Kamran Karimi (Ed.)

Proceedings of the Workshop on Causality and Causal Discovery

In Conjunction with the Seventeenth Canadian Conference on Artificial Intelligence (AI'2004)

London, Ontario, Canada, 16 May 2004

Technical Report CS-2004-02
April 2004

Department of Computer Science
University of Regina
Regina, Saskatchewan
Canada S4S 0A2

ISSN 0828-3494
ISBN 0-7731-047-1

Preface

This volume contains papers selected for presentation at the Workshop on Causality and Causal Discovery, in conjunction with the Seventeenth Canadian Conference on Artificial Intelligence (AI'2004), held in London, Ontario, Canada on 16 May 2004.

Causality and discovering causal relations are of interest because they allow us to explain and control systems and phenomena. There have been many debates on causality and whether it is possible to discover causal relations automatically. Different approaches to solving the problem of mining causality have been tried, such as utilising conditional probability or temporal approaches. Discussing, evaluating, and comparing these methods can add perspective to the efforts of all the people involved in this research area. The aim of this workshop is to bring researchers from different backgrounds together to discuss the latest work being done in this domain.

The occurrence of this workshop is the result of the joint efforts of the authors, the programme committee members, and the Canadian AI organisers. This volume would not have been possible without the help of the members of the programme committee who reviewed the papers attentively. The Canadian AI'2004 organisers, General Chair, Kay Wiese (Simon Fraser University), Program Co-Chairs Scott Goodwin and Ahmed Tawfik (both from the University of Windsor), and Local Organiser Bob Mercer (University of Western Ontario), supported the workshop from the beginning to the end. Thanks to Weiming Shen for hosting the workshops at National Research Council Canada (NRC) facilities and helping with the co-ordination. The Department of Computer Science at the University of Regina, and especially Howard Hamilton contributed their time and resources towards the preparation of this volume. The efforts of all the people not mentioned by name, who in any way helped in making this workshop possible, are greatly appreciated.

April 2004 Kamran Karimi

Editor

Kamran Karimi, University of Regina

Programme committee

Cory Butz, University of Regina
Eric Neufeld, University of Saskatchewan
Richard Scheines, Carnegie Melon University
Steven Sloman, Brown University

Table of contents

$_{CAUSATI}$O^{NT} and DOLCE

Jos Lehmann and Aldo Gangemi

Laboratory for Applied Ontology
Institute of Cognitive Science and Technology
Italian National Research Council
http://www.loa-cnr.it/

Abstract. This paper offers an overview of CausatiOnt, a semi-formal ontology conceived as a basis for (automatic) legal reasoning about causation in fact. Moreover, a preliminary axiomatization in DOLCE upper ontology is provided of part of CausatiOnt. This axiomatization is a step toward making CausatiOnt, or at least part of it, more rigorous and toward enabling the automatic discovery of causal relations in the model of a legal case.

1 Introduction

In the context of a research in Artificial Intelligence and Law (AI&Law), extensively reported in [1] and, more concisely, in [2], the problems posed by the automation of legal responsibility attribution are thoroughly analyzed and (partially) reduced to the problems posed by automatic reasoning about causation. Based on such reduction, the main contribution delivered by this research is an analytical subsumption hierarchy - an ontology, in Artificial Intelligence (AI) terms - which semi-formally represents the knowledge (i.e. the concepts and the conceptual relations) used in the legal domain as the basis for reasoning about causation. We call such ontology CausatiOnt[1].

This paper offers a description of a work in progress, which aims at axiomatizing CausatiOnt within DOLCE upper ontology [3]. This merging is being tried because, despite a preliminary specification in Protégé-2000, CausatiOnt is still too complex for use in automatic reasoning, as it comprises knowledge which is, logically speaking, rather ambiguous. DOLCE, on the contrary, has a well founded first order characterization [4], which may help in making CausatiOnt more rigorous and, therefore, potentially useful for the automatic discovery of causal relations in the model of a legal case. We proceed as follows: section 2 discusses the causal relation typically employed in legal reasoning, causation in fact; section 3 presents the theoretical basis and the class hierarchy of CausatiOnt; section 4 introduces the preliminary results of the axiomatization of CausatiOnt in DOLCE; section 5 draws a conclusion.

[1] From CAUSATIon ONTology.

2 From legal responsibility to causation in fact

Legal Theory provides various arguments (see [2], section 1.1) in favor of the following legal theoretical position: reasoning about the attribution of legal responsibility to a person involved in a case largely rests on causal reasoning. From an AI&Law perspective, this strongly suggests that the automation of legal responsibility attribution in one way or another requires the automation of legal causal reasoning. This may be achieved by adopting, among other things, a suitable ontology of causal concepts, such as the one presented in sections 3 and 4 of this paper.

Before presenting the ontology, we first spend some words on the relation between the notion of legal responsibility and the underlying causal knowledge. This is meant to clarify the nature of such knowledge and of the causal relation that CausatiOnt is meant to capture: causation in fact.

Consider the following example, from [5].

Example 1 (The Desert Traveler). A desert traveler T has two enemies. Enemy 1 poisons T's canteen and Enemy 2, unaware of Enemy 1's action, shoots and empties the canteen. A week later, T is found dead and the two enemies confess to action and intention.

If a jury were asked to attribute the legal responsibility for T's death, it would probably have to consider the following additional information, which is left implicit in Example 1: T never drank from the canteen, T was found dead by dehydration.

Based on such information, the jury would very probably come to an unanimous decision and indicate Enemy 2 as the responsible person for T's death. If asked why, the jury may answer: because Enemy 2 caused T's death. If asked in what sense Enemy 2 caused what he caused, the jury would probably say that Enemy 2's action is a *counterfactual condition* of T's death, which makes it a cause. In other words, had Enemy 2 not shot the canteen, T would still be among us. But this is not true - it should be replied. Had Enemy 2 not shot the poisoned canteen, T would have drunk from it and he would not be among us anyway. Therefore, Enemy 2's action is not a counterfactual condition of T's death. Is it still its cause? - the jury should be asked. Again its answer would probably be unanimous and indicate Enemy 2's action as the cause of T's death in the sense that he is the most *proximate cause* of T's death. If asked to give a definition of such proximity, the jurors would probably give a temporal definition: Enemy 2's action is the latest cause of T's death. But, then again, it could be replied that from a strictly physical point of view the heat of the Sun was definitely a temporally more proximate cause than Enemy 2's action.

This "cat and mouse game" with the jury could go on for a long time because Example 1 is no real-life case. It is just a tricky and underspecified combination of circumstances devised by some smart philosopher on some lazy day, with the explicit purpose of fooling imaginary juries. The example, though, does show the following: a "short circuit" in our causal understanding of a series of events has major consequences on our capacity to attribute (legal) responsibility.

[2] provides a legal theoretical bridge between the legal concept of responsibility and the causal notions that support its attribution. Such bridge consists of five elements: first, the distinction between causation in fact and legal causation; second, the distinction between the ontological problems posed by causation in fact and the procedural problems posed by legal evidence and the burden of proof; third, the definition of legal responsibility in terms of liability and accountability; fourth, the definition of the grounds for legal responsibility attribution, among which causation in fact; fifth, the definition of causation in fact. In the following we briefly illustrate the first and the last of these elements.

The legal language makes a distinction between *causation in fact* and *legal causation*. On the one hand, the problem of causation in fact is the problem of understanding what *actually* happened (i.e. what caused what) in a case. Such factual interpretation is something legal experts usually take for granted and mostly see as unproblematically achieved by common sense. In Example 1 the connection between the shooting of the canteen and T's death by dehydration is an instance of causation in fact, because Enemy 2 had the intention to kill T, he believed that by shooting the canteen T would die (rather than be saved from poisoning), he shot the canteen, T died. On the contrary, the connection between the poisoning of the canteen and T's death is not an instance of causation in fact, because T never drank from the canteen[2]. On the other hand, legal causation is the set of criteria that should be applied either when a clear factual interpretation of the case is missing or when legal policy considerations should be applied, therefore adopting a causal interpretation that is different from the factual causal one. In Example 1, supposing that, after the poisoning but before the shooting of the canteen, T had drunk from it and supposing impossible to establish the temporal priority between the effects of poisoning and the effects of dehydrating on T's body, the attribution of legal responsibility should be based on legal causation (for instance, by accepting that both Enemy 1's and Enemy 2's conducts legally caused T's death).

Now, how to give a sufficiently general *definition of causation in fact*? There are various traditional legal theoretical approaches to the problem of giving this definition, most notably approaches based on the notion of causal proximity or on counterfactuals[3]. Traditional approaches, though, suffer of a lack of an *explicit* account of the elements of a case that a judicial authority should consider when assessing causation in fact. This jeopardizes consistency of application of such tests over large corpora of cases. In order to overcome the common shortcoming of traditional approaches, Hart and Honoré propose in [6] to base legal causal assessment on an explicit definition of causation in fact, like the following one.

Definition 1 (Causation in fact). *Agent A causes an event e, that might involve agent B, if either of the following holds:*

1. *A starts some physical process that leads to e;*

[2] Legally speaking Enemy 1's action may be considered just as an attempt at murdering T.

[3] Typical examples of counterfactual tests used in the legal domain are the *sine qua non* and the *but for* tests. For detailed overviews of these approaches see [2] or [6].

2. *A provides reasons or draws attention to reasons which influence the conduct of B, who causes e;*
3. *A provides B with opportunities to cause e.*
4. *All the important negative variants of clauses 1, 2, 3*

For what concerns Example 1 the causal connection between Enemy 2 shooting and T dying is non linear and may be considered either as a case of the negative variant of clause 1 above (Enemy 2's conduct prevents the physical process of hydration which leads to T's death by dehydration) or as a case of clause 3 above (Enemy 2's conduct provides T with the opportunity of causing his own death by dehydration).
In conclusion, Definition 1 carves a portion of causal knowledge that is very relevant to AI&Law research.

3 An overview of CausatiOnt

In order to make Definition 1 more rigorous and possibly useful to *automatic* classification and/or interpretation, it should be reconfigured along clear ontological lines and restructured by means of a subsumption hierarchy, i.e. a so called is-a hierarchy. This is exactly the original purpose of CausatiOnt, the ontology presented in this section. It should be noticed that the presentation of CausatiOnt given here is rather theoretical. We only occasionally exemplify the intuition behind each newly introduce notion by referring to a subset of Example 1 (namely: E_1 = the bullet is shot; E_2 = the canteen is broken). But neither in this section nor in the following ones do we provide a complete model of E_1, E_2 and of their causal connection, as this would require many more pages than available or a drastic cut in the theoretical treatment of the introduced notions.

3.1 Philosophical preliminaries

The first and most obvious restructuring distinguishes in Definition 1 four main ontological levels, corresponding to four main types of causation, as usually described in the philosophical literature: physical causation, agent causation, interpersonal causation, negative causation[4]. *Physical causation* is described by the final part of clause 1 of Definition 1, where the definition mentions *a physical process that leads to an event. Agent causation* is described by the initial part of clause 1, where Definition 1 mentions *an agent starting a physical process.* The agreement around cases of agent causation is not reached as easily as in cases of physical causation. This is due to the problem of detecting the beliefs,

[4] Distinguishing between varieties of causation is the pragmatic answer of the philosophy of causation to the (temporary?) lack of stable scientific theories of some fundamental phenomena. For instance, without a stable neuropsychological solution of the mind-body problem, it is impossible to choose in a principled way between a reduction of agent causation to physical causation and a reduction of physical causation to agent causation.

desires and intentions of the agent that starts the physical process. Things become even more complex when considering *interpersonal causation*, described by clauses 2 and 3. One might be tempted to consider interpersonal causation just as a subcase of agent causation, where the psychological state of an agent exerts a causal influence on another agent. Things are not that simple, though. The causal influence that an agent may exercise on someone else may be physical in nature or psychological or a combination of the two. Finally, the most elusive case of causation is *negative causation.* Definition 1 refers to negative causation in clause 4 as to *all the important negative variants of the preceding clauses.* It is ontologically very difficult, almost paradoxical, to accept the general idea that something that does not exist can cause anything. For reasons of space we can not analyze the subtleties of this fascinating problem here.

In [2] definitions are given for physical and agent causation within the wider structure of CausatiOnt and some analytical material is provided on interpersonal and negative causation, which are both left as research objectives. In this paper we limit the scope of the presentation of CausatiOnt to the knowledge needed for defining physical causation (shown in figure 1). In other words, we present only the knowledge needed for assessing causal relations between events, without considering actions.

Before starting with the detailed presentation of the class hierarchy shown in figure 1, the following general philosophical biases of CausatiOnt with respect to physical causation should be highlighted:

Cognitivism CausatiOnt is based on the assumption that causal relations are neither purely ontological nor purely epistemological. Therefore, the representation of causal knowledge cannot be limited to the ontological elements of causal relations (i.e. the entities). It must be extended to the epistemological elements (i.e. the categories) and to the phenomenological relations between them (i.e. the dimensions). This extension might seem as a non parsimonious scientific practice. But it gives us some room to explain what in causal reasoning pertains to us as observing entities and what pertains to the world as observed entity. Furthermore, by not limiting ourselves to ontology we provide a clear way of distinguishing semantically similar terms (e.g., matter, a category; mass, a dimension; object, an entity). In a similar fashion, we are able to adopt the distinction defined in [7] between causality (a category, representing *general causal principles*) and causation (a reified relation, i.e. an entity, representing *particular causal relations*). All this will further be explained in section 3.2.

Singularism According to singularism, physical causation relates events, i.e. particular changes of the world located in space and time[5] [8].

Functionalism Functionalism [9], [10], [11] may be seen as the continuation of singularism by other means. The main difference from singularism is that functionalism seeks sharper tools than the notion of change for detecting

[5] Ducasse would for instance say that the cause of the particular change E_2 is E_1 if E_1 alone occurred in the immediate environment of E_2 immediately before. This, of course, begs the question - what is the definition of ‘immediate environment’?

physical causation. The various functionalist views proposed so far try to reduce the notion of causation to physical notions, such as energy or momentum transfer between physical processes, in accordance to contemporary Physics[6].

Formalism According to CausatiOnt, like according to most treatments of causal relations, physical causation has the formal properties of transitivity, asymmetry and non reflexivity.

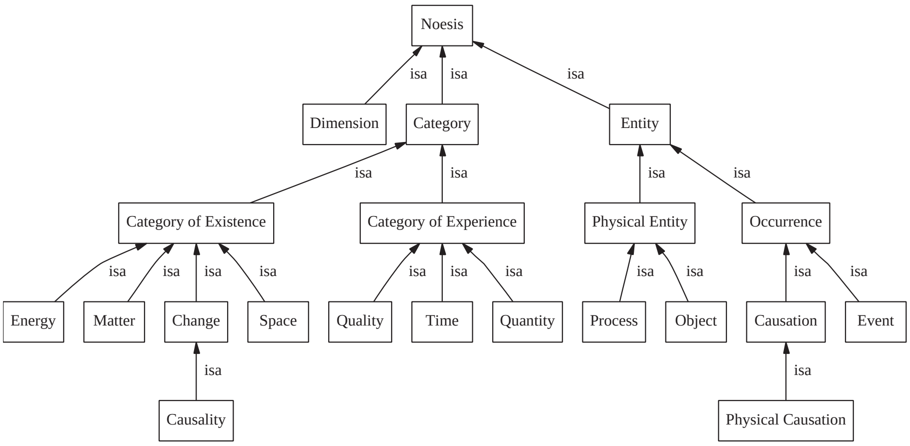

Fig. 1. General hierarchy of CausatiOnt

3.2 CausatiOnt

We present here the class hierarchy shown, at different levels of detail, in figures 1 and 2. This hierarchy is an image of a preliminary specification of CausatiOnt in Protégé-2000, a fairly liberal knowledge representation tool, based on the classical is-a relation. Protégé-2000's liberalism includes the possibility of distinguishing among the following data types in an ontology. *Class*, i.e. a set of (prototypical) individuals (so called instances). A class has a name, that uniquely identifies it and, possibly, a number of slots that intensionally describe it; it is related by is-a relations to its subclasses and by i-o (instance of) relations to its instances. *Slot*, i.e. a (user defined) binary relation between the instances of a class and the instances of another class, or a literal (symbolic or numeric). *System*

[6] For instance, a functionalist would consider a relation between E_1 and E_2 as causal, if the actual physical intersection between E_1 and E_2 involves exchange of a conserved quantity (e.g. energy). Such exchange may be seen as a criterion for further specification of the 'immediate environment' used by singularists

class, i.e. a class that has classes as its instances (i.e. a metaclass). The creation of system classes is usually used in order to expand Protégé-2000's knowledge model because classes and slots are all instances of system class. *Constraint*, i.e. an assertion that restricts the domain and the range of slots.

Protégé-2000 variegated data types allow to represent knowledge that pertains to, at least, three logical orders (instances, classes, system classes). Such specifications may then be subject to further specification in order to fully express them at the first order. In the rest of this section we provide exactly the first liberal specification of CausatiOnt. For each introduced notion we provide a synthetic natural language definition, some comments and the indication of how the notion is implemented in Protégé-2000. Next section provides indications of how CausatiOnt has been imported into DOLCE, in order to axiomatize it in a semantically well founded model.

Definition 2 (Noesis). *Noesis is the psychological counterpart of experience (i.e. perception, learning and reasoning).*

The notion of noesis has a rather long philosophical tradition, which dates back to Greek Philosophy. As far as we are concerned, we adopt here the notion of noesis in its broadest cognitive sense. We consider all the experiences of an individual human being to be physical phenomena. On the one hand, perceptual experiences (e.g. perceiving the form of the canteen) are the result of the interaction between the physical world (i.e. light) and an individual's sensory system (e.g. his optic nerve and other parts of his brain). On the other hand, intellectual experiences (e.g. thinking about the notion of form) occur in the brain, i.e. they too are physical phenomena. Besides their physical nature, though, both perceptual and intellectual experiences generally seem to have a psychological counterpart, i.e. a part of which the individual is aware (i.e. the form of the canteen, in the example of perceptual experiences, and the notion of form, in the example of intellectual experiences). Any such psychological counterpart of an experience is noesis. Noesis is represented in Protégé-2000 as a standard class, with no slots.

