Analysing chromatographic data using data mining to monitor petroleum content in water

Geoffrey Holmes, Dale Fletcher, Peter Reutemann and Eibe Frank

Computer Science Department, University of Waikato, New Zealand.

Abstract

Chromatography is an important analytical technique that has widespread use in environmental applications. A typical application is the monitoring of water samples to determine if they contain petroleum. These tests are mandated in many countries to enable environmental agencies to determine if tanks used to store petrol are leaking into local water systems.

Chromatographic techniques, typically using gas or liquid chromatography coupled with mass spectrometry, allow an analyst to detect a vast array of compounds—potentially in the order of thousands. Accurate analysis relies heavily on the skills of a limited pool of experienced analysts utilising semi-automatic techniques to analyse these datasets—making the outcomes subjective.

The focus of current laboratory data analysis systems has been on refinements of existing approaches. The work described here represents a paradigm shift achieved through applying data mining techniques to tackle the problem. These techniques are compelling because the efficacy of preprocessing methods, which are essential in this application area, can be objectively evaluated. This paper presents preliminary results using a data mining framework to predict the concentrations of petroleum compounds in water samples. Experiments demonstrate that the framework can be used to produce models of sufficient accuracy—measured in terms of root

mean squared error and correlation coefficients—to offer the potential for significantly reducing the time spent by analysts on this task.

Keywords: Gas Chromatography Mass Spectrometry, GC-MS, BTEX, Data Mining, Model Trees, Regression, Data Preprocessing, Correlation Optimized Warping, Petroleum Monitoring.

1 Introduction

Chromatography is an important analytical technique that has widespread use in environmental applications (Christensen and Tomasi 2007; Hupp et al. 2008; Pérez Pavón et al. 2004). A typical application is the monitoring of water samples to determine if they contain petroleum. These tests are mandated in many countries to enable environmental agencies to determine if petroleum tanks are leaking into local water systems. In this paper we outline a data mining approach to analysing chromatographic data and apply it to the detection of toluene, a compound found in petroleum.

Chromatographic techniques, typically using gas or liquid chromatography coupled with mass spectrometry, allow an analyst to detect a vast array of compounds—potentially in the order of thousands. Accurate analysis relies heavily on the skills of a limited pool of experienced analysts utilising semi-automatic techniques, making the outcomes subjective.

The focus of current laboratory data analysis systems has been on refinements of existing approaches. The work described here represents a paradigm shift achieved through applying data mining techniques to tackle the problem. These techniques are compelling because data mining offers an experimental framework where the efficacy of methods can be evaluated and compared systematically. This is especially important in chromatography because many data pre-processing techniques are needed for this type of data.

This need for a paradigm shift has been seen by others in related domains such as proteomics and metabolomics (Taylor and King 2002), where data mining has been used for pattern discovery purposes in an attempt to answer fundamental research questions. It is often the case however, that due to the effort required to obtain data, the datasets contain only

a small number of samples with a vast number of compounds, a combination where spurious patterns are highly likely to be uncovered. This becomes particularly problematic when no degree of certainty is associated with the predictions made from those patterns.

Environmental data differs from this. The number of compounds is much smaller and the number of samples is much larger. In addition, historic data is available to provide a rich source for experimentation and verification. Nevertheless it would be very useful to be able to quantify the uncertainty associated with individual predicted measurements obtained from a data mining tool. There is some emerging work in this area (Nouretdinov et al. 2001) but it remains an open area for research. Consequently, we focus on average measurement error in this paper.

Currently, chromatograms are processed semi-automatically in two stages. First, a computer program processes the data and then an analyst, looking at all the data, manually corrects mistakes. If the use of data mining techniques could enable samples to be processed at levels that are at least as accurate as the current system and uncertainty measurements were possible, then they could be used to isolate those aspects that require manual attention. This methodology would lead to significant gains in productivity because the expectation is that the *bulk* of the task can be handled by the data mining framework.

Much of the processing of a chromatogram is dedicated to pre-processing. Once again, data mining evaluation methodologies can be used to determine the efficacy of each pre-processing method at each stage by observing the effect on average measurement error. Data management is an essential component of the framework as data mining techniques all learn from experience and the datasets are large and cumbersome. Once a large collection of accurately labelled data—produced by trained analysts—is available for mining, then existing data mining techniques can be applied to establish a state-of-the-art for the framework.

