Use of Round Shape Bootstrap Samples for a Classifier Design

Yoshihiro Mitani, Yusuke Fujita, and Yoshihiko Hamamoto

***Abstract*—In order to improve the error rate estimation method, a bootstrap method was proposed. We focus on improvement of not the error rate estimation but a classification performance of a classifier. We explore a bootstrap approach for designing a classifier. In this paper, in order to improve the classification performance of a classifier, we propose a round shape bootstrap method. The areas of bootstrap samples generated by the round shape bootstrap method are expected to be more smoothed. Experimental results show the proposed method is effective for a nearest neighbor (1-NN) classifier.**

***Index Terms*—Bootstrap samples, artificial sample generation, classifier design, 1-NN, pattern recognition, classification performance.**

I. Introduction

In a small training sample size situation, designing a classifier which has a high classification performance is important. Hence, a considerable amount of effort has been made to improve the classification performance, particularly in small training sample size situations. In the field of an artificial neural network(ANN), there are many discussion of designing an ANN classifier in small training sample size situations [1][2][3]. In the literature [3], the classification performance of the ANN classifier has improved by adding the small noise to the training samples. The results show that the size of the noise influences the classification performance of the ANN classifier. The noise injection approach is not always the improvement of the classification performance of the ANN classifier. Because the outlier may influence the classification performance.

In statistics, Efron introduced the bootstrap method [4]. Efron [5], Jain *et al*. [6], Chernick *et al*. [7], and Hand [8] show that a bootstrap method improves than the conventional methods in estimating the error probability, particularly when the available sample size is small. In not an error rate estimation but a classifier design, we show that the use of bootstrap samples is effective in designing an ANN classifier [9]. In order to improve the ANN classifier generalization ability, we proposed a bootstrap method which used a local training sample and combined them linearly [9]. However, the previously proposed method is only applied to the ANN classifier. And the bootstrap samples generated by the previously proposed method seem not to be smoothed because the shape is convex. In this paper, we propose a round shape bootstrap method for improvement of a classification performance of a classifier. The round shape bootstrap approach is based on the basic idea of generating a convex bootstrap sample approach [9]. The round shape bootstrap method seems smoothed than the convex one. Experimental results show the effectiveness of the proposed method. Thus, the proposed method may be promising in designing a classifier.

Manuscript received November 25, 2015; revised March 1, 2016.

Yoshihiro Mitani is with the Department of Intelligent System Engineering, National Institute of Technology, Ube College, Ube, Japan (e-mail: mitani@ube-k.ac.jp).

Yusuke Fujita, and Yoshihiko Hamamoto are with Yamaguchi University, Ube, Japan.

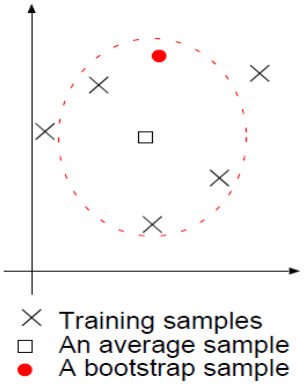

Fig. 1. An illustration of a round shape bootstrap sample generation.

II. Round Shape Bootstrap Samples

We propose a round shape bootstrap method in order to improve the classification performance of a classifier. Let $X^c = \{x_1^c, x_2^c,...,x_N^c\}$ be an original training sample set from class c. Here we focus on the generation of the round shape bootstrap for each class. A bootstrap sample y^c is generated by the following:

Step 1. Select x_{k0} from X^c randomly.

Step 2. Select the nearest r - 1 training samples $x_{k1},\ldots, x_{kr-1}$ of x_{k0} , using the Euclidean distance. Here, the value of r must be selected from 2 to N.

Step 3. Compute a local average vector $\hat{x}$, using r training samples.

Step 4. Compute the average of the Euclidean distance d among r training samples and the local average vector $\hat{x}$.

Step 5. Generate a bootstrap sample, $y_c=S\delta+\hat{x}$, where δ is given by a random vector which distance is within d, and the value of S must be selected as a non-negative real number.

