Knowledge Discovery in Databases: Application to Chromatography

C.H.Bryant[12], R.C.Rowe[3]

[1]School of Computing and Mathematics, The University of Huddersfield, HD1 3DH, United Kingdom.

[2]Address correspondence to C.H.Bryant, Department of Computer Science, University of York, Heslington, York, YO1 5DD.

[3]Zeneca Pharmaceuticals, Alderley Park, Macclesfield, Cheshire, SK10 2NA, United Kingdom.

Abstract

This paper reviews emerging computer techniques for discovering knowledge from databases and their application to various sets of separation data. The data-sets include the separation of a diverse range of analytes using either liquid, gas or ion chromatography. The main conclusion is that the new techniques should help to close the gap between the rate at which chromatographic data is gathered and stored electronically and the rate at which it can be analysed and understood.

1 Introduction

The spread of laboratory automation and growth in the use of chemical databases has dramatically increased the amount of chromatographic data which is available electronically. The complexity, terse nature or sheer volume of such data can make it difficult to discover patterns, trends or relationships hidden within it which may be important for scientific or commercial reasons. Further, the software which is currently used to acquire and manage chromatographic data is not capable of discovering such knowledge. This is an example of one of the ominous problems of the age of digital information, namely data overload. The ability of humans to analyse and understand large data-sets lags behind their ability to gather and store that data. This paper reviews emerging computer techniques for discovering knowledge hidden in data and describes how they have been applied to chromatography.

2 Knowledge Discovery in Databases

The process of using these techniques has become known as Knowledge Discovery in Databases (KDD). Eminent researchers in the KDD field have defined the process [1] as:

> The non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data.

In this definition *data* comprises a set of facts and *pattern* is an expression in some language describing a subset of the data. The term *process* implies that there are many steps. The discovered patterns are *valid* for new data with some degree of certainty and are *novel* (at least to the system and preferably to the user). The phrase *ultimately understandable* implies that the patterns should be understandable, although this may require some post processing.

A wide range of computer techniques are used in KDD. These originate from statistics, Pattern Recognition, Databases, Data Visualisation and branches of Artificial Intelligence such as Machine Learning, Machine Discovery and Knowledge Acquisition for expert systems. This paper focuses on the use of Artificial Intelligence for chromatography. During

the 1980s this was mainly confined to the manual development of expert systems [2]. Since then however Machine Learning has been used to discover knowledge from chromatographic data.

Machine Learning is the study and computer modelling of learning processes. It is concerned with understanding the process of learning and providing computers with the ability to learn. Research on the provision of learning abilities conducted over the last twenty years has resulted in techniques which are now being utilised for KDD.

The classification of learning strategies shown in Table 1 allows Machine Learning techniques to be compared in terms of the types of external information that they use and their strategies and methods. The inference capabilities of machine learning systems vary. No inference is needed in rote learning as the environment provides information exactly at the level needed to perform the task. In learning from instruction the information provided by the environment is general or abstract and the learning system must perform some inference to fill in the details. The deductive and inductive learning strategies must be capable of performing their particular modes of inference but they place a smaller burden on the external environment than the strategies mentioned above. Analogical strategies require both inductive and deductive capabilities: finding common substructure involves induction whereas performing analogical mapping is a form of deduction.

There are two types of machine induction: supervised and unsupervised. In supervised learning, or learning from examples, classes are defined before induction begins and the learning system is given examples of each class. The system uses induction to find a description for each class from the examples. In unsupervised learning the classes are not predefined. Instead classes must be discovered and descriptions found for them.

The next section describes each of the techniques shown in Table 1.

3 KDD Techniques

This section describes some Machine Learning techniques which have been used to discover knowledge from chromatographic data. A large number of tools for Machine Learning are commercially available. Table 2 shows a selection of multi-paradigm Machine Learning tools. In addition there are single paradigm tools; some of these are listed later in this

section.

There is often a mismatch between the input requirements of a KDD tool and the representation of data in a particular database. Such a mismatch requires that the data is transformed before it is input to the tool. This transformation is referred to as *preprocessing* because it is performed prior to induction. Preprocessing can be very time consuming e.g. see Section 4.4. Two examples of preprocessing that are often required for a chromatographic database are 1) selecting a subset of the original set of attributes and 2) merging values of attributes.