Definition 3 (Category). *Category is knowledge-related (i.e. epistemological) noesis.*

A category is a kind of noesis, which cannot be (philosophically) reduced to any other kinds. It must therefore be postulated. Categories form the intellectual *background* of our noetic experience of the world (i.e. of our perception, learning and reasoning about the world). Even though categories play a crucial role in noesis, we are hardly aware of them in our experience. When perceiving, learning or reasoning we are not fully aware of the categories that are supporting our effort. For instance, when reasoning about (i.e. having an intellectual experience of) or perceiving (i.e. having a physical experience of) an entity (e.g. an object, say, the bullet or the canteen), a number of categories (e.g. matter and quantity) make our experience possible, even though they are not immediately present to our mind and/or to our sensory system. Categories are, therefore, here understood as in (Kantian) Epistemology: as the basic notions on which our (intellectual and perceptual) experience builds up[7]. Our intent is to use categories as purely descriptive notions that clarify the intuitive meaning of the terms that are used in reasoning about entities (which we call the dimensions, see below). As shown in figure 1 we distinguish between two main groups of categories: the categories of existence and the categories of experience. The opposition between these two types of categories is the epistemological equivalent of the opposition, within noesis, between entity (or Ontology) and category (or Epistemology). In other words, just like in noesis, where we distinguish existence (the entity) from knowledge (the category), in category we distinguish between the knowledge of what exists (category of existence) from the knowledge of the modes of knowledge (category of experience). These second categories describe how we know what exists (or, rather, how we know the categories of existence). Categories of existence encompass notions such as space, matter, energy, change, causality; whereas category of experience encompass notions such as quantity, quality and time[8]. Categories are all represented in Protégé-2000 as subclasses of noesis, with no slots.

Two categories of existence that deserve some attention here are change and causality. On the one hand, we postulate change as a separate category from time following the philosophical position [13] according to which change must be assumed as distinct from time in order for objects to keep their identity through the occurrence of events (i.e. temporal individuals) that change them. Furthermore, following [7] we propose to distinguish causality from causation and to see the former as a kind of change. In other words we propose to see causality as an *ur*-element of our knowledge of what exists: causality is a piece of our knowledge of *how what exists can change.* For instance, in Example 1 there is a causality relation between, on the one hand, the shooting of the bullet or the poisoning of the canteen (possible causes) and, on the other hand, the death of the traveler (possible effect). But there is a relation of causation only between the shooting of the bullet (actual cause) and the death of the traveler (actual effect). We therefore propose to see causality as the epistemological counterpart of an ontological dependence. In other words, the build up of experience by means of causality requires the concurrent presence of certain categories of existence. For instance, we propose here to adopt the following ontological dependence between categories of existence as the standard notion of causality: energy cannot exist without matter, matter cannot exist without space.

Definition 4 (Dimension). *Dimension is experience-related (i.e. phenomenological) noesis. A dimension relates two categories.*

[7] We want to avoid to use here the expression *a priori* for describing the status of categories. As a matter of fact, under a noetical perspective nothing is *a priori* and one may see categories as the result of evolution, both of individuals and of species.

[8] The main philosophical rationale behind having time as a category of experience is the idea that when we talk about time we do not connote an entity or a natural dimension that exists with independence of what we are as (human) observers. The foundation of the notion of time rests on the biology of the observer [12].

The cognitive build up provided by the categories allows dimensions to emerge. The standard example of a dimension is mass. By experience, all physical objects have a mass, which is the quantity of matter they comprise. We never have, though, a concrete experience of either matter or quantity as such. Therefore, we must assume their existence as categories, rather than as entities, and employ them in the definition of the notion of mass. In other words, the concrete notion of mass relates the epistemological to the ontological part of our noetic experience. We experience objects (ontology) as having mass (phenomenology), which relates two categories: matter and quantity (epistemology). In the definitions of dimensions, we associate categories to one another with the expression 'experienced by means of'. This is to underline the fact that the definition of dimensions in terms of categories is not an ontological but a phenomenological definition. We therefore say, for instance, that mass is matter *experienced by means of* quantity (rather than mass *is* a quantity of matter), where the experience of matter by means of quantity is a purely *intellectual* one, as both matter and quantity are categories, not entities. Furthermore, it should be noticed that we use the expression 'experienced by means of' also in the definition of entities in terms of dimensions. In this case, the expression 'experienced by means of' refers to the *perceptual* (rather than the intellectual) experience of an entity (e.g. an object) through a dimension (e.g. mass).

The following dimensions have been defined: volume (i.e., space experienced by means of quantity), form (i.e. space experienced by means of quality), location (i.e., space experienced by means of time); mass (i.e., matter experienced by means of quantity), material (i.e., matter experienced by means of quality), state (i.e., matter experienced by means of time); work (i.e., energy experienced by means of quantity), energy-form (i.e., energy experienced by means of quality), power (i.e., energy experienced by means of time); direction (change experienced by means of quantity), transition (change experienced by means of quality), period (change experienced by means of time).

All dimensions are represented in Protégé-2000 as instances of the class dimension. This, in turn, is a subclass both of noesis and of standard slot, which is a type of system class. In other words, the instances of the class dimension are particular kinds of slots, which by definition associate a category of existence with a category of experience.

Definition 5 (Entity). *Entity is existence-related (i.e. ontological) noesis.*

The notion of entity indicates something that exists separately from other things and has a clear identity. In Example 1 everything is an entity. Entity is represented in Protégé-2000 as a subclass of noesis with no slots.

Definition 6 (Physical entity). *Physical entity is an entity experienced by means of* one or more *of the following dimensions: volume, form, location, mass, material, state, work, energy-form, power, direction, transition, period.*

Physical entity is represented in Protégé-2000 as a subclass of entity with no slots.

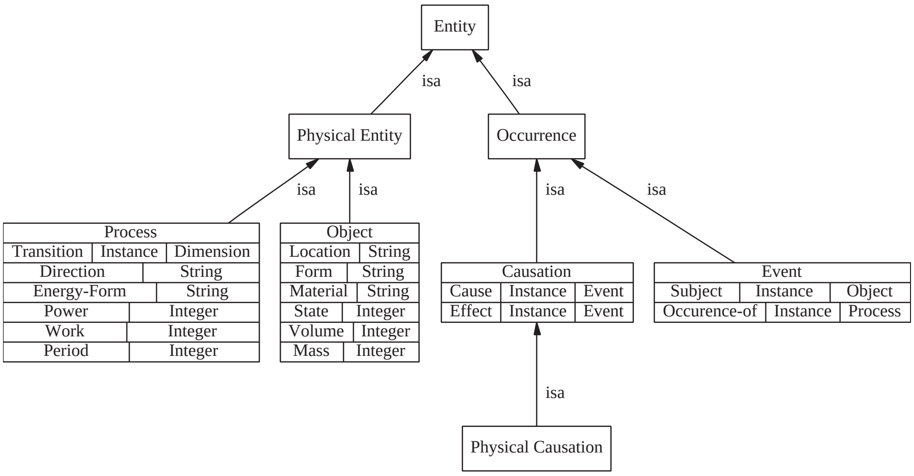

Fig. 2. Entities in CausatiOnt

Definition 7 (Object). *Object is a physical entity which is experienced by means of* all *of the following dimensions: volume, form, location, mass, material, state.*

In Example 1 objects are the bullet and the canteen. Object is represented in Protégé-2000 as a subclass of entity with slots (its dimensions).

Definition 8 (Process). *Process is a physical entity experienced by means of* all *of the following dimensions: work, energy-form, power, direction, transition, period.*

In Example 1 being shot and being broken are processes. Process is represented in Protégé-2000 as a subclass of entity with slots (its dimensions).

Definition 9 (Occurrence). *Occurrence is a* reified relation *between objects, processes and/or occurrences.*

Occurrence is represented in Protégé-2000 as a subclass of entity with no slots.

Definition 10 (Event). *Event is an occurrence of a process (the occurrence of) which changes the value of a dimension of an object (the subject).*

In Example 1 an example of event is the trigger being pulled.
Finally, the notion of causation may be defined.

Definition 11 (Causation). *Causation is an occurrence of two events, the cause and the effect.*

Definition 11 is the counterpart within CausatiOnt of definition 1. It is very broad and it is needed as a definitional node in the ontology. In other words, all the clauses that provide the sufficient conditions for more restrictive (and therefore more interesting) causal relations are provided in the definitions subsumed by Definition 11. This does not mean that the relation introduced in Definition 11 is indistinguishable from simple sequencing of events. Definition 11 introduces a *type of occurrence*. This has, of course, a rather strong implication: *by definition* all *reified* relations between events are causal relations.

Definition 12 (Physical causation). *Physical causation is causation between an event E_1, which is an occurrence of a physical process P_1 (the occurrence of) involving an object O_1 (the subject), and event E_2, which is an occurrence of a physical process P_2 (the occurrence of) involving an object O_2 (the subject). A relation of physical causation holds between E_1, the cause, and E_2, the effect, if the following conditions are met:*

1. *O_1 and O_2 are not the same object, according to the adopted identity criterion for objects.*
   Comment: the subjects must be truly distinguished objects.
2. *P_1 and P_2 are not the same process, according to the adopted identity criterion for processes.*
   Comment: an event cannot cause itself. By this clause we adopt the view that causation is a non reflexive relation.
3. *P_1's period precedes P_2's period.*
   Comment: the cause temporally precedes the effect. Even for processes that are temporally distributed (i.e. continuous) the causing process starts before the caused one. By this clause we adopt the view that causation is a temporally asymmetric relation.
4. *P_1's energy-form is the same as P_2's energy-form or E_2 is reducible to events $E_{2,1} \ldots E_{2,n}$ such that:*
   (a) *$E_{2,1} \ldots E_{2,n}$ are occurrences of processes $P_{2,1} \ldots P_{2,n}$, which all have the same energy form of P_1.*
   (b) *$E_{2,1} \ldots E_{2,n}$ have as their subjects objects $O_{2,1} \ldots O_{2,n}$, which are the grains of O_2, according to the adopted structural constraints.*
   Comment: in the interaction between two objects energy is transferred or transformed. In this latter case, the transformation of energy should be reducible to a transfer of energy between the cause and the events occurring to the structural components of the object of the effect (its grains according to a chosen granularity).
5. *P_1's direction is the same as P_2's direction or P_1's power is greater or equal to P_2's power or P_1's work is greater or equal to P_2's work.*
   Comment: this clause accounts for the fact that usually changes of one sign cause changes of the same sign (i.e. an increase can usually only be caused by an increase and a decrease by a decrease). If this condition cannot be tested (which might be the case when lack of information makes it impossible to establish the directions of either P_1 or P_2) or if it is not satisfied, one may want to use the principle of the dispersion of energy in order to distinguish the cause from the effect.

6. *The category of existence of P_2's transition can not exist without the category of existence of P_1's transition, according to the adopted causality constraint.* Comment: changes in O_1's dimensions can only affect those dimensions of O_2 that are ontologically dependent on the dimensions changed in O_1, according to the adopted causality constraint between categories of existence.

It should be added that we take physical causation to be a *transitive* relation. Definition 12 is represented in Protégé-2000 as a subclass of causation with slots. The conditions listed in the definition should be implemented as a series of constraints.
The information given on E_1 and E_2 so far may be used by the reader for an intuitive testing of clauses 1, 2, 3, 6 of definition 12. Clauses 4 and 5 are more difficult to test, not only for what concerns the information given here on E_1 and E_2, but in general for any two couples of non repeatable events. In conclusion, the most important characteristic of definition 12 is its use of a controlled vocabulary, which defines terms that pertain to three distinct philosophical levels: epistemology, phenomenology and ontology. Such modularity makes it possible to define causation by means of several types of traditionally distinct criteria employed within the same one definition: formalism (clauses 1, 2), singularism and functionalism (clauses 3, 4, 5), cognitivism (clause 6).

4 Preliminary axiomatization of CausatiOnt in DOLCE

DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering) is an ontology of particulars, as shown in the top class of Figure 3. DOLCE is based on a fundamental distinction between four types of entities: Endurants, Perdurants, Qualities and Abstract entities. *Endurants* are wholly present (i.e., all their proper parts are present) at any time they are present. Endurants roughly correspond to objects in CausatiOnt. *Perdurants*, on the other hand, just extend in time by accumulating different temporal parts, so that, at any time they are present, they are only partially present, in the sense that some of their proper temporal parts (e.g., their previous or future phases) may be not present. Perdurants roughly correspond to processes in CausatiOnt. DOLCE's third branch is Quality. Qualities can be seen as the basic entities we can perceive or measure: shapes, colors, sizes, sounds, smells, as well as weights, lengths, electrical charges, etc. Qualities may be clustered in quality types. The term 'quality' is often used as a synonymous of 'property', but this is not the case in DOLCE: qualities are particulars, properties are universals. Qualities inhere to entities: every entity (including qualities themselves) comes with certain qualities, which exist as long as the entity exists. DOLCE's qualities are not comparable to CausatiOnt's dimensions, because the latter are not entities. DOLCE distinguishes between a quality (e.g., the capacity of the canteen in Example 1), and its value (e.g., 1 liter). Values are Abstracts, called qualia in DOLCE, and describe the position of an individual quality within a certain conceptual space, called here quality space. Such quality spaces are subsumed by the fourth branch of DOLCE, i.e. abstract entities, and they are called Regions. So when we say that two canteens

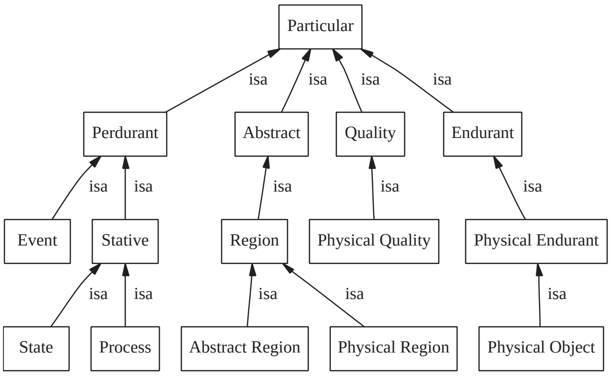

Fig. 3. General hierarchy of DOLCE

have (exactly) the same capacity, in DOLCE we mean that their capacity qualities, which are distinct entities, have the same position in the measure-for-fluids space, that is they have the same capacity quale. This distinction between qualities and qualia is inspired by the so-called trope theory. Its intuitive rationale is mainly due to the fact that natural language - in certain constructs - often seems to make a similar distinction. Each quality type has an associated quality space with a specific structure. For example, lengths are usually associated to a metric linear space, and colors to a topological 2D space etc. For a full specification and formal characterization of DOLCE refer to [4][9]. Our first effort in axiomatizing CausatiOnt in DOLCE[10] has been directed at importing CausatiOnt's epistemological and phenomenological branches into DOLCE. As shown in figure 4 and in the following set of definitions, categories are Abstract regions (definitions 1-10). By (11) we have defined CausantiOnt's relation ExperiencedByMeansOf in terms of DOLCE's relation ExactLocation, which generically locates any type of particular in a region. In (12-13) we have hooked up categories and DOLCE's qualities, by means of DOLCE's relation QLocation, which relates qualities to regions. In (14) we have defined the ontological constraint for causality. Finally in (15) we give an example of how dimensions should be defined in DOLCE as a relation between a particular and a region.

$$Category^{\mathbf{c}}(x) \rightarrow AbstractRegion(x) \tag{1}$$

$$\begin{aligned} &Category^{\mathbf{c}}(x) \equiv \\ &\quad CategoryOfExistence^{\mathbf{c}}(x) \vee \\ &\quad \vee CategoryOfExperience^{\mathbf{c}}(x) \end{aligned} \tag{2}$$

[9] Available on http://wonderweb.semanticweb.org/deliverables/D18.shtml

[10] In order to avoid confusion with DOLCE's original predicates, in the following all the predicates introduced in DOLCE from CausatiOnt are distinguished by the superscript $\mathbf{c}$.

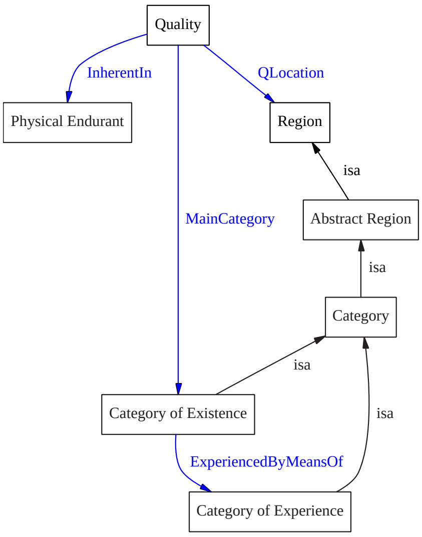

Fig. 4. Import of CausatiOnt into DOLCE

$$CategoryOfExistence^{\mathbf{C}}(space^{\mathbf{C}}) \tag{3}$$

$$CategoryOfExistence^{\mathbf{C}}(matter^{\mathbf{C}}) \tag{4}$$

$$CategoryOfExistence^{\mathbf{C}}(energy^{\mathbf{C}}) \tag{5}$$

$$CategoryOfExistence^{\mathbf{C}}(change^{\mathbf{C}}) \tag{6}$$

$$CategoryOfExperience^{\mathbf{C}}(quantity^{\mathbf{C}}) \tag{7}$$

$$CategoryOfExperience^{\mathbf{C}}(quality^{\mathbf{C}}) \tag{8}$$

$$CategoryOfExperience^{\mathbf{C}}(time^{\mathbf{C}}) \tag{9}$$

$$ExactLocation(change^{\mathbf{C}}, causality^{\mathbf{C}}) \tag{10}$$

$$\begin{aligned} &ExperiencedByMeansOf^{\mathbf{C}}(x,y) =_{def} \\ &\quad CategoryOfExistence^{\mathbf{C}}(x) \wedge CategoryOfExperience^{\mathbf{C}}(y) \wedge \\ &\quad \wedge ExactLocation(x,y) \end{aligned} \tag{11}$$

$$\begin{aligned} &HasCategory^{\mathbf{C}}(x,y) =_{def} \\ &\quad Quality(x) \wedge Category^{\mathbf{C}}(y) \wedge QLocation(x,y) \end{aligned} \tag{12}$$

$$\begin{aligned} &MainCategory^{\mathbf{C}}(x,y,z) =_{def} \\ &\quad Quality(z) \wedge HasCategory^{\mathbf{C}}(z,x) \wedge HasCategory^{\mathbf{C}}(z,y) \wedge \\ &\quad \wedge ExperiencedByMeansOf^{\mathbf{C}}(x,y) \end{aligned} \tag{13}$$

$$\begin{aligned} &CausalityOrder^{\mathbf{C}}(x,y,z,w) =_{def} \\ &\quad Quality(z) \wedge Quality(w) \wedge \\ &\quad \wedge \exists x^{*} MainCategory^{\mathbf{C}}(x,x^{*},z) \wedge \end{aligned} \tag{14}$$

$$\begin{aligned}
&\wedge\exists y^* MainCategory^{\mathbf{c}}(y,y^*,w)\wedge\\
&\wedge(x=space^{\mathbf{c}}\rightarrow(y=space^{\mathbf{c}}\vee y=matter^{\mathbf{c}}\vee y=energy^{\mathbf{c}})\\
&\wedge(x=matter^{\mathbf{c}}\rightarrow(y=matter^{\mathbf{c}}\vee y=energy^{\mathbf{c}})\\
&\wedge(x=energy^{\mathbf{c}}\rightarrow(y=energy^{\mathbf{c}}))
\end{aligned}$$

$$\begin{aligned}
&Volume^{\mathbf{c}}(x,y)=_{def} \qquad (15)\\
&PhysicalEndurant(x)\wedge\exists zInherentIn(z,x)\wedge QLocation(z,y)\wedge\\
&\wedge MainCategory^{\mathbf{c}}(space^{\mathbf{c}},quantity^{\mathbf{c}},z)
\end{aligned}$$

5 Conclusion

Based on axioms (1-15) further research efforts will be directed at defining the relation of causation in DOLCE by means of a representation paradigm called Descriptions and Situations, which extends DOLCE and is now under development. Once this definitional phase is complete, an implementation of the resulting knowledge structure will be attempted. All this is aimed at creating the conceptual basis of a tool for automatic testing, relative to Definition 1, of (legal) models of causation in fact.

