# Supercomputer Data Mining - Project Overview

*Firat Tekiner and Mike Pettipher*
*Research Support Services, Manchester Computing, University of Manchester*

The aim of this project is to produce a supercomputing data mining tool for use by the UK academic community which utilises a number of advanced machine learning and statistical algorithms for large datasets. This is a joint project with the University of West England (UWE), Bristol where Dr. Larry Bull is the PI and the University of East Anglia (UEA) and is funded by EPSRC under the new applications to HPC.

Dealing with the massive quantity and diversity of data generated by research and industry presents one of the defining challenges to data mining. The huge size of many databases presents an opportunity to discover previously unobserved patterns. However, analysing large data is the major issue in data mining as processing such data in a timely manner would not be possible. Therefore, the aim of the supercomputer data mining project is to produce a supercomputing data mining resource for use by the UK academic community. In particular, a number of evolutionary computing-based algorithms and the ensemble machine approach will be used to exploit the large-scale parallelism possible in supercomputing. Moreover, the issue of the scalability of data mining algorithms is often not addressed experimentally and evaluation is most commonly conducted on relatively small datasets [1]. It is not fully understood how significant differences in data mining algorithms on these small data sets translate to large data sets, therefore, we aim to explore the relationship between algorithms and data size as part of this project [2].

Parallelisation at a high level will be considered where large number of independent tasks (data mining algorithms) are performed on different chunks of data rather than parallelising the algorithm itself. In other words data parallelism is adopted instead of task parallelism. The simultaneous consideration of outputs from more than one algorithm for a given input presented to all algorithms, e.g. by majority voting, has been found to give improved performance over the traditional use of a single algorithm [3]. This ensemble approach would appear to be ideally suited for use on a supercomputing resource since each algorithm operates independently and so can be executed in parallel on different processors before a final output is determined; with a large number of processors available, the potential use of very large (heterogeneous) ensembles becomes possible for large datasets.

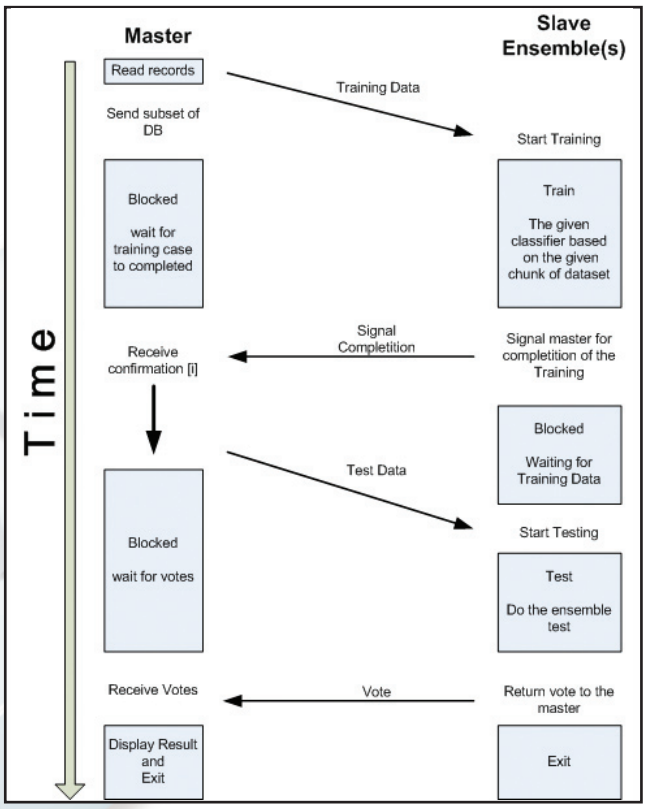


Figure 1: Master-slave implementation of data mining flow chart.

In Figure 1, master process instantiates the data miners on the other processors and sends a subset of the data to each learner (slave) (ultimately MPI/IO will be used to avoid this initial communication). The Learner contains a data mining technique which has two methods; train and test. The Learner causes the algorithm to repeat the train/test cycle until some threshold or timeout is reached, Figure 2. When all Learners complete the training method and signal termination, master sends the reserved data to all learners to run the ensemble test and receives their output. Initially four data mining techniques [4], namely, naïve bayes, kth nearest neighbour, decision trees (c4.5 and c5.0) and support vector machines will be implemented in serial algorithms that will be combined through the use of ensemble techniques.

