Genetic Fuzzy Discretization for Classification Problems*

Yoon-Seok Choi and Byung-Ro Moon

School of Computer Science & Engineering, Seoul National University
Shillim-dong, Gwanak-gu, Seoul, 151-742, Korea
{yschoi, moon}@soar.snu.ac.kr

Many real-world classification algorithms can not be applied unless the continuous attributes are discretized and the interval discretization methods are used in many machine learning techniques. It is hard to determine the intervals for the discretization of numerical attributes that has an infinite number of candidates. And interval discretization methods are based on a crisp set, a value in a continuous attribute must belong to only one interval. They are often not proper for describing a value located around the boundaries of intervals. Fuzzy partioning is an attractive method for those cases in classification problems. An important decision in fuzzy partitioning is about the positions of interval boundaries and the degrees of overlapping in the fuzzy sets. We optimize the parameters that specify fuzzy partitioning by genetic algorithms.

We divide the range of a continuous attribute into k intervals and represent each value by a k-bit string where each bit corresponds to one interval. The i^{th} bit of the binary string represents whether the value belongs to the i^{th} interval or not. While a value belongs to only one interval in a simple discretization, it can belong to more than one interval in fuzzy discretization. Thus a value can be represented by a binary mask. For example, in Fig. 1, the value 0.595 belongs to the third and fourth intervals and is represented by a binary mask 00110. It provides more flexibility in machine learning algorithms for pattern classification. We optimize the boundaries of intervals and the degrees of overlapping in fuzzy discretization. We use four parameters for each interval I_i: t_i, t_{i+1}, l_i and u_i. The genetic fuzzy membership function is defined as follows:

$$G_0(a) = \begin{cases} 1, \text{ if } \quad v_{min} \le a < u_0, \\ 0, \text{ otherwise.} \end{cases}$$

$$G_i(a) = \begin{cases} 1, \text{ if } \quad l_i \le a < u_i, \qquad i = 1, \ldots, k-2, \\ 0, \text{ otherwise.} \end{cases}$$

$$G_{k-1}(a) = \begin{cases} 1, \text{ if } \quad l_{k-1} \le a \le v_{max}, \\ 0, \text{ otherwise.} \end{cases}$$

where $l_i \le t_i$, $i = 1, \ldots, k-1$, and $t_{i+1} \le u_i$, $i = 0, \ldots, k-2$. In the above, t_i and t_{i+1} indicate the "base" boundary points of interval i. l_i and u_i indicates the left and right boundaries of the interval i, respectively. t_i, l_i and u_i are determined by genetic optimization. Fig.1 shows an example of discretization results.

* This work was supported by Brain Korea 21 Project. The ICT at Seoul National University provided research facilities for this study.

K. Deb et al. (Eds.): GECCO 2004, LNCS 3103, pp. 1303–1304, 2004.

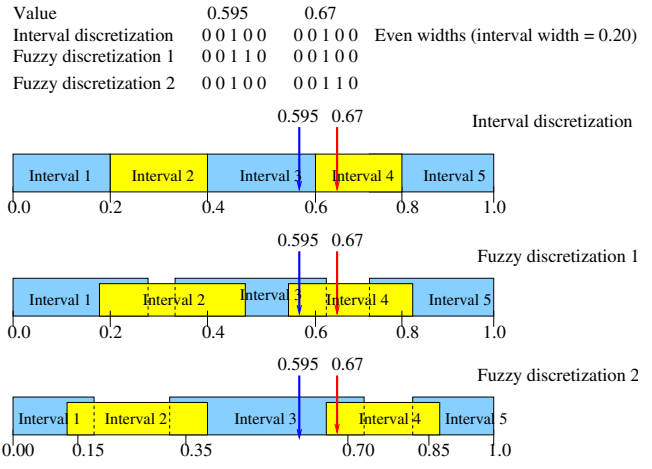

Fig. 1. Three different example discretizations and the corresponding binary masks

We used two well known datasets from the UCI Machine Learning Database Repository[1] to measure the performance of the proposed approach. The domain of each attribute is discretized into five intervals. For robust comparison, we applied the n-fold cross-validation. In order to evaluate the generated data sets, we use artificial neural networks and C4.5 that is a typical top-down method for inducing decision trees. From the simulation results, we can observe that the GAFD improved the classification performance over ID and FD. "ID" indicates the conventional discretization method and "FD" indicates a conventional fuzzy discretization with a pre-specified overlap degree. "GAFD" indicates the suggested genetic fuzzy discretization that evolves the widths of intervals and the degrees of overlap. Considering the group standard deviation($\sigma/\sqrt{n}$) in the case of ANN, we can say that GAFD was better than the others with a confidence level higher than 99% (Table 1).

Table 1. Result on the dataset

Database	WDBC†			WDG‡		
Method	ANN (n=1000)		C4.5	ANN (n=1000)		C4.5
	Avg(%)	$\sigma/\sqrt{n}$	Accuracy(%)	Avg(%)	$\sigma/\sqrt{n}$	Accuracy(%)
ID	94.08912	0.029414	90.32864	81.21342	0.025860	73.66527
FD	95.38696	0.028718	92.26217	81.79459	0.022058	73.56529
GAFD	**96.38463**	0.033088	**93.14554**	**83.11773**	0.024903	**73.98520**

† Wisconsin Diagnostic Breast Cancer.
‡ Waveform Database Generator.

[1] http://www.ics.uci.edu/∼mlearn/MLRepository.html