines) have been proposed in the last few years [2], [3], and have been used to study how combining machines can influence the generalization performance [5]. Numerous theoretical and empirical studies have been published to establish the advantages of combining learning machines, for example using Boosting or Bagging methods, [4] [9] very often leads to improved generalization performance, and a number of theoretical explanations have been proposed [5], [9].

---

\* Corresponding author.

Subbagging [6] is a particular ensemble architecture of bagging, consisting of a voting combination of a number of learning machines. Each machine is obtained by training the same underline learning algorithm, e.g. a decision tree, on a dataset drawn randomly with replacement from an initial training set with having some part of it (50%). In the double bagging framework proposed by Hothorn and Lausen [8], the out-of-bag sample is used to genrate an additional classifier model to integrate with the base learning model. In the setup of Hothorn and Lausen, the double-bagging uses the values of the linear discriminant functions trained on the out-of-bag sample as additional predictors for bagging classification trees only. The discriminant variables are computed for the bootstrap sample and a new classifier is constructed using the original variables as well as the discriminant variables. Using subbagging in double bagging rather than bagging, will result in larger out-of-bag sample for the additional classifier and if the additional classifier is a strong classifier it could result in an efficient ensemble method [10].

In this paper we have presented our experimental results, where we have constructed ensemble of three stable (linear) classifiers using different subsample rate (SSR) and check their accuracy in several benchmark datasets. We also compared their performance with aingle and bagged stable classifiers. In [7] authors analyzed the characteristics of subbagging. In particular, they developed probabilistic bounds on the distance between the empirical and the expected error that, unlike in the standard analysis of bagging are not asymptotic. These bounds formally suggest that subbagging can increase the stability of the learning machines when these are not stable and decrease otherwise. Keeping this in mind we develop our experiments. The results indicate that subbagging can significantly increase the stability of these machines whereas the test error is close (and sometimes better) to that of the same machines trained on the full training set. Another important practical consequence of our theoretical results is that, since subbagging improves the stability, it can be easier to control the generalization error. The paper is organized as follows: in Section 2 we have reviewed the theory of stability of subsampled ensemble. In Section 3, have presented the experiment and the results of the experiments with some discussion. It is followed by the Conclusion of the work.

## 2 Theory

Let  $D_l = \{(x_i, y_i) \in \mathcal{X} \times \{-1, 1\}\}_{i=1}^l$  be the training set. A learning algorithm is a function  $\mathcal{A} : (\mathcal{X} \times \{-1, 1\})_{i=1}^l \rightarrow (\mathcal{X})^{\mathbb{R}}$  which give an training set  $D_l$ , returns a real valued classifier  $f_{D_l} : \mathcal{X} \rightarrow \mathbb{R}$ . Classification of a new input  $\mathbf{x}$  is done by taking the sign of  $f_{D_l}(\mathbf{x})$ . Now we discuss the notion of  $\beta_l$  stability and relate it to the generalization performance of a subsampled ensemble based on [7]. The output of a larning machine on a new (test) point  $\mathbf{x}$  should not change more than  $\beta_l$  when we train the machine with any training set of size  $l$  and when we train the machine with the same training set but one training point (any point) removed. Now let  $F_{D_l}^r = E_{D_r}[f_{D_r}]$  be the expected combinations of machines

trained on subsamples of size  $r$ . Next, we state a theorem modified from [7] to support our findings

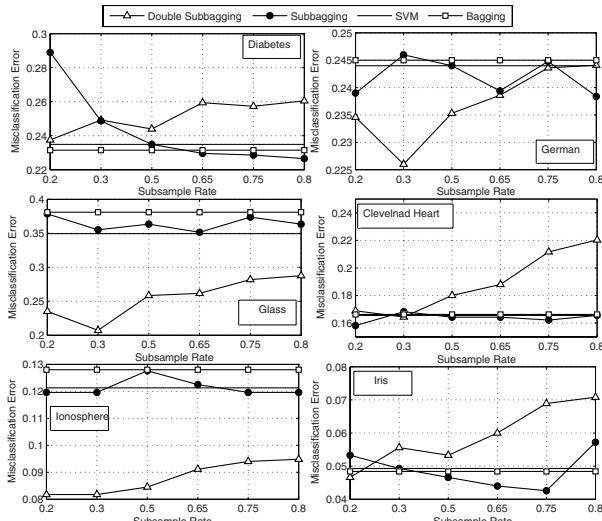
**Theorem 1.** For any given  $\delta$  with probability  $1 - \eta$  the generalization misclassification error of the expected combination  $F_{D_l}^r$  of machines each using a subsample of size  $r$  of the training set and each having stability  $\beta_r$  is bounded by:

$$R_{emp}^\delta(F_{D_l}^r) + \frac{4r}{l} \frac{\beta_r}{\delta} + \sqrt{\frac{1}{2l} \left( \frac{4r\beta_r}{\delta} + 1 \right)^2 \ln\left(\frac{1}{\eta}\right)}$$