3.1 Decision-Tree Induction

Decision-trees are a formalism for representing knowledge of how to classify examples, where each example comprises attributes, the values of those attributes and a classification (or decision). The leaf nodes of a decision-tree represent the classes and the internal nodes (branches) represent questions concerning the values of the attributes.

A family of algorithms has been developed for generating decision-trees which is known as the Top-Down-Induction-of-Decision-Tree (TDIDT) family. The most influential member of this family is ID3. ID3 is available as part of several commercially available tools (see Table 3).

3.2 Inductive Logic Programming

Inductive Logic Programming (ILP) [3] is an active area of research in computer science which has given rise to a number of general purpose tools that can be applied to chemistry. One of these, CProgol, has now been licensed to several companies (including Smith-Kline Beecham) by Oxford University [4]. ILP has been defined as the intersection between machine induction and Logic Programming [5].

The most widely used language in Logic Programming is Prolog (**Pro**gramming in **log**ic) [6]. Most ILP systems use a subset of Prolog as the representational formalism for both hypotheses and observations. In doing so ILP overcomes two of the main limitations of Machine Learning techniques such as the TDIDT family:–

1. The use of a limited knowledge representation formalism.

2. The difficulty in using substantial background knowledge in the learning process.

The greater representative power provided by Prolog allows ILP to induce rules which express relationships that cannot be represented by decision-trees. For example, rules can be induced that reason not only about properties of observations but also about the relationship between those observations, where an observation corresponds to a leaf in decision tree.

The second limitation is also important because one of the well-established findings of Artificial Intelligence is that the use of domain knowledge is essential for achieving intelligent behaviour. Logic offers an elegant formalism to represent knowledge and hence incorporate it in the induction task. ILP offers the opportunity to use both specialist knowledge on particular problems in chemistry and general chemical knowledge during induction. General knowledge refers to knowledge which is common-place amongst chemists.

3.3 Case Based Reasoning

Case Based Reasoning (CBR) solves a current problem by seeking a similar case from the past and then adapting the solution to that previous case so that it may be applied to the current problem. Several tools for CBR are commercially available (see Table 4). CBR requires that a case-structure be defined which is capable of describing the relevant features of cases. A collection of cases represented in this way is known as a case-base. CBR works as follows.

1. Previous cases which partly match the current case are retrieved from the case-base. The retrieval process consists of two steps: recalling previous cases and selecting a best subset of these recalled cases. During the recall step salient features of the current case are used as indices to recall cases which have been labelled by those same features or by other features which can be derived from them. The set of cases recalled is reduced to a small number (possibly only one) of the most relevant cases for subsequent consideration.

2. An approximate solution is then proposed by taking the relevant parts of the solution(s) to the subset of cases identified in the previous step.

3. The approximate solution is then adapted so that it is more suited to the current problem.

4. The adapted solution is then evaluated in the real world. Attempts are made to identify the causes behind any failings it may have: such feedback can be used in subsequent reasoning.

5. The new solution, together with the problem and any associated useful reasoning, is stored in the case-base. It is indexed so that it can be used for subsequent cases.

A CBR system learns as a result of its reasoning activity because it becomes more efficient and more competent by storing its learning experiences. Thus a CBR system for chromatography will not only retrieve (and when necessary adapt) previous separations but will also improve its performance over time as new separations are added to the case-base.

3.4 (Artificial) Neural Networks

Neural Networks emulate the learning behaviour of living nerve cells in animal physiology. The basic processing unit is the neuron which takes one or more inputs and produces an output. Each input to a neuron has an associated weight which modifies the strength of that input. The neuron simply addes together all the inputs and calculates an output to be forwarded to another neuron. The number of neurons in a Neural Network can range to many thousands. The methods by which the neurons are organised are referred to as the network architectures. The most popular architecture comprises three layers of neurons in which the output of each neuron is passed to all the neurons in the next layer. Data flows in via the input layer, passes through one or more hidden layers and finally exits via the output layer. This is the so-called feed-forward network or multi-layer perceptron. In theory any number of neurons can be connected in any number of layers. In practice, however, there are limitations.

Unlike conventional computer programs which have to be explicitly programmed, Neural Networks are trained with previous examples. During the training process the values

of the weights at each neuron are adjusted to bring the output of the network closer to the desired output. The method used to adjust the weights is known as the training algorithm. There are a number of these algorithms in use, the most common being the back propagation of errors. Training can be a very time consuming process. However after training is completed Neural Networks operate quickly on new examples.

Further details of Neural Networks can be found in [7]. A number of industrial Neural Network packages are commercially available (see Table 5).