References

[1] Lehmann, J.: Causation in Artificial Intelligence and Law - A modelling approach. PhD thesis, University of Amsterdam - Faculty of Law - Department of Computer Science and Law (2003)

[2] Lehmann, J., Breuker, J., Brouwer, B.: Causation in ai&law (to appear). AI and Law (2004)

[3] Gangemi, A., Guarino, N., C., Oltramari, A., Schneider, L.: Sweetening ontologies with dolce. In: Proceedings of EKAW 2002: 166-181. (2002)

[4] Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A.: Wonderweb deliverable d18 - final report. Technical report, National Research Council - Institute of Cognitive Science and Technology (2003)

[5] Pearl, J.: Causality. Cambridge University Press (2000)

[6] Hart, H., Honore, T.: Causation in the Law. Oxford University Press (1985)

[7] Hulswit, M.: A semeiotic account of causation - The cement of the Universe from a Peircean perspective. PhD thesis, Katholieke Universiteit Nijmegen (1998)

[8] Ducasse, C.: On the nature and observability of the causal relation. Journal of Philosophy 23 57-68 (1926)

[9] Russell, B.: Human Knowledge. Simon and Schuster (1948)

[10] Salmon, W.: Scientific Explanation and the Causal Structure of the World. Princeton: Princeton University Press (1984)

[11] Dowe, P.: Causality and conserved quantities: A reply to salmon. Philosophy of Science 62, 321-333 (1995)

[12] Maturana, H.: The nature of time. http://www.inteco.cl/biology/nature.htm (1995)

[13] Lombard, L.: Event - A metaphysical study. Routledge and Kegan Paul (1986)

Inference Systems Derived from Additive Measures

Bassem Sayrafi and Dirk Van Gucht [*]

{bsayrafi,vgucht}@cs.indiana.edu
Computer Science Department,Indiana University,
Bloomington, IN 47405-4101, USA

Abstract. We establish a link between measures and certain types of inference systems and we illustrate this connection on examples that occur in computing applications, especially in the areas of databases and data mining.

1 Introduction

The main contribution of our paper is the establishment of a link between set-based additive measures and certain types of inference systems. To show the applicability of our result, we apply it to particular measures, especially some that occur in the areas of of databases and data mining. Our work significantly generalizes that of Malvestuto [9], Lee [14], and Dalkilic and Robertson [5], where it was shown how Shannon's entropy measure [12] can be used to derive inference systems for functional and multivalued dependencies in relational databases [6].

Our measure framework can be used to find evidence of presence or absence of relationships (possibly causal) [10]. For example, if $\mathcal{M}$ is a measure, and X and Y are sets, then the quantity $\mathcal{M}(X \cup Y) - \mathcal{M}(X)$, i.e., *the rate of change of* $\mathcal{M}$ *in going from* X *to* $X \cup Y$, plays a crucial role in this regard. Depending on its value, this rate can capture interesting relationships. For example, when this rate is 0, it can interpreted as "X fully determines Y according to $\mathcal{M}$", and if it is $\mathcal{M}(Y)$, it can be interpreted as X *and* Y *are independent according to* $\mathcal{M}$.

As a simple, motivating example consider the cardinality measure $|.|$ defined over all subsets of some set S. The cardinality measure has some important properties: for all X, Y, and Z subsets of S, it holds that

$$|X| \leq |X \cup Y| \qquad \textbf{isotonicity}\text{, and}$$
$$|X \cup Y \cup Z| + |X| \leq |X \cup Y| + |X \cup Z| \qquad \textbf{subadditivity}.$$

From these properties follow some others. For example, we can deduce the following "transitivity" property:

$$(|X \cup Y| - |X|) + (|Y \cup Z| - |Y|) \geq (|X \cup Z| - |Z|). \qquad (1)$$

[*] The authors were supported by NSF Grant IIS-0082407.

Given this, we can consider constraints on cardinalities. For example, the constraint $|X| = |X \cup Y|$ states that $X \supseteq Y$. Well-known inference rules for set containment can then be derived from the rules about the cardinality measure. For example, if the constraints $|X \cup Y| = |X|$ and $|Y \cup Z| = |Z|$ are true, then, by the transitivity and the isotonicity rules, $|X \cup Z| = |Z|$. A simpler way of writing this is an inference rule about the set-inclusion relation:

$$\frac{X \supseteq Y \qquad Y \supseteq Z}{X \supseteq Z}$$

The paper is organized into several sections. In Section 2, we introduce additive measures and give examples. In Section 3, we introduce finite differentials for such measures and study the properties of these differentials. In Section 4, we introduce measure constraints and derive inference systems for these constraints from the rules of differentials. We illustrate our approach by deriving some specific inference systems from measures. Finally, in Section 5, we establish a duality between measures and differentials similar to the one that exists between integrals and derivatives in calculus.

2 Additive measures

In this section, we define additive measures. We then give several examples of such measures that occur in practice.

In the rest of the paper, S denotes a finite set, $\mathcal{S}$ denotes 2^S, U, V, X, Y, and Z (possibly subscripted) denote subsets of S, $\mathcal{Y}$ and $\mathcal{Z}$ denote subsets of $\mathcal{S}$ and $\mathcal{M}$ denotes a real-valued function over $\mathcal{S}$. Furthermore, we use the following abbreviations:

$$\begin{aligned} XY &= X \cup Y; \\ X \cdot \mathcal{Y} &= \{XY \mid Y \in \mathcal{Y}\}; \\ \sqcup\mathcal{Y} &= \textstyle\bigcup_{Y \in \mathcal{Y}} Y; \\ \sqcap\mathcal{Y} &= \textstyle\bigcap_{Y \in \mathcal{Y}} Y; \\ \mathcal{Y}[Y \leftarrow Z] &= \mathcal{Y} - \{Y\} \cup \{Z\}. \end{aligned}$$

Our definitions for measures are inspired by the inclusion-exclusion principle for counting finite sets [2]. In light of this, we define the following function $\mathcal{D}$:

Definition 1. *Let f be a function from $\mathcal{S}$ into the reals, let $X \subseteq S$, and let $\mathcal{Y}$ be subset of $\mathcal{S}$. Then the function $\mathcal{D}_f$ at X and $\mathcal{Y}$ is defined as follows:*

$$\mathcal{D}_f(X, \mathcal{Y}) = \sum_{\substack{\mathcal{Z} \subseteq \mathcal{Y} \\ \texttt{odd}(\mathcal{Z})}} f(X \sqcup \mathcal{Z}) - \sum_{\substack{\mathcal{Z} \subseteq \mathcal{Y} \\ \texttt{even}(\mathcal{Z})}} f(X \sqcup \mathcal{Z}). \tag{2}$$

We illustrate this definition in the following table.

X	$\mathcal{Y}$	$\mathcal{D}_f(X,\mathcal{Y})$
X	$\emptyset$	$-f(X)$
X	$\{Y\}$	$f(XY) - f(X)$
X	$\{Y_1, Y_2\}$	$f(XY_1) + f(XY_2) - f(XY_1Y_2) - f(X)$
X	$\{Y_1, Y_2, Y_3\}$	$f(XY_1) + f(XY_2) + f(XY_3) + f(XY_1Y_2Y_3)$ $-f(XY_1Y2) - f(XY_1Y_3) - f(XY_2Y_3) - f(X)$
$Y_1 \cap Y_2$	$\{Y_1, Y_2\}$	$f(Y_1) + f(Y_2) - f(Y_1Y_2) - f(Y_1 \cap Y_2)$

With the use of the function $\mathcal{D}$, we can now define subadditive and superadditive measures.

Definition 2. *Let S be a finite set, let $\mathcal{M}$ be a function from $\mathcal{S}$ into the reals, and let n be a positive natural number. $\mathcal{M}$ is called a n-*subadditive *(n-*superadditive*)* measure *if for each $X \subseteq S$, and each nonempty set $\mathcal{Y}$ of subsets of S, with $|\mathcal{Y}| \leq n$, $\mathcal{D}_{\mathcal{M}}(X,\mathcal{Y}) \geq 0$ ($\mathcal{D}_{\mathcal{M}}(X,\mathcal{Y}) \leq 0$, respectively).*

Example 1. Mathematical measures [4] are n-subadditive for each $n \geq 1$. For such measures, $\mathcal{M}(\emptyset) = 0$. When also $\mathcal{M}(S) = 1$ these measures are called probability measures.

The following proposition, the proof of which is straightforward, relates subadditive measures with superadditive measures. This proposition allows us to focus on subadditive measures.

Proposition 1. *Let $\mathcal{M}$ be a function from $\mathcal{S}$ into the reals and define $\overline{\mathcal{M}}$ (also a function from $\mathcal{S}$ into the reals) as follows:*

$$\overline{\mathcal{M}}(X) = [\mathcal{M}(S) - \mathcal{M}(X)] + \mathcal{M}(\emptyset).\text{[1]} \tag{3}$$

$\mathcal{M}$ is an n-subadditive measure if and only if $\overline{\mathcal{M}}$ is an n-superadditive measure.

2.1 Frequently used measures

In this subsection, we describe a variety of application areas in databases and data mining where measures occur naturally. We identify these measures and fit them in the our measures framework. In the area of databases, we consider aggregate functions and relational data-uniformity measures. In the area of data mining, we focus on measures that occur in the context of the item sets problems.

[1] Notice that $\overline{\mathcal{M}}(S) = \mathcal{M}(\emptyset)$ and $\overline{\mathcal{M}}(\emptyset) = \mathcal{M}(S)$.

Databases - aggregation functions Computations requiring aggregate functions occur frequently in database applications such as query processing, data cubes [8], and spreadsheets. Among these, the most often used are `count`, `sum`, `min`, `max`, `avg`, `variance`, *order statistics*, and `median`. Each of these functions operates on finite sets (`count` on arbitrary finite sets, and the others on finite sets of (nonnegative) numbers) and each returns a nonnegative number. Thus they are measures. We elaborate on how they fit precisely in our framework.

1. Define `count(X)` to be the cardinality of X. From the inclusion-exclusion principle, it follows that `count` is n-subadditive for each $n \geq 1$. (Similar reasoning demonstrates that `sum` is n-subadditive for all $n \geq 1$.)
2. Let S consist of positive integers. Define `max(X)` to be equal to the largest integer in X, for $X \neq \emptyset$, and $\texttt{max}(\emptyset)$ to be equal to the smallest element in S. Then `max` is an n-subadditive measure for $n \geq 1$. The key to showing that `max` is n-subadditive for $n \geq 1$ is the observation that $\texttt{max}(\mathcal{Y}) = maximum(Y)$ for some set $Y \in \mathcal{Y}$. (Similar reasoning demonstrates that `min` is n-superadditive for all $n \geq 1$.)
3. Let S consist of positive integers. Order-statistics are used to determine the i^{th} smallest element of S. For example, the 2^{nd} order statistics, denoted $\texttt{min2}(X)$, returns the second smallest element in X. Clearly, `min2` is 1-superadditive. However, it is not 2-superadditive (e.g. let $Y_1 = \{1, 4, 5\}$, $Y_2 = \{2, 4, 5\}$ and $X = Y_1 \cap Y_2$).
4. The functions `avg`, `variance`, and `median` are neither n-subadditive nor n-superadditive for any $n \geq 1$. However, observe that in the case of `avg` both the numerator and the denominator come from n-subadditive measures (`sum` and `count`, respectively). It follows that the quotient of two subadditive measures is not necessarily a subadditive measure.

Databases - data uniformity Consider the values occurring under an attribute of a relation in a relational database. These values can occur uniformly (e.g. the values 'male' and 'female' in the gender attribute of a census), or skewed (e.g. the values for the profession attribute in the same census). Measuring these degrees of uniformity can influence how data is stored or processed. When data is numeric, a common way to measure uniformity is to use the variance statistic. This statistic computes the average of the distances between data values and their average. To measure data uniformity for categorical data we consider the Simpson measure [13], and the Shannon entropy measure [12]. Unlike variance, these measures are specified in terms of probability distributions defined over the data sets. We show that, unlike variance, the Shannon measure is n-subadditive for $n \leq 2$ and the Simpson measure is n-subadditive for $n \geq 1$.

Let T be a nonempty finite relation over the relation schema S and let p be a probability distribution over T. For $X \subseteq S$, define p_X to be the marginal probability distribution of p on X. Thus if $x \in \Pi_X(T)$ then $p_X(x) = \sum_{\{t \in T | t[X] = x\}} p(t)$.

The *Simpson* measure $\mathcal{S}$ and the *Shannon* measure $\mathcal{H}$ are defined as follows:[2]

$$\mathcal{S}(X) = \sum_{x\in\Pi_X(T)} p_X(x)(1-p_X(x)) = 1 - \sum_{x\in\Pi_X(T)} p_X^2(x), \tag{4}$$

$$\mathcal{H}(X) = -\sum_{x\in\Pi_X(T)} p_X(x)\log p_X(x). \tag{5}$$

It can be shown that the Simpson measure ($\mathcal{S}$) is an n-subadditive measure for all $n \geq 1$ [11]. The Shannon Entropy measure ($\mathcal{H}$) is a 2-subadditive measure but it is not a 3-subadditive measure. Indeed, for the following relation over attributes A, B, C, $\mathcal{D}_{\mathcal{H}}(\emptyset, \{\{A\},\{B\},\{C\}\}) < 0$.

A	B	C
1	1	1
1	1	2
1	2	1
2	1	1

Data mining - frequent item sets

An prominent problem in data mining is discovering frequent item sets. In this problem, a set of baskets is given. Each basket contains a set of items. In practice, the items may be products sold at a grocery store, and baskets correspond to items bought together by customers. The frequent items sets problem is to find the item sets that occur frequently within the baskets.

More formally, let S be a set of items and let $\mathcal{B}$ be a subset of $\mathcal{S}$ consisting of the baskets. Define $\mathcal{B}(X) = \{B \mid X \subseteq B \text{ and } B \in \mathcal{B}\}$ and define the *frequency* measure $\texttt{freq}$ as $\texttt{freq}(X) = \frac{|\mathcal{B}(X)|}{|\mathcal{B}|}$. It can be shown that $\texttt{freq}$ is an n-superadditive measure for $n \geq 1$ [3].

3 Measure Differentials

Some natural issues that arise for measures is (1) to calculate their rate of change and (2) to determine where these rate changes reach optima. Typically, these issues are considered for functions over continuous domains by using traditional calculus techniques, in particular *derivatives.* In our framework for additive measures, we have discrete, set-based functions, and thus reasoning about derivatives must be done with the methods of finite differences and finite difference equations [7].

Definition 3. *Let f be a function from $\mathcal{S}$ into the reals, let X be a subset of S, let $\mathcal{Y}$ be a subset of $\mathcal{S}$, and let Y be in $\mathcal{Y}$. We define the* finite difference *of f at X relative to $\mathcal{Y}$ as follows:*

[2] In ecology, $\mathcal{S}$ is known as the Simpson rarity function.

$$\Delta_f(X,\mathcal{Y}) = f(X) \text{ if } \mathcal{Y} = \emptyset, \tag{6}$$

and

$$\Delta_f(X,\mathcal{Y}) = \Delta_f(XY,\mathcal{Y}-\{Y\}) - \Delta_f(X,\mathcal{Y}-\{Y\}) \text{ otherwise.} \tag{7}$$

Notice that the definition is dependent on the choice for Y in $\mathcal{Y}$. We will show however that each possible choice of Y leads to the same result, i.e., $\Delta_f(X,\mathcal{Y})$ is well defined.

Proposition 2. *Let f be a function from $\mathcal{S}$ into the reals. Then, for each $X \subseteq S$ and for each set $\mathcal{Y} \subseteq \mathcal{S}$, $\Delta_f(X,\mathcal{Y})$ is well-defined.*

Proof. Trivially, $\Delta_f(X,\mathcal{Y})$ is well-defined when $0 \leq |\mathcal{Y}| \leq 1$. When $|\mathcal{Y}| \geq 2$, $\mathcal{Y}$ contains two different sets Y and Y'. We need to show $\Delta_f(XY,\mathcal{Y}-\{Y\}) - \Delta_f(X,\mathcal{Y}-\{Y\}) = \Delta_f(XY',\mathcal{Y}-\{Y'\}) - \Delta_f(X,\mathcal{Y}-\{Y'\})$. We show this by induction on $|\mathcal{Y}|$.

