Information Research, Vol. 3 No. 2, September 1997

Discovering the hidden secrets in your data - the data mining approach to information

Michael Lloyd-Williams
Department of Information Studies
University of Sheffield

Introduction

Nowadays, digital information is relatively easy to capture and fairly inexpensive to store. The digital revolution has seen collections of data grow in size, and the complexity of the data therein increase. Advances in technology have resulted in our ability to meaningfully analyse and understand the data we gather lagging far behind our ability to capture and store these data (Fayyad, *et al.* 1996). A question commonly arising as a result of this state of affairs is, having gathered such quantities of data, what do we actually do with it? (Fayyad & Uthurasamy, 1996)

It is often the case that large collections of data, however well structured, conceal implicit patterns of information that cannot be readily detected by conventional analysis techniques (Lloyd-Williams, *et al.*, 1995). Such information may often be usefully analysed using a set of techniques referred to as *knowledge discovery* or *data mining*. These techniques essentially seek to build a better understanding of data, and in building characterisations of data that can be used as a basis for further analysis (Limb & Meggs, 1995), extract *value* from *volume* (Scarfe & Shortland, 1995). This paper describes a number of empirical studies of the use of the data mining approach to the analyse of health information. The context is unimportant, the examples described serving to highlight the factors perceived as influencing the success or otherwise of the data mining approach in each case, and to illustrate the generic difficulties that may be encountered during the data mining process, and how these difficulties may be overcome.

Knowledge discovery and data mining

It is generally accepted that the reason for capturing and storing large amounts of data is due to the belief that there is valuable information implicitly coded within it (Fayyad & Uthurasamy, 1996). An important issue is therefore how is this hidden information (if it exists at all) to be revealed? Traditional methods of knowledge generation rely largely upon manual analysis and interpretation (Fayyad, et al, 1996). However, as data collections continue to grow in size and complexity, there is a corresponding growing need for more sophisticated techniques of analysis (Fayyad, *et al*, 1996). One such innovative approach to the knowledge discovery process is known as *data mining*.

Data mining is essentially the computer-assisted process of information analysis (Limb & Meggs, 1995) . This can be performed using either a top-down or a bottom-up approach. Bottom-up data mining analyses raw data in an attempt to discover hidden trends and groups, whereas the aim of top-down data mining is to test a specific hypothesis (Hedberg, 1995). Data mining may be performed using a variety of techniques, including intelligent agents, powerful database queries, and multi-dimensional analysis tools (Watterson,1995). Multi-dimensional analysis tools include the use of neural networks, as described in this work.

The data mining approach expedites the initial stages of information analysis, thereby quickly providing initial feedback that may be further and more thoroughly investigated if appropriate. The results obtained are not (unless otherwise specified) influenced by preconceptions of the semantics of the data undergoing analysis. Patterns and trends may therefore be revealed that may otherwise remain undetected, and/or not considered.

It should be stated at this juncture that this paper advocates the use of data mining techniques in *conjunction* with traditional approaches to analysis, and not as a direct replacement.

The knowledge discovery process

According to researchers such as Fayyad, *et al* (1996), the process of knowledge discovery via data mining can be divided into four basic activities; *selection*, *pre-processing*, *data mining*, and *interpretation*. These stages are discussed in the following text. A graphical representation of the general process is presented in Figure 1.

Selection

Selection involves creating the target data set, i.e. the data set to about undergo analysis. As discussed previously, modern datasets may be both large and complex. Large datasets which are not particularly complex may generally be subjected in their entirety to the analysis process (subject to technical constraints). Indeed, the larger the amount of available data, the greater the likelihood that an identifiable trend or pattern may be identified and statistically validated. However, if the dataset is relatively complex, it is often considered impractical to attempt to subject the complete dataset for analysis. It is a common misconception to assume that the complete dataset should be submitted to the data mining software, which in turn will automatically resolve any problems and make sense of any inconsistencies. This is not in fact the case, and is partly due to the probability that the data represents a number of different aspects of the domain which may not be directly related. Subjecting such data to automated analysis may result in the identification of meaningless patterns or trends, which in turn wastes time and effort. Careful thought should therefore be given as to the purpose of the analysis exercise, and a target dataset created which contains data that reflects this purpose.