4 Applications of KDD to Chromatography

This section describes how the techniques described above have been applied to various types of chromatography.

4.1 HPLC

4.1.1 Enantioseparations

KDD techniques have been used in an attempt to automatically acquire the knowledge needed to select a chiral stationary phase (CSP) for an enantioseparation by HPLC.

Three KDD tools were applied to data from a database of published enantioseparations performed on commercially available Pirkle-type (i.e. the *brush* or *multiple interaction* type) CSPs [9]. The aim was to induce rules that recommend particular CSP chiral selectors based on the structural features of an enantiomer pair. Two of the tools that were used, Golem and Progol, are from the field of Inductive Logic Programming (see Section 3.2). The other, DataMariner [10], is a commercially available tool whose learning algorithm has similarities to that of ID3 (see Section 3.1). The application of each of the tools is described, in turn, below.

DataMariner induced a set of rules which had a high degree of accuracy [11]. A cross-validation performed on it suggested that it would recommend as its first choice a correct CSP chiral selector for 63% of enantiomer pairs that can be separated on Pirkle-type CSPs. This is more than ten times greater than the accuracy that would

have resulted from choosing one of the selectors at random. Another validation, which used test data that had not ben input to DataMariner , supported this result and suggested that either the first or second choice recommendation of the optimal rule-set would be correct for 79% of enantiomer pairs that can be separated on Pirkle-type CSPs.

Golem was used to generate rules from published data on the attempted separation of a series of 3-substituted phthalide enantiomer pairs on (R)-N-(3,5-dinitrobenzoyl)-phenylglycine [12]. These rules predict, with a high accuracy (82%), which enantiomer pairs in the data-set can be separated on this CSP chiral selector. The rules are justified in that they reflect some of the findings of the analysts who performed the separations.

Progol was applied to published data on 197 attempted separations on seven commercially available Pirkle-type CSP chiral selectors [13]. Progol induced a set of rules for each of five of these selectors. All of these rules are very concise which facilitates both their interpretation and the comprehension of their coverage. The two sets of rules which were induced for the two CSPs occurring most frequently in the data, reflect advice given by a commercial supplier of CSPs [14] . The training accuracy and test accuracy for the union of these two data-sets are 93% and 61% respectively.

The results suggest that the application of ILP to enantioseparations may prove fruitful and that this line of research should be pursued further.

CHIRULE [15] was a case-based reasoning (see Section 3.3) system which used similarity searching on molecular properties to retrieve a list of enantiomer pairs that were chemically similar to a given enantiomer pair, together with CSPs that have been reported in the literature as having successfully separated them. CHIRULE's original case base comprised data on successful separations involving the 14 CSPs described in a Daicel application guide [16]. Each of these CSPs was of one of the following types: Chiral Ligand Exchange Chromatography (CLEC), Crown Ethers, and Natural and Synthetic Polymers. CHIRULE was validated in three ways.

Leave one out cross-validation Each enantiomer pair in the case base of CHIRULE

was posed, in turn, as the one to be separated. The accuracy figures that were reported were high. The first choice of CSP recommended by CHIRULE had successfully been used by Daicel in 79% of cases. Either the first or second choice had been used in 88% of cases.

Comparison with a Separations Scientist The first choice recommendation of CHIRULE was compared with the first choice of a separations scientist for each of four enantiomer pairs that were not stored in the case base of CHIRULE. CHIRULE agreed with the scientist in three of the four cases.

Use of CHIRBASE as the case-base Further testing was performed for those enantiomer pairs where the first choice of CSP recommended by CHIRULE had not been used by Daicel. For each such case a database called CHIRBASE [17] was used to determine whether a separation involving the enantiomer pair and the CSP recommended by CHIRULE had been recorded in the literature. These tests showed that in 88% of cases the first choice of CSP recommended by CHIRULE had been used by Daicel or in another separation stored in CHIRBASE. Either the first or second choice had been used in 91% of cases.

The validation proved that CHIRULE can recommend with a high accuracy a suitable CSP chiral selector for those enantiomers where chiral recognition is achievable using either the CLEC, Crown ether or Polymer types of CSPs.

4.1.2 Peak-Shape Classification

Peak-shape distortion reduces the accuracy and precision of HPLC methods. When this problem arises it can be rectified. However the on-set of this problem may not always be apparent because the change in the peak-shape may be subtle. Neural Networks have been generated that classify peak-shapes [18]. After training, the performance of an optimised Neural Network was compared to that of a human expert by presenting both with 396 individual peak profiles. Although both exhibited an overall success rate of 85%, the Neural Network performed the task in 5.6s where as the expert took 8 hours.