Before describing the software tool that we have implemented to manage the data pre-processing and presenting some results for petroleum detection, we introduce the basic concepts behind this application.

1.1 Gas Chromatography Mass Spectrometry (GC-MS)

In *gas chromatography* a sample is injected into a heated *column*—for example, a long glass capillary tube. Due to the different chemical properties in the sample, the *time of flight* of the sample passing through the column is different (and known) for different compounds. The time taken by a compound to pass through the column is called the *retention time*, and compounds are said to *elute* from the column.

At the end of the column a detector, in this case a *mass spectrometer* (mass-spec), ionizes, accelerates, deflects and detects the separated ionized compounds. The important action here is the detection of molecular fragments using their mass to charge ratio.

Thus, both units work to produce a fine-grained identification of the components of the sample. Having a mass spectrometer after gas chromatography is essential for some samples as they may contain compounds that have the same retention time: two or more compounds may co-elute from the column and the mass-spec must be used to differentiate them. In this context, a *chromatogram* is essentially a sequence of mass-spec scans over time. It is often viewed as a two-dimensional plot of the total ion count of a scan against time (see, for example, the top panel of Figure 1) alongside a mass-spec plot at a given retention time (bottom panel of Figure 1).

2 Data Mining Framework for Chromatographic Data

In this section we outline the integrated environment we have developed to convert a chromatogram into a form suitable for machine learning techniques for data mining. The end product of this process is the production of data *instances*. In the section that follows we describe a learning experiment involving instances that have been prepared using the tool. The application is the monitoring of petroleum in water.

Chromatographic data can be extremely noisy and must undergo a number of transformations in order to make it amenable to analysis by data mining algorithms. Outlining all of the possible pre-processing steps is beyond the scope of this article, so we restrict attention to some of the more important steps. The goal of pre-processing is to produce learning instances that can be used to train a data mining model for future use in predicting the concentration of a compound in an unlabelled chromatogram.

We have developed an experimental framework which allows users to visualise, pre-process, and generate instances for data mining. Due to the typical size of chromatographic data the framework is designed to take advantage of multi-core machine architectures. It utilises a workflow interface called Kepler (see, www.kepler-project.org) that enables developers to visualise the total pre-processing workflow from raw data to instances. Having the ability to generate instances after each pre-processing step provides the opportunity to use data mining to objectively evaluate the efficacy of each step. This is done by building a predictive model from the generated instances and computing the measurement error.

Chromatographic data is complicated by a number of factors including noise, sample impurities and instrument maintenance. These factors can lead to two consecutive samples producing significantly different chromatograms. This type of noise is particularly difficult to handle with data mining algorithms because two such samples should produce similar instances. In particular, retention time shifts are a major problem.

