Advances in Knowledge-Based Technologies

Proceedings of the
Master and PhD Seminar
Summer term 2015, part 1

Softwarepark Hagenberg
SCCH, Room 0/2
8 May 2015

Software Competence Center Hagenberg
Softwarepark 21
A-4232 Hagenberg
Tel. +43 7236 3343 800
Fax +43 7236 3343 888
www.scch.at

Fuzzy Logic Laboratorium Linz
Softwarepark 21
A-4232 Hagenberg
Tel. +43 7236 3343 431
www.flll.jku.at

Program

Chair: Susanne Saminger-Platz

9:00 Robert Pollak and Roland Richter:
Fingerprint indexing via BRIEF minutia descriptors
9:30 Patrick Traxler, Pablo Gómez, Tanja Grill:
A Robust Alternative to Correlation Networks for Identifying Faulty Systems
10:00 Mario Pichler:
Bayesian Networks: A short intro and exemplary use cases

Fingerprint indexing via BRIEF minutia descriptors

Robert Pollak and Roland Richter
Department of Knowledge-Based Mathematical Systems
Johannes Kepler University, Linz, Austria
robert.pollak@jku.at, roland.richter@jku.at

April 21, 2015

Abstract

We use BRIEF binary local image descriptors as minutia descriptors for indexing of biometric fingerprint databases. Tests with varying descriptor size and parametrization are performed on a proprietary database. Compared with the speed of a proprietary implementation of conventional minutiae matching, we find that BRIEF descriptors are fast enough for database indexing. The tested descriptors outperform two other image descriptors (LBP, HoG) from recent literature with respect to matching rates and average penetration rates.

A Robust Alternative to Correlation Networks for Identifying Faulty Systems

Patrick Traxler [1] and **Pablo Gómez**[2] and **Tanja Grill**[1]
[1]Software Competence Center Hagenberg, Austria
e-mail:patrick.traxler@scch.at
tanja.grill@scch.at
[2]Institute of Applied Knowledge Processing, Johannes Kepler University, Linz, Austria
e-mail: pablo.gomez@faw.jku.at

Abstract

We study the situation in which many systems relate to each other. We show how to robustly learn relations between systems to conduct fault detection and identification (FDI), i.e. the goal is to identify the faulty systems. Towards this, we present a robust alternative to the sample correlation matrix and show how to randomly search in it for a structure appropriate for FDI. Our method applies to situations in which many systems can be faulty simultaneously and thus our method requires an appropriate degree of redundancy. We present experimental results with data arising in photovoltaics and supporting theoretical results.

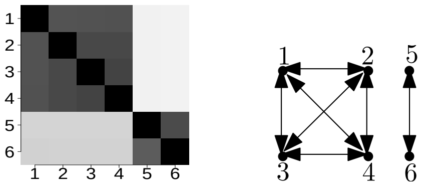
1 Introduction

The increasing number of technical systems connected to the Internet raises new challenges and possibilities in diagnosis. Large amount of data needs to be processed and analyzed. Faults need to be detected and identified. Systems exist in different configurations, e.g. two systems of the same type that have different sets of sensors. Knowledge about the system design is often incomplete. Data is often unavailable due to unreliable data connections. Besides these and other difficulties, the large amount of data also opens new possibilities for diagnosis based on machine learning.

The idea of our approach is to conduct fault detection and identification (FDI) by comparing data of similar systems. We assume to have data of machines, devices, systems of a similar type and want to know if some system is faulty and if so, to identify the faulty systems. This situation may deviate from classic diagnosis problems in that we just have limited information (e.g. sensor or control information) of system internals. Moreover, we may have incomplete knowledge about the system design. This makes manual system modeling hard or even impossible. The problem is then to compare the limited information of the working systems (perhaps only input-output information) to identify faulty systems.

In this work we tackle one concrete problem of this kind. It is motivated by photovoltaics. We describe it in more detail below. The problem that arises in our and other applications is that not every two systems can be compared. We thus need to learn relations between systems.

There are different approaches to learn structure, e.g. learning Bayesian networks, Markov random fields, or similar concepts. The concept that fits our needs are correlation networks. A correlation network is some structure in the correlation matrix, e.g. a minimum spanning tree or a clustering. In our application we have n variables which represent the produced energy per photovoltaic system. Given that a single system correlates strongly with enough other systems, we use this information for FDI via applying a median.



(a) Fitness matrix (b) Digraph

Figure 1: Learning relations between 6 systems. We draw an edge between two systems if there is a strong linear relation between them. First, we compute the fitness matrix, 1(a), our robust alternative to the sample correlation matrix. Darker colors mean a stronger linear relation. Going from Fig. 1(a) to 1(b) is a discretization step via thresholding. The digraph is the input for conducting FDI.

