Application of Knowledge-based Network Processing to Automated Gas Chromatography Data Interpretation

Alan P. Levis
Department of Computer Science
Florida State University
Tallahassee, Florida

Robert G. Timpany
Department of Computer Science
Florida State University
Tallahassee, Florida

Wayne E. Austad
Robotics Department
Idaho National Engineering Laboratory
Idaho Falls, Idaho

John W. Elling
Measurement Technology Group
Los Alamos National Laboratory
Los Alamos, New Mexico

Jamie J. Ferguson
Robotics Department
Idaho National Engineering Laboratory
Idaho Falls, Idaho

Douglas A. Klotter
Department of Computer Science
Florida State University
Tallahassee, Florida

Susan I. Hruska
Department of Computer Science
Florida State University
Tallahassee, Florida

RECEIVED
OCT 20 1995
OSTI

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

MASTER

ABSTRACT

A method of translating a two-way table of qualified symptom/cause relationships into a four layer Expert Network for diagnosis of machine or sample preparation failure for Gas Chromatography is presented. This method has proven to successfully capture an expert's ability to predict causes of failure in a Gas Chromatograph based on a small set of symptoms, derived from a chromatogram, in spite of poorly defined category delineations and definitions. In addition, the resulting network possesses the advantages inherent in most neural networks: the ability to function correctly in the presence of missing or uncertain inputs and the ability to improve performance through data-based training procedures.

Acquisition of knowledge from the domain experts produced a group of imprecise cause-to-symptom relationships. These are reproduced as parallel pathways composed of Symptom-Filter-Combination-Cause node chains in the network representation. Each symptom signal is passed through a Filter node to determine if the signal should be interpreted as positive or negative evidence and then modified according to the relationship established by the domain experts. The signals from several processed symptoms are then combined in the Combination node(s) for a given cause. The resulting value is passed to the Cause node and the highest valued Cause node is then selected as the most probable cause of failure.

Keywords: Laboratory Automation, Gas Chromatography, Expert System, Expert Network, Knowledge Acquisition, Knowledge Representation

1 MOTIVATION

The U.S. Department of Energy (DOE) is currently involved in the cleanup of several mixed-waste dump sites. This task involves location of the wastes, identification of the soil contaminants via laboratory analysis and then treatment of the specific contaminants thus identified. In the interest of human safety, minimal exposure to the contaminants is desired, and thus automated methods are preferred.

Laboratory analysis of the waste samples is by far the most human-intensive and financially expensive aspect of this task. Raw soil samples are typically prepared by laboratory technicians. The prepared sample is then processed through the analytical equipment. At this point, highly trained and experienced analytical staff are required to interpret the results. If some failure has occurred in either sample preparation or equipment operation, the resulting data are useless and the process must be repeated. This problem is compounded when the analytical equipment is automated to run multiple samples in succession. Should a fault occur and go undetected or uncorrected in an automated laboratory, it is often the case that all data obtained subsequent to the failure must be discarded. Thus, automated interpretation of the data remains the bottleneck in the analytical process.¹

The DOE initiated the Contaminant Analysis Automation (CAA) program to develop standards and technology which promote the integration of analytical equipment into a fully automated laboratory that meets EPA standards.² One goal of the Data Interpretation Module of the CAA program is to detect and diagnose sample preparation and equipment failure during the individual sample analysis rather than after multiple samples have been processed. With the appropriate feedback and control circuits, the fault may then be corrected and the analysis of multiple samples may proceed with minimal waste and expense.