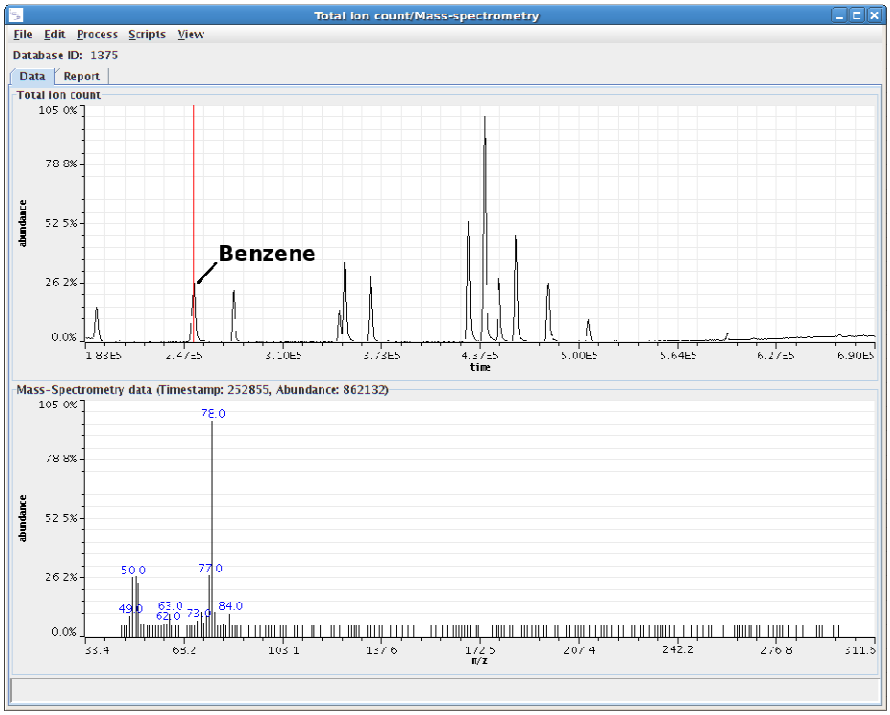

Fig. 1. Single Chromatogram indicating Benzene.

Figure 1 contains a three-dimensional view of a single chromatogram as provided by our tool. The time dimension (x-axis of the top panel) represents retention time. Successive mass-spec scans are performed and these are summarized into total ion count in the top panel. The detail of a mass-spec scan at the peak corresponding to the expected retention time of *Benzene* is shown in the bottom panel. The area under the peak is the basis for calculating the concentration of the compound in the sample. In the application described here, the purpose of the analysis is to extract the concentrations of six specific compounds found in petroleum, namely: *Benzene*, *Toluene*, *Ethylbenzene*, *O-xylene* and *M&P-xylene*.

We are able to label the plot for a given compound like *Benzene* because we know its expected retention time (from laboratory standards) and we know its expected mass-spec fingerprint from, for example, the National Institute of Standards and Technology library (see www.nvl.nist.gov). The bottom panel in Figure 1 depicts the mass-spec scan at the peak and this matches the library fingerprint for *Benzene*.

The goal of pre-processing in the framework is to produce a data instance that represents a single chromatogram. This is complex because GC-MS instrumentation is susceptible to various forms of noise. Two significant problems are particularly difficult for data mining applications: retention time shifts and co-elution. Retention time shifts make it difficult to standardize chromatograms—in an attribute-value context as applied in data mining, the values of attributes should have the same meaning across instances. Co-elution, the emergence of two or more compounds at the same retention time, creates superimposed peaks causing difficulties for the data mining algorithms. Figure 2 depicts multiple superimposed chromatograms and clearly shows the problem of retention time shifts.

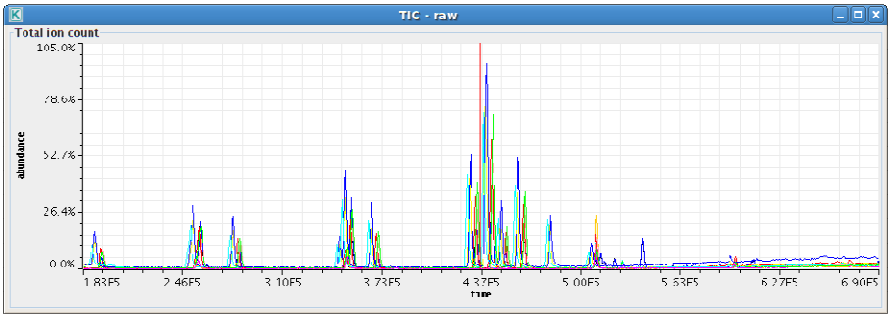

Fig. 2. Several unaligned chromatograms

There are many possible alignment methods in the literature (Johnson et al. 2003; Nederkassel et al. 2006; Pravdova et al. 2002), none of which are ideal for all applications. Consequently, our framework has been designed to allow experimentation with different methods, each of which can be evaluated via a data mining experiment.