1. When $|\mathcal{Y}| = 2$ this equation becomes
$$\Delta_f(XY,\{Y'\}) - \Delta_f(X,\{Y'\}) = \Delta_f(XY',\{Y\}) - \Delta_f(X,\{Y\}).$$
Further expansion leads to the equation $f(XYY') - f(XY) - f(XY') + f(X) = f(XY'Y) - f(XY') - f(XY) + f(X)$ which is clearly true.
2. When $|\mathcal{Y}| \geq 3$, by induction, we are allowed to expand the left hand side of the equation, i.e., the expression $\Delta_f(XY,\mathcal{Y}-\{Y\}) - \Delta_f(X,\mathcal{Y}-\{Y\})$, into the expression $\Delta_f(XYY',\mathcal{Y}-\{Y,Y'\}) - \Delta_f(XY,\mathcal{Y}-\{Y,Y'\}) - \Delta_f(XY',\mathcal{Y}-\{Y,Y'\}) + \Delta_f(X,\mathcal{Y}-\{Y,Y'\})$. Similarly, the right-hand of the equation can be expanded to expression $\Delta_f(XY'Y,\mathcal{Y}-\{Y',Y\}) - \Delta_f(XY',\mathcal{Y}-\{Y',Y\}) - \Delta_f(XY,\mathcal{Y}-\{Y',Y\}) + \Delta_f(X,\mathcal{Y}-\{Y',Y\})$. Clearly both expressions are equal.

□

It turns out that the functions $\mathcal{D}$ and Δ are closely related:

Proposition 3. *Let f be a function from $\mathcal{S}$ into the reals. Then for each $X \subseteq S$ and for each set $\mathcal{Y} \subseteq \mathcal{S}$*

$$\mathcal{D}_f(X,\mathcal{Y}) = (-1)^{|\mathcal{Y}|-1}\Delta_f(X,\mathcal{Y}). \tag{8}$$

Proof. The proof is by induction on $|\mathcal{Y}|$. For $\mathcal{Y} = \emptyset$, we have $\mathcal{D}_f(X,\emptyset) = -f(X) = -\Delta_f(X,\emptyset)$. For $\mathcal{Y} = \{Y\}$, we have $\mathcal{D}_f(X,\{Y\}) = f(XY) - f(X) = \Delta_f(X,\mathcal{Y})$, and the claim follows.
For $|\mathcal{Y}| \geq 2$, and $Y \in \mathcal{Y}$, we have by the definition of Δ

$$(-1)^{|\mathcal{Y}|-1}\Delta_f(X,\mathcal{Y}) = (-1)(-1)^{|\mathcal{Y}|-2}(\Delta_f(XY,\mathcal{Y}-\{Y\}) - \Delta_f(X,\mathcal{Y}-\{Y\})),$$

which, by induction, is equal to

$$(-1)(\mathcal{D}_f(XY,\mathcal{Y}-\{Y\}) - \mathcal{D}_f(X,\mathcal{Y}-\{Y\})).$$

By the definition of $\mathcal{D}$, we have that $\mathcal{D}_f(X,\mathcal{Y}-\{Y\})-\mathcal{D}_f(XY,\mathcal{Y}-\{Y\})$ is equal to $\sum_{\substack{\mathcal{Z}\subseteq\mathcal{Y}-\{Y\}\\ \texttt{odd}(\mathcal{Z})}} f(X\sqcup\mathcal{Z})-\sum_{\substack{\mathcal{Z}\subseteq\mathcal{Y}-\{Y\}\\ \texttt{even}(\mathcal{Z})}} f(X\sqcup\mathcal{Z})-(\sum_{\substack{\mathcal{Z}\subseteq\mathcal{Y}-\{Y\}\\ \texttt{odd}(\mathcal{Z})}} f(XY\sqcup\mathcal{Z})-\sum_{\substack{\mathcal{Z}\subseteq\mathcal{Y}-\{Y\}\\ \texttt{even}(\mathcal{Z})}} f(XY\sqcup\mathcal{Z}))$ which, after rearranging terms and realizing that $|\mathcal{Z}|$ is even if and only if $|\mathcal{Z}\cup\{Y\}|$ is odd, is equal to

$$\sum_{\substack{\mathcal{Z}\subseteq\mathcal{Y}-\{Y\}\\ \texttt{odd}(\mathcal{Z})}} f(X\sqcup\mathcal{Z})+\sum_{\substack{\mathcal{Z}\subseteq\mathcal{Y}-\{Y\}\\ \texttt{even}(\mathcal{Z})}} f(XY\sqcup\mathcal{Z})$$
$$-\sum_{\substack{\mathcal{Z}\subseteq\mathcal{Y}-\{Y\}\\ \texttt{even}(\mathcal{Z})}} f(X\sqcup\mathcal{Z})-\sum_{\substack{\mathcal{Z}\subseteq\mathcal{Y}-\{Y\}\\ \texttt{odd}(\mathcal{Z})}} f(XY\sqcup\mathcal{Z})$$

This is equal to $\sum_{\substack{\mathcal{Z}\subseteq\mathcal{Y}\\ \texttt{odd}(\mathcal{Z})}} f(X\sqcup\mathcal{Z})-\sum_{\substack{\mathcal{Z}\subseteq\mathcal{Y}\\ \texttt{even}(\mathcal{Z})}} f(X\sqcup\mathcal{Z})=\mathcal{D}_f(X,\mathcal{Y})$. □

In the following proposition we summarize some important properties of $\mathcal{D}$. These properties are specified as equalities and inequalities, but it is more useful here to view them as inference rules.

Proposition 4. *Let $\mathcal{M}$ be an n-subadditive measure ($n\geq 1$). Let $\mathcal{Y}$ be a subset of $\mathcal{S}$. Then $\mathcal{D}_{\mathcal{M}}$ satisfies following properties:*

$$\frac{1\leq|\mathcal{Y}|\leq n}{\mathcal{D}_{\mathcal{M}}(X,\mathcal{Y})\geq 0}\qquad\textbf{sign rule;}$$

$$\frac{Y\in\mathcal{Y}}{\mathcal{D}_{\mathcal{M}}(X,\mathcal{Y}-\{Y\})=\mathcal{D}_{\mathcal{M}}(X,\mathcal{Y})+\mathcal{D}_{\mathcal{M}}(XY,\mathcal{Y}-\{Y\})}\ \textbf{reduction.}$$

When $\mathcal{M}$ is an n-superadditive measure, the reduction rule remains valid. The sign rule however needs to altered by replacing $\mathcal{D}_{\mathcal{M}}(X,\mathcal{Y})\geq 0$ with $\mathcal{D}_{\mathcal{M}}(X,\mathcal{Y})\leq 0$.

Proof. The sign rule follows from the fact we always define $\mathcal{D}_{\mathcal{M}}(X,\mathcal{Y})\geq 0$ for $1\leq|\mathcal{Y}|\leq n$. Reduction follows from (7) and (8). □

Using Proposition 4, we derive interesting rules about measure differentials in the next proposition.

Proposition 5. *Let $\mathcal{M}$ be an n-subadditive measure (when $\mathcal{M}$ is n-superadditive, the inequalities change direction) and $n\geq 1$. Let $\mathcal{Y}$ be a subset of $\mathcal{S}$. Then the rules displayed in Figure 1 follow from Proposition 4.*

$$\frac{Y \in \mathcal{Y}}{\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}[Y \leftarrow YZ]) = \mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) + \mathcal{D}_{\mathcal{M}}(XY, \mathcal{Y}[Y \leftarrow Z])} \qquad \textbf{general chain rule;}$$

$$\frac{Y \in \mathcal{Y} \qquad Y \subseteq X}{\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) = 0} \qquad \textbf{triviality;}$$

$$\frac{U \subseteq X \sqcup \mathcal{Y}}{\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) \geq \mathcal{D}_{\mathcal{M}}(XU, \mathcal{Y})} \qquad \textbf{weak augmentation;}$$

$$\frac{0 \leq |\mathcal{Y}| < n}{\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) \geq \mathcal{D}_{\mathcal{M}}(XU, \mathcal{Y})} \qquad \textbf{augmentation;}$$

$$\frac{X \subseteq Z \qquad |\mathcal{Y}| = 1}{\mathcal{D}_{\mathcal{M}}(X, \{Y\}) + \mathcal{D}_{\mathcal{M}}(Y, \{Z\}) \geq \mathcal{D}_{\mathcal{M}}(X, \{Z\})} \qquad \textbf{weak transitivity;}$$

$$\frac{Y \in \mathcal{Y} \qquad |\mathcal{Y}| < n}{\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) + \mathcal{D}_{\mathcal{M}}(Y, \mathcal{Y}[Y \leftarrow Z]) \geq \mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}[Y \leftarrow Z])} \qquad \textbf{transitivity};$$

$$\frac{Y \in \mathcal{Y} \qquad 2 \leq |\mathcal{Y}| \leq n}{\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y} - \{Y\}) \geq \mathcal{D}_{\mathcal{M}}(X, \mathcal{Y})} \qquad \textbf{replication};$$

$$\frac{Y \in \mathcal{Y} \qquad |\mathcal{Y}| \leq n}{\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) + \mathcal{D}_{\mathcal{M}}(Y, \mathcal{Y} - \{Y\}) \geq \mathcal{D}_{\mathcal{M}}(X, \mathcal{Y} - \{Y\})} \qquad \textbf{coalescence}.$$

Fig. 1. Additional rules for $\mathcal{D}$

4 Measure Constraints

In this section, we consider the situations wherein measure differentials are minimized. In particular, for subadditive (superadditive) measures, we consider when $\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) = 0$ for $0 \leq |\mathcal{Y}| \leq n$. This leads us to introduce *level-n constraints* and to derive inference rules for them. By applying these results to particular measures, we uncover certain classes of constraints in databases and data mining, as well as corresponding inference systems.

Definition 4. Let $\mathcal{M}$ be an n-subadditive (n-superadditive) measure. We call $\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) = 0$ for $0 \leq |\mathcal{Y}| \leq n$ a *level-n constraint* and we say that $\mathcal{M}$ satisfies $X \Rightarrow \mathcal{Y}$ if $\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) = 0$.

It turns out that Definition 4 and Propositions 4 and 5 yield the inference rules for level-n constraints. These rules are a direct consequence of the rules in Propositions 4 and 5 although care must be taken regarding rules when $\mathcal{Y} = \emptyset$.

Proposition 6. *Let $\mathcal{M}$ be an n-subadditive (n-superadditive) measure. Let $\mathcal{Y}$ be a set of subsets of S such that $0 \leq |\mathcal{Y}| \leq n$, and U and Z be subsets of S. Then the level-n constraint of $\mathcal{M}$ satisfies the inequalities in Figure 2.*

$$\frac{Y \in \mathcal{Y} \qquad X \Rightarrow \mathcal{Y}[Y \leftarrow ZY]}{X \Rightarrow \mathcal{Y} \qquad XY \Rightarrow \mathcal{Y}[Y \leftarrow Z]} \quad \textbf{general chain rule (a)};$$

$$\frac{Y \in \mathcal{Y} \qquad X \Rightarrow \mathcal{Y} \qquad XY \Rightarrow \mathcal{Y}[Y \leftarrow Z]}{X \Rightarrow \mathcal{Y}[Y \leftarrow ZY]} \quad \textbf{general chain rule (b)};$$

$$\frac{Y \in \mathcal{Y} \qquad X \Rightarrow \mathcal{Y} \qquad XY \Rightarrow \mathcal{Y} - \{Y\}}{X \Rightarrow \mathcal{Y} - \{Y\}} \quad \textbf{reduction;}$$

$$\frac{Y \in \mathcal{Y} \qquad Y \subseteq X}{X \Rightarrow \mathcal{Y}} \quad \textbf{triviality};$$

$$\frac{U \subseteq X \sqcup \mathcal{Y} \qquad X \Rightarrow \mathcal{Y}}{XU \Rightarrow \mathcal{Y}} \quad \textbf{weak augmentation};$$

$$\frac{1 \leq |\mathcal{Y}| < n \qquad X \Rightarrow \mathcal{Y}}{XU \Rightarrow \mathcal{Y}} \quad \textbf{augmentation};$$

$$\frac{X \subseteq Z \qquad X \Rightarrow \{Y\} \qquad Y \Rightarrow \{Z\}}{X \Rightarrow \{Z\}} \quad \textbf{weak transitivity};$$

$$\frac{Y \in \mathcal{Y} \qquad |\mathcal{Y}| < n \qquad X \Rightarrow \mathcal{Y} \qquad Y \Rightarrow \mathcal{Y}[Y \leftarrow Z]}{X \Rightarrow \mathcal{Y}[Y \leftarrow Z]} \quad \textbf{transitivity};$$

$$\frac{Y \in \mathcal{Y} \qquad 2 \leq |\mathcal{Y}| \leq n \qquad X \Rightarrow \mathcal{Y} - \{Y\}}{X \Rightarrow \mathcal{Y}} \quad \textbf{replication};$$

$$\frac{Y \in \mathcal{Y} \qquad 2 \leq |\mathcal{Y}| \leq n \qquad X \Rightarrow \mathcal{Y} \qquad Y \Rightarrow \mathcal{Y} - \{Y\}}{X \Rightarrow \mathcal{Y} - \{Y\}} \quad \textbf{coalescence}.$$

Fig. 2. Constraint rules for $\mathcal{D}$

Proof. The proof of these rules follows directly from Propositions 4 and5. However, in instances where $\mathcal{Y} = \emptyset$ is possible, care must be taken to deduce these inference rules. We show for example how this applies to the reduction rule and coalescence when $\mathcal{Y} - \{Y\} = \emptyset$. For reduction, we have $\mathcal{D}_{\mathcal{M}}(X, \{Y\}) + \mathcal{D}_{\mathcal{M}}(XY, \emptyset) = \mathcal{D}_{\mathcal{M}}(X, \emptyset)$. Since the left hand side of the equation is zero, then we must have that $\mathcal{D}_{\mathcal{M}}(X, \emptyset) = 0$. The converse is not true however, i.e. $\mathcal{D}_{\mathcal{M}}(X, \emptyset) = 0$ does not imply $\mathcal{D}_{\mathcal{M}}(X, \{Y\}) = 0$ and $\mathcal{D}_{\mathcal{M}}(XY, \emptyset) = 0$, since one of the terms maybe positive and the other negative. Furthermore, coalescence for $|\mathcal{Y}| = 1$ does not hold, that is $\mathcal{D}_{\mathcal{M}}(X, \{Y\}) = 0$ and $\mathcal{D}_{\mathcal{M}}(Y, \emptyset) = 0$ does not imply $\mathcal{D}_{\mathcal{M}}(X, \emptyset) = 0$. Even though $\mathcal{D}_{\mathcal{M}}(X, \{Y\}) + \mathcal{D}_{\mathcal{M}}(Y, \emptyset) = 0 \geq \mathcal{D}_{\mathcal{M}}(X, \emptyset)$, yet $\mathcal{D}_{\mathcal{M}}(X, \emptyset) = -\mathcal{M}(X)$ which can be less than zero. □

4.1 Case studies

It turns out that when we apply Definition 4, and Propositions 4 and 5 to specific measures we uncover useful inference systems that can be used to reason about the relationships between the sets involved. Here we briefly cover the inference systems that can be uncovered when we use the measures `count` for counting sets, the Shannon entropy and the Simpson measure for data uniformity in databases, and finally `freq` in data mining.

