Parallel Knowledge Discovery from Large Complex Databases

Diane J. Cook and Lawrence B. Holder

Department of Computer Science Engineering
University of Texas at Arlington
Box 19015, Arlington, TX 76019
cook@cse.uta.edu, holder@cse.uta.edu

NASA Grant NAS5-32337 (5555-16)
Final Report

1 Research Orientation

This report on grant NAS5-32337 begins by restating the objectives, approach, and arguments from the original proposal *Parallel Knowledge Discovery from Large Complex Databases.* The following section will describe progress made before this report period, progress made during the current report period, and planned activities for the next period. The last section will describe the educational benefits that have been derived from this effort.

1.1 Parallel Knowledge Discovery

NASA is focusing on grand challenge problems in Earth and space sciences. Within these areas of science, new instrumentation will be providing scientists with unprecedented amounts of unprocessed data. Our goal is to design and implement a system that takes raw data as input and efficiently discovers interesting concepts that can target areas for further investigation and can be used to compress the data. Our approach will provide an intelligent parallel data analysis system.

This effort will build upon two existing data discovery systems: the SUBDUE system used at the University of Texas at Arlington, and NASA's AUTOCLASS system. AUTOCLASS has been used to discover concepts in several large databases containing real or discrete valued data. SUBDUE, on the other hand, has been used to find interesting and repetitive *structure* in the data. Although both systems have been successful in a variety of domains, they are hampered by the computational complexity of the discovery task. The size and complexity of the databases expected from Earth and Space Science (ESS) program will demand processing capabilities found in parallel machines.

This effort focuses on combining the two approaches to concept discovery and speeding up the discovery process by developing a parallel implementation of the systems. The two discovery systems will be combined by using SUBDUE to compress the data fed to AUTOCLASS, and letting SUBDUE evaluate the interesting structures in the classes generated by AUTOCLASS. The parallel implementation of the resulting AUTOCLASS/SUBDUE system will be run on a massively-parallel machine (nCUBE), and will be tested for speedup on a number of large databases used by the ESS program.

1.2 Goals Stated in the Proposal

A general goal of the proposed research was to develop a combined discovery system (SUBDUE + AUTOCLASS) to be applied to a variety of earth and space science databases. Our objective was to improve overall system speed, discovery power, and generality, by

- Improving the existing SUBDUE discovery system for application to structural earth and space databases;
- Using the SUBDUE system as a pre- and post-processor for NASA's AUTOCLASS discovery system; and
- Speeding up the combined discovery process through improved algorithms and parallel implementations on MIMD machines such as the nCUBE and Connection Machine 5.

2 Recent Work and Planned Work

This section organizes the previous four months' work (already done under the current grant) and the next four months' work (to be done under the grant) around issues in machine discovery.

2.1 Parallel Subdue / AutoClass

The first step to parallelizing the SUBDUE / AutoClass discovery system is to ensure that both systems use the same programming language, and that a language be chosen that is supported by a majority of parallel systems. Joe Potts, one of our students who worked with the NASA Ames group in summer 1994, completed a port of AutoClass from Lisp to C. That port is now finished and is available from our anonymous ftp site (csr.uta.edu). The NASA Ames group has announced this distribution and is quite happy with the results and with the ability to distribute the code, as they cannot distribute the code from NASA. We have already received several requests for the C version of AutoClass from Lockheed, NASA Ames, and the Jet Propulsion Laboratory.

One benefit of Mr. Potts' stay at Ames was a redesign of the parallelization strategy for AutoClass. The initial strategy for parallelizing at a high level was abandoned after discussions with members of the Ames group indicated the size of typical data sets (e.g., Landsat images) would require too much data redundancy across processors. The high-level approach was set aside in favor of a low level approach that would spread the large number of observations out over the processors rather than distributing classes, classifications or attributes.

The SIMD parallelization of AutoClass on the Connection Machine 5 is complete. Joe Potts will be defending his Masters thesis soon on the design and evaluation of AutoClass' parallelization. Through an NSF equipment grant awarded to Dr. Cook, the department has acquired a 128-processor nCUBE. Ports of both parallel AutoClass and Subdue will be performed for this machine as well.

A MIMD parallelization of Subdue is now complete, using a 128-processor nCUBE. Gehad Galal is working on MIMD-parallel and distributed version of the Subdue system. The parallel version of the system allocates distinct sets of substructures for each processor to evaluate and grow. Load balancing is performed as processors run out of substructures to consider. The parallel system is completed and we are currently performing speedup experiments. A distributed version of the system, in which the database itself is distributed over multiple workstation connected with a high-speed switch, will be implemented this summer.

