Super Computer Heterogeneous Classifier Meta-Ensembles

Anthony Bagnall, University of East Anglia, UK
Gavin Cawley, University of East Anglia, UK
Ian Whittley, University of East Anglia, UK
Larry Bull, University of West of England, UK
Matthew Studley, University of West of England, UK
Mike Pettipher, University of Manchester, UK
Firat Tekiner, University of Manchester, UK

ABSTRACT

This article describes the entry of the Super Computer Data Mining (SCDM) Project to the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) 2006 Data Mining Competition. The SCDM project is developing data mining tools for parallel execution on Linux clusters. The code is freely available; please contact the first author for a copy. We combine several classifiers, some of them ensemble techniques, into a heterogeneous meta-ensemble, to produce a probability estimate for each test case. We then use a simple decision theoretic framework to form a classification. The meta-ensemble contains a Bayesian neural network, a learning classifier system (LCS), attribute selection based-ensemble algorithms (Filtered Attribute Subspace based Bagging with Injected Randomness [FASBIR]), and more well-known classifiers such as logistic regression, Naive Bayes (NB), and C4.5.

Keywords: data mining; high performance computing; machine learning; PAKDD competition

INTRODUCTION

This article describes the entry of the SCDM Project to the PAKDD 2006 Data Mining Competition. The SCDM project is funded by the Engineering and Physical Sciences Research Council of the UK government and began in January 2005. The objective of the project is to develop data mining tools, implemented in C++ and MPI, for parallel execution on super computer architecture. The main super computer facility we use is based at the University of Manchester and is called CSAR. It forms part of the UK high performance computing (HPC) service. However, the SCDM code will run on any cluster and will be
freely available to the academic community. The SCDM toolkit has already made a valuable contribution to several research projects (Bagnall, Ratnamahatana, Keogh, Lonardi, & Janacek, in press; Bagnall, Whittley, Janacek, et al., 2006; Bagnall, Whittley, Studley, et al., 2006; Whittley et al., 2006) and we are encouraging the more widespread usage of the software. More details can be found at the project Web site (still under development) (Tekiner, 2006) or by contacting the first author of this article. Our motivation for this project is to develop tools that will be able to perform complex analysis of very large data with many attributes. We have assembled several large attributes, many cases data sets to form a standard test suite for the assessment of data mining algorithms on this type of data (Bagnall, 2006). The main algorithmic focus of the project is on ensemble techniques, with a particular emphasis on attribute selection and evolutionary algorithms. This competition has been particularly useful for us in developing algorithms and implementations. The PAKDD competition data set:

- makes a useful addition to the data collection;
- provides a test bed for our implementations of existing algorithms (for this work k-NN, C4.5, Naive 3, Bayes, Neural Network, Support Vector Machine [SVM], and Logistic Regression); and
- allows us to assess new variations of classifiers (LCSs) and attribute selection based ensemble algorithms (FASBIR).

The structure of this article is as follows. We first describe our understanding of the problem and the data. Then we detail the classifiers used in the ensemble. In the following section we analyse the individual and ensemble performance of the classifiers, and finally we summarise the conclusions we can draw from the results and outline the future direction of the SCDM project.

This research is supported by EPSRC grants GR/T18479/01, GR/T18455/01, and GR/T/18462/01

**Problem Understanding**

The PAKDD 2006 Data Mining Competition involves building classifiers to predict whether customers will choose a third generation (3G) or second generation (2G) phone. Our approach to this type of problem is normally to adopt a data mining methodology such as Cross Industry Standard Process for Data Mining (CRISP) (Shearer, 2000). Our ability to gain a good business understanding through interaction with the customer is obviously limited due to the nature of the competition. Nevertheless, a structured approach is always beneficial.

**Business Understanding**

Our only source of business understanding other than the data itself comes from the competition Web site. “An Asian telco operator which has successfully launched a third generation (3G) mobile telecommunications network would like to make use of existing customer usage and demographic data to identify which customers are likely to switch to using their 3G network.” This makes it clear that predicting 3G usage is, in itself, just the first stage in the data analysis, as predicting switching is a more complex task requiring repeated measurements on the same users. However, the outputs from predicting usage could easily be used to make decisions about how to market for switching. There