4.2 GC

Two projects have applied Decision-tree induction (see Section 3.1) to the problem of classifying organic pollutants given their GC-MS data. Both describe the use of commercially available tools that incorporate induction algorithms based on ID3. Scott [20] successfully used 1^{st}-Class to induce classification and identification decision trees. Derde *et al.* [19] used Ex-Tran for a classification problem.

4.3 Thin-Layer Chromatography

Ex-Tran has also been applied to a Thin-Layer chromatography data-set [21]. A decision tree was generated which predicated the retention time of 22 substituted benzoic acids with a high accuracy. The data-set comprised the retention time and the values of 12 physico-chemical properties for each derivative.

4.4 Ion Chromatography

Mulholland *et al.* [22] used the C4.5 algorithm, an extension of ID3, to induce a decision tree for choosing a detector when performing ion interaction chromatography. The decision tree was validated in two ways. Firstly a similar tree was generated using only 90% of the data for training and this tree was tested using the other 10% of the data. Secondly by using another test-set which was provided by a domain-expert and comprised 52 pertinent examples of the ideal choice of detector, as selected by that expert. The validation showed that 70% of the recommendations made by the decision tree were an exact match with the published methods and a further 22% were acceptable to the domain expert in that s/he thought that they would perform well for the given separation.

The data used by Mulholland *et al.* originated from a database of published methods for ion chromatography. The database contained information on almost 4000 applications, including most of the chromatographic conditions employed. Part of this data was input to the C4.5 algorithm after being preprocessed. Mulholland *et al.* reported that this preprocessing was the most time consuming part of the work.

Recently further results of this work have been published [23] in which another machine

induction tool was applied to *all* of the data in the database. The validation of the resulting rules showed that over 85% of the methods recommended by the rules worked and almost 62% of them were considered ideal.

5 Conclusion

The field of Machine Learning has now matured to the extent that a wide range of its techniques have become accessible to industry because they have been implemented as commercially available tools. These techniques have been used to discover knowledge from various data-sets which cover the separation of a diverse range of analytes using several types of chromatography. The classification accuracies for the applications reviewed are all high. This suggests that the use of Machine Learning techniques should help to close the gap between the rate at which chromatographic data is gathered and stored electronically and the rate at which it can be analysed and understood.

References

[1] V.Fayyad, G.Piatetsky-Shapiro and P.Smyth, *Communications of the ACM*, 39(11), 27-34 (November 1996).

[2] C.H. Bryant, A.E. Adam, D.R. Taylor and R.C. Rowe, *Analytica Chimica Acta*, 297, 317-347 (1994).

[3] N.Lavrac and S.Dzeroski, *Inductive Logic Programming: Techniques and Applications*, Ellis Horwood, (1994).

[4] Personal Communication with S.Muggleton, Oxford University Computing Laboratory, Wolfson Building, Parks Road, Oxford, OX1 3QD, United Kingdom. (1996).

[5] J.W.Lloyd, *Foundations of Logic Programming*, Springer-Verlag, Berlin, (1984).

[6] W.F. Clocksin and C.S. Mellish, *Programming in Prolog*, 1^{st} Ed, Springer-Verlag, (1981).

[7] J.R.M. Smits, W.J. Melssen, L.M.C. Buydens and G. Kateman, *Chemometrics and Intelligent Laboratory Systems*, 22, 165-189 (1994).

[8] C.B.Lucasius and G.Kateman, *Trends in Analytical Chemistry*, 10(8), 254-261 (1991).

[9] C.H. Bryant, *Data Mining for Chemistry: the Application of Three Machine Induction Tools to a Database of Enantioseparations*. Ph.D. Thesis. University of Manchester Institute of Science and Technology, UK. (1996).

[10] Logica UK Ltd., Cambridge, UK.

[11] C.H. Bryant, A.E.Adam, D.R. Taylor and R.C. Rowe, *Chemometrics and Intelligent Laboratory Systems*, 34, 21-40 (1996).

