Genetic Programming for Attribute Construction in Data Mining

Fernando E. B. Otero **Monique M. S. Silva** **Alex A. Freitas**

Pontificia Universidade Catolica do Parana (PUC-PR)
Postgraduate program in applied computer science
Rua Imaculada Conceicao, 1155
Curitiba – PR. 80215-901. Brazil
{fbo, mmonique, alex}@ppgia.pucpr.br
Tel./Fax: (55) (41) 330-1669 (c/o Alex Freitas)
Web page: www.ppgia.pucpr.br/~alex

This paper addresses the classification task of data mining. The goal of attribute construction is to construct new attributes out of the original ones, transforming the original data representation into a new one where regularities in the data are more easily detected.

We use a standard tree-structure representation for each individual. The GP constructs new attributes out of the continuous (real-valued) attributes of the data set being mined. Each individual corresponds to a candidate new attribute. The terminal set consists of all the continuous attributes in the data being mined. The function set consists of four arithmetic operators (+, -, *, /) and two relational operators, namely "≤", "≥". We used tournament selection, standard tree crossover and point mutation. The fitness function was information gain ratio.

The experiments were performed with four public-domain data sets from the UCI data set repository, available at: `http://www.ics.uci.edu/~mlearn/MLRepository.html`. We compared the classification error rate of C4.5 using only the original attributes with the error rate of C4.5 using both the original attributes and the new attribute constructed by the GP. We did experiments with 3 values of the tournament size k and 3 values of maximum tree size (number of nodes). For each of the 9 combinations of these parameters values we ran a 10-fold cross-validation.

The results are reported in Tables 1, 2, 3, and 4. In the title of each table we report the data set and the error rate (in %) obtained by C4.5 using only the original attributes. In the tables themselves, each cell contains the error rate obtained by C4.5 using the attribute constructed by GP. The value of a cell is in bold if the error rate obtained using the attribute constructed by the GP is smaller than the error rate obtained using only the original attributes. The numbers after "±" denote standard deviations. In general, the results can be summarized as follows. In the Abalone data set the attribute constructed by GP led to a slight increase in error rate, but this increase was not significant. In the Wine data set the attribute constructed by GP lead to some reduction in error rate, but again this reduction was not significant. In the other two data sets (Balance-Scale and Waveform) the attribute constructed by GP led to a reduction in error rate which was significant – the corresponding error rate intervals (considering the standard deviations) do not overlap. In general, in the four data sets the differences in error rates associated with different combinations of parameter values was not significant, showing that GP was quite robust to variations in these two parameters.

Table 1: (Abalone) error rate of original attr.: 79.2 ± 0.37

	Maximum tree size (number of nodes)		
k	31	63	127
2	79.31 ± 0.36	**79.18 ± 0.35**	79.21 ± 0.41
4	79.21 ± 0.33	79.21 ± 0.33	**79.16 ± 0.36**
8	79.21 ± 0.33	79.21 ± 0.33	**79.16 ± 0.36**

Table 2: (Balance-scale) error of orig. attr.: 22.42 ± 1.34

	Maximum tree size (number of nodes)		
k	31	63	127
2	**11.47 ± 2.13**	**7.78 ± 0.66**	**8.58 ± 0.58**
4	**9.06 ± 0.42**	**8.26 ± 0.65**	**8.26 ± 0.44**
8	**8.58 ± 0.67**	**8.74 ± 0.68**	**8.10 ± 0.63**

Table 3: (Waveform) error rate of orig. attr.: 25.06 ± 0.66

	Maximum tree size (number of nodes)		
k	31	63	127
2	**23.04 ± 0.40**	**22.48 ± 0.45**	**22.68 ± 0.57**
4	**22.44 ± 0.59**	**22.86 ± 0.50**	**22.56 ± 0.57**
8	**22.22 ± 0.49**	**22.60 ± 0.42**	28.86 ± 2.74

Table 4: (Wine) error rate of original attr.: 6.48 ± 2.05

	Maximum tree size (number of nodes)		
k	31	63	127
2	**5.31 ± 1.38**	**4.72 ± 1.47**	**4.72 ± 1.47**
4	**3.54 ± 1.30**	**5.31 ± 1.63**	**3.54 ± 1.30**
8	**3.54 ± 1.30**	**4.72 ± 1.47**	**5.30 ± 1.62**