We can also think of correlation networks as a method for knowledge discovery. It has been applied in areas such as biology [7; 4] and finance [5] to analyze gene co-expression and financial markets. In our situation, the first step is to learn linear relations between systems. For learning we need historical data. A sample result of this step is depicted in Fig. 1. In Fig. 1(a) the fitness matrix, our robust alternative to the correlation matrix, is shown. It represents the degree of linearity between any two systems. For FDI, the second step of our method, we work with the result as depicted in 1(b) and current data. In the example, we derive for every of the six systems an estimation $\hat{m}_i$ of its current value y_i from its neighbors current values, e.g. for system 1 we get an estimate from the current values of the systems $2, 3, 4$ and for system 5 from system 6. Finally, we test for a fault by checking if $|\hat{m}_i - y_i|$ is large.

The major difficulty we try to tackle with this approach is the presence of many faults. Faults influence both the learning problem and the FDI problem. Robustness is an essential property of our algorithms. Our result can be seen as a robust structure learning algorithm for the purpose of FDI. Robustness is a preferable property of many learning and estimation algorithms. However, the underlying optimization problems unlike their non-robust variants are often NP-hard. This is for example the case for computing robust and non-robust estimators for linear regression, e.g. Least Median of Squares versus Ordinary Least Squares [6]. We avoid NP-hardness by a careful modeling of our problem. In particular, our algorithms are computationally efficient. Under some conditions, FDI can be done in (almost) linear time in the number of systems n.

To summarize our contributions, we introduce a novel alternative to the sample correlation matrix and present a first use of it to discover structure appropriate for general FDI and in particular for identifying faulty photovoltaic systems. Our method works in the presence of many faults. Our algorithms are computationally efficient. Our method incorporates a couple of techniques from machine learning and statistics: (Repeated) Theil-Sen estimation for robust simple linear regression. Trimming to obtain a robust fitness measure. Randomized subset selection for improved running time. And a median mechanism to conduct FDI.

1.1 Motivating Application: Identifying Faulty Photovoltaic Systems

Faults influence the performance of photovoltaic systems. PV systems produce less energy than possible if faults occur. We can distinguish between two kinds of faults. Faults caused by an exogenous event such as shading, (melting) snow, and tree leafs covering solar modules. And faults caused by endogenous events such as module defects and degradation, defects at the power inverter, and string disconnections.

We are going to detect faults by estimating the drop in produced energy. Most of the common faults result in such a drop. The particular problem is given by the sensor setup. We just assume to know the produced energy and possible but not necessarily the area (e.g. the zip code) where the PV system is located.

We apply our method to PV system data. Difficulties in the application are different system types and deployments of systems. For example, different number of strings and modules per string and differing orientation (north, west, south, east) of the modules. Moreover, the lack of information due to the lack of sensors and incomplete data due to unreliable data connections. Faults occur frequently, in particular exogenous faults during winter.

The novelty of our work in the context of photovoltaics is that it works in an extremely restrictive sensor setting. To the best of our knowledge, we are the first to consider this restrictive sensor setting. We only need to know the produced energy of a PV system. There is also the implicit assumption, which is tested by the learning algorithm, that the systems are not too far from each other so that we can observe them in similar working (environmental) conditions. Distances of a couple of kilometers are possible. Systems which are very close to each other and have the same orientation such as systems in a solar power plant yield the best results.

1.2 Related Work

Correlation networks have applications in biology and finance. See e.g. [5; 7; 4] and the references therein. In biology [7; 4], they are applied to study gene interactions. The correlation matrix is the basis for clustering genes and the identification of biologically significant clusters. In [7; 4], a scale-free network is derived via the concept of topological overlap. Scale-free networks tend to have few nodes (genes) with many neighbors, so called hubs.

Correlation networks are primarily used for knowledge discovery. In particular, concepts such as clusters, hubs, and spanning trees are interpreted in the context of biology and finance. In our work, we introduce a robust alternative to correlation networks.

Other structural approaches are based on Bayesian networks, Markov random fields and similar concepts. Gaussian Markov random fields are loosely related to correlation networks. Their structure is described by the precision matrix, the inverse covariance matrix (ch. 17.3, [3].)

Another structural approach is FDI in sensor networks [2; 1; 8; 9]. The current approach [2; 1; 8] mainly deals with wireless sensor networks. The algorithms usually use the median for FDI such as we do. The difference is that FDI in wireless sensor networks uses a geometric model similar to interpolation methods. It requires the geographic location of the sensors. It is assumed that two sensors close to each other have a similar value. This cannot be assumed in general. To overcome these problems of manual modeling, we apply machine learning techniques.

References

[1] Jinran Chen, Shubha Kher, and Arun Somani. Distributed fault detection of wireless sensor networks. In *Proc. of the 2006 Workshop on Dependability Issues in Wireless Ad Hoc Networks and Sensor Networks*, pages 65–72, 2006.

