http://doi.org/10.4038/cjms.v45i1.4861

Mining for treasures of information: data mining

W. N. Wickremasinghe[1]

The Ceylon Journal of Medical Science 2002; 45: 39-45

Abstract

Data mining is relatively new in data analysis. It is a process that is used to explore and analyze large quantities of data in order to discover meaningful patterns or rules by automatic or semi-automatic means. Currently information-rich sectors benefit from this and most significant and hence profitable applications are in marketing.

Data mining is introduced in this article. Some important applications are discussed and common techniques described. A small application in agricultural research is illustrated and current status of data mining discussed.

Key words: Data mining, statistical data mining, data analysis, knowledge discovery.

Introduction

As far as statistics is concerned, data analysis is not a new concept. However, last decade saw the beginning of a new era for data analysis, called data mining. While there is no "standard definition" for data mining at present, it is essentially a process that uses a mixture of statistical and modern computing tools to explore and analyze large amounts of data in order to find meaningful patterns or treasures of information. Originally, data mining was taught and discussed as an area in Computer Science. However, no analysis is complete without a formal scientific approach that gives valid conclusions. Therefore, the need for more and more statistical input into data mining techniques was evident. Today, statisticians have come forward to contribute to this new area while typical data mining tools have also got added statistical features in them (1). For a data mining job to be meaningful and profitable, the data should be available in electronic form that could be easily accessible through inexpensive means. In the recent past, much attention has been focused on this area and more articles and text books appeared in that context (2,3,4,5,6,7). However, as far as Sri Lanka is concerned, there is some lack of awareness in this direction. This may be mainly due to the unavailability of resources in this connection at present.

The objecitve of this article is to introduce this new concept and to briefly discuss some of the important applications and techniques of data mining. An illustration of a small scale application of data mining in the field of agricultural research is given and the current status of data mining is also mentioned (8).

The need

A small business builds relationships with its customers by noticing their needs and remembering their preferences, and learning from the past on how to serve them better. On the other hand, in a large establishment such as a supermarket or a grocery store most customers may never interact personally with its employees. Even when there is customer interaction, it is likely to be with a different sales clerk each time. Therefore, there has to be a method to notice, remember, and learn, from these interactions. "Market Basket Analysis" is one simple data mining tool that can be used in such situations, for example, to find out which items sell together, to suggest new store lay-outs, and to determine which products to put on special, etc. Here, instead of interacting with the customer, the company looks at the automated data about customer sales, and analyzes

1. Senior Lecturer, Department of Statistics, University of Colombo.

them automatically to get useful information. Likewise, the need is there in other sectors as well to get valuable information out of vast amounts of automated data. In the next section, some important applications of data mining in different sectors will be discussed.

Some applications

Criminal investigations

When a large number of investigators are employed to investigate a certain criminal act, the independent reports submitted by them can be explored automatically to find some useful links (if any) if the information is automated. The United States government adopted a data mining technique called "Automatic Link Analysis" in early 90's to sift through thousands of reports submitted by investigators in the Oklahoma City bombing case (1).

Banking

Data mining can be used to promote customer retention, when the customer is free to change supplier, for example, in the banking sector and cellular phone companies. An interesting example on how a bank had used data mining to find out a cheap way to hold on to profitable customers can be found in Berry and Linoff (2).

Agricultural research

Christensen and Cook discuss an application of statistical data mining to understand the relationship between 10 soil characteristics and corn yield in a field experiment (3). They have used the common menu of any statistical data analysis, *i.e.* descriptive analysis, formal analysis, and advanced analysis, in that order, but with the help of modern data mining tools such as "XGobi", "Grand Tour", and some SAS tools that help one to produce results by exploring the data automatically.

Cancer research

Relatively new data mining techniques such as Support Vector Machines (SVMs) have been successfully used in cancer classification. In a recent study conducted by Ramaswamy *et al.*, it has been demonstrated that it is feasible to accurately classify cancer patients through multi-class molecular classifications with the help of microarray gene expression analysis based on SVM algorithm (9). The authors have obtained an accuracy of 78% as compared to 9% through random classification and have suggested future clinical implementation of molecular cancer diagnostics when conventional methods are difficult or impossible.

Manufacturing

Data mining techniques have been successfully used in manufacturing areas like production scheduling and planning, quality control, and optimization of resources allocation (4).

