Lecture Notes in
Computer Science 1280

Xiaohui Liu Paul Cohen
Michael Berthold (Eds.)

Advances in Intelligent Data Analysis

Reasoning about Data

Second International Symposium, IDA-97
London, UK, August 1997
Proceedings

Springer

Springer
Berlin
Heidelberg
New York
Barcelona
Budapest
Hong Kong
London
Milan
Paris
Santa Clara
Singapore
Tokyo

Xiaohui Liu Paul Cohen
Michael Berthold (Eds.)

Advances in Intelligent Data Analysis

Reasoning about Data

Second International Symposium, IDA-97
London, UK, August 4-6, 1997
Proceedings

Springer

Volume Editors

Xiaohui Liu
Birkbeck College, University of London, Department of Computer Science
Malet Street, London WC1E 7HX, UK
E-mail: hui@dcs.bbk.ac.uk

Paul Cohen
University of Massachusetts, Amherst, Department of Computer Science
Amherst, MA 01003-4610, USA
E-mail: cohen@cs.umass.edu

Michael Berthold
Universität Karlsruhe, IRF
Am Zirkel 2, D-76128 Karlsruhe, Germany
E-mail: berthold@ira.uka.de

Cataloging-in-Publication data applied for

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Advances in intelligent data analysis : reasoning about data ; second international symposium ; proceedings / IDA-97, London, UK, August 4 - 6, 1997. X. Liu ... (ed.). - Berlin ; Heidelberg ; New York ; Barcelona ; Budapest ; Hong Kong ; London ; Milan ; Paris ; Santa Clara ; Singapore ; Tokyo : Springer, 1997
(Lecture notes in computer science ; Vol. 1280)
ISBN 3-540-63346-4

CR Subject Classification (1991): H.3, I.2, G.3, I.5.1, I.4.5, J.2, J.1, J.3

ISSN 0302-9743
ISBN 3-540-63346-4 Springer-Verlag Berlin Heidelberg New York

Typesetting: Camera-ready by author
SPIN 10547800 06/3142 – 5 4 3 2 1 0 Printed on acid-free paper

Foreword

This volume is a collection of papers presented at the Second International Symposium on Intelligent Data Analysis (IDA-97) held at Birkbeck College, University of London, August 4-6, 1997. Our community is growing fast: The first Symposium, held in Baden-Baden in 1995, received 69 extended abstracts for review; this time we received 107 full papers. We accepted 50, divided equally between oral and poster presentations. Each paper was considered by at least two independent reviewers and the results were discussed in a meeting of program committee members.

IDA-97 was a single-track conference consisting of oral and poster presentations, invited speakers, demonstrations, and exhibitions. The conference Call for Papers introduced a theme, "Reasoning About Data", and many papers complement this theme, but other, exciting topics have emerged, including exploratory data analysis, data quality, knowledge discovery, and data-analysis tools, as well as the perennial technologies of classification and soft computing. A new and exciting theme involves analyzing time series data from physical systems, such as medical instruments, environmental data, and industrial processes.

We are very grateful to Professor David Hand, of the Open University, and Dr. Larry Hunter, of the National Library of Medicine in Washington D.C., for agreeing to give keynote talks. Each has made major contributions and has helped to define the interdisciplinary field of intelligent data analysis.

The symposium owes much to many hard-working individuals. Michael Berthold handled publicity and publications, and is responsible for this volume. The program committee and additional volunteer reviewers read the papers with great care and wrote excellent, informative, helpful reviews. Terrie Korpita and Peggy Weston at the University of Massachusetts coordinated the international reviewing process, a huge job performed flawlessly. The organization at Birkbeck has been tremendous. Phil Gregg was responsible for the IDA-97 homepage, which served to keep the community abreast of developments and was responsible in no small part for the increasing size of the symposium. Betty Walters, Sylvie Jami, Trevor Fenner, Claude Gierl, and Steven Swift were responsible for daily operations, enquiries, registrations, and the details that grew exponentially as the conference date drew nearer. The Systems group in Computer Science at Birkbeck College provided valuable computing support for the reviewing process and conference exhibitions. The symposium would not exist were it not for the confidence and financial backing of Birkbeck College, particularly Roger Johnson, Ken Thomas, and Guy Fitzgerald.

June 1997

Xiaohui Liu
Paul Cohen

Organization Committee

General Chair:

Xiaohui Liu
Department of Computer Science
Birkbeck College, Malet Street
London WC1E 7HX, UK
E-mail: hui@dcs.bbk.ac.uk

Technical Program:

Paul Cohen
Department of Computer Science
University of Massachusetts, Amherst
Amherst, MA 01003-4610, USA
E-mail: cohen@cs.umass.edu

Publicity&Publication:

Michael Berthold
Universität Karlsruhe, IRF
Am Zirkel 2, 76128 Karlsruhe, Germany
E-mail: berthold@ira.uka.de

Local Arrangements:

Trevor Fenner
Department of Computer Science
Birkbeck College, Malet Street
London WC1E 7HX, UK
E-mail: trevor@dcs.bbk.ac.uk

Finance:

Sylvie Jami
Department of Computer Science
Birkbeck College, Malet Street
London WC1E 7HX, UK
E-mail: s.jami@dcs.bbk.ac.uk

Sponsorship:

Mihaela Ulieru
Simon Fraser University
2357 Riverside Drive
North Vancouver, B.C.
Canada V7H 1V8
Email: ulieru@cs.sfu.ca

Exhibition:

