# Genetic Search for Accurate Feature Sets

**Brendan Burns**
Williams College, 2127 Baxter Hall, Williamstown, Mass. 01267
98bdb@cs.williams.edu

This abstract describes a feature selection system, INDiGENT. INDiGENT has been to designed to enhance knowledge based neural networks by genetically searching the set of input features for an optimum subset. This search is designed to enable INDiGENT to make more accurate classifications. Signifigantly, INDiGENT has shown that it can obtain similar increases in accuracy as more complicated theory revision systems.

Expert systems have proven themselves effective decision makers for many types of problems. However, the accuracy of such systems is highly dependent upon the accuracy of the human expert's domain theory. To escape this dependency, many machine learning systems have been developed to automatically refine and correct an expert's domain theory.

Classification systems rely heavily upon having the best (i.e., most relevant) set of input data. Many domains have a variety of potential input features. Therefore, choosing the appropriate set of inputs is critical to the performance of any expert system, and specifically in this case, to the creation of accurate neural networks.

There are a number of systems which use neural network encodings to refine a variety of rule bases including finite state automata (Maclin & Shavlik 1993), certainty-factor rule bases (Mahoney 1996) and first-order horn clause logic (Towell, Shavlik, & Noordewier 1990). INDiGENT uses the KBANN method for encoding domain theories. There has also been work done on genetically refining neural network topologies (Opitz 1995), (Mahoney 1996). There are also a number of systems that utilize a variety of other search methods to find optimum feature sets (e.g. (Kira & Rendell 1992), (Asker & Maclin 1997) ).

INDiGENT operates in the following manner. An initial set of input features is selected. These features are those which are explicitly referenced in the domain theory as well as those features which are not mentioned in the domain but have been identified by the inductive learner C4.5 as significant. This initial feature set is then mutated to create a population for the genetic search. All members of this population are then evaluated using ten-fold cross validation. Two parents are then selected proportional to their measured accuracy. The parents are crossed over and mutated to create a new child. This new child is then evaluated. If its accuracy is better than the worst member of the current population the child replaces the worst member. This process continues for an arbitrary number of generations.

Presently, experiments are complete for three domains. These domains involve the identification of signifigant locations in strands of DNA. INDiGENT's performance on these domains relative to several other existing systems is shown below.

All evaluations were reported using ten-fold cross validation to insure generalization. The results are signifigant because they show that simple feature set selection has the ability to perform on par with a more extensive theory revision system like REGENT. This result is important because it allows for the improvement of accuracy while maintaining the simplicity, and therefore legibility, of the initial domain information. TNT-INDiGENT (Total Neural Topology INDiGENT) has been designed to combine input feature set revision with the hidden layer theory revision performed by REGENT and is currently being evaluated to determine if even greater imrpvements in accuracy are possible.

| System | Promoters | RBS | Splice Junctions |
|---|---|---|---|
| INDiGENT | 95.56% | 92.76% | 94.96% |
| REGENT | 95.83% | 92.17% | 95.92% |
| KBANN | 93.70% | 91.05% | 94.75% |
| C4.5 | 88.0% | 84.8% | |

Table 1: INDiGENT's accuracy compared to several other systems

## References

Asker, L., and Maclin, R. 1997. Feature engineering and classifier selection: A case study in venusian volcano detection. In *The Proceedings of ICML-1997.*

Kira, K., and Rendell, L. 1992. The feature selection problem: Traditional methods and a new algorithm. In *The Proceedings of AAAI-92.*

Maclin, R., and Shavlik, J. W. 1993. Using knowledge-based neural networks to improve algorithms: Refining the Chou-Fasman algorithm for protein folding. *Machine Learning* 11:195–215.

Mahoney, J. 1996. *Combining Symbolic and Connectionist Learning Methods to Refine Certaintty-Factor Rule-Bases.* Ph.D. Dissertation, University of Texas, Austin.

Opitz, D. W. 1995. *An Anytime Approach to Connectionist Theory Refinement: Refining the Topologies of Knowledge-Based Neural Networks.* Ph.D. Dissertation, Department of Computer Sciences, University of Wisconsin-Madison. (Also appears as UW Technical Report.

Towell, G. G.; Shavlik, J. W.; and Noordewier, M. O. 1990. Refinement of approximate domain theories by knowledge-based neural networks. In *Proceedings of AAAI-1990*, 861–866.