Catalogue shopping

The concept of a personal code be used for catalogue shopping is a potentially new application. The idea is to see if one single code that covers all relevant "body sizes" can be assigned to individuals. Usually, each individual knows his or her hat size, waist size, shirt/blouse size, length of pants, shoe size etc., but can we assign a code say, MX01, to represent all males with a certain range of hat sizes, shirt sizes, pant sizes, and shoe sizes, etc? If possible, this will help someone to order many things by just giving one simple code! A similar study had been conducted by some researchers for the US army to redesign the uniforms of female soldiers so as to reduce the number of different uniform sizes while still providing each soldier with well-fitting khakis. Using the clustering technique called K-means algorithm, they were able to come up with sizes that fit particular body types (4). We are currently working on a pilot study of this nature as part of an undergraduate project, and have obtained some interesting results, initially.

The above are some, but not all, areas where data mining has been successfully applied or data mining is a potenital tool. Some of the commonly used data mining techniques will be discussed in the next section.

Common techniques

One main feature in data mining is the discovery of knowledge. It starts with the data and tries to tell us something we didn't know. Knowledge discovery has basically two approaches: Directed (or supervised), and undirected (or unsupervised) learning. In directed learning, the task is to explain (estimate, classify, predict, etc.) the value of some "target field" in terms of all other variables. Statistical procedures such as multiple regression and time series analysis are some examples of this nature. In undirected learning, there is no target field. One simply asks the computer to identify patterns or recognize relationships. Some techniques are designed for directed and some are for undirected learning, while some others work on both.

Market Basket Analysis (MBA)

MBA is a simple descriptive tool of undirected learning which can be used to find out which items sell together, which items to put on specials etc., in big grocery stores or supermarkets. Here, information given in a "market basket" is analyzed. Typically, each customer purchases different sets of items, in different quantities, at different times of the week. So, one can imagine the variation as well as valuable information contained in say, 100,000 records! By looking at co-occurrence of items, the chance of a formal rule of association such as "If A is purchased, then B and C are also purchased" can be worked out using large number of records. Basically, the analysis uses statistical and probabilistic ideas based on prior information. Fortunately, many supermarkets and stores have automated data on point-of-sale transactions which can be explored automatically if relevant tools are available.

Memory Based Reasoning (MBR)

MBR uses the concept of "identifying similar cases from experience and then applying the information from the cases to the problem at hand". One important feature is that MBR does not depend on the format of data. It is a powerful tool that works on situations such as classification of news items, fraud detection, customer response prediction, etc., where most other tools do not do well or cannot be used. Berry and Linoff discuss a case study where news stories have been classified using MBR with the help of "distance functions" based on a notion called "relevance feedback" (4). This is an interesting example which shows how free-text can be analyzed using MBR.

Automatic Cluster Detection (ACD)

This technique is a tool for unsupervised learning. In clustering, there is no pre-classifed data and we let the computer search for groups of records (or "clusters") that are similar to one another, assuming that similar records represent similar customers or suppliers or products. The technique uses modern clustering algorithms that are essentially based on multivariate statistical theory.

Link Analysis (LA)

LA is based on the mathematical area called "graph theory", and it exploits various relationships that exist among people, sources, or services, to solve real problems. In practice, we observe certain links such as, airlines linking cities together, phone calls linking people together, and referral patterns linking physicians to certain pharmaceutical companies. A graph will consist of nodes and relationships between nodes called edges. For example, these nodes may be cities linked by an airline. A typical problem in LA is to find the shortest path between 2 nodes in a graph. Another difficult problem is to find the shortest "Hamiltonian path", *i.e.* the shortest path that visits all nodes exactly once in a graph. Recent findings (1) have shown that sequences of DNA in a test tube (a biological computer?) can be analyzed to find a Hamiltonian path for a small graph, with speed much more faster than the fastest supercomputer! (4).

Decision Trees (DT)

These are powerful tools for classification and prediction, and are traditionally drawn with the root at the top and the leaves at the bottom. The

process starts with a record entering the tree at the root node. At the root, a test is applied to determine which "child node" the record will next belong to. This process is repeated until the record arrives at a "leaf node". All the records that end up at a given leaf are classified the same way and there is a unique path from the root to each leaf. Various algorithms are used to build DT and probabilistic ideas are used to measure the accuracy of classifications.

