Constructing X-of-N Attributes with a Genetic Algorithm

Otavio Larsen Alex A. Freitas Julio C. Nievola

Postgraduate Program in Applied Computer Science
Pontificia Universidade Catolica do Parana
Rua Imaculada Conceicao, 1155. Curitiba – PR. 80215-901. Brazil.
{larsen, alex, nievola}@ppgia.pucpr.br
http://www.ppgia.pucpr.br/~alex

Abstract

The predictive accuracy obtained by a classification algorithm is strongly dependent on the quality of the attributes of the data being mined. When the attributes are little relevant for predicting the class of a record, the predictive accuracy will tend to be low. To combat this problem, a natural approach consists of constructing new attributes out of the original attributes. Many attribute construction algorithms work by simply constructing conjunctions and/or disjunctions of attribute-value pairs. This kind of representation has a limited expressiveness power to represent attribute interactions. A more expressive representation is X-of-N [Zheng 1995]. An X-of-N condition consists of a set of N attribute-value pairs. The value of an X-of-N condition for a given example (record) is the number of attribute-value pairs of the example that match with the N attribute-value pairs of the condition. For instance, consider the following X-of-N condition: X-of-{“Sex = male”, “Age < 21”, “Salary = high”}. Suppose that a given example has the following attribute-value pairs: {“Sex = male”, “Age = 51”, “Salary = high”}. This example has 2 out of the 3 attribute-value pairs of the X-of-N condition, so that the value of the X-of-N condition for this example is 2.

In our GA an individual represents a X-of-N attribute, i.e. the set of N attribute-value pairs composing a X-of-N attribute. Each attribute-value pair is of the form $A_i = V_{ij}$, where A_i is the i-th attribute and V_{ij} is the j-th value belonging to the domain of the A_i. The current version of our GA can cope only with categorical attributes. (Continuous attributes are discretized in a preprocessing step.) The value of N is an integer number varying from 2 to 7. The fitness function is the information gain ratio of the constructed attribute.

In order to evaluate how good the new attributes constructed by the GA are, we have compared the performance of the C4.5 algorithm using only the original attributes with the performance C4.5 using both the original attributes and the new attributes constructed by the GA. Hereafter we refer to the former and to the latter as the original data set and the extended data set, respectively. The performance of C4.5 in both the original data set and the extended data set was measured with respect to the classification error rate. The experiments were done by using public-domain data sets available from http://www.ics.uci.edu/~mlearn/MLRepository.html.

The results are shown in Table 1. The results for the first four data sets of Table 1 were produced by a 10-fold cross-validation procedure. The results for the last three data sets (the monks data sets) were obtained by using the predefined partition of the data into training and test sets. The second and third columns of Table 1 show the error rate obtained by C4.5 in the original data set and the extended data set (with the new X-of-N attributes), respectively. The numbers after the symbol “?” denote standard deviations. For each data set, the difference in the error rates of the second and third columns is deemed to be significant when the two error rate intervals (taking into account the standard deviations) do not overlap. When the error rate of the “original + X-of-N” attributes is significantly better (worse) than the error rate of the “original” attributes, there is a “(+)” ((“-“)) sign in the third column. Note that the X-of-N attribute constructed by the GA significantly improved the performance of C4.5 in three data sets (tic-tac-toe, promoters and monks-2), and it significantly degraded the performance of C4.5 in just one data set (monks-3). In the other three data sets the difference in the error rates was not significant.

Table 1: Error rate obtained by C4.5 in seven data sets

Data Set	Error Rate (%)	
	Original attributes	original + X-of-N attrib.
hepatitis	22.84 ? 1.83	26.34 ? 4.99
Wisc. breast cancer	4.57 ? 0.64	4.83 ? 0.78
tic-tac-toe	14.31 ? 1.14	5.34 ? 0.47 (+)
promoters	18.88 ? 2.19	13.25 ? 1.98 (+)
monks-1	0.00 ? 0.00	0.00 ? 0.00
monks-2	29.60 ? 0,04	0.00 ? 0.00 (+)
monks-3	0.00 ? 0.00	2.80 ? 0,01 (–)

Acknowledgment

This research project is financially supported by Motorola, Jaguariuna –SP, Brazil.

Reference

[Zheng 1995] Z. Zheng. Constructing nominal X-of-N attributes. *Proc. 14th Int. Joint Conf. on Artificial Intelligence,* 1064-1070. Morgan Kaufmann, 1995.