By repeating step1 to Step 5 n times, we get n bootstrap samples. In the proposed method, we use n bootstrap samples and original training samples. If $S < 1$, a bootstrap sample is generated in a shrunk area compared with the distribution of training samples. On the other hand, if $S > 1$, a bootstrap sample is generated in an extended area. When S = 0, a bootstrap sample becomes the average vector of r

DOI: 10.7763/IJET.2017.V9.966

training samples selected. Fig. 1 shows an illustration of a round shape bootstrap sample generation. This is an example of the area by the proposed method when S = 1 and r = 5. A bootstrap sample is randomly generated within the dashed line circle made by 5 training samples. Note that a square is an average vector of 5 training samples.

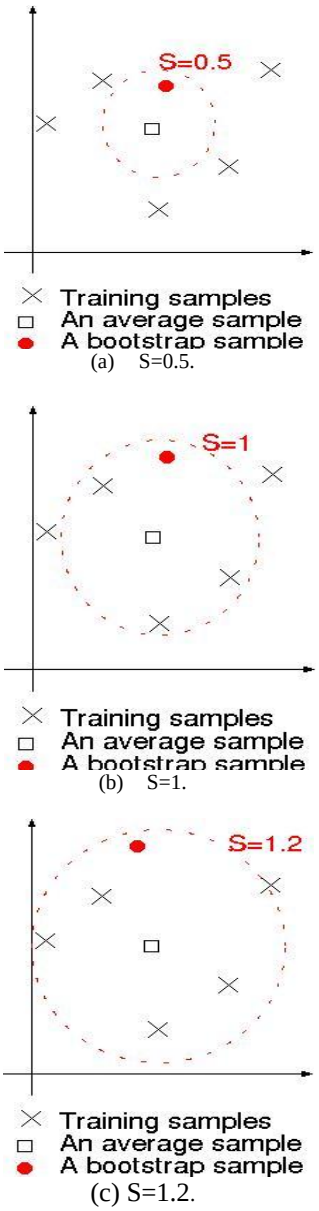

Fig. 2. Bootstrap sample generation areas by a parameter value S.

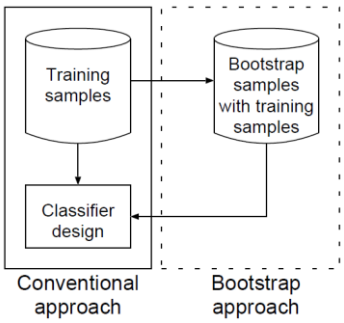

Fig. 3. An illustration of the bootstrap approach.

Let's see influence of the parameter value S. Fig. 2 shows bootstrap sample generation areas by a parameter value S. Fig. 2 (a), (b), and (c) denote S=0.5, 1, and 1.2, respectively. The bootstrap sample generation areas varies as the value S changes.

According to values of r and S, the distribution of bootstrap samples varies. This influences the classification performance of a classifier. Hence, parameter values of r and S must be optimized in terms of the classification performance.

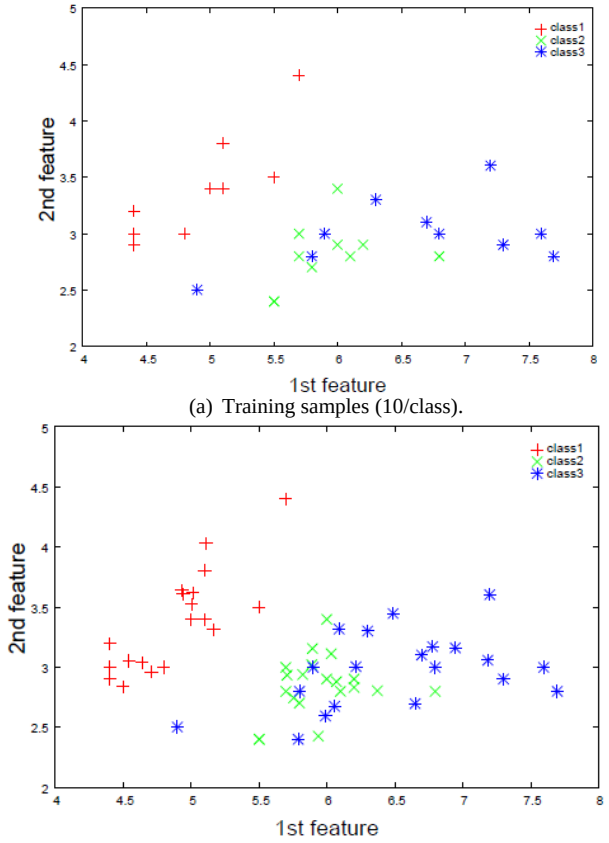

Fig. 4. An example of round shape bootstrap samples on the Iris data set with 1st- and 2nd- features.

