Innovative Massively Parallel AI Applications

David L. Waltz
Thinking Machines Corporation
Cambridge, MA and
Brandeis University
Waltham, MA

Abstract

Massively parallel applications must address problems that will be too large for workstations for the next several years, or else it will not make sense to expend development costs on them. Suitable applications include one or more of the following properties: 1) large amounts of data; 2) intensive computations; 3) requirement for very fast response times; 4) ways to trade computations for human effort, as in developing applications using learning methods. Most of the suitable applications that we have found come from the general area of very large databases. Massively parallel machines have proved to be important not only in being able to run large applications, but in accelerating development (allowing the use of simpler algorithms, cutting the time to test performance on realistic databases) and allowing many different algorithms and parameter settings to be tried and compared for a particular task. This presentation summarizes four such applications.

The applications described are: 1) prediction of credit card "defaulters" (non-payers) and "attritters" (people who didn't renew their cards) from a credit card database; 2) prediction of the continuation of time series, e.g. stock price movements; 3) automatic keyword assignment for news articles; and 4) protein secondary structure prediction. These add to a list identified in an earlier paper [Waltz 90] including: 5) automatic classification of U.S. Census Bureau long forms, using MBR – Memory-Based Reasoning [Creecy et al. 92, Waltz 89, Stanfill & Waltz 86]; 6) generating catalogs for a mail order company that maximize expected net returns (revenues from orders minus cost of the catalogs and mailings) using genetically-inspired methods; and 7) text-based intelligent systems for information retrieval, decision support, etc.

1 Data Analysis – Predicting defaulters and attritters

In a cooperative project with Citicorp, Craig Stanfill of Thinking Machines led an effort that tested the performance of a number of different AI learning as well as statistical methods for predicting behavior of credit card holders. In each of the experiments, a system was trained on a portion of the database, and then tested on another portion. Training samples for defaulters used fifteen variables (total charges, total paid, number of charges, balance, months overdue, bankruptcy indicator, etc.) for a card holder as input, and the binary data on pay/default three months ahead as the target

output. About ten different methods were tested in all. The best methods tested, in order of effectiveness on this problem, included CART (Classification and Regression Trees), 2-dimensional additive regression, 100 nearest neighbor MBR, and 1-dimensional additive regression. The major advantage of massive parallelism for these tasks is the ability to run a large number of large-scale experiments in a short time. Without fast response times, one might well be tempted to use a small subset of a database, and to stop once a method provided what seemed like reasonable results.

2 Time series prediction

In the fall of 1991, Xiru Zhang of Thinking Machines and Jim Hutchinson, a grad student at MIT and part-time Thinking Machines employee, entered the Santa Fe Institute time series prediction contest. Two main time series, 100,000 points from a chaotic physical system and the other 30,000 points from a financial time series, were made available on the Internet. The goals were to provide the most accurate prediction for the next 500 points for each series. Zhang and Hutchinson used backpropagation neural nets to solve both these problems. To decide on the size and structures of the networks, they analyzed each by computing the autocorrelation function for each time series, and finding the pattern of the mean value for each time series, averaged over fairly large windows. The autocorrelations gave an idea of how large a window would be appropriate: quite wide (20-30 input units) for the chaotic time series, and narrow (seven input units) for the financial time series. The financial time series also used ternary inputs (–1, 0, +1) since nearly all adjacent samples stayed the same or moved up or down by a fixed quantum. 500 different nets were trained for each of the series. Zhang and Hutchinson's entries proved to be the winners for each of the contests. (The physical system was the motion of a particle in a time-varying four-dimensional force field, sampled at very short evenly spaced intervals; the financial times series was the exchange rate of the Swiss franc vs. the dollar over several weeks.) The entire project was completed in two weeks, again indicating the value of being able to do a number of large-scale experiments in a short time. For more information, see [Zhang and Hutchinson 1992].

3 Automatic Keyword Assignment

This project was done in conjunction with Dow Jones. The goal was to assign keywords to each news article. A training sample was provided of 32,000 articles, each typically assigned from six to eight keywords by human editors. The method used here was a variant of MBR. A vector similarity text retrieval system (CMDRS – Connection Machine Data Retrieval System) was used to find the 16 articles nearest to an article to be keyworded. Each of these near neighbors was assigned a score, and had some number of keywords attached to it. The union of the keywords for all 16 near neighbors was formed, and each keyword was given a score equal to the sum of the scores of all the near neighbors in which it occurred. The system then saved the eight highest scored keywords, provided each exceeded a threshold, and assigned them to the new article.