Richard Weber
MIT GmbH, Promenade 9
52076 Aachen, Germany
E-mail: rw@mitgmbh.de

Program Committee

Eric Backer	Delft University of Technology, The Netherlands
Riccardo Bellazzi	University of Pavia, Italy
Michael Berthold	University of Karlsruhe, Germany
Carla Brodley	Purdue University, USA
Gongxian Cheng	Birkbeck College, UK
David Dowe	Monash University, Australia
Fazel Famili	National Research Council, Canada
Julian Faraway	University of Michigan, USA
Thomas Feuring	WWU Münster, Germany
Alex Gammerman	Royal Holloway London, UK
David J Hand	Open University, UK
Rainer Holve	Forwiss Erlangen, Germany
Wenling Hsu	AT&T Consumer Lab, USA
Larry Hunter	National Library of Medicine, USA
David Jensen	University of Massachusetts, USA
Frank Klawonn	University of Braunschweig, Germany
David Lubinsky	University of Witwatersrand, South Africa
Ramon Lopez de Mantaras	Artificial Intelligence Research Institute, Spain
Sylvia Miksch	Stanford University, USA
Rob Milne	Intelligent Applications Ltd, UK
Gholamreza Nakhaeizadeh	Daimler-Benz Forschung und Technik, Germany
Claire Nedellec	Université Paris-Sud, France
Erkki Oja	Helsinki University of Technology, Finland
Henri Prade	University Paul Sabatier, France
Daryl Pregibon	AT&T Research, USA
Peter Ross	University of Edinburgh, UK
Steven Roth	Carnegie Mellon University, USA
Lorenza Saitta	University of Torino, Italy
Peter Selfridge	AT&T Research, USA
Rosaria Silipo	University of Florence, Italy
Evangelos Simoudis	IBM Almaden Research, USA
Derek Sleeman	University of Aberdeen, UK
Paul Snow	Consultant, USA
Rob St. Amant	North Carolina State University, USA
Lionel Tarassenko	Oxford University, UK
John Taylor	King's College London, UK
Loren Terveen	AT&T Research, USA
Hans-Jürgen Zimmermann	RWTH Aachen, Germany

Additional Reviewers: Gilles Bisson, Mark Derthick, Peter Edwards, G.C. van den Eijkel, Yves Kodratoff, Nigel Martin, Andrew McQuatt, F. Mitchell, Marjorie Moulet, Ch. Vrain, Simon White, Mark Winter.

Table of Contents

Intelligent Data Analysis: Issues and Opportunities 1
David J. Hand (UK)

Section I: Exploratory Data Analysis, Preprocessing, and Tools

Decomposition of Heterogeneous Classification Problems 17
C. Apte, S. J. Hong, J. R. M. Hosking, J. Lepre, E. P. D. Pednault and B. K. Rosen (USA)

Managing Dialogue in a Statistical Expert Assistant with a Cluster-Based User Model 29
M. W. Muller (South Africa)

How to Find Big-Oh in Your Data Set (and How Not to) 41
C. C. McGeoch, D. Precup and P. R. Cohen (USA)

Data Classification Using a W.I.S.E. Toolbox 53
I. Berry and P. Gough (UK)

Mill's Methods for Complete Intelligent Data Analysis 65
T. A. O. Cornish (UK)

Integrating Many Techniques for Discovering Structure in Data 77
D. E. Gregory and P. R. Cohen (USA)

Meta-Reasoning for Data Analysis Tool Allocation 89
R. Levinson and J. Wilkinson (USA)

Navigation for Data Analysis Systems 101
R. St. Amant (USA)

An Annotated Data Collection System to Support Intelligent Analysis of Intensive Care Unit Data 111
C. L. Tsien and J. C. Fackler (USA)

A Combined Approach to Uncertain Data Analysis 123
H. Yu and A. Ramer (Australia)

Section II: Classification and Feature Selection

A Connectionist Approach to the Distance-Based Analysis of Relational Data 137
K. Schädler and F. Wysotzki (Germany)

Efficient GA Based Techniques for Automating the Design of Classification Models ... 149
R. Glover and P. Sharpe (UK)

Data Representations and Machine Learning Techniques ... 161
C. P. Lam, G. A. W. West and T. M. Caelli (Australia)

Development of a Knowledge-Driven Constructive Induction Mechanism ... 173
S. Lo and A. Famili (Canada)

Oblique Linear Tree ... 187
J. Gama (Portugal)

Feature Selection for Neural Networks through Functional Links Found by Evolutionary Computation ... 199
S. Haring, J. N. Kok and M. C. van Wezel (The Netherlands)

Building Simple Models: A Case Study with Decision Trees ... 211
D. Jensen, T. Oates and P. R. Cohen (USA)

Exploiting Symbolic Learning in Visual Inspection ... 223
M. Piccardi, R. Cucchiara, M. Bariani and P. Mello (Italy)

Forming Categories in Exploratory Data Analysis and Data Mining ... 235
P. D. Scott, R. J. Williams and K. M. Ho (UK)

A Systematic Description of Greedy Optimisation Algorithms for Cost Sensitive Generalisation ... 247
M. van Someren, C. Torres and F. Verdenius (The Netherlands)

Dissimilarity Measure for Collections of Objects and Values ... 259
P. Valtchev and J. Euzenat (France)

Section III: Medical Applications

ECG Segmentation Using Time-Warping ... 275
H. J. L. M. Vullings, M. H. G. Verhaegen and H. B. Verbruggen (The Netherlands)

Interpreting Longitudinal Data through Temporal Abstractions: An Application to Diabetic Patients Monitoring ... 287
R. Bellazzi, C. Larizza and A. Riva (Italy)

Intelligent Support for Multidimensional Data Analysis in Environmental Epidemiology ... 299
V. Kamp and F. Wietek (Germany)

Section IV: Soft Computing

Network Performance Assessment for Neurofuzzy Data Modelling ... 313
S. R. Gunn, M. Brown and K. M. Bossley (UK)

A Genetic Approach to Fuzzy Clustering with a Validity Measure Fitness Function 325
S. Nascimento and F. Moura-Pires (Portugal)

The Analysis of Artificial Neural Network Data Models 337
C. M. Roadknight, D. Palmer-Brown and G. E. Mills (UK)