1. The level-n constraint for count, $X \Rightarrow \mathcal{Y}$ holds if and only if $\sqcap \mathcal{Y} \subseteq X$ (for $|\mathcal{Y}| \geq 1$) or $X = \emptyset$ (for $\mathcal{Y} = \emptyset$). This is a direct consequence of the inclusion-exclusion principle for counting finite sets. The resulting inference system for count follows directly from Proposition 6 and is shown in Figure 3. The case where $\mathcal{Y} = \emptyset$ deserves special consideration, for example, when $\mathcal{Y} - \{Y\} = \emptyset$, the reduction rule becomes $Y \subseteq X$ and $count(XY) = 0$ imply $count(X) = 0$.
2. For a relation T, the level-1 constraint for the Shannon entropy measure holds if and only if a functional dependency $X \rightarrow Y$ holds in T. This was shown in [9][14][5]. The corresponding inference system rules that can be derived correspond to the well known rules of functional dependencies. Some of the inference system rules are shown in Figure 4.
The level-2 constraint for the Shannon entropy measure holds if and only if a multivalued dependency $X \twoheadrightarrow Y$ holds in T. In this case, $X \twoheadrightarrow Y$ holds if and only if $\mathcal{H}(X \cup Y) + \mathcal{H}(X \cup Z) = \mathcal{H}(X \cup Y \cup Z) + \mathcal{H}(X)$ and $Z = R - Y$ [5]. The corresponding inference system rules that can be derived using our measure framework correspond directly to the well known rules of multivalued dependencies. Some of the inference system rules are shown in Figure 5.
3. For a relation T, the level-1 constraint for the Simpson measure holds if and only if a functional dependency $X \rightarrow Y$ holds in T. The corresponding inference system rules that can be derived correspond to the well known rules of functional dependencies. Some of the inference system rules are shown in Figure 4.
The level-2 constraint for the Simpson measure $X \Rightarrow Y$ holds if and only if a special multivalued dependency $X \twoheadrightarrow Y$ holds for a relation T such

general chain rule (a)

$$\frac{Y \in \mathcal{Y} \qquad \sqcap\mathcal{Y}[Y \leftarrow ZY] \subseteq X}{\sqcap\mathcal{Y} \subseteq X \qquad \sqcap\mathcal{Y}[Y \leftarrow Z] \subseteq XY}$$

triviality

$$\frac{Y \in \mathcal{Y} \qquad Y \subseteq X}{\sqcap\mathcal{Y} \subseteq X}$$

augmentation

$$\frac{1 \leq |\mathcal{Y}| < n \qquad \sqcap\mathcal{Y} \subseteq X}{\sqcap\mathcal{Y} \subseteq XU}$$

transitivity $(Y \in \mathcal{Y}, |\mathcal{Y}| < n)$

$$\frac{\sqcap\mathcal{Y} \subseteq X \qquad \sqcap\mathcal{Y}[Y \leftarrow Z] \subseteq Y}{\sqcap\mathcal{Y}[Y \leftarrow Z] \subseteq X}$$

replication $(Y \in \mathcal{Y})$

$$\frac{2 \leq |\mathcal{Y}| \leq n \qquad \sqcap\mathcal{Y} - \{Y\} \subseteq X}{\sqcap\mathcal{Y} \subseteq X}$$

general chain rule (b)

$$\frac{Y \in \mathcal{Y} \qquad \sqcap\mathcal{Y} \subseteq X \qquad \sqcap\mathcal{Y}[Y \leftarrow Z] \subseteq XY}{\sqcap\mathcal{Y}[Y \leftarrow ZY] \subseteq X}$$

weak transitivity

$$\frac{X \subseteq Z \qquad Y \subseteq X \qquad Z \subseteq Y}{Z \subseteq X}$$

weak augmentation

$$\frac{U \subseteq X \sqcup \mathcal{Y} \qquad \sqcap\mathcal{Y} \subseteq X}{\sqcap\mathcal{Y} \subseteq XU}$$

reduction

$$\frac{Y \in \mathcal{Y} \qquad \sqcap\mathcal{Y} \subseteq X \qquad \sqcap\mathcal{Y} - \{Y\} \subseteq XY}{\sqcap\mathcal{Y} - \{Y\} \subseteq X}$$

coalescence $(2 \leq |\mathcal{Y}| \leq n)$

$$\frac{Y \in \mathcal{Y} \qquad \sqcap\mathcal{Y} \subseteq X \qquad \sqcap\mathcal{Y} - \{Y\} \subseteq Y}{\sqcap\mathcal{Y} - \{Y\} \subseteq X}$$

Fig. 3. Inference system derived for the count measure.

reflexivity

$$\frac{Y \subseteq X}{X \rightarrow Y}$$

augmentation

$$\frac{X \rightarrow Y}{XU \rightarrow Y}$$

transitivity

$$\frac{X \rightarrow Y \qquad Y \rightarrow Z}{X \rightarrow Z}$$

Fig. 4. Inference system for function dependencies derived from the Shannon entropy measure.

reflexivity

$$\frac{Y \subseteq X}{X \twoheadrightarrow Y}$$

replication

$$\frac{X \rightarrow Y}{X \twoheadrightarrow Y}$$

augmentation

$$\frac{X \twoheadrightarrow Y}{XU \twoheadrightarrow Y}$$

coalescence

$$\frac{X \twoheadrightarrow Y \qquad Y \rightarrow Z}{X \rightarrow Z}$$

transitivity

$$\frac{X \twoheadrightarrow Y \qquad Y \twoheadrightarrow Z}{X \twoheadrightarrow Z}$$

Fig. 5. Inference system for multivalued dependencies derived from the Shannon entropy measure.

that $|Y_x| = 1$ or $|Z_x| = 1$ (where $Z = S - XY$, $Y_x = \Pi_Y(\sigma_{X=x}(T))$ and $Z_x = \Pi_Z(\sigma_{X=x}(T)))$. This can be shown by expanding $X \rightarrow Y|Z = 0$ for Simpson's measure, which works out to be

$$S(X \cup Y) + S(X \cup Z) = S(X \cup Y \cup Z) + S(X).$$

The corresponding inference system rules that can be derived using our measure framework correspond directly to the well known rules of multivalued dependencies. Some of the inference system rules are shown in Figure 5.

4. For a family of baskets $\mathcal{B}$, the level-1 constraint of the `freq` measure holds if and only if $\texttt{freq}(X \cup Y) = \texttt{freq}(X)$ if and only if $\mathcal{B}(X \cup Y) = \mathcal{B}(X)$ if and only if there is a pure association rule from X to Y, denoted $X \rightarrow Y$, in $\mathcal{B}$. (A pure association rule is an association rule with confidence 100% [1].) The inference rules of our framework hold for such association rules.

   The level-n constraint for the `freq` measure can be interpreted to yield weaker forms of association rules. For example, $X \Rightarrow \{Y_1, Y_2\}$ holds if and only if $freq(XY_1) + freq(XY_2) = freq(X) + freq(XY_1Y_2)$. To illustrate the use of the inclusion-exclusion principle in this interpretation, referring to Figure 6, we have $freq(X) = |a| + |b| + |c| + |d|$, $freq(XY_1Y_2) = |c|$, $freq(XY_1) = |b|+|c|$, and $freq(XY_2) = |c|+|d|$. Putting everything together, we must have $|a| = 0$. This implies that X can only be bought with Y_1 or Y_2. The inference rules in Figure 2 also hold for these rules. The case where $\mathcal{Y} = \emptyset$ deserves special consideration as it implies $freq(X) = 0$ which implies that all items in X are not bought together. For example, when $\mathcal{Y} - \{Y\} = \emptyset$, the reduction rule becomes $freq(XY) = freq(X)$ and $freq(XY) = 0$ imply $freq(X) = 0$.

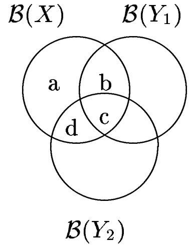

Fig. 6. Frequency constraints example

5 Duality

In this section we will establish a duality between measures and differentials. This duality is similar to the one that exists between derivatives and integrals in calculus:

$$\int_{x}^{x+y} F'(u)du = F(x+y) - F(x).$$

In our setting this duality is captured by the expression

$$\mathcal{D}_{\mathcal{M}}(X, \{X \cup Y\}) = \mathcal{M}(X \cup Y) - \mathcal{M}(X).$$

In other words, one can reasonably think about the expression $\mathcal{D}_{\mathcal{M}}(X, \{X \cup Y\})$ as stating the integration of the function $\mathcal{D}_{\mathcal{M}}$ "from" X "to" $X \cup Y$.

We wish to explore this duality in more depth. To do so, we consider functions satisfying the properties of measure differentials (Proposition 4) and "integrate" them. We can show that the resulting functions are measures and that their measure differentials are the original functions. These results establish that it is possible go back and forth between measures and differentials.

Definition 5. *Let $\mathcal{D}$ be a function from $2^S \times 2^{2^S}$ into the reals and let $n \geq 1$. We call $\mathcal{D}$ an n-differential if it has the following property:*

$$\frac{Y \in \mathcal{Y}}{\mathcal{D}(X, \mathcal{Y} - \{Y\}) = \mathcal{D}(X, \mathcal{Y}) + \mathcal{D}(XY, \mathcal{Y} - \{Y\})}\ \textbf{reduction}$$

We call $\mathcal{D}$ a positive n-differential *if $\mathcal{D}$ is an n-differential and $\mathcal{D}$ satisfies the property:*

$$\frac{X \subseteq S \qquad 1 \leq |\mathcal{Y}| \leq n}{\mathcal{D}(X, \mathcal{Y}) \geq 0}\ \textbf{positive}.$$

We call $\mathcal{D}$ a negative n-differential if $\mathcal{D}$ is an n-differential and $\mathcal{D}$ satisfies the following property:

$$\frac{X \subseteq S \qquad 1 \leq |\mathcal{Y}| \leq n}{\mathcal{D}(X, \mathcal{Y}) \leq 0}\ \textbf{negative}.$$

The following proposition formulates the duality between measures and differentials.

Proposition 7. *Let $\mathcal{D}$ be a n-differential ($n \geq 1$) and let $\mathcal{M}$ be the function from 2^S into the reals defined as follows:*

$$\mathcal{M}(X) = -\mathcal{D}(X, \emptyset). \tag{9}$$

Then for each $X \subseteq S$ and for each nonempty set $\mathcal{Y}$ of subsets of S such that $|\mathcal{Y}| \leq n$

$$\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) = \mathcal{D}(X, \mathcal{Y}). \tag{10}$$

If $\mathcal{D}$ is a positive (negative) n-differential then $\mathcal{M}$ is an n-subadditive (an n-superadditive) measure.

Proof. We prove this by induction on $|\mathcal{Y}|$. For $|\mathcal{Y}| = 0$, $\mathcal{D}_{\mathcal{M}}(X, \emptyset) = -\mathcal{M}(X) = \mathcal{D}(X, \emptyset)$. When $|\mathcal{Y}| \geq 1$, by the properties of $\mathcal{D}_{\mathcal{M}}$ (Proposition 4), $\mathcal{D}_{\mathcal{M}}(X, \mathcal{Y}) = \mathcal{D}_{\mathcal{M}}(X, \mathcal{Y} - \{Y\}) - \mathcal{D}_{\mathcal{M}}(XY, \mathcal{Y} - \{Y\}) = \mathcal{D}(X, \mathcal{Y} - \{Y\}) - \mathcal{D}(XY, \mathcal{Y} - \{Y\})$, by the induction hypothesis. By the reduction rule for $\mathcal{D}$ this is equal to $\mathcal{D}(X, \mathcal{Y})$.

It immediately follows from the definition of measure that when $\mathcal{D}$ is a positive (negative) n-differential, $\mathcal{M}$ is an n-subadditive (an n-superadditive) measure. □

Acknowledgments: We thank Marc Gyssens, Paul Purdom, and Edward Robertson for helpful discussions on topics covered in this paper.

References

1. R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487–499, 1994.
2. R.A. Brualdi. *Introductory Combinatorics (3rd edition).* Prentice-Hall, 1999.
3. T. Calders. *Axiomatization and Deduction Rules for the Frequency of Itemsets.* PhD dissertation- University of Antwerp, 2003.
4. D. Cohn. *Measure Theory.* Birkhäuser-Boston, 1980.
5. M. Dalkilic and E. Robertson. Information dependencies. In *Symposium on Principles of Database Systems*, pages 245–253, 2000.
6. R. Fagin. Multivalued dependencies and a new normal form for relational databases. *ACM Trans. Database Syst.*, 2(3):262–278, 1977.
7. S. Goldberg. *Introduction to Difference Equations.* Dover, 1986.
8. J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, and H. Pirahesh. Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals. *J. Data Mining and Knowledge Discovery*, 1(1):29–53, 1997.
9. F. M. Malvestuto. Statistical treatment of the information content of a database. *Information Systems*, 11:211–223, 1986.
10. L. Mazlack. Imprecise causality in mined rules. In *Proceedings:Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing: 9th International Conference, RSFDGrC*, pages 581–588, 2003.
11. B. Sayrafi and D. Van Gucht. Reasoning about additive measures (full version). *in preparation*, 2004.
12. C. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, July and October 1948.
13. E. Simpson. Measurement of diversity. In *Nature*, volume 163, page 688, 1949.
14. T. Lee. An Information-Theoretic Analysis of Relational Databases – Part I: Data Dependencies and Information Metric. *IEEE Transactions on Software Engineering*, SE-13:1049–1061, 1987.

From Temporal Rules to One Dimensional Rules

Kamran Karimi and Howard J. Hamilton

Department of Computer Science
University of Regina
Regina, Saskatchewan
CANADA S4S 0A2
{karimi, hamilton}@cs.uregina.ca

Abstract. In this paper we propose a new algorithm, called 1DIMERS (One Dimensional Investigation Method for Enregistered Record Sequences), to mine rules in any data of sequential nature, temporal or spatial. We assume that each record in the sequence is at the same temporal or spatial distance from others, and we do not constrain the rules to follow any monotonic direction, meaning that the rules can involve condition attributes in previous and next records relative to the decision attribute. Removing the conceptual temporal limitations makes 1DIMERS a generalised form of TIMERS (Temporal Investigation Method for Enregistered Record Sequences). TIMERS merges consequent records together, and then finds causal or acausal relationships among the variables in the merged records. The kind of rules discovered by TIMERS has also been called sequential rules. In general the passage of time is limited to one direction, which has been used in our previous work to distinguish between causality and acausality. Since in principle it is possible to move back and forth along a sequence, with general sequential data we can no longer intuitively speak of causality and acausality based on a direction. As a result, in 1DIMERS we substitute the terms "causality" and "acausality" with "forward-predictive" and "backward-predictive," respectively. 1DIMERS and TIMERS may each be applicable to a different problem depending on the user's choice, and we give examples of each program's applicability. In previous work we have been using C4.5 as the classifier for creating temporal rules. Here we employ CART as well, which has the ability to regress as well as classify, and show that the results are independent of the underlying rule discovery program.

1. Introduction

Given a sequence of records, the problem we are considering is finding rules for predicting the value of a decision attribute that appears in each record. The traditional approach is to look for a relationship among the decision attribute and other attributes within the same record. One example rule would be [If(a = 6) then (*decision* = *false*)]. This method may not produce good results if there is an *inter-record* relationship among the attributes. Bounded by temporal constraints, in such a case we usually expect only the previous records to affect the current decision attribute, but here we investigate the possibility that the decision attribute's value is determined by attributes not only in previous records, but in next records, or both previous and next records. One example rule in this context would be: [If{($a_{\text{current-1}}$ = 2) AND ($b_{\text{current+1}}$ = 2)} then ($\textit{decision}_{\text{current}}$ = *false*)].

The attributes are now qualified with their position *relative* to the position of the decision attribute. In this example, "current-1" could be read as "previous," and "current+1" could be read as "next."

Relying on data records that appear one after the other in a sequence from a single source (so they are related), sets our approach apart from methods such as [16] than do not consider any ordering among the input records. In this paper we use the term "one-dimensional" instead of "sequential" to avoid confusion with common terminology. Though the data we deal with is sequential in nature, it need not obey any temporal order. Also, the attributes in each record in the sequence can represent any number of dimensions, temporal or spatial. "One-dimensional" here refers to the fact that there is a single temporal or spatial ordering among the records. Also, a sequence implies an unbreakable order, while in this paper we present rules that go back or forth (or both) in the sequence to predict the value of a decision attribute. This usage may not be suitable for real-time execution of temporal rules (we can't give a verdict until sometime in the future). We expect them to be of value in cases where predictive power is of more importance, or we are processing stored temporal data where at any given instant the future and past are available. Alternatively we may be processing spatial data, where future and past are simply substituted by notions of nearby locations (neighbourhood), and considered available. Lifting the restriction of following an inherent order in rules opens the door to new methods of analysing data. Other than that, there are hints that in spite of the intuitive appeal of finding patterns and rules in temporal sequences of data such as time series in a fixed temporal direction, in some cases the results may not be useful [6].

Sequential data and sequential rules have been studied before [1, 4, 20]. For example, in [4] the authors provide a genetic algorithm solution to the problem of detecting rules that manoeuvre a plane that is being chased by a missile in a two dimensional space. Discrete attributes such as speed, direction of the missile, turning rate of the plane, etc. are measured during 20 time steps. It is assumed that after 20 steps the missile will stop the chase. The rules discovered in that paper form part of a plan, and the genetic algorithm changes parts of the plan to make them better suitable to solving the problem. Since the aim of the plans is to prevent a hit, the system is developed to produce rules that come up with evasive actions. The rules are then used in a simulator to measure their effectiveness. Time is obviously the sequencing factor in this example. In this paper we provide one example of sequential data that resembles this application in the sense that it consists of 15 measurements made after a failure is detected in a robot. The observations are then used to classify failure types. However in general we are interested in predicting the value of an attribute that is included in each record. "Being hit," or "failure" does not appear in any of the records, while an attribute such as soil temperature can be measured at regular intervals along with other related variables. Other examples in this paper address the problem of predicting the value of such an attribute.

The remainder of the paper is organised as follows. Section 2 provides background for our work and also presents intuitive examples of the concepts used in the paper. Section 3 formally presents the 1DIMERS (One Dimensional Investigation Method for Enregistered Record Sequences) algorithm, which is a generalised version of TIMERS (Temporal Investigation Method for Enregistered Record Sequences). In Section 4 we present the results of experiments with TIMERS, showing its effectiveness in solving temporal problems. The results of running TIMERS on a Robot learning problem, involving fixed number of relevant records ($w = n = 15$), are presented. It is a classification problem, with discrete values for the decision attribute. Other experiments in Section 4 show that TIMERS is effective for solving regression problems, where the decision attribute is continuous, and provide the results of running CART on two datasets. C4.5's results are provided for comparison purposes. 1DIMERS overlaps with TIMERS in its method, and for comparison's sake its results are provided in each case after trying TIMERS. Section 5 looks at another application domain that closely resembles the temporal domain, and that is spatial sequential data, and shows that the same techniques are effective there. TIMERS and 1DIMERS' results are presented and compared. Section 6 concludes the paper.

2. Background

Our previous work focuses on discovering temporal rules [7, 8] that allow us to predict what happens next, given previous observations. A temporal rule is a rule that involves time, i.e. the condition attributes appear at different times than the decision attribute. We divide temporal rules into two possible categories, causal, and acausal. A *causal* relation is one that involves attributes in the past affecting the decision attribute in the future [19]. The past affecting the future is the normal direction of time, and provides our definition of causality with an intuitive sense. We also consider the case where the future observations affect the past. We consider future affecting the past as a sign of *acausality* [15], or *temporal co-occurrence*. In an acausal relationship, the attributes just happen to be observed together over time, while none is causing the other. In this case there may be a hidden cause that has escaped our observation. Other than causality and acausality, the third possibility is that the relationship between the condition attributes and the decision attribute is *instantaneous*, meaning that value of the decision attribute is best determined by the condition attributes at the same time.

We proposed the TIMERS method to detect a causal or acausal relation among temporal sequences of data [11, 13, 14]. TIMERS provides a set of tests and guidelines, for judging the nature of a relationship. It is partly performed by software, and partly by the domain expert who is analysing the data. Following this algorithm, we generate classification rules from the data, using an operation called *flattening*. Flattening merges consecutive records in the normal, forward, direction of time (for the causality test) or the backward direction of time (for the acausality test). The number of records merged is determined by a time window *w*, and represents our guess as to how many records may be involved in an inter-record relationship. The quality of the rules, determined by their training or predictive accuracy, allows us to judge the data as containing a causal or acausal relation. TIMERS performs three tests: One without flattening, to test the instantaneous hypothesis, and two others to determine the temporal characteristics of the data. The order to consider goes from instantaneous, to acausal, to causal. So if the results of an instantaneous test is about the same or better than the other two, then we declare the relationship among the decision and condition attributes to be instantaneous. Otherwise if the results of the acausality test is about the same, or better than the causality test, then we

.

declare the relationship as acausal. Otherwise the relationship is causal. This order implies that when dealing with temporal relationships, the tendency is to declare it as acausal. More explanation is provided in [13], where an algorithm for flattening data in both forward and backward directions is provided.