Data management is one of the key issues in any data mining application and it will be one of the most important factors that affect the problems that can be studied. The amount of data may be extremely large ranging from hundreds of Megabytes to Terabytes. The Newton service provides 1 Terabyte of shared memory and it will be fully exploited in this project. However, despite the advantages of this environment, data access will be a major challenge, and one of the algorithmic objectives will be to mask the data access time from the disk. The best scenario would be to read the data in advance so that it is available when the processor is ready for it. Then, it does not matter how long it takes to access the data.

Once the initial objectives are achieved advanced techniques will be used to improve data management issues as well as the performance of the algorithms themselves.

## References


[1] C.L. Blake and C.J. Merz. UCI repository of machine learning databases, 1998.

[2] I. M. Whittley, A. J. Bagnall, L. Bull, M. Pettipher, M. Studley and F. Tekiner, "On the evaluation of MFS and FASBIR k nearest neighbour algorithms on large data sets", Submitted to IDA05, 2005.

[3] T. Dietterich, An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. Machine Learning 40(2): 139-158, 2000.

[4] M. Dunham, Data mining introductory and advanced topics, Prentice Hall, 2002.


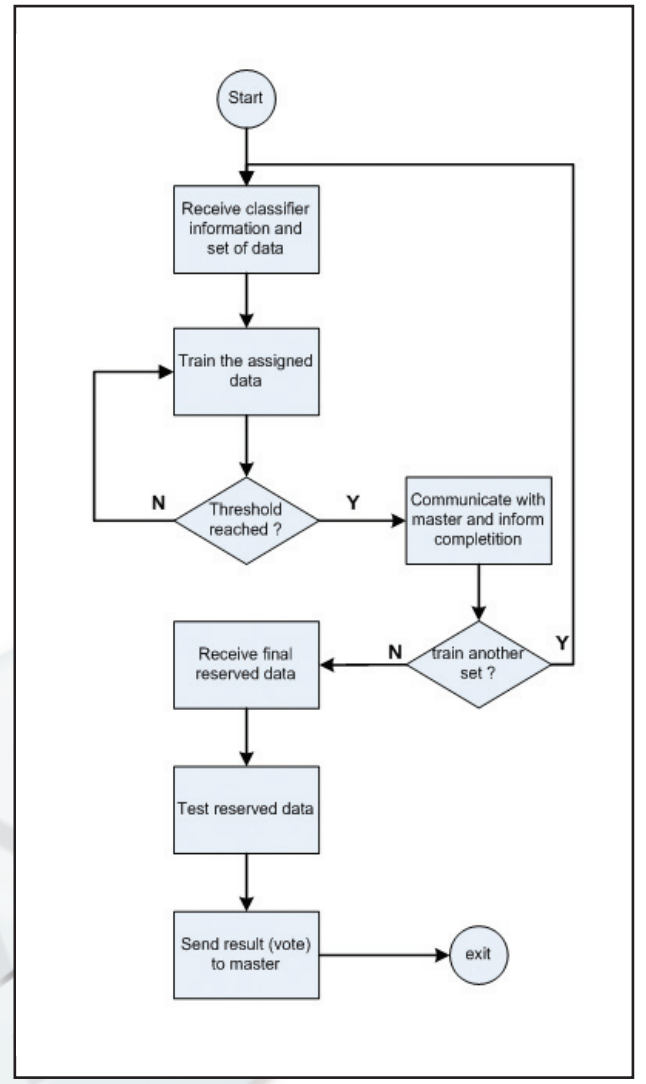


Figure 2: Data miner flow chart.


CRAY Europe 2005 TECHNICAL WORKSHOP

Cray Technical Workshop Europe
September 20-22, 2005

**Registration is FREE!**

Register online at www.cray.com/europe
or email workshop@cray.com

We look forward to seeing you in Switzerland!

Hosted by CSCS in Manno in beautiful Ticino, Switzerland, the 2nd annual Cray Technical Workshop offers the opportunity to hear directly from users and Cray experts about current scientific advancements using Cray systems. You will also receive a preview into the technical details on current and upcoming Cray products. The prime focus of the 2005 Technical Workshop will be MPP computing, experiences with Red Storm and the Cray XT3 system.

CSCS Swiss National Supercomputing Centre

**Featured Speakers include:**

Bill Camp, Sandia National Laboratories
Thomas Zacharia, ORNL
Marie-Christine Sawley, CSCS
Michael Levine, Pittsburgh Supercomputing Center
Sanzio Bassini, CINECA
Jack Dongarra, University of Tennessee