Soft Computing for Intelligent Data Analysis

X Liu, R Johnson, G Cheng, S Swift, A Tucker
Birkbeck College, Department of Computer Science
University of London, London WC1E 7HX, United Kingdom

Abstract

Intelligent data analysis (IDA) is an interdisciplinary study concerned with the effective analysis of data. This paper will briefly look at some of the key issues in intelligent data analysis, discuss the opportunities for soft computing in this context, and present several IDA case studies in which soft computing has played key roles. These studies are all concerned with complex real-world problem solving, including consistency checking between mass spectral data with proposed chemical structures, screening for glaucoma and other eye diseases, forecasting of visual field deterioration, and diagnosis in an oil refinery involving multivariate time series. Bayesian networks, evolutionary computation, neural networks, and machine learning in general are some of those soft computing techniques effectively used in these studies.

1. Introduction

Intelligent data analysis (IDA) is an interdisciplinary study concerned with the effective analysis of data. IDA requires careful thinking at every stage of an analysis process, intelligent application of relevant domain expertise regarding both data and subject matters, and critical assessment and selection of relevant analysis methods. Although statistics has been the traditional method for data analysis, the challenge of extracting useful information from large quantities of online data has called for advanced computational analysis methods. Soft computing provides a wealthy set of methodologies that can contribute substantially to the effective analysis of large data sets.

This paper will briefly look at some of the key issues in intelligent data analysis, discuss the opportunities for soft computing in this context, and present several IDA case studies in which soft computing has played key roles. These studies, conducted at Birkbeck, are all concerned with complex real-world problem solving, including consistency checking between mass spectral data with proposed chemical structures, screening for glaucoma and other eye diseases, forecasting of visual field deterioration, and diagnosis in an oil refinery involving multi-variate time series. Bayesian networks, evolutionary computation, neural networks, and machine learning in general are some of those soft computing techniques effectively used in these studies.

2. Key IDA Issues

For the last decade or so, the size of machine-readable data sets has increased dramatically and the problem of "data explosion" has become apparent. On the other hand, recent developments in computing have provided the basic infrastructure for fast data access; processing power and storage devices are continuing to become cheaper and more powerful, networks are providing more bandwidth and higher reliability, personal computers and workstations are widespread, and On-Line Analytic Processing (OLAP) allows rapid retrieval of data from data warehouses [2]. In addition, many of the advanced computational methods for extracting information from large quantities of data, or "data mining" methods, are beginning to mature, e.g. artificial neural networks, Bayesian networks, decision trees, genetic algorithms, and statistical pattern recognition [12]. These developments have created a new range of problems and challenges for the analysts, as well as new opportunities for intelligent systems in data analysis [6, 7].

The following issues have been found particularly important in the quest for intelligent data analysis: **strategies, data quality** and **scalability**. Firstly, data analysis in a problem-solving context is typically an *iterative* process involving problem formulation, model building, and interpretation of the results. The question of how data analysis may be carried out effectively should lead us to looking closely not only at those individual components in the data analysis process, but also at the process as a whole, asking what would con-

0-7803-5211-4/99/$10.00 © 1999 IEEE

stitute a sensible data analysis **strategy**. This strategy should describe the steps, decisions and actions which are taken during the process of analyzing data to build a model or answer a question, and one might define a good strategy as being the hallmark of intelligent data analysis [6].

Secondly, data are now viewed as a key organizational resource and the use of high-quality data for decision making has received increasing attention [13]. It is commonly accepted that one of the most difficult and costly tasks in large-scale data analysis is trying to obtain clean and reliable data, and many have estimated that as much as 50% to 70% of a project's effort is typically spent on this part of the process. In response to this opportunity, "data cleaning" companies are being created, and "data quality groups" are being set up in corporations. Since the use of the "wrong" kind of data or very "low-quality" data often leads to useless analysis results, research on **data quality** has attracted a significant amount of attention from different communities including information systems, management, computing, and statistics.

Thirdly, one of the key issues involved in large-scale data analysis is **scalability**, e.g. if a method works well for a task involving a dozen variables and 10,000 cases, is it still going to perform well for one with 100 variables and one million cases? There has been a lot of intensive research on the development of efficient, approximate or heuristic algorithms that are able to scale well [3]. Among many related methods are sampling, feature selection, restriction of model search space, and application of domain knowledge. However, if the number of variables involved is large because data is highly-structured, or there is simply a huge amount, say terabytes, of data to deal with in an application, then some basic hardware support is essential. In this connection, it is pleasing to see that the High Performance Computing Community have been taking initial steps in this direction, asking how a desktop user might be able to utilize the computer power offered by a set of supercomputers and massive data stores interconnected by a high-speed, high bandwidth communication network, and to use resources from such a network as though it were from a single machine.