Simulation Data Analysis Using Fuzzy Graphs 347
K.-P. Huber and M. R. Berthold (Germany)

Mathematical Analysis of Fuzzy Classifiers 359
F. Klawonn and E.-P. Klement (Germany; Austria)

Neuro-Fuzzy Diagnosis System with a Rated Diagnosis Reliability and Visual Data Analysis 371
A. Lapp and H.-G. Kranz (Germany)

Genetic Fuzzy Clustering by Means of Discovering Membership Functions 383
M. Turhan (Turkey)

Section V: Knowledge Discovery and Data Mining

A Strategy for Increasing the Efficiency of Rule Discovery in Data Mining 397
D. McSherry (UK)

Intelligent Text Analysis for Dynamically Maintaining and Updating Domain Knowledge Bases 409
K. Schnattinger and U. Hahn (Germany)

Knowledge Discovery in Endgame Databases 423
M. Schlosser (Germany)

Parallel Induction Algorithms for Data Mining 437
J. Darlington, Y.-K. Guo, J. Sutiwaraphun and H. W. To (UK)

Data Analysis for Query Processing 447
J. Robinson and B. G. T. Lowden (UK)

Datum Discovery 459
L. Siklóssy and M. Ayel (France)

A Connectionist Approach to Extracting Knowledge from Databases 465
Y. Zhou, Y. Lu and C. Shi (China)

Section VI: Estimation, Clustering

A Modulated Parzen-Windows Approach for Probability Density Estimation ... 479
G. C. van den Eijkel, J. C. A. van der Lubbe and E. Backer (The Netherlands)

Improvement on Estimating Quantiles in Finite Population Using Indirect Methods of Estimation ... 491
M. R. García, A. A. Cebrián and E. A. Rodríguez (Spain)

Robustness of Clustering under Outliers ... 501
Y. Kharin (Belarus)

The BANG-Clustering System: Grid-Based Data Analysis ... 513
E. Schikuta and M. Erhart (Austria)

Section VII: Data Quality

Techniques for Dealing with Missing Values in Classification ... 527
W. Z. Liu, A. P. White, S. G. Thompson and M. A. Bramer (UK)

The Use of Exogenous Knowledge to Learn Bayesian Networks from Incomplete Databases ... 537
M. Ramoni and P. Sebastiani (UK)

Reasoning about Outliers by Modelling Noisy Data ... 549
J. X. Wu, G. Cheng and X. Liu (UK)

Section VIII: Qualitative Models

Reasoning about Sensor Data for Automated System Identification ... 561
E. Bradley and M. Easley (USA)

Modelling Discrete Event Sequences as State Transition Diagrams ... 573
A. E. Howe and G. Somlo (USA)

Detecting and Describing Patterns in Time-Varying Data Using Wavelets ... 585
S. Boyd (Australia)

Diagnosis of Tank Ballast Systems ... 597
B. Schieffer and G. Hotz (Germany)

Qualitative Uncertainty Models from Random Set Theory ... 609
O. Wolkenhauer (UK)

Author Index ... 621

Diagnosis of Tank Ballast Systems

Björn Schieffer and Günter Hotz

Fachbereich 14 – Informatik, Universität des Saarlandes
Postfach 15 11 50, 66041 Saarbrücken, FRG

URL: http://www-hotz.cs.uni-sb.de/schieffer/
email: schieffer@cs.uni-sb.de

Abstract. The paper deals with the diagnosis problem of hybrid systems. A new two-level approach for that problem is introduced and discussed with the domain example of tank ballast systems. The first level determines the possible defects while the second one calculates their real valued degree. The new approach is shown to be very powerful by 3000 randomly generated single and double faults. In fact, for all of these defects the approach is able to compute the correct diagnoses.

1 Introduction

The control unit of a hybrid system and its environment consisting of mechanical and hydrodynamical components influence each other in a closed loop via sensors and actuators. Such a situation may be found in modern car management units, air planes or manufacturing plants.

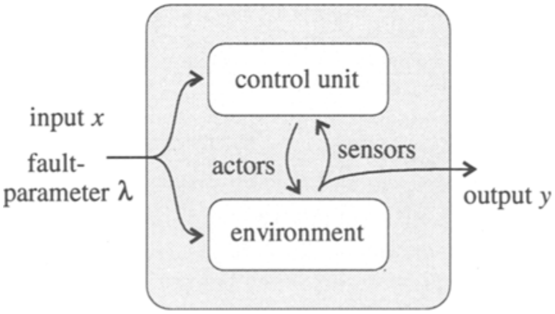

Fig. 1. Scheme of a hybrid system

According to figure 1, the behavior of a hybrid system depends on an input $x \in X$ of the input space X and a fault parameter $\lambda \in C$ of the configuration

X. Liu, P. Cohen, M. Berthold (Eds.): "Advances in Intelligent Data Analysis" (IDA–97)
LNCS 1280, pp. 597–608, 1997.

space C. The output $y = (y_1, y_2, \ldots, y_r) \in Y$ of the output space Y is measured by the r sensors of the system.

The problem to determine the output $y = sim(x, \lambda)$ with known input x and fault parameter λ is called *simulation.* Reality performs some kind of simulation, it *solves* the underlying differential equations when the system is running. If one is able to control the arising numerical difficultys, he can imitate this natural simulation. Much harder is the problem of *diagnosis* which is an inverse problem to simulation. To compute the diagnoses one has to determine the possible fault parameters λ when the input x and the output y are known. If *sim* is not injective, a diagnosis does not have to be unique and more than one fault parameters may explain the behavior of the system. Therefore, a *complete diagnosis* is given by

$$diag : \ X \times Y \to \mathcal{P}(C)$$

Much interest has been spent on inverse problems in the linear case [CW84, Ise84, Mas86, EM94]. But the closed loop of the control unit with its environment leads to a non-linear behavior in the case of hybrid systems. Because there exist only domain specific solutions in the case of non-linear inverse problems, we restrict to the diagnosis of tank ballast systems as introduced in [DBMB93].