Pre-processing

Pre-processing involves preparing the dataset for analysis by the data mining software to be used. This may involve resolving undesirable data characteristics such as missing data (non-complete fields), irrelevant fields, non-variant fields, skewed fields, and outlying data points. The pre-processing activities may result in the generation of a number of (potentially overlapping) subsets of the original target dataset.

Data fields are generally viewed as being complete if 70% or more of the records contain values (SPSS Inc., 1995). In cases where the field is considered theoretically complete, but in fact is less than 100% complete, various techniques such as estimation, or assigning the category mode are available for producing *synthetic* data. The generation of accurate values to represent missing data is currently one of the main research areas occupying the data mining community (Fayyad, *et al*, 1996).

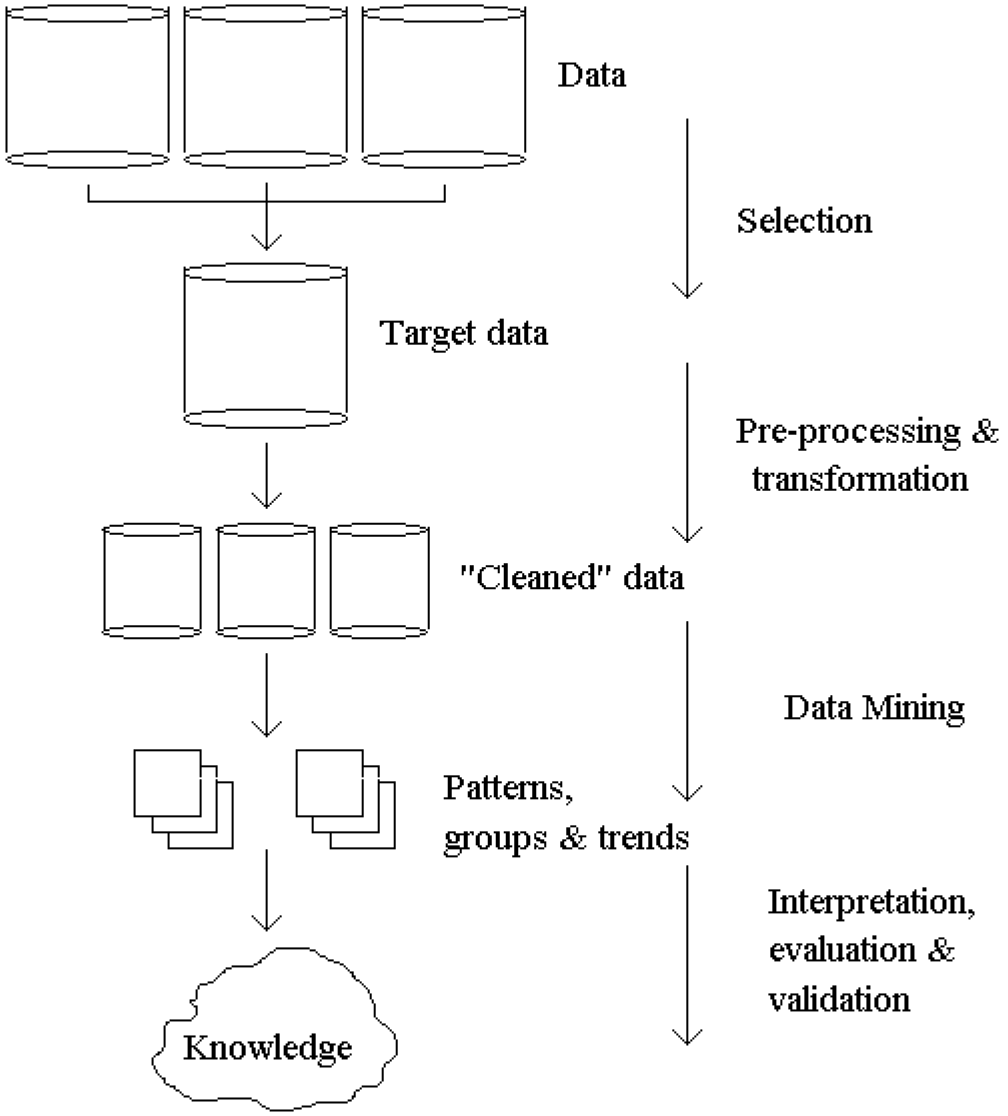
Data
Selection
Target data
Pre-processing & transformation
"Cleaned" data
Data Mining
Patterns, groups & trends
Interpretation, evaluation & validation
Knowledge

Figure 1: Knowledge Discovery Overview

Depending upon the original source of the data, and the storage format employed, pre-processing may also involve converting the data into a format acceptable to the data mining software being used. This initial collection and manipulation of data (during the selection and pre-processing stages) in the data mining process is sometimes referred to as *collection* and *cleaning* [6].

Data mining

This involves subjecting the cleaned data to analysis by the data mining software in an attempt to identify hidden trends or patterns, or to test specific hypotheses. It is recommended that any (apparently) significant results obtained are validated using traditional statistical techniques at this stage.

Interpretation

This involves the analysis and interpretation of the results produced. This may well involve returning to previous stages to carry out additional activities in order to provide further information if necessary.

Artificial Neural Networks

Artificial neural networks are defined as information processing systems inspired by the structure or architecture of the brain (Caudill & Butler, 1990) . They are constructed from interconnecting processing elements which are analogous to neurones. The two main techniques employed by neural networks are known as *supervised learning* and *unsupervised learning*.

Supervised learning is essentially a two stage process; firstly training the neural network to recognise different classes of data by exposing it to a series of examples, and secondly, testing how well it has learned from these examples by supplying it with a previously unseen set of data.

Unsupervised learning is so called as the neural network requires no initial information regarding the correct classification of the data it is presented with. The neural network employing unsupervised learning is able to analyse a multi-dimensional data set in order to discover the natural clusters and sub-clusters that exist within that data. Neural networks using this technique are able to identify their own classification schemes based upon the structure of the data provided, thus reducing its dimensionality. Unsupervised pattern recognition is therefore sometimes called *cluster analysis* (Backer, 1978).

The empirical studies presented in this paper were carried out primarily using the unsupervised learning technique based upon the Kohonen Self Organising Map. The principles underpinning the Kohonen Map have been described in some detail in works such as those by Kohonen (1988) and Zupan &: Gasteiger (1993) , therefore only a brief description is provided here. The Kohonen Map is essentially constructed as a two dimensional grid of nodes. data are presented to each node in the grid simultaneously. Each node then competes in order to represent the data pattern currently being presented. One node will *win* this competition, and will have its parameters altered so that if this same data pattern were to appear again, the same node would have the best chance of winning. The overall effect is a two dimensional grid of nodes, that each respond to some subset or cluster of the incoming multidimensional data. Differing patterns are represented by nodes that are separated on the grid by varying distances, according to the level of difference. In summary, the Kohonen Map maps the high-dimensional continuous space occupied by the input data patterns into a lower-dimensional discrete space formed by an organised lattice of nodes (Wilkins, *et al.*, 1993). The Kohonen Map approach has been widely and successfully used in multivariate data analysis, and being closely related to other methods (such as the k-means type), is often seen as the best choice of analysis method from an interpretational point of view (Murtagh &: Hernández-Pajares, *et al.*, 1995).

Emprical studies

This section present details of three empirical studies of the use of the data mining approach in the analysis of health information. The information analysed during these studies was extracted from the World Health Organisation's *Health for All Database, the Babies at Risk of Intrapartum Asphyxia* database, and a series of databases containing infertility information. These studies represent both successful and unsuccessful instances of data mining activities, and are intended to assist in highlighting the factors perceived as influencing the outcome of each of the projects concerned.

The *Health for All* database

The World Health Organisation's (WHO) *Health for All (HFA)* Database was created to make health-related data collected by the WHO available to outside users. The Database contains statistical indicators for the WHO HFA targets relating to Health-for-All in Europe by the year 2000. These include: better health, mortality/morbidity, lifestyles conducive to health, healthy environment, health services, and health policies. Other statistical indicators relate to demographics, socio-economics and other supplementary health-related information. Data are present for all European countries and includes those countries in existence before, and after the recent political changes in central and eastern Europe as well as the former Soviet Republics. Data are included from 1970 to 1993 inclusive, although it is important to note that there are gaps in the data that are available.

The approach taken to analyse the HFA database generally corresponded with that illustrated in Figure 1. During the selection process, mortality data relating to the following conditions was extracted form the HFA database; life expectancy at birth; probability of dying before five years of age; infant mortality; post-neonatal mortality; standardised death rates (SDR) for circulatory diseases; SDR for malignant neoplasms; SDR for external causes of injury and poisoning; SDR for suicide and self-inflicted injury.

Data were extracted for 39 European countries. Although a far higher number of countries are represented on the Database, the data could only be considered complete for 39 of them. The data were then converted into a format acceptable to the software being used. An underlying aim of the study was to track any changes in the data that may have occurred over the years for the same samples of countries in order to examine whether any patterns identified remained consistent over time. This necessitated the selection of data that were separated by intervals of three to four years (where possible) since it was expected that changes, if any, would be more apparent on this time scale. Three years were chosen, giving three subsets of data (1982, 1986, and 1989). The extracted data was then analysed by custom written Kohonen Self Organising Map software in order to identify possible groupings. Finally, standard statistical techniques were used to evaluate the validity of the groupings. This investigation therefore made use of the bottom-up data mining approach described previously.

Preliminary work resulted in two distinct groups or clusters of countries in each year being apparent (Lloyd-Williams, *et al.*, 1996). It was observed that all countries in the first of the groups were from Central and Eastern Europe or from the former Soviet Republics, while all countries in the second group were from Northern (i.e., the Nordic countries), Western, or Southern Europe. A two-sample t-test was used to validate the groupings identified. The results obtained confirmed that the identified groups were significantly different or separated from each other.

In addition to the geographical division, the classification also appeared to reflect differences in wealth. Countries in the first of the groups were relatively poor; all countries (excluding the former Soviet Union) had GNP per capita of less than US$5,000 in 1989 (the latest year for which GNP figures were available). Countries in the second of the groups were relatively wealthy; all countries had GNP per capita exceeding US$5,000 (except Portugal, whose 1989 GNP per capita is just under US$4,300). A t-test was performed, and the result indicated a significant difference between the two groups in terms of GNP per capita. The observation that the classification appeared to reflect two different GNP groups suggested that GNP could be inter-related with the health indicators. In order to further explore this possibility, the coefficient of correlation was calculated between GNP and all seven HFA indicators used in the initial analysis. Results obtained indicated that GNP is strongly and positively correlated with life expectancy, and strongly but negatively correlated with the SDR for diseases of the circulatory system. A small and positive correlation was found to exist between GNP and the SDR for malignant neoplasms, but fairly large negative correlations between GNP and all the other indicators, except the SDR for suicide and self-inflicted injury. Overall, the available data indicated that death rates for malignant neoplasms tended to rise with increasing affluence, while death rates for other diseases tended to fall.

Further work was then performed in order to obtain a finer classification. This work resulted in the identification of six groups, which were essentially sub-divisions of the two groups produced by the preliminary work (Lloyd-Williams & Williams, 1996). Characteristics of the groups ranged from the lowest life expectancy, coupled with the highest probability of dying before five and infant mortality rate, to the highest mean life expectancy, coupled with the lowest probability of dying before five, and infant mortality rate. Over time, mean life expectancy increased, while the probability of dying before five and infant mortality rate both decreased for all groups. Detailed summary statistics of each group may be found elsewhere (Williams, 1995)

Group membership of all six groups remained relatively stable over the period under consideration. Eight of the 39 countries did experience movement between groups over the years, however in all such cases, cumulative movement was limited to an adjacent group.

In order to validate the groupings identified, Euclidean distances were calculated, and the generalised t-test used to test if the underlying means of each group pair were significantly different from each other. Results obtained confirm that the identified groups were significantly different or separated from each other. The classifications obtained can therefore be said to be valid based upon the available data.

The success of the application of the data mining approach to the analysis of the HFA database may be attributed to a number of factors as follows. Careful initial data selection and associated pre-processing ensured that that the target dataset was largely complete, containing only highly relevant data appropriate to the investigation. No non-variant or skewed fields were included. Similarly, no fields in the available data contained outlying data points. The dataset processed could therefore in this instance be viewed as being an ideal subject for the data mining.

The *Babies at Risk of Intrapartum Asphyxia* database

The *Babies at Risk of Intrapartum Asphyxia* database was analysed in conjunction with staff at the Sheffield Children's Hospital. This database contains data collected from a wider study on the relationship between intrapartum asphyxia and neonatal encephalopathy. Neonatal encephalopathy is a condition characterised by impairment of consciousness, abnormalities of muscle tone and of feeding. There is also frequent evidence of injury to other organs such as the heart, kidney, and gastrointestinal system. Neonatal encephalopathy is thought to arise when there is significant intrapartum asphyxia, that is, the foetus is deprived of oxygen during labour.

The *Babies at Risk of Intrapartum Asphyxia* database contains detailed obstetric data, including cardiotocogram (CTG) traces taken during labour. Both the foetal heart rate (FHR) and the uterine contractions are represented on the trace produced. Also present on the database are pre-labour assessments of maternal and antenatal risk factors that might influence the outcome of the labour. Factors relating to the quality of intrapartum obstetric care are also recorded.

In order to construct the database, the CTG traces were analysed by dividing each trace into thirty-minute epochs, and then examining each epoch for abnormal FHR patterns. A *severity score* was then applied to each abnormality detected. Each of the 128 patients represented on the database is therefore associated with a number of epochs, the average being fourteen (the minimum being one, and the maximum being forty). Each epoch in turn is further described using a number of parameters presenting various aspects of the associated FHR.

All the activities relating to data capture, analysis, and initial database construction were carried out by medical staff involved in the wider study on the relationship between intrapartum asphyxia and neonatal encephalopathy. The target dataset was therefore created primarily for use within this study, rather than for analysis using neural networks.

Due to the purpose of the main study, much of the data represented on the database are encoded, rather than being specifically represented. For instance, the baseline FHR in beats per minute (bpm) recorded on the database varies from 100-109 at the lowest level, to >180 at its highest. The baseline FHR is recorded on a scale of -2 (representing 100-109 bpm) to +3 (representing >180 bpm). These figures represent whether or not the baseline FHR is slower or faster than would be expected. The bradycardia parameter is used to indicate a slow FHR (< 100 bpm). In the absence of bradycardia, this parameter is set to zero (normal). A period of greater than three minutes with recovery is represented by a value of +3, and a value of +4 if there is no recovery. This use of a zero value to indicate a normal reading is employed in a large proportion of the parameters of the database.

The Kohonen Self Organising Map software employed to analyse the dataset expects the data therein to be presented on an interval or ordinal scale. That is, that the range of parameter values stand in some relationship to each other. The software also expects values that lie adjacent to each other within a parameter range to be similar to each other. However, some parameters on the database occupy a nominal scale where parameter values do not relate to each other. For instance, the uterine contractions parameter is represented as zero (normal), +3 (over-contracting), or +1 (the recording of the contraction on the trace is technically poor). This parameter was therefore considered to be inappropriate for use in its existing form, and removed from the target dataset.

Initial analysis work was not without its difficulties due to the nature of the data involved. At one stage of processing, 52% of the available dataset presented parameters that were all set to zero (indicating a normal response), apart from the baseline FHR. This extremely high level of non-variant data resulted in no discernible patterns being detectable at the early stage of work.

After performing further selection and pre-processing activities (including the removal of the majority of non-variant data), some success was achieved. Two distinct groups were identified within the database, and their existence statistically validated using a two-sample t-test. Further work involved the training of a neural network employing the radial basis function (RBF) approach to recognise the two groups, and to differentiate between members of these groups. During this period, an overall probability of correct recognition of 85% was achieved, providing further evidence that the groups identified did exist within the database.

Further work on the database resulted in a refinement of the original two groups being achieved. The resulting four groups were again found to be statistically different from each other, each exhibiting specific combinations of FHR patterns. The use of the RBF software again resulted in successful identification of the groups, with 100% of records within the first group, 97% of records within the second group, and 99% of records within the third and fourth groups being correctly identified.

Although the identification of statistically valid groups may indicate that this data mining exercise was successful, it should be noted that the data used to obtain these groups represented only a proportion of that available. The high proportion of non-variant data present precluded much of the database from being used in the analysis process. Further limiting factors in this case included the use of a nominal scale to represent certain parameters. The fact that the data provided had already undergone pre-coding also restricted the ability to perform any subsequent meaningful data manipulation, as the original data values were in many instances unknown.

Infertility databases

Despite recent improvements in infertility diagnosis and the increase in sophistication and variety of treatment techniques, there still appears to be great difficulty in successfully predicting how a particular patient will respond to a specific course of treatment. Although many types of data can be collected which in theory are relevant to the likelihood of successful treatment, in reality the complexity of the interactions between these parameters appear to be beyond the capabilities of conventional methods of analysis. The primary aim of this investigation was therefore to investigate the use of the data mining approach in assisting in the prediction of whether a specific patient would be successfully treated using a particular treatment pathway. Against this background, three databases holding data relating to three different aspects of infertility diagnosis and treatment were provided for analysis using the data mining approach.

The first of these databases contained details of patients who had undergone ovulation induction with gonadotrophins. This database was analysed in conjunction with staff at the Jessop Hospital, Sheffield. This database holds 17 parameters for each of the 122 patients represented, including information relating to the treatment outcome, i.e. whether the patient became pregnant or not. The main objective of this study was to attempt to identify the combination of characteristics that appeared to indicate a successful treatment outcome.

A second database containing details of patients who had undergone stimulated cycles of IVF treatment was analysed in conjunction with staff at St. Michael's Hospital, Bristol. Each of the 403 patients represented on this database had three fertilised eggs implanted (the optimal number to achieve one successful pregnancy whilst avoiding multiple pregnancies). For each patient, there were therefore four possible outcomes (0-3 pregnancies). The main aim of this study was to identify the characteristics of patients who are most likely to achieve a single successful pregnancy, the clinicians wishing to avoid the potential for multiple (and hence possibly complicated) pregnancies.

A third database containing details of 65 patients who had undergone natural cycle IVF treatment was also analysed in conjunction with staff at St. Michael's Hospital, Bristol. Each patient is represented on the database by 12 parameters and a corresponding diagnostic category (endometriosis, tubal damage, or unexplained infertility). The aim of this study was slightly different fro the previous two described in attempting to determine whether the specific combination of 12 parameter values could be associated with the diagnostic category of the patient.

The data mining approach failed to provide the information that was specifically required in all three of the cases described. During the initial stages of investigation, all three datasets appeared initially to be highly suitable for analysis using the data mining approach. Despite the fact that some parameters represented data values using a nominal scale, the original databases exhibited little missing data, few irrelevant fields, no non-variant fields, few skewed fields, and few outlying data points. Any limited undesirable characteristics originally present were removed during the selection and pre-processing activities.

The main problem encountered during this investigation appears to be fundamental in that the patterns representing the information required were not evident within the available datasets. It was therefore concluded that the data mining approach was unable to assist in the analysis of the data provided. This appeared to be primarily due to the fact that the range of factors that fully determine a couple's ability to conceive is not known. It is therefore reasonable to assume that in this case, the pattern (or a recognisable proportion) formed by these factors is not present in the data currently being accumulated by the clinical staff.

Conclusions

This paper has provided an introduction to the concepts of knowledge discovery and data mining. The empirical studies presented are intended to highlight the factors perceived as contributing to the success or otherwise of the data mining approach.

It should be evident from the studies described that the potential for success is largely determined prior to the actual data mining activity, i.e. during the activities performed in the production of the cleaned data. Extreme care should therefore be taken during the selection and pre-processing activities in order to ensure wherever possible that the target dataset actually contains relevant and usable data, and that these data are in a form suitable for use.

This latter point is particularly relevant, as the form required by the providers may not be compatible with that required for the data mining process. This was evidenced by the analysis of the Babies at Risk of Intrapartum Asphyxia database, where coding activities performed as a result of a wider study precluded much of the data from analysis by neural networks. The likelihood of success of a data mining project is therefore highly dependent upon the quality and format of the data made available.