The selected testbed application is fault detection for sample preparation and equipment failure of a Varian Gas Chromatograph (GC) equipped with an Electron Capture Detector (ECD). Single point failures are assumed.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

2 KNOWLEDGE ACQUISITION

2.1 Prior systems

Several knowledge-based fault diagnosis systems for GC analytical systems have been implemented as rule-based Expert Systems.^{3,4} These systems are typically based on True/False relationships between symptoms and causes. This requirement preempts the ability of the systems to cope with uncertain data. In addition, these systems frequently require that each cause be identified by a unique set of symptoms. Where this is not possible, ranges of severity of the symptoms are typically used to partition a single symptom into multiple symptoms and provide the required unique symptom sets for the given causes. This requirement is at odds with the goal of building systems that are able to reach conclusions in the event that some symptoms are missing or degraded.

These limitations pose a challenge to the implementation of these systems in a fully automated analytical laboratory. Additionally, the system should be flexible enough to allow for the possibility of changing analytical systems or data. One method of accomplishing this is to include the capability of data-based refinement of the system.

2.2 The knowledge base

A panel of experts in analytical chemistry was assembled and undertook the task of identifying the relationships between symptoms and causes using a natural representation. The results exemplify the problems typically associated with the process of acquiring knowledge from human experts.

Experts were first asked to identify symptoms and causes that were likely to be observed in an automated environment with the analytical equipment configuration selected for the project. This resulted in the following lists of symptoms:

Retention time change	Spike/Surrogate precision and reproducibility
Sensitivity change	Peak tailing
Unresolved peaks	Band broadening
Clipped peaks	Rounded peaks
Negative dip after peak	Baseline drift - irregular
Baseline drift - rising	Cannot zero baseline
High background noise	Irregular spikes
Ghost peaks	Extra peaks
No peaks	Replicate precision and reproducibility
Leading peaks	

and causes:

Leaking syringe	Dirt in syringe	Sample decomposition
Late elution	Column bleed	Column degradation
Contaminated column	Contaminated detector	Make-up gas
Dirt in the injector	Leaking septum	Sample too concentrated
Sample size too large	Carrier gas low	

The nature of the relationships between symptoms and causes was then addressed. The experts were most comfortable viewing the relationships as being from causes to symptoms. The system operation, on the other hand, uses symptoms to diagnose causes. Further, experts selected the semantic qualifiers: "Always", "Usually", "Sometimes", "Infrequently", and "Never", to define the relationships. An example of a typical relationship identified by the experts would be, "When the carrier gas is low, you always see a retention time change, usually see a sensitivity change, sometimes see unresolved peaks and sometimes see band broadening." This resulted in a

two-way table of symptoms and causes where entries correspond to the semantic qualifier chosen by the experts to relate a symptom to its cause.

The definition and use of these qualifiers is not inherently clear. The immediate indication is that these qualifiers refer to the frequency of observation of symptoms given causes, however, this is not strictly the case. In some cases the combination of symptoms represents completely separate or overlapping symptom sets each of which indicates a common cause. Little indication of the distinctions of definition or usage was given during the knowledge elicitation sessions. In the early phases of implementation some symptoms were identified as being mutually exclusive and yet appeared within the symptom set of a single cause. One such example is the case of the symptoms "No peaks" and "Extra peaks" both being related to the cause "Leaking syringe". It should be obvious that a chromatogram could not exhibit both a lack of peaks and extra peaks, therefore these two symptoms must participate in parallel pathways for concluding the cause of "Leaking syringe".

3 IMPLEMENTATION

Based on the imprecise nature of the semantic qualifiers and the requirement that correct conclusions be drawn even in the face of missing information, we chose to implement the knowledge-based inferencing system as an Expert Network⁵ rather than as a rule-based Expert System.

The idea of using Expert Networks and the associated connectionist training algorithms to refine this type of knowledge in an Expert System from data is well established.⁶⁻⁹ Typically, an Expert Network takes its initial architectural components directly from the rule-base of an Expert System, and then refines them using algorithms similar to neural network training algorithms. Information solicited from the front-end user becomes the input pattern vector, rule assertions become nodes, logical chains become connections, and output decisions are represented as output pattern vectors. Measures of uncertainty attached to rules in the Expert System can be used as initial values for the weights on the network connections. The network is represented as a directed acyclic graph that mirrors the topology of the initial Expert System rule base. Note that the network and its components take on much higher level functional meaning than in traditional low-level neural networks. The processing of information through the network is determined by the mechanisms of the inference engine of the Expert System. Expert Networks may therefore exploit several different types of knowledge, including rules and symptom/cause relationships solicited from the expert.