In a sequential spatial dataset it is reasonable to assume that there may be connections between records at the neighbouring positions, before and after the current record. In this paper we introduce the *sliding position* flattening method which includes forward and backward flattening as special cases. The principle behind the sliding position method is that both previous and next records can be influential in determining the current value of the decision attribute. In a temporal domain this means considering both past and future observations. With any fixed window size *w*, the new flattening algorithm first places the current decision attribute at position one, and uses the next *w*-1 records to predict its value. This corresponds to a *backward flattening* in TIMERS, where future values are used to predict the past. Then the current attribute is set at position 2, and the previous record (position one) and the next *w*-2 records are used for prediction. This case has no correspondence in our previous algorithm in [13]. This movement of the current position continues and at the end it is set to *w*, and the previous *w*-1 records are used for prediction. This corresponds to *forward flattening* in TIMERS.

As an example consider four temporally consecutive records, each with four fields: <1, 2, 4, true>, <2, 3, 5, false>, <6, 7, 8, true>, <5, 2, 3, true>. Suppose we are interested in predicting the value of the last (Boolean) variable. Using a window of size 3, we can merge them as in Table 1. The decision attribute is indicated in **bold** characters. When it comes to the record involving the decision attribute, we do not consider any condition attributes in the same record as the decision [13]. The *Record.value* notation in Table 1 means that we are only including the decision attribute. For example, <R_1, R_2, R_3.**false**> would contain <1, 2, 4, true, 2, 3, 5, true, **false**>, where *false* is the decision attribute in R_3. This is to make sure that minimum amount of data is shared between the original record and the flattened record.

Instantaneous. *w* = 1 (original data)	Forward (Causality). *w* = 3	Backward (Acausality). *w* = 3	Sliding position. *w* = 3
R_1 = <1, 2, 4, **true**>	<R_1, R_2, R_3.**false**>	<R_3, R_2, R_1.**true**>	<R_2, R_3, R_1.**true**>
R_2 = <2, 3, 5, **true**>	<R_2, R_3, R_4.**true**>	<R_4, R_3, R_2.**true**>	<R_1, R_3, R_2.**true**>
R_3 = <6, 7, 8, **false**>			<R_1, R_2, R_3.**false**>
R_4 = <5, 2, 3, **true**>			<R_3, R_4, R_2.**true**>
			<R_2, R_4, R_3.**false**>
			<R_2, R_3, R_4.**true**>

Table 1. Results of flattening using the forward, backward, and sliding position methods

Normal flattening (vs. sliding position flattening) with a window size of *w* reduces the number of records by *w*-1. It is possible that not all the data in a dataset follow each other temporally, but only every *n* records. For example, every two records were generated one after the other, but there is no relationship between the first record and the third record, or any other record. In this case we consider the window size *w* to be the same as *n*, and perform flattening so that every consecutive *n* records are merged into one, and thus the number of flattened records is divided by *n*. Sliding position flattening increases the number of records.

TIMERS and 1DIMERS perform a series of pre-processing operations, notably flattening, then provide the processed data to another software to generate rules or trees and evaluate them. A post-processing phase can then follow, in which the data are presented to the user in a sequentially meaningful way. We have tried C4.5 [17] before, and because of its availability of source code, have been able to integrate it into our TimeSleuth software [9, 12]. TimeSleuth performs all processing before and after running C4.5 and hence partially implements both the TIMERS and 1DIMERS algorithms. Results of comparing TimeSleuth with other causality miners appear in [14]. Being a classifier, to apply data with continuous variables to C4.5, one has to perform discretisation on the decision attribute, which is not always reasonable with continuous data. In this paper we use another package called CART [2] which can classify as well as perform regression. We evaluate our method with CART as the underlying rule-discoverer, and we show that it performs with little basic variation with different rule-discovery programs.

To use CART we had to perform many of the pre- and post-processing operations "by hand," i.e., using tools that were not integrated into CART. We performed flattening with TimeSleuth, and then deleted the current-time attributes using Mcrosoft Excel. The results for CART were not presented in a temporally valid way since CART, like most other data mining and machine learning algorithms, does not consider any order among its input records. So in the output a variable from the future could precede a variable from the past, for example.

3. The 1DIMERS Algorithm

Modern physics has established time and space as a unity, where one is inconceivable without the other. However, time remains an anomaly because unlike the spatial dimensions, it seems that one cannot move back in time, although experiments have shown that at the particle level, this is in fact possible [5]. Discovering temporal associations that predict the future, based on past observations, is possible, and one can conceptually use the same idea for one-dimensional space as well. Our previous work has used the distinction between moving back and forth in time as the basis of distinguishing causality on one side, and acausality (or temporal co-occurrence) on the other. Since we do not consider an explicit representation of time as necessary (time is implicitly present in the order of the records), a measure such as length can be substituted for time.

Consider the problem of drilling a well. The well can be regarded as a one-dimensional entity. As the drill is making its way through the ground, new points are explored and registered. When we stop, we have a series of records that follow each other along the line. While it seems that the data was produced in a certain temporal order, one could argue that if the drilling were started from the opposite side, then we would be encountering the points from the reverse direction of time. It makes perfect sense to analyse the drilling data in any direction of time, with the results being valid in both cases.

.

Of course now it is not possible to talk about cause and effect because what happens to precede something in one direction, will be following it in the opposite direction.

1DIMERS is an evolution of TIMERS. It changes the terminology from a temporal domain to a spatial domain and provides a more general flattening method, as intuitively described in Section 2. In 1DIMERS an instantaneous rule becomes a *punctual* rule (happens on a point instead of at an instant). A causal rule becomes a *forward-predictive* rule. An acausal rule becomes a *backward-predicitve* rule. The intuitive distinction of causal vs. acausal rules does not exist here. "Forward" and "backward" in 1DIMERS simply refer to the original direction of the data. The lack of distinction between the two possible directions of movement on a line side steps some conceptual problems and debates about causality [3]. In 1DIMERS two new categories are added to one-dimensional rules. The first one is called "linearly extended." A relation is called linearly extended when it is not punctual, and there is no strong evidence that either direction (forward or backward) result in a better predictive ability. The second new category is called *bidirectional predictive* and applies to cases where both directions result in similar results. In TIMERS such cases would be labelled acausal in a temporal domain because. The bias towards acausality in TIMERS is because causality is a strong assumption about any relation. In the general one-dimensional domain, where the restrictions and implications of time do not hold, we can employ a more varied terminology. The sliding position flattening operator is presented in Algorithm 1.

```
For ($i$ = 0; $i \leq |\boldsymbol{D}| - w$; $i$++)
{
        flattenedRecord = <>
        for($j$ = 1; j < $pos$, $j$++)              // previous records
            flattenedRecord += $\boldsymbol{D}_{\mathrm{i+j}}$
        for($j = pos + 1$; $j \leq w$, $j$++)          // next records
            flattenedRecord += $\boldsymbol{D}_{\mathrm{I+j}}$
        flattenedRecord += $Field(\boldsymbol{d}, \boldsymbol{D}_{\mathrm{I+}pos})$     // the decision attribute
        output(flattendRecord)
}
```

Algorithm 1. The Sliding position flattening method

Formally, the flattening operator F(***w***, ***pos***, ***D, d***) takes as input a window size ***w***, The position of the decision attribute within the window ***pos***, the input records ***D***, and the decision attribute ***d***, and outputs flattened records according to Figure 1. D_i returns the *i*th record in the input **D**. *Field*() returns a single field in a record, as specified by its first variable. The += operator stands for concatenating the left hand side with the right hand side, with the results going to the right hand side variable. <> denotes an empty record. This flattening algorithm is simpler than the one presented in [13]. The 1DIMERS algorithm is presented in Figure 2 below. *F*() is the flattening operator as defined in Algorithm 1.

.

Input: A sequence of sequentially ordered data records D, minimum and maximum flattening window sizes α and β, where $\alpha \leq \beta$, a minimum accuracy threshold Ac_{th}, tolerance values ε_i, and a decision attribute d. The attribute d can be set to any of the observable attributes in the system, or the algorithm can be tried on all attributes in turn.

Output: A verdict as to whether the relation among the decision attribute and the condition attributes is punctual, forward-predictive, backward-predictive, bidirectional predictive, or linearly extended.

RuleGenerator() is a function that receives input records, generates decision trees, rules, or any other representation for predicting the decision attribute, and returns the training or predictive accuracy of the results.

```
1DIMERS($D, \alpha, \beta, Ac_{th}, \varepsilon$, d)
$ac_p$ = RuleGenerator($D$, $d$);  // punctual accuracy. window size = 1
for ($w = \alpha$ to $\beta$)
        for($pos$ = 1 to $w$)
             $ac_{w,pos}$ = RuleGenerator($F(w, pos, D, d), d$)
        end for
end for

$ac_b = \max(ac_{\alpha,1}, \ldots, ac_{\beta,1})$   // The best value with the decision attribute in the past (backward)
$ac_f = \max(ac_{\alpha,\alpha}, \ldots, ac_{\beta,\beta})$  // The best value with the decision attribute in the future (forward)
$ac_l = \max(ac_{\alpha,\theta}, \ldots, ac_{\beta,\theta})$,  $\forall\, w, \alpha \leq w \leq \beta$, then $1 < \theta < w$  // Best value in-between

// Maybe there is not enough related information, or the variables are random
if ($Ac_{th} >_{\varepsilon_1} \max(ac_p, ac_l, ac_f, ac_f)$ then discard results and stop.

verdict = "for attribute " + $d$ + ", "

if ($ac_p \geq_{\varepsilon_2} \max(ac_l, ac_f, ac_b)$) then verdict += "the relation is punctual"
else if($ac_l >_{\varepsilon_3} \max(ac_f, ac_b)$) then verdict += "the relation is linearly extended"
else if ($ac_f \approx_{\varepsilon_4} ac_b$) then verdict += "the relation in bidirectional-predictive"
else if ($ac_b >_{\varepsilon_5} ac_f$) then verdict += "the relation is backward-predictive"
else verdict += "the relation is forward-predictive"

return verdict.
```

Algorithm 2. The 1DIMERS method

We use the ε subscripts in the comparison operators to allow the domain expert to ignore small differences. We define $a >_\varepsilon b$ as $a > b + \varepsilon$ and $a \approx_\varepsilon b$ as $|a - b| \leq \varepsilon$. The value of ε, a non-negative number, is determined by a domain expert. If the results for different window values are about the same, we suggest using the smallest window size.

4. Experimental Results

This section provides the results of experiments with the TIMERS and 1DIMERS methods. There is a clear temporal element in the datasets used in the experiments. The results confirm that regression and classification both give consistent results. In all experiments we use training accuracy values.

Even though 1DIMERS is more general and includes TIMERS, in some cases its additional analysis methods may not be applicable or needed. As shown in the experiments below, this includes cases where the semantics of the data do not allow a sliding position flattening (robot failure data in section 4.1), and hence 1DIMERS is not applicable. Another example is in strictly temporal datasets where there is a strong causal relationship (artificial robot data 4.2.1), where 1DIMERS's new flattening method would not produce any better results than TIMERS.

4.1 Classifying Robot Failures

In this section we attend to the problem of failure detection in a robot that grabs, moves and puts down objects. Upon encountering a failure, the force and torque values in the *x*, *y*, and *z* axis (a total of 6 values) are recorded 15 times at regular intervals. The whole process takes 315 ms. The results are then used to classify the type of error that occurred. In [18], five strategies have were to create decision rules for solving the problem. The first one uses the 6 sensor values as they are, while in others these values are processed first, and then used in the decision making process. The 5th strategy combines all the data available to other strategies. The observations have been divided into 5 learning problem datasets. LP1 (failure in approach to grasp), LP2 (failure in transfer), LP3 (failure in positioning a part after a transfer), LP4 (failures in approach to ungrasp), and LP5 (failure in motion with part). Here we observed the force and torque values after an error had already occurred. This kind of data must be processed by the standard TIMERS method, as it does not make sense to place the decision attribute (occurrence of failure) within the observed records. Thus a sliding position flattening is impossible.

To obtain the results in Table 2, we used TIMERS to merge every 15 consecutive records into a single one, and used C4.5 to create decision trees. We tried both the forward and the backward directions of time. C4.5 was invoked with default parameters, with the exception of the values marked with a *, where the -g (use gain) option was used to generate the decision tree. In these cases the values obtained by default arguments appear in parenthesis. The window size was fixed at 15. The five strategies covered in [18] are presented as S1 to S5. The best value for each learning problem among the 5 strategies is presented in bold.

Problem	S1	S2	S3	S4	S5	TIMERS (forward)	TIMERS (backward)
LP1	78%	80%	**96%**	85%	89%	97.7%	97.7%
LP2	45%	57%	51%	**68%**	64%	95.7%	95.7%
LP3	49%	75%	**87%**	85%	83%	85.1%* (48.0)	97.9%
LP4	65%	60%	**95%**	77%	83%	100%* (94.9)	100%* (99.1)
LP5	69%	63%	72%	49%	**77%**	90.9%* (89.0)	90.9%* (82.3)

Table 2. Accuracy values for the robot learning problem

We see that TIMERS gives either better or nearly the same accuracy values as the best of the 5 strategies in [18]. It is also more consistent compared to the other 5 strategies in terms of the quality of results. While TIMERS and S1 both use the original values of force and torque, TIMERS performs considerably better without requiring the user to come up with ways to process data. This is a desirable quality because it frees the user from having to guess which processing method should be used in any particular case.

4.2. Evaluation of Regression on Temporal data

In this subsection we compare the effectiveness of our TIMERS and 1DIMERS methods using two different rule discovery approaches, that of C4.5 and CART. We see that the results are consistent in both cases. The data under investigation is temporal, hence using TIMERS with its temporal terminology is more appropriate. We also provide the results of 1DIMERS' sliding position flattening, and compare the results with those of TIMERS.

We will use two temporal datasets. The first one is from an artificial life program called URAL [21], and involves an artificial robot moving through a two-dimensional board. It can move to left, right, up and down. The goal is for us to discover the effects of moving, on the robot's position, expressed by a *x* and *y* pair. The board is 8 × 8 and there are 1000 observed records. This data comes from a controlled environment with no exceptions, and hence the rules are easy to learn. We consider the results of this test as a form of "sanity check" and have been using them as such in our papers. The second dataset is from a weather station in Louisiana. It includes 342 records of air temperature, the soil temperature, humidity, wind speed and direction and solar radiation, gathered hourly.

In the following tables, the values under "classification" represent the percentage of correct classifications done on training data (training accuracy) while the values for "regression" represent the error (mean square error). So higher values for classification are better, while lower values for regression are desired. We did not change the presentation to stay closer to the actual output of the programs we use.

4.2.1 The Artificial Robot

Each record in this dataset contains a *x* and *y* position value, the direction of movement at the time, and also a binary variable indicating the presence or absence of food. We set the decision attribute to be the current value of *x*, and the other three attributes are set as the condition attributes. There is no relationship between the current value of *x*, and the current values of *y*, direction of the movement, or the presence of food, so we predict that an instantaneous test (no flattening, or setting the window size to 1) will give poor results. Intuitively we know that the current value of *x* depends on the previous value of *x*, and the previous direction of movement. This temporal relationship makes us consider the relationship as a causal one. The acausal hypothesis says that you can tell where you were before if you know where you are now. This hypothesis is clearly wrong, as we could have ended at the current position from a different number of previous positions. Hence we do not expect to get good results with our acausality test. Results of using normal flattening are shown in Table 3, where the "Classification" column indicates the percentage of correct classifications, while the "Regression" column (applicable only to CART) shows the mean square error.

Window	Normal Flattening	CART		C4.5
		Classification	Regression	Classification
1	N/A	24.6%	1.687	46.0%
2	Forward	100%	0	**100%**
	Backward	71.7%	0.469	70.6
3	Forward	100%	0	100%
	Backward	74.1%	0.435	71.7%
4	Forward	100%	0	100%
	Backward	76.2%	0.408	72.8%
5	Forward	100%	0	100%
	Backward	79%	0.378	74.5%

Table 3. CART and C4.5's results with the robot data

As shown in Table 3, CART and C4.5 behave similarly when provided with the same data. After the data have been flattened, the difference in results between the two programs diminishes significantly. The conclusion is the same in both cases: value of *x* is in a causal relation with other attributes, because a causality test provides better results than either the instantaneous or acausal tests. More specifically, the previous *x* and direction of movement causally determine the current value of *x*. This trend (100% accuracy for the causal test) is continued with window sizes higher than 5.

Using the sliding position flattening gives the results shown in Table 4, which are consistent with our expectations. With any position bigger than 1, the previous record, containing the relevant information for accurate prediction of current *x* value, is included in the flattened data. C4.5 discovers the correct temporal relation between the current value of *x* and the previous *x* and movement direction, and results are 100% accuracy with sliding positions of 2 or more.

In Tables 3 and 4 we see that there are slight differences between the results obtained with normal flattening in the acausal mode on one hand, and with sliding position flattening when the position is one, on the other hand. As shown in Table 1, the same information is provided to C4.5 in both cases, so one may expect the same results. However, the order of the attributes in the flattened records is different. This different ordering is evident in Table 1, and causes the results to vary.

Window	Position	Accuracy
2	1	70.6%
2	2	**100%**
3	1	71.5%
3	2	100%
3	3	100%
4	1	72.7%
4	2	100%
4	3	100%
4	4	100%
5	1	75.1%
5	2	100%
5	3	100%
5	4	100%
5	5	100%

Table 4. The results of using the sliding position flattening window to predict the value of *x*

4.2.2 The weather data

The subject of experiments in this subsection is a real-world dataset from weather observations in Louisiana [23], and hence interpreting the dependencies and relationships is harder. The main aim, however, is to compare CART and C4.5's results so as to evaluate TIMERS' consistency in giving a verdict based on the quality of the rules. We have set the soil temperature to be the decision attribute. The results obtained with normal flattening are shown in Table 5.

Window	Normal Flattening	CART		C4.5
		Classification	Regression	Classification
1	N/A	47.8%	1.375	27.7%
2	Forward	58.5%	441.48	82.78%
	Backward	60.5%	441.47	75.1%
3	Forward	78.0%	0.41	86.8%
	Backward	79.5%	0.47	**87.1%**
4	Forward	80.6%	0.37	84.4%
	Backward	80.3%	0.45	84.7%
5	Forward	64.0%	3.05	86.7%
	Backward	63.4%	470.65	82.9%

Table 5. CART and C4.5's results with Louisiana weather data.

The relationship between the soil temperature and other variables is not instantaneous, as observed by relatively poor results with a window of 1 (instantaneous test). The accuracy goes up after flattening, implying that there is a temporal relationship at work (the current value of the soil temperature has a close relationship with the previous values of the soil temperature, among others). TIMERS allows the user to use his domain knowledge when labelling a relationship, especially when the results are similar. In this case we decide to declare the relationship as acausal, because the accuracy values in the

.

two directions of time are not much different. With different time window values, CART displayed more variation than C4.5, but the user is still able to make a decision as to the acausal nature of the relationship.

The results of trying the same data with 1DIMERS' sliding position flattening, as implemented with TimeSleuth are shown in Table 6.

Window	Position	Accuracy
2	1	75.1%
2	2	82.7%
3	1	85.3%
3	2	82.4%
3	3	86.8%
4	1	85.3%
4	2	85.9%
4	3	83.2%
4	4	84.4%
5	1	85.0%
5	2	**87.0%**
5	3	85.0%
5	4	83.8%
5	5	86.7%

Table 6. Results of Sliding position flattening on the weather data.

All values in Table 6 are similar. 1DIMERS gives the verdict of "bidirectional predictive" for these results, which is in contrast to TIMERS' verdict of "acausal." TIMERS would lump both bidirectional and backward predictive verdicts under the verdict of acausal. TIMERS has less resolution in its verdicts, but given the temporal nature of the data, we would choose TIMERS' temporal verdict.

5. Spatial Data

For the experiments described in this section, we used data generated while drilling an oil well [22]. It includes observations about the characteristics of the rock being pierced, including the porosity of the rock (its capacity to hold oil) and different resistance values. The records were registered every 0.5 metres, between the depths of 7400 and 8907.5 metres. The decision attribute was set to be the porosity. TIMERS and 1DIMERS are both tried on this dataset.

To produce results with C4.5, we discretised the value of the porosity to 20 different values. In this case, CART was more effective with regression than classification. Classification took a much longer time to finish, and was done mainly for comparison with C4.5. Table 7 shows the results of the normal flattening methods.

.

Window	Normal Flattening	CART		C4.5
		Classification	Regression	Classification
1	N/A	35.9%	0.010	42.0%
2	Forward	40.4%	0.010	**45.3%**
	Backward	40.1%	0.009	44.0%
3	Forward	38.0%	0.009	41.0%
	Backward	38.4%	0.009	42.0%
4	Forward	37.5	0.009	42.8%
	Backward	31.5%	0.009	41.4%
5	Forward	36.9%	0.009	39.9%
	Backward	29.5%	0.009	39.3%

Table 7. CART and C4.5's results on drilling-sample data. TIMERS method

We get very similar results with the previous and next samples (backward and forward). TIMERS declares the relationship between the porosity and the condition variables to be instantaneous because, the instantaneous test gives about the same results as the temporal tests, and TIMERS gives precedence to being instantaneous. Hence the porosity at each point depends on the current values of the other variables at the same point. This result matched our expectations, because many of the fields in the data, including the resistance values, are related to porosity.

We also applied 1DIMERS, as implemented in TimeSleuth, to this data. The results are given in Table 8.

Window	Position	Accuracy
2	1	44.0%
2	2	45.3%
3	1	42.2%
3	2	**49.0%**
3	3	41.0%
4	1	43.1%
4	2	47.6%
4	3	48.2%
4	4	42.8%
5	1	38.9%
5	2	46.9%
5	3	46.4%
5	4	46.8%
5	5	39.9%

Table 8. C4.5's results with the 1DIMERS method.

According to the results in Table 8, and assuming that a 49.0% result is sufficiently better than 44.0%, the relationship among the porosity and the other variables is best described as linearly-extended, with a window size of 3 and a sliding position of 2. In other words, to predict the porosity at a given depth, the two neighbouring values, half a metre above and below, should be used.

Throughout Table 8 we get better results with a window position bigger than 1 and smaller than the window size, implying a definite neighbourhood relationship between porosity and the other attributes.

6. Concluding Remarks

We introduced 1DIMERS as a conceptual evolution of TIMERS for application on any one dimensional data, and gave the example of a dataset containing samples taken at regular intervals from an oil well. The similarities between a spatial line and temporal line made this generalisation intuitive. We also demonstrated that this method can be used with different underlying rule discoverers, and provide consistent results. CART was employed as an alternative to C4.5, and its ability to generate regression trees allowed us to work with datasets that C4.5 could not handle efficiently. TIMERS/1DIMERS can be implemented both at the rule level and at the tree level [10]. TimeSleuth is an example of a software package that implements the TIMERS and 1DIMERS methods at the rule level.

TimeSleuth, the program that partially implements TIMERS/1DIMERS, can be freely downloaded from http://www.cs.uregina.ca/~karimi/downloads.html. Its user interface employs a temporal/causal terminology.

References

1. Antunes, C. and Oliveira, A., Using Context-Free Grammars to Constrain Apriori-based Algorithms for Mining Temporal Association Rules, *Workshop on Temporal Data Mining (KDD2002)*, Edmonton, Canada. July, 2002.
2. Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. *Classification and Regression Trees*. Wadsworth Inc., 1984.
3. Freedman, D. and Humphreys, P., *Are There Algorithms that Discover Causal Structure?*, Technical Report 514, Department of Statistics, University of California at Berkeley, 1998.
4. Grefenstette, J.J., Ramsey, C.L., Schultz, A.C, Learning Sequential Decision Rules Using Simulation Models and Competition, *Machine Learning 5(4)*, 1990, pp. 355-381.
5. Hawking, S.W., *The Universe in a Nutshell*, Bantam Books, 2001.
6. Lin, J, Keogh, E. and Truppel, W., Clustering of streaming time series is meaningless, The eighth ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, San Diego, California, USA, 2003, pp. 56-65.
7. Karimi, K. and Hamilton, H.J., Finding Temporal Relations: Causal Bayesian Networks vs. C4.5, *The Twelfth International Symposium on Methodologies for Intelligent Systems (ISMIS'2000)*, Charlotte, NC, USA, October 2000, pp. 266-273.
8. Karimi, K. and Hamilton, H.J., Learning With C4.5 in a Situation Calculus Domain, *The Twentieth SGES International Conference on Knowledge Based Systems and Applied Artificial Intelligence (ES2000)*, Cambridge, UK, December 2000, pp. 73-85.
9. Karimi, K. and Hamilton, H.J., RFCT: An Association-Based Causality Miner, *The Fifteenth Canadian Conference on Artificial Intelligence (AI'2002)*, Calgary, Alberta, Canada, May 2002, pp. 334-338.
10. Karimi, K. and Hamilton, H.J., Temporal Rules and Temporal Decision Trees: A C4.5 Approach, *Technical Report CS-2001-02*, Department of Computer Science, University of Regina, Regina, Saskatchewan, Canada, December 2001.

.

11. Karimi, K. and Hamilton, H.J., Discovering Temporal Rules from Temporally Ordered Data, *The Third International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2002)*, Manchester, UK, August 2002, pp. 334-338.
12. Karimi, K., and Hamilton, H.J. TimeSleuth: A Tool for Discovering Causal and Temporal Rules, *The 14th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2002)*, Washington DC, November, 2002, pp. 375-380.
13. Karimi, K., and Hamilton, H.J., Distinguishing Causal and Acausal Temporal Relations, *The Seventh Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'2003)*, Seoul, South Korea, April/May 2003.
14. Karimi, K. and Hamilton H.J., Using TimeSleuth for Discovering Temporal/Causal Rules: A Comparison, *The Sixteenth Canadian Artificial Intelligence Conference (AI'2003)*, Halifax, Nova Scotia, Canada, June 2003.
15. Krener, A. J. Acausal Realization Theory, Part I; Linear Deterministic Systems. *SIAM Journal on Control and Optimization*. 1987. Vol 25, No 3, pp. 499-525.
16. Pearl, J., *Causality: Models, Reasoning, and Inference*, Cambridge University Press. 2000.
17. Quinlan, J. R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann, 1993.
18. Seabra Lopes, L. and Camarinha-Matos, L.M. Feature Transformation Strategies for a Robot Learning Problem, *Feature Extraction, Construction and Selection. A Data Mining Perspective*, H. Liu and H. Motoda (edrs.), Kluwer Academic Publishers, 1998.
19. Schwarz, R. J. and Friedland B., *Linear Systems*. McGraw-Hill, New York. 1965.
20 Sætrom, P. and Hetland, M.L., Unsupervised Temporal Rule Mining with Genetic Programming and Specialized Hardware, *International Conference on Machine Learning and Applications* (ICMLA'2003), 2003.
21. http://www.cs.uregina.ca/~karimi/downloads.html/URAL.java
22. http://explorer.ndic.state.nd.us/. Our data came from a sample CD.
23. http://typhoon.bae.lsu.edu/datatabl/current/sugcurrh.html. Contents change with time.

.

Empirical Investigation of Equilibration-Manipulation Commutability

Denver Dash

Intel Research, SC12-303, 3600 Juliette Lane, Santa Clara, CA 95054, USA,
denver.h.dash@intel.com

Abstract. I consider two operators that are used to transform causal models: the *Do* operator for modeling manipulation and the *Equilibration* operator for modeling a system that has achieved equilibrium. I present an experiment which tested whether or not these two operations commute, i.e., whether or not an equilibrated-manipulated model is necessarily equal to the corresponding manipulated-equilibrated model. My results provide evidence that these operators do not commute. I propose that this result has strong implications for causal discovery from equilibrium data.

1 Introduction

In the study of artificial intelligence, an explicit representation of causality creates the potential for developing an agent that can perform extremely sophisticated reasoning tasks. Constructing a causal model provides an agent with a robust means to diagnose symptoms, and to perform prediction given a current observed state of the system. Most importantly, a causal model releases an agent from the need to store a combinatorially large set of pairs $\{action \Rightarrow effect\}$, allowing the result of external manipulation on various system components to be predicted directly from the model using the *Do* operator [Wold, 1954; Goldszmidt and Pearl, 1992]. By accepting the assumption of *causal faithfulness* [Pearl, 1988; Pearl and Verma, 1991; Spirtes *et al.*, 1993], it is possible in principle to recover causal models from data using constraint-based [Spirtes *et al.*, 1993; Verma and Pearl, 1991; Cheng *et al.*, 2002] or Bayesian [Cooper and Herskovits, 1992; Heckerman *et al.*, 1995; Bouckaert, 1995] causal discovery methods. Causal reasoning plus the ability to learn causal models from data could potentially enable an intelligent agent to build and test hypotheses about its environment and could help automate the process of scientific discovery from data. These are topics that sit on the forefront of artificial intelligence research.

It has been shown by Iwasaki and Simon [1994] that, given assumptions about the form of the causal model, the causal relations governing a dynamic system can change as the time-scale of observation of the system is increased. In particular, they introduce the *Equilibration* operator that produces the causal relations of a system in *equilibrium* given the dynamic (non-equilibrium) causal system.

The *Do* operator, $Do(M, \mathbf{U} = \mathbf{u})$, transforms a causal model M to a new causal model M' where a subset of variables $\mathbf{U}$ in M' are fixed to specific values independent of the causes of $\mathbf{U}$. On the other hand, the *Equilibration* operator, $\mathit{Equilibrate}(M, X)$, transforms the model M with a dynamic (time-varying) variable X to a new causal model M' where X is static. This paper considers the relationship between these two operators. In particular I am interested in the following property:

Definition 1 (Equilibration-Manipulation Commutability). *Let $M(\mathbf{V})$ be a causal model over variables $\mathbf{V}$. M satisfies the Equilibration-Manipulation Commutability (EMC) property iff*

$$\mathit{Equilibrate}(Do(M, \mathbf{U} = \mathbf{u}), X) = Do(\mathit{Equilibrate}(M, X), \mathbf{U} = \mathbf{u}),$$

for all $\mathbf{U} \subseteq \mathbf{V}$ and all $X \in \mathbf{V}$.

I use the shorthand EMC to denote Equilibration-Manipulation Commutability.

In this paper, I ask the question (hereafter referred to as *the EMC question*): "Does the EMC property hold for all dynamic causal models?" This question is important for at least the following reason: Very often in practice a causal model is first built from equilibrium relationships, and then causal reasoning is performed on that model. This common approach takes path A in Figure 1.

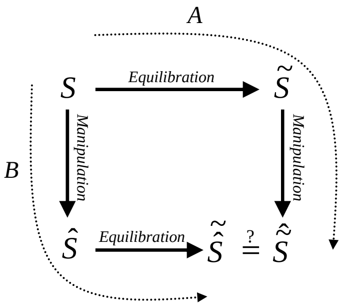

Fig. 1. The EMC Question asks whether or not the *Do* operator commutes with the Equilibration operator operating on a dynamic causal model S.

When a manipulation is performed on a system, however, the state of the system in general becomes "shocked" taking the system out of equilibrium, a situation which is modeled by path B in Figure 1. The validity of the common approach of taking path A thus hinges on the answer to the EMC Question.

The EMC Question has implications for causal discovery from data. A very similar question can be posed in terms of the causal faithfulness condition: "Given a causally faithful dynamic model S, does the new model $\tilde{S}$ resulting from some equilibration of S obey causal faithfulness?" This question can be viewed in terms of Figure 1: if path $S \rightarrow \tilde{S}$ leads to the only graph that is faithful to the equilibrium probability distribution, and if the manipulated equilibrium graph $\hat{\tilde{S}}$ is not equal to the true causal graph defined by $\tilde{\hat{S}}$, then $\tilde{S}$ does not obey the causal faithfulness assumption.

Previously, Dash and Druzdzel [2001] have argued that care must be taken when using equilibrium models for causal reasoning. In this paper, I introduce empirical studies that verify this fact by showing that the EMC question can be answered in the negative.

2 Motivating Example: the Ideal Gas System

Here I briefly restate the example provided in Dash and Druzdzel [2001] showing that the *Do* operator does not commute with the *Equilibration* operator. Consider in Figure 2-(a) the example of an ideal-gas trapped in a chamber with a movable piston, on top of which sits a mass, M. The temperature, T, of the gas is controlled externally by a temperature reservoir placed in contact with the chamber. H is the height of the piston, F_b the total force exerted on the bottom face of the piston, and P is the pressure of gas. In this example, M and T can be controlled directly and so will be exogenous variables. When the values of either M or T are altered, the height of the piston will change: If M is increased then the height will decrease; whereas if T is increased then H will increase. In words the causal ordering can be described as follows: *"In equilibrium, the force applied to the bottom of the piston must equal the weight of the mass on top of the piston. Given the force on the bottom of the piston, the pressure of the gas must be determined, which together with the temperature determines the height of the piston through the ideal-gas law."*

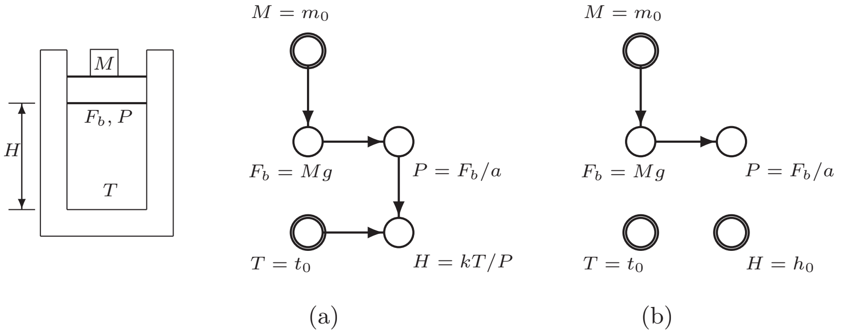

Fig. 2. The causal model of the ideal gas system in equilibrium.

By applying the *Do* operator to Figure 2-(a), one can derive Figure 2-(b) when manipulating the height of the piston to some constant value h_0. Letting I_D denote the underlying dynamic causal model (not shown) for the ideal gas system, then Figure 2-(b) corresponds to the model $Do(\mathit{Equilibrate}(I_D, H), H)$ resulting from manipulating the equilibrium ideal gas model. Next I will derive the model *Equilibrate*$(Do(I_D, H), H)$, resulting from equilibrating the manipulated model. In later paragraphs I will then argue on physical grounds that *Equilibrate*$(Do(I_D, H), H)$ is the model that corresponds to our intuition of this equilibrium manipulated system.

To derive *Equilibrate*$(Do(I_D, H), H)$, I must first derive the dynamic model I_D of the ideal gas system. Imagine dropping a mass M on the piston, simultaneously altering the temperature of the gas, and shortly after measuring the values of all the remaining variables. The physics of this system is comprised of a few fundamental equations: The force on the top of the piston F_t is given by the weight of the mass M:

$$F_t = Mg. \tag{1}$$

The acceleration A of the piston is given by Newton's second law:

$$\Sigma_i F_i = MA. \tag{2}$$

The pressure of the gas P is related to the temperature T and the height of the piston H through the ideal gas law:

$$P = kT/H, \tag{3}$$

where k is a constant. The force on the bottom of the piston is determined by the pressure and the cross-sectional area a of the cylinder:

$$P = F_b/a \tag{4}$$

The height H and the velocity V are determined by recurrence relations (integrals):

$$V_{(t)} = V_{(t-1)} + A_{(t-1)}\Delta t \tag{5}$$

$$H_{(t)} = H_{(t-1)} + V_{(t-1)}\Delta t \tag{6}$$

A shorthand depiction of the causal graph of this system is shown in Figure 3-(a). Since I_D is a dynamic model, it should in principle express a structure at multiple time slices. The graph in Figure 3-(a) represents such a graph: the dashed arcs in this figure denote causation from time slice i to $i+1$, and the solid arcs denote intra-time-slice causation. The dashed arcs were called *integration links* by Iwasaki and Simon [1994].

