THE LH5 MODEL FOR DATA MINING

Dimitris Parapadakis
Dpt. of Artificial Intelligence and Interactive Multimedia
Elia El-Darzi
Dpt. of Information Software Systems
Harrow School of Computer Science,
University of Westminster, UK
Watford Road, HA1 3TP
E-MAIL: D.Parapadakis@wmin.ac.uk
Tel: (+44020) 79115000 (ext. 4422)
Fax: (+44020) 79115906

ABSTRACT

In the age of E-Business many companies are faced with massive data sets that must be analysed for gaining a competitive edge. These data sets are in many instances incomplete and quite often not of very high quality. Although statistical analysis can be used to pre-process these data sets, this technique has its own limitations. In this paper we are presenting a system – and its underlying model – that can be used to investigate the integrity of existing data and pre-process the data into clearer data sets to be mined. LH5 is a rule-based system, capable of self-learning and is illustrated using a medical data set.

INTRODUCTION

The strong competition between electronic businesses employing and analysing large amounts of electronic data has led to an increase in the use of intelligent methods to extract useful 'meaning' from raw data. This meaning may take the form of associations, sequential patterns, classifiers, or clusters, but, in the end, the function used is one of finding common patterns among seemingly different people, products, customer records, and events. Data mining is a relatively new science, building on the strengths of machine learning – extracting classifying rules from data – statistical analysis, and operational research, to allow decision makers to find useful patterns in large data 'mines'. In this, it emulates human intelligence where, for example, in the case of 70% of customers who bought product A but also product B an association is identified to help future decision making.

Mining Sets from Multiple Sources

A common problem in learning methods is that the quality of the analysis depends on the quality of data collected. Many companies are now interested in mining data; for example, a multinational company can be mining its Internet sales data to identify classifications of geographical patterns; the information being sought is whether customers from Europe are buying a particular product under certain conditions in larger quantities than US customers. Applying data mining tools on the data can extract this information with ease; assuming that the company keeps accurate sales records with consistent rigour in all departments, the answer will be pretty accurate. However, searching for the same information on sales data collected from departments that record with different standards of quality or even searching for the same information on sales data collected from other companies through the World-Wide-Web will not produce results of similar accuracy. The answer sought may be found through the same statistical analysis and collation but, being partially based on low quality data, how trustworthy is this answer? The problem is that, quite often, the larger the distribution of the data collection points in an organisation the greater the chances that some of these sources of data are likely to be sources that do not keep accurate records. The result of this problem is that mining this data can lead to wrong decision making in the organisation. Some of these organisations are likely to be departments (internally) or companies (externally) that will succeed, some are likely to be departments or companies that will go bankrupt within a year, and some are likely to be departments or companies that publish misleading data available for marketing reasons.

Cost of Decision Making

As data mining techniques mature, information analysts have turned their attention to the wealth of knowledge that can be extracted by mining large data sets such as medical records. Analysis of medical records can identify and present trends, support theories, and answer 'what if' questions. The results merit the support of the field: faster identification of epidemics, more accurate links to causation, better management of medical resources [3].

Yet, as said earlier, the success of decision making is related, not only to the methods applied, but also to the quality of the data used. In addition, knowledge acquisition methods – regardless if they are used with experts or not – can never attain a true representation of what is in a person's mind. In the case of collection of medical records there are a number of factors that can affect how a certain symptom can be recorded and how impartially and accurately a diagnosis is reached (see figure 1). At best, following a lengthy series of stages, the number of misrepresented or missed facts can be minimised – not eliminated. In medical records even this attempt to approximate accuracy is abandoned as resources can limit the length of discussion between doctor and patient to as little as 15 minutes. Therefore, the collection of patient records from which information is mined can contain errors, and the mined information will be an

approximation to the truth in the hope that, in the 'larger picture', most erroneous entries will cancel each other out.