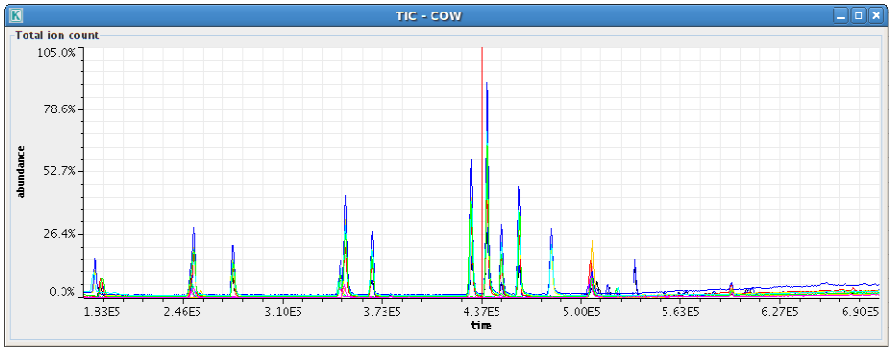

Fig. 3. Chromatograms aligned by correlation optimised warping.

One alignment method uses correlation optimised warping (COW) to align a chromatogram with a reference (Nielsen et al. 1998). The method shrinks or stretches a window to maximise the correlation coefficient with the reference chromatogram. The problematic retention time shifts depicted in Figure 2 are aligned in Figure 3 and the busy region around 4.37 is now clearly differentiated. Methods like this contain parameters such as *window size* and *maximum warp*. Traditionally, such parameters have been estimated by trial and error. Our framework supports experimentation employing data mining to determine optimal parameter settings.

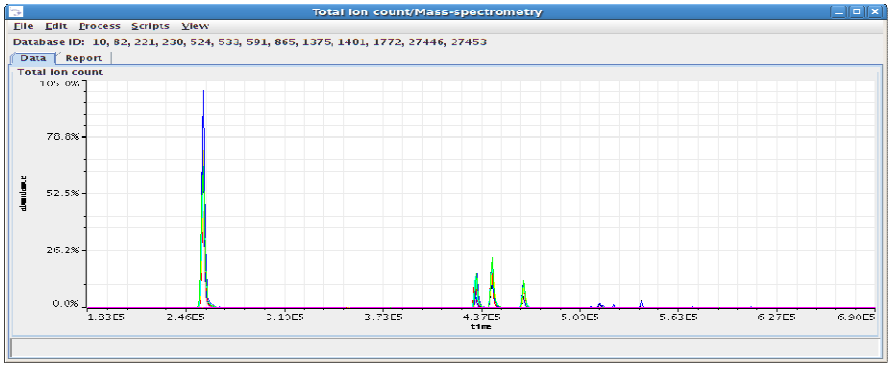

Fig. 4. Chromatograms normalised to Benzene.

The co-elution problem can be handled either by pre-processing or by including the mass-spec data in a data instance and letting the data mining algorithm distinguish co-eluting compounds. The example in Figure 4 uses a pre-processing technique based on mass-spec compound normalisation. This method adjusts mass-spec ratios to match those of a target compound. Figure 4 depicts a normalisation to *Benzene*. As can be seen, all chromatograms are aligned by COW and most unrelated mass-spec data have been removed by normalization so that clean data instances can now be created for data mining.

The final step in the pre-processing stage is instance generation, which is not entirely straightforward. All instances are required to have the same number of data points. Thus, a three-dimensional chromatogram must be transformed to a two-dimensional attribute-value format where each attribute has the same meaning. Retention time alignment is necessary, but if mass-spec data is included in the instance then additionally some form of standardisation for the mass-spec is also required. This is because the mass-spec data often have differing numbers of ion counts.

This series of examples is just one combination of the many pre-processing steps that can be trialled ahead of building an actual predictive data mining model for the concentration of the target compound. It is generally advisable to attempt to remove as much noise from the data as possible ahead of mining. For experimentation purposes we have implemented several algorithms for baseline correction, retention time alignment, noise removal, peak detection and instance generation.

3 Application to Monitoring Petroleum in Water

The specific application considered in this paper involves building models from chromatograms in order to predict the concentrations of six compounds of petroleum in water. Currently, these chromatograms are first analysed by commercial software that determines the concentrations of the compounds. An analyst then examines the chromatogram to ensure that the software has integrated the peaks correctly. This process is time consuming and error-prone. The data mining application is to build a model of compound concentrations from historically labelled data in order to automatically predict the concentrations of compounds in future (unseen) chromatograms.