III. Bootstrap Samples Based Learning

In order to improve a classification performance of a classifier, we propose a bootstrap approach. In a small training sample size situation, the difficulty of a classifier design always remains. To overcome this difficulty, we explore a round shape bootstrap method. Fig. 3 shows an illustration of the bootstrap approach. In the bootstrap samples based learning, we use bootstrap samples and training samples.

An Iris data set is well known and is widely used for pattern recognition filed [10]. In the experiments, we use the Iris data set. It is a real data. The number of classes is 3, and the dimensionality is 4. The number of samples per a class is 50. Fig. 4 shows an example of round shape bootstrap samples on the Iris data set with 1st- and 2nd- features. Since the dimensionality of the Iris data set is 4, we use 2 features: 1st- and 2nd- features, for visualization. Fig. 4(a) shows only 10 training samples per a class. On the other hand, Fig. 4(b) shows 10 training samples and 10 bootstrap samples generated by the proposed method. From the distribution of Fig. 4(b), it is expected that the proposed method improves the classification performance. The distribution of the Iris data set per a class may be smoothed.

TABLE I: INFLUENCE OF VALUES OF R AND S IN THE PROPOSED METHOD. THE UPPER IS THE AVERAGE ERROR RATE (%) AND THE LOWER IS THE 95% CONFIDENCE INTERVAL

Values of r	Values of S								
	0.0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6
3	4.92 4.54,5.30	4.69 4.31,5.06	4.47 4.09,4.85	4.45 4.09,4.82	4.50 4.13,4.86	4.67 4.28,5.06	4.86 4.47,5.25	5.15 4.77,5.54	5.51 5.11,5.91
5	4.88 4.50,5.26	4.57 4.21,4.93	4.35 3.98,4.71	4.61 4.19,5.03	4.58 4.22,4.94	4.90 4.50,5.29	5.45 5.02,5.89	6.00 5.57,6.44	6.71 6.24,7.18
7	4.97 4.58,5.36	4.79 4.40,5.18	4.68 4.28,5.07	4.69 4.32,5.05	4.95 4.57,5.32	5.30 4.91,5.70	6.04 5.62,6.46	6.89 6.45,7.32	7.43 6.97,7.89

IV. EXPERIMENTAL RESULTS

In order to investigate the effectiveness of the proposed method, we used an Iris data set. The effectiveness of the proposed method is examined in terms of the error rate. The error rate *Pe* is defined as follows.

Pe = 100 x No. of test samples misclassified / No. of all test samples [%]

In error rate estimation literature, the holdout method [11] has been successfully used, because it maintains the statistical independence between the training and test samples. In order to evaluate the proposed method, the average error rate is obtained by the holdout method. Fig. 5 shows the error rate estimation. First, for each class, we randomly divide 50 available samples into 10 training samples and 40 test samples. Second, for each class, we generate 40 round shape bootstrap samples from 10 training samples. Then the training sample size is 50 per a class, which contains 40 round shape bootstrap samples and 10 training samples. Third, the error rate is computed by a classifier. Finally, by 100 repetitions, the average error rate and the 95% confidence interval are obtained. Here, we adopt a nearest neighbor (1-NN) classifier [6] as a classifier since it is well known and widely used in the pattern recognition field.

The purpose of the experiment 1 is to investigate the classification performance of the conventional and the proposed method. The parameter values of r and S of the proposed method vary. Table II is the comparison of the proposed and conventional method. The error rate of the proposed method outperforms the conventional one. Then the parameter values of r and S are 5 and 0.4, respectively. The 95 % confidence intervals of the proposed method do not overlap that of the conventional method. Therefore, the proposed method seems effective in terms of the error rate.