3. Soft Computing for IDA

The wealthy set of methodologies in soft computing can contribute substantially to the effective analysis of a variety of data. This section describes our experience in applying various techniques to different analysis tasks. It will become clear that certain techniques are particularly suited to some specific tasks, and a combination of several techniques is sometimes required to tackle complex problems. And often the analysis of data "intelligently" lies in the appropriate application of relevant domain expertise regarding both data and subject matters, and in the critical assessment and selection of relevant analysis methods. Below four case studies will be outlined.

3.1 Analysing mass spectral data

Correlating mass spectral data with proposed chemical structures is the major task of a large number of mass spectrometry groups around the world. This correlation task requires considerable human expertise and competent spectral data analysis.

A simple way of viewing this correlation problem would be a mapping from a set of input variables representing the mass spectral data and structural information to a binary output variable representing "consistent" or "inconsistent/don't know". The problem of achieving such mappings is that there are potentially a large number of input variables but only one output variable. When submitted, the sample is automatically subjected to the analysis by a mass spectrometer using a robot system. The result typically consists of a data set of 120 mass spectra (scans) for the sample. Each spectrum will contain up to 1200 peaks in terms of its mass relative to the proposed molecular mass. So the mass spectral data alone could have 120 X 1200 data points and this is in addition to the corresponding information regarding the proposed structure. An expert mass spectrometrist typically hypothesizes how the proposed structure may be fragmented in the mass spectrometer and uses a set of resultant predicted fragments (peaks) to correlate with the actual peaks from the original data set.

Therefore the challenging task here is to significantly reduce the number of possible input variables to make the above-mentioned mapping possible. Two sets of data are essential to enable the automation of the correlation task: the mass spectral data collected by mass spectrometers and the predicted fragments of sample by the mass spectrometrist. The former may be directly available after a few experiments, but the latter is not readily available; it needs to be generated from the proposed structure using relevant chemical knowledge and expertise acquired from the experts. To transform the sample and proposed structure into a report of whether the chemist has made the right kind of product, the system needs to perform the following tasks.

Data pre-processing: Apart from choosing a suitable spectrum range and ensuring the quality of mass spectral data, two software modules are developed to select most significant data or features. The first selects "significant" data (peaks) from a mass spectral data set and the heuristic employed in deciding whether a particular peak is significant considers the mass and intensity of the peak relative to the proposed molecular mass. The list of significant peaks obtained this way will be guaranteed to contain peaks from a single structure only in cases where the sample is pure. Frequently this is not the case and hence the list of significant peaks needs to be partitioned such that those peaks which are derived from the proposed structure can be identified for structure-spectrum correlation. This is the job of the second software module, which uses statistical techniques to analyse a data set from a mixture, determine the number of components present, and resolve pure spectra for these components.

Inferring the predicted fragments of the molecule: We have developed a knowledge-based system for predicting how the proposed structure may fragment in a mass spectrometer using expert knowledge of mass spectral events such as bond breakages and rearrangements. A hypothesize-and-test method has been used. In particular, a set of rules have been developed which detect the presence of characteristic sites in the molecule at which breakage may occur and these rules are applied to the proposed structure to obtain a number of hypotheses regarding the likely breaking sequences of the molecule.

Performing the main confirmation operation: We have developed a neural network to perform this task. A number of features have been extracted by combining the significant data generated and the predicted peaks and by applying domain-specific knowledge. These limited number of features are then used as input variables for the neural network and the results obtained are very encouraging [4].

This is an application where careful thinking is required at each stage of the analysis process, and where a variety of techniques are utilised and integrated into a single problem-solving **strategy**, including knowledge-based systems, neural networks and statistical methods.

3.2 Analysing screening test data

In the last few years, we have developed a software-based visual field screening system that integrates a visual stimuli generating program with a number of machine learning components [9]. This system was developed in response to the practical need to screen subjects in various public environments where the specialised instruments for examining the visual field cannot be made available. In particular, the system was designed to detect glaucoma and optic neuritis effectively.

To evaluate the performance of this screening system, various characteristics affecting software quality were systematically studied. In particular, three of the most important characteristics for screening applications: functionality, reliability, and efficiency, were carefully evaluated. *Functionality* in a screening system is used to refer to the system's capability of detecting as many as possible of those subjects in the community who suffer from an eye disease at an early stage, and at the same time, to minimise the number of "false positives" - those who failed the test, but have no eye disease. *Reliability* in the context of screening applications is how reliably the data collected by the system reflect subject's visual functions or damages. *Efficiency* in the same context is concerned about how to minimise the amount of time a subject has to spend on a single test visit, while maintaining the quality of the test results.