[12] C.H. Bryant, A.E.Adam, D.R. Taylor and R.C. Rowe, *Chemometrics and Intelligent Laboratory Systems*, 36, 111-123 (1997)

[13] C.H. Bryant, *Accepted for publication in the Springer Verlag Proceedings of The Seventh International Workshop on Inductive Logic Programming, Prague, 17-20 September 1997.*

[14] Phenomenex Ltd. UK, Macclesfield, Cheshire, United Kingdom.

[15] S.T. Stauffer. *Expert System Shells in Chemistry: CHIRULE, a Chiral Chromatographic Column Selection System Using Similarity Searching and Personal Construct Theory.* PhD Thesis. Virginia Polytechnic Institute State University, USA. (1993)

[16] Daicel Chemical Industries Ltd., 8-1, Kasumigaseki 3-chome, Chigoda-ku, Tokyo 100, Japan. (1989)

[17] B.Koppenhoefer, R.Graf, H.Holzschuh, A.Nothdurft and U.Trettin, *Journal of Chromatography*, 666, 557-563 (1994).

[18] R.C.Rowe, V.J.Mulley, J.C.Hughes, I.T.Nabney and R.M.Debenham *LC-GC*, 7(1), 36-42 (1994) .

[19] M.Derde, L.Buydens, C.Guns and D.L.Massart, *Analytical Chemistry*, 59, 1868-1871, (1987)

[20] D.R.Scott, *Analytica Chimica Acta*, 223, 105-121 (1989).

[21] M.A-Razzak and R.C.Glen, *in H.Van De Waterbeemd,* Advanced Computer Assisted Techniques in Drug Discovery*, Weinheim:VCH, (1995).*

[22] M.Mulholland, D.B.Hibbert, P.R.Haddad and C.Sammut, *Chemometrics and Intelligent Laboratory Systems*, 27, 95-104 (1995).

[23] M.Mulholland, P.Preston, D.B.Hibbert, P.R.Haddad and P.Compton, *Journal of Chromatography*, 739, 15-24 (1996).

Strategies	Methods	Computer Techniques[a]
Rote	Learning by being programmed	
	Learning by memorisation	
Instruction	Learning by being told	
Deductive	Learning by deriving proofs	
Inductive	Learning from examples	Decision-tree Induction Inductive Logic Programming Artificial Neural Networks
	Learning from observation and discovery	Artificial Neural Networks
Analogical	Learning by making analogies	Case Based Reasoning

[a]Only computer techniques described in this paper are listed in this table.

Table 1: Classification of Learning by Underlying Strategies, Methods and Techniques.

Tool	Supplier	Address
Clementine	Integral Solutions Ltd.	Basingstoke, UK.
KATE	AcknoSoft	Paris, France.
RECALL	ISoft SA.	Gif sur Yvette, France.
Information Havester	Information Havesting Inc.	Mt. Kisto, New York, USA.
IDIS	Information Discovery Inc.	Hermosa Beach, California, USA.
Knowledge Seeker	Angross Software International Ltd.	Ontario, Canada.

Table 2: A Selection of Multi-Paradigm Data Mining Tools.

Tool	Supplier	Address
Crystal	Intelligent Environments Ltd.	Sunbury-on-Thames, UK.
Insight 2+	Level Five Research	Indialantic, FL, USA.
1^{st} Class	Trinzic Corp.	Redwood, CA, USA.
Ex-Tran	Intelligent Terminals Ltd.	San Francisco, California, USA.

Table 3: A Selection of Commercially Available Tools which can Generate Decision-trees.

Tool	Supplier	Address
CBR2	Inference Ltd.	Slough, UK.
ESTEEM	Esteem Software Inc.	Cambridge City, IN, USA.
CASECRAFT	AcknoSoft	Palo Alto, CA, USA.
RECALL	ISoft SA.	Gif sur Yvette, France.
REMIND	Cognitive Systems Ltd.	Boston, MA, USA.
	Intelligent Applications Ltd.	Livingston, Scotland, UK.
S_3-CASE	techno GmbH	Kaiserlautern, Germany.

Table 4: A Selection of Commercially Available Case-Based Reasoning Tools.

Tool	Supplier	Address
Neural Works Professional	Neural Ware Inc.	Pittsburg, PA, USA.
Neuro Shell Neuro Windows	Ward Systems Inc	Frederick, M.D. USA.
Genesis	Neural Systems Inc.	Vancouver, Canada.
Neural Desk	Neural Computer Sciences	Southampton, UK.
BrainMaker Professional	California Scientific Software	Grass Valley, California, USA.

Table 5: A Selection of Commercially Available Neural Network Tools.