In [Dav84] it is shown how to obtain a diagnosis of single faults in the case of a discrete configuration space C as for integrated circuits. A generalization to multiple faults is given in [Rei87], improved in [dKMR92] and put into practice in [MNTMQ96]. In [dKW87] and [FS92] strategies to select the sensor positions are introduced. The need of a quantitative system model is shown in [DK89].

The diagnosis of tank ballast systems leads to a continuous configuration space C. This is caused by the fact, that one is not only interested in determining *which* defects are present, but also wants to know the *degree* of these defects in order to decide whether the current operation has to be stopped or not. A defect of a system may occur suddenly or in a smooth way. The determination of the degree of a fault is also necessary to detect smooth defects early.

In this paper, we present a two level approach for the diagnosis problem. At the first level the defects that may cause the measured behavior are determined. In the second level the degrees of these defects are calculated.

Not only single values as a snapshot of the system are considered to perform diagnosis, but functions over a space of time that describe the behavior. Therefore, all available information is used.

The methods are based on two properties of the domain. As the first property, different degrees of the same defect result in the same effects. This is used to combine system configurations of different degrees of the same defect into one fault candidate. The second property is the monotony of the system behavior – with increasing degree of a fault its effect increases too. This property is used to determine the degree with a binary search. In the domain of tank ballast systems these two propertys are not always but generally fulfilled. We will discuss how to treat violations. This may exponentially increase the runtime of the methods. We will show how to slow down the exponential growth with the help of penalty functions.

In section 2, the domain of tank ballast systems is introduced and a clear example is extracted. In section 3, the approach is sketched with some idealizations. In section 4, we discuss the problems if these idealizations do not hold. In the last section, we give some experimental results to demonstrate the applicability of the methods.

Fig. 2. The Brent Spar

2 Domain

Tank ballast systems are used in huge ships or swimming platforms as the *brent spar* shown in figure 2 or the *micopery 7000* which can crane up loads of a weight up to 14 000 tons. To keep the balance it contains 57 ballast tanks that can be filled by sea water with the aid of two strong pumps. It is also possible to change the center of gravity by pumping water from some ballast tanks to others. These operations are directed by an electronical control station. Defects in the ballast tank system may lead to heavy trouble including the sinking of the platform. Therefore, automatic diagnosis tools are needed to assist the control officer.

Tank ballast systems consists of very different components such as tanks, pipes, valves, filters, pumps, vent pipes or float switches. Some typical defects result in a plugging of pipes or filters. Others cause leaks in the pipes, leaking valves or some air in a pump. We want to determine not only which of these defects may be present in the system but the degree of the defect as it may be necessary to know when the control officer has to decide if the current operation has to be stopped or may be continued.

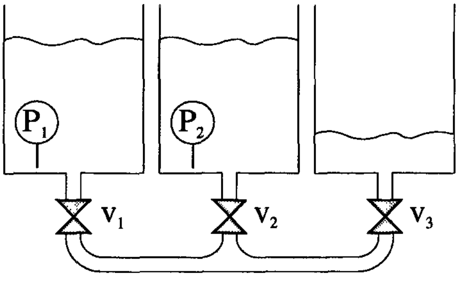

Fig. 3. Running Example

We extract a clear example of that domain. According to figure 3, it contains three tanks that are connected via some pipes and valves. The pressure inside of two of the tanks is measured by two pressure sensors.

An input $x \in X$ consists of the initial water height in the tanks and the positions of the three valves. Therefore, $X := \mathbb{R}^3 \times \{\mathit{open}, \mathit{closed}\}^3$.

Let the possible defects of the system be arbitrary shifts of the valve positions. Therefore, the number n of possible defects is $n = 3$ and $C := [0,1] \times [0,1] \times [0,1]$. The configuration $(0,0,0) \in C$ corresponds to the nominal value when there is no fault in the system.

We observe the two sensors P_1 and P_2 over the interval $[t_{start}, t_{end}]$ of time. Therefore, the sensors result in $y_1, y_2 \in M := \{f : [t_{start}, t_{end}] \to \mathbb{R}\}$ and the output space is $Y := M^2$.

We assume to have a simulator $sim : X \times C \to Y$ that computes the progress of the pressure corresponding to an input x and a fault parameter λ. We denote $(y_1^{x,\lambda}, y_2^{x,\lambda}) := sim(x, \lambda)$, where $y_1^{x,\lambda}$, $y_2^{x,\lambda}$: $[t_{start}, t_{end}] \to \mathbb{R}$.

Our task is to determine which valve positions are shifted and to compute the degree of these shifts. We assume that the input $x \in X$ and a measurement $\hat{y} = (\hat{y}_1, \hat{y}_2) \in Y$ of the sensors are known. That means, we try to decide whether $\lambda_i \neq 0$ and we try to compute the value of λ_i for $i \in \{1,2,3\}$ and a $\lambda = (\lambda_1, \lambda_2, \lambda_3) \in C$ with

$$(y_1^{x,(\lambda_1,\lambda_2,\lambda_3)}, y_2^{x,(\lambda_1,\lambda_2,\lambda_3)}) = (\hat{y}_1, \hat{y}_2) \tag{1}$$

We restrict to one single input x corresponding to figure 3 and therefore define $x := (1, 1, 0.4, \mathit{open}, \mathit{open}, \mathit{open})$. Let the product of the liquid density and the gravitational constant be 10000 Pa per meter. Then, in the fault free case the pressure of both sensors goes down from 10000 Pa to 8000 Pa. That is the pressure in the situation that all three tanks have the same water height. According to figure 4, the measured behavior $\hat{y} := (\hat{y}_1, \hat{y}_2)$ differs from the nominal value $(y_1^{x,(0,0,0)}, y_2^{x,(0,0,0)})$. We are looking for a diagnosis to explain this difference.

