New cancer diagnosis method
on the basis of fuzzy theory and boosting

Hiro Takahashi[1]
h041305d@mbox.nagoya-u.ac.jp

Hiroyuki Honda[1]
honda@nubio.nagoya-u.ac.jp

[1] Department of Biotechnology, School of Engineering, Nagoya University,
Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan

Keywords: cancer diagnosis, boosting, fuzzy theory, reliability evaluation

1 Introduction

For adequate treatment of patients, it is important that the accurate and reliable algorithm is developed for construction of diagnosis system. Artificial neural network (ANN) [4], and fuzzy neural network (FNN) combined with SWEEP operator method (FNN-SWEEP method) [1] are useful for constructing cancer class prediction model with high accuracy. However, ANN and FNN-SWEEP method have overfitted the data used in the model construction process because these methods are not able to avoid the "curse of dimensionality". We have developed boosted fuzzy classifier with SWEEP operator (BFCS) method on the basis of fuzzy theory and boosting algorithm. This method has been applied to construct class predictors for cancer diagnosis using clinical data with respect to microscopic observation such as clump thickness or marginal adhesion for breast cancer. The model performance has been evaluated through comparison with a conventional method such as support vector machine (SVM) [5] and fuzzy neural network combined with SWEEP operator (FNN-SWEEP) method previously proposed by us.

2 Method and Results

2.1 Data processing

In this study, we used breast cancer data set from Wisconsin Breast Cancer Database from a study of Wolberg *et al*. [6]. These data consisted of 9 clinical parameters (clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitoses) and 699 examples. Sixteen samples of this data set are including uncompleted data set. We used remained 683 examples in which 444 are benignant examples and 239 are malignant examples. Wolberg *et al*. reported that all nine parameters may be important risk factor for development of breast cancer. In this experiment, this data set was partitioned into two groups: 100 samples (65 benignant, 35 malignant) as a modeling data set for constructing the class prediction model (predictor) and 583 (379 benignant, 204 malignant) as a blinded data set for evaluating the constructed predictor.

2.2 Comparison of the performance of BFCS and other method

Performance of BFCS models as the class predictors was investigated. For comparison, two kinds of class predictors, FNN-SWEEP and SVM, were constructed. Since it is necessary to construct the class predictors, which can correctly classify not only modeling data but also new blinded data, the performance of the predictors was compared on accuracy using blinded data set which was never used for modeling.

2.3 BFCS model

In the present paper, we have developed boosted fuzzy classifier with SWEEP operator (BFCS) method on the basis of AdaBoost [2] which is the most basic and practical boosting algorithm. Our algorithm has features which are fast, accurate, and reliable. The structure of a BFCS model was showed in Figure 1. Figure 1 shows the structure of two types of BFCS, BFCS-1 and BFCS-1,2, in which is composed of 1-input type I FNN models [3] and 1-input FNN or 2-input FNN models, respectively. In addition, BFCS is able to calculate reliability index (RI) on the basis of the fuzzy theory, for prediction result of each example.

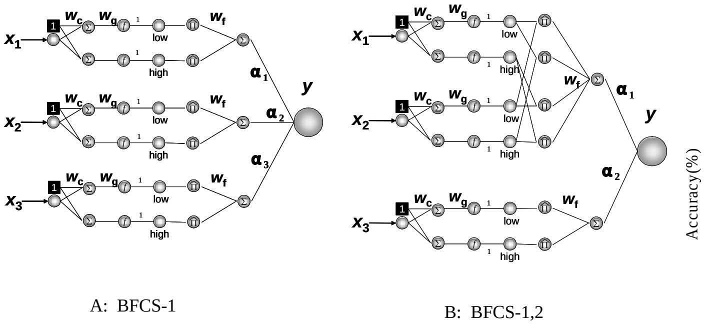

Figure 1: Concept of BFCS model.