The main goal of this work is to demonstrate that the framework can be used to produce models of sufficient accuracy—measured in terms of root mean squared error and correlation coefficients—to offer the potential for significantly reducing the time spent by analysts on this task. Further, in order to demonstrate the utility of the pre-processing methods in the experimental framework, we designed an experiment to determine the accuracy of different predictive models at various stages of pre-processing. Four attribute-value datasets, containing 200 instances from the same GC-MS instrument, were randomly selected from a large database provided by a routine testing laboratory. In each case the values were the total ion counts at a certain retention time. The target compound in each case was *Toluene*. The first dataset (RAW) contains raw chromatographic data, segmented into fixed interval timestamps to give the same number of data points. This is essentially equivalent to no pre-processing. The second dataset (COW) contains aligned data using correlation optimised warping. The third dataset (COW-NORM) contains data that has been mass-spec normalised following COW alignment. The fourth dataset (FULL) has been subjected to a set of comprehensive pre-processing techniques, including COW, NORM, chromatogram cropping to a retention time window, and the addition of internal standard data for concentration calibration. This last dataset simply demonstrates the flexibility of the framework to combine techniques.

Table 1 shows a list of the data mining regression methods used in the experiment. The instances are generated in the ARFF file format in order to utilise the WEKA data mining workbench (Witten and Frank 2005). The methods vary in complexity but were all able to produce results on the 200 instances in reasonable time. None of the methods were optimised for their parameters, which could make a significant difference to the results. The methods are basic linear regression [1], partial least squares regression [2] (a commonly used technique in chemometric applications), support vector machine regression using a radial basis function kernel [3], locally weighted learning [4] utilising partial least squares regression to construct a model from the twenty nearest instances to the test instance, and model trees [5] with linear regression models at the leaves. Method [6] is as per [5], but utilises a target value transformed via the natural logarithm, which is transformed back once a prediction has been obtained from the tree.

Table 1. Regression Methods

	Regression Method
[1]	Linear Regression
[2]	PLS Regression
[3]	SVM Regression
[4]	Locally Weighted Regression
[5]	Model Trees
[6]	Model Trees using log(1 + target)

Table 2 shows the average correlation coefficients from running ten times ten-fold cross-validation using each method on the four datasets. The values in parentheses are the average standard deviations of the correlation coefficients across the 100 runs. These measure the variation in the correlation coefficients of the models constructed by the methods. As can be seen, the methods with high correlation coefficients also have low standard deviations, indicating that good models are produced in each run.

Table 2. Correlation coefficients of data mining algorithms on each dataset

Dataset	[1]	[2]	[3]	[4]	[5]	[6]
RAW	0.684	0.882	0.844	0.915	0.664	0.768
RAW std dev	(0.306)	(0.118)	(0.101)	(0.084)	(0.236)	(0.214)
COW	0.656	0.901	0.865	0.944	0.908	0.903
COW std dev	(0.325)	(0.080)	(0.096)	(0.068)	(0.131)	(0.097)
COW+NORM	0.803	0.821	0.817	0.847	0.924	0.922
COW+NORM std dev	(0.190)	(0.161)	(0.158)	(0.124)	(0.071)	(0.069)
FULL	0.869	0.917	0.925	0.988	0.935	0.998
FULL std dev	(0.103)	(0.065)	(0.051)	(0.019)	(0.068)	(0.003)

Table 3. Root mean squared error of data mining algorithms on each dataset (the range of the target is approximately 0 to 1600)

Dataset	[1]	[2]	[3]	[4]	[5]	[6]
RAW	129.95	69.51	97.97	68.15	136.73	129.30
RAW std dev	(74.24)	(31.34)	(73.43)	(45.37)	(80.91)	(118.44)
COW	185.02	69.25	89.60	48.73	72.42	76.53
COW std dev	(170.41)	(36.97)	(64.34)	(27.85)	(74.48)	(66.02)
COW+NORM	109.61	96.68	93.28	101.39	69.49	70.51
COW+NORM std dev	(54.63)	(57.40)	(58.89)	(68.54)	(46.46)	(51.59)
FULL	80.957	63.19	79.61	21.80	59.04	13.05
FULL std dev	(30.73)	(28.98)	(63.74)	(15.88)	(39.22)	(11.17)

A high correlation coefficient is not a guarantee for a good predictive model as it is possible to have high correlation and large errors. Table 3

shows the average root mean squared (RMS) error values from running ten times ten-fold cross-validation using each method on the four datasets. The values in parentheses are the average standard deviations of the RMS values across the 100 runs. As can be seen, the methods with high correlation coefficients also have low RMS values given the range, and low variances between runs, once again indicating that good models are produced in each run.