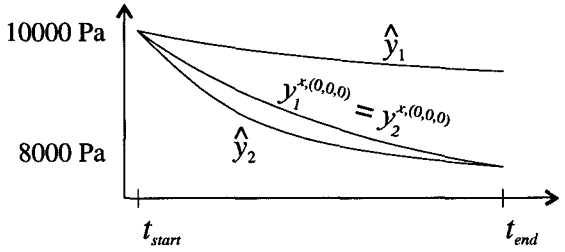

Fig. 4. Nominal value $\left(y_1^{x,(0,0,0)}, y_2^{x,(0,0,0)}\right)$ and measured behavior $\hat{y} = (\hat{y}_1, \hat{y}_2)$

3 Idealized Diagnosis

In this section, we present an approach for the diagnosis problem assuming some idealizations. In the next section, we will discuss how to use it to compute a real diagnosis.

A fault candidate $fc \in Cand := \mathcal{P}\big(\{1, 2, \ldots, n\}\big)$ is a selection out of the n possible defects. As a simplification one may be interested only in single or double faults. The set of fault candidates is then reduced to $Cand_1 := \{1, 2, \ldots, n\}$ or $Cand_2 := \big\{\{a, b\} \mid 1 \leq a < b \leq n\big\}$.

Our interest lies in the set of fault candidates that may explain the measurements as well as in the real valued degree of the defects of these fault candidates. This suggests a two level approach. At the first level, we try to eliminate as many fault candidates as possible with qualitative arguments. In the second level we either determine the degrees of the remaining faults or further eliminate them with quantitative methods.

3.1 First Level

At the first level, we use qualitative attributes over the behavior of the system, that

- are hopefully invariable against the degree of the faults,
- but vary for different fault candidates.

Such attributes may be the sign, the monotony or the curvature of a measurement $m \in M$ or the comparison of the *size* of two measurements $m_1, m_2 \in M$. The size $\|m\|$ of a measurement $m \in M$ is defined by

$$\|m\| := \int_{t_{start}}^{t_{end}} |m| \, dt \tag{2}$$

This implies a partial ordering of M by $m_1 < m_2 \; :\Longleftrightarrow \; \|m_1\| < \|m_2\|$.

A *spot* denotes a sequence of such attributes and a *profile* denotes the evaluation of a spot. Now, we are able to describe the qualitative fault reduction:

1. Choose a spot sp.
2. Determine the profile of each fault candidate with respect to sp.
3. Determine the profile $\hat{p}$ of the measurement $\hat{y}$.
4. Eliminate all fault candidates with a profile unequal to $\hat{p}$.

To give an example, we consider the system introduced in the last section. To keep it simple, we restrict to the fault candidates $\{fc_1, fc_2, fc_3\}$ of the three single faults.

1. Let $Comp(m_1, m_2)$ denote the comparison of the size of the measurements m_1 and m_2. Then, we choose the spot sp by

$$sp := \left(Comp(\hat{y}_1, y_1^{x,(0,0,0)});\ Comp(\hat{y}_2, y_2^{x,(0,0,0)})\right)$$

2.fc_1: If valve v_1 is shifted, there is a bigger resistance for the water. Therefore, the flow out of the first tank is lower and the measurement of P_1 is greater than expected. As a further consequence the flow out of the second tank is higher and the measurement of P_2 is smaller than expected. To conclude, we get the profile $p_1 := (greater, smaller)$.

fc_2: If valve v_2 is shifted, we get the profile $p_2 := (smaller, greater)$ because of a symmetry to the fault candidate fc_1.

fc_3: If valve v_3 is shifted, the water of the first and the second tank can not sink as fast as expected. Thus, we get the profile $p_3 := (greater, greater)$.

3. According to figure 4, the measured behavior $\hat{y}$ implies the profile $\hat{p} := (greater, smaller)$.
4. Because $\hat{p} \neq p_2$ and $\hat{p} \neq p_3$, the only remaining fault candidate is the first one.

3.2 Second Level

There are two motivations for the need of a second level. In the first place the information known from the first level is not enough and one may need the knowledge of the degree of a fault – for example to determine if the current operation has to be stopped or may be continued. And in the second place the qualitative attributes of the first level could not be able to eliminate all but one fault candidate. In that case we need the second level to further reduce the set of fault candidates.

A fault candidate given by the first level corresponds to a subspace $S := [\mu_1, \mu'_1] \times [\mu_2, \mu'_2] \times \cdots \times [\mu_n, \mu'_n] \subset C$ of the configuration space C, where $0 \leq \mu_i \leq \mu'_i \leq 1$ for $1 \leq i \leq n$. A configuration $(\nu_1, \nu_2, \ldots, \nu_n) \in C$ with $\nu_i \in \{\mu_i, \mu'_i\}$ for every $1 \leq i \leq n$ denotes a *corner* of the search space S. The *dimension* $d := \#\{i \mid \mu_i \neq \mu'_i\}$ of a search space S is the number of unfixed parameters.

Now, the task of the second level is to either determine a configuration $\lambda \in S$ for a given search space S that may be responsible for the measurement $\hat{y}$ or to reject S if no such λ exists.

One-dimensional Search Spaces To continue the example, we try to determine the degree of the shift of valve v_1. That means, we search for $\lambda = (\lambda_1, 0, 0) \in S_1$ that fulfills equation (1) where $S_1 := [0,1] \times [0,0] \times [0,0]$.

