

FINAL TECHNICAL REPORT

GRANT #: N00014-00-1-0769

PRINCIPAL INVESTIGATORS: William H. Hsu (bhsu@cis.ksu.edu), Shing I Chang (changsi@ksu.edu)

INSTITUTION: Kansas State University

Department of Computing and Information Sciences (CIS)

Department of Industrial and Manufacturing Systems Engineering (IMSE)

GRANT TITLE: *Real-time Temporal Probabilistic Inference using Bayesian Networks: Decision Support in Manpower and Personnel Management*

AWARD PERIOD: 01 Jul 2000 – 01 Sep 2001

OBJECTIVE: (1) To analyze historical data on personnel assignment for simulation-based decision support and optimization, by building Bayesian network models from data using known **structure learning** algorithms (K2), **variable selection wrappers**, and **parameter estimation** algorithms such as expectation maximization (EM) for calibrating a Bayesian network with known structure; (2) to develop fast, parallel, and approximate algorithms for inference using the learned networks; (3) to formulate a **simulation-based monitoring** test bed for assignment and distribution; (4) to apply **sampling-based inference** in Bayesian networks to decision support problems in selection and classification.

APPROACH: We have formulated several machine learning and probabilistic inference problems pertaining to simulation-based monitoring in personnel science. Specifically, we have developed probabilistic inference specifications for decision support problems in the following areas:

- *Assignment*
 - o Problem definition: hybrid qualitative/quantitative inference in Bayesian networks
 - o Observables: personnel demographics, vocational battery and other aptitude test scores, qualitative evaluation (instructor, supervisor, detailer)
 - o Unknowns: probabilistic ranking measures (softmax, other generalized linear models)
 - o Approaches: **sampling-based approximate inference** (completed 2000-2001), adaptive dynamic programming, multi-attribute decision making (multiple intermediate criteria)
- *Distribution*
 - o Problem definition: inference in Markov models, dynamic Bayesian networks
 - o Observables: *as above*; in addition: transfer records, introspective evaluation, elicited utility or preference (not yet implemented)
 - o Unknowns: ranking measures for multiple personnel
 - o Approaches: **first-order Markov model learning** (completed Summer, 2000), sampling-based approximate inference, adaptive dynamic programming, multi-objective decision making (multiple intermediate criteria)
- *Selection*
 - o Problem definition: as for assignment, with different (more general) observables and criteria
 - o Observables: as for assignment, but with less historical evaluation data
 - o Unknowns: probabilistic ranking measures *for selection*
 - o Approaches: **BN structure learning** (completed Summer, 2000; improved and reimplemented Summer, 2001); sampling-based approximate inference; comparison with **fuzzy outranking** (completed Spring, 2001)
- *Classification*

- Problem definition: as for selection and assignment, with additional observables and more specific criteria
- Observables: as above; in addition: task-specific aptitude tests, qualifications, qualitative interview data, professional credentials, work history
- Unknowns: probabilistic ranking measures *for selection*
- Approaches: BN structure learning; sampling-based approximate inference; comparison with **decision tree induction** (completed May, 2001) and **Simple Bayesian inference** (completed July, 2001)

The above specifications facilitate application, to personnel management, of the generic probabilistic reasoning tools we have developed for decision support and expert systems. The penultimate goal of developing stochastic sampling (simulation) algorithms for inference has been to **predict and monitor** suitability of personnel for specified tasks or task groups under real-world conditions:

- changing requirements and preferences (utility)
- skill set migration (due to reassignment, attrition)

The ultimate goal is to use this information to generate decisions using techniques such as:

- Kalman filtering
- Multi-attribute decision making (e.g., value iteration)
- Multi-objective decision making (e.g., Pareto optimization)

Parameter estimation for these procedures can be done using error backpropagation in feedforward artificial neural networks, as well as other gradient-based maximum likelihood estimation techniques and expectation-maximization (EM). During preliminary research in 2000, however, we found that like many decision support problems, the above selection and assignment problems have ill-defined independent variables. That is, **relevant** variables are not fully identified. Traditional approaches to this problem include principal components analysis, independent components analysis, factor analysis, and other dimensionality-reducing transforms (e.g., self-organizing maps). Another approach uses **wrappers** for supervised inductive machine learning, which we implemented between May, 2000 and June, 2001.

ACCOMPLISHMENTS (throughout award period):

Based upon the above requirements, we have developed a suite of reusable Java-based software modules for **reasoning under uncertainty** in the *personnel science domain* – specifically, outranking measures for selection and predictive models for assignment and distribution. The results of this distributed data mining application will be compared to known historical observations and Bayesian optimal classification and prediction over these observations.

We have focused on development of input, data preparation (sampling, aggregation, and online analytical processing), structure learning, and sampling-based inference, especially adaptive importance sampling, and **visualization and evaluation** modules for constructing graphical models of probability – specifically, discrete Bayesian networks (BNs). During 2000-2001, we have developed modules in Java that implement *K2*, a BN structure learning algorithm, a gradient-based parameter estimation algorithm for learning conditional probability tables from data, and five stochastic importance sampling algorithms:

1. Forward simulation – completed March 2001
2. Probabilistic logic sampling (rejection sampling) – completed April 2001
3. Backward sampling – completed May 2001
4. Heuristic importance sampling and self importance sampling – completed June 2001
5. Adaptive importance sampling – completed July 2001

Based upon the first prototype, a redesign phase (July – August 2001), produced a reimplementations of K2 and modules 1-5 in September – October 2001.

