Chapter 4.22
Comparing Four-Selected Data Mining Software

Richard S. Segall
Arkansas State University, USA

Qingyu Zhang
Arkansas State University, USA

INTRODUCTION

This chapter discusses four-selected software for data mining that are not available as free open-source software. The four-selected software for data mining are SAS® Enterprise Miner™, Megaputer PolyAnalyst® 5.0, NeuralWare Predict® and BioDiscovery GeneSight®, each of which was provided by partnerships with our university. These software are described and compared by their existing features, characteristics, and algorithms and also applied to a large database of forest cover types with 63,377 rows and 54 attributes. Background on related literature and software are also presented. Screen shots of each of the four-selected software are presented, as are future directions and conclusions.

BACKGROUND

Historical Background

Han and Kamber (2006), Kleinberg and Tardos (2005), and Fayyad et al. (1996) each provide extensive discussions of available algorithms for data mining.

Algorithms according to StatSoft (2006b) are operations or procedures that will produce a particular outcome with a completely defined set of steps or operations. This is opposed to heuristics that according to StatSoft (2006c) are general recommendations or guides based upon theoretical reasoning or statistical evidence such as "data mining can be a useful tool if used appropriately."

The Data Intelligence Group (1995) defined data mining as the extraction of hidden predictive

information form large databases. According to The Data Intelligence Group (1995), "data mining tools scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations."

Brooks (1997) describes rules-based tools as opposed to algorithms. Witten and Frank (2005) describe how data mining algorithms work including covering algorithms, instance-based learning, and how to use the WEKA, an open source data mining software that is a machine learning workbench.

Segall (2006) presented a chapter in the previous edition of this Encyclopedia that discussed microarray databases for biotechnology that included a extensive background on microarray databases such as that defined by Schena (2003), who described a microarray as "an ordered array of microscopic elements in a planar substrate that allows the specific binding of genes or gene products." The reader is referred to Segall (2006) for a more complete discussion on microarray databases including a figure on the overview of the microarray construction process.

Piatetsky-Shapiro (2003) discussed the challenges of data mining specific to microarrays, while Grossman et al. (1998) reported about three NSF (National Science Foundation) workshops on mining large massive and distributed data, and Kargupta at al. (2005) discussed the generalities of the opportunities and challenges of data mining.

Segall and Zhang (2004, 2005) presented funded proposals for the premises of proposed research on applications of modern heuristics and data mining techniques in knowledge discovery whose results are presented as in Segall and Zhang (2006a, 2006b) in addition to this chapter.

Software Background

There is a wealth of software today for data mining such as presented in American Association for Artificial Intelligence (AAAI) (2002) and Ducatelle (2006) for teaching data mining, Nisbet (2006) for CRM (Customer Relationship Management) and software review of Deshmukah (1997). StatSoft (2006a) presents screen shots of several softwares that are used for exploratory data analysis (EDA) and various data mining techniques. Proxeon Bioinformatics (2006) manufactures bioinformatics software for proteomics the study of protein and sequence information.

Lazarevic et al. (2006) discussed a software system for spatial data analysis and modeling. Leung (2004) compares microarray data mining software.

National Center for Biotechnology Information (NCBI) (2006) provides tools for data mining including those specifically for each of the following categories of nucleotide sequence analysis, protein sequence analysis and proteomics, genome analysis, and gene expression. Lawrence Livermore National Laboratory (LLNL) (2005) describes their Center for Applied Scientific Computing (CASC) that is developing computational tools and techniques to help automate the exploration and analysis of large scientific data sets.

MAIN THRUST

Algorithms of Four-Selected Software

This chapter specifically discusses four-selected data mining software that were chosen because these software vendors have generously offered their services and software to the authors at academic rates or less for use in both the classroom and in support of the two faculty summer research grants awarded as Segall and Zhang (2004, 2005).

SAS Enterprise Miner™ is a product of SAS Institute Inc. of Cary, NC and is based on the SEMMA approach that is the process of Sampling (S), Exploring (E), Modifying (M), Modeling (M), and Assessing (A) large amounts of data. SAS