Having decided what to evaluate, clinical data from laboratory-based and field-based investigations in different communities were then carefully collected. Here are a very brief description of two such field studies. The first is the World Health Organisation programme for preventing optic neuritis in the Kaduna State, Nigeria [1]. The subjects were from a farming community in Kaduna, who were largely computer-illiterate. The visual field tests were carried out in village huts on consenting subjects aged 15 years and over in several rural communities that were endemic for optic neuritis in the guinea savannah of Kaduna State, Northern Nigeria. In all, 3182 test records from 2388 different eyes were collected using six notebook computers operated by ophthalmic nurses. The other is a pilot study to detect people with glaucoma, a common disease with the elderly in the UK, sponsored by UK's Medical Research Council [9]. The test was offered during routine attendance at a large urban general practice in North London and was conducted in a corner of the main waiting room separated by a cotton screen. More than 900 people were screened and over 2000 test records were collected during the screening period.

Since the main *functionality* for the screening system is its "discriminating power", we aim to establish the system's capability in maximising the chance of detecting those in the community who suffer from an eye disease at an early stage, while minimising the number of "false positives". Consequently we have used the *Receiver Operator Characteristic* (ROC) analysis and associated methods for this purpose. To assess the *reliability* of the system, we have proposed the follwing two criteria: 1) the consistency between the repeated test results from the same subject; 2) the agreement between disease patterns discovered from our test and those from other established screening instruments. The *efficiency* of the system was assessed by the question of "what is the minimum number of repeated measurements during a test to maintain the quality of test results".

One of the most challenging issues in this application is aspects of **data quality**, particularly the management of outlying data. *Outliers* in screening data are often measurement errors but some of them can be caused by pathological conditions of the subject. To distinguish between these two requires a careful application of relevant domain knowledge. Two general strategies have been suggested for knowledge-based outlier analysis, and soft computing techniques such as neural networks and rule inductions have been found particularly useful in this context [8, 14].

3.3 Learning multivariate forecasting models

The common trait of the conditions for a glaucoma sufferer is a functional abnormality in the optic nerve, leading to loss of visual field. The prediction of visual field deterioration in patients who are suffering from glaucoma plays an important role in the management, treatment and control of the disease's progress. For example, if the deterioration is slowing down, it might be appropriate to reduce the medication; or if the deterioration is speeding up, an increase in medication might be needed or surgery might be necessary.

Visual field tests are normally performed on a clinical machine, which collects test data from patients. The particular test machine we are concerned with examines 76 points in each eye. The corresponding data set being considered contains information on patients' eyes tested approximately every six months for between five and 22 years. Therefore, the length of time series corresponding to some of patients' visual field tests can be rather short. Data pre-processing has demonstrated a strong interdependency among these 76 variables, especially for those lying on the same nerve fibre bundle.

The Vector Auto-Regressive (VAR) [10] process appears to be an appropriate way of modelling the multivariate time series data from the patient's visual fields. For the VAR process to be of use, the order must be identified and the associated parameters must be estimated, for example using the standard method of solving the appropriate set of Yule-Walker equations. However, this technique places constraints on the minimum number of time series observations in the dataset.

We have used a genetic algorithm to overcome these problems by learning both the order and corresponding parameters. Preliminary results have shown that this approach provides a better way for fitting a VAR process than the conventional statistical methods. This approach has been found to be competent in modelling short-length multivariate time-series data such as the visual field data. This gives the approach a wider range of applications than the standard statistical methods (e.g. than with the Yule-Walker equations).

3.4 Learning causal structures from time-series

Many complex chemical processes record multivariate time-series data every minute. This data is characterised by a large number of interdependent variables, though some may have no substantial impact on any others. There can be large time lags between causes and effects (over 120 minutes in some chemical processes). Take the oil refinery process as an example. In reviewing oil refinery data, process engineers often come across trends with unexpected characteristics. In many cases these anomalous events have a significant adverse economic impact, whether in terms of reduced yield, excessive equipment stress, or violation of environmental constraints. The identification of such events is important but of greater importance still are adequate casual explanations of them, which could then be used to modify operating practices, retrain operators or conduct anticipatory planning.