[2] M. Ding, Dechang Chen, Kai Xing, and Xiuzhen Cheng. Localized fault-tolerant event boundary detection in sensor networks. In *Proc. of the 24th Annual Joint Conference of the IEEE Computer and Communications Societies*, volume 2, pages 902–913, 2005.

[3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer, 2008.

[4] Steve Horvath. *Weighted network analysis: applications in genomics and systems biology*. Springer Science & Business Media, 2011.

[5] Dror Y. Kenett, Michele Tumminello, Asaf Madi, Gitit Gur-Gershgoren, Rosario N. Mantegna, and Eshel Ben-Jacob. Dominating clasp of the financial sector revealed by partial correlation analysis of the stock market. *PLoS ONE*, 5(12):e15032, 12 2010.

[6] Peter J Rousseeuw and Annick M Leroy. *Robust regression and outlier detection*, volume 589. John Wiley & Sons, 2005.

[7] Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(17), 2005.

[8] Chongming Zhang, Jiuchun Ren, Chuanshan Gao, Zhonglin Yan, and Li Li. Sensor fault detection in wireless sensor networks. In *Proc. of the IET International Communication Conference on Wireless Mobile and Computing*, pages 66–69, 2009.

[9] Yang Zhang, N. Meratnia, and P. Havinga. Outlier detection techniques for wireless sensor networks: a sur-

vey. *Communications Surveys and Tutorials, IEEE*, 12(2):159–170, 2010.

BAYESIAN NETWORKS: A SHORT INTRO AND EXEMPLARY USE CASES

Mario Pichler
Software Competence Center Hagenberg GmbH (SCCH), Hagenberg, Austria

Credible computerized approaches to, e.g., decision making or situation/risk assessment strongly depend on exploitable expert knowledge. A key problem, however, is to find a suitable knowledge representation method that a) is easy to understand and usable by domain experts of different disciplines, and b) is seamlessly usable by computer-based reasoning techniques.

This talk introduces a promising approach of formal knowledge representation. We are modeling domain knowledge by means of Probabilistic Graphical Models. Especially, we are investigating Bayesian Networks for modeling mutual influences of factors originating from heterogeneous data sources including implicit knowledge of domain experts, and the integration of associated uncertainties in a single model.

Three use cases of formal representations of expert knowledge by means of Bayesian Networks for decision support and situation/risk assessment are presented: a) a formal model of the *Stop of Go*© avalanche decision strategy, b) environmental modeling for early warning systems, and c) tourism knowledge model generation.

References

Antonucci, A.; Salvetti, A. & Zaffalon, M. (2004): Hazard assessment of debris flows by credal networks, Technical report IDSIA-02-04, IDSIA, available online: http://www.idsia.ch/idsiareport/IDSIA-02-04.pdf.

Bayes, T. (1763): An Essay towards Solving a Problem in the Doctrine of Chances. Philosophical Transactions, 53:370–418.

Conrady, S. & Jouffe, L. (2013). Tutorial on Driver Analysis and Product Optimization with BayesiaLab. Available online: http://library.bayesia.com/display/whitepapers/Driver+Analysis+and+Product+Optimization [last access: 2014/09/06].

Korb, K.B. & Nicholson, A.E. (2010): Bayesian Artificial Intelligence. CRC Press, 2. Ed.

Larcher, M. (1999): Stop or Go: Entscheidungsstrategie für Tourengeher. Berg&Steigen, 99(4):18–23.

Larcher, M., Mössmer, G. & Würtl, W. (2013): Sicher am Berg: Skitouren Risikomanagement Stop or Go und Notfall Lawine. Österreichischer Alpenverein, Innsbruck.

Lee, K., Lee, H. & Ham, S. (2013): The Effects of Presence Induced by Smartphone Applications on Tourism: Application to Cultural Heritage Attractions . In Xiang, Z. & Tussyadiah, I. (Eds.) Information and Communication Technologies in Tourism 2014, Springer International Publishing, 59-72.

Nunkoo, R. & Ramkissoon, H. (2011): Developing a community support model for tourism. Annals of Tourism Research, 38:964-988.

Pearl, J. (1985): Bayesian networks: a model of self-activated memory for evidential reasoning. In: Cognitive Science Society 1985. UC Irvine, 329–334.

Pichler, M., Eder, S. & Larcher, M. (2014): About Probabilistic Graphical Models in Probabilistic Avalanche Science: The Case of Stop or Go. In: International Snow Science Workshop. Banff, Canada, 1071–1078.

Pichler, M. & Leber, D. (2014): On the formalization of expert knowledge: A disaster management case study. In: 25th International Workshop on Database and Expert Systems Applications - DEXA 2014. Munich, Germany, 149–153.

Pichler, M., Steiner, L. & Neiß, H. (2015): Probabilistic Modelling of Influences on Travel Decision Making. e-Review of Tourism Research (eRTR). ENTER 2015 Vol. 6 Short Papers.