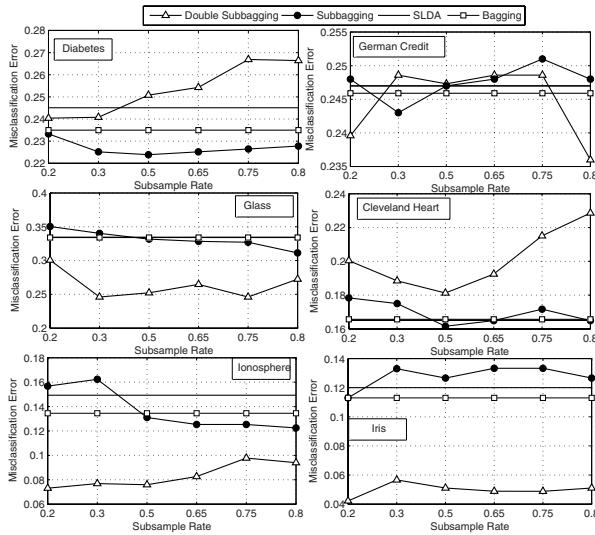
Thus we see from the theorem that the control of the test error by empirical error ( $R_{emp}^\delta(F_{D_l}^r)$ ) is mainly due to the subsampling effect rather than the size of the ensemble. This indicates that increasing the number of base classifiers should not improve this control as much as reducing the value of  $r$ . This we experimentally verified in Section 3. It should also be noted from Theorem 1 that if the stability of a subsampled machine  $F_{D_l}^r$  is better than a single machine then the ensemble will give better bound otherwise the bound will be worse and subbagging should be avoided.

### 3 Experiments and Discussion of Results

We conducted a number of experiments using six datasets from UCI [1]: Diabetes, German Credit, Glass Identification, Cleveland Heart, Ionosphere and Iris. We focused mainly on the stable classifiers Linear Support Vector Machine



**Fig. 1.** Performance of LSVM, Bagging LSVM, Subbagging LSVM and Double Subbagging LSVM with different SSR. In X axis the SSR are given, in Y axis, average of 10 repetitions of 10-fold cross-validation misclassification error is given.



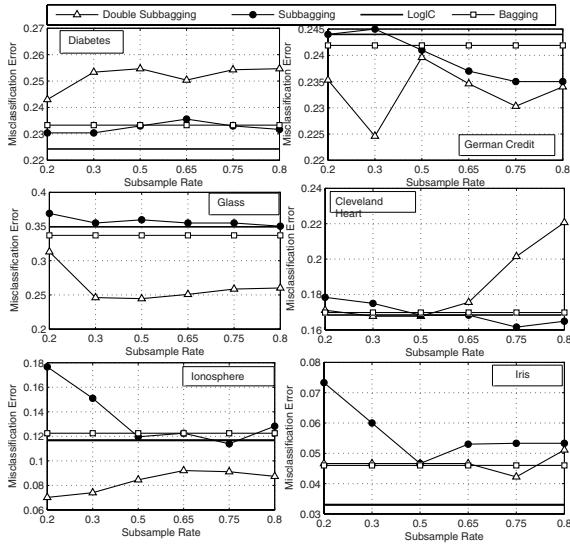
**Fig. 2.** Performance of *SLDA*, *Bagging SLDA*, *Subbagging SLDA* and *Double Subbagging SLDA* with different SSR. In *X* axis the SSR are given, in *Y* axis, average of 10 repetitions of 10-fold cross-validation misclassification error is given.

(LSVM), Stable Linear Discriminant Analysis (SLDA) and Logistic Linear Classifier (LOGLC), we have also checked the performance of Classification and Regression Tree (CART) in different SSR. The first goal of the experiments was to check whether in subbagging with low SSR, the performance of the linear classifiers is better than single classifier and bagged version of the classifier. Secondly we wanted to check whether smaller SSR improve the accuracy of double subbagging with linear classifiers as the additional classifiers. Thirdly, we wanted to check the performance of CART with smaller SSR; whether with smaller SSR, it produce competent accuracy.