The knowledge table representation used in the diagnosis problem lends itself to a similar type of translation to network form. The inference mechanism employed by the human experts as they use the table is not clearly specified and this presented yet another challenge to automation of the diagnostic process. Details of the implementation appear in the following sections.

3.1 Architecture

The nature of the knowledge table produced during the knowledge acquisition process indicates that information is processed at two distinct inferencing levels: that of the semantic qualifier, which acts to filter inputs into categorized positive and negative evidence, and that of the combination of several pieces of categorized evidence. This suggests the implementation of a four-layer Expert Network consisting of Input, Filter, Combination, and Output nodes. Inference chains are constructed connecting all symptoms related to a given cause to that cause through appropriate Filter and Combination nodes. The result is an Expert Network connecting symptoms to their related causes.

First symptom sets are identified for each cause based on the knowledge derived from the domain experts.

Each Input node in the symptom set is then connected to a Filter node based on the semantic qualifier used by the experts to relate the symptom to the cause. These pathways are then connected to a Combination node that is in turn connected to the Cause node which is related to the symptom set.

Figure 1 shows the architecture derived from a set of four symptoms related to a single cause.

Figure 1: Simple Network Architecture

When parallel pathways are required, additional Combination nodes are added. Inputs common to all pathways are connected to all Combination nodes while inputs specific to only one path are connected to only one Combination node. All Combination nodes related to the cause are linked to the Cause node.

Figure 2 shows the architecture arising from the relationship of four symptoms to a single cause where the symptoms "No peaks" and "Extra peaks" necessitate the creation of parallel pathways.

Figure 2: Network Architecture for Parallel Pathways

3.2 Inferencing mechanism

As evidenced by the network architecture, information processing occurs in parallel through all sub-networks and in sequence along each pathway in each sub-network. Signals are propagated from Input nodes through Filter nodes, detailed in section 3.2.2. Filtered symptoms are then combined in Combination nodes, detailed in section 3.2.3. The resulting values of the Combination nodes represent a relative ordering of the sub-networks or the parallel pathways within a sub-network, if any exist. The Cause nodes then take on the highest value of any Combination nodes which serve as input to that Cause node. The value of the Cause node can be taken as a relative ranking among all possible causes given the symptoms observed.

3.2.1 Input nodes

In work by other teams of scientists on the Data Interpretation Module, automated symptom detection algorithms are being built to feed data on input symptoms to the Input nodes in the Expert Network. The Input nodes are used to scale the input signals to a 0.0 to 1.0 range. The outgoing connections of these nodes carry a weight of 1.0 to transmit the adjusted input signal to the inferencing layers.

3.2.2 Filter nodes

Filter node functionality embodies the first level of inferencing. Processing at this level consists of first determining whether the input signal indicates the presence or absence of a symptom. This is accomplished by testing the signal against a fixed threshold value. Those signals which exceed the threshold are taken to represent positive evidence for that cause. Those which do not exceed the threshold are subtracted from 1.0 and the result taken as negative evidence for the cause.

Filter nodes fall into five categories, one per semantic qualifier. These categories are differentiated by three parameters, the Good-dog factor, the Bad-dog factor, and the outgoing connection weight. Based on the semantic qualifiers used by the experts, it is logical to assume that "Always" symptoms are, in general, better predictors than "Infrequently" symptoms for a given cause. In light of this, "Always" symptoms carry a higher contribution, whether positive or negative, to the conclusion than do "Infrequently" symptoms. This is accomplished by the three parameters associated with the Filter nodes.

Currently, the Good-dog factor serves to separate positive and negative evidence. All signals exceeding the threshold are multiplied by the Good-dog factor which is 1.0 for all Filter node classes except "Never". In effect, this passes the full-strength, positive evidence to the Combination node. "Never" symptoms which exceed the threshold are taken as strong evidence against a given cause and thus have Good-dog factors of -1.0. In the future, we will be investigating methods of using the Good-dog factor to filter evidence by allowing the training algorithms to modify the values of this parameter.

The Bad-dog factor serves to convert small positive signal values into negative evidence. Values range from -0.50 for "Always" symptoms to -0.001 for "Infrequently" symptoms. "Always" Bad-dog factors are high since these symptoms are expected to be good predictors, while "Infrequently" Bad-dog factors are small due to lack of reliability of these symptoms as predictors. "Never" Filter nodes have a Bad-dog factor of 0.0 to insure that the system does not produce a false-positive failure identification.

The weight of the outgoing connection from the Filter nodes is also qualifier dependent and serves to refine the contribution of a given symptom to the combination of all symptoms in a symptom set. Initial values of these weights range from 0.85 for "Always" nodes and "Never" nodes to 0.25 for "Infrequently" nodes. Currently the outgoing connection weights are the only trained parameters in the network. Modification of these weight values

during training results in an overall increase in average separation of the ranking values for Cause nodes.

3.2.3 Combination nodes

Combination nodes combine positive and negative evidence via an EMYCIN-like formula.¹⁰

$$R = y^+ + y^- \quad (1)$$

where

$$y^+ = 1 - \prod_{a_i > 0} (1 - a_i) \quad (2)$$

$$y^- = 1 + \prod_{a_i < 0} (1 + a_i) \quad (3)$$

$$a_i = w_{C_j F_k} * v_{F_k} \quad (4)$$

and $w_{C_j F_k}$ is the weight of the connection from Filter node k to Cause node j and v_{F_k} is the outgoing value of Filter node k .

Outgoing connections from the Combination nodes have a weight of 1.0.

3.2.4 Cause nodes

Typically, there is only one input to each Cause node and the Cause node simply takes the product of the value of the connected Combination node and the weight of the connection. However, in the case of parallel pathways for a single cause, the Cause node selects the highest value calculated by all pathways which terminate at that node.

4 RESULTS

Data from the Varian GC with induced faults is currently being collected. To date, the system has been tested using simulated symptom sets which match the experts' best guesses of the symptom/cause relationships. For a given cause related to five non-conflicting symptoms, for example, testing was performed with these five symptoms present. This testing has been performed with the related symptoms at full strength and with the symptoms at strengths degraded by random amounts up to 25%.

For the fourteen causes represented in the current network, the system correctly predicts the most likely cause of failure using the simulated data. This verifies that the system has adequately captured the current state of the experts' reasoning related to this problem. As real instrument failure data becomes available, it is highly likely that the system will need to be modified in order to predict correctly in all situations. It is rarely the case that systems built strictly from expertise gathered in the knowledge elicitation process perform perfectly when tested with real data. The strength of our representation of this system as an Expert Network lies in its ability to be modified using neural-network style training algorithms. Work is currently underway on prototype algorithms

which will adjust not only the outgoing connection weights of the Filter nodes, but also other connection strengths, Good-dog, Bad-dog, and threshold values, and the architectural topology of the network itself using training data gathered from the instrument's operation.

5 CONCLUSIONS

Creation of a knowledge-based, fault diagnosis system for automated GC systems as an Expert Network is a method which has several benefits over previous solution implementations. Representation of knowledge as an Expert Network allows the domain experts to use natural semantic qualifiers to identify relationships between symptoms and causes. Further, it exploits the distinctions in the predictive reliability of the symptoms while preserving the ambiguity inherent in the meaning of the semantic qualifiers used by the experts. In addition, it preserves the expert's reasoning mechanism by maintaining two layers of functional processing: filtering and evidence combination. Due to the inherent capabilities of knowledge-based networks, and Expert Networks in particular, it also introduces the potential for improved performance via data-based training.