Consider now fixing the height of the piston using this model to describe the result. To fix the piston in the dynamic model, we must set H to some constant value for all time, $H_{(t)} = h_0$. We also must stop the piston from moving, so we must set $V_{(t)} = 0$ and $A_{(t)} = 0$. Thus, in the dynamic graph with integration links, we can think of this one action of setting the height of the piston as three separate actions. Applying the *Do* operator to these three variables results in the causal graph shown in Figure 3-(b). Since H is being held constant, the graph in Figure 3-(b) is already an equilibrium graph with respect to H, so applying the *Equilibration* operator results in no change to the graph.

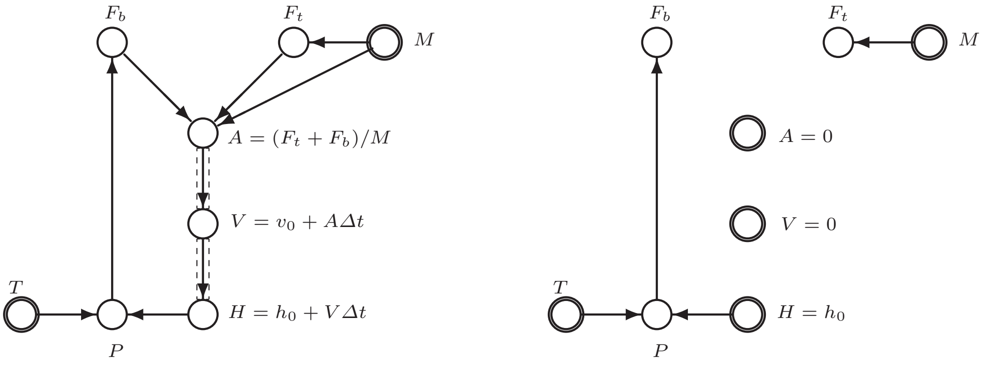

Fig. 3. The graph corresponding to the *Equilibrate*$(Do(I_D, H), H)$ operation on the ideal-gas dynamic model is identical to the intuitive causal graph obtained by manipulating the equilibrium ideal gas system.

Finally, consider the *true* causal graph that results when the height of the piston is set to a constant value: $H = h_0$. Physically this can be achieved by setting the piston to the desired height, and inserting pins into the walls of the chamber, locking it into place. In words, the *true* causal ordering for this system can be described thus: *Since H and T are both fixed, P is determined by the ideal-gas law, $P = kT/H$. Since the gas is the only source of force on the bottom of the piston, F_b is determined by P: $F_b = Pa$.* Thus, P is no longer determined by F_b, and F_b is independent of M. This description is precisely the model *Equilibrate*$(Do(I_D, H), H)$, shown in Figure 3-(b).

3 Discovery from Data: Empirical Results

Section 2 presented an example that implies that the answer to the EMC Question was "no". This section addresses the EMC Question using empirical studies. I performed numerical simulations of some dynamic systems to demonstrate that as the time scale was increased enough so that an equilibration could occur, the causal structure that was learned from data corresponds to the structure obtained by applying the *Equilibration* operator to the dynamic model. This fact is significant because it indicates that whenever a causal structure that is learned from equilibrium data is used for causal reasoning, then Path A of Figure 1 is being taken: if the EMC property does not hold for the model being used then subsequent causal reasoning will produce incorrect results. These experiments provide an empirical answer to the EMC Question because it has been proven [Spirtes *et al.*, 1993] that, in the absence of latent variables, assuming a faithful model to a distribution exists, then the PC algorithm will recover the graph that is faithful to the distribution that generated the data. Furthermore Spirtes *et* *al.* [1993] also argue that the probability of generating a non-faithful model by chance is zero.

In order to simulate and learn the causal structure of the ideal gas system, two minor adjustments to the system were made. First, in order for this dynamic system to achieve equilibrium, there must exist a damping force. In this case, I added a linear damping term: $F_v = -\gamma V$ which is proportional to the negative of the velocity of the piston.

The second adjustment to this system was made due to the fact that the causal discovery algorithm used for this task (the PC algorithm [Spirtes *et al.*, 1993]), uses linear independence tests. The ideal gas law $H = P/T$ involves a non-linear relationship between T and H, and the presence of non-linear associations, together with the assumption of linearity and a large database of records, could allow the significance test to return low p-values if the relation is severely underfit by a straight line. Thus, to avoid artifacts in the learning process due to nonlinear relations in the system, I performed a simulation on the linearized version of the ideal gas system.

This linear system is identical to the original ideal gas system, except the ideal gas law is replaced by the linear relationship $P = -k(H - T - \hat{h})$. Physically, this change corresponds to replacing the ideal gas with a spring whose base can be adjusted with a constant offset T, and where the compression of the spring is given by $\hat{h} - H$ ($\hat{h}$ is the relaxed height of the piston when $M = 0$ and $T = 0$). It appears that the equation for A in the original system is also non-linear because of the inverse dependence on M; however, this relation does not come into play when learning S_1 (because A is not included in the causal model), and the M drops out of the equation in equilibrium, leaving only a linear relationship between the forces in S_2. For this reason I refer to this system as the *pseudo-linear ideal gas system*.

The values of the constants in the ideal gas system were determined by trial-and-error to ensure that the velocity of the piston remained much less than H (to avoid numerically-induced instabilities) and that the height of the piston would never approach zero (which would cause a singularity in the ideal-gas law: $P = T/H$). The values that were used were: $h_0 = 6$, $v_0 = 1$, $m_0 = 6$, and $t_0 = 50$. Each γ_i term was assumed to be a Gaussian random variable with mean zero. It was observed that the ability to correctly recover the expected causal structures depended strongly on the relative noise levels of the variables. To illustrate this fact, I introduce an additional parameter ρ which links the standard deviations (denoted as σ_i) of the noise-terms. The following values were used: $\sigma_H = 0.75$, $\sigma_m = 0.5$, $\sigma_T = 5$, $\sigma_t = 0.5\rho$, $\sigma_a = 0.6\rho$, $\sigma_p = 0.9\rho$, $\sigma_b = 0.9\rho$. Since ρ has a constant value for all records in any given database, it will not violate causal sufficiency for this system. The frictional force was treated as a latent variable (no attempt was made to include it into the learning), and was treated as deterministic for simplicity—its only purpose was to damp out oscillations. The coefficient of friction γ was set to 0.25 to allow lightly damped oscillatory motion of the piston. A few typical equilibrations of the piston are illustrated in Figure 4.

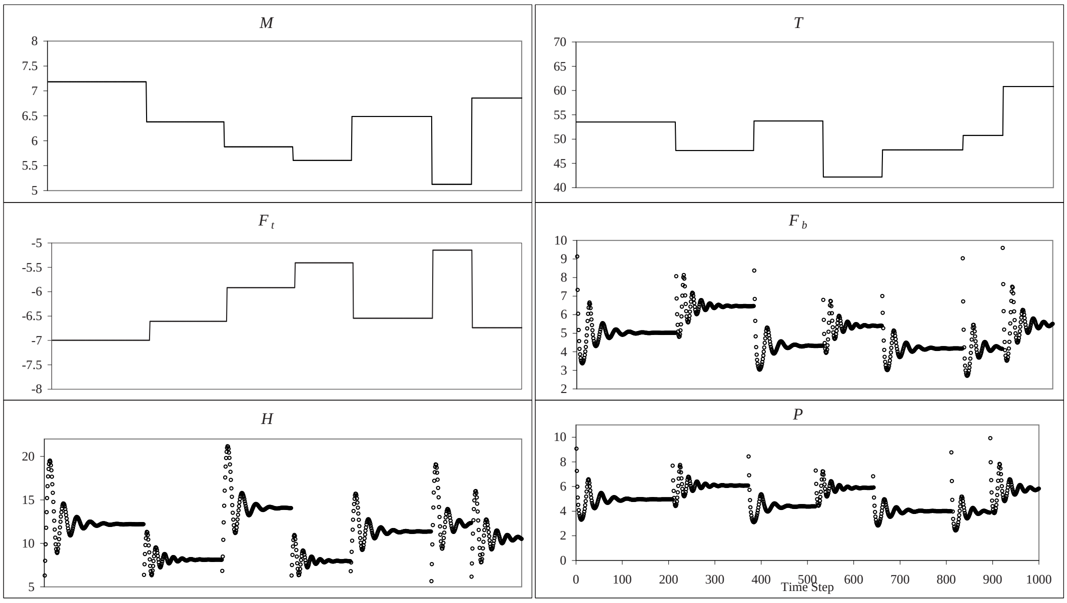

Fig. 4. A few typical equilibrations of the pseudo-linear ideal-gas system.

Distinct runs were generated by repeatedly sampling the noise terms of each variable (i.e., "shocking" the system) and allowing the equation system to guide the evolution of the variables. In order for the system to converge, it was noted that an assumption of stationary noise terms was required. That is, all error terms are sampled once at time step $t = 0$, and thereafter the system was allowed to evolve deterministically until equilibrium, as opposed to sampling the noise terms anew at each time step. This was necessary because randomly shocking the system close to equilibrium will continuously bring it out of equilibrium again.

Each run was allowed to go up to 1000 time steps or until the system was determined to be in an equilibrium state, whichever came first. The system was deemed to be in the equilibrium state if the absolute difference in the change of H from one time step to the next was less than 0.0001. Given the mean value of H: $\langle H \rangle = \langle T \rangle / \langle M \rangle \simeq 10$, this amounts to a change of about 1/1000 of 1 percent. Thus, we can be confident that if the system was stopped prematurely, the values will be nearly identical to the those at time step $t = 1000$.

Using this procedure, two databases D_{dyn} and D_{equ} were generated. Each complete run to equilibrium corresponded to a single record in the databases: a snapshot of the system state at time step $t = 0$ produced a single record for D_{dyn}, and a snapshot at $t = 1000$ defined a record of D_{equ}. This was repeated until two databases of some size N were generated. These two databases were used with the PC algorithm to learn the causal structures observed on short (D_{dyn}) and long (D_{equ}) time-scales. A modified version of PC was used which forbade cycles or bi-directional arrows and randomized the order in which independencies were checked [Dash and Druzdzel, 1999]. Data for each variable took on a continuous range of values, and in all cases the Fisher's-z statistic was used to test for conditional independence using a significance level of $\alpha = 0.05$.

I restricted structure learning to the variables $\{M, T, H, P, F_t, F_b\}$, namely the variables relevant to the static analysis of this system. Over this subset of variables we expect to recover the two structures S_1 and S_2 shown in Figure 5: S_1 when $t = 0$ and S_2 when $t = 1000$. N was systematically varied from the set {100, 500, 1000, 2000, 4000, 10000}, and ρ was varied from the set {0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4}. 100 measurements were taken for each (N, ρ) combination, and the probability P_{hit}, the fraction of times that precisely the correct structure was learned, was calculated. We expected that as N was increased, P_{hit} for both S_1 and S_2 would increase, ideally approaching unity. Figure 6 shows the probability of recovering the correct structure as a function of N, averaged over values of ρ. When the linear equation system is used, the learned graphs converge neatly to S_1 and S_2.

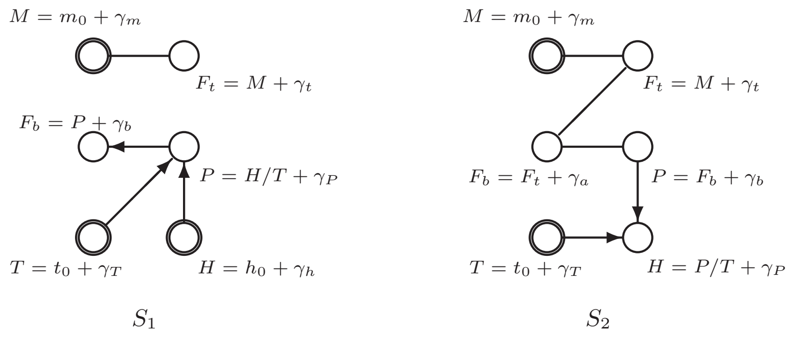

Fig. 5. The two patterns expected to be recovered from the simulation of the ideal-gas system. S_1 is the expected pattern for $t = 0$ (D_{dyn}), and S_2 is the expected pattern for $t = 1000$ (D_{equ}).

The important observation about these simulations is this: If we alter the ideal gas system by setting $A = V = 0$ for all time and setting $H = h_0$, we can simulate the ideal-gas system under the assumption that H is being manipulated to the value h_0. However, this manipulation will produce data from a distribution identical to that of the model S_1, and therefore, we would learn S_1 from the data generated by manipulating H. This of course, is not the same graph that we would get by applying the *Do* operator to S_2, verifying exactly the observations of Section 2.

Considered from the standpoint of causal discovery these results are disheartening. Using data from the equation system of Figure 2 with independent error terms, the causal graph shown there (S_1) would be learned by a constraint-based discovery algorithm such as the PC algorithm. On the other hand, using data

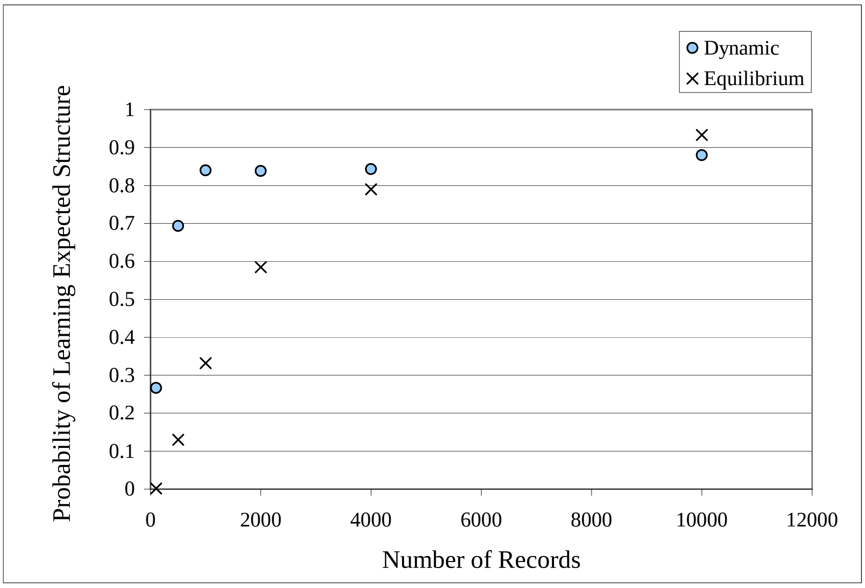

Fig. 6. The probability of learning the expected dynamic (S_1) and equilibrium (S_2) graphs as the number of records increases for the pseudo-linear ideal-gas system, averaged over all values of ρ.

from the equations governing the manipulated system would yield the causal model S_2. The end result is clear: a causal graph learned based on the equilibrium ideal-gas system and altered with the *Do* operator will yield the incorrect causal graph of Figure 2-(b).

4 Conclusions

The main conclusions of this experiment are two-fold: (1) The causal graph recovered from data depends strongly on the time-scale at which the data was generated. (2) The causal graph taken from long-time-scale data will not in general produce the correct distribution when used to predict the effect of manipulations on the system. These conclusions support the assertions presented by Dash and Druzdzel [2001] that equilibrium models do not support causal reasoning.

Complicating this situation is the fact that many systems possess multiple time-scales. In the present case, only one significant time-constant were present. In systems with multiple relevant time-scales, modeling and/or learning causal interactions will be even more difficult. In a single sentence: These results imply that caution is advised when attempting to learn causal models from equilibrium data.

Bibliography

R. Bouckaert. *Bayesian belief networks: From construction to inference.* PhD thesis, University Utrecht, 1995.

Jie Cheng, Russell Greiner, Jonathan Kelly, David Bell, and Weiru Liu. Learning Bayesian networks from data: an information-theory based approach. *Artificial Intelligence*, 137(1):43–90, May 2002.

Gregory F. Cooper and Edward Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, 1992.

Denver H. Dash and Marek J. Druzdzel. A hybrid anytime algorithm for the construction of causal models from sparse data. In *Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI–99)*, pages 142–149, San Francisco, CA, 1999. Morgan Kaufmann Publishers, Inc.

Denver Dash and Marek J. Druzdzel. Caveats for causal reasoning with equilibrium models. In Salem Benferhat and Philippe Besnard, editors, *Proceedings of the Sixth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU 2001)*, volume 2143 of *Lecture Notes in Artificial Intelligence*, pages 192–203, Toulouse, France, 2001. Springer-Verlag.

Moises Goldszmidt and Judea Pearl. Ranked-based systems: A simple approach to belief revision, belief update and reasoning about evidence and actions. In B. Nebel, C. Rich, and W. Swartout, editors, *Proceedings of the 3rd International Conference on Knowledge Representation and Reasoning*, pages 661–672, San Mateo, CA, 1992. Morgan Kaufmann.

David Heckerman, Dan Geiger, and David M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

Yumi Iwasaki and Herbert A. Simon. Causality and model abstraction. *Artificial Intelligence*, 67(1):143–194, May 1994.

Judea Pearl and Thomas S. Verma. A theory of inferred causation. In J.A. Allen, R. Fikes, and E. Sandewall, editors, *KR–91, Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pages 441–452, Cambridge, MA, 1991. Morgan Kaufmann Publishers, Inc., San Mateo, CA.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann Publishers, Inc., San Mateo, CA, 1988.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search.* Springer Verlag, New York, 1993.

T.S. Verma and Judea Pearl. Equivalence and synthesis of causal models. In P.P. Bonissone, M. Henrion, L.N. Kanal, and J.F. Lemmer, editors, *Uncertainty in Artificial Intelligence 6*, pages 255 –269. Elsevier Science Publishing Company, Inc., New York, N. Y., 1991.

Herman Wold. Causality and econometrics. *Econometrica*, 22(2):162–177, April 1954.