Intelligent Data Analysis

Xiaohui Liu
Brunel University, UK

INTRODUCTION

Intelligent Data Analysis (IDA) is an interdisciplinary study concerned with the effective analysis of data. IDA draws the techniques from diverse fields, including artificial intelligence, databases, high-performance computing, pattern recognition, and statistics. These fields often complement each other (e.g., many statistical methods, particularly those for large data sets, rely on computation, but brute computing power is no substitute for statistical knowledge) (Berthold & Hand 2003; Liu, 1999).

BACKGROUND

The job of a data analyst typically involves problem formulation, advice on data collection (though it is not uncommon for the analyst to be asked to analyze data that have already been collected), effective data analysis, and interpretation and report of the finding. Data analysis is about the extraction of useful information from data and is often performed by an iterative process in which exploratory analysis and confirmatory analysis are the two principal components.

Exploratory data analysis, or data exploration, resembles the job of a detective; that is, understanding evidence collected, looking for clues, applying relevant background knowledge, and pursuing and checking the possibilities that clues suggest.

Data exploration is not only useful for data understanding but also helpful in generating possibly interesting hypotheses for a later study—normally a more formal or confirmatory procedure for analyzing data. Such procedures often assume a potential model structure for the data and may involve estimating the model parameters and testing hypotheses about the model.

Over the last 15 years, we have witnessed two phenomena that have affected the work of modern data analysts more than any others. First, the size and variety of machine-readable data sets have increased dramatically, and the problem of data explosion has become apparent. Second, recent developments in computing have provided the basic infrastructure for fast data access as well as many advanced computational methods for extracting information from large quantities of data. These developments have created a new range of problems and challenges for data analysts as well as new opportunities for intelligent systems in data analysis, and have led to the emergence of the field of Intelligent Data Analysis (IDA), which draws the techniques from diverse fields, including artificial intelligence (AI), databases, high-performance computing, pattern recognition, and statistics. What distinguishes IDA is that it brings together often complementary methods from these diverse disciplines to solve challenging problems with which any individual discipline would find difficult to cope, and to explore the most appropriate strategies and practices for complex data analysis.

MAIN THRUST

In this paper, we will explore the main disciplines and associated techniques as well as applications to help clarify the meaning of intelligent data analysis, followed by a discussion of several key issues.

Statistics and Computing: Key Disciplines

IDA has its origins in many disciplines, principally statistics and computing. For many years, statisticians have studied the science of data analysis and have laid many of the important foundations. Many of the analysis methods and principles were established long before computers were born. Given that statistics are often regarded as a branch of mathematics, there has been an emphasis on mathematics rigor, a desire to establish that something is sensible on theoretical ground before trying it out on practical problems (Berthold & Hand, 2003). On the other hand, the computing community, particularly in machine learning (Mitchell, 1997) and data mining (Wang, 2003) is much more willing to try something out (e.g., designing new algorithms) to see how they perform on real-world datasets, without worrying too much about the theory behind it.

Statistics is probably the oldest ancestor of IDA, but what kind of contributions has computing made to the subject? These may be classified into three categories. First, the basic computing infrastructure has been put in place during the last decade or so, which enables large-scale data analysis (e.g., advances in data warehousing

and online analytic processing, computer networks, desktop technologies have made it possible to easily organize and move the data around for the analysis purpose). The modern computing processing power also has made it possible to efficiently implement some of the very computationally-intensive analysis methods such as statistical resampling, visualizations, large-scale simulation and neural networks, and stochastic search and optimization methods.

Second, there has been much work on extending traditional statistical and operational research methods to handle challenging problems arising from modern data sets. For example, in Bayesian networks (Ramoni et al., 2002), where the work is based on Bayesian statistics, one tries to make the ideas work on large-scale practical problems by making appropriate assumptions and developing computationally efficient algorithms; in support vector machines (Cristianini & Shawe-Taylor, 2000), where one tries to see how the statistical learning theory (Vapnik, 1998) could be utilized to handle very high-dimensional datasets in linear feature spaces; and in evolutionary computation (Eiben & Michalewicz, 1999) one tries to extend the traditional operational research search and optimization methods.

Third, new kinds of IDA algorithms have been proposed to respond to new challenges. Here are several examples of the novel methods with distinctive computing characteristics: powerful three-dimensional virtual reality visualization systems that allow gigabytes of data to be visualized interactively by teams of scientists in different parts of the world (Cruz-Neira, 2003); parallel and distributed algorithms for different data analysis tasks (Zaki & Pan, 2002); so-called any-time analysis algorithms that are designed for real-time tasks, where the system, if stopped any time from its starting point, would be able to give some satisfactory (not optimal) solution (of course, the more time it has, the better solution would be); inductive logic programming extends the deductive power of classic logic programming methods to induce structures from data (Mooney, 2004) ; Association rule learning algorithms were motivated by the need in retail industry where customers tend to buy related items (Nijssen & Kok, 2001), while work in inductive databases attempt to supply users with queries involving inductive capabilities (De Raedt, 2002). Of course, this list is not meant to be exhaustive, but it gives some ideas about the kind of IDA work going on within the computing community.

IDA Applications

Data analysis is performed for a variety of reasons by scientists, engineers, business communities, medical and government researchers, and so forth. The increasing size and variety of data as well as new exciting applications such as bioinformatics and e-science have called for new ways of analyzing the data. Therefore, it is a very difficult task to have a sensible summary of the type of IDA applications that are possible. The following is a partial list.

- **Bioinformatics:** A huge amount of data has been generated by genome-sequencing projects and other experimental efforts to determine the structures and functions of biological molecules and to understand the evolution of life (Orengo et al., 2003). One of the most significant developments in bioinformatics is the use of high-throughput devices such as DNA microarray technology to study the activities of thousands of genes in a single experiment and to provide a global view of the underlying biological process by revealing, for example, which genes are responsible for a disease process, how they interact and are regulated, and which genes are being co-expressed and participate in common biological pathways. Major IDA challenges in this area include the analysis of very high dimensional but small sample microarray data, the integration of a variety of data for constructing biological networks and pathways, and the handling of very noisy microarray image data.
- **Medicine and Healthcare:** With the increasing development of electronic patient records and medical information systems, a large amount of clinical data is available online. Regularities, trends, and surprising events extracted from these data by IDA methods are important in assisting clinicians to make informed decisions, thereby improving health services (Bellazzi et al., 2001). Examples of such applications include the development of novel methods to analyze time-stamped data in order to assess the progression of disease, autonomous agents for monitoring and diagnosing intensive care patients, and intelligent systems for screening early signs of glaucoma. It is worth noting that research in bioinformatics can have significant impact on the understanding of disease and consequently better therapeutics and treatments. For example, it has been found using DNA microarray technology that the current taxonomy of cancer in certain cases appears to group together molecularly distinct diseases with distinct clinical phenotypes, suggesting the discovery of subgroups of cancer (Alizadeh et al., 2000).
- **Science and Engineering:** Enormous amounts of data have been generated in science and engineering (Cartwight, 2000) (e.g., in cosmology, chemical engineering, or molecular biology, as discussed previously). In cosmology, advanced computational tools are needed to help astronomers understand the origin of large-scale cosmological structures