Considering figure 5 we notice that the effect of a shifted position of valve v_1 increases with an increasing degree of the fault:

$$\mu_1 \leq \mu_1' \Rightarrow \begin{cases} y_1^{x,(\mu_1,0,0)} \leq y_1^{x,(\mu_1',0,0)} \\ y_2^{x,(\mu_1,0,0)} \geq y_2^{x,(\mu_1',0,0)} \end{cases}$$

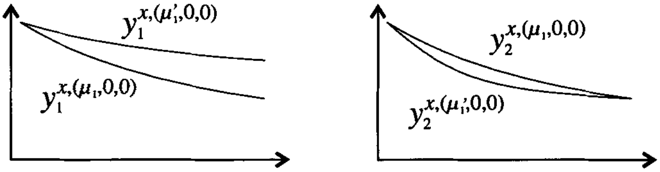

Fig. 5. Monotony of the behavior with $\mu_1 < \mu_1'$

In the case of an existing $\lambda \in S_1$ that fulfills (1), we may derive two bounding conditions:

$$\lambda_1 \in]\mu_1, \mu_1'[\iff y_1^{x,(\mu_1,0,0)} < \hat{y}_1 < y_1^{x,(\mu_1',0,0)} \tag{3}$$

$$\lambda_1 \in]\mu_1, \mu_1'[\iff y_2^{x,(\mu_1,0,0)} > \hat{y}_2 > y_2^{x,(\mu_1',0,0)} \tag{4}$$

These bounding conditions may help to approximate the value of λ_1 with a binary search. Dividing the initial search space S_1 into the two subspaces $S_1^l := [0, \frac{1}{2}] \times [0,0] \times [0,0]$ and $S_1^h := [\frac{1}{2}, 1] \times [0,0] \times [0,0]$, we are able to decide with the two bounding conditions whether $\lambda \in S_1^l$ or $\lambda \in S_1^h$. Thus, we may choose the one that contains λ and again divide that search space into two subspaces. If we continue that, we get a $\mathbb{R}$-analytic computation in the sense of [HVS95] for which the corners of the required subspaces approximate λ_1.

If there is no $\lambda \in S_1$ with property (1), it is possible that one of the bounding conditions (3) and (4) does not hold for any of the two subspaces or that they hold in different subspaces. In one of theses cases the binary search may be stopped and the corresponding fault candidate may be eliminated. But it is also possible that none of these cases appears and the binary search approximates a value λ' with $(y_1^{x,\lambda'}, y_2^{x,\lambda'}) \neq (\hat{y}_1, \hat{y}_2)$.

To explain this, we consider M and define an equivalence relation $\sim$ of non comparable elements in M by

$$m_1 \sim m_2 :\iff \int_{t_{start}}^{t_{end}} |m_1|\, dt = \int_{t_{start}}^{t_{end}} |m_2|\, dt$$

This implies a residual class $M_{/\sim}$ that may take over the ordering $<$ of M in a well defined way.

$$\begin{aligned}[m_1] &:= \{m_2 \in M \mid m_1 \sim m_2\} \\ M_{/\sim} &:= \{[m_1] \mid m_1 \in M\} \\ [m_1] < [m_2] :&\Longleftrightarrow m_1 < m_2\end{aligned}$$

The decisions of the binary search described above are all made by comparisons of the size of the behavior at corners of subspaces with the size of $\hat{y}_1$ and $\hat{y}_2$. Therefore, these decisions are the same for all measurements $([\hat{y}_1], [\hat{y}_2])$ and the binary search approximates a $\lambda' \in S_1$ with

$$(y_1^{x,\lambda'}, y_2^{x,\lambda'}) \in ([\hat{y}_1], [\hat{y}_2])$$

This implies the need to check whether the result of the binary search is a diagnosis of the measurement $\hat{y}$ or not.

Two-dimensional Search Spaces The search space S of a fault candidate corresponding to a double fault is two-dimensional. To determine $\lambda \in S$ we adapt the binary search of the one-dimensional case. Now, the search spaces are divided into four subspaces. For each of these we have to decide whether to continue the search in the subspace or to discard it. Again, these decisions are made by bounding conditions that we conclude from an observed monotonic behavior of the system.

Generalization To avoid restriction to the example we give a general bounding condition for systems with monotonic behavior.

> Let $y^{x,\lambda} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_r)$ for $x \in X$ and $\lambda \in C$ and let $S \subseteq C$ be a search space of arbitrary dimension $d \leq n$ with $\lambda \in S$. Then, for each measurement $\hat{y}_i$ with $1 \leq i \leq r$ there are two corners $\nu_{\min}, \nu_{\max} \in S$ of the search space S so that $y^{x,\nu_{\min}} \leq \hat{y}_i \leq y^{x,\nu_{\max}}$.

To see this, one should imagine to reach the corners $\nu_{\min}$ and $\nu_{\max}$ starting from λ by successively moving a single parameter λ_i in the direction of one of its two limits μ_i or μ'_i. In doing so, the behavior is changed monotonical. With the right choice between μ_i and μ'_i it is always possible to increase or decrease the behavior. If done for all n parameters, the corners $\nu_{\min}$ and $\nu_{\max}$ of the search space S are found.

As the bounding condition holds for search spaces of arbitrary dimensions, we may search for arbitrary multiple faults in systems with monotonic behavior.

Concluding, we sketch the analytical algorithm. Its input is a search space $S \subseteq C$, a measurement $\hat{y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_r) \in Y$ and the system input $x \in X$.

1. Let $\mathcal{S}$ be a set of search spaces, initialized with $\{S\}$

2. If $\mathcal{S} = \emptyset$ then discard the search. No solution $\lambda \in S$ can be found.
3. Delete a search space S' from $\mathcal{S}$ and create the subspaces of S'.
4. For each of these subspaces check the bounding conditions for the measurement $\hat{y}$. Include those subspaces in $\mathcal{S}$ that fulfill the bounding conditions.
5. Continue the calculation with the second step.

If the search is not discarded, the corners of the search spaces in $\mathcal{S}$ converge to a configuration $\lambda \in S$. If the behavior $y^{x,\lambda}$ of the system corresponding to λ is equal to the measurement $\hat{y}$, then a diagnosis is found. As the behavior of the system is monotonical, all diagnoses for $\hat{y}$ build a connected set including λ.