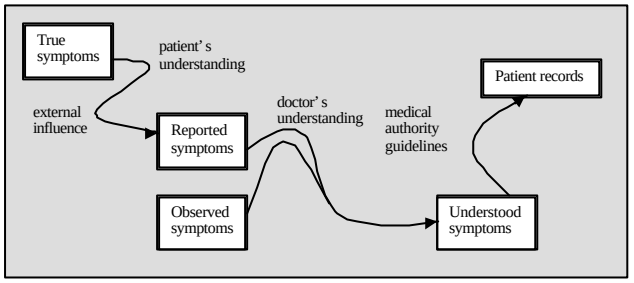

FIGURE 1

Whereas we can accept that the bias that comes into capturing and recording patient data cannot be removed, it is necessary to measure the quality decay within the data before data mining takes place. Data mining should then be based not only on the data but also on the quality of the data as well [6].

In the case of mining medical data the cost of mining an impure sample is evident. Centralised decision-making – affecting the channelling of funds towards specific problems – is made based on aggregate data collected and assumed to be mostly correct. The result is the often seen reversal of decision 'in light of new evidence'.

Many medical authorities are now making decisions based on data collected from a large set of doctors. Increasingly, web-based questionnaires are used to collect data and an authority will often allocate funds based on statistical analysis of the doctors' information, analyse and identify geographical trends, and, in short, try to maximise the health service provided within the budget available. This electronic analysis of data is used to improve the management of funds in the same way as with any other company relying on electronic data. However the frequency of cases where 'new evidence has shown that the previous assumption was wrong' suggest that not all sources of information are of the same quality. This is not difficult to accept; as all information providers will differ in their understanding of the rigour required for their work, doctors will differ in the quality and completeness of the information they provide. A careful human-based analysis of each doctor's record-keeping can identify the 'good' cases from the 'bad' cases so that the classification and statistical analysis is only done from the best sources. However, as with all companies, the cost of having one more employee to monitor the quality of each information provider is prohibiting, thus ensuring that erroneous data will continue to pollute the data mine.

THE LH5 MODEL

The LH5 model – a development of the Hydra Learning System [2] – has been designed to extract knowledge from data employing models of belief that mirror those learned by a human. When presented with new facts a computer-based classifying system will either confirm/expand its knowledge or correct it. Positive instances will confirm or add rules, negative instances will enhance rules to ensure discrimination. But in learning systems the data have to come from accurate and consistent records representing a correct 'view' of how the world is, otherwise it is accepted that the system will most likely fail [6]. Yet, unlike learning done by computers, human learning cannot afford to fail. It can handle an infinite number of sources with varied degrees of quality and still extract knowledge that can be used for decision making. To aid this, human learning has an ever-increased capacity of discriminating sources and detecting bias [4]. For example a 4-year old child can tell whether one person is more trustworthy than another, or whether other children are less to be believed than adults [5].

LH5 represents this model of learning by extending any normal rule-base with information as to the source of each fact – and rules derived from these facts – and with a belief set defining the trustworthiness of each source (see figure 2). At any one moment LH5 makes a distinction between the system's accepted view of the world, and the system's perceived views of others (which may or may not match the sources' real view of the world). The system's view must remain a consistent set of rules. But as this set is dependent on what LH5 has received from different sources, and on how much LH5 trusts these sources, rules in the set may change by the introduction of relevant facts, but also of irrelevant facts which change the belief set.

