

Data Mining and Complex Problems: Case Study in Composite Materials

Luis Rabelo and Mario Marin
University of Central Florida

Lisa Huddleston
NASA, Kennedy Space Center

Copyright © 2009 SAE International

ABSTRACT

Data mining is defined as the discovery of useful, possibly unexpected, patterns and relationships in data using statistical and non-statistical techniques in order to develop schemes for decision and policy making. Data mining can be used to discover the sources and causes of problems in complex systems. In addition, data mining can support simulation strategies by finding the different constants and parameters to be used in the development of simulation models. This paper introduces a framework for data mining and its application to complex problems. To further explain some of the concepts outlined in this paper, the potential application to the NASA Shuttle Reinforced Carbon-Carbon structures and genetic programming is used as an illustration.

INTRODUCTION

Data mining is a very useful technique that has become dominant and well developed in the last twenty years. Data mining has been defined as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data." [3] Another interesting definition is the "Discovery of useful, possibly unexpected, patterns in data using statistical and non-statistical techniques." [20] This definition makes a strong emphasis on the "non-statistical" techniques. Our experience working with complex problems, incomplete data, high noise levels have provided us with a more comprehensive methodology where data mining base-models can be used with other type of empirical models. Our methodology has the following steps:

1. Understand Complex Problem from different viewpoints: We have to understand the problem and the goals and objectives assigned to the data mining task. We have to view complex problems from different dimensions. It is usual a multi-disciplinary effort. Some of the important viewpoints are:
 - a. Basic Theory – First principles are very important to understand. The data mining team must be networked with the scientists and experts from the different dimensions. The data mining team has to be conversant with the different contributions of the different disciplines involved (materials, optics, etc.).
 - b. Experiments and Visits – The different experiments must be understood by the data mining team. How they relate to each other? How was the equipment calibrated? Who are the owners of the data?
 - c. Organizational/Cultural/Political – The program environment must be investigated. Do the participants understand the goals/objectives and procedures of the data mining task? Is there an institutional culture of sharing ideas, information, and data? Is top management championing the data mining team?
2. Gather Information from current databases/files: This step is very important. Complex problems in large/global organizations have distributed databases and other type of repositories of data and information with different formats, different

computing/IT platforms, and levels of details and accuracy.

3. Develop map of databases from the different points in the life-cycle: It is important to have a clear picture and guidance of the different variables, experiments and data available. A map of this is very important with the flexibility to integrate different databases and create new ones. An ontology will be very important (if budget and sophistication of the project permit) to increase agility and interoperability.
4. Develop map of "models" (analytical and Empirical) from the different points in the life-cycle: Usually, this step is totally forgotten from the data mining task (it was difficult to find an article on data mining with this philosophy). The traditional data miners go directly to the database to start playing with the data and the variables. Not only are the results from experiments very important for the data mining task but so are previously developed models based on statistics, non-statistical techniques, finite element analysis, simulations, and first principle models. These models have important information. We must be able to explore their fusion with the predictive models to be developed by the data mining task.
5. Build databases from current ones: Now that we know the goals/objectives, of the different environments, we can create comprehensive databases with the relevant data and variables. The different procedures can be used to start preparing the data for the mining.
6. Knowledge Discovery and Predictive Modeling: Develop the different models, discovery of relationships, according to the goals/objectives of the data mining task. It is important to explore the information fusion of the different models developed.
7. Deployment of the models developed. This not only includes the development of a user interface but also includes the interpretation of the models' answers in the corresponding technical language. An integrity management plan must be implemented with the appropriate documentation.

DATA MINING

Data mining has several variations and many techniques. Two processes of data mining are knowledge discovery and predictive modeling.

KNOWLEDGE DISCOVERY AND PREDICTIVE MODELING - Some problems just require knowledge discovery. On the other hand, if the objective is to build a computationally efficient and accurate predictive model it usually requires both views.

Knowledge discovery provides information about the relationships of the data. Procedures for the preparation, formatting, transformation, and the design of experiments are important steps in this process.

Predictive modeling is a method used to create a model of behavior. These models can be statistical (e.g., multiple-linear regression) or non-statistical (e.g., genetic programming). As explained by [8] "A predictive model is made up of a number of predictors, which are variable factors that are likely to influence future behavior or results." The preliminary steps can also involve procedures for the preparation, formatting, and the design of experiments.

Complex problems usually require both processes. Knowledge discovery is important to find relationships and causalities which can support the development of predictive models. These "discovered" relationships can represent "new" variables to be used in the predictive models. Knowledge discovery is used first and then the predictive model is developed based on the information generated by knowledge discovery. However, an existing predictive model can also be mined using knowledge discovery in order to provide "new" relationships and augment the knowledge base. For example, a predictive model based on neural networks can be built for a particular complex problem. Then, the neural network model (that is usually used for predictions) can be analyzed to provide some new relationships of the complex process that were unknown.

GENETIC PROGRAMMING - Genetic programming is an automated methodology inspired by Darwinian evolution. Genetic programming assigns the computer the ability to develop computer programs/models (represented as tree structures). Genetic programming is based on the Darwinian ideas of "survival of the fittest" and the natural genetic operators of reproduction, mating, and mutation [5, 9, 10, 11]. As stated by Ratner [15], "The process begins with a fitness function and a set of user-selectable mathematical and logical functions" from which a predictive model can be formed. "A first generation of as many as 250 - 1000 models is randomly generated using the functions and variables available; the "fitness" of each model is evaluated using collected data." [15]

The next generations of models are created following the processes of mating, reproduction, and mutation. When two models "mate", the offspring are combinations of the parents' genetic material. Thus, each parent probabilistically contributes "good" genetic material to the child. After a number of generations, the process provides a predictive model adapted to the objective.

Genetic programming is considered a non-statistical methodology for data mining. Its major use is predictive modeling [16]. However, genetic programming can also be used for knowledge discovery (Koza et al. [10, 11]).