4 Real Diagnosis

In the last section, we simplified the task of diagnosis. What has to be considered if one wants to use the developed ideas for a real diagnosis problem?

1. Sensors do not output the real valued function of behavior of the measured variable in a time interval, but a sequence of single values as points of support of it.
2. - A model of a system is never complete.
   - Due to numerics, the simulation of that model is not an exact but a rough computation.
   - The sensors only have finite precision.

   This leads to an unavoidable difference between the calculated and the measured behavior even in the case of a fault free system. Such differences are called *noise.*
3. In systems bigger than the used example the observed monotony of the behavior may be violated.

These facts lead to some consequences:
From 1: Instead of $M := \{f : [t_{start}, t_{end}] \to \mathbb{R}\}$ we define $M := \mathbb{R}^s$. Now, we need a new definition for the size of a measurement $m = (m^1, m^2, \ldots, m^s) \in \mathbb{R}^s$. Therefore, equation (2) is substituted with some $p \in \mathbb{N}$ by

$$||m||_p := \sqrt[p]{\sum_{i=1}^{s} |m^i|^p} \tag{5}$$

From 2: In the presence of noise we are not able to find a configuration λ with a behavior $y^{x,\lambda}$ equal to the measurement $\hat{y}$ but only similar to it. Therefore, equation (1) has to be changed. With a proper definition of the relation $\approx$ we get

$$y^{x,\lambda} \approx \hat{y} \tag{6}$$

From 2: Noise may change the value of the attributes used in the first level and therefore leads to wrong results.
From 2 and 3: Noise and violations of the monotony of the behavior may violate the bounding conditions used in the second level and therefore lead to wrong decisions in the search so that the solution λ is not found.

Remedy To deal with noisy measurements, we need a measure $L \in [0,1]$ of the precision we use. The extreme value of $L = 0$ stands for absolute precision without any noise and $L = 1$ means, that there is no precision at all and the noise dominates.

$$m_1 \approx_0 m_2 \iff m_1 = m_2$$

$$m_1 \approx_1 m_2 \qquad \forall m_1, m_2 \in M$$

Now, we define a continuous change from one extreme to the other. Thereby, we distinguish absolute similarity $\approx^a$ and relative similarity $\approx^r$. The latter one takes the size of the measurements into account.

$$a \approx_L^a b :\iff (1-L)\,||a-b|| \;\leq\; L$$

$$a \approx_L^r b :\iff ||a-b|| \;\leq\; L\,(||a||+||b||)$$

This similarity implies a robustness against noise.

> An attribute of a measurement m is said to be ***robust*** against noise of level L, iff it holds for all measurements similar to m with level L.

For some attributes, like the monotony or the curveness of a measurement and for the comparison of the size of two measurements we developed methods to determine the robustness. From that, we gain robustness of the first level of the approach.

At the second level, the robust comparison of two measurements prevents us from wrong decisions whether a subspace should be discarded or not. The price is a possibly exhaustive search instead of a binary search. For the same price violations of monotony of the behavior may be handled by a system of penalties. A penalty $\pi(S, \hat{y}, i)$ of a subspace S expresses the belief that S contains a diagnosis of the measurement $\hat{y}_i \in M$ of the sensor $i \in \{1, 2, \ldots, r\}$.

For an exhaustive search, the number of search spaces inquired may increase exponentially. Therefore, we examined different methods to calculate the penalties with the goal to get robustness against noise and violation of monotony while dealing with a tractable number of search spaces. Best results could be achieved with a penalty function $\pi(S, \hat{y}, i)$ that uses the distances $||\hat{y}_i - y_i^{x,\nu}||$ for all corners ν of the subspace S in relation to the activity of S. The activity of a subspace S is defined by the maximum distance $||y_i^{x,\nu} - y_i^{x,\nu'}||$ for all corners ν, ν' of S. As one can see in the next section, the results are nearly as good as the unreachable optimal search. Thus, the exhaustive search could be avoided.

5 Results

The power of the new approach is demonstrated by some empirical examinations we made for a system much more complex than the example introduced. It also consists of three tanks, but in addition there are more pipes and valves, floating

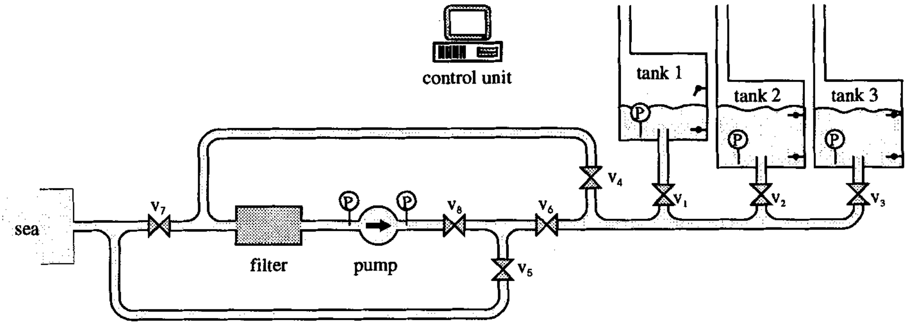

Fig. 6. Scheme of a ballast tank system

switches, bend pipes, a pump, a filter, some additional pressure sensors, a throttle valve and a control unit (figure 6). We designed a fault model that pays attention to the typical defects given by a manufacturer of ballast systems and randomly created 3000 single and double faults with additional noise. An implementation of the two-level approach was able to diagnose all these faults correctly. In about one of three cases the first level leads to a minimal diagnosis in the sense, that for each resulting fault candidate there exists a configuration that may explain the measurements. In the remaining cases the second level was able to eliminate the additional fault candidates for which there exists no such configuration. It appeared that such a elimination is nearly as costly as the computation of a configuration that is a diagnosis. We also omitted the first level and only used the second level. In that case, the runtime roughly doubles.