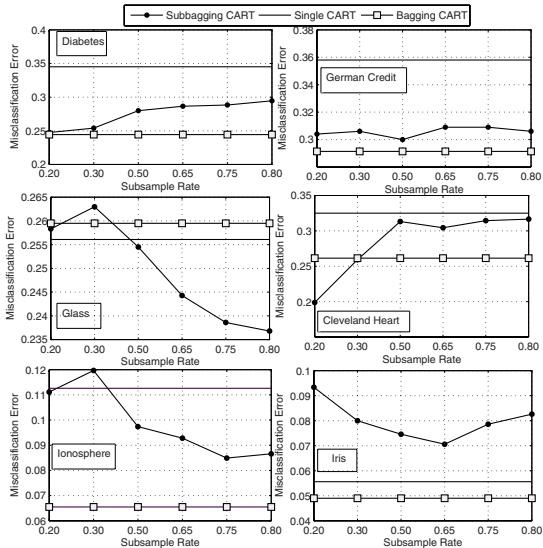
In all the experiment we have fixed the size of the ensemble size to 20, for all the classifiers. For LSVM, first we have optimized the cost parameter with 5-fold cross-validation. Then we use that parameter to create the ensembles. For each data set, ten stratified ten-fold cross-validation was performed.

### 3.1 Discussion of the Results

In case of LSVM we see that the performance of double subbagging is best among all the classifiers. In case of subbagging it showed superior performance with relatively low SSR in case most of the datasets. It showed better accuracy than bagging and single LSVM in all datasets except Ionosphere dataset, (Fig. 1). This is because the test performance of the subbagged machines are bounded by the empirical error and its stability does not depend on the training size [7]. In case SLDA the scenario is nearly same, with double subbagging performing poorly



**Fig. 3.** Performance of *LOGLC*, *Bagging LOGLC*, *Subbagging LOGLC* and *Double Subbagging LOGLC* with different SSR. In X axis the SSR are given, in Y axis, average of 10 repititions of 10-fold cross-validation misclassification error is given.



**Fig. 4.** Performance of *CART*, *Bagging CART*, *Subbagging CART* with different SSR. In X axis the SSR are given, in Y axis, average of 10 repititions of 10-fold cross-validation misclassification error is given.

in two datasets (Diabetes and Cleveland Heart), check Fig. 2. For LOGLC, the single LOGLC performed better than the ensemble of LOGLC in three datasets and it is apparent from the Fig. 3. In case of CART the performance of subbagging is satisfactory, as it showed competent performance with lower SSR in case of four datasets, than bagging CART in Fig. 4.

## 4 Conclusion

In this paper we have experimented with ensemble of several stable (linear) classifiers, where the ensembles are constructed with different subsampling rates. It is interesting to notice that ensemble with low SSR (0.2 – 0.5) producing better accuracy with approximately all the linear classifiers, compared with bagging and single linear classifier. This implies that, one can get similar performance similar to a single classifier (machine), by training faster machines using smaller training sets. This is an important practical advantage of ensembles of machines to reduce the computational time significantly, specially in the case of large datasets. In case of decision tree it also produced better accuracy with low SSR (0.2 – 0.5). The main direction which emerges from this work is that, time complexity of ensemble of machines can be reduced without deteriorating its accuracy.

## References

1. Blake, C.L., Merz, C.J.: UCI Repository of Machine Learning Databases, <http://www.ics.uci.edu/mlearn/MLRepository.html>
2. Bousquet, O., Elisseeff, A.: Stability and generalization. *J. Mach. Lear. Res.* 2, 499–526 (2002)
3. Elisseeff, A., Evgeniou, T., Pontil, M.: Stability of Randomized Algorithm. *J. Mach. Lear. Res.* 6, 55–79 (2005)
4. Breiman, L.: Bagging predictors. *Machine Learning* 24(2), 123–140 (1996a)
5. Breiman, L.: Heuristics of instability and stabilization in model selection. *Annals of Statistics* 24(6), 2350–2383 (1996c)
6. Bühlman, P.: Bagging, subbagging and bragging for improving some prediction algorithms. In: Arkatas, M.G., Politis, D.N. (eds.) *Recent Advances and Trends in Nonparametric Statistics*, pp. 9–34. Elsevier, Amsterdam (2003)
7. Evgeniou, T., Pontil, M., Elisseeff, A.: Leave one out error, stability, and generalization of voting combinations of classifiers (Preprint) (2001)
8. Hothorn, T., Lausen, B.: Double-bagging: combining classifiers by bootstrap aggregation. *Pattern Recognition* 36(6), 1303–1309 (2003)
9. Shapire, R., Freund, Y., Bartlett, P., Lee, W.S.: Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics* (1998)
10. Zaman, M.F., Hirose, H.: A New Double Bagging via the Support Vector Machine with Application to the Condition Diagnosis for the Electric Power Apparatus. In: International Conference on Data Mining and Applications (ICDMA 2009), pp. 654–660 (2009)