TABLE II. COMPARISON OF THE PROPOSED AND CONVENTIONAL METHOD. THE UPPER IS AN AVERAGE ERROR RATE (%) AND THE LOWER IS THE 95% CONFIDENCE INTERVAL

Proposed method	Conventional method
4.35 3.98,4.71	5.40 5.01,5.80

The purpose of the experiment 2 is to investigate the influence of parameter values of r and S. In the experiments, we vary that r are 3, 5, and 7, and S are from 0.0 to 1.6 every 0.2. Table I shows the influence of values of r and S in the proposed method. When the values of r and S are 5 and 0.4, the error rate shows the minimum. The parameter values of r and S directly influence the error rate. Particularly when the parameter value of S becomes larger, the error rate seems larger. Therefore, the optimization of parameter values of r and S must be considered.

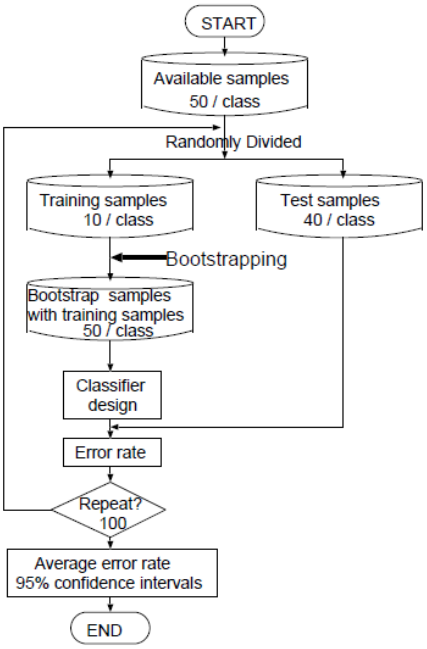

Fig. 5. Error rate estimation.

V. CONCLUSIONS

In this paper, we have proposed a round shape bootstrap method for the improvement of a classification performance of a classifier. Experimental result on the Iris data set shows the proposed method using 1-NN classifier is effective. In the future, the performance of the proposed method should be investigated on various classifiers. Furthermore, another type of the data set should be investigated.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 25330357.

REFERENCES

[1] S. Raudys and A. K. Jain, "Small sample size problems in designing artificial neural networks," *Artificial Neural Networks and Statistical Pattern Recognition: Old and New Connections*, pp. 33–50, 1991.

[2] S. Raudys, "Why do multilayer perceptron have favorable small sample properties?" *Pattern Recognition in Practice IV*,\pp. 287–298, 1994.

[3] Y. Hamamoto, Y. Mitani, and S. Tomita, "On the effect of the noise injection in small training sample size situations," in *Proc. International Conference on Neural Information Processing*, pp. 626–628, 1994.

[4] B. Efron, "Bootstrap method: Another look at the jackknife," *The Annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979.

[5] B. Efron, "Estimating the error rate of a prediction rule: Improvement on cross-validation," *J. American Statistic Association*, vol. 78, no. 382, pp. 316–331, 1973.
[6] A. K. Jain, R. C. Dubes, and C. C. Chen, "Bootstrap techniques for error e stimation," *IEEE Trans*, vol. 9, no. 5, pp. 628–633, 1987.
[7] D. J. Hand, "Recent advances in error rate estimation," *Pattern Recognition Letters*, vol. 4, pp. 335–346, 1986.
[8] M. R. Chernick, V. K. Murthy and C. D. Nealy, "Application of bootstrap and other resampling techniques: Evaluation of classifier performance," *Pattern Recognition Letters*, vol. 3, pp. 167–178, 1985.
[9] Y. Mitani, Y. Hamamoto, and S. Tomita, "Use of bootstrap samples in designing artificial neural network classifier," in *Proc. 1995 IEEE International Conference on Neural Networks*, vol. 4, pp. 2103–2106, 1995.
[10] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, pp. 179–188, 1936.
[11] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Second Edition, Academic Press, 1990.

Yoshihiro Mitani is currently a professor at National Institute of Technology, Ube College, Japan. His research interests include pattern recognition and image processing techniques. He is a member of IEEE. Ph. D.

Yusuke Fujita is currently an assistant professor at Yamaguchi University, Japan. His research interests include pattern recognition and image processing techniques. Ph. D.

Yoshihiko Hamamoto is currently a professor at Yamaguchi University, Japan. His research interests include pattern recognition. He is a member of IEEE. Ph. D.