Introduction

The thesis describes two fundamental contributions towards the induction of predictive models for dynamical systems, based on a pattern recognition approach. One contribution involves the enrichment of an experimental software tool for non-parametric system identification, called SAPS [Cellier 1991]. The other concerns the establishment of a promising link of SAPS' underlying framework, called GSPS [Klir 1985], with a relatively young research domain that has the necessary methodology to aid in GSPS' problem solving. The contributions proposed in this thesis help to overcome the limited power of SAPS to the identification of complex black-box systems. As the thesis demonstrates, the most promising perspectives of achieving a good predictive model, is via the use of data mining methods.
The first contribution remains entirely in the domain of general system theory (GST), from which SAPS originates. The second contribution implies a total paradigm shift that goes back to the re-use of basic principles in GSPS. The second modification transcends the boundaries of GST towards the domain of Knowledge Discovery in Databases (KDD); system identification approach in GST is hereby considered as supervised learning applied to dynamical systems. Consequently, this thesis bridges a gap between the two domains. The GST domain deals with the identification and modelling of dynamical systems, while the KDD domain tries to identify useful and non-trivial patterns from observed, mostly static, data.
In general system theory, many methodologies exist, but one of them, called General System Problem Solving (GSPS) [Klir 1985], is of particular importance in the sub-domain of system identification, because it forms the backbone of SAPS. A central principle of GSPS that maps observed input-output data to a state space via a time-invariant dependency pattern forms the basis of a transformation that rephrases the system identification problem in GST to a data mining problem in KDD. In the latter domain, a plethora of existing techniques can be used to solve the deficiency of SAPS, towards the identification of complex systems, indirectly. This results in an extremely powerful approach to system identification for large complex systems where many variables come into play.
This thesis provides an overview of the two domains involved in two separate parts. Each part moves gradually from the general underlying framework towards a specific implementation tool. The first two chapters of each part are based on current literature. The remaining two chapters in each part are new contributions of the dissertation. A summary of the major contributions of this thesis is found in the conclusion at the end.

Part I: A System Theoretic Approach

Chapter 1 describes the underlying framework for the non-parametric system identification approach in the domain of systems theory. It balances a formalised approach with a more descriptive approach. The chapter moves from General System Theory towards General System Problem Solving. The latter permits non-parametric identification of black-box systems via an inductive approach based on pattern-recognition. The General System Problem Solving framework is elaborated and more formalisations are given. They help to place the SAPS tool, which is described in chapter 2, in the right context. The very important concept of a 'mask'

is introduced and formalised in chapter 1. Throughout the chapter, some initial assumptions with regard to the goal of the dissertation are highlighted.
Chapter 2 elaborates on a popular version of SAPS, which is developed by Cellier [1991] and is called SAPS-II. It has proven to have a good potential for practical applications and as such it will serve as a reference for the research developed in this dissertation. Although SAPS-II is based on GSPS, some methodological differences and restrictions apply that are explained in this chapter. Attaching a quality to a mask provides the possibility to compare masks and to search for the best one. Two existing forecasting methods, which are compared with the new methods introduced in chapter 8, are explained.
Chapter 3 establishes a link between problem types in SAPS and in hidden Markov models via a rigorous formalisation of 'state-transition' matrix construction. It shows that correlation in the system memory does not play a role for SAPS and that induction problems are similar in SAPS as in hidden Markov models [Van Welden 20xx].
The first part of chapter 4 is theoretical and concentrates on the introduction and justification of a newly introduced sub-optimal approach. The order complexity of the subsequent algorithm is not exponential as it is for the existing optimal one, but polynomial. This results in the possibility of tackling systems that are more complex [Van Welden and Vansteenkiste 1996]. The second part of chapter 4 gives a brief overview of the newly designed tool that supports the sub-optimal mask search. This tool, which is called SAPS-ST, is intended as a prototype to explore new model construction techniques [Van Welden and Vansteenkiste 1994]. Compatibility between the new software tool, SAPS-ST, and the existing tool SAPS-II is not described in this dissertation. The reader is referred to [Van Welden 1999].

Part II: A Data Mining Approach

Chapter 5 provides an overview of the emerging domain of Knowledge Discovery in Databases (KDD) and data mining. It follows the same structure as chapter 1. The field of machine learning is briefly touched upon, because some concepts help in understanding the material chapter 5 contains. It shows that system identification by SAPS belongs to a supervised learning paradigm. The latter encompasses classification and regression. A formalisation of classification is undertaken and tree classifiers are introduced at the end of this chapter. Chapter 5 is indispensable for showing the plethora of methods that can be applied when doing system-identification via supervised learning.
Chapter 6 concentrates on tree classifiers and shows their underlying principles. This chapter supplies the knowledge to understand chapter 8 where the use of tree classifiers for SAPS is demonstrated. It elaborates on a specific statistical approach for classification and regression trees on which the software tool CART© resides.
Chapter 7 explicitly establishes the link between GST and KDD. GSPS can be seen as a data-mining approach, and, consequently, it can be embedded in KDD [Van Welden et al. 1998]. Relationships and emphasis shifts are explained at different levels of abstraction. A sharper focus to directed systems (in GST) and to supervised learning (in KDD) is done. The resulting paradigm shift for SAPS is outlined.
Chapter 8 illustrates how to apply data mining to SAPS with the aid of tree classifiers. The latter are chosen because they give comprehensible models and because they have much in common with the extension made in chapter 4. The advantages of using tree classifiers are shown for recoding and for the handling of missing values [Van Welden et al. 1999]. An extra benefit is the ranking of variable importance. In this chapter, the implementation of the nearest neighbour algorithm in SAPS-II is modified and simplified. The new nearest neighbour methods should be seen in the context of data mining.

© CART is a product from Salford Systems Inc., see www.salford-systems.com