

Title:

THE SI-COMBINER: DETECTION AND USING
SCENARIO INFORMATION FOR SMART SYSTEMS

CONF-990402--

Author(s):

Bob K. Den Hartog, ESA-EPE
John W. Elling, BioReason, Inc
Roger M. Kieckhafer, University of
Nebraska

Submitted to:

ANS 8th Topical Meeting on Robotics &
Remote Systems
April 25-29, 1999
Pittsburgh, PA USA

RECEIVED
MAR 12 1999
OSTI

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

MASTER

Los Alamos
NATIONAL LABORATORY

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the University of California for the U.S. Department of Energy under contract W-7405-ENG-36. By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. The Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

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.

DISCLAIMER

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

THE SI-COMBINER: DETECTING AND USING SCENARIO INFORMATION FOR SMART SYSTEMS

Bob K. Den Hartog (Los Alamos National Laboratory), John W. Elling (BioReason, Inc.) and Roger M. Kieckhafer (University of Nebraska)

Mail Stop J580, Los Alamos, NM 87545/USA

e-mail: BobiKai@aol.com

Tel: 505-665-3231

ABSTRACT

Many problems, such as Aroclor Interpretation, are ill-conditioned problems in which trained programs, or methods, must operate in scenarios outside their training ranges because it is intractable to train them completely. Consequently, they fail in ways related to the scenarios.

Importantly, when multiple trained methods fail divergently, their patterns of failures provide insights into the true results. The SI-Combiner solves this problem of Integrating Multiple Learned Models (IMLM) by automatically learning and using these insights to produce a solution more accurate than any single trained program. In application, the Aroclor Interpretation SI-Combiner improved on the accuracy of the most accurate individual trained program in the suite.

For Smart Systems, the implication of trusting a single computation or sensor that is not able to report its own accuracy is devastating. Once one begins to compare multiple results, majority rule may not make sense, especially for computations that are not provably independent. With *approximate* understanding of the conditions that confound an individual computation or sensor, the presented IMLM method allows sensible interpretation of multiple, possibly disagreeing, results.

This paper presents a new fuzzy IMLM method called the SI-Combiner and its application to Aroclor Interpretation. Additionally, this paper shows the improvement in accuracy that the SI-Combiner's components show against Multicategory Classification (MCC), Dempster-Shafer (DS), and the best individual trained program in the Aroclor Interpretation suite (iMLR).

1. Introduction

The SI-Combiner uses a suite of trained programs, or methods, which must operate in conditions that are outside of their training domains and in which they may fail. The SI-Combiner incorporates application-specific knowledge, called scenarios, about conditions in the testing domains that can distort the inputs and thereby cause divergent behavior among the methods.

Since scenarios are defined as signal conditions, during periodic retraining the method training set (which does not cover the testing domain) can be synthetically extended and distorted to provide training samples throughout the entire testing domain.

From the extended training set, the SI-Combiner learns two types of information:

Scenario Identification (SI) rules, which capture how patterns of method opinions give information about a sample's scenario, and Method Weighting rules, which capture for each scenario, the accuracy of a method's opinion.

Armed with the SI rules and the learned biases in the Method Weighting rules, the SI-Combiner first analyzes a test sample and assigns values between zero and one for each possible scenario for the sample. Higher numbers indicate higher likelihood that the sample belongs to that scenario. This SI step is key and gives the SI-Combiner its name.

Finally, the SI-Combiner blends the scenario-based accuracy information, the method opinions and the scenario information producing final results, confidences, context information and explanations.

2. Application

2.1 Chemistry

Certain chemical compounds named *Aroclor 1242*¹, *Aroclor 1254* and *Aroclor 1260* have become environmental concerns as carcinogens. One method of detecting them is to use gas chromatography, which produces data in the form of chromatograms, shown in Figure 1.

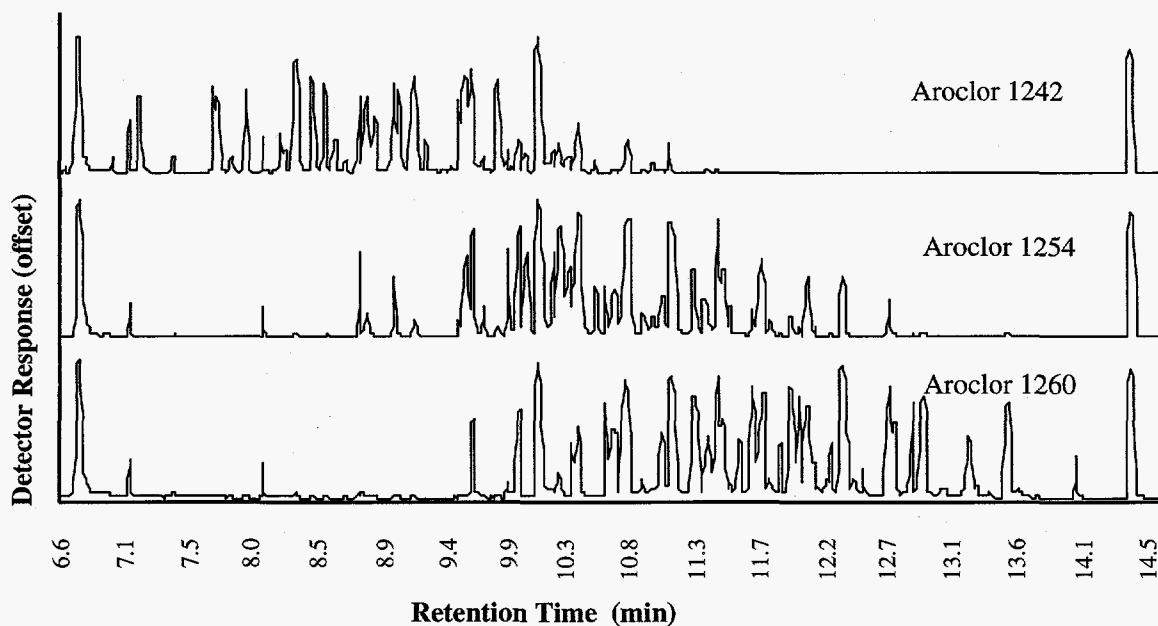


Figure 1: Chromatograms of Aroclor 1242, Aroclor 1254 and Aroclor 1260

¹ *Aroclor* is a trademark of Monsanto Corporation.

2.1.1 Detecting and quantifying Aroclors in chromatograms

Aroclor Interpretation examines a chromatogram, such as shown in Figure 1 and determines the presence or absence of each Aroclor. For each Aroclor present, the concentration in micrograms per milliliter (ug/ml) is measured. Each peak represents a separate component. The x-axis value for a peak reveals the component's identity, while the y-axis value indicates the concentration.

The y-axis value, however, has no units: interpretation of a *test* chromatogram with unknown content must be based on comparisons with *standard* chromatograms (or *standards*) with known Aroclor concentrations.

2.2 The Training Domain For Aroclor Interpretation

Before processing a batch of tests, operators produce a set of five standard chromatograms with standard concentrations for each of the three Aroclors to create fifteen standards. By comparing the chromatogram of a test to the fifteen standard chromatograms, a chemist or an automated pattern recognition method estimates the concentration or quantity of an Aroclor believed present. These single, pure Aroclors constitute the *training domain*.

2.2.1 The Training Domain Cannot Cover The Testing Domain

When the automated method or the chemist encounters a sample with conditions like those in the training domain, its chromatogram can be accurately compared with the standards. However, many real-life samples contain multiple Aroclors, no Aroclors, aged or weathered Aroclors, and materials that are not Aroclors but appear in the chromatogram. The lower chromatogram in Figure 2 contains the same Aroclor content as the upper chromatogram, but the added presence of oil distorts the baseline and creates nonAroclor peaks. It is intractable to train automated methods or humans on all possible conditions encountered in the *testing domain*. However, they must work on samples with conditions outside their training domain. The SI-Combiner enables this by exploiting the variations in conditions through the testing domain.

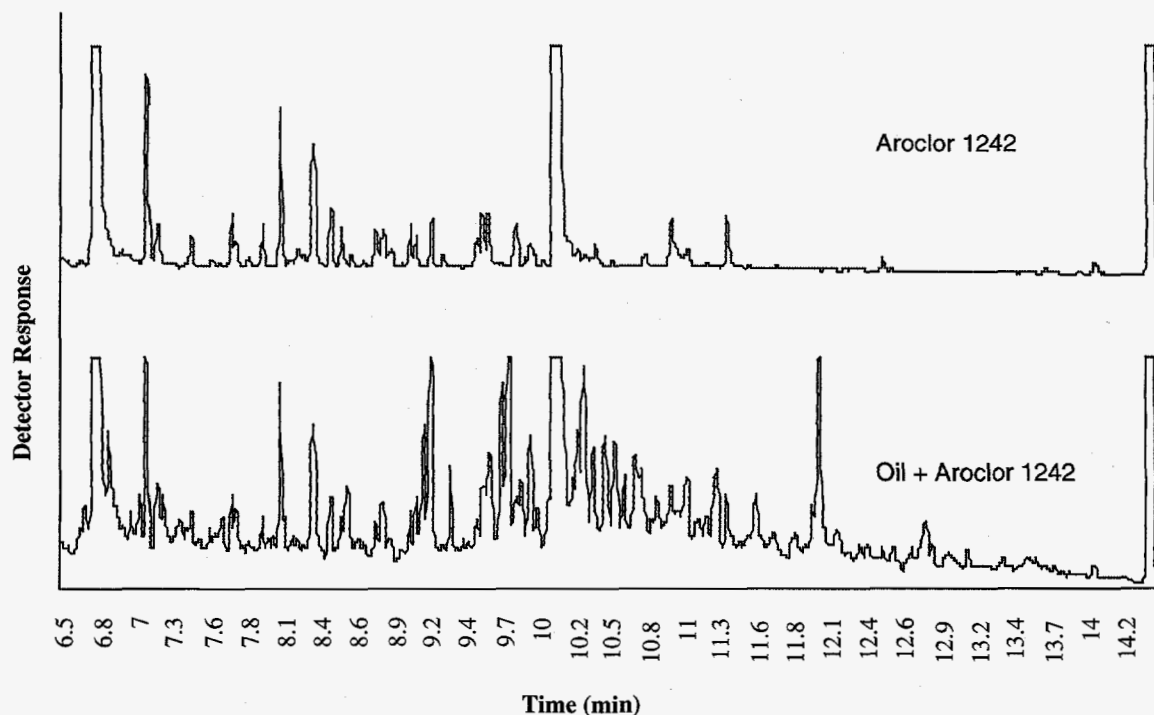


Figure 2: Chromatograms of Aroclor 1242 with and without oil contamination

2.3 Automated Aroclor Interpretation

The Aroclor Interpretation Data Interpretation Module (DIM)[i] is shown in Figure 3.

2.3.1 SI-Combiner Uses Opinions From Trained Methods

Five automated pattern recognition methods were implemented in the Aroclor Interpretation DIM. Each method, like a chemist, receives fifteen standards for each retraining.

Four methods produce quantity estimates for each Aroclor:

- Iterative Multiple Linear Regression (iMLR)
- Neural Network (NN)
- Principle Components Regression (PCR)
- Peak Area Analysis (PA)

The fifth method produces a value between zero and one for each Aroclor reflecting its belief that the Aroclor is present:

- Neural Network for Identification (NNI) [ii]

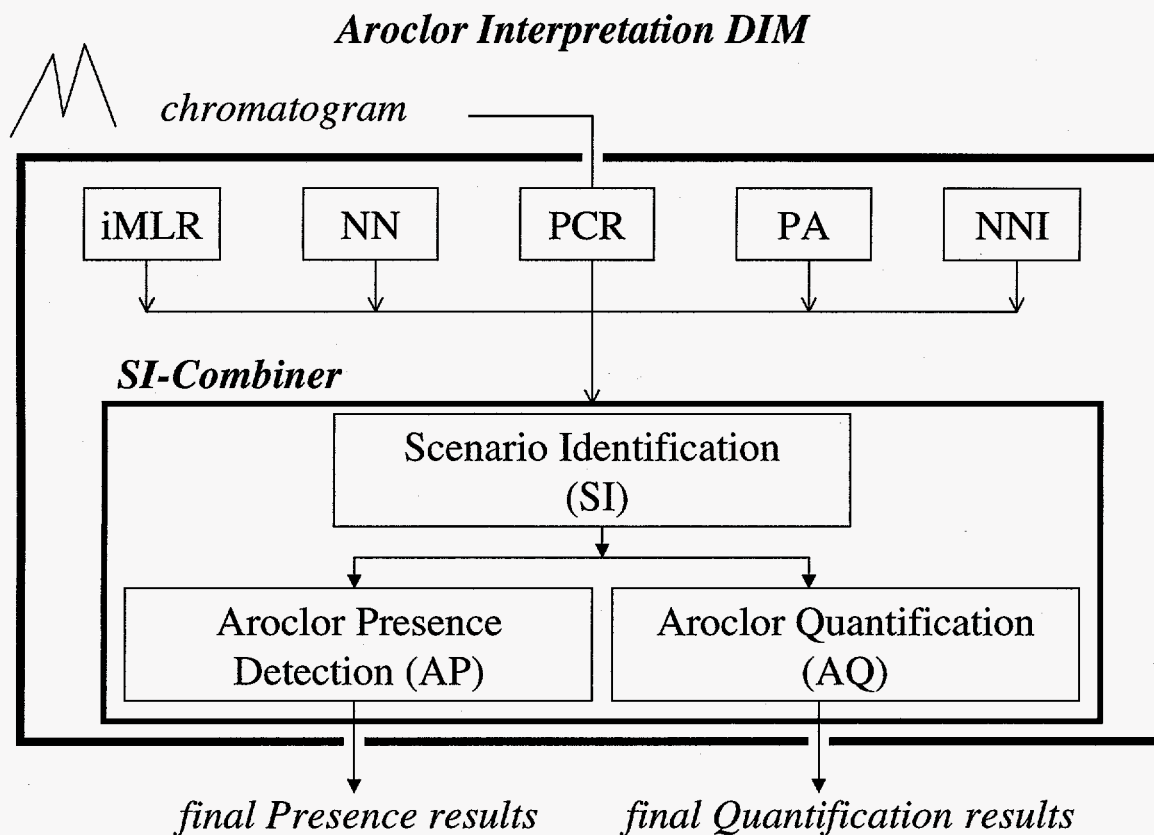


Figure 3: Aroclor Interpretation DIM including the SI-Combiner.

Each method produces an opinion for each of the three target Aroclors, so that the SI-Combiner receives an *opinion set* containing fifteen views of a sample.

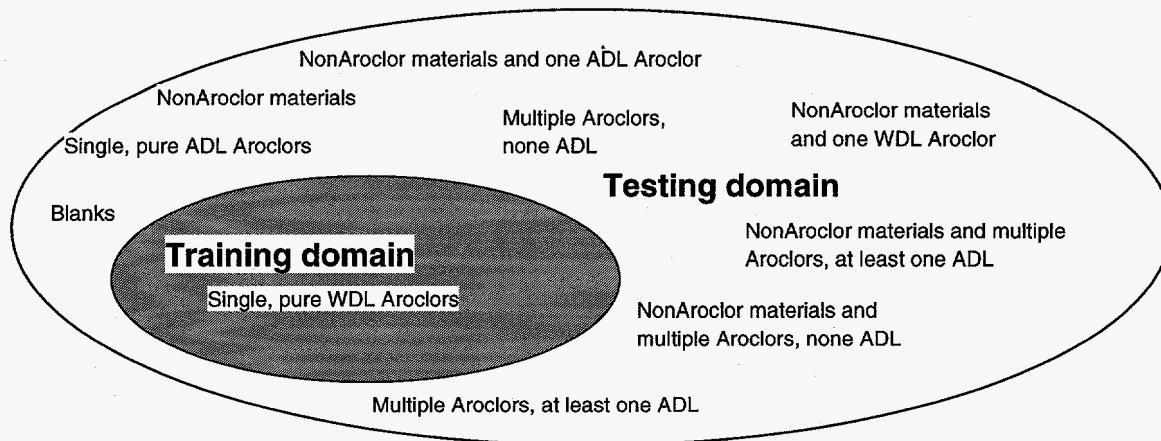
2.3.2 Scenarios Are Signal Conditions That Distort Method Opinions

Methods disagree due to conditions in the signal. In Aroclor Interpretation, and in other problems, these conditions can be *approximately* identified.

Scenario elements are

1. conditions in the method's inputs that affect the methods responses, which in turn, are
2. related to patterns of opinion sets.

For Aroclor Interpretation, five scenario elements are used to partition the testing domain into scenarios. These elements include one contamination descriptor, three presence descriptors (*none*, *single* and *multiple*), and one overconcentration descriptor. Together, these can combine to partition the entire testing domain into ten scenarios, shown in Figure 4. Only the samples with a single, pure WDL Aroclor in them have conditions that match the Training domain for these particular Methods.



Aroclor quantities are

- **BDL** Below Detection Limits
- **WDL** Within Detection Limits
- **ADL** Above Detection Limits

Figure 4: Samples with conditions inside and outside the Methods' Training domain

The effect of contamination is seen in the lower chromatogram of Figure 2.

An opinion set is fifteen opinions, in a specified order, for a sample. Scenario changes cause observable changes in these patterns.

Evidence of these observable changes in patterns of opinion sets, which are directly linked to changes in scenario, is shown in Figure 5. For simplicity, the opinion sets contain information for one Aroclor at a time, so five-tuples are plotted instead of fifteen-tuples. In these radar plots, each opinion is mapped on a separate spoke, creating a spatial view of the relationships among the opinions. The left plot of Figure 5 shows opinion sets for samples in the training domain, i.e., single uncontaminated Aroclors. The radial symmetry reflects the identical performances by the four quantification Methods for each test sample, which are accurate since they are operating in their training domain.

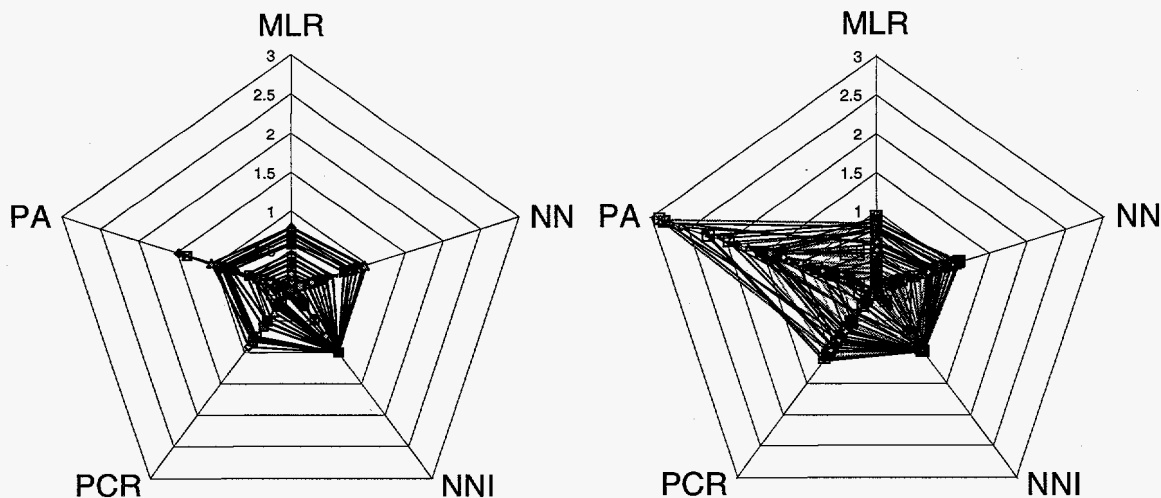


Figure 5: Opinion set patterns vary according to scenario

The plot on the right side of Figure 5 shows the opinions sets for samples with *multiple* pure Aroclors. The pattern for samples in this scenario shows distortion along the upper left axis. Each of the ten scenarios for Aroclor Interpretation has a pattern. These patterns are learned automatically and matched approximately by the SI step.

2.3.3 Artificially Extend the SI-Combiner's Training Domain

To learn how method accuracy depends on scenario, the SI-Combiner needs training data consisting of method opinions for samples in all scenarios. The specific definition of scenario requires that the method opinions vary according to the scenario and that scenarios are due to conditions in the input signal. This allows the explicit manipulation of signals to synthesize artificial training data. The SI-Combiner artificially extends the Method training domain through the Synthetic Chromatogram Generator. The Synthetic Chromatogram Generator uses the fifteen genuine standards that constitute the Method Training Set and a Noise Engine to generate chromatograms in all scenarios. These chromatograms are *representative*; much of the testing domain is still not covered by the training samples.

2.3.4 Automatically Learning SI And Method Weighting Rules

Two types of rule sets are in the SI-Combiner:

1. Scenario Identification: fuzzy classification produces *scenario weights*, one per scenario, which measure the test sample's similarities to training examples in each scenario. The key to the SI-Combiner is the Scenario Identification (SI) step, as indicated in the technique's name.
2. Method-Weighting: For Aroclor Interpretation, there are two types of accuracy measured since two questions need to be answered. The first Method-Weighting rule set is the Aroclor Presence Detection (AP) rule set, which is used to answer the question "Is the Aroclor Present?". The second Method-Weighting rule set is the Aroclor Quantification (AQ) rule set, which characterizes each quantitative method's bias in each scenario, and is used to answer "What quantity of Aroclor is present?"

2.3.5 Approximately Match Opinion Patterns from Test Samples to Training Samples

The SI-Combiner exploits very sparse and representative training data, through fuzzy logic, by *approximately* matching an SI rule.

Every testing sample will match every rule to some degree, and a rule's firing weight clearly reflects how closely the test sample resembles a sample seen in the training set. As shown in Figure 3, the SI-Combiner works in steps. First, SI produces fuzzy belief weights for each scenario by approximately matching the SI rules.

2.3.6 Weight Method Opinions By Scenario Likelihood And Method Accuracy

The SI-Combiner's second step, AP, classifies each of three Aroclors as *present* or *absent* in a sample.

The third step, AQ, produces quantitative values for the concentration of each Aroclor in a sample.

AP and AQ use automatically learned empirical biases for each of the Methods in each scenario. Through fuzzy logic, AP and AQ combine scenario weights, automatically learned biases for each of the Methods in each scenario, and Methods' results to determine results for a sample. AP and AQ use well-established techniques internally; their novel features are the use of SI's scenario weights, and the nature of the bias estimations learned by each step.

3. Results

The SI-Combiner improves on two existing approaches to IMLM, which combine either

- multiple experts each viewing a portion of the problem, or
- selection of *one* expert's view of the entire problem.

The SI-Combiner, in contrast, blends multiple experts, each viewing the whole problem, according to likely accuracies for a given sample. The SI-Combiner uses:

- all available information
- biases found through empirical learning
- application-specific knowledge to predict and exploit variations in the scenarios

3.1 Components of the SI-Combiner

Figure 6 summarizes these results.

Scenario Identification : <i>Class</i>	Combiner vs. Multicategory Classification (MCC) <i>Combiner improves on MCC by:</i>
Contamination	15%
None	1%*
Single	55%
Multiple	55%
Overconcentration	6%

* MCC misclassified two samples; the Combiner correctly classified all samples

Aroclor Presence Detection: <i>Error Type</i>	Combiner vs. Dempster Shafer (DS) <i>Combiner improves on DS by:</i>
False Positive	58%
False Negative	78%

Aroclor Quantification:	<i>Combiner improves on listed technique by:</i>
MLR	14%
Averaging	92%
Peak Area	95%

Figure 6: Summary of SI-Combiner improvements

For SI, the SI-Combiner was compared to Multicategory Classification (MCC)[iii]. The SI-Combiner handled the nonlinear, poorly separated space far better than MCC did.

Furthermore,

- as a fuzzy classifier, the SI-Combiner accrues evidence of a test sample's membership in all scenarios, which is passed on without loss to the SI-Combiner's method-weighting steps.
- SI rule firing weights are available for explanations, accountability and tracability.

For AP, the SI-Combiner was compared to Dempster-Shafer (DS)[iv]. As summarized in Figure 6, the SI-Combiner produced fewer errors than DS did, and referred fewer samples with indeterminate results to the operator. Furthermore,

- the SI-Combiner uses empirical biases; DS uses errors reported by the Methods, which, like opinions, are not guaranteed accurate for samples outside the training domain.

For AQ, the SI-Combiner greatly outperformed a scenario-ignorant combining method of Averaging and the human representative Method of PA. The SI-Combiner improved on the best individual method, iMLR.

4. Conclusion

The SI-Combiner uses a suite of trained methods, which must operate in conditions that are outside of their training domains and in which they may fail. The SI-Combiner incorporates application-specific knowledge, called *scenarios*, about conditions in the testing domains that can distort the inputs and thereby cause divergent behavior among the methods. This problem is a version of Integrating Multiple Learned Models (IMLM).

The SI-Combiner improves on existing approaches to IMLM, and its components perform better than traditional approaches to the same tasks. The SI-Combiner is a new and powerful solution for IMLM problems.

ACKNOWLEDGEMENTS

This work is supported in part by the U.S. Department of Energy under the Defense Programs Technology Transfer Initiative Program. Portions of this work were performed under a Cooperative Research and Development Agreement with Varian Chromatography Systems Business, Lockheed Missiles and Space Company, and NeuralWare, Inc.

REFERENCES

i J.W.Elling, S.M.Mniszewski, J.D.Zahrt, L.N.Klatt. "Automated Chromatographic Data Interpretation Using an Expert System." *Journal of Chromatographic Science*, Vol. 32 No. 6.

ii G.R.Magelssen, J.W.Elling "Chromatography Pattern Recognition using Iterative Probabilistic Neural Networks", *Journal of Chromatography A*, to be published.

iii J.T.Tou, R.C.Gonzalez. *Pattern Recognition Principles*, Addison-Wesley 1974.