In addition, a MIMD-parallel version of AutoClass is underway. In this version, distinct processors are given a distinct number of classes into which to fit the data. The MIMD-parallel version of AutoClass will be finished this spring, at which time we can compare performance between MIMD and SIMD versions of the system and combine parallel version of Subdue and AutoClass.

2.2 Improvements to the Subdue System

The two ongoing directions for the improvement of Subdue are the inclusion of facilities for accepting background knowledge to guide the discovery process and the evaluation and refinement of Subdue to

work with structural features extracted from image data.

2.2.1 Domain-Dependent Discovery in Subdue

SUBDUE is designed as a domain-independent tool for discovering concepts in structured data. To make SUBDUE's discovery process more useful across a wide variety of domains, domain knowledge can also be used to guide the discovery process. We expect that compressing the graph using combinations of domain-independent and domain-dependent knowledge will increase the chance of finding interesting substructures and realize even greater compression.

Work continues on Subdue's ability to include both domain-dependent and domain-independent background knowledge. The domain knowledge can be input using a hierarchical collection of substructures known to be interested for the given domain, using a collection of domain-specific graph match rules, or using a set of feature extraction functions. All forms of domain-dependent search guidance have been incorporated into the Subdue system. Experiments have been performed in the domains of program analysis and CAD analysis to compare the results of discovery with domain-independent knowledge to discovery with domain-dependent knowledge and discovery with both types of knowledge.

Recently, a probabilistic version of Subdue in which background knowledge can be encoded as probabilistic models was completed. The results of this work have been submitted to the Artificial Intelligence Journal. Surnjani Djoko, the main student responsible for integrating domain knowledge into Subdue, successfully defended her Ph.D. dissertation on this topic in August 1995.

2.2.2 Processing Images with Subdue

With the eventual goal of processing NASA imagery using our proposed integration of Subdue and AutoClass, we are evaluating Subdue's ability to discover patterns from structural features extracted from images. The computer vision portion of the project involves processing images using standard image processing techniques (e.g., edge detection). Low level image data is extracted from the images, and then this data is transformed into higher level symbols which are input to the Subdue system. The main region features that will be extracted are lines and angles between lines.

The results of these experiments have been submitted to the Florida AI Research Symposium. Stephen Poe, the student mainly responsible for evaluating the use of Subdue on image data, successfully defended his Masters thesis on this topic in August 1995.

2.3 Integration of Subdue and AutoClass

The integration of the structural discovery process in Subdue and the attribute-value clustering process in AutoClass will yield a system capable of discovering previously-undiscoverable patterns using the combination of structural and non-structural data features of the data. Our initial design for the integration was to use SUBDUE as a pre-processor of the structural component of the data in order to construct new non-structural attributes for addition to the set of existing, non-structural attributes. The new data, augmented with this non-structural information about the structural regularities in

the data, can now be passed to the AUTOCLASS system. The structural information will bias the classifications preferred by AUTOCLASS towards those consistent with the structural regularities in the data.

Discussions with the AutoClass group at NASA Ames have suggested a different approach to the integration in which Subdue is included as a parameterized model with AutoClass. AutoClass evaluates classes by tuning the parameters of statistical models of the classes (e.g., single normal distribution for each attribute). Expressing Subdue as a parameterized model whose parameters select the types of substructures preferred by Subdue would allow a more seamless integration of the two systems.

When both Subdue and AutoClass have been ported to the nCUBE machine, we plan to evaluate both approaches to the integration of SUBDUE and AUTOCLASS by comparing the results of AUTOCLASS alone on data with a weak, non-structural classification and a strong, structural classification. We also plan to compare previous AUTOCLASS results with the classifications obtained with the integrated discovery system using the same data augmented with natural structural information such as temporal and spatial orientations.

2.4 Graphical User Interface

In order to make Subdue easier to use and the results easier to interpret, a Master's student will be developing a graphical user interface to Subdue this summer. This interface will display the input database (or selected portions of the database) as a graph, and visually display substructures as they are discovered.

3 Educational Benefits of This Work

One Ph.D. student and two master's students have complete theses based on research funded by this NASA grant. In addition, we are currently supporting one Ph.D. student and one Master's student to work on this project.

Surnjani Djoko finished her Ph.D. dissertation in August 1995 on investigating methods for incorporating domain-dependent knowledge into a discovery system, and is currently working at Bell Northern Research. Joe Potts, a masters student, used a portion of the NASA money to visit NASA Ames in summer 1994. Mr. Potts has completed the conversion of AutoClass from Lisp to C and the parallelization of AutoClass on the CM 5. Mr. Potts defended his masters thesis in December of 1996 and plans to continue investigating related issues for his doctoral work. Stephen Poe obtained his master's degree during this report period on the topic of substructure discovery in image data. Gehad Galal will finish his Master's thesis this May on the topic of designing parallel and distributed versions of Subdue.