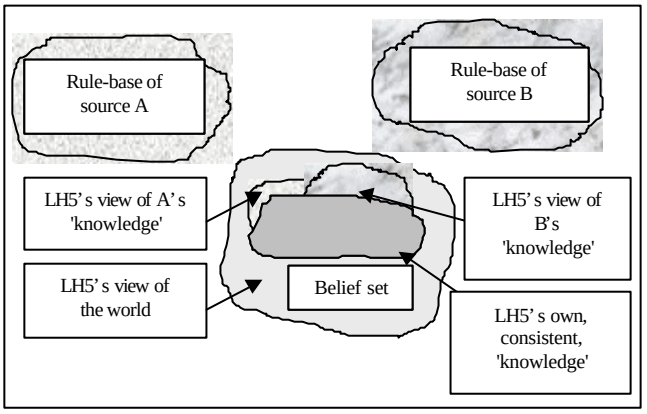

FIGURE 2

Improving the Quality of the Data Set

LH5 represents this model of learning by extending any normal set of learned information (rules, classifiers, sequences) with information about the source of each fact – and knowledge derived from these facts – and with a calculated belief set defining the trustworthiness of each source (see figure 3). At any one moment LH5 makes a distinction between the system's accepted view of the world, and the system's perceived views of others (which may or may not match the sources' real view of the world). The system's view must remain a consistent set of rules on which decisions can be based. However, as this set is dependent on what LH5 has received from different sources, and on how much LH5 trusts these sources, rules in the set may change by the introduction of relevant facts, but also of irrelevant facts which change the belief set.

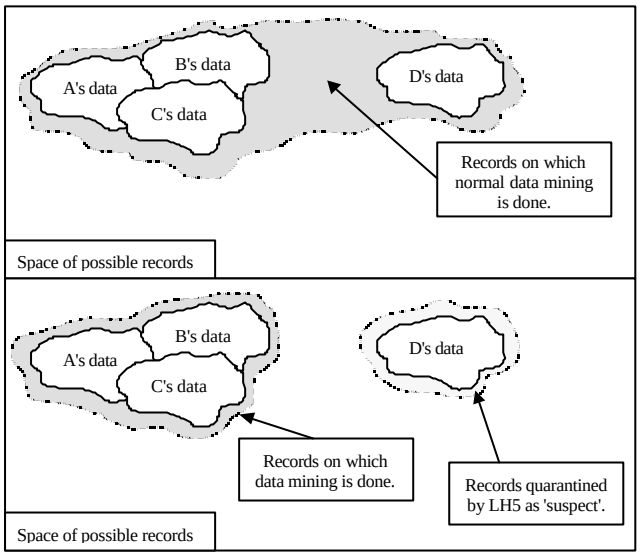

FIGURE 3

The strength of this model comes from turning the 'search for patterns' to the discriminated sources as well. When a child learns not to trust other children as much as adults, it can apply this knowledge to discriminate against newly encountered children, thus improving the efficiency of future learning. Similarly, the LH5 model can look for useful patterns on the discriminated source – in our case doctors – and look for common elements of training, locality, number of patients, to identify useful patterns that can be used to discriminate other sources and improve the overall quality of the data mine.

LH5's Discriminating Process

The LH5 system assumes a core 'image' of a source's knowledge, as derived from the data supplied by that source. For example, assume that we have the set of all doctors providing data on patients diagnosed to suffer (or not suffer) from panic disorder:

$Sources_{doctors}$*: {Dr. Blue, Dr. Verdi, Dr. Blanche, Dr. Mauve, Dr. Black, Dr. Scarlet, ...}*

Each doctor provides a number of patient records where, for each patient, we have the symptoms and a diagnosis (see figure 4).

As the data set does not include two doctor's diagnoses of the same patient, a data mining system can be used to collect all patient records to extract from them an abstract picture of what symptoms would normally be present when a patient suffers from panic disorder.

Though the experiment presented here is quite trivial and used for demonstration purposes, it is easy to see how including doctors' data without first filtering out those who may not follow the required rigour in their diagnosis and data recording can lead to wrong decisions being made. The results of this can affect funding, research and future diagnoses.

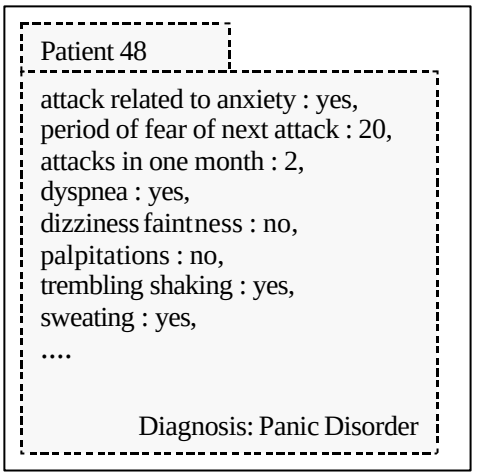

FIGURE 4

LH5 employs a customised learning algorithm, originally based on ID3 [7] to extract rules from the data it is given:

$$\text{Learn}(Data) \text{ ---> } Knowledge \quad (1)$$

However the knowledge gained could only represent the collective knowledge of doctors if it were possible for all doctors to be working in 'one mind', i.e. with the same methods, knowledge, experience, rigour. The LH5 model caters for this in two steps:

- Removal of singly untrustworthy sources (SUS)
- Removal of deviating sources (DS)

The first step is simply mirroring the human process of raising suspicion over any source that is not consistent with itself, on the grounds that any source that records the same facts but opposite conclusions at different times is, in the absence of other evidence, unreliable:

$$\text{RemoveSUS}(Sources) \text{ ---> } Sources' \quad (2)$$

where *Sources'* are in this case all doctors that have not recorded diagnoses contradicting other diagnoses from the same source.

For the second step LH5 isolates the data provided for each source and repeats the learning process for each source separately, creating a representation of each doctor's knowledge as it appears from the data recorded.

$$\left.\begin{array}{l}\text{Learn}(Data_{\text{Dr. Blue}}) \text{ ---> } Knowledge_{\text{Dr. Blue}} \\ \text{Learn}(Data_{\text{Dr. Verdi}}) \text{ ---> } Knowledge_{\text{Dr. Verdi}} \\ \text{etc.}\end{array}\right\} \quad (3)$$

In the case where all doctors were exposed to the same patients and symptoms and followed the same methods for diagnosis the knowledge representation that LH5 records for each doctor should be the same:

$$Knowledge_i = Knowledge_j \quad (4)$$

for all i,j in $Sources_{doctors}$2. However, as each $Knowledge_i$ represents a different doctor's knowledge – gained from

different experiences – a deviation can be expected between the representation of knowledge:

$$Knowledge_i \cap Knowledge_j = Deviation_{ij} \quad (5)$$

The LH5 model is calculating this deviation by measuring the degree of 'coverage' between one doctor's knowledge and the other doctor's data, i.e. the likelihood that one doctor would diagnose the other doctor's patients' symptoms as caused by the same illness. In cases of a complete match the deviation should be null:

$$Knowledge_i \cap Knowledge_j = Deviation_{ij} = null \quad (6)$$

Due to the fact that each doctor will see only a subset of all possible combinations of symptoms it is highly unlikely that for any two doctors (6) would be true. However the deviation between any two doctors should vary within a range of acceptable deviation without exceptions, allowing the creation of 'clusters of medical opinion'. LH5 uses this property to identify the clusters and, in turn, uses this to isolate doctors who deviate from the generally accepted opinions by more than what it finds to be the normal within the training data set.

In the experiment performed, a separate application was created generating 100 random patient records allocated to any one of 6 different doctors. The application then proceeded to generate three training sets for LH5, selecting at random one doctor and converting the diagnoses to misdiagnoses for 75%, 50% and 25% of his cases respectively. LH5 has consistently detected the erroneous source and removed it from the data set in all cases, using for its decision no other knowledge except the data set that was analysed (see figure 5).

The LH5 System

The system used (see figure 5) has been developed in the last 5 years around the LH5 model and has been tested with large data sets involving more than 2000 facts. The system has been originally developed in LPA Prolog and then redesigned in SWI-Prolog as a background knowledge base handling tool, involving about 5000 lines of Prolog rules, and interfacing with users through an independent, Windows-based RAD front-end, using OLE.

The learning in the core of LH5 is done by the Hydra system [2], a substantially enhanced version of Quinlan's ID3 algorithm [7] specifically customised for the needs of LH5 to allow for the handling of variances in representations of data between different sources, including linguistic issues such as synonyms. Following the central concepts of ID3 the system produces a tree of rules extracted from the given data set. As with ID3, the system can be used for unsupervised learning, classifying data belonging to multiple categories producing rules that are always consistent with the data set and can be used to prove any piece of data within the data set. Nonetheless, the LH5 model is independent from the learning process applied; any learning algorithm can be used in the place of Hydra, including genetic algorithms or neural nets.

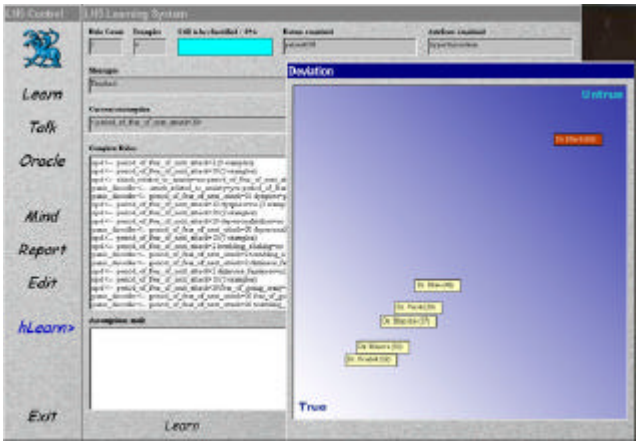

FIGURE 5

Given a set of data D provided by multiple sources S the implementation of the LH5 model's process for clearing the data is as follows:

i. Collect all data in D_{raw}
ii. harmonise representations in D_{raw} producing D
iii. extract sources from D producing set of sources S
iv. for each source s in S
v. if $D(s)$ is inconsistent with itself remove s from S giving S' and $D(s)$ from D giving D'
vi. if remaining S' contains only one source then
vii. apply Hydra on D' producing rules R and stop
viii. else
ix. for each s in S'
x. remove s from S' giving S''
xi. apply Hydra on $D(s)$ giving rules $R(s)$
xii. for each source s_2 in S''
xiii. calculate deviation(s,s_2) comparing $D(s_2)$, with $R(s)$
xiv. calculate totaldeviation(s) from deviation(s,s_i) for all s_i in S''
xv. calculate threshold deviation from all totaldeviation(s_i) for all s_i in S'
xvi. for each s in S'
xvii. if totaldeviation(s) >threshold remove s from S' producing $S'_{cleared}$ and remove $D(s)$ from D' producing $D'_{cleared}$
xviii. apply Hydra on $D'_{cleared}$ giving $R_{cleared}$

In the cases where all sources are found to be trustworthy the knowledge mined from the data set ($R_{cleared}$) will be the same with the knowledge that would have been mined by applying a traditional algorithm. In the case however that the sources are likely to be untrustworthy $R_{cleared}$ is going to be a cleaner set of rules to base decisions on than what would otherwise have been mined.

CONCLUSION

The results have shown promise for the LH5 model for the process of removing from a data mine possibly suspect sources and producing a clearer data set without using any external supervision. As data mining in E-Business starts looking at larger and larger sets of data that is collected with reduced rigour in quality it is evident that there are numerous applications of the LH5 model. By detecting

which sources to learn from and which sources to 'quarantine' with their data out of the data mine, LH5 can lead to better data mining results and, in turn, better decision making.

REFERENCES

[1] C. Ford, *Lies! Lies!! Lies!!! : The Psychology of Deceit*, American Psychiatric Press, 1996

[2] D. Parapadakis *The HYDRA system: A Machine Learning System for Multiple, Impure Sources*, CIMCA '99, Vienna, Feb 1999. Published in Computation Intelligence for Modelling, Control and Automation, IOS Press 1999 pp183-188

[3] D. M. Komp, *Anatomy of a Lie*, Zondervan Publishing House, 1998

[4] E. MacLean, *Between the Lines: How to Detect Bias and Propaganda in the News and Everyday Life*, Black Rose Books, 1981

[5] J. Holt, *How Children Learn*, Penguin Books, 1983

[6] D. Partridge, Databases that Learn, *Machine Learning Principles and Techniques*, ed. R. Forsyth, Chapman and Hall, 1989

[7] A. Shapiro, Structure Induction in Expert Systems, Addison-Wesley, 1987