Artificial Neural Networks (NN)

Artificial Neural Networks, or simply Neural Nets (NN) are probably the most common data mining technique. NN are based on simple models describing neural interconnections in brains, to be used on computers. These tools learn from existing data and generalize the patterns inside these data for future prediction or classification. NN can also be used for time series prediction. One advantage with NN is their wide applicability and the availability of supporting tools on many platforms. However, one drawback is the difficulty in understanding what is happening inside the "black box" that produces models.

Genetic Algorithms (GA)

GA use ideas from genetics and natural selection to find the optimal sets of parameters that describe a predictive function. Thus, it is used for directed learning. The area is relatively new and therefore the applications are not so widespread, but is attracting a lot of attention at present.

Support Vector Machines (SVMs)

SVMs are tools that incorporate advanced mathematical and statistical ideas. These are powerful classification systems based on a variation of an existing statistical data mining tool, "regularization techniques for regression" (10). SVMs have shown very successful applications particularly in biological classification tasks including gene expression analysis. At present, much attention is focused on this promising data mining tool.

An illustration

For a given data mining problem, one or more of the above may be used. For most techniques there are supporting tools available. Standard statistical approach also works in many situations. Statistics is undoubtedly the best tool for data analysis. In some cases, statistics produces results as good as any of the above data mining techniques.

The following is an illustration of a small scale data mining application in agricultural research. This is based on the work previously published by us (8).

Background of the study

The data come from an experiment carried out by the Rubber Research Institute of Sri Lanka with the aim of studying the sensitivity of some *Hevea* clones to environmental differences. The experiment consist of 7 sites and 10 varieties each with 10 randomized plots (replicates) per site. A previous study had indicated the existence of variety × site interaction with respect to growth of these 10 varieties. Therefore, one objective was to find suitable mathematical models to describe the growth over the period 1975-1986 (12 years). It was also of importance to find any natural grouping of different variety – site combinations with respect to growth (8). The latter is an unsupervised data mining exercise.

Methods and Results

Descriptive analysis on the growth over the 12 years was done using basic statistical tools such as scatter plots and summary tables. Using a purely statistical exercise, "logistic growth curve" of the form

$$y_i = \frac{\theta_1}{1 + \theta_2 e^{-\theta_3 x_i}} + \varepsilon_i,$$

where y_i is the average tree girth (cm), x_i is the time (years) θ's are model parameters, and ε_i is an error term, was found to be a reasonable fit to growth data of all varieties.

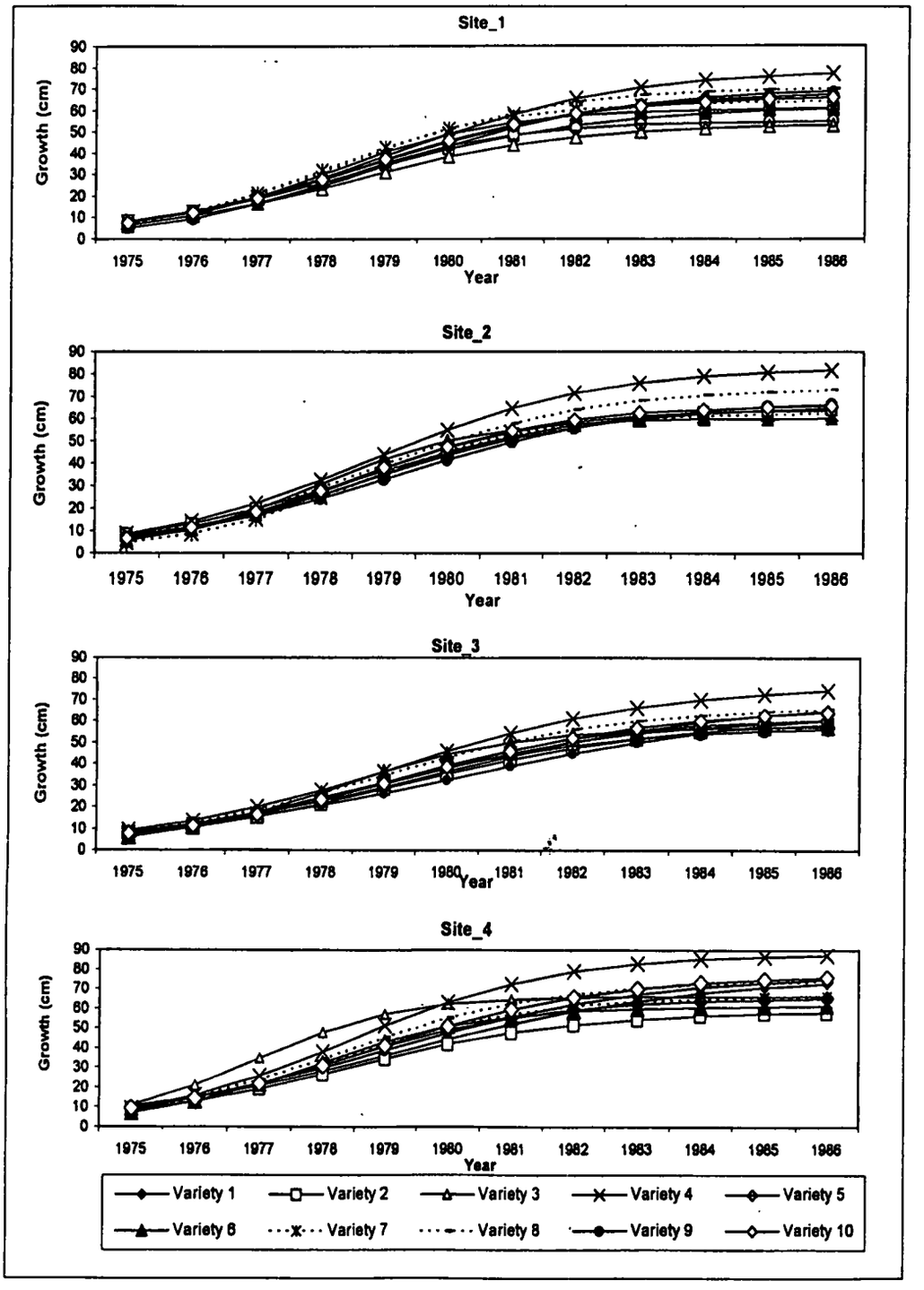

Figure 1. Growth curves of the 10 varieties in Sites 1-4

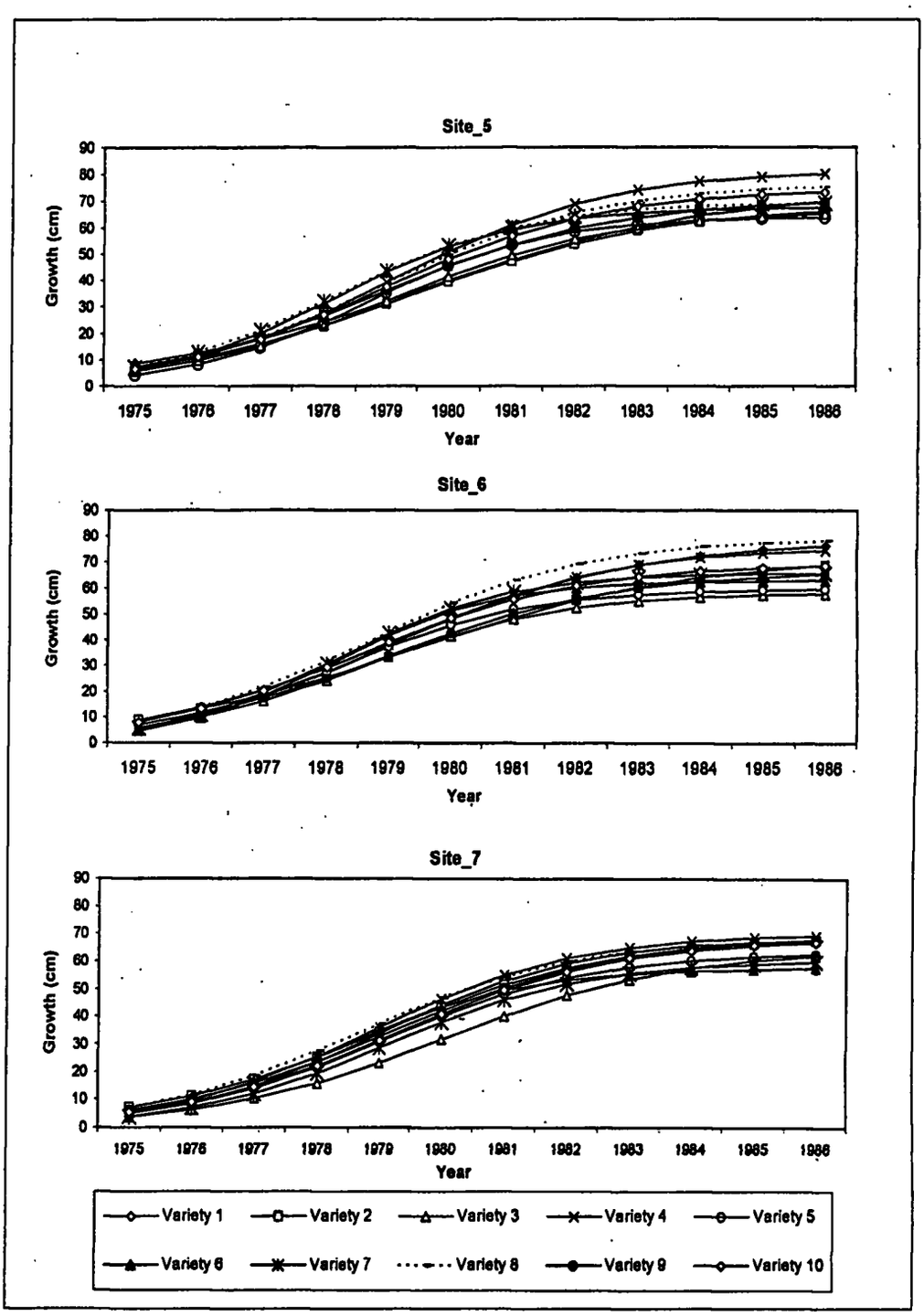

Figure 2. Growth curves of the 10 varieties in Sites 5-7

It is interesting to note that θ_1 represents girth at the final year, θ_2 the inital girth, and θ_3 growth rate. Seventy (70) nonlinear models of this type with different parameter estimates were obtained for the 10 varieties × 7 sites using the tool PROC NLIN of SAS with the iterative Gauss-Newton option (See Fig. 1 and Fig. 2). Up to this point, it was a typical statistical data analysis problem. However, after this point some unsupervised data mining tools had to be employed to group the varieties with respect to their growth pattern. Namely, 3 approaches: Single Linkage Clustering method, Confidence Regions, and Confidence Intervals, were employed. It was found that a reasonable grouping with fewer identifiable groups was given only by the Confidence Interval approach in this case (6).

Conclusion

The area of data mining is rapidly growing. Today, supporting tools are being developed for most data mining techniques, but they are relatively expensive. A reason for this may be that most significant and hence profitable applications of data mining are in marketing. However, not all sectors can benefit from data mining unless they have large quantities of automated data that could be explored at low cost. Increased statistical input into most unsupervised data mining techniques is certainly going to help make the end results more acceptable. Big vendors of statistical software have now come forward to add data mining tools to their products, probably due to the increasing demand. With the rapid development of the Information Technology, many enterprises can be expected to rely more on data analysis and statistical models to extract treasures of information from large pools of data, and transform them into profitable actions.

References

1. Hastie T. (http://www.stat.stanford.edu/~hastie/MRC/sldm.html) 2000.
2. Berry M.J.A., Linoff G. Data Mining Techniques for Marketing, Sales, and Customer Support, Wiley Publication. 1997.
3. Christensen W.F., Cook D. Data Mining Soil Characteristics Affecting Corn Yield, Technical Report. Iowa State University. 1999; 6: 1-33.
4. Guohua, W., Francis T.E.H. Data Mining: Concepts, Applications, and Techniques, ASEAN Journal on Science and Technology for Development 2000; 17: 77-86.
5. Maindonald J. (http://wwwmaths.anu.edu.au/~johnm/dm/dmpaper.html) 2000.
6. Milliken G.A. Applications of Nonlinear Statistical Models, Part 1; Technical Report, Department of Statistics, Kansas State University.
7. Pryke A. (http://www.andypryke.com/university/dm_docs/dm_intro.html) 2001.
8. Padmika K.D.T., Jayasekera N.E.M., Wickremasinghe W.N., Karunasekera K.B. A study on modelling early growth of some *Hevea* clones, Proceedings of the 56th Annual Session, Sri Lanka Association for the Advancement of Science 2000: 202.
9. Ramaswamy S., Tamayo P., Rifkin R. Mukherjee S., Yeang C., Angelo M., Ladd C., Reich M., Latulippe E., Mesirov J.P., Poggio T., Gerald W., Loda M., Lanfer E.S., Golub T. R. Multiclass cancer diagnosis using tumor gene expression signatures. Proceedings of the National Academy of Science 2001; 98, 26: 15149-15154.
10. Vapnik V.N. Statistical Learning Theory. John Wiley & Sons, N.Y. 1998.