The research sponsored by this NASA grant has contributed to the course Dr. Cook is teaching on Parallel Algorithms for Artificial Intelligence. The results of the parallel discovery algorithms has been incorporated into the new syllabus so that additional students can benefit from ongoing research and contribute to the state-of-the-art in efficient machine discovery methods.

Many of the issues involved in discovery of patterns in NASA-related data sets have been included in Dr. Holder's course on Machine Learning. These issues have been incorporated into the projects done by the students in the class. Knowledge of these issues has given the students a better appreciation for the difficulties and potential benefits of machine discovery in such domains containing both structural and non-structural information. Master's student Stephen Poe became interested in this project through discussions in Dr. Holder's class.

4 Publications

Publications supported by this NASA project.

1. S. Djoko, D. J. Cook, and L. B. Holder. An Empirical Study of Domain Knowledge and its Benefits to Substructure Discovery, to appear in *IEEE Transactions on Knowledge and Data Engineering.*

2. G. Galal and D. J. Cook, "Exploiting Parallelism in a Scientific Discovery System to Improve Scalability", to appear in *Proceedings of the Tenth Annual Florida AI Research Symposium*, 1997.

3. S. Djoko, D. J. Cook, and L. B. Holder. Discovering Informative Structural Concepts Using Domain Knowledge. *IEEE Expert*, 10, pages 59–68, 1996.

4. D. J. Cook, L. B. Holder, and S. Djoko. Knowledge Discovery from Structural Data, *Journal of Intelligence and Information Sciences*, Volume 5, Number 3, pages 229–245, 1995.

5. S. Djoko, D. J. Cook and L. B. Holder. Analyzing the Benefits of Domain Knowledge in Substructure Discovery. In the *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, pages 75–80, 1995.

6. D. J. Cook and L. B. Holder. Substructure Discovery Using Minimum Description Length and Background Knowledge. In *Journal of Artificial Intelligence Research*, Volume 1, pages 231–255, 1994.

7. S. Djoko. Guiding Substructure Discovery with Minimum Description Length and Background Knowledge, Student Abstract in *Proceedings of the Twelfth National Conference on Artificial Intelligence*, page 1442, 1994.

8. L. B. Holder, D. J. Cook, and S. Djoko. Substructure Discovery in the Subdue System, in *Proceedings of the AAAI Workshop on Knowledge Discovery in Databases*, pages 169–180, 1994.

Related publications submitted prior to this report period.

1. L. B. Holder and D. J. Cook. Discovery of Inexact Concepts from Structural Data. In *IEEE Transactions on Knowledge and Data Engineering*, Volume 5, Number 6, pages 992–994, 1993.

2. L. B. Holder, D. J. Cook, and H. Bunke, Fuzzy Substructure Discovery, *Ninth International Machine Learning Conference*, Aberdeen, Scotland, pages 218–223, 1992.

3. L. B. Holder. Empirical substructure discovery. In *Proceedings of the Sixth International Workshop on Machine Learning*, pages 133–136, 1989.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
September 1996

3. REPORT TYPE AND DATES COVERED
Contractor Report

4. TITLE AND SUBTITLE
Parallel Knowledge Discovery from Large Complex Databases
(Final Report)

5. FUNDING NUMBERS
NAS5-32337
Task: 5555-16

6. AUTHOR(S)
Diane J. Cook and Lawrence B. Holder

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS (ES)
Department of Computer Science Engineering
University of Texas at Arlington
Box 19015
Arlington, TX 76019

8. PEFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS (ES)
National Aeronautics and Space Administration
Washington, DC 20546-0001

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
NASA/CR-1999-209466

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Unclassified–Unlimited
Subject Category: 43
Report available from the NASA Center for AeroSpace Information,
7121 Standard Drive, Hanover, MD 21076-1320. (301) 621-0390.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

NASA is focusing on grand challenge problems in Earth and space sciences. Within these areas of science, new instrumentation will be providing scientists with unprecedented amounts of unprocessed data. Our goal is to design and implement a system that takes raw data as input and efficiently discovers interesting concepts that can target areas for further investigation and can be used to compress the data. Our approach will provide an intelligent parallel data analysis system.

14. SUBJECT TERMS
SUBDUE system, AUTOCLASS system, nCUBE, discovery system

15. NUMBER OF PAGES
7

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL