

An Overview of Heuristic Knowledge Discovery for Large Data Sets Using Genetic Algorithms and Rough Sets

Alina Lazar, PhD
Youngstown State University

H E U R I S T I C S

Uninformed or blind search, which processes and evaluates all nodes of a search space in the worst case, is not realistic for extracting knowledge from large data sets because of time constraints that are close related to the dimension of the data. Generally, the search space increases exponentially with problem size thereby limiting the size of problems which can realistically be solved using exact techniques such as exhaustive search. An alternative solution is represented by heuristic techniques, which can provide much help in areas where classical search methods failed.

The word "heuristic" comes from Greek and means "to know", "to find", "to discover" or "to guide an investigation". Specifically, "Heuristics are techniques which seek good (near-optimal) solutions at a reasonable computational cost without being able to guarantee either feasibility or optimality, or even in many cases to state how close to optimality a particular feasible solution is." (Russell, Norvig, 1995)

Heuristic refers to any techniques that improves the average-case performance on a problem-solving task but does not necessarily improve the worst case performance. Heuristic techniques search the problem space "intelligently" using knowledge of previously tried solutions to guide the search into fruitful areas of the search space. Often, search spaces are so large that only heuristic search can produce a solution in reasonable time. These techniques improve the efficiency of a search process, sometimes by sacrificing the completeness or the optimality of the solution. Heuristics are estimates of the distance remaining to the goal, estimates computed based on the domain knowledge.

The domain knowledge provides help to heuristics in guiding the search and can be represented in a variety of knowledge formats. These formats include patterns, networks, trees, graphs, version spaces, rule sets, equations, and contingency tables. With regard to heuristics there are a number of generic approaches such as greedy, A* search, tabu search, simulating annealing, and population-based heuristics. The heuristic methods can be applied to a wide class of problems in optimization, classification, statistics, recognition, planning and design.

Of special interest is the integration of heuristic search principles with the dynamic processes in which data becomes available in successive stages, or where

data and inputs are subjects to uncertainties or with large-scale data sets. The integration is a vehicle to generate data driven hypotheses.

The kind of knowledge produced, and the heuristic search algorithm selected, will reflect the nature of the data analysis task. The hypotheses are being represented as sets of decision rules and the extracted rules will be represented in terms of rough sets. Rough sets were selected because of the nature of our data sets.

From a mathematical point of view the problems, can be formulated in terms of the well known, minimal set cover problem, which is a combinatorial optimization problem.

Traditional methods for combinatorial optimization problems are not appropriate here for several reasons. These methods are NP-hard in the worst case and would be costly to use given the size of the data sets. Also, since large data sets are dynamical in nature, adding new data would require running the traditional combinatorial approach again.

The techniques used to solve these difficult optimization problems have slowly evolved from constructive methods, like uniformed search, to local search techniques and to population-based algorithms. Our research goal was to use blend population-based algorithms with methods dealing with uncertainty in order to induce rules from large data sets.

U N C E R T A N T Y A N D E V O L U T I O N

Population-based heuristic methods are iterative solution techniques that handle a population of individuals which are evolving according to a given search strategy. At each iteration, periods of self-adaptation (mutations) alternate with periods of cooperation (crossover), and periods of competition (selection). The population-based heuristic search (Conrad, 1978) is dependent of the following components: the knowledge representation for the specific problem we want to solve and the search strategy or the evolution process. The adaptability of an individual represents its ability to survive in an uncertain environment. Artificial Intelligence researchers have explored different ways to represent uncertainty (Russell, Norvig, 1995): belief networks, default reasoning, Dempster-Shafer theory, fuzzy sets theory, rough sets theory.

For the problems we want to solve, the learning task will require a representation that explicitly deals with uncertainty. The evolutionary learning methods that are employed must be able to work with such a representation. In this chapter we look first at basic ways to represent uncertainty in developing rules. And, then we will investigate how that uncertain knowledge can be used to direct evolutionary search and learning.

Uncertainty, as well as evolution, is a part of nature. When humans describe complex environments, they use linguistic descriptors of real-world circumstances, which are often not precise, but rather "fuzzy". The theory of fuzzy sets (Zadeh, 1965) provides an effective method of describing the behavior of a system which is too complex to be handled with the classical precise mathematical analysis.

