Chapter XXI
Using Data Mining for Forecasting Data Management Needs

Qingyu Zhang
Arkansas State University, USA

Richard S. Segall
Arkansas State University, USA

ABSTRACT

This chapter illustrates the use of data mining as a computational intelligence methodology for forecasting data management needs. Specifically, this chapter discusses the use of data mining with multidimensional databases for determining data management needs for the selected biotechnology data of forest cover data (63,377 rows and 54 attributes) and human lung cancer data set (12,600 rows of transcript sequences and 156 columns of gene types). The data mining is performed using four selected software of SAS® Enterprise Miner™, Megaputer PolyAnalyst® 5.0, NeuralWare Predict®, and BioDiscovery GeneSight®. The analysis and results will be used to enhance the intelligence capabilities of biotechnology research by improving data visualization and forecasting for organizations. The tools and techniques discussed here can be representative of those applicable in a typical manufacturing and production environment. Screen shots of each of the four selected software are presented, as are conclusions and future directions.

INTRODUCTION

Mining biological, medical, or behavioral data is an emerging area for research on bioinformatics (Cohen & Hersh, 2005; Wang & Yang, 2005). This chapter illustrates the use of data mining as a computational intelligence methodology for forecasting data management needs. Specifically, this chapter discusses the use of data mining with multidimensional databases for determining data

management needs for the selected biotechnology data of forest cover data and human lung cancer data sets. The analysis and results will be used to enhance the intelligence capabilities of biotechnology research by improving data visualization and forecasting for organizations. The tools and techniques discussed here can be representative of those applicable in a typical manufacturing and production environment. The chapter also helps organizations to choose proper data mining software for their forecasting data management needs.

The data mining is performed using four selected software of SAS® Enterprise Miner™, Megaputer PolyAnalyst® 5.0, NeuralWare Predict®, and BioDiscovery GeneSight®. One of the databases is that of forest cover type's data that is a very large database composed of 63,377 rows and 54 attributes. The other database is that composed of human lung carcinomas cancer data and is a smaller database with data elements at the human gene level that comprise a microarray database consisting of 12,600 rows of transcript sequences and 156 columns of gene types. Background on related literature and software are also presented. Screen shots of each of the four selected software are presented, as are conclusions and future directions.

BACKGROUND

The study of forecasting started in the 1960s with two categories of linear (e.g., regression) and nonlinear forecasting techniques (e.g., artificial neural network and self-organizing map). Most data mining techniques combine both linear and nonlinear models (He & Xu, 2005) and use different data analysis tools to discover relationships and knowledge in data that may be used to make valid classification and predictions (Chen, Diao, Dulong, et al., 2005; Nielson, 2005).

Data mining has been used in many fields for effective discovery and prediction of new knowledge. Neaga and Harding (2005) present an enterprise integration and management framework based on data mining. Alverez-Macias and Mata-Vazquez (2004) use data mining for the management of software development process. Wu, Chen, and Chian (2006) implement data mining techniques in a product quality control system. Padmanabhan, Zheng, and Kimbrough (2006) use data mining techniques for analyzing user data tracked online for e-business. Rubinov, Soukhorokova, and Ugon (2006) discuss the clusters and classification in data analysis. To improve the prediction accuracy, Li and Ye (2006) present a supervised clustering and classification algorithm for mining data with both numeric and nominal variables in medical diagnosis.

As discussed by Segall (2006) in a chapter in *Encyclopedia of Data Warehousing and Mining*:

Microarray informatics is a rapidly expanding discipline in which large amounts of multi-dimensional data are compressed into small storage units. Data mining of microarrays can be performed using techniques such as drill-down analysis rather than classical data analysis on a record-by-record basis. Both data and metadata can be captured in microarray experiments.

An important issue of this chapter is what benefits your organization can drive from a properly implemented storage management policy and specifically for databases of varying dimensionalities such as at the microarray level.

Segall (2006) further discusses the following background on microarray databases by Schena (2003) and National Center for Biotechnology Information (NCBI):

A Microarray has been defined by Schena (2003) as 'an ordered array of microscopic elements in a planar substrate that allows the specific binding of genes or gene products.' Schena (2003) claims microarray databases as "a widely recognized