In order to perform causal explanation, some method is required for reasoning about relationships between these variables back in time. For example, the reason for a particular temperature becoming extremely high may be that a flow rate dropped ten minutes ago and the flowed dropped because, one minute before that, a valve was closed by a control engineer. A number of approaches to causal explanation have been proposed by different AI communities over the years. Early proposals include the use of *rule-based* and *model-based* systems. A more recent paradigm for performing causal inference is the Bayesian Network [11] and its dynamic counterpart can model a system over time [5]. Most of the research on dynamic networks, however, has focused on small models or models with small time lags. It would be desirable to learn a Dynamic Bayesian Network (DBN) for large datasets with large possible time lags such as the refinery data.

The search for the casual structure from a large multivariate time series with large time lags is a daunting task, particularly if it must be found quickly. In some applications such as diagnosis in an oil refinery, the casual explanation may be required in a very short space of time. We have suggested a general methodology which attempts to overcome some of the key problems associated with learning such a model. In particular, a representation that produces a dynamic Bayesian network is proposed, based on a reasonable assumption. An algorithm for finding a good structure in as short a time as possible using evolutionary computation is then developed.

The compromise we are trying to make in finding a good DBN quickly is between quality and efficiency. That is, we want as good a DBN as possible for diagnosis (quality) but we want it in as short a time as possible (efficiency). A number of experiments using both oil refinery data and synthetic data sets have been performed and it has been found that the proposed algorithm does show a good compromise between time demand and quality in the automatic generation of explanations. Further research is being conducted, especially on the **scalability** of such algorithms.

4 Concluding remarks

The evolution of computing technology and the ever-increasing size and variety of data sets have created a new range of problems and challenges for data analysts, as well as new opportunities for intelligent systems in data analysis. In this paper we have discussed some of the key IDA issues and introduced several case studies where it is important that these issues be properly addressed. For example, careful thinking of the possible analysis strategies is particularly important for the consistency-checking application, data quality control is especially challenging in the analysis of screening data, and scalability is one of the key issues in the learning of casual structures from multivariate time series data. Soft computing techniques such as Bayesian networks, genetic algorithms, neural networks and rule induction have been found very useful in addressing these IDA issues.

5 Acknowledgements

We thank other members of the IDA Group at Birkbeck and our industrial partners for their contribution.

References

[1] G. Cheng, X. Liu, J. Wu, and B. Jones. Establishing a reliable visual function test and applying it to screening optic nerve disease in onchocercal communities. *International Journal of Bio-Medical Computing*, 41:47–53, 1996.

[2] E. F. Codd. *Providing OLAP (On-Line Analytic Processing) to User-Analysts: An IT Mandate.* E F Codd and Associates, 1994.

[3] W. W. Cohen. Fast effective rule induction. *Proc. of the Twelfth International Conference on Machine Learning*, pages 115–123, 1995.

[4] H. Dettmar, X. Liu, R. Johnson, and A. Payne. Knowledge-based data generation. *Knowledge-Based Systems*, 11:167–177, 1998.

[5] N. Friedman, K. Murphy, and S. Russell. Learning the structure of dynamic probabilistic networks. In *Proc. of the 14th Conference on Uncertainty in Artifical Intelligence*, pages 139–147, 1998.

[6] D. J. Hand. Intelligent data analysis: Issues and opportunities. In X. Liu, P. Cohen, and M. Berthold, editors, *Advances in Intelligent Data Analysis: Reasoning about Data, LNCS 1280*, pages 1–14. Springer-Verlag, 1997.

[7] X. Liu. Intelligent data analysis: issues and challenges. *The Knowledge Engineering Review*, 11:365–371, 1996.

[8] X. Liu, G. Cheng, and J. Wu. Noise and uncertainty management in intelligent data modeling. *Proc. of the 12th National Conference on Artificial Intelligence (AAAI-94)*, pages 263–268, 1994.

[9] X. Liu, G. Cheng, and J. Wu. Ai for public health: Self-screening for eye diseases. *IEEE Intelligent Systems*, 13:5:28–35, 1998.

[10] H. Lutkepohl. *Introduction to Multivariate Time series Analysis.* Springer-Verlag, 1993.

[11] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference.* Morgan Kaufmann, 1988.

[12] G. Piatetsky-Shapiro and W. J. Frawley. *Knowledge Discovery in Databases.* AAAI Press / The MIT Press, 1991.

[13] G. Tayi and T. Ballou. Examining data quality. *Communications of the ACM*, 41:2, 1998.

[14] J. Wu, G. Cheng, and X. Liu. Reasoning about outliers by modelling noisy data. In X. Liu, P. Cohen, and M. Berthold, editors, *Advances in Intelligent Data Analysis (IDA-97) LNCS 1280*, pages 549–558. Springer-Verlag, 1997.