The theory of rough sets (Pawlak, 1991) emerged as another mathematical approach for dealing with uncertainty that arises from inexact, noisy or incomplete information. Fuzzy sets theory assumes that the membership of the objects in some set is defined as a degree ranging over the interval $[0,1]$. Rough sets theory focuses

on the ambiguity caused by the limited distinction between objects in a given domain.

Fuzzy sets have been employed to represent rules generated by evolutionary learning systems. Using fuzzy concepts, Valenzuela-Rendon (1997) tried to overcome the limitations of the conventional rule-based classifier system (Holland, 1975) when representing continuous variables. He used fuzzy logic to represent the results of the genetic-based search of the classifier system.

Likewise, fuzzy functions have been used to describe and update knowledge in Cultural Algorithms. First, Reynolds (1994) employed a fuzzy acceptance and influence function in the solution of real-valued constrained optimization problems. Following the same idea Zhu designed a fully fuzzy cultural algorithm (Zhu, Reynolds, 1998) which included a fuzzy knowledge representation scheme in order to deal with the continuous variables (Zhu, Reynolds, 1998) in the belief space, as well as a fuzzy acceptance and influence function. All these approaches were tested on real-values function optimization problems. More recently, Jin (2000) used a "fuzzy" knowledge representation for normative knowledge in the belief space of cultural algorithms, to solve the real-valued constrained function optimization.

The design of a fuzzy representation system is not an easy job, because of the membership functions should be carefully chosen, and the procedures that use these functions should specified precisely. The problem is to optimize the fuzzy membership functions for a problem and to find optimum plans related to the fuzzy performance measures. It is natural approach to use heuristics (i.e. evolutionary algorithms) to solve this task.

Another approach to represent uncertainty is with rough sets. Rough sets are based on equivalence relations and set approximations, and the algorithms for computing rough set properties are combinatorial in nature. Wroblewski (1995) implemented a genetic algorithm for computing reducts, based on permutation code as well as a "greedy" algorithm. Another approach for building reducts is described by Vinterbo (2000) and it is based on the set cover problem, in particular on finding minimal hitting sets using a classical genetic algorithm. Finding a minimal set of decision rules or a satisfactory set is an NP-complete problem. Agotnes (1999) used a genetic algorithm to build an optimal set of decision rules, where the fitness function was based on the quality of each rule. In conclusion, there are many hybrid methods that integrate evolutionary algorithms and other methods from soft computing, methods such as rough sets.

Evolution can be defined in one word, "adaptation" in an uncertain environment. Nature has a robust way of dealing with the adaptation of organisms to all kind of changes and to evolve successful organisms. According to the principles of natural selection, the organisms that have a good performance in a given environment, survive and reproduce, whereas the others die off. After reproduction, a new generation of offspring, derived from the members of the previous generation is formed. The selection of parents from these offspring is often based upon fitness. Changes in the environment will affect the population of organisms through the random mutations. Mayr said that "Evolution is a dynamic, two-step process of random variation and selection" (Fogel, 1995). Using examples from natural systems and theories of adaptive behavior researchers have been trying to build heuristic evolutionary learning systems.

Evolutionary algorithms are heuristic optimization methods inspired from natural evolution processes. Currently there are three basic population-only mechanisms that model evolution: genetic algorithms, evolutionary strategies and evolutionary programming. Each of the methods models the evolution of a

population of individuals at a different scale and applies election and reproduction operators to find an individual that is fit with regard of the fitness function. The genetic algorithm models evolution at the gene scale, but evolutionary strategies and evolutionary programming, model evolution at the species level.

The cultural algorithms (Reynolds, 1994) approach adds another level to the evolutionary process inspired from the human societies and cultural evolution. It adds to the population space, belief space. The belief space will be a collection of symbolic knowledge that will be used to guide the evolution of the population.

Besides the rule based methods, decision trees are well known for their inductive learning capabilities. Any decision tree can be reformulated as a set of rules. One of the problems related to the decision trees is finding the smallest decision tree. Simple heuristics can solve the problem. Researchers have tried to integrate Genetic algorithms with decision tree learning in order to solve complex classification problems. Bala (1997) applied the above methodology for difficult visual recognition problems involving satellite and facial image data. Other researchers combined the genetic algorithms or evolutionary strategies with neural networks.

Reynolds (2000) investigated the use of cultural algorithms to guide decision tree learning. The data was taken from a real world archeological database, with a collection of sites found in Valley of Oaxaca, Mexico. The problem was to localize the sites that present evidence of warfare as opposed with those that did not.

The goal was to employ evolution-based techniques to mine a large-scale spatial data set describing the interactions of agents over several occupational periods in the ancient valley of Oaxaca, Mexico. Specifically, we want to extract from the data set spatial constraints on the interaction of agents in each temporal period. These constraints will be used to mediate the interactions of agents in a large-scale social simulation for each period and will need to be checked many times during the course of the simulation.

One of the major questions was how to represent the constraint knowledge. Popular data mining methods such as decision trees work well with data collected in a quantitative manner. However, the conditions under which the surface survey data was collected here introduced some uncertainty into the data. Would a representation that explicitly incorporated uncertainty into its structure produce a more efficient representation of the constraints here that one that did not? This is important since the complexity of the constraint set will impact the complexity of the simulation that uses those rules.

Here, we use genetic algorithms to guide the search for a collection of rough set rules to describe constraints on the location of particular types of warfare in the Valley. Since warfare was a major factor in the social evolution in the Valley, the constraints reflecting its spatial and temporal patterning are important ingredients in the model. The rules generated are compared with those produced by a decision tree (Reynolds, 2000) algorithm. In each of the phases examined, the best rule set that used the Rough Set representation always had fewer conditions in it, and the average rule length was less than that for the decision tree approach in every case but one. In that case they were equal. The differences were most marked in those periods where the warfare patterns were most complex. It was suggested that the differences reflect the inclusion of noise factors as explicit terms in the Decision tree representation and their exclusion in the rough sets approach.

A comparison (table 1) of two decision systems from the first period where the two approaches begin to show larger differences in rule and condition number, Rosario, demonstrates that the Rough Set approach has a fewer percentage of

inconclusive rules and a larger percentage of conclusive ones than for the decision tree approach.

Table 1: Comparison between Decision Trees and Rough Set Rule Induction

	Decision Trees	Rough Set Rules
Advantages	Easy to understand	Very Expressive Modular knowledge Good with missing data They handle imprecise data
Disadvantages	May be difficult to use with continuous data They look at simple combinations of attributes They need to break numeric fields into fixed ranges Not very good with inexact data Not flexible No way to handle missing data Can not easily approach large data sets May have over fitting Less accurate predictions	Can be memory intensive Can be computational intensive

In addition, the rough set approach needs to evaluate fewer conditions relative to the inconclusive ones than the decision tree approach. These differences, it is argued, result from the explicit consideration of uncertainty into a period that is more complex and more prone to the introduction of such uncertainty than previous periods.

F U T U R E T R E N D S

The focus of the comparisons here was on the syntactic or structural differences in the decision systems produced. In future work a comparison of the semantic differences will be accomplished by using the approaches to produce alternative ontologies in the agent-based simulation and assess the differences that are produced. In other words, do the syntactic differences reflect semantic differences in simulation model performance? And, what impact does the use of uncertainty to represent ontological knowledge of the agents have on the basic simulation results.

C O N C L U S I O N

Genetic algorithms, as population-based algorithms, are good vehicles in which to build meta-level heuristics to guide the search more efficiently. That knowledge, here we will use rough sets concepts, or rules, can be employed to direct the evolutionary search. The rules can reflect spatial and temporal patterns that will guide the generation of new candidate search objects by the evolutionary engine. The spatial and temporal continuity of the data will facilitate this process.

REFERENCES

Agotnes, T., "Filtering Large Propositional Rule Sets while Retaining Classifier Performance," Master's thesis, Norwegian University of Science and Technology, Department of Computer and Information Science, Feb. 1999.

Bala, J., Jong, K.D., Huang, J., Vafaie, H. and Wechsler, H., "Using Learning to Facilitate the Evolution of Features for Recognizing Visual Concepts," *Evolutionary Computation*, vol. 4, no. 3, pp. 297–311, 1997.

Conrad, M., "Evolution of Adaptive Landscape," in *Theoretical Approaches to Complex Systems* (R. Heim and G. Palm, eds.), vol. 21 of Springer Lecture Notes in Biomathematics, pp. 147–169, Springer-Verlag, 1978.

Fogel, D.B., *Evolutionary Computation - Toward a New Philosophy of Machine Learning*. IEEE PRESS, 1995.

