Genetic Algorithm Wrappers for Feature Subset Selection in Supervised Inductive Learning

William H. Hsu **Cecil P. Schmidt** **James A. Louis**

Laboratory for Knowledge Discovery in Databases, Kansas State University
234 Nichols Hall, Manhattan, KS 66506
{bhsu | cps4444 | jal8334}@cis.ksu.edu http://www.kddresearch.org

We derive a validation-based genetic algorithm for feature selection in supervised inductive learning, based upon the following loss functions:

1. **Inferential loss**: Quality of the model produced by an inducer as detected through inferential loss evaluated over a holdout validation data set $D_{val} \equiv D \setminus D_{train}$
2. **Model loss**: "Size" of the model under a specified coding or representation
3. **Ordering loss**: Inference/classification-independent and model-independent measure of data quality given only training and validation data D and hyperparameters $\boldsymbol{\alpha}$

$$f\left(\boldsymbol{\alpha}, D, \bar{\mathbf{I}}_{\mathbf{e}}\right) = a \cdot f_a\left(\boldsymbol{\alpha}, D, \bar{\mathbf{I}}_{\mathbf{e}}\right) + b \cdot f_b\left(\boldsymbol{\alpha}, D\right) + c \cdot f_c\left(\boldsymbol{\alpha}, D\right) \quad (1)$$

$$f_a^{BN}\left(\boldsymbol{\alpha}, D, \bar{\mathbf{I}}_{\mathbf{e}}\right) = 1 - \sqrt{\frac{1}{\sum_{X_i \in \mathbf{X} \setminus \mathbf{E}} a_i} \sum_{X_i \in \mathbf{X} \setminus \mathbf{E}} \sum_{j=1}^{a_i} \left(P'\left(x_{ij}\right) - P\left(x_{ij}\right)\right)^2} \quad (2)$$

$$f_a^{DT}\left(\boldsymbol{\alpha}, D\right) = 1 - \frac{m_{correct}}{m_{val}} \quad (3)$$

$$\textbf{where } m_{correct} \equiv h.classification - accuracy\left(D_{val}.select\left(\boldsymbol{\alpha}\right)\right)$$
$$h \equiv h_0.train\left(D_{train}.select\left(\boldsymbol{\alpha}\right)\right)$$
$$m_{val} \equiv \left|D_{val}\right|$$

$$f_b^{BN}\left(\boldsymbol{\alpha}, D\right) = 1 - \frac{\sum_{i=1}^{n} \left(a_i \cdot max\left(\prod_{\left|X_j \in \mathbf{Pa}_{X_i}\right|} a_j, 1 \right) \right)}{\prod_{i=1}^{n} a_i} \quad (4)$$

$$\textbf{where } a_i \equiv arity\left(X_i, B = \left(\chi, E, \Theta\right)\right)$$
$$\left(E, \Theta\right) = K2\left(\boldsymbol{\alpha}, D_{train}\right)$$

$$f_b^{DT}\left(\boldsymbol{\alpha}, D\right) = 1 - \frac{h.size(\,)}{s_{max}} \quad e.g., s_{max} = m$$

$$f_c^{DT}\left(\boldsymbol{\alpha}\right) = 1 - \frac{\left|\boldsymbol{\alpha}\right|}{n} \quad (5)$$

$$a + b + c = 1 \quad (6)$$

(7)

In related work on genetic wrappers for variable selection in supervised inductive learning, we adapted Equation (3) [HWRC02] from similar fitness functions developed by Cherkauer and Shavlik for decision tree pre-pruning and by Guerra-Salcedo and Whitley for connectionist learning [GW99]. This breadth of applicability demonstrates the generality of simple genetic algorithms as wrappers for performance tuning in supervised inductive learning.

In experiments using the UC Irvine Machine Learning Database repository, this system is shown to be competitive with search-based feature selection wrappers.

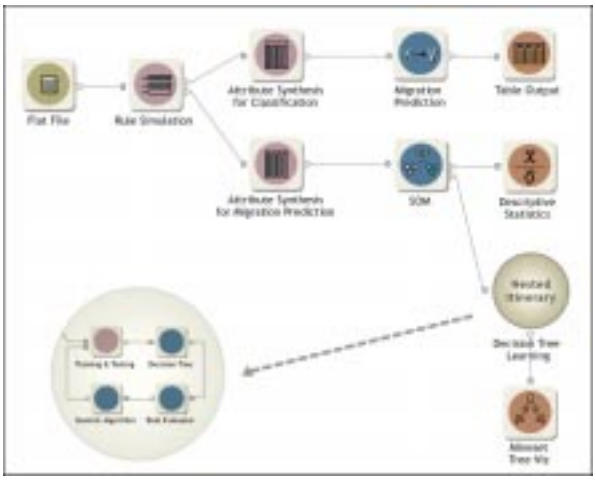

Figure 1. Itinerary for MLJ-CHC

Figure 1 illustrates a real-world application [HWRC02] – automobile insurance risk analysis – that uses the GA wrapper system (depicted in the lower-left inset). Preliminary results on this test bed also indicate that the system is competitive with search-based wrappers.

References

[Be90] D. P. Benjamin, editor. *Change of Representation and Inductive Bias*. Kluwer Academic Publishers, Boston, 1990.

[GW99] C. Guerra-Salcedo and D. Whitley. Genetic Approach to Feature Selection for Ensemble Creation. In *Proceedings of the 1999 International Conference on Genetic and Evolutionary Computation (GECCO-99)*. Morgan-Kaufmann, San Mateo, CA, 1999.

[HWRC02] W. H. Hsu, M. Welge, T. Redman, and D. Clutter. Constructive Induction Wrappers in High-Performance Commercial Data Mining and Decision Support Systems. *Knowledge Discovery and Data Mining*, Kluwer, 2002.