



Non-Confidential Description

Portfolio of Data Mining Tools for Extremely Large Datasets

Technology Case: RFT-0075, 0159, and 0203

Invention Summary

Standard data mining techniques have been sufficient for some areas of information analysis where the datasets are small enough that analysis can be performed relatively quickly and efficiently. However, these standard data mining techniques, such as association rule mining (ARM), have not been as successful in areas such as bioinformatics, nanotechnology, VLSI design, and spatial data, which each have extremely large data sets and where mining implicit relationships among the data can be prohibitively time-consuming.

These NDSU-developed data mining technologies are designed specifically for organizing extremely large datasets into an efficiently usable form. The organizational format of the data takes into account the fact that different bits of data can have different degrees of contribution to value. For example, in some applications, high-order bits alone may provide the necessary information for data mining, making the retention of all data unnecessary.

Benefits

- Highly efficient data mining approach specifically targeted for extremely large data sets.
- Dramatic decreases in processing time and increases in system performance.
- Data mining operations reduced from hours in duration to virtually instantaneous.
- Applicable to bioinformatics, nanotechnology, VLSI design, spatial data, remote sensed imagery, and other applications with large data sets.

Invention Premise

1. **System and Method for Organizing, Compressing and Structuring Data for Data Mining Readiness (RFT-075).** A method of structuring data in a data-mining-ready format represented by a basic tree structure, as shown in Figure 1. This is done by creating one tree structure for each bit position of the binary form of the data contained in the database. Once this bit-sequential format of the data is achieved, the formatted data is structured into a tree format that can be mined for patterns virtually instantaneously. **This invention has one issued US patent and one US patent pending.**

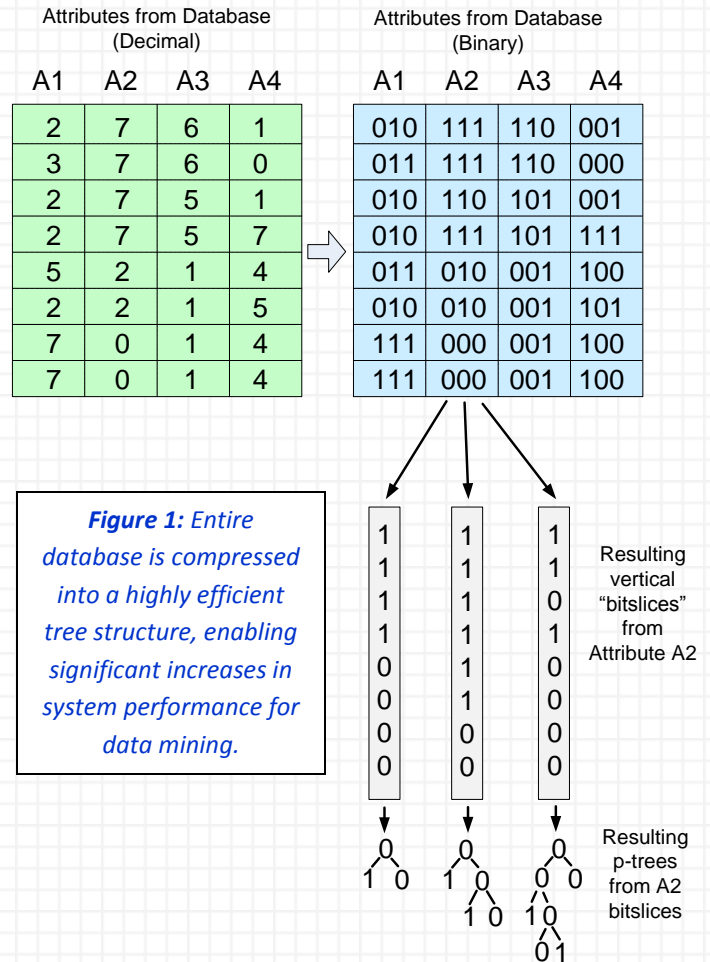
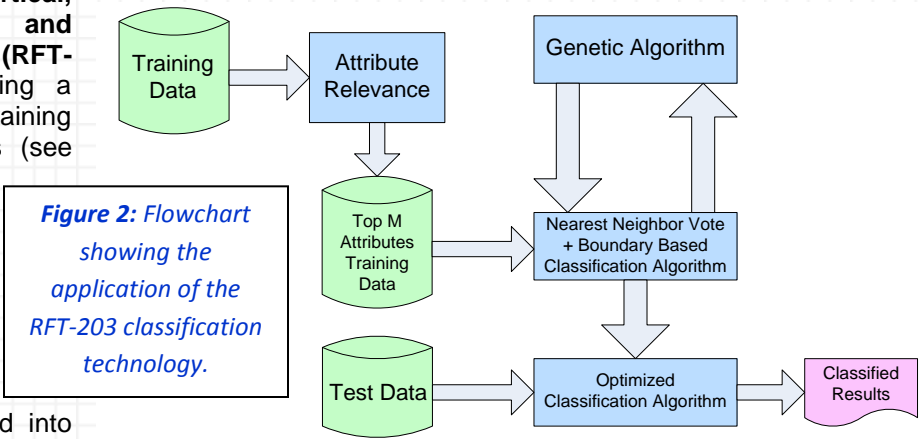


Figure 1: Entire database is compressed into a highly efficient tree structure, enabling significant increases in system performance for data mining.

2. **Vertical Set Inner Product Technology (VSIP) with Predicate Trees (RFT-159).** This invention builds upon the data mining tree structure described above (in RFT-075) by then operating on the basis data tree structure with algebraic techniques to create clusters of related data items that can be easily analyzed. **This invention has one US patent pending and one pending PCT application.**

3. **Parameter Optimized, Vertical, Nearest-Neighbor-Vote and Boundary-Based Classification (RFT-0203).** A system of classifying a subject data item based on a training set of pre-classified data items (see Figure 2). A piecewise-linear approximation of a local boundary between different classes of data items is automatically computed. The local boundary is approximated by a neighborhood set of data items selected from the training set that have been pre-classified into different classes and have points similar to a point of the subject data item. **This invention has one issued US patent.**



The Lead Inventor



William Perrizo, Ph.D.
NDSU Distinguished Professor
Department of Computer Science

Dr. William Perrizo is North Dakota State University Distinguished Professor and Fargo-Moorhead Chamber of Commerce Distinguished Professor of Computer Science at North Dakota State University. He received his Ph.D. from the University of Minnesota, Minneapolis, in 1972, his M.S. from the University of Wisconsin, Madison, in 1967 and bachelor's degree from St. John's University in 1965. Dr. Perrizo has over 200 refereed publications including over 50 journal papers. He has been a Research Scientist at IBM ABS in Rochester, MN, a Research Scientist at the U.S. Air Force Electronic Systems Division at Hanscom Air Force Base, MA, a Visiting Professor at the University of Minnesota, and a Visiting Assistant Professor at Oregon State University. Dr. Perrizo's expertise is in Database Systems, Data Mining, Knowledge Discovery, Distributed Database Systems, High Performance Computer Systems, Communications Networks, Precision Agriculture, Bio-informatics, and Remotely Sensed Imagery Analysis.

Inquiries

Jonathan Tolstedt, Licensing Associate/Patent Agent
NDSU Research Foundation, Fargo, ND 58108-6050
Phone: 701-231-8173 Fax: 701-231-6661
Email: jtolstedt@ndsurf.org
Web: www.NDSUResearchFoundation.org