Holland, J.H., *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI: University of Michigan Press, 1975.

Jin, X. and Reynolds, R.G., "Using Knowledge-based Systems with Hierarchical Architecture to Guide Evolutionary Search," *International Journal of Artificial Intelligence Tools*, vol. 9, pp. 27–44, March 2000.

Lazar, A. and Sethi, I.K., "Decision Rule Extraction from Trained Neural Networks Using Rough Sets," in *Intelligent Engineering Systems Through Artificial Neural Networks* (C. H. Dagli, A. L. Buczak, and J. Ghosh, eds.), vol. 9, (New York, NY), pp. 493–498, ASME Press, Nov. 1999.

Lazar, A. and Reynolds, R.G., (2001) "Evolution-based Learning of Ontological Knowledge for a Large-scale Multi-agent Simulation", submitted at The Fourth International Workshop on Frontiers in Evolutionary Algorithms (FEA 2002), Research Triangle Park, North Carolina, USA, March 8-13, 2002

Nazzal, A.H., (1997) *Learning Site-Settlement Patterns From Large-Scale Spatial-Temporal Databases With Cultural Algorithms*, Wayne State University, Ph.D. Thesis.

Pawlak, Z., *Rough Sets - Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers, 1991.

Reynolds, R.G., "An Introduction to Cultural Algorithms," in *Proceedings of the Third Annual Conference on Evolutionary Programming*, River Edge, NJ (A. V. Sebald and L. J. Fogel, eds.), pp. 131–139, World Scientific Publishing, 1994.

Reynolds, R.G., "The Impact of Raiding on Settlement Patterns in the Northern Valley of Oaxaca: An Approach Using Decision Trees", in *Dynamics in Human and Primate Societies*, Ed. T. Kohler, and G. Gummerman, Oxford University Press, 2000, pp: 251-274.

Russell, S.J., and Norvig, P., *Artificial Intelligence a Modern Approach*. Prentice Hall, Upper Saddle River, New Jersey, 1995.

Valenzuela-Rendon, M., "Reinforcement Learning in the Fuzzy Classifier System," tech. rep., Monterrey: ITESM, Campus Monterrey, Centro de Inteligencia Artificial, 1997.

Vinterbo, S., and Øhrn, A., "Approximate Minimal Hitting Sets and Rule Templates," *International Journal of Approximate Reasoning*, 25(2), pp. 123–143, 2000.

Wroblewski, J., "Finding Minimal Reducts Using Genetic Algorithms," in *Proceedings of Second International Joint Conference on Information Science*, pp. 186–189, Sept. 1995.

Zadeh, L., "Fuzzy Sets," *Information and Control*, vol. 8, pp. 338–353, 1965

Terms and Definitions

Knowledge Discovery: in data sets is the process of identifying valid, novel, potentially useful, and ultimately understandable patterns/models in data.

Data mining: is a step in the knowledge discovery process that, under some acceptable computational efficiency limitations, finds patterns or models in data.

Heuristics: A rule of thumb, simplification, or educated guess that reduces or limits the search for solutions in domains that are difficult and poorly understood. Unlike algorithms, heuristics do not guarantee optimal, or even feasible, solutions and are often used with no theoretical guarantee.

Evolutionary Computation: Computer-based problem solving systems that use computational models of evolutionary processes as the key elements in design and implementation.

Genetic Algorithms: An evolutionary algorithm which generates each individual from some encoded form known as a "chromosome" or "genome". Chromosomes are combined or mutated to breed new individuals. "Crossover", the kind of recombination of chromosomes found in sexual reproduction in nature, is often also used in GAs. Here, an offspring's chromosome is created by joining segments chosen alternately from each of two parents' chromosomes which are of fixed length.

Uncertainty: Information or data that is often imprecise, incoherent, and incomplete.

Fuzzy Set Theory: Fuzzy set theory replaces the two-valued set-membership function with a real-valued function, that is, membership is treated as a probability, or as a degree of truthfulness.

Rough Set Theory: Rough set theory is a new mathematical tool to deal with vagueness and uncertainty. Any vague concept is replaced by a pair of precise concepts - called the lower and the upper approximation of the vague concept. The lower approximation consists of all objects which surely belong to the concept and the upper approximation contains all objects which possibly belong to the concept.