This implementation has proven effective in the prediction of cause from symptoms with complete symptom sets and shows promise in the ability to function with uncertain and missing data. In addition, while this system was designed for fault diagnosis in automated GC systems, the methods employed in the creation of the system are designed to be readily extensible to fault diagnosis in other equipment.

6 FUTURE WORK

Future investigation is centered on three goals. The first is the use of connectionist-style learning techniques to improve the performance of the system. The second area of investigation is the automation of the creation of the GC fault diagnosis system and extension to other automated laboratory analytical systems. The third area of investigation encompasses methods to refine the symptom-to-cause relationships based on archival data. This includes introduction of relationships undiscovered by the domain experts as well as destruction of relationships that are not verifiable by the data.

7 ACKNOWLEDGEMENTS

This work is supported in part by the U.S. Department of Energy, Assistant Secretary for Environmental Management, Under DOE-Idaho Operations Office, Contract DE-AC07-94ID13223, and the Florida High Technology and Research Council.

8 REFERENCES

- [1] J. W. Elling, L. K. Klatt and W. P. Unruh, "Automated Data Interpretation in an Automated Environmental Laboratory", Laboratory Robotics and Automation, Vol. 6 Num. 2, pp. 73-78, 1994
- [2] T. H. Erkkila, R. M. Hollen and T. J. Beugelsdijk, "The Standard Laboratory Module: An Integrated Approach to Standardization in the Analytical Laboratory", Laboratory Robotics and Automation, Vol. 6 Num. 2, pp. 57-64, 1994

- [3] M. J. Stillman, G. Huang, S. Lahiri and Q. Zhu, "ACexpert. Design and Implementation of ACselect, AAexpert and GC-MSexpert: Expert Systems that aid in the Analysis of Environmental Samples", The World Congress on Expert Systems Proceedings, pp. 2645-2453, 1991
- [4] R. Bach, J. Karnicky, S. Abbott, "An Expert System for High Performance Liquid Chromatography Methods Development", Artificial Intelligence Applications in Chemistry, T. H. Pierce, B. A. Hohne, ACS Symposium Series 306, pp. 278-296, American Chemical Society, Washington D. C., 1986
- [5] D. C. Kuncicky, S. I. Hruska, R. C. Lacher, "Hybrid Systems: The Equivalence of Rule-Based Expert System and Artificial Neural Network Inference", International Journal of Expert Systems, Vol. 4 Num. 3, pp. 281-297, 1992
- [6] R. C. Lacher, S. I. Hruska, and D. C. Kuncicky, "Backpropagation Learning in Expert Networks", IEEE Transactions on Neural Networks, Volume 3, Number 1, January 1992, pp. 62-72.
- [7] S. I. Hruska and D. C. Kuncicky, "Automated Knowledge Refinement for Control Chart Selection", Heuristics, The Journal of Knowledge Engineering, Volume 5, Number 3, Fall 1992, pp. 101-107.
- [8] S. I. Hruska, "Using Expert Networks For Fault Tolerance in a Control Chart Selection System", Intelligent Engineering Systems Through Artificial Neural Networks, C. H. Dagli, L. I. Burke, B. R. Fernández, and J. Ghosh, Volume 3, pp. 695-700, ASME Press, New York, NY, 1993.
- [9] R. C. Lacher, "Expert Networks: Paradigmatic Conflict, Technological Rapprochement", Minds and Machines, Volume 3, 1993, pp. 53-71.
- [10] E. H. Shortliffe, B. G. Buchanan, "Chapter 11: A model of Inexact Reasoning in Medicine", Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project, B. G. Buchanan, E. H. Shortliffe, pp. 233-262, Addison-Wesley, Reading, MA, 1984.