It is also apparent from these studies (and in particular, the analysis of the infertility databases), that if the data provided does not contain useful information within the context of the focus of the investigation, then the use of neural networks cannot generate such information any more than traditional analysis techniques can. However, it may well be the case that the use of neural networks for data mining allows this conclusion to be reached more quickly than might ordinarily be the case.

Finally, it should be stated once more that this paper advocates the use of data mining as an approach that can be used to expedite the initial stages of information analysis in order that the results obtained may be more thoroughly investigated. It should be used in conjunction with traditional approaches, not in direct competition.

References

- Backer, E. (1978) *Cluster Analysis by Optimal Decomposition of Induced Fuzzy Sets.* Delftse Universitaire Pers. Thesis: Delft University.
- Caudill, M. & Butler, C. (1990) *Naturally Intelligent Systems*. Cambridge: MIT Press.
- Fayyad, U. & Uthurasamy, R. (1996) "Data Mining and Knowledge Discovery in Databases" *Communications of the ACM*, 39(11), 24-26.
- Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996) "The KDD Process for Extracting Useful Knowledge from Volumes of Data" *Communications of the ACM*, 39(11), 27-34.
- Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. & Uthurasamy, R. (1996) *Advances in Knowledge Discovery and Data Mining*. Cambridge: AAAI/MIT Press.
- Hedberg, S.R. (1995) "The Data Gold Rush" *Byte*, 20(10), 83-88.
- Kohonen, T. (1988) *Self-Organisation and Associative Memory*. New York: Springer-Verlag.
- Limb, P.R., & Meggs, G.J. (1995) "Data Mining -Tools and Techniques" *British Telecom Technology Journal*, 12(4), 32-41.
- Lloyd-Williams, M., Jenkins, J., Howden-Leach, H., Mathur, M., Morris, C & Cooke, I. (1995) "Knowledge Discovery in an Infertility Database Using Artificial Neural Networks" *IEE Colloquium on Knowledge Discovery in Databases*. IEE Digest No:1995/021(B).
- Lloyd-Williams, M. & Williams, S. (1996) "A Neural Network Approach to Analysing Healthcare Information", *Topics in Health Information Management*, 17(2), 26-33.
- Lloyd-Williams, M., Williams, S, Bath, P. & Morris, C. (1996) "Knowledge Discovery in the WHO *Health for All* Database: Developing A Taxonomy of Mortality Patterns for European Countries", In: Richards, B. & de Glanville,H. (eds) *Current Perspectives in Healthcare Computing. Proceedings of HC96,* pp. 551-556. Weybridge: BJHC.
- Murtagh, F. & Hernández-Pajares, M. (1995) "The Kohonen Self-Organizing Map Method: An Assessment" *Journal of Classification* 12, 165-190.
- Scarfe, R. & Shortland, R.J. (1995) "Data Mining Applications in BT" *IEE Colloquium on Knowledge Discovery in Databases*. IEE Digest No:1995/021(B).
- SPSS Inc. (1995) *Neural Connection Applications Guide*. Chicago: SPSS Inc.
- Watterson, K. (1995) "A Data Miner's Tools" *Byte* 20(10), 91-96. 4.
- Wilkins, M. F., Boddy, L., & Morris, C. W. (1993) "Kohonen Maps and Learning Vector Quantization: Neural Networks for Analysis of Multivariate Biological Data" *Binary*, 6, 64-72.
- Williams, T.S. "*Knowledge Discovery in the WHO Health for All Database: Developing A Taxonomy of Mortality Patterns for European Countries*" MSc Thesis, University of Sheffield, Sheffield, England, 1995
- Zupan, J. & Gasteiger, J. (1993) *Neural Networks for Chemists*. Weinheim: VCH.

This paper was originally presented at the SCGISA & RRL.net Workshop on Health and Crime Data Analysis, held at the Novotel,

Sheffield, May 1997

How to cite this paper:

Lloyd-Williams, Michael (1997) "Discovering the hidden secrets in your data - the data mining approach to information" *Information Research*, **3**(2) Available at: http://informationr.net/ir/3-2/paper36.html

Check for citations, using Google Scholar

Contents

11112
Web Counter
Counting only since 28 December 2002

Home