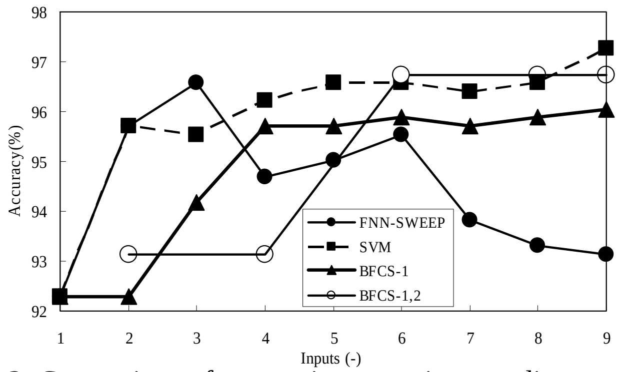

Figure 2: Comparison of accuracies on various predictors.

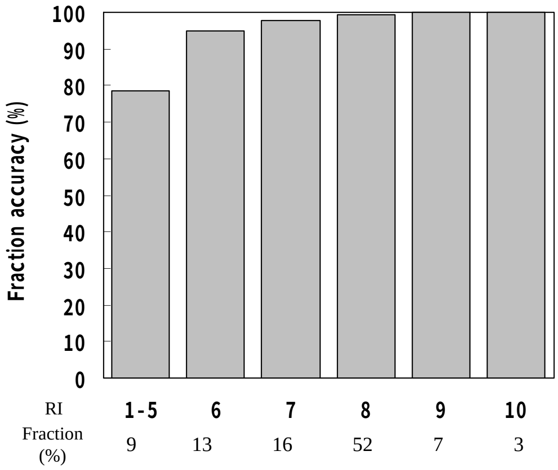

Figure 3: Fraction accuracies on various RIs.

Table 1: Comparison of calculation time for each method.

Method	Time(sec)
FNN-SWEEP	1355
SVM	3.51
BFCS-1	0.1
BFCS-1,2	0.12

Results and Discussion

We have developed BFCS method and applied it to clinical data of cancers to construct reliable class predictors for diagnosis. The results show that the BFCS class predictors are superior to other method, such as FNN-SWEEP and SVM, with regards that high prediction accuracy (Figure 2) and ability to evaluate reliability of predictions for each patient (Figure 3). Furthermore, BFCS is extremely faster algorithm than FNN-SWEEP and SVM (Table 1). As an another data, we applied the construction of class predictors using proteomic pattern data set taken from MALDI-TOF mass spectrometry in the serum of patient with ovarian cancer. Similar good results were obtained in the case of BFCS modeling. From these results, it was suggested that BFCS has the potential to function as a new method of cancer diagnosis.

References

[1] Ando, T., Suguro, M., Hanai, T., Kobayashi, T., Honda, H. and Seto, M., Fuzzy neural network applied to gene expression profiling for predicting the prognosis of diffuse large B-cell lymphoma, *Jpn. J. Cancer Res.*, 93:1207-1212, 2002.

[2] Freund,Y. and Schapire,R., A decision-theoretic generalization of online learning and an application to boosting, *J. Comput. System Sci.*, 55:119-139, 1997.

[3] Horikawa, S., Furuhashi, T. and Uchikawa, Y., On fuzzy modeling using fuzzy neural networks with the back-propagation algorithm, *IEEE T. Neural Networ.* 3:801-806, 1992.

[4] Tomita, Y., Tomida, S., Hasegawa, Y., Suzuki, Y., Shirakawa, T., Kobayashi, T. and Honda, H., Artificial neural network approach for selection of susceptible single nucleotide polymorphisms and construction of prediction model on childhood allergic asthma, *BMC Bioinformatics*. 5:120, 2004.

[5] Vapnik, V., *Statistical Learning Theory*, Wiley, New York, NY, 1998.

[6] Wolberg, W. H. and Mangasarian, O.L., Multisurface method of pattern separation for medical diagnosis applied to breast cytology, *P. Natl. Acad. Sci. USA*, 87:9193-9196, 1990.