In an independent component of this effort, we have developed the following Java-based implementations of machine learning codes:

1. *ID3* (decision tree induction), ported from MLC++ – completed January 2001
2. *C4.5* (decision tree induction), ported from code by J. Ross Quinlan – completed May 2001, revised July 2001
3. Simple (Naïve) Bayes, adapted from MLC++ – completed June 2000, redesigned and reimplemented June 2001, third and final revision October 2001
4. Search-based wrapper for feature subset selection (ported from MLC++) – completed June 2001
5. Genetic wrapper for feature subset selection (ported from MLC++) – first prototype November 1999, second prototype September 2001
6. Wrapper for genetic programming hyperparameter (inductive bias) optimization – completed January 2001

Experiments using these modules have led to one (1) journal paper based on 1, 3, and 5 in the second list, two (2) conference papers based on the same modules, one (1) workshop paper based on module 4, one (1) workshop paper and two (2) conference papers based on module 6, one (1) refereed book chapter based on the entire second list, and two (2) conference papers in preparation based on the first list and on a genetic wrapper for variable selection and ordering in *integrative* Bayesian network structure learning, parameter estimation, and inference (*aka* fusion, propagation, and structuring). This represents a total of

- 2 journal papers and book chapters (1 accepted, 1 under review)
- 6 refereed workshop and conference papers (4 published, 2 in preparation)

CONCLUSIONS: Adaptive importance sampling has been shown to be competitive with other importance sampling algorithms in efficiency and accuracy, with the added benefit of being very robust in the presence of unlikely evidence. Preliminary studies of the Enlisted Master File completed since the conclusion of this project indicate that this problem environment is typical in personnel science. A genetic wrapper has been shown to provide improvement in variable selection and hyperparameter optimization and we further hypothesize that it can be used to provide data-driven adaptation in Bayesian network learning and inference (e.g., in finding good variable orderings and subsets for structure learning).

SIGNIFICANCE: Our studies have provided information on to how an integrated learning and Bayesian network inference system can be applied to classification and prediction in a decision support system for personnel management.

PATENT INFORMATION: N/A

AWARD INFORMATION: N/A

REFEREED PUBLICATIONS (for total award period):

[GH00] S. M. Gustafson and W. H. Hsu. Genetic Programming for Strategy Learning in Soccer-Playing Agents: A KDD-Based Architecture. In *Proceedings of the [Genetic and Evolutionary Computation Conference \(GECCO-2000\) Workshop Program](#)*, Las Vegas, NV, July, 2000.

[GH01] S. M. Gustafson and W. H. Hsu. Layered Learning in Genetic Programming for a Cooperative Robot Soccer Problem. In *Proceedings of the [4th European Conference on Genetic Programming \(EuroGP-2001\)](#)*, Lake Como (Milan), Italy, April, 2001.

[HWRC00] W. H. Hsu, M. Welge, T. Redman, and D. Clutter. Genetic Wrappers for Constructive Induction in High-Performance Data Mining. In *Proceedings of the [Genetic and Evolutionary Computation Conference \(GECCO-2000\)](#)*, Las Vegas, NV, July, 2000.

[HCGG00] W. H. Hsu, Y. Cheng, H. Guo, and S. Gustafson. Genetic Algorithms for Reformulation of Large-Scale KDD Problems with Many Irrelevant Attributes. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2000)*, Las Vegas, NV, July, 2000.

[HG01] W. H. Hsu and S. M. Gustafson. Genetic Programming for Layered Learning of Multi-agent Tasks. In *Late-Breaking Papers of the [Genetic and Evolutionary Computation Conference \(GECCO-2001\)](#)*, San Francisco, CA, June, 2001.

[HWRC02] W. H. Hsu, M. Welge, T. Redman, and D. Clutter. Constructive Induction Wrappers for High-Performance Data Mining. *International Journal of Knowledge Discovery and Data Mining*, Kluwer Academic Publishers, to appear.

[Sc01] C. P. Schmidt. A Filter Approach using a Committee Machine of Wrappers. In W. H. Hsu, H. Kargupta, H. Liu, and N. Street, eds. *Working Notes of the Workshop on Wrappers for Performance Enhancement in Knowledge Discovery in Databases (ML-5), International Joint Conference on Artificial Intelligence (IJCAI-01)*. Seattle, WA, 04 August 2001.

BOOK CHAPTERS, SUBMISSIONS, ABSTRACTS AND OTHER PUBLICATIONS (for total award period)

[HKLS01] W. H. Hsu, H. Kargupta, H. Liu, and N. Street, eds. *Working Notes of the Workshop on Wrappers for Performance Enhancement in Knowledge Discovery in Databases (ML-5), International Joint Conference on Artificial Intelligence (IJCAI-01)*. Seattle, WA, 04 August 2001.

[Hs02] W. H. Hsu. Control of Inductive Bias in Supervised Learning using Evolutionary Computation: A Wrapper-Based Approach. Book chapter, in revision for J. Wang, ed., *Data Mining: Opportunities and Challenges*, IDEA Press, 2002.

[GHHS02] H. Guo, E. Horvitz, W. H. Hsu, and E. Santos, eds. *Working Notes of the Joint Workshop on Real-Time Decision Support and Diagnosis, AAAI/UAI/KDD-2002*. Edmonton, Alberta, Canada, 28 July 2002, to appear.