IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Hershey PA 17033-1117, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com
ITB8132

Chapter III

Steps to Success for the Mining Process

The previous chapters have given you some background on the core components of corporate IT systems along with software technology that promotes "business intelligence" throughout an enterprise. This included a good foundation on the high end analytical portion of information systems, namely data mining technology. All this sounds fantastic, state-of-the-art software that helps increase the flow of value-added information which leads to a reduction of uncertainty in a given business environment. However, the bottom line to the productivity enhancing process from IT implementation really entails proper management and utilization of this technology. In other words, an organization can spend huge sums of dollars on the best systems available, but if they are not implemented properly, their value and dollars invested become useless.

Data mining technology is no exception. In fact, because of the more complex nature of the technology (e.g., statistics and mathematic underpinnings), the potential for underutilization or improper utilization is probably greater than other types of analytical applications. The following chapter provides some helpful hints on how to manage the mining process as it illustrates some common pitfalls that exist in conducting a high-end analysis. Remember, today's technology is good, but it doesn't do all the work for you.

MINING THE RIGHT DATA (Garbage In, Garbage Out)

If this book teaches you anything about data mining, let it be this: If you gather your data poorly, you are destined to failure. No data mining technique or tweak can rescue you. If, on the other hand, you gather your data well,

success will come so easily that you will wonder why everybody thinks data mining is so complicated.

The data you gather must:

- contain information,
- be relevant,
- be in a format which data mining can use effectively.

How Much Is My Existing Data Worth?

It is not uncommon to look to data mining to see if there is an opportunity to squeeze interesting nuggets of information out of an existing database. This is sometimes possible, sometimes not. If you are interested in visualization or slice and dice, your chances are good. If the database is in any way poorly designed for the task you wish to perform, you must accept that it will be easier to start gathering the right data from scratch. Avoid the temptation to mine a database simply to justify gathering the data in the first place: you will be throwing good money after bad.

A database is a heap of numbers, but that does not mean it necessarily contains useful information. Many of the numbers could be zero, for example, which would not be much use.

The first test for a database is to see whether it contains much information of any kind. Information theory studies this topic in depth using arguments which center around how predictable the data is. (See for example Price, 1980.) The conclusion of the theory is this: your database contains no more information than the smallest file you can compress it to. So, using a file compression routine such as PKZIP, try compressing the database. If it compresses to 100KB, it has ten times as much information as a file that compresses to 10KB, and so on.

The file compression test is a simple way of determining whether your database contains a lot of information. If it does not, you'll have to start gathering the data again.

In order to determine whether the information in the database is relevant, you will first need to decide what you want to do with the data. This book outlines some potential applications of data mining. As you go through the book, first ask whether the technique is applicable to your business. Then ask whether your database contains the data you will require. If it does not, you will have to start from scratch.

The final issue to consider when gathering data is whether the data has been gathered in a format which a data mining algorithm can use. Since the database was not designed with data mining in mind, it is unlikely to be in the correct form and some re-jigging will be necessary. For example, in order to