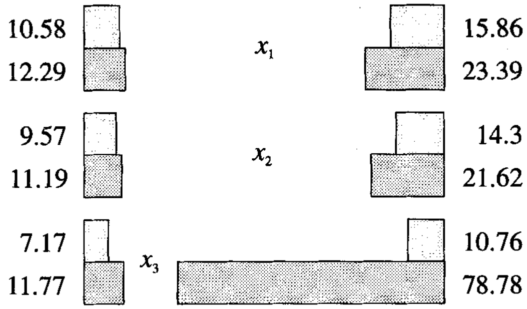

Fig. 7. Mean number of inquired search spaces for three different operations

Figure 7 compares the mean number of inquired search spaces at the second level to the one of a unknown optimal search process. Optimal in the sense that it always knows which subspaces to discard or not. Therefore, it only inquires the minimal number of search spaces needed. The figure distinguishes between three different inputs $\{x_1, x_2, x_3\}$ and the single faults (on the left side) and double faults that are injected. It demonstrates that the approach is very near to the unknown optimal strategy.

The success of the approach for that fixed example encouraged us to start developing a design tool for ballast tanks. With the help of that tool we will test the approach for arbitrary ballast tank systems. We hope that we then will be able to automatically select the sensors in a new system design that are useful to simplify the computation of its diagnoses. The key word of these future plans is *design for diagnosis*.

References

[CW84] E. Y. Chow and A. S. Willsky. Analytical redundancy and the design of robust failure detection systems. *IEEE Transactions on Automatic Control*, 29:603–614, 1984.

[Dav84] R. Davis. Diagnostic Reasoning based on structure and behavior. *Artificial Intelligence*, 24:347–410, 1984.

[DBMB93] O. Dressler, C. Böttcher, M. Montag, and A. Brinkop. Qualitative and Quantitative Models in a Model-based Diagnosis System for Ballast Tank Systems. In *Proc. of the International Conference on Fault Diagnosis TOOLDIAG '93, Toulouse, France*, pages 397–405, 1993.

[DK89] D. Dvorak and B. Kuipers. Model-Based Monitoring of Dynamic Systems. In *Proceedings IJCAI*, pages 1238–1243, 1989.

[dKMR92] J. de Kleer, A. K. Mackworth, and R. Reiter. Characterizing diagnosis and systems. *Artificial Intelligence*, 56:197–222, 1992.

[dKW87] J. de Kleer and B. Williams. Diagnosing multiple faults. *Artificial Intelligence*, 32:97–130, 1987.

[EM94] H. W. Engl and J. McLaughlin. *Inverse Problems and Optimal Design in Industry*. Teubner press, 1994.

[FS92] B. Faltings and P. Struss, editors. *Recent Advances in Qualitative Physics*, chapter Sensor Selection in Complex System Monitoring Using Information Quantification and Causal Reasoning. MIT press, 1992.

[HVS95] G. Hotz, G. Vierke, and B. Schieffer. Analytic machines. Technical Report TR95-025, Electronic Colloquium on Computational Complexity (http://www.eccc.uni-trier.de/eccc), 1995.

[Ise84] R. Isermann. Process fault detection based on modeling and estimation methods: A survey. *Automatica*, 20:387–404, 1984.

[Mas86] M. A. Massoumnia. A geometric approach to the synthesis of failure detection filters. *IEEE Transactions on Automatic Control*, 31:839–846, 1986.

[MNTMQ96] R. Milne, C. Nicol, L. Travé-Massuyès, and J. Quevedo. TIGER: knowledge based gas turbine condition monitoring. *AI Communications*, 9:92–108, 1996.

[Rei87] R. Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32:57–96, 1987.

Author Index

Apte 17
Ayel 459
Backer 479
Bariani 223
Bellazzi 287
Berry 53
Berthold 347
Bossley 313
Boyd 585
Bradley 561
Bramer 527
Brown 313
Caelli 161
Cebrián 491
Cheng 549
Cohen 41, 77, 211
Cornish 65
Cucchiara 223
Darlington 437
Easley 561
Eijkel, van den 479
Erhart 513
Euzenat 259
Fackler 111
Famili 173
Gama 187
García 491
Glover 149
Gough 53
Gregory 77
Gunn 313
Guo 437
Hahn 409
Hand 1
Haring 199
Ho 235
Hong 17
Hosking 17
Hotz 597
Howe 573
Huber 347
Jensen 211
Kamp 299
Kharin 501
Klawonn 359
Klement 359
Kok 199
Kranz 371
Lam 161
Lapp 371
Larizza 287
Lepre 17
Levinson 89
Liu, W. 527
Liu, X. 549
Lo 173
Lowden 447
Lu 465
Lubbe, van der 479
McGeoch 41
McSherry 397
Mello 223
Mills 337
Moura–Pires 325
Muller 29
Nascimento 325
Oates 211
Palmer–Brown 337
Pednault 17
Piccardi 223
Precup 41
Ramer 123
Ramoni 537
Riva 287
Roadknight 337
Robinson 447
Rodríguez 491
Rosen 17
Schädler 137
Schieffer 597
Schikuta 513
Schlosser 423
Schnattinger 409
Scott 235
Sebastiani 537
Sharpe 149
Shi 465
Siklóssy 459
Someren, van 247
Somlo 573
St. Amant 101
Sutiwaraphun 437
Thompson 527
To 437
Torres 247
Tsien 111
Turhan 383
Valtchev 259
Verbruggen 275
Verdenius 247
Verhaegen 275
Vullings 275
West 161
Wezel, van 199
White 527
Wietek 299
Williams 235
Wilkinson 89
Wolkenhauer 609
Wu 549
Wysotzki 137
Yu 123
Zhou 465