We were able to test this method against human keyworders. We mixed keywords assigned automatically to articles with keywords assigned by human editors, and gave the randomized collection to expert human editors to evaluate. These expert editors graded each keyword as relevant, irrelevant, or borderline. Counting borderline as irrelevant, the automatic system achieved a recall

of .8 and a precision of .72, which compares quite well to human performance of .82 recall and .88 precision [Masand et al. 1992]. More recent work has used a variant of Koza's genetic algorithm methods [Koza 92] to evolve an expression that is used to decide whether to acceptor reject each keyword assignment. Using this method, [Masand 93] has shown that if it is allowed to refer 8% of the articles to humans for keywording, this system can achieve performance better than human editors on the remaining 92% of the articles.

The original system required two months of effort by a two person team. The genetic algorithm enhancement has required an additional person-month.

4 Protein secondary structure prediction

In his Ph.D. thesis work at Brandeis University, much of it done while a part-time employee of Thinking Machines, Xiru Zhang implemented a system that now holds the world's record for accuracy in protein secondary structure prediction [Zhang et al. 92]. It is now easy to analyze amino acid sequences; the sequences for over 20,000 proteins are now known. However, it is very difficult to find the three-dimensional structures of proteins; fewer than 1000 are known, and a large fraction of these are hemoglobins. The reason is that, to find three-dimensional ("tertiary") structure, researchers must crystallize the protein, and then perform X-ray diffraction analysis. This process typically requires 2-3 years time for two or three researchers; moreover, some proteins apparently cannot be crystallized, and so resist tertiary structure analysis. Computer methods for finding tertiary structures for unknown proteins has also proved intractable to date, and most researchers have worked on the simpler problem of finding the portions of an amino acid sequence that correspond to helices, to sheets (two-dimensional structures that consist of parallel bonded strand sections of the amino acid sequence), or to "coil" (anything that isn't a helix or sheet).

Until recently the best secondary structure results had used a backpropagation neural net [Sejnowski & Qian 89]. Zhang redid this work, but with a larger training set (about 110 non-homologous proteins with helix-sheet-coil labelings) drawn from the Brookhaven database [Kabsch & Sander 85], and achieved similar results. He also tried other methods, including a statistical method he devised, which was about as accurate as the neural net, and MBR, which was a little better than the other two. However, Zhang noticed that these three methods only agreed with each other about 80% of the time. He thus devised a hybrid architecture that used all three methods (neural nets, statistics and MBR) separately, and then combined the results of these three using another neural net. The resulting system performed with about 3% greater accuracy than any other system built to date, an increase that is highly significant statistically. Again, the ability to do a number of experiments, each of a large size, led to significantly increased performance.

5 Summary

Massively parallel machines have proved useful for a number of large database-related tasks. It has generally been important to be able to perform many experiments, and to do so on large amounts of data. Learning and memory-based methods have required modest amounts of time to program,

and have yielded excellent performance when compared to other programs as well to to humans.

References

Creecy, R., B. Masand, S. Smith & D. L. Waltz. "Trading MIPS and Memory for Knowledge Engineering." CACM 35, 8, August 1992, 48-64.

Kabsch, W. & C. Sander. "Dictionary of protein secondary structures: pattern recognition off hydrogen-bonded and geometrical features." Biopolymers 22, 1983, 2577-2637.

Koza, J. Genetic Programming: On the Programming of Computers by Means on Natural Selection. Cambridge: MIT Press, 1992.

Masand, B. "Effects of query and database sizes on classification of news stories using memory based reasoning." AAAI Spring Symposium on CBR, Stanford, March 1993.

Masand, B., G. Linoff, and D. L. Waltz. "Classifying news stories using memory-based reasoning." Proc. SIGIR Conf., Copenhagen, July 1992.

Qian, N. & T. J. Sejnowski. "Predicting the secondary structure of globular proteins using neural network models." J. Molecular Biology 202, 1988, 865-884.

Stanfill, C. & D. L. Waltz. "Toward Memory-Based Reasoning." CACM 29, December 1986, 1213-1228.

Waltz, D. L. "Memory-Based Reasoning." In M. Arbib and A. Robinson (eds.) Natural and Artificial Parallel Computation, Cambridge: MIT Press, 1989, 251-276.

Waltz, D. L. "Massively Parallel AI." Proc. AAAI-90, Boston, 1117-1122.

Zhang, X. & J. Hutchinson. "Practical Issues in Nonlinear Time Series Prediction." To appear in A. Weigend and N Gershenfeld (eds.), Predicting the Future and Understanding the Past: Proceedings of the 1992 Santa Fe Institute Time Series Competition, Addison-Wesley, 1993.

Zhang, X., J. Mesirov, & D. L. Waltz. "A Hybrid System for Protein Secondary Structure Prediction." J. Molecular Biology 225, 1992, 1049-1063.