The models that produce the best results are locally weighted learning and model trees using the log transform on the target. These results are extremely promising in terms of indicating that this problem may well be amenable to automation. Similar results using these methods can be obtained for the other four compounds although *Xylenes* and *Ethylbenzene* pose a more difficult alignment problem as their peaks have minimal separation and their mass-spec fingerprints are similar.

In terms of pre-processing, RAW is usually the worst format, and in general successive pre-processing leads to better outcomes. However, this is not always the case as can be seen with the COW+NORM combination. Here, normalisation removes the internal standard peaks that some algorithms rely on to improve accuracy.

4 Conclusion

We have presented a data mining approach to analysing data from a GC-MS instrument. These instruments are commonly used for a range of environmental analyses. The experimental framework we have developed is oriented towards producing data that can be used as input to data mining algorithms. Data mining provides the ability to objectively measure the impact of pre-processing techniques, which are essential in this domain. They also provide algorithms that can be used to make predictions on new data, thus offering the potential for automating many of these tasks.

The problem of monitoring petroleum levels in water is a test often mandated in many countries to enable environmental agencies to determine if petroleum tanks are leaking into local water systems. We have demonstrated that our framework can be used to produce models of sufficient accuracy—measured in terms of root mean squared error and correlation coefficients—to offer the potential for significantly reducing the time spent by analysts on this task. In the space of GC-MS problems, measuring

petroleum concentration in water is considered straightforward, but to our knowledge there does not currently exist a fully automatic solution.

In order to demonstrate the usefulness of our framework, it will be important to tackle new, more challenging tasks in the future. Moreover, while we have presented promising results from a cross-validation study, it is essential that the application is assessed operationally alongside the current methodology.

Acknowledgements

The authors would like to thank the Environmental GC-MS team at R. J. Hill Laboratories for their support.

References

Christensen, JH, Tomasi, G (2007). Practical aspects of chemometrics for oil spill fingerprinting. Journal of Chromatography A, Volume 1169, Issues 1-2, Pages 1-22

Hupp, AM, Marshall LJ, Campbell DI, Waddell Smith R, McGuffin VL (2008). Chemometric analysis of diesel fuel for forensic and environmental applications. Analytica Chimica Acta, Volume 606, Issue 2, Pages 159-171.

Johnson KJ, Wright BW, Jarman KH, Synovec RE (2003) High-speed peak matching algorithm for retention time alignment of gas chromatographic data for chemometric analysis, Journal of Chromatography A, Volume 996, Issues 1-2, Pages 141-155.

Nederkassel AM, Daszykowski M, Eilers PHC, Van der Heyden A (2006) A comparison of three algorithms for chromatograms alignment, Journal of Chromatography A, Volume 1118, Issue 2, Pages 199-2.

Nielsen N-P, Carstensen JM, Smedsgaard J (1998) Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping, Journal of Chromatography A, Volume 805, Issues 1-2, Pages 17-35.

Nouretdinov, I, Melluish, T and Vovk, V (2001) Ridge Regression Confidence Machine, Proceedings of the 18th International Conference on Machine Learning, USA, pp 385-392, Morgan Kaufmann.

Pérez Pavón JL, Peña AC, Pinto CG, Cordero, BM (2004) Detection of soil pollution by hydrocarbons using headspace–mass spectrometry and identification of compounds by headspace–fast gas chromatography–mass spectrometry. Journal of Chromatography A, Volume 1047, Issue 1, Pages 101-109

Pravdova V, Walczak B, Massart DL (2002) A comparison of two algorithms for warping of analytical signals, Analytica Chimica Acta, Volume 456, Issue 1, Pages 77-92.

Taylor J, King RD, Altmann T, and Fiehn O (2002) Application of metabolomics to plant genotype discrimination using statistics and machine learning. Bioinformatics, 18: S241 - S248.

Witten IH, Frank E (2005) Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Francisco, 2nd Edition.