The e-Science and Data Mining Special Interest Group: Launch, Aims and Preliminary Requirements Analysis

Niall Adams

Department of Mathematics, Imperial College

Jim Austin

Department of Computer Science, University of York

Lisa Blanshard

CCLRC e-Science Centre, Daresbury Laboratory

Ken Brodlie

School of Computing, University of Leeds

Yike Guo

Department of Computing, Imperial College

Bob Mann

Institute for Astronomy and National e-Science Centre, University of Edinburgh

Bob Nichol

Institute for Cosmology and Gravitation, University of Portsmouth

Adrian Shepherd

School of Crystallography, Birkbeck College

Amos Storkey

School of Informatics, University of Edinburgh

Abstract

The mini-workshop on "e-Science and Data Mining" at AHM2004 was proposed by the authors of this paper in their guise of the steering group of the nascent e-Science and Data Mining Special Interest Group (esdm-sig) as an opportunity to seek the participation of the UK e-science community in its work. In this paper we summarise the goals of the esdm-sig and the oral version presented at AHM2004 will outline preliminary results from the requirements analysis that the steering group will conduct ahead before then, as a means of initiating discussion of the place of data mining within UK e-Science.

1 From SDMIV to *esdm-sig*

One of the principal motivations for the e-Science programme is the data avalanche being experienced in many scientific disciplines. A consequence of this is the necessity for a change in the way that science is done. Many conventional analysis techniques will not scale to the exponentially increasing data volumes found in many scientific domains, nor do they typically take advantage of the multitude of distributed data sources that may be relevant to a given science problem. The data mining and visualization communities within computer science may have much to offer here, and much progress has been made in recent years in developing techniques for extracting salient features from large, multi-dimensional datasets.

This work has more often found motivations and applications from the commercial world than from science, and, to redress this, NeSC hosted a workshop on Scientific Data Mining, Integration and Visualization (*SDMIV*)[1] in October 2002. The *SDMIV* workshop brought together about fifty people, ranging from software engineers developing Grid infrastructure software, to computer scientists with expertise in data mining and visualization, to applications specialists from a wide range of disciplines, including astronomy, atmospheric science, bioinformatics, chemistry, digital libraries, engineering, environmental science, experimental physics, marine sciences, oceanography and statistics.

The *SDMIV* workshop identified a set of common problems related to data mining facing the e-Science community, but few potential solutions. A year on, the situation is different, as a number of e-Science projects (notably DAME[2] and DiscoveryNet[3]) have working data mining systems operating in an e-Science environment, and it is timely to bring their experience to bear on the more general problem of data mining in e-Science. This has led to the setting up of the *e-Science Data Mining Special Interest Group* (*esdm-sig*), which seeks to broaden the bandwidth between the data mining community within academic computer science and those within the e-Science communities who wish to make use of data mining techniques to aid their research, plus the software engineers building the computational infrastructure within which this activity will take place. The interest engendered by the *SDMIV* workshop illustrates the interest in this topic within the e-Science community, as does the awarding of two of the four *e-Science International Sister Project* grants to collaborations in this area.

2. esdm-sig Requirements Analysis

One of the initial tasks of the *esdm-sig* will be to conduct a thorough study of the data mining requirements and expertise within the e-Science community, and the implications this has for the further development of e-Science middleware. The *esdm-sig* steering group will initiate this process by way of a questionnaire circulated throughout the UK e-Science community, and the oral version of this mini-workshop paper will present the preliminary results from the analysis of that survey, as well as seeking further input to it.

We need to understand the kinds of data mining activity being planned and undertaken within the UK e-Science community, the practical problems being experienced and the solutions that have been found. We also need to identify where further computer science research is needed – e.g. perhaps in improving the scalability of data mining algorithms so that they can handle the volume and dimensionality of the data becoming available within e-Science applications – as well as the software engineering challenges involved in mining distributed data, as often found within the e-Science domain. We also need to understand better the role of visualization in guiding the mining process, and in understanding the results from it.

Those interested in the work of the *esdm-sig* are welcome to contact its convenor Bob Mann (*rgm@roe.ac.uk*).

References:

1. Scientific Data Mining, Integration and Visualization (SDMIV) workshop report: www.nesc.ac.uk/talks/sdmiv/report.pdf
2. DAME project: www.cs.york.ac.uk/dame
